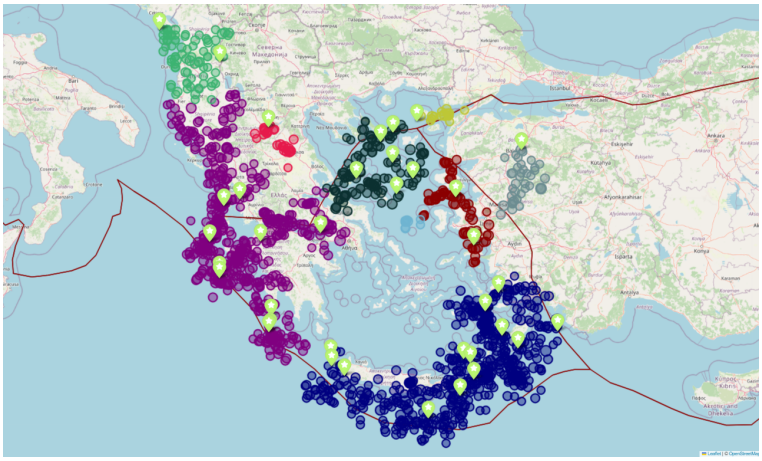
Διεθνές Πανεπιστήμιο Ελλάδος
Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Δεδομένων σεισμών με τεχνικές Εξόρυξης Γνώσης και Ανάλυσης Χρονοσειρών



Φοιτήτρια:
Βασιλική Δέλλιου
Αριθμός Μητρώου: 02/2023

Επιβλέπων:
Στέφανος Ουγιάρογλου

15-06-2025

Ανάλυση Δεδομένων σεισμών με τεχνικές Εξόρυξης Γνώσης και Ανάλυσης Χρονοσειρών
Κωδικός Δ.Ε.: 24316
Ονοματεπώνυμο φοιτητή: Βασιλική Δέλλιου
Ονοματεπώνυμο εισηγητή: Στέφανος Ουγιάρογλου
Ημερομηνία ανάληψης: 07-11-2024
Ημερομηνία περάτωσης: 15-06-2025

Βεβαιώνω ότι είμαι η συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια, την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

«Αφιέρωση»

Στους γονείς μου

Ευχαριστίες

Πρώτα απ' όλα, ένα μεγάλο ευχαριστώ στα αδέρφια μου, Κατερίνα και Βαγγέλη. Η στήριξή τους – συχνά σιωπηλή αλλά πάντοτε ουσιαστική – υπήρξε για μένα θεμέλιο δύναμης και σταθερότητας σε κάθε στιγμή αυτής της διαδρομής.

Ευχαριστώ θερμά την Κατερίνα για την αμέριστη αγάπη της, την υποστήριξη και την ενθάρρυνσή της. Για κάθε φορά που ήταν εκεί, έτοιμη να με ακούσει και να μου υπενθυμίζει πόσο πιστεύει σε μένα.

Ευχαριστώ από καρδιάς τον Βαγγέλη για την αγάπη του, την άνευ ορίων υπομονή και ανεκτικότητά του. Για την συνεχή παρουσία του σε κάθε βήμα της πορείας μου. Η ηρεμία του και οι συμβουλές του έκαναν τις δυσκολίες να φαίνονται ασήμαντες.

Τέλος, θα ήθελα να εκφράσω τις ειλικρινείς ευχαριστίες μου στον επιβλέποντα καθηγητή μου, κ. Στέφανο Ουγιάρογλου για την πολύτιμη καθοδήγησή του, την αδιάκοπη υποστήριξή του. Οι συμβουλές και η παρότρυνσή του να εξερευνήσω νέες ιδέες και προκλήσεις υπήρξαν καθοριστικής σημασίας για την επιτυχή ολοκλήρωση αυτής της εργασίας.

Abstract

This thesis focuses on the analysis of seismic activity in Greece, one of the most seismically active regions in Europe and worldwide. It utilizes the open availability of seismological data, provided by the Geodynamic Institute of the National Observatory of Athens. The research applies state-of-the-art Machine Learning methods, with a particular emphasis on Unsupervised Learning techniques for clustering. Several algorithms were implemented using the Python programming language and compared, including K-Means, DBSCAN, HDBSCAN, and OPTICS. The analysis was conducted using both two-dimensional and five-dimensional data, examining the geographical and physical characteristics of earthquakes. A wide range of visualizations and interactive maps were also created. These maps include tectonic plate boundaries, enhancing the interpretation of the results. The findings reveal the spatio-temporal distribution of earthquakes and confirm the presence of known seismic zones. Notably, density-based algorithms outperformed the partition-based K-Means, as they identified naturally shaped clusters, that closely align with established seismogenic areas. Furthermore, time series forecasting models were applied, with the M5P model emerging as the most effective in estimating future seismic activity. Overall, this study makes a significant contribution to the understanding of earthquake events in Greece and provides a concrete framework for seismological data analysis, which may serve as a foundation for future extensions, such as identifying spatio-temporal patterns and correlating them with geological faults.

Keywords
Seismology, Machine Learning, Unsupervised Learning, Clustering, Time series analysis, Seismological data, K-Means, DBSCAN, HDBSCAN, OPTICS, M5P, WEKA, Tectonic plates, Python

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στη μελέτη της σεισμικής δραστηριότητας στην Ελλάδα, μια από τις πλέον σεισμογενείς περιοχές της Ευρώπης, αλλά και παγκοσμίως. Αξιοποιεί την ανοιχτή διάθεση σεισμολογικών δεδομένων από το Γεωδυναμικό Ινστιτούτο του Εθνικού Αστεροσκοπείου Αθηνών. Η μελέτη εφαρμόζει σύγχρονες μεθόδους Μηχανικής Μάθησης. Έμφαση δίνεται στις τεχνικές Μη Εποπτευόμενης Μάθησης, ειδικά στη συσταδοποίηση. Υλοποιήθηκαν, μέσω της γλώσσας Python και συγκρίθηκαν αλγόριθμοι όπως οι: K-Means, DBSCAN, HDBSCAN και OPTICS. Η ανάλυση έγινε σε δεδομένα δύο αλλά και πέντε διαστάσεων. Εξετάστηκαν γεωγραφικά και φυσικά χαρακτηριστικά των σεισμών, ενώ παράλληλα δημιουργήθηκε μια ποικιλία οπτικοποιήσεων και διαδραστικοί χάρτες. Αυτοί οι χάρτες ενσωματώνουν και τα όρια των τεκτονικών πλακών για καλύτερη ερμηνεία των αποτελεσμάτων. Τα ευρήματα αποκαλύπτουν τη χωροχρονική κατανομή των σεισμών και επιβεβαιώνουν γνωστές σεισμογενείς ζώνες. Ειδικότερα, οι αλγόριθμοι που βασίζονται στην πυκνότητα επέδειξαν την υπεροχή τους, έναντι του διαμεριστικού αλγόριθμου K-Means, καθώς εντόπισαν φυσικά διαμορφωμένες συστάδες, που ταυτίζονται με γνωστές σεισμογενείς ζώνες. Επιπλέον, εφαρμόστηκαν μοντέλα πρόβλεψης χρονοσειρών. Το μοντέλο M5P αναδείχθηκε ως το πιο αποτελεσματικό στην εκτίμηση της μελλοντικής σεισμικής δραστηριότητας. Συνολικά, η εργασία συμβάλλει σημαντικά στην κατανόηση των σεισμικών γεγονότων στην Ελλάδα και παρέχει συγκεκριμένο πλαίσιο ανάλυσης σεισμολογικών δεδομένων, το οποίο μπορεί να αποτελέσει βάση για μελλοντικές επεκτάσεις, όπως ο εντοπισμός χωροχρονικών μοτίβων και η συσχέτισή τους με γεωλογικά ρήγματα.

Λέξεις-Κλειδιά
Σεισμολογία, Μηχανική Μάθηση, Μη Εποπτευόμενη Μάθηση, Συσταδοποίηση, Ανάλυση χρονοσειρών, Σεισμολογικά δεδομένα, K-Means, DBSCAN, HDBSCAN, OPTICS, M5P, WEKA, Τεκτονικές πλάκες, Python

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Σεισμοί

Η Γη είναι ένας ζωντανός πλανήτης με συνεχή, γεωλογική δραστηριότητα [1]. Από την αρχαιότητα έως και σήμερα ο σεισμός αποτελεί ένα φυσικό φαινόμενο, που εξακολουθεί να προκαλεί δέος, ανησυχία και φόβο στον άνθρωπο. Σεισμός καλείται η αιφνίδια και απρόβλεπτη απελευθέρωση συσσωρευμένης ενέργειας στο εσωτερικό της Γης, η οποία μεταδίδεται στην επιφάνεια, μέσω σεισμικών κυμάτων. Το φαινόμενο του σεισμού οφείλεται σε φυσικές, γεωλογικές διεργασίες, που λαμβάνουν χώρα και πιο συγκεκριμένα στη ρήξη πετρωμάτων, ή την κίνηση των τεκτονικών πλακών. Ο πολύπλοκος μηχανισμός γένεσης σεισμικών δονήσεων φέρνει στο προσκήνιο την απρόβλεπτη φύση και τη δυναμική του πλανήτη μας. Έτσι λοιπόν, παρά την τεχνολογική ανάπτυξη, η ανθρωπότητα είναι ευάλωτη και εκτεθειμένη στις ισχυρές, εσωτερικές δυνάμεις της Γης.

Ειδικά στην σημερινή εποχή οι σεισμοί δύνανται να προκαλέσουν μεγάλες καταστροφές, καθώς οι αστικές περιοχές είναι, πλέον, πυκνοκατοικημένες. Εκτός από σοβαρές και εκτεταμένες ζημιές που μπορεί να επέλθουν στον υλικό κόσμο - όπως για παράδειγμα, η κατάρρευση κτιρίων, οι βλάβες σε κρίσιμες υποδομές και η διακοπή δικτύων ηλεκτροδότησης, ύδρευσης και επικοινωνιών - είναι δυστυχώς σύνηθες, να καταγράφονται και ανθρώπινες απώλειες ή τραυματισμοί, έπειτα από μια ισχυρή, επιφανειακή σεισμική δόνηση. Οι κοινωνικές, οικονομικές και ψυχολογικές επιπτώσεις τέτοιων γεγονότων είναι εξίσου σημαντικές, καθώς ο σεισμός δεν πλήττει μόνο το δομημένο περιβάλλον, αλλά και το αίσθημα ασφάλειας των πληγέντων κοινοτήτων.

Παρά την εντατική ερευνητική δραστηριότητα και την πρόοδο της σεισμολογικής επιστήμης, η αξιόπιστη πρόβλεψη καταστροφικών σεισμών παραμένει, έως σήμερα, μη εφικτή. Αυτό είναι ένα γεγονός, που αφορά τόσο τις βραχυπρόθεσμες (λίγες ώρες έως μήνες), όσο και τις μακροπρόθεσμες (έτη) προσπάθειες πρόγνωσης. Στη βραχυχρόνια πρόβλεψη, η πολυπλοκότητα των σεισμικών φαινομένων εμποδίζει την ακριβή πρόβλεψη του χρόνου, του τόπου και του μεγέθους. Στη μακροχρόνια από την άλλη, ενώ υπάρχει η δυνατότητα να εκτιμηθεί η σεισμική επικινδυνότητα μιας περιοχής - με βάση την ιστορικότητα και τα γεωλογικά δεδομένα - δεν μπορεί να καθοριστεί με ακρίβεια η στιγμή ενός μελλοντικού σεισμού. Η αλληλεπίδραση πολλαπλών αστάθμητων παραγόντων, καθιστούν οποιαδήποτε απόπειρα, ανεξαρτήτως χρονικού ορίζοντα, ένα εξαιρετικά δυσεπίλυτο επιστημονικό ζήτημα.

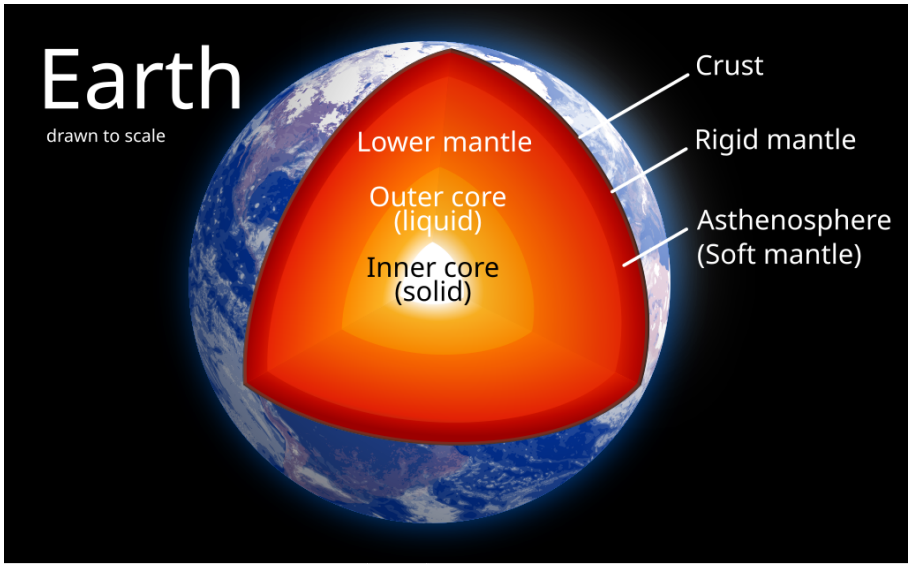
Οι σεισμοί εντοπίζονται κυρίως στα όρια των τεκτονικών πλακών. Η συσσώρευση τάσης, ως αποτέλεσμα της κίνησης των πλακών, οδηγεί τελικά σε θραύση ρηγμάτων, όταν ξεπεραστεί το όριο αντοχής τους. Ωστόσο, τα ρηγματικά συστήματα στη φύση δεν είναι απλές, μεμονωμένες, γεωλογικές δομές, αλλά σύνθετα δίκτυα, που αλληλεπιδρούν μεταξύ τους. Τα συστήματα αυτά εμφανίζουν ιδιαίτερη πολυπλοκότητα και χαοτική δυναμική, γεγονός που καθιστά εξαιρετικά δύσκολο να προβλεφθεί η σεισμική συμπεριφορά. Επιπλέον, υπάρχει εκτεταμένη αναζήτηση για πιθανούς προδρόμους σεισμών, που θα μπορούσαν να "βοηθήσουν" στην έγκαιρη πρόβλεψη - ενημέρωση και αντίδραση. Δυστυχώς, και σε αυτή την περίπτωση, έχουν βρεθεί μόνο ελάχιστα αξιόπιστα δείγματα. Ο μοναδικός πρόδρομος σεισμού, που αναντίρρητα έχει αποδειχθεί και καθιερωθεί, είναι η περιστασιακή εμφάνιση προσεισμών, δηλαδή γεγονότων που συμβαίνουν σε στενή χρονική και χωρική εγγύτητα έναντι ενός κύριου σεισμού [3].

Η σεισμική δραστηριότητα των μικρών σεισμών μπορεί να ελεγχθεί και να αναλυθεί αξιόπιστα και σε βάθος, μέσω στατιστικών μοντέλων. Ωστόσο, οι ισχυροί και καταστροφικοί σεισμοί, εμφανίζονται με πολύ μικρότερη συχνότητα. Συχνά μεσολαβούν χρονικά διαστήματα εκατοντάδων ετών, μεταξύ δύο μεγάλων σεισμών σε ένα συγκεκριμένο ρήγμα. Συνεπώς, αυτή η περιορισμένη χρονική έκταση των σεισμολογικών καταλόγων αποτελεί έναν περιοριστικό παράγοντα στην αξιολόγηση τέτοιων σπάνιων γεγονότων [3].

Στη σημερινή εποχή, έχει ανανεωθεί το επιστημονικό ενδιαφέρον για την ανάλυση σεισμολογικών συμβάντων, με την έλευση τεχνολογιών, όπως το Διαδίκτυο των Πραγμάτων (Internet of Things – IoT), αλλά και της Μηχανικής Μάθησης (Machine Learning – ML). Σε αυτό το πλαίσιο, η χρήση του Διαδικτύου των Πραγμάτων (IoT) επιτρέπει τη συνεχή και σε πραγματικό χρόνο παρακολούθηση της σεισμικής δραστηριότητας. Η ανάπτυξη δικτύων από συνδεδεμένους αισθητήρες (σεισμογράφοι, GPS και αισθητήρες πίεσης ή παραμόρφωσης), συμβάλλει στην απρόσκοπτη συλλογή και μετάδοση δεδομένων [4]. Οι αισθητήρες IoT μπορούν να τοποθετηθούν τόσο σε απομακρυσμένες γεωλογικές περιοχές, όσο και σε κρίσιμες υποδομές (πχ. γέφυρες, φράγματα, κτίρια), παρέχοντας πληροφορίες για τη δυναμική συμπεριφορά κατασκευών, την τοπική σεισμική απόκριση και τη σταδιακή συσσώρευση παραμορφώσεων.

Η Μηχανική Μάθηση έχει σημειώσει ραγδαία πρόοδο τα τελευταία χρόνια. Ο συμβολή της είναι καθοριστική και ενισχύει το ρόλο της σε διάφορες επιστημονικές περιοχές. Σε αντίθεση με τις παραδοσιακές μεθόδους, η Μηχανική Μάθηση προσφέρει δυνατότητες για την επίλυση σύνθετων προβλημάτων και αυξάνει την υπολογιστική αποδοτικότητα [5]. Οι αλγόριθμοι Μη Εποπτευόμενης Μάθησης, που είναι και το αντικείμενο αυτής της Διπλωματικής Εργασίας, αξιοποιούν τις τεράστιες ποσότητες των πολυδιάστατων σεισμολογικών δεδομένων, για να ανιχνεύσουν μοτίβα και λεπτομέρειες, που συχνά διαφεύγουν στις παραδοσιακές μεθοδολογίες. Αυτή η προσέγγιση ανοίγει νέους ορίζοντες στην κατανόηση των πολύπλοκων μηχανισμών, που διέπουν τη σεισμική δραστηριότητα. Τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν για την αυτόματη ανίχνευση συμβάντων, την αναγνώριση σεισμικών προτύπων και την υποστήριξη έγκαιρης προειδοποίησης. Φυσικά, λόγω της χαοτικής και πολύπλοκης συμπεριφοράς των σεισμών, δεν έχουν, ακόμη τουλάχιστον, τη δυνατότητα να εγγυηθούν την ακριβή πρόβλεψη ενός μεμονωμένου γεγονότος. Ωστόσο, η μελέτη των αποτελεσμάτων, προσφέρει νέες, σημαντικές προοπτικές στον τρόπο αντίληψης και κατανόησης της σεισμικής συμπεριφοράς.

1.1.1 Θεωρία των Λιθοσφαιρικών Πλακών – Τεκτονικές Πλάκες



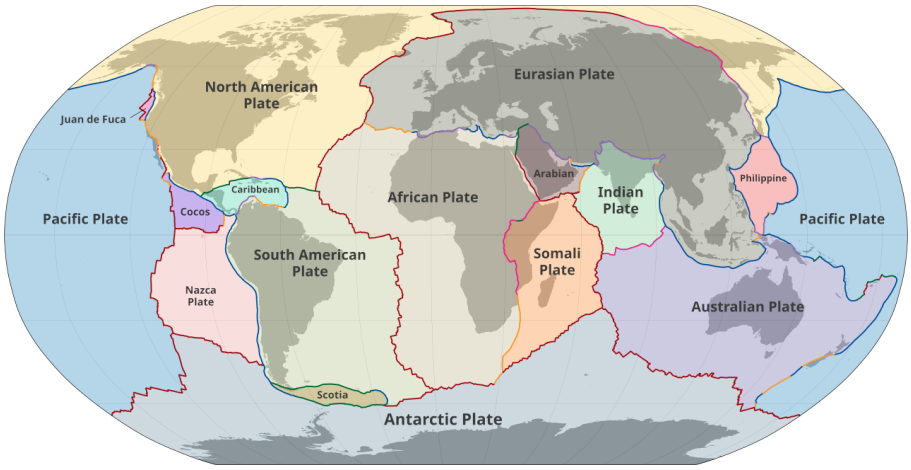
Σχήμα 1.1: Η δομή της Γης. Πηγή: Wikipedia [1]

Η Γη αποτελείται από τρία στρώματα (Σχήμα 1.1): τον φλοιό, τον μανδύα και τον πυρήνα. Ο φλοιός είναι το εξωτερικό στρώμα της Γης. Έχει βάθος από 5 έως 70 χιλιόμετρα και χωρίζεται σε ωκεάνιο και ηπειρωτικό. Ο μανδύας είναι το μεσαίο στρώμα. Έχει βάθος 2.900 χιλιόμετρα. Αποτελεί το παχύτερο στρώμα και χωρίζεται σε ανώτερο και κατώτερο μανδύα. Ο πυρήνας αποτελεί το βαθύτερο στρώμα της Γης. Έχει βάθος από 2.900 χιλιόμετρα έως 6.400 χιλιόμετρα και χωρίζεται σε εξωτερικό και εσωτερικό πυρήνα [1].

Ήταν το 1912, όταν ο μετεωρολόγος Alfred Wegener διατύπωσε τη θεωρία, ότι στο παρελθόν οι ήπειροι αποτελούσαν μια ενιαία ήπειρο, την Πανγαία. Ισχυρίστηκε, πως οι ήπειροι της Γης δεν κατέχουν σταθερή θέση και μετακινούνται συνεχώς. Αρχικά, η θεωρία του Wegener δε βρήκε αποδοχή. Η επιστημονική τεκμηρίωση των απόψεων του ήρθε αρκετές δεκαετίες αργότερα. Χρειάστηκαν πενήντα χρόνια (1960), για να δοθεί μια πειστική εξήγηση για τη κίνηση των λιθοσφαιρικών πλακών. Οι κινήσεις στο εσωτερικό του μανδύα, που ευθύνονται για τη δυναμική συμπεριφορά των λιθοσφαιρικών πλακών, οφείλονται στην ανάγκη αποβολής της θερμότητας από το εσωτερικό της Γης. Η θεωρία αυτή εξελίχθηκε και διαμορφώθηκε πλήρως ως η Θεωρία των Λιθοσφαριρικών Πλακών, και πλέον, αποτελεί το δόγμα των Γεωεπιστημών [6].

Η λιθόσφαιρα, είναι το εξωτερικό και πιο στερεό τμήμα του φλοιού της Γης. Είναι κατακερματισμένη σε μικρές πλάκες, οι οποίες "επιπλέουν" και μετακινούνται ανεξάρτητα πάνω στο πιο ελαστικό υπόστρωμα του μανδύα, την ασθενόσφαιρα. Οι πλάκες αυτές είναι γνωστές ως τεκτονικές ή λιθοσφαιρικές πλάκες (Σχήμα 1.2) και όπως είναι σαφές, αποτελούν τα κύρια δομικά στοιχεία της επιφάνειας της Γης.

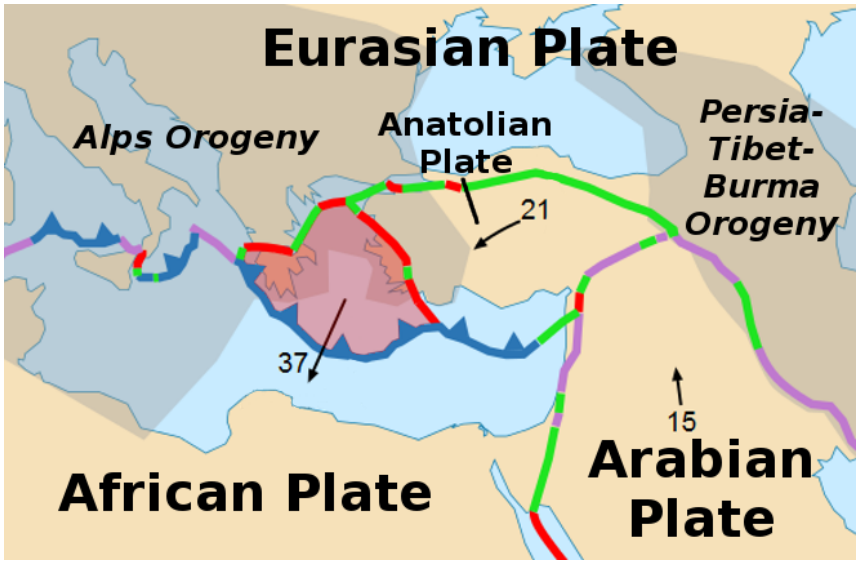
[1] https://en.wikipedia.org/wiki/Internal_structure_of_Earth



Σχήμα 1.2: Οι τεκτονικές πλάκες της Γης. Πηγή: Wikipedia [2]

Υπάρχουν επτά μεγάλες τεκτονικές πλάκες (94% της γήινης επιφάνειας) και αρκετές μικρότερες. Οι κύριες τεκτονικές πλάκες είναι οι: Αφρικανική, Ανταρκτική, Ευρασιατική, Ινδο-Αυστραλιανή, Β. Αμερικής, Ν. Αμερικής και Ειρηνική. Κάποιες από τις μικρότερες είναι οι: Αραβική, Φιλιππίνων. Υπάρχουν και ακόμη μικρότερες, στις οποίες συγκαταλέγονται η πλάκα του Αιγαίου (γνωστή ως Ελληνική), η πλάκα της Ανατολίας και η Αδριατική πλάκα (Απουλία).

Η πλάκα του Αιγαίου (Σχήμα 1.3) θεωρείται τμήμα της Ευρασιατικής πλάκας, από την οποία βρίσκεται σε διαδικασία απόκλισης (απομάκρυνσης). Αντίστοιχα, η πλάκα της Ανατολίας είναι ηπειρωτική, και χωρίζεται από την Ευρασιατική με το ρήγμα της Βόρειας Ανατολίας και από την Αραβική με το ρήγμα της Ανατολικής Ανατολίας.



Σχήμα 1.3: Η πλάκα του Αιγαίου. Πηγή: Wikipedia [3]

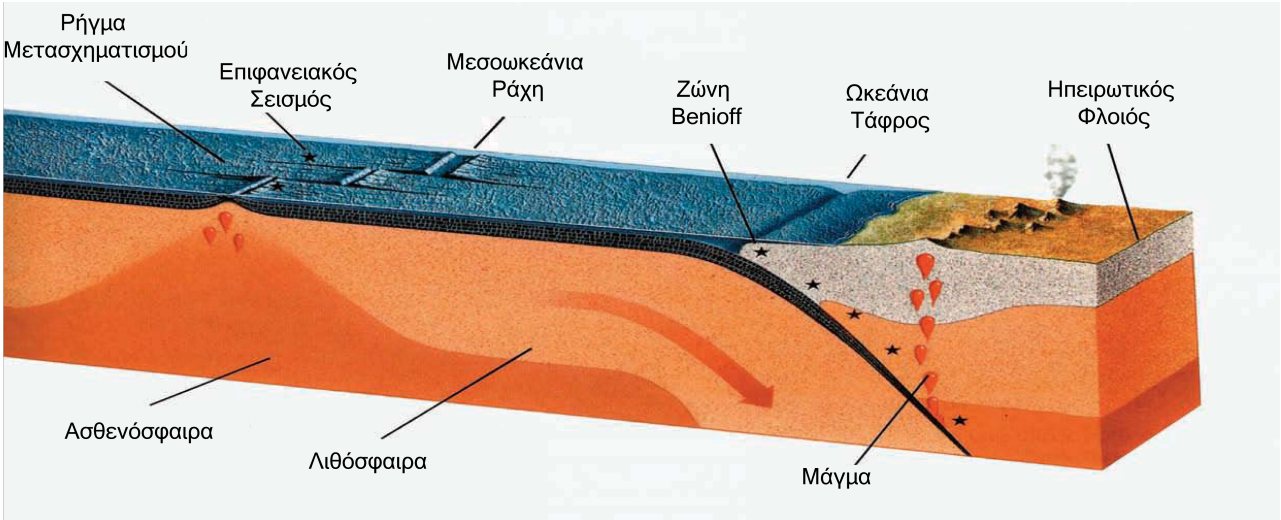
[2] https://en.wikipedia.org/wiki/Plate_tectonics
[3] https://en.wikipedia.org/wiki/Aegean_Sea_plate

Η πλάκα της Αδριατικής είναι μια μικρή, ηπειρωτική τεκτονική πλάκα. Κατά την Κρητιδική περίοδο αποσπάστηκε από την Αφρικανική πλάκα. Το βόρειο τμήμα της έχει υποστεί παραμόρφωση, κατά την σύγκρουσή της με την Ευρασιατική πλάκα και την ορογένεση των Άλπεων.

Είναι ιδιαίτερα σημαντικό να τονιστεί, ότι σε παγκόσμιο επίπεδο, η συντριπτική πλειονότητα - που αγγίζει το 80% των σεισμικών δονήσεων και των ηφαιστειακών εκρήξεων - δεν κατανέμεται ομοιόμορφα στην επιφάνεια της Γης. Αντίθετα, παρατηρείται μια ισχυρή συσχέτιση με τα όρια των λιθοσφαιρικών πλακών. Αυτή η συγκέντρωση υπογραμμίζει τον καθοριστικό ρόλο της τεκτονικής των πλακών ως τον πρωταρχικό μηχανισμό πρόκλησης των πιο έντονων γεωλογικών φαινομένων, που διαμορφώνουν τον πλανήτη μας και επηρεάζουν άμεσα το περιβάλλον και τις ανθρώπινες κοινωνίες.

Η γεωλογική δράση στα όρια των λιθοσφαιρικών πλακών είναι έντονη και εκδηλώνεται με γενέσεις ηφαιστείων, τάφρων, οροσειρών και φυσικά σεισμών. Υπάρχουν τρεις βασικοί τύποι τεκτονικών ορίων (Σχήμα 1.4) [6]:

- **Αποκλίνοντα όρια:** Στην περίπτωση αυτή οι πλάκες απομακρύνονται, με συνέπεια τη δημιουργία νέου φλοιού (πχ. μέσοωκεάνια ράχη – Ατλαντικός ωκεανός). Η σεισμική δραστηριότητα είναι μέτρια με επιφανειακούς σεισμούς.
- **Συγκλίνοντα όρια:** Στην περίπτωση αυτή οι πλάκες συγκρούονται, προκαλώντας υποβύθιση (καταστροφή φλοιού), ή ορογένεση (πχ Άνδεις, Ιαπωνία). Παρατηρείται έντονη σεισμικότητα, που φτάνει σε πολύ μεγάλα βάθη.
- **Πλευρικώς ολισθαίνοντα όρια:** Στην περίπτωση αυτή οι πλάκες ολισθαίνουν πλευρικά η μία ως προς την άλλη (πχ. ρήγμα San Andreas). Δεν υπάρχει ούτε παραγωγή, αλλά ούτε καταστροφή φλοιού, όπως στις παραπάνω περιπτώσεις. Οι οριζόντιες μετακινήσεις προκαλούν τη συσσώρευση τάσεων, οι οποίες απελευθερώνονται με τη μορφή ισχυρών σεισμών.



Σχήμα 1.4: Οι κινήσεις των λιθοσφαιρικών πλακών. Πηγή: [1]

.

1.1.2 Βασικές Έννοιες Σεισμών

Η εστία ή υπόκεντρο του σεισμού είναι το σημείο εντός της γης, όπου ξεκινά η ρήξη, ενώ το επίκεντρο είναι η κάθετη προβολή της εστίας στην επιφάνεια της Γης. Εστιακό βάθος καλείται η απόσταση μεταξύ της εστίας και του επικέντρου του σεισμού.

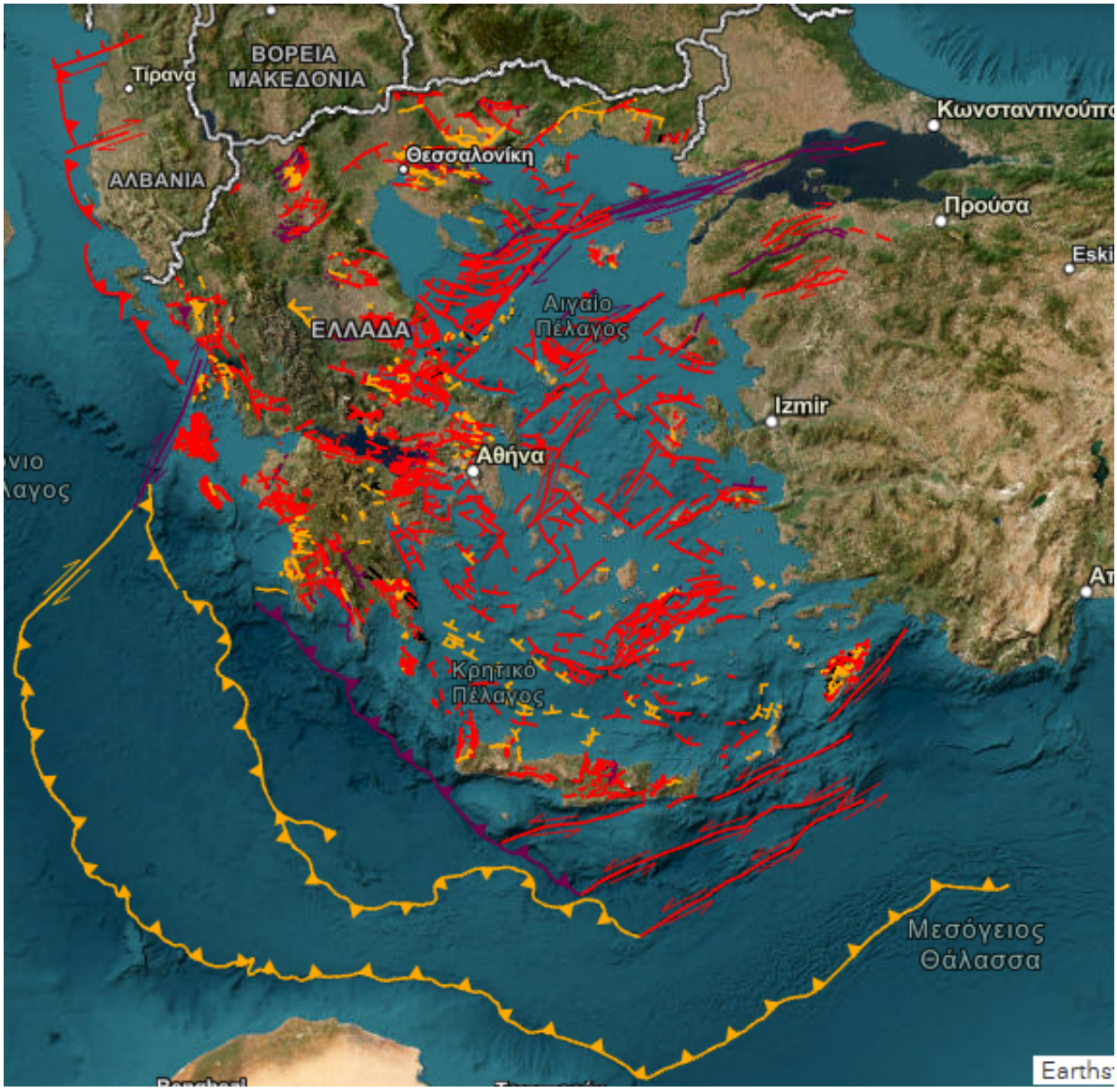
Στο εσωτερικό της Γης, και με τις σχετικές κινήσεις των λιθοσφαιρικών πλακών τα πετρώματα υφίστανται συνεχείς πιέσεις και τάσεις. Η συσσώρευση τεράστιας ποσότητας δυναμικής ενέργειας στα πετρώματα μόλις φτάσει σε ένα συγκεκριμένο όριο, οδηγεί στην απότομη θραύση των πετρωμάτων [7]. Η επιφάνεια πάνω στην οποία γίνεται αυτή η θραύση και η σχετική κίνηση των δύο τεμαχίων του πετρώματος ονομάζεται σεισμικό ρήγμα. Στο χρονικό αυτό σημείο έχουμε τη γένεση ενός σεισμού [1]. Η αποθηκευμένη ελαστική ενέργεια απελευθερώνεται μερικώς ή πλήρως και μετατρέπεται σε κινητική ενέργεια, προκαλώντας ταλάντωση στο πέτρωμα. Αυτή η ταλάντωση διαδίδεται με τη μορφή σεισμικών κυμάτων, τα οποία ταξιδεύουν μέσα από διάφορα γεωλογικά στρώματα και φτάνουν σε μεγάλες αποστάσεις από το επίκεντρο του σεισμού [7].

Τα ρήγματα ταξινομούνται ανάλογα με τη διεύθυνση και τη φορά της σχετικής κίνησης των τεμαχών κατά τη θραύση σε [6]:

- **Κανονικά ρήγματα:** Στην περίπτωση αυτή, τα δύο μέρη απομακρύνονται μεταξύ τους. Τα ρήγματα αυτά σχηματίζονται λόγω εφελκυστικών δυνάμεων, που προκαλούν διαστολή της λιθόσφαιρας. Συχνά απαντώνται σε περιοχές όπου ο φλοιός εκτείνεται.

- **Ανάστροφα ρήγματα:** Στην περίπτωση αυτή, το ένα μέρος ωθείται επάνω στο άλλο. Τα ρήγματα αυτά δημιουργούνται από θλιπτικές δυνάμεις, που συμπιέζουν τον φλοιό. Είναι χαρακτηριστικά των συγκλινόντων ορίων.

- **Οριζόντιας μετατόπισης ρήγματα:** Στην περίπτωση αυτή, τα τεμάχη μετακινούνται οριζόντια (παράλληλα) το ένα σε σχέση με το άλλο.



Σχήμα 1.5: Τα ρήγματα της Ελλάδας. Πηγή: Ε.Α.Γ.Μ.Ε. HeDBAF [4]

Στο (Σχήμα 1.5) αποτυπώνονται τα ρήγματα που υπάρχουν στην Ελλάδα. Μέσω της συγκεκριμένης γεωχωρικής Βάσης Δεδομένων, παρέχονται ουσιαστικές και τεκμηριωμένες πληροφορίες για τα ενεργά και σεισμικά ρήγματα του ελλαδικού χώρου. Η Εθνική Βάση Δεδομένων Ενεργών Ρηγμάτων της Ελλάδας (HeDBAF)[5]συνιστά ένα ιδιαιτέρως σημαντικό και αξιόλογο εθνικό εγχείρημα. Είναι ο καρπός της συνεργασίας όλων των ακαδημαϊκών και ερευνητικών φορέων της χώρας. Υλοποιήθηκε υπό την αιγίδα του Οργανισμού Αντισεισμικού Σχεδιασμού και Προστασίας (ΟΑΣΠ)[6].

[4] https://activefaults.eagme.gr/el/

[5] https://gaia.igme.gr/portal/apps/webappviewer/index.html?id=f141b9da08f34107806d227cb0d6afe9

[6] https://oasp.gr/

Το μέγεθος (Magnitude) ενός σεισμού εκφράζει την ενέργεια που εκλύεται, λόγω της θραύσης των πετρωμάτων (σεισμική δόνηση). Υπάρχουν πολλές κλίμακες για το μέγεθος των σεισμών. Οι μετρήσεις στο Γεωδυναμικό Ινστιτούτο του Εθνικού Αστεροσκοπείου Αθηνών[7] αναφέρονται σύμφωνα με το M_L (Local Magnitude), δηλαδή την ευρέως γνωστή, κλίμακα Richter, που είναι τοπικού μεγέθους. Συγκεκριμένα, η λογαριθμική φύση της κλίμακας Richter σημαίνει, ότι μια αύξηση του μεγέθους, κατά μία μόλις μονάδα αντιστοιχεί σε μια εντυπωσιακή αύξηση της ενέργειας που απελευθερώνεται κατά περίπου 31,5 φορές. Αυτή η εκθετική αύξηση της ενέργειας με κάθε μονάδα μεγέθους εξηγεί, γιατί ακόμα και μικρές διαφορές στην κλίμακα Richter ερμηνεύονται ως σημαντικά διαφορετικές επιπτώσεις στο έδαφος και στις κατασκευές.

Δε θα πρέπει φυσικά, να συγχέεται το μέγεθος ενός σεισμού, με την ένταση. Το μέγεθος είναι μία και συγκεκριμένη τιμή, που χαρακτηρίζει τον σεισμό. Αντίθετα, η ένταση ενός σεισμού μεταβάλλεται από περιοχή σε περιοχή. Εξαρτάται από το μέγεθος, αλλά και την απόσταση από το επίκεντρο του σεισμού και από άλλους παράγοντες [6]. Για τη μέτρηση της έντασης ενός σεισμού χρησιμοποιείται η κλίμακα Mercalli (MM), η οποία αποτυπώνει τις επιπτώσεις ενός σεισμού στους ανθρώπους και στο δομημένο περιβάλλον.

Οι σεισμοί μπορούν να κατηγοριοποιηθούν, έτσι ώστε να γίνει πιο εύκολη η κατανόηση της σεισμικής συμπεριφοράς. Βάσει του εστιακού βάθους διακρίνονται σε [1]:

- **Επιφανειακούς σεισμούς (0 - 70 km):** Το εστιακό σημείο βρίσκεται σε μικρό βάθος από την επιφάνεια της Γης και συνήθως προκαλούν τις μεγαλύτερες καταστροφές.
- **Ενδιάμεσους σεισμούς (70 - 300 km):** Το εστιακό σημείο βρίσκεται σε ενδιάμεσο βάθος.
- **Μεγάλου βάθους σεισμούς (>300 km):** Το εστιακό σημείο βρίσκεται σε μεγάλο βάθος και συνήθως δεν προκαλούν μεγάλες επιφανειακές ζημιές.

Οι σεισμοί ενδιάμεσου και μεγάλου βάθους καλούνται πλουτώνιοι. Οι περισσότεροι σεισμοί που προκαλούν σοβαρές καταστροφές έχουν μικρά εστιακά βάθη, δηλαδή λιγότερο από 30 χιλιόμετρα. Γενικά, όσο πιο μικρό είναι το εστιακό βάθος ενός σεισμού, τόσο μεγαλύτερες είναι οι επιπτώσεις του στην επιφάνεια της Γης και, κατά συνέπεια, μεγαλύτερη και η καταστροφική του δύναμη [7].

Βάσει της αιτίας έχουμε τους [1]:

- **Τεκτονικούς σεισμούς:** Αποτελούν τους πιο συχνούς σεισμούς. Προκαλούνται από την απότομη κίνηση των τεκτονικών πλακών. Αυτές οι κινήσεις μπορεί να είναι συμπιεστικές, εφελκυστικές ή οριζόντιες. Το 90% των σεισμών παγκοσμίως είναι τεκτονικοί σεισμοί.
- **Ηφαιστειακούς σεισμούς:** Σχετίζονται με ηφαιστειακή δραστηριότητα. Προκαλούνται από την εισροή/εκροή του μάγματος μέσα σε ένα ηφαίστειο. Δεν είναι πάντα ισχυροί σε μέγεθος, μπορούν όμως να γίνουν ιδιαιτέρως καταστροφικοί. Το 7% παγκοσμίως είναι ηφαιστειακοί σεισμοί. Επιπλέον, σε αυτή την περίπτωση συχνά συναντάται σεισμική ακολουθία διάρκειας ημερών, εβδομάδων ή μηνών. Η ακολουθία αυτή ονομάζεται σμηνοσειρά και κανένας σεισμός δε μπορεί εμφανώς να χαρακτηριστεί ως κύριος.

[7] `https://www.gein.noa.gr/`

- **Εγκατακρημνισιγενείς σεισμούς:** Το 3% της σεισμικής δραστηριότητας παγκοσμίως οφείλεται σε αυτή την κατηγορία σεισμών. Οι συγκεκριμένοι σεισμοί συμβαίνουν από την πτώση οροφών σπηλαίων, λόγω διάβρωσης. Είναι τοπικοί και μικροί σεισμοί και βρίσκονται μακριά από τα όρια των λιθοσφαιρικών πλακών.

1.1.3 Σεισμοί στον Ελλαδικό χώρο

Η Ελλάδα διασχίζεται από τη μία άκρη ως την άλλη, από ένα μεγάλο πλήθος ρηγμάτων. Ενδεχομένως, να μην υπάρχει περιοχή στη χώρα μας, που να μην έχει αισθανθεί, έστω και μία μικρή σεισμική δόνηση. Η Ελλάδα συγκαταλέγεται στις πλέον σεισμογενείς περιοχές στον κόσμο. Καταλαμβάνει την πρώτη θέση στην Ευρώπη και τη Μεσόγειο και την έκτη θέση παγκοσμίως. Αυτό οφείλεται στη γεωγραφική της θέση. Όπως έχει ήδη αναφερθεί στην προηγούμενη ενότητα, η πλειονότητα των σεισμών έχει ως κύρια αιτία τις κινήσεις των λιθοσφαιρικών πλακών. Επομένως, η έντονη σεισμικότητα στην Ελλαδικό χώρο, είναι στην ουσία το αποτέλεσμα της σύγκλισης και επαφής των ορίων των τεκτονικών πλακών, της Ευρασιατικής και της Αφρικανικής [1].

- **Το Ελληνικό ή Αιγιακό Τόξο** (Σχήμα 1.6) είναι ένας γεωλογικός σχηματισμός, που διαμορφώνεται από τις κινήσεις των λιθοσφαιρικών πλακών (Ευρασιατικής και Αφρικανικής), και πιο συγκεκριμένα στο σημείο σύγκλισης της ωκεάνιας (τμήμα της αποτελεί η Ανατολική Μεσόγειος) με την ηπειρωτική πλάκα (τμήμα της αποτελεί το Αιγαίο). Λόγω της μεγαλύτερης

Σχήμα 1.6: Το Ελληνικό τόξο. Πηγή: [2]

πυκνότητας, η ωκεάνια πλάκα της Ανατολικής Μεσογείου βυθίζεται κάτω από την ηπειρωτική Αιγιακή πλάκα [1]. Το τόξο είναι κατά βάση θαλάσσιο. Πρόκειται για μια οροσειρά στο νότιο τμήμα της Ελληνικής πλάκας. Ουσιαστικά οι βουνοκορφές είναι τα νησιά του Ιονίου, η Κρήτη, και τα Δωδεκάνησα. Εκτείνεται από τη Δυτική Αλβανία, τα νησιά του Ιονίου, την Κρήτη, την Κάρπαθο, τη Ρόδο, και φτάνει μέχρι τη νότια Τουρκία.

- **Η Ελληνική Τάφρος** διαμορφώνεται κατά μήκος της ζώνης επαφής μεταξύ των δύο λιθοσφαιρικών πλακών. Πρόκειται για ένα σύνθετο σύστημα από βαθιές θαλάσσιες λεκάνες. Εκτείνεται από την Κεφαλονιά έως και τη Ρόδο και το μεγαλύτερο βάθος της έχει καταγραφεί νοτιοδυτικά της Πελοποννήσου, στο Ιόνιο Πέλαγος, φτάνοντας περίπου τα 4.500 μέτρα. Αυτό αποτελεί και το βαθύτερο σημείο σε ολόκληρη τη Μεσόγειο [1].

- **Το Ελληνικό Νησιωτικό Τόξο** εκτείνεται παράλληλα με την τάφρο και πολύ κοντά σε αυτήν. Περιλαμβάνει μια ακολουθία νησιών (Λευκάδα, Κεφαλονιά, Ζάκυνθο, Κύθηρα, Κρήτη, Ρόδο), και τη νοτιοδυτική Πελοπόννησο.

- **Το Ελληνικό Ηφαιστειακό Τόξο** περιλαμβάνει ένα σύστημα από ηφαίστεια — τόσο ενεργά όσο και ανενεργά — όπως αυτά του Σουσακίου, των Μεθάνων, της Μήλου, της Σαντορίνης και της Νισύρου. Ο σχηματισμός τους οφείλεται στη μερική τήξη υλικών, που προέρχονται από την υποβυθιζόμενη Αφρικανική λιθοσφαιρική πλάκα [1].

- **Η Οπισθοτάφρος**, ταυτίζεται με το Κρητικό Πέλαγος. Αποτελεί μια θαλάσσια λεκάνη μικρότερου βάθους, σε σχέση με την κύρια τάφρο. Το μέγιστο βάθος της αγγίζει περίπου τα 2.000 μέτρα. Βρίσκεται μεταξύ του νησιωτικού τόξου και του ηφαιστειακού τόξου.

- **Η Τάφρος του Βορείου Αιγαίου** είναι το κυριότερο μορφολογικό χαρακτηριστικό στην περιοχή. Έχει βάθος περίπου 1.500 μέτρα και φτάνει ως τη θάλασσα του Μαρμαρά. Είναι δε, το πιο δυναμικό ρήγμα, καθώς αποτελεί τη φυσική συνέχεια του ρήγματος της Βόρειας Ανατολίας.

Η παραμόρφωση των πετρωμάτων εμφανίζεται εντονότερα στις περιοχές, όπου συγκλίνουν οι τεκτονικές πλάκες. Οι σεισμικές εστίες εντοπίζονται κυρίως σε μια σεισμική ζώνη γνωστή ως "ζώνη Benioff", όπως φαίνεται στο (Σχήμα 1.7).

Σχήμα 1.7: Σχηματική απεικόνιση του Ελληνικού τόξου. Πηγή: [2]

1.2 Ανάλυση Σεισμολογικών Δεδομένων – Συσταδοποίηση

Η φύση δρα ανεξέλεγκτα, απροειδοποίητα, και η καταγραφή των σεισμικών δονήσεων παράγει τεράστιες ποσότητες πληροφορίας από διάφορες μετρήσεις. Η μελέτη της σεισμικής δραστηριότητας απαιτεί τη συνεχή συλλογή, επεξεργασία και αποθήκευση ενός μεγάλου όγκου δεδομένων. Τα δεδομένα αυτά σχετίζονται με τη χωροχρονική εξέλιξη της σεισμικής δραστηριότητας. Είναι προφανές, ότι αποτελούν πολύτιμο υλικό για έρευνα, καθώς είναι εκτενή και ακατέργαστα. Στην Ελλάδα, η ενόργανη καταγραφή των σεισμών ξεκίνησε το 1897 [8], σηματοδοτώντας την απαρχή μιας νέας εποχής, που αφορά στη μελέτη της σεισμικής δραστηριότητας. Σήμερα, τα σεισμολογικά δεδομένα διατίθενται, σχεδόν, σε πραγματικό χρόνο μέσω ιστοσελίδων, όπως για παράδειγμα του Γεωδυναμικού Ινστιτούτου Αθηνών[8]και άλλων σεισμολογικών εργαστηρίων πανεπιστημιακών ιδρυμάτων.

Κάθε σεισμικό γεγονός, που καταγράφεται από τα σεισμογραφικά δίκτυα συνοδεύεται από ένα πλούσιο σύνολο ποσοτικών και ποιοτικών παραμέτρων, οι οποίες συνιστούν το αποτύπωμά του. Πέρα από τον ακριβή χρόνο γένεσης (με ακρίβεια δευτερολέπτου), την λεπτομερή τοποθεσία του επικέντρου (μέσω του γεωγραφικού μήκους και πλάτους) και το μέγεθος, που αντικατοπτρίζει την απελευθερωμένη ενέργεια, καταγράφεται επίσης το εστιακό βάθος, πληροφορία κρίσιμη για την κατανόηση της γεωδυναμικής διαδικασίας, που τον προκάλεσε και των πιθανών επιπτώσεων στην επιφάνεια. Επιπρόσθετα, συλλέγονται και άλλες παράμετροι, όπως ο τύπος των σεισμικών κυμάτων, η διάρκεια της δόνησης, η διεύθυνση της ρήξης και οι φασματικές ιδιότητες του σεισμικού σήματος, προσφέροντας μια ολοκληρωμένη εικόνα του φαινομένου. Συνεπώς, κάθε σεισμός αποτελεί μια πηγή σύνθετης και πρωτογενούς πληροφορίας, η οποία δεν είναι άμεσα κατανοητή και χρήσιμη χωρίς την κατάλληλη επεξεργασία και ερμηνεία. Η Σεισμολογία, ως επιστήμη που στοχεύει στην κατανόηση των σεισμικών φαινομένων και στην εκτίμηση της σεισμικής επικινδυνότητας, αξιοποιεί συστηματικά τις ισχυρές μεθόδους της Στατιστικής Ανάλυσης για την αναγνώριση προτύπων και συσχετίσεων στα σεισμικά δεδομένα, καθώς και της Θεωρίας Πιθανοτήτων για τη μοντελοποίηση της τυχαιότητας και την εκτίμηση της πιθανότητας εμφάνισης σεισμών (χαρακτηριστικό παράδειγμα αποτελεί η Κατανομή Poisson, η οποία χρησιμοποιείται για την ανάλυση της συχνότητας εμφάνισης ανεξάρτητων γεγονότων, όπως οι σεισμοί, σε ένα δεδομένο χρονικό διάστημα ή χώρο) [3].

Λαμβάνοντας υπόψιν τον τεράστιο όγκο και την πολυπλοκότητα των δεδομένων που συσσωρεύονται από τα σύγχρονα σεισμογραφικά δίκτυα για κάθε σεισμικό γεγονός, η χρήση προηγμένων τεχνικών Εξόρυξης Γνώσης (Knowledge Mining) και Μηχανικής Μάθησης (ML) κρίνεται πλέον επιτακτική. Οι παραδοσιακές στατιστικές μέθοδοι, αν και θεμελιώδεις, συχνά αδυνατούν να ανακαλύψουν λεπτές συσχετίσεις, κρυμμένα πρότυπα ή μη γραμμικές σχέσεις μέσα σε αυτόν τον πλούτο πληροφοριών. Η Εξόρυξη Γνώσης μπορεί να αυτοματοποιήσει την ανακάλυψη νέων γνώσεων και χρήσιμων πληροφοριών από τα σεισμικά δεδομένα, ενώ η Μηχανική Μάθηση προσφέρει εξελιγμένους αλγορίθμους, ικανούς να αναλύσουν μεγάλους όγκους δεδομένων, να αναγνωρίσουν σύνθετα πρότυπα και να δημιουργήσουν προβλεπτικά μοντέλα με αυξημένη ακρίβεια.

Ειδικότερα, η συσταδοποίηση (clustering) επιδιώκει να εντοπίσει ομάδες (clusters) σεισμικών γε-

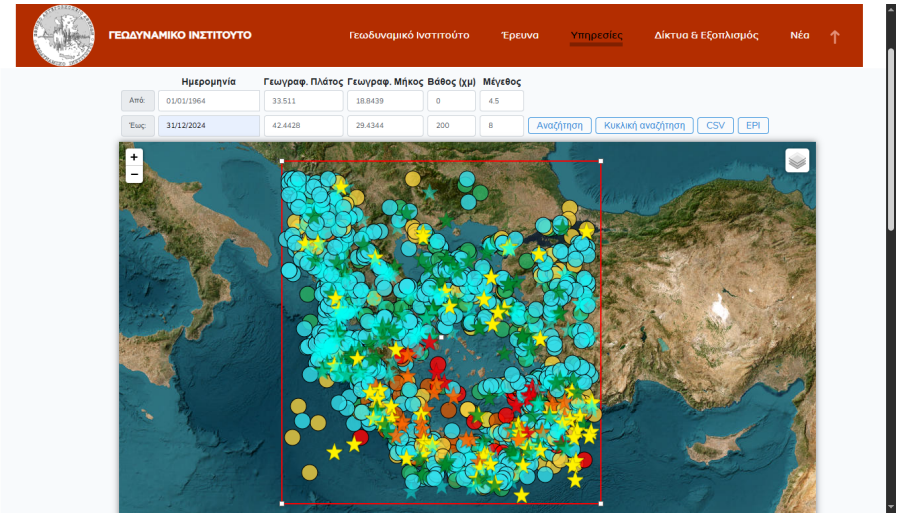
[8] https://www.gein.noa.gr/

γονότων, που παρουσιάζουν κοινά χαρακτηριστικά, όπως για παράδειγμα, εγγύτητα στον χώρο. Επομένως, μπορεί να αξιοποιηθεί, έτσι ώστε να αντληθεί συγκεκριμένη πληροφορία για τους σεισμούς, που σαφέστατα είναι εξέχουσας σημασίας. Μέσω της υλοποίησης ποικίλλων αλγορίθμων επιτυγχάνεται η αποκάλυψη κρυφής γνώσης, μοτίβων και σχέσεων. Με τον τρόπο αυτό, καθίσταται εφικτή η δυνατότητα να εντοπιστούν περιοχές, στις οποίες υπάρχει αυξημένη δραστηριότητα, ή παρουσιάζουν κάποια ιδιαίτερη σεισμική συμπεριφορά. Η ενσωμάτωση αυτών των τεχνολογιών αναμένεται να οδηγήσει στην ανάπτυξη πιο αποτελεσματικών μεθόδων έγκαιρης προειδοποίησης. Αναπτύσσεται βαθύτερη αντίληψη της σεισμικής δραστηριότητας και των πιθανών συσχετίσεων της με γεωλογικές δομές (ρήγματα, τεκτονικές πλάκες κτλ).

Στην παρούσα εργασία, η συσταδοποίηση χρησιμοποιείται ως μέσο διερεύνησης της κατανομής των σεισμών στον Ελλαδικό χώρο, για τη χρονική περίοδο των τελευταίων 61 ετών. Μέσα από αυτή τη διαδικασία, επιδιώκεται η ανάδειξη σημαντικών χωροχρονικών προτύπων, η ταυτοποίηση συστάδων σεισμικής δραστηριότητας και η απομόνωση πιθανών ακραίων γεγονότων, που εμφανίζονται ως θόρυβος.

1.3 Σεισμολογικά Δεδομένα του Γεωδυναμικού Ινστιτούτου του Εθνικού Αστεροσκοπείου Αθηνών

Το Γεωδυναμικό Ινστιτούτο του Εθνικού Αστεροσκοπείου Αθηνών αποτελεί τον κύριο φορέα συλλογής και διάθεσης σεισμολογικών δεδομένων στην Ελλάδα. Ιδρύθηκε το 1893, και η λειτουργία του από τότε είναι συνεχής. Από την ίδρυσή του, το Ινστιτούτο έχει αναπτύξει ένα εκτεταμένο δίκτυο σεισμολογικών σταθμών. Η συστηματική παρακολούθηση της σεισμικής δραστηριότητας εκτείνεται στα γεωγραφικά πλάτη 34°- 42°N και γεωγραφικά μήκη 19°- 30°E [8]. Τα δεδομένα που συλλέγονται και διατίθενται, περιλαμβάνουν το Γεωγραφικό Πλάτος και Μήκος του σεισμικού γεγονότος, το Βάθος (χλμ), τον Χρόνο Γένεσης (ημερομηνία και ώρα γένεσης), το Μέγεθος και την Τοποθεσία.



Σχήμα 1.8: Η Βάση Αναζήτησης για το επιλεγμένο Σύνολο Δεδομένων. Πηγή: Γεωδυναμικό Ινστιτούτο Αθηνών

	Standard	Standard	Standard	Standard	Standard	Standard
1	Origin Time (GMT)	Latitude	Longitude	Depth (km)	Magnitude (ML)	Location
2	2024-12-22 19:09:56	34.7571	26.2610	41	4.8	120.7 km ESE of Iraklion
3	2024-12-21 16:29:51	38.5579	21.6216	11	4.7	26.9 km NE of Mesolonghi
4	2024-12-13 08:34:12	40.2791	24.1292	14	4.6	10.7 km WNW of Kariai
5	2024-11-03 17:03:51	40.1353	23.2475	15	5.2	31.1 km SSW of Poliyiros
6	2024-10-19 07:08:30	34.8715	26.3132	8	4.8	116.2 km SW of Karpathos
7	2024-10-15 16:54:22	35.6323	28.6743	15	4.8	98.7 km SSE of Rodhos
8	2024-10-07 06:18:45	38.2828	23.3047	14	4.5	32.6 km SW of Chalkida

Σχήμα 1.9: Το αρχείο σε μορφή .csv. Πηγή: Γεωδυναμικό Ινστιτούτο Αθηνών

Θα πρέπει να τονιστεί σε αυτό το σημείο, η ανοικτή διάθεση των δεδομένων, μέσω της επίσημης ιστοσελίδας του Ινστιτούτου. Οι κατάλογοι και η βάση αναζήτησης των σεισμών καλύπτουν την περίοδο από το 1964 έως και σήμερα. Στο πλαίσιο της παρούσας εργασίας, θα χρησιμοποιηθούν δεδομένα από την ιστοσελίδα του Γεωδυναμικού Ινστιτούτου[9], για την περίοδο 01/01/1964 - 31/12/2024 και μέγεθος από 4.5R και πάνω (Σχήμα 1.8). Τα δεδομένα είναι σε μορφή CSV και η υλοποίηση του πειράματος γίνεται με τη γλώσσα Python. Το συγκεκριμένο σύνολο δεδομένων περιέχει 1.826 εγγραφές και 6 στήλες με τις εξής μεταβλητές (Σχήμα 1.9):

- **Origin Time (GMT) (Χρόνος Γένεσης σεισμού)**,
- **Latitude (Γεωγραφικό Πλάτος)**,
- **Longitude (Γεωγραφικό Μήκος)**,
- **Depth (km) (Βάθος σε χλμ)**,
- **Magnitude (ML) (Μέγεθος σε κλίμακα Richter)** και
- **Location (Τοποθεσία)**.

Είναι σημαντικό να υπογραμμιστεί, πως η επιλογή ενός κατώτατου ορίου μεγέθους (4.5R) για την ανάλυση των σεισμών δεν είναι αυθαίρετη, αλλά βασίζεται σε σαφή σεισμολογικά κριτήρια και την ιδιαίτερη φύση της σεισμικότητας της περιοχής. Σύμφωνα με την επίσημη ταξινόμηση των σεισμών, όπως αυτή καθορίζεται από αναγνωρισμένους φορείς, όπως για παράδειγμα το Michigan Tech University[10], οι σεισμοί με μέγεθος κάτω από 4.9R χαρακτηρίζονται γενικά ως ασθενείς. Πιο συγκεκριμένα, σεισμοί μεγέθους 4.0 - 4.9R θεωρούνται "Light" (Ασθενείς), ενώ αυτοί που κυμαίνονται από 5.0 - 5.9R κατατάσσονται ως "Moderate" (Μέτριοι) και σεισμοί από 6R και πάνω χαρακτηρίζονται ως "Strong" (Ισχυροί).

Το συγκεκριμένο σύνολο δεδομένων περιλαμβάνει κατά κύριο λόγο ένα μεγάλο πλήθος ασθενών και μέτριων σε μέγεθος σεισμών. Είναι η βάση για την Στατιστική Ανάλυση (Κεφ. 4) και την Συσταδοποίηση στα Σεισμολογικά Δεδομένα (Κεφ. 5). Η επιλογή να εξεταστούν μόνο σεισμοί μεγέθους 4.5R και άνω δεν οδηγεί σε απώλεια κρίσιμης πληροφορίας. Αντιθέτως, αυτή η στρατηγική επιλογή επιτρέπει την αποτελεσματικότερη εστίαση στους σεισμούς εκείνους, που έχουν τη δυναμική να επηρεάσουν και να αναδείξουν πιο ξεκάθαρες σεισμικές τάσεις και μοτίβα.

[9] https://www.gein.noa.gr/ypiresies-proionta/vasi-anazitisis/

[10] https://www.mtu.edu/geo/community/seismology/learn/earthquake-measure/magnitude/

1.3.1 Δεδομένα ορίων Τεκτονικών Πλακών

Η κατανόηση της σεισμικής δραστηριότητας σε μια περιοχή είναι άρρηκτα συνδεδεμένη με τη γεωδυναμική της δομή και ειδικότερα, με την αλληλεπίδραση των τεκτονικών πλακών. Προκειμένου να ενισχυθεί η ερμηνευτική αξία των γεωχωρικών οπτικοποιήσεων, που παράγονται από τον κώδικα υλοποιείται η ενσωμάτωση των ορίων των τεκτονικών πλακών στους χάρτες της σεισμικής δραστηριότητας (βλ. Παράρτημα Α, Συνάρτηση προσθήκης τεκτονικών πλακών στον χάρτη). Για τον σκοπό αυτό, χρησιμοποιήθηκε ένα ξεχωριστό σύνολο δεδομένων, που περιέχει τα γεωγραφικά δεδομένα των ορίων των πλακών. Το dataset αυτό αποτελείται από 12.321 εγγραφές. Περιλαμβάνει τρεις στήλες: plate, lat, και lon, οι οποίες αντιστοιχούν στην ονομασία της τεκτονικής πλάκας και στις γεωγραφικές συντεταγμένες (γεωγραφικό πλάτος και μήκος) που ορίζουν τις γραμμές των ορίων της. Το συγκεκριμένο πρόσθετο dataset παρέχεται από το Kaggle [11].

Η ενσωμάτωση αυτών των δεδομένων στους παραγόμενους χάρτες είναι ζωτικής σημασίας για τους ακόλουθους λόγους:

- **Γεωδυναμική Ερμηνεία:** Η συντριπτική πλειονότητα των σεισμών στον πλανήτη συμβαίνει στα όρια των τεκτονικών πλακών, όπου συγκεντρώνονται οι τάσεις λόγω της σχετικής τους κίνησης (σύγκλιση, απόκλιση ή οριζόντια ολίσθηση). Η ταυτόχρονη απεικόνιση των σεισμών και των ορίων των πλακών επιτρέπει την άμεση οπτική συσχέτιση των σεισμικών γεγονότων με τις περιοχές, όπου οι γεωλογικές δυνάμεις είναι πιο ενεργές.

- **Αναγνώριση Σεισμογενών Ζωνών:** Μέσω αυτής της οπτικοποίησης, γίνεται ευκολότερη η αναγνώριση και η επιβεβαίωση των κύριων σεισμογενών ζωνών στην περιοχή μελέτης. Για παράδειγμα, στον ελλαδικό χώρο η επικάλυψη των σεισμών με το Ελληνικό Τόξο (όπου η Αφρικανική πλάκα καταβυθίζεται κάτω από την Ευρασιατική) αναδεικνύεται με σαφήνεια, εξηγώντας τη φύση της εντατικής σεισμικής δραστηριότητας.

- **Πρόσθετο Επίπεδο Πληροφορίας:** Η προσθήκη των ορίων των πλακών λειτουργεί ως ένα κρίσιμο γεωλογικό υπόβαθρο, παρέχοντας ένα πλαίσιο για την ερμηνεία των κατανομών των σεισμών. Μετατρέπει τους απλούς χάρτες σημείων σε ένα δυναμικό εργαλείο για την κατανόηση των θεμελιωδών γεωλογικών διεργασιών, που καθορίζουν τη σεισμικότητα.

Συνεπώς, η συμπερίληψη των δεδομένων των τεκτονικών πλακών στους χάρτες που παράγονται από τον κώδικα δεν είναι απλώς μια αισθητική βελτίωση, αλλά μια αναγκαία προσθήκη που εμπλουτίζει την ανάλυση και προσφέρει βαθύτερη γεωδυναμική κατανόηση των σεισμικών φαινομένων, που εξετάζονται.

1.4 Κίνητρο

Η Ελλάδα κατατάσσεται στις πλέον σεισμογενείς περιοχές της Ευρώπης και παρουσιάζει έντονη και συχνή σεισμική δραστηριότητα. Το γεγονός αυτό καθιστά τη μελέτη των σεισμών ιδιαίτερα ζωτικής σημασίας. Η κατανόηση της σεισμικής συμπεριφοράς και η δυνατότητα έγκαιρης και έγκυρης εκτίμησης σεισμικών φαινομένων αποτελούν διαχρονικά ζητήματα μεγάλου επιστημονικού και φυσικά, κοινωνικού ενδιαφέροντος. Η ανοιχτή διάθεση σεισμολογικών δεδομένων (Open Data) από

[11] `https://www.kaggle.com/datasets/cwthompson/tectonic-plate-boundaries`

το Γεωδυναμικό Ινστιτούτο του Εθνικού Αστεροσκοπείου Αθηνών προσφέρει μια μοναδική ευκαιρία για ανάλυση της σεισμικότητας του ελληνικού χώρου, μέσω σύγχρονων τεχνικών Μηχανικής Μάθησης.

Αν και στο παρελθόν έχουν εκπονηθεί διπλωματικές εργασίες, που αξιοποιούν δεδομένα σεισμών, η πλειονότητά τους βασίστηκε σε εργαλεία όπως το WEKA, εστιάζοντας κυρίως σε τεχνικές Κατηγοριοποίησης (Classification) ή βασικής Συσταδοποίησης (Clustering). Μέχρι σήμερα δεν έχει παρουσιαστεί αντίστοιχη εργασία, που να εφαρμόζει σύγχρονες μεθόδους Μη Εποπτευόμενης Μάθησης με χρήση Python, ειδικά στο συγκεκριμένο σύνολο δεδομένων (dataset) του Γεωδυναμικού Ινστιτούτου.

Ο κυρίαρχος στόχος της παρούσας εργασίας είναι η συμβολή της στην κατανόηση της χωρικής και χρονικής κατανομής, των σεισμών στον ελληνικό χώρο. Επιπροσθέτως, επιχειρείται η ανάδειξη της προστιθέμενης αξίας της Μηχανικής Μάθησης και συγκεκριμένα της Μη Εποπτευόμενης Μάθησης στη σεισμολογική έρευνα.

1.5 Συνεισφορά

Στο πλαίσιο της παρούσας μελέτης, υλοποιείται ανάλυση σεισμολογικών δεδομένων με χρήση της γλώσσας Python. Ο κατάλογος των σεισμών (dataset) αντλείται από το Γεωδυναμικό Ινστιτούτο Αθηνών. Πραγματοποιείται ο απαραίτητος έλεγχος και προεπεξεργασία των δεδομένων, έτσι ώστε να καταστούν κατάλληλα προς χρήση. Στο αμέσως επόμενο στάδιο, λαμβάνονται στατιστικά στοιχεία και παρουσιάζεται με αυτόν τον τρόπο η αρχική πληροφορία, που μπορεί να δοθεί από το σύνολο δεδομένων (dataset).

Ακολούθως, αξιοποιούνται σύγχρονες βιβλιοθήκες Μηχανικής Μάθησης και οπτικοποίησης των αποτελεσμάτων. Εφαρμόζονται μέθοδοι Μη Εποπτευόμενης Μάθησης, με έμφαση στην Συσταδοποίηση (Clustering). Συγκεκριμένα, υλοποιούνται και συγκρίνονται οι αλγόριθμοι K-Means, DBSCAN, OPTICS και HDBSCAN, με στόχο τον εντοπισμό ομάδων (clusters) σεισμικών γεγονότων. Μεγάλη βαρύτητα δίνεται επίσης, στην οπτικοποίηση των αποτελεσμάτων. Οι οπτικοποιήσεις αναπαρίστανται με γραφήματα και τρισδιάστατες απεικονίσεις. Με αυτό τον τρόπο, είναι εφικτή η αξιολόγηση της πληροφορίας, που παρέχουν οι αλγόριθμοι. Επιπλέον, δημιουργούνται διαδραστικοί χάρτες, στους οποίους ενσωματώνονται τα όρια των τεκτονικών πλακών. Έτσι, επιτυγχάνεται μια πληρέστερη και γεωλογικά τεκμηριωμένη ερμηνεία των αποτελεσμάτων.

Η εργασία προσφέρει, ένα συγκεκριμένο πλαίσιο ανάλυσης σεισμολογικών δεδομένων. Το πλαίσιο αυτό μπορεί να αποτελέσει βάση για μελλοντικές επεκτάσεις, όπως για παράδειγμα: εντοπισμός χωροχρονικών μοτίβων και συσχέτιση με γεωλογικά ρήγματα.

1.6 Οργάνωση Διπλωματικής Εργασίας

Η παρούσα Διπλωματική Εργασία οργανώνεται σε έξι κεφάλαια, τα οποία καλύπτουν τόσο τη θεωρητική ανάλυση, όσο και την πρακτική εφαρμογή των μεθόδων Μη Εποπτευόμενης Μάθησης

στις σεισμολογικές καταγραφές. Στο πρώτο κεφάλαιο, γίνεται μια θεωρητική προσέγγιση των σεισμών γενικά. Δίδεται ιδιαίτερη έμφαση στον Ελλαδικό χώρο, μιας και αποτελεί το αντικείμενο της μελέτης αυτής. Εξετάζεται η έννοια της συσταδοποίησης και η χρησιμότητά της στην ανάλυση σεισμικών δεδομένων. Παρουσιάζονται, επίσης, το σύνολο δεδομένων (dataset), που προέρχεται από το Γεωδυναμικό Ινστιτούτο Αθηνών (κατάλογος των σεισμών), αλλά και ένα συμπληρωματικό σύνολο δεδομένων για τα όρια των Τεκτονικών Πλακών.

Το δεύτερο κεφάλαιο είναι αφιερωμένο στη Συσταδοποίηση. Ξεκινά με μια εισαγωγή στους βασικούς αλγορίθμους, κατηγοριοποιώντας τους σε: βασιζόμενους σε Διαμέριση (Partition-based), Ιεραρχικούς (Hierarchical) και βασιζόμενους σε Πυκνότητα (Density-based). Ακολουθεί λεπτομερής ανάλυση σε θεωρητικό επίπεδο δημοφιλών αλγορίθμων. Αρχικά, αναλύεται ο K-Means, καθώς και η μέθοδος εύρεσης της βέλτιστης τιμής k, μέσω της μεθόδου του αγκώνα (Elbow Method). Στη συνέχεια, παρουσιάζονται οι αλγόριθμοι συσταδοποίησης DBSCAN, HDBSCAN και OPTICS. Για τον αλγόριθμο DBSCAN, γίνεται επίσης αναφορά στη χρήση της μεθόδου των k-Πλησιέστερων γειτόνων (k-Nearest Neighbors) για τη βέλτιστη επιλογή των παραμέτρων ε και $MinPts$. Στην περίπτωση του OPTICS, εξετάζεται το γράφημα προσβασιμότητας (Reachability Plot), το οποίο προσφέρει χρήσιμες ενδείξεις για τη δομή των συστάδων. Επιπλέον, παρουσιάζονται μέθοδοι αξιολόγησης της ποιότητας της συσταδοποίησης, όπως ο Δείκτης Silhouette, ο Δείκτης Davies–Bouldin και ο Δείκτης Calinski–Harabasz.

Στο τρίτο κεφάλαιο αναλύονται οι βασικές τεχνολογίες, που χρησιμοποιούνται για την υλοποίηση της εργασίας. Αρχικά, παρουσιάζεται με συνοπτικό τρόπο η γλώσσα προγραμματισμού Python, καθώς και οι κυριότερες βιβλιοθήκες, που υιοθετούνται. Συγκεκριμένα αναφέρονται: η βιβλιοθήκη Scikit-learn, η βιβλιοθήκη Folium για τη δημιουργία διαδραστικών χαρτών κ.τ.λ. Εν συνεχεία, γίνεται αναφορά στο Google Colab, που αποτελεί το περιβάλλον ανάπτυξης και εκτέλεσης του κώδικα. Στο κεφάλαιο αυτό περιλαμβάνονται και άλλες χρήσιμες βιβλιοθήκες, που υποστηρίζουν τη διαδικασία ανάλυσης. Τέλος, γίνεται μια συνοπτική παρουσίαση του λογισμικού WEKA.

Το τέταρτο κεφάλαιο επικεντρώνεται στη στατιστική ανάλυση των δεδομένων. Παρουσιάζεται με διεξοδικό τρόπο η Διερευνητική Ανάλυση των Δεδομένων (Exploratory Data Analysis - EDA). Η Διερευνητική Ανάλυση Δεδομένων (EDA) χρησιμοποιείται για τη σύνοψη των κύριων χαρακτηριστικών ενός συνόλου δεδομένων με διάφορες οπτικοποιήσεις [9]. Επίσης, στο κεφάλαιο αυτό, αναλύεται στατιστικώς και η πρόσφατη σεισμική δραστηριότητα της Σαντορίνης.

Στο πέμπτο κεφάλαιο παρουσιάζονται τα πειράματα, που εφαρμόζονται, αξιοποιώντας τον διαμεριστικό (Partition-based) αλγόριθμο K-Means και τους βασιζόμενους σε πυκνότητα (Density-based) αλγορίθμους DBSCAN, HDBSCAN και OPTICS. Τα δεδομένα αναλύονται τόσο σε δύο (2D) όσο και σε πέντε (5D) διαστάσεις (K-Means και DBSCAN). Επιπλέον, γίνεται εφαρμογή τεχνικών πρόβλεψης χρονοσειρών, με βάση τα αποτελέσματα της συσταδοποίησης, αξιοποιώντας το λογισμικό WEKA..

Στο έκτο και τελευταίο κεφάλαιο, αναφέρονται τα κύρια συμπεράσματα που προκύπτουν από την ανάλυση των σεισμολογικών δεδομένων, καθώς και οι προτάσεις για μελλοντικές βελτιώσεις και επεκτάσεις. Η Διπλωματική Εργασία ολοκληρώνεται με το Παράρτημα Α, στο οποίο ενσωματώνεται ο συνολικός κώδικας και στο τέλος παρατίθεται η Βιβλιογραφία.

Κεφάλαιο 2

Συσταδοποίηση

Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD) αποτελεί μια αυτοματοποιημένη διαδικασία αναζήτησης μέσα σε μεγάλους όγκους δεδομένων, με σκοπό τον εντοπισμό υποκείμενων μοτίβων και σχέσεων, τα οποία μπορούν να ερμηνευθούν ως χρήσιμη και αξιοποιήσιμη γνώση. Ουσιαστικά, πρόκειται για τη διαδικασία εξαγωγής γνώσης από τα αρχικά δεδομένα, καθιστώντας τα πιο κατανοητά και λειτουργικά για τη λήψη αποφάσεων.

Ο τομέας της Ανακάλυψης Γνώσης αναδύθηκε από την Εξόρυξη Δεδομένων (Data Mining) και παραμένει στενά συνδεδεμένος με αυτήν. Συγγενής έννοια αποτελεί και η Μηχανική Μάθηση (Machine Learning), η οποία ωστόσο συνιστά υποσύνολο της KDD. Η Μηχανική Μάθηση εστιάζει στην ανάπτυξη αλγορίθμων και συστημάτων ικανών να "μαθαίνουν" από δεδομένα και να βελτιώνονται με την εμπειρία [10]. Αντιθέτως, η KDD περιλαμβάνει ένα ευρύτερο φάσμα σταδίων, που εκτείνεται από την προεπεξεργασία και επιλογή των δεδομένων, έως την εξόρυξη μοτίβων και την ερμηνεία ή αξιολόγηση των αποτελεσμάτων. Με τον τρόπο αυτό, η KDD αξιοποιεί τεχνικές της Μηχανικής Μάθησης για την αναγνώριση γνώσης, εντάσσοντας τες σε ένα συνολικό και οργανωμένο πλαίσιο αναλυτικής διαδικασίας.

2.1 Εισαγωγή στη Συσταδοποίηση

Η Μη Εποπτευόμενη Μάθηση (Unsupervised Learning) είναι μια προσέγγιση της Μηχανικής Μάθησης, κατά την οποία ένας αλγόριθμος ανακαλύπτει δομές και μοτίβα σε δεδομένα χωρίς να υπάρχουν ετικέτες ή εξαρτημένες μεταβλητές. Ουσιαστικά, ο μαθησιακός μηχανισμός καλείται να εξάγει μια χρήσιμη αναπαράσταση, να αποκαλύψει σχέσεις και μοτίβα στα δεδομένα, δίχως να υπάρχουν προκαθορισμένες έξοδοι [11]. Θεωρείται μάθηση μέσω παρατήρησης, όχι μέσω παραδειγμάτων.

Στην πράξη, η Μη Εποπτευόμενη Μάθηση είναι συχνά συνώνυμη με την συσταδοποίηση (clustering) [12]. Στόχος της είναι η ανακάλυψη εγγενών κλάσεων ή δομών μέσα στα δεδομένα, όταν οι ετικέτες αυτών των κλάσεων είναι άγνωστες, και ακόμη και ο αριθμός τους μπορεί να μην είναι προκαθορισμένος. Η διαδικασία βασίζεται αποκλειστικά στα χαρακτηριστικά των ίδιων των δεδομένων: τα δεδομένα είναι, σε μεγάλο βαθμό, παρόμοια μεταξύ τους και συγκεντρώνονται στην ίδια συστάδα, ενώ ταυτόχρονα είναι διαφορετικά - όσο γίνεται περισσότερο - από δεδομένα που ανήκουν σε διαφορετικές συστάδες.

2.1.1 Αλγόριθμοι βασιζόμενοι σε Διαμέριση (Partition - based)

Οι αλγόριθμοι βασιζόμενοι σε Διαμέριση είναι μια μέθοδος Μη Εποπτευόμενης Μάθησης, η οποία αποσκοπεί στην ανακάλυψη συστάδων εντός ενός συνόλου δεδομένων, μέσω της βελτιστοποίησης μιας αντικειμενικής συνάρτησης. Η προσέγγιση αυτή βασίζεται στην ομαδοποίηση των δεδομένων σε *k* συστάδες, με τον επιθυμητό αριθμό *k* να καθορίζεται εκ των προτέρων από τον χρήστη. Δεδομένου ενός συνόλου *n* αντικειμένων, δημιουργούνται *k* διαμερίσεις, όπου κάθε διαμέριση αντιστοιχεί σε μία συστάδα και ισχύει ότι $k \leq n$, με την κάθε συστάδα να περιέχει τουλάχιστον ένα αντικείμενο. Συνήθως, υιοθετείται η αποκλειστική εκχώρηση των αντικειμένων στις συστάδες, δηλαδή κάθε αντικείμενο ανήκει σε μία μόνο ομάδα [12]. Η διαδικασία ξεκινά από μια αρχική διαμέριση και βελτιώνεται επαναληπτικά, μετακινώντας αντικείμενα από τη μία συστάδα στην άλλη, με σκοπό τη βελτίωση της συνολικής ποιότητας της συσταδοποίησης.

Μια από τις πιο συνηθισμένες και διαδεδομένες αντικειμενικές συναρτήσεις, που χρησιμοποιούνται είναι το Άθροισμα των Τετραγωνικών Σφαλμάτων (Sum of Squared Errors - SSE) [13]. Ο στόχος είναι η ελαχιστοποίηση του SSE, δηλαδή του αθροίσματος των τετραγωνικών αποστάσεων μεταξύ των αντικειμένων και των αντίστοιχων κέντρων των συστάδων τους, διασφαλίζοντας έτσι, τη μέγιστη ομοιότητα (intraclass similarity) εντός κάθε συστάδας και την ελάχιστη ομοιότητα (interclass similarity) μεταξύ διαφορετικών συστάδων. Οι περισσότερες μέθοδοι διαμεριστικής συσταδοποίησης βασίζονται στον υπολογισμό αποστάσεων μεταξύ των αντικειμένων.

Παρά το γεγονός ότι η εύρεση της ολικής βέλτιστης λύσης είναι υπολογιστικά δύσκολη, οι διαμεριστικοί αλγόριθμοι παραμένουν ιδιαίτερα δημοφιλείς, χάρη στην απλότητα, την υπολογιστική αποδοτικότητα και την ευκολία υλοποίησής τους [9]. Για τον λόγο αυτό, εφαρμόζονται ευρετικές προσεγγίσεις, όπως οι άπληστοι αλγόριθμοι K-Means και K-Medoids, οι οποίοι στοχεύουν στην προσέγγιση ενός τοπικού βέλτιστου λύσης.

2.1.2 Ιεραρχικοί Αλγόριθμοι (Hierarchical)

Η Ιεραρχικοί αλγόριθμοι αποτελούν μια μέθοδο Μη Εποπτευόμενης Μάθησης, η οποία δημιουργεί μια ιεραρχική δομή συστάδων, χωρίς να απαιτείται εκ των προτέρων ο καθορισμός του αριθμού *k* των συστάδων. Η διαδικασία αυτή μπορεί να απεικονιστεί με τη μορφή δενδρογράμματος (dendrogram), το οποίο αποδομεί ένα σύνολο δεδομένων σε διαδοχικά επίπεδα συστάδων. Η τελική συσταδοποίηση επιτυγχάνεται "κόβοντας" το δενδρόγραμμα στο επιθυμητό επίπεδο ομοιότητας, όπου κάθε συνδεδεμένο υποδέντρο αντιστοιχεί σε μία συστάδα. Υπάρχουν δύο βασικές προσεγγίσεις στους Ιεραρχικούς Αλγορίθμους [13]:

- **Αθροιστική (agglomerative), ή bottom-up προσέγγιση:** Ξεκινά με κάθε αντικείμενο να αποτελεί ξεχωριστή συστάδα και προοδευτικά συγχωνεύει τις πιο όμοιες μεταξύ τους συστάδες μέχρι να προκύψει μία ενιαία συστάδα ή μέχρι να ικανοποιηθεί κάποιο κριτήριο τερματισμού.
- **Διαχωριστική (divisive), ή top-down προσέγγιση:** Ξεκινά με όλα τα αντικείμενα ενωμένα σε μία μακροσυστάδα και διαχωρίζει επαναληπτικά τις συστάδες σε μικρότερες, μέχρι να επιτευχθεί η επιθυμητή διάσπαση ή κάθε αντικείμενο να αποτελέσει ξεχωριστή συστάδα.

Η Ιεραρχική συσταδοποίηση υποστηρίζει την ύπαρξη υποσυστάδων μέσα σε συστάδες, οδηγώντας στη δημιουργία μιας δενδροειδούς δομής, που αντανακλά τις σχέσεις ομοιότητας σε πολλαπλά επίπεδα. Η διαδικασία συσταδοποίησης βασίζεται στη σχετική εγγύτητα μεταξύ των αντικειμένων, η οποία υπολογίζεται με χρήση διαφόρων μετρικών αποστάσεων, όπως για παράδειγμα η Ευκλείδεια απόσταση. Η απόφαση για συγχώνευση ή διαχωρισμό συστάδων καθορίζεται από κάποιο μέτρο ομοιότητας ή απόστασης, το οποίο επιλέγεται ώστε να βελτιστοποιεί ένα συγκεκριμένο κριτήριο, όπως το άθροισμα των τετραγωνικών σφαλμάτων [9].

Ένα βασικό μειονέκτημα των ιεραρχικών μεθόδων είναι η μη αναστρεψιμότητα των αποφάσεων: μόλις πραγματοποιηθεί μια συγχώνευση ή ένας διαχωρισμός, δεν μπορεί να ανακληθεί. Ωστόσο, αυτή η ιδιότητα συμβάλλει στη μείωση του υπολογιστικού κόστους, καθώς περιορίζει τον αριθμό των ενδεχόμενων εναλλακτικών διαμερίσεων που πρέπει να εξεταστούν [12]. Αντιπροσωπευτικοί Ιεραρχικοί αλγόριθμοι είναι οι: BIRCH, CURE και CHAMELEON.

2.1.3 Αλγόριθμοι βασιζόμενοι σε Πυκνότητα (Density-based)

Οι αλγόριθμοι που βασίζονται στην Πυκνότητα αποτελούν μια ισχυρή κατηγορία αλγορίθμων Μη Εποπτευόμενης Μάθησης, οι οποίοι είναι ιδιαίτερα αποτελεσματικοί στην ανακάλυψη μη γραμμικών δομών και συστάδων με αυθαίρετο σχήμα [9]. Η βασική ιδέα είναι, ότι οι συστάδες αντιστοιχούν σε περιοχές υψηλής πυκνότητας σημείων, οι οποίες διαχωρίζονται από περιοχές χαμηλής πυκνότητας.

Η διαδικασία βασίζεται στην έννοια της πυκνότητας των σημείων, η οποία καθορίζεται από τον αριθμό των γειτονικών αντικειμένων εντός μιας δεδομένης ακτίνας. Μια συστάδα συνεχίζει να αναπτύσσεται, όσο η πυκνότητα στη γειτονιά της υπερβαίνει ένα προκαθορισμένο κατώφλι. Αυτή η προσέγγιση επιτρέπει την αποτελεσματική διαχείριση του θορύβου (noise) και των ακραίων τιμών, καθώς τα σημεία που βρίσκονται σε αραιές περιοχές, ή δεν ανήκουν σε καμία πυκνή συστάδα χαρακτηρίζονται ως θόρυβος [12]. Επιπλέον, δεν απαιτείται εκ των προτέρων ορισμός του αριθμού των συστάδων, γεγονός που καθιστά τη μέθοδο περισσότερο ευέλικτη και αυτόνομη.

Τέλος, οι αλγόριθμοι συσταδοποίησης βάσει πυκνότητας διαχωρίζουν φυσικά τα δεδομένα σε συστάδες και σημεία θορύβου, κάτι που δεν είναι εφικτό με άλλες τεχνικές. Έτσι, θεωρούνται ιδιαίτερα αποτελεσματικοί σε εφαρμογές που απαιτούν ανίχνευση σύνθετων δομών και αντοχή σε θόρυβο. Αντιπροσωπευτικοί αλγόριθμοι αυτής της κατηγορίας είναι οι DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points To Identify the Clustering Structure) και DENCLUE (DENsity-based CLUstEring).

2.2 K-Means

Ο Αλγόριθμος K-Means είναι ένας από τους πιο γνωστούς και ευρέως χρησιμοποιούμενους αλγορίθμους διαμέρισης (Partition-based). Ο Stuart Lloyd πρότεινε τον πρότυπο αλγόριθμο το 1957, γι' αυτό και ονομάζεται συχνά ως ο αλγόριθμος του Lloyd. Ο K-Means χρησιμοποιεί πρωτότυπα (prototypes), για να αναπαραστήσει τις συστάδες, που δημιουργεί. Το πρωτότυπο ή κεντρικό σημείο (centroid) είναι συνήθως ο μέσος μιας ομάδας σημείων. Αρχικά, γίνεται μια τυχαία κατανομή

των σημείων σε διακριτές συστάδες, που δεν επικαλύπτονται μεταξύ τους και επιλέγονται *k* κεντρικά σημεία (cendroids). Το *k* είναι μια τιμή, που καθορίζεται από τον χρήστη, εκ των προτέρων. Η παράμετρος αυτή αντιπροσωπεύει τον επιθυμητό αριθμό συστάδων, που θα δημιουργηθούν.

Ο αλγόριθμος αξιολογεί την ποιότητα της διαμέρισης χρησιμοποιώντας μια συνάρτηση στόχου. Τα στοιχεία, που ανήκουν στην ίδια συστάδα πρέπει να παρουσιάζουν υψηλή ομοιότητα. Επιπλέον, τα στοιχεία, που ανήκουν σε διαφορετικές συστάδες πρέπει να διαφέρουν αισθητά μεταξύ τους [12]. Η ιδέα του αλγορίθμου αφορά, ουσιαστικά, μια επαναληπτική διαδικασία. Κάθε σημείο τοποθετείται στη συστάδα με το πλησιέστερο κεντροειδές (cendroid). Τα σημεία, που έχουν εκχωρηθεί στο ίδιο κεντροειδές, σχηματίζουν μια συστάδα. Ο σκοπός, σε αυτό το σημείο, είναι η επιλογή και ανάθεση των σημείων σε ομάδες να γίνεται με τρόπο, ώστε να ελαχιστοποιείται το άθροισμα των τετραγωνικών αποστάσεων, εντός των συστάδων (WCSS - Within-Cluster Sum-of-Squares criterion). Για την Σχέση 2.1 ισχύει ότι: *n* είναι το σύνολο δειγμάτων *x* και μ_j ο μέσος όρος των σημείων της συστάδας C. Το κριτήριο WCSS καλείται και inertia (αδράνεια) και αποτελεί μία ποσότητα, που αξιολογεί την ποιότητα της συσταδοποίησης K-Means[1].

$$WCSS = \sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2) \tag{2.1}$$

Ακολούθως, τα κεντρικά σημεία (cendroids) κάθε συστάδας επαναπροσδιορίζονται, ως ο μέσος όρος των σημείων, που ανήκουν σ' αυτήν. Η διαδικασία αυτή επαναλαμβάνεται μέχρι τα σημεία να μην αλλάζουν πλέον συστάδα ή μέχρι τα κεντροειδή να παραμείνουν αμετάβλητα. Η παραπάνω διαδικασία αποτυπώνεται στον παρακάτω Αλγόριθμο 1 [11], [14].

Αλγόριθμος 1 K-Means

1: Επιλογή k αρχικών κεντρικών σημείων (centroids) των συστάδων.
2: **Επανάληψη**:
3: Ανάθεση κάθε σημείου στο πλησιέστερο κεντρικό σημείο της συστάδας.
4: Αναπροσαρμογή του κεντρικού σημείου κάθε συστάδας.
5: **Μέχρι** σύγκλιση

Ο αλγόριθμος K-Means αποτελεί μια ιδιαίτερα δημοφιλή επιλογή για προβλήματα συσταδοποίησης σε διάφορους τομείς εφαρμογών. Η δημοφιλία του οφείλεται κυρίως, στην απλότητα της υλοποίησής του και στο χαμηλό υπολογιστικό κόστος, που απαιτεί. Είναι ιδιαίτερα αποτελεσματικός, όταν πρόκειται για συστάδες με σφαιρικό ή κυκλικό σχήμα και μπορεί να εφαρμοστεί ακόμη και σε πιο σύνθετες μορφές δεδομένων, όπως κείμενα ή χρονοσειρές. Ο αλγόριθμος οδηγείται πάντοτε σε κάποια λύση, και η σύγκλιση, συνήθως, επιτυγχάνεται ήδη από τα πρώτα στάδια της διαδικασίας. Ωστόσο, όπως συμβαίνει με κάθε μεθοδολογική προσέγγιση, έτσι και ο K-Means συνοδεύεται από ορισμένα εγγενή μειονεκτήματα. Ένα βασικό ζήτημα αφορά την ανάγκη καθορισμού, εκ των προτέρων, του αριθμού των συστάδων, κάτι που μπορεί να επηρεάσει την ποιότητα της ανάλυσης, αν δεν επιλεγεί κατάλληλα. Επιπλέον, είναι ευαίσθητος στην αρχική επιλογή των κεντροειδών, γεγονός που μπορεί να οδηγήσει σε τοπικά και όχι σε καθολικά βέλτιστα αποτελέσματα. Ο αλγόριθμος,

[1] `https://scikit-learn.org/stable/modules/clustering.html#k-means`

επίσης, δυσκολεύεται να εντοπίσει συστάδες με μη σφαιρικά σχήματα ή με ανομοιόμορφες πυκνότητες, ενώ επηρεάζεται ιδιαίτερα από την παρουσία θορύβου ή μεμονωμένων ακραίων τιμών (outliers) στα δεδομένα [13].

Η υπολογιστική πολυπλοκότητα του K-Means εξαρτάται κυρίως από το πλήθος των σημείων δεδομένων *n*, τον αριθμό των συστάδων *k*, τη διάσταση *d*, αλλά και από τον αριθμό των επαναλήψεων *i*, έως ότου επιτευχθεί σύγκλιση. Στη βασική του μορφή, η πολυπλοκότητα ανά επανάληψη είναι *O(nkd)*. Κάθε σημείο συγκρίνεται με κάθε κεντροειδές, για να υπολογιστούν οι αποστάσεις και τελικά να ανατεθεί στη σωστή συστάδα. Εφόσον, η διαδικασία επαναλαμβάνεται, έως ότου τα κεντροειδή σταθεροποιηθούν, ο συνολικός χρόνος εκτέλεσης είναι *O(nkdi)*. Παρά τη θεωρητικά υψηλή πολυπλοκότητα, στην πράξη ο K-Means είναι αρκετά αποδοτικός, καθώς συγκλίνει γρήγορα σε λίγες επαναλήψεις, ειδικά σε καλά διαχωρισμένα δεδομένα. Ωστόσο, σε πολύ μεγάλα σύνολα δεδομένων ή όταν υπάρχουν πολλές διαστάσεις, η αποδοτικότητα μπορεί να μειωθεί, κάτι που έχει οδηγήσει στην ανάπτυξη βελτιστοποιημένων εκδόσεων όπως K-Means++ για καλύτερη αρχικοποίηση [14].

2.2.1 Επιλογή Βέλτιστου Αριθμού Συστάδων του Αλγορίθμου K-Means

Η επιλογή της σωστής τιμής της παραμέτρου *k*, στον K-Means για παράδειγμα, αποτελεί μια πρόκληση, καθώς απαιτεί την εξέταση διαφόρων μεθόδων και την κατανόηση των εγγενών χαρακτηριστικών των δεδομένων. Κάθε μετρική αξιολόγησης οφείλει να εξετάζει και να συγκρίνει, αν η συσταδοποίηση που προκύπτει:

- διαχωρίζει τα δεδομένα με τρόπο αντίστοιχο με κάποια υπάρχουσα κατηγοριοποίηση (αν υπάρχει),
- ικανοποιεί την αρχή ότι τα σημεία μιας συγκεκριμένης συστάδας εμφανίζουν μεγάλη ομοιότητα μεταξύ τους και τα σημεία που ανήκουν σε διαφορετικές συστάδες διαφέρουν μεταξύ τους σε υψηλό βαθμό.

Ο αλγόριθμος K-Means έχει ως προϋπόθεση την παράμετρο *k*, που ορίζεται από τον χρήστη. Η επιλογή της ιδανικής τιμής της παραμέτρου είναι κρίσιμη για την αποτελεσματική ανάλυση των συστάδων.

Η Μέθοδος του Αγκώνα (Elbow Method): αποτελεί μια δημοφιλή, αλλά όχι πάντοτε αξιόπιστη μετρική για τον προσδιορισμό του ιδανικού αριθμού συστάδων *k*. Η επιλογή της βέλτιστης τιμής *k* εμπεριέχει αναπόφευκτα ένα βαθμό υποκειμενικότητας, καθώς βασίζεται στο επίπεδο λεπτομέρειας, που επιθυμεί ο χρήστης. Επίσης, εξαρτάται από παράγοντες, όπως η κατανομή των δεδομένων. Η μέθοδος αυτή βασίζεται στην παρακολούθηση της μεταβολής της ενδοσυσταδικής διακύμανσης, όσο αυξάνεται ο αριθμός των συστάδων (παράμετρος *k*). Πιο συγκεκριμένα, κατασκευάζεται ένα γράφημα, όπου στον οριζόντιο άξονα αποτυπώνεται η τιμή του *k* και στον κατακόρυφο το άθροισμα των τετραγώνων των αποστάσεων (Sum of Squared Errors - SSE ή WCSS) των σημείων από τα αντίστοιχα κεντροειδή τους. Όπως φαίνεται και στο (Σχήμα 5.1) στην αρχή της καμπύλης παρατηρείται έντονη μείωση του σφάλματος (ή inertia), ωστόσο από ένα σημείο και έπειτα η μείωση αυτή επιβραδύνεται, δημιουργώντας μια χαρακτηριστική γωνία στο γράφημα - τον αποκαλούμενο "αγκώνα" [12].

2.3 DBSCAN

Ο DBSCAN (Density-Based Spatial Clustering of Applications with Noise)[2]είναι ένας αλγόριθμος συσταδοποίησης, που προτάθηκε το 1996 από τους M. Ester, H.-P. Kriegel, J. Sander και X. Xu. Ενδεικτικά πεδία εφαρμογής του συγκεκριμένου αλγορίθμου περιλαμβάνουν την ανάλυση γεωχωρικών δεδομένων, την επεξεργασία εικόνας και βίντεο, τον εντοπισμό ανωμαλιών (anomaly detection), χρονοσειρές και δεδομένα αισθητήρων κ.α. Αποτελεί μια ισχυρή μέθοδο Μη Εποπτευόμενης Μάθησης, που ανακαλύπτει συστάδες με βάση την πυκνότητα των σημείων δεδομένων. Η βασική του αρχή είναι να διαχωρίζει περιοχές υψηλής πυκνότητας από περιοχές χαμηλής πυκνότητας [15], [16]. Ο DBSCAN χρησιμοποιεί δύο βασικές παραμέτρους:

- ***ε (epsilon):*** η ακτίνα της γειτονιάς γύρω από ένα σημείο,
- ***MinPts:*** ο ελάχιστος αριθμός σημείων που απαιτούνται εντός της *ε-γειτονιάς* , για να θεωρηθεί ένα σημείο ως σημείο πυρήνα (core point). Διευκρινίζεται ότι *ε-γειτονιά* ενός σημείου p είναι το σύνολο των σημείων, που βρίσκονται σε απόσταση μικρότερη ή ίση από ε από το σημείο p.

Με βάση αυτές τις παραμέτρους, τα σημεία δεδομένων κατηγοριοποιούνται ως:

- **Σημεία Πυρήνα (Core points):** Σημεία που έχουν τουλάχιστον *MinPts* γείτονες εντός της ακτίνας ε. Αυτά αποτελούν τους "πυλώνες" των πυκνών περιοχών.
- **Σημεία Ορίου (Border points):** Σημεία που δεν είναι σημεία πυρήνα, αλλά βρίσκονται εντός της *ε-γειτονιάς* ενός σημείου πυρήνα.
- **Σημεία Θορύβου (Noise points):** Σημεία που δεν είναι ούτε σημεία πυρήνα, ούτε σημεία ορίου. Αυτά παραλείπονται από τη συσταδοποίηση.

Ο αλγόριθμος DBSCAN ακολουθεί μια δομημένη διαδικασία για τον εντοπισμό και τον σχηματισμό συστάδων, βασιζόμενος στις έννοιες των σημείων πυρήνα, ορίου και θορύβου. Αρχικά, ο DBSCAN κατηγοριοποιεί κάθε σημείο του συνόλου δεδομένων ως σημείο πυρήνα (core point), σημείο ορίου (border point), ή σημείο θορύβου (noise point). Αυτή η αρχική ταξινόμηση γίνεται με βάση τις παραμέτρους ε (μέγιστη ακτίνα γειτονιάς) και *MinPts* (ελάχιστος αριθμός σημείων στη γειτονιά).

Στη συνέχεια, απορρίπτονται τα σημεία που έχουν χαρακτηριστεί ως θόρυβος, καθώς δεν αποτελούν μέρος καμίας πυκνής περιοχής. Άρα, όπως είναι κατανοητό ο DBSCAN δεν παράγει μια πλήρη συσταδοποίηση, καθώς ορισμένα σημεία δεν ανήκουν σε καμία συστάδα. Έπειτα, ο αλγόριθμος εστιάζει στα σημεία πυρήνα. Δημιουργούνται ακμές (συνδέσεις) μεταξύ όλων των σημείων πυρήνα, που βρίσκονται σε απόσταση μικρότερη ή ίση του ε μεταξύ τους.

Ακολούθως, δημιουργούνται οι συστάδες. Κάθε ομάδα συνδεδεμένων σημείων πυρήνα σχηματίζει μια ξεχωριστή συστάδα. Τέλος, τα σημεία ορίου αντιστοιχίζονται σε αυτές τις συστάδες. Ένα σημείο ορίου ενσωματώνεται στη συστάδα του σημείου πυρήνα (ή ενός εκ των σημείων πυρήνα), στο οποίο είναι αρκετά κοντά (εντός απόστασης ε). Σε περίπτωση που ένα σημείο ορίου βρίσκεται

[2] `https://scikit-learn.org/stable/modules/clustering.html#dbscan`

κοντά σε σημεία πυρήνα από διαφορετικές συστάδες, ενδέχεται να χρειαστεί επίλυση διαφορών για την τελική του ανάθεση. Η παραπάνω διαδικασία αποτυπώνεται στον παρακάτω Αλγόριθμο 2 [14].

Αλγόριθμος 2 DBSCAN

1: Επισήμανση όλων των σημείων ως σημεία πυρήνα, ορίου ή θορύβου.
2: Εξάλειψη των σημείων θορύβου.
3: Τοποθέτηση μια ακμής μεταξύ όλων των σημείων πυρήνα, που βρίσκονται εντός απόστασης epsilon το ένα από το άλλο.
4: Μετατροπή κάθε ομάδας συνδεδεμένων σημείων πυρήνα σε ξεχωριστή συστάδα.
5: Αντιστοίχιση κάθε σημείου ορίου σε μία από τις συστάδες των σχετιζόμενων σημείων πυρήνα του.

Σε αντίθεση με άλλους αλγορίθμους (όπως ο K-Means, για παράδειγμα), ο DBSCAN δεν απαιτεί να καθοριστεί εκ των προτέρων ο αριθμός των συστάδων. Ένα από τα βασικά πλεονεκτήματα του DBSCAN είναι η ικανότητά του να ανιχνεύει συστάδες αυθαίρετου σχήματος και μεγέθους, καθώς δεν βασίζεται σε γεωμετρικές παραδοχές, αλλά στη χωρική εγγύτητα και την πυκνότητα των σημείων. Επιπλέον, διαχειρίζεται αποτελεσματικά τον θόρυβο και τα απομονωμένα σημεία (outliers), ταξινομώντας τα ως μη συσχετισμένα με κάποια συστάδα [9]. Η απόδοσή του είναι γενικά καλή σε δεδομένα δύο ή τριών διαστάσεων, ειδικά όταν χρησιμοποιείται σε συνδυασμό με δομές χωρικής ευρετηρίασης, όπως τα k-d trees (Unsupervised Nearest Neighbors)[3].

Ωστόσο, ο DBSCAN παρουσιάζει και ορισμένους περιορισμούς. Η αποτελεσματικότητά του εξαρτάται σε μεγάλο βαθμό από τη σωστή επιλογή των παραμέτρων ε και *MinPts*. Λανθασμένες τιμές μπορεί να οδηγήσουν είτε στην αποτυχία σχηματισμού συστάδων, είτε στον χαρακτηρισμό πολλών σημείων ως θόρυβο. Επιπλέον, ο αλγόριθμος αντιμετωπίζει δυσκολίες, όταν οι συστάδες έχουν σημαντικά διαφορετική πυκνότητα, καθώς μία ενιαία τιμή για το *ε (epsilon)* δεν μπορεί να εξυπηρετήσει όλες τις περιπτώσεις επαρκώς. Τέλος, σε περιβάλλοντα υψηλής διαστασιμότητας (high-dimensional data), η απόδοση του DBSCAN υποβαθμίζεται (curse of dimensionality), καθώς οι μετρικές αποστάσεων χάνουν τη διακριτική τους ικανότητα [9], [17].

Εάν χρησιμοποιείται ένας χωρικός δείκτης (spatial index) η υπολογιστική πολυπλοκότητα του αλγορίθμου DBSCAN είναι *O(nlogn)*, όπου *n* είναι ο αριθμός των σημείων της βάσης δεδομένων. Σε διαφορετική περίπτωση, η πολυπλοκότητα είναι O(n^2) [13].

2.3.1 Καθορισμός Παραμέτρων του Αλγορίθμου DBSCAN

Ένα κρίσιμο βήμα για την επιτυχή εφαρμογή του αλγορίθμου είναι ο σωστός καθορισμός των παραμέτρων του: της ακτίνας γειτονιάς (ε) και του ελάχιστου αριθμού σημείων (*MinPts*). Η βέλτιστη επιλογή αυτών των παραμέτρων είναι καθοριστική για την ποιότητα της συσταδοποίησης.

[3] https://scikit-learn.org/stable/modules/neighbors.html#unsupervised-nearest-neighbors

Εύρεση *k*-Πλησιέστερων Γειτόνων: Αρχικά, γίνεται υπολογισμός των *k*-Πλησιέστερων Γειτόνων για όλα τα σημεία του συνόλου δεδομένων, για μια συγκεκριμένη τιμή του *k*. Κατόπιν, πραγματοποιείται ταξινόμηση των υπολογισμένων τιμών των *k*-Πλησιέστερων Γειτόνων σε αύξουσα σειρά και στη συνέχεια παράγεται μια γραφική απεικόνιση. Στο γράφημα των ταξινομημένων τιμών (οριζόντιος άξονας), λογικά, πρέπει να υπάρχει μια απότομη αλλαγή, η οποία αντιστοιχεί σε κάποια τιμή του ε (κατακόρυφος άξονας). Ακολούθως, γίνεται επιλογή της τιμής *k* ως τιμή της παραμέτρου του αλγορίθμου *MinPts* και της τιμής ε, που προκύπτει από το γράφημα.

Σημεία, για τα οποία το αποτέλεσμα του υπολογισμού των *k*-Πλησιέστερων Γειτόνων είναι μικρότερο από το επιλεγμένο ε θα χαρακτηριστούν ως σημεία πυρήνα, ενώ τα υπόλοιπα ως σημεία θορύβου ή ορίου. Θα πρέπει να τονιστεί σε αυτό το σημείο, ότι η τιμή του ε, που προσδιορίζεται με αυτή τη μέθοδο, εξαρτάται από το *k*, αν και δεν αλλάζει δραματικά με μικρές μεταβολές του *k*. Ωστόσο, ο καθορισμός του *k* είναι καίριας σημασίας γιατί [14]:

- Εάν το *k* είναι υπερβολικά μικρό, υπάρχει ο κίνδυνος ακόμη και λίγα σημεία που ενδέχεται να αποτελούν θόρυβο ή ακραίες τιμές να χαρακτηριστούν εσφαλμένα ως συστάδα.
- Εάν το *k* είναι πολύ μεγάλο, τότε μικρού μεγέθους συστάδες (δηλαδή με λιγότερα σημεία από την τιμή του *k*) ενδέχεται να αγνοηθούν και να χαρακτηριστούν ως θόρυβος, οδηγώντας σε απώλεια σημαντικής πληροφορίας.

2.4 HDBSCAN

Ο αλγόριθμος HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)[4]αποτελεί μια βελτιωμένη εκδοχή του γνωστού DBSCAN και έχει σχεδιαστεί, για να υπερβαίνει κάποιους από τους περιορισμούς του, ιδιαίτερα όταν πρόκειται για συστάδες διαφορετικής πυκνότητας. Σε αντίθεση με τον DBSCAN, ο οποίος απαιτεί σταθερή πυκνότητα για όλες τις συστάδες, ο HDBSCAN μπορεί να εντοπίσει συστάδες με διαφορετικά επίπεδα πυκνότητας και να διαχειριστεί με πιο ευέλικτο τρόπο τα σημεία θορύβου. Ο αλγόριθμος στηρίζεται σε ιεραρχική ομαδοποίηση και εφαρμόζει μια διαδικασία εξαγωγής επίπεδων συστάδων από ένα δέντρο πυκνοτήτων, αποφεύγοντας έτσι, την ανάγκη καθορισμού της παραμέτρου *epsilon*, που απαιτεί ο DBSCAN. Οι βασικές παράμετροι που καθορίζουν τη λειτουργία του HDBSCAN είναι το *min_cluster_size*, που ορίζει το ελάχιστο μέγεθος μιας συστάδας και προαιρετικά το *min_samples*, που υπολογίζει την πυκνότητα πυρήνα (Απόσταση Πυρήνα - Core Distance) και ρυθμίζει τον ορισμό της τοπικής πυκνότητας.

Η βασική ιδέα πίσω από τον HDBSCAN ξεκινά με τη μετατροπή του χώρου σε έναν γράφο, στον οποίο κάθε κόμβος αντιπροσωπεύει ένα σημείο δεδομένων και τα βάρη των ακμών βασίζονται σε μία τροποποιημένη έννοια απόστασης, που ονομάζεται Αμοιβαία Προσβάσιμη Απόσταση (Mutual Reachability Distance). Ακολούθως, κατασκευάζεται ένα Ελάχιστο Επικαλυπτικό Δέντρο (Minimum Spanning Tree (MST)), που αποτυπώνει τη βασική δομή πυκνότητας των δεδομένων. Στη συνέχεια, ο αλγόριθμος αφαιρεί ακμές με αυξανόμενα βάρη, παράγοντας μια ιεραρχία συστάδων. Τέλος, εφαρμόζεται μία διαδικασία συμπύκνωσης (condensation) και εξαγωγής (extraction),

[4] `https://scikit-learn.org/stable/modules/clustering.html#hdbscan`

μέσω της οποίας επιλέγονται οι πιο σταθερές συστάδες, με βάση μια έννοια γνωστή ως Σταθερότητα Συστάδων (Cluster Stability), οδηγώντας στη δημιουργία του τελικού επιπέδου συστάδων [18], [19]. Ο HDBSCAN βασίζεται ακριβώς στην ιδέα του πόσο σταθερή είναι μια συστάδα, καθώς μεταβάλλεται το κατώφλι πυκνότητας στην ιεραρχία που δημιουργεί. Μια σταθερή συστάδα είναι αυτή, που διατηρεί τη συνοχή της σε ένα ευρύ φάσμα πυκνοτήτων, χωρίς να διασπάται ή να συγχωνεύεται εύκολα με άλλες.Η παραπάνω διαδικασία αποτυπώνεται στον παρακάτω Αλγόριθμο 3.

Αλγόριθμος 3 HDBSCAN

1: Υπολογισμός της Αμοιβαίας Προσβάσιμης Απόστασης για κάθε ζεύγος σημείων.
2: Κατασκευή του MST, βάσει αυτών των αποστάσεων.
3: Δημιουργία ιεραρχίας συστάδων, μέσω προοδευτικής αφαίρεσης ακμών (κατασκευή dendrogram).
4: Συμπύκνωση της ιεραρχίας με βάση το min_cluster_size και καταγραφή της σταθερότητας κάθε συστάδας.
5: Εξαγωγή των τελικών επιπέδων συστάδων με βάση τη μέγιστη σταθερότητα.

Ένα από τα σημαντικά πλεονεκτήματα του HDBSCAN είναι η ικανότητά του να ανιχνεύει συστάδες με διαφορετική πυκνότητα, καθώς και το ότι μπορεί να εντοπίζει σημεία θορύβου με μεγαλύτερη ακρίβεια, αποφεύγοντας την ανάγκη αυθαίρετου καθορισμού της παραμέτρου *eps*. Επιπλέον, δεν απαιτεί εκ των προτέρων τον αριθμό των συστάδων, ενώ η έννοια της σταθερότητας παρέχει έναν πιο θεμελιωμένο τρόπο αξιολόγησης της ποιότητας των συστάδων [20].

Παρά τα πλεονεκτήματά του, ο HDBSCAN παρουσιάζει και ορισμένες προκλήσεις. Η εφαρμογή του σε πολύ μεγάλους όγκους δεδομένων μπορεί να είναι υπολογιστικά απαιτητική, ενώ η ερμηνεία της ιεραρχίας και της σταθερότητας ενδέχεται να δυσκολεύει την κατανόηση των αποτελεσμάτων για μη ειδικούς χρήστες. Επιπλέον, η σωστή ρύθμιση των παραμέτρων *min_cluster_size* και *min_samples* εξακολουθεί να απαιτεί πειραματισμό και κατανόηση της φύσης των δεδομένων.

Η υπολογιστική πολυπλοκότητα του HDBSCAN εξαρτάται κυρίως από την κατασκευή του Minimum Spanning Tree (MST), η οποία σε γενικές γραμμές, έχει πολυπλοκότητα *O(nlogn)* όταν χρησιμοποιούνται κατάλληλες δομές δεδομένων. Παρά το αυξημένο υπολογιστικό κόστος σε σχέση με τον DBSCAN, ο HDBSCAN προσφέρει πιο εύρωστα και ευέλικτα αποτελέσματα, ειδικά σε δεδομένα με πολύπλοκη δομή ή ετερογένεια πυκνότητας.

2.5 OPTICS

Ο αλγόριθμος OPTICS (Ordering Points To Identify the Clustering Structure)[5]προτάθηκε από τους Ankerst, Breunig, Kriegel και Sander (SIGMOD, 1999) και αποτελεί μια προέκταση του DBSCAN, σχεδιασμένη, για να αντιμετωπίζει έναν από τους βασικούς περιορισμούς του: την ανάγκη ύπαρξης μιας ενιαίας πυκνότητας σε όλες τις συστάδες [13]. Επιπλέον, δεν υπάρχει ανάγκη καθορισμού

[5]https://scikit-learn.org/stable/modules/clustering.html#optics

σταθερών τιμών για τις παραμέτρους ε και *MinPts*. Σε αντίθεση με τον DBSCAN, που επιστρέφει συγκεκριμένες συστάδες για συγκεκριμένες παραμέτρους, ο OPTICS δημιουργεί μια συνολική αναπαράσταση της τοπικής πυκνότητας των δεδομένων, από την οποία μπορούν να εξαχθούν συστάδες μεταβαλλόμενης πυκνότητας. Ουσιαστικά, δεν επιστρέφει απευθείας συστάδες, αλλά μια ταξινόμηση των σημείων, η οποία απεικονίζει τη δομή πυκνότητας του συνόλου δεδομένων [12].

Βασική ιδέα του αλγορίθμου είναι ότι οι συστάδες υψηλότερης πυκνότητας ενυπάρχουν μέσα σε συστάδες χαμηλότερης πυκνότητας. Συνεπώς, τα σημεία με μεγαλύτερη πυκνότητα επεξεργάζονται πρώτα, ώστε να αποκαλυφθεί σε αρχικό στάδιο η πιο συνεκτική δομή των δεδομένων. Αντί να κατασκευάζει άμεσα συστάδες, ο αλγόριθμος παράγει μία σειρά ταξινόμησης των σημείων (cluster ordering), που αποτυπώνει τη σχετική προσβασιμότητά τους σε συστάδες διαφόρων πυκνοτήτων. Ο αλγόριθμος ταξινομεί όλα τα σημεία σε μία σειρά, σύμφωνα με το πόσο εύκολα μπορεί να προσεγγιστεί το καθένα, και η πληροφορία αυτή μπορεί στη συνέχεια να χρησιμοποιηθεί για την εξαγωγή συστάδων μέσω γραφημάτων ή άλλων τεχνικών. Για κάθε σημείο, υπολογίζονται δύο βασικές ποσότητες:

- **Απόσταση Πυρήνα (Core Distance):** Είναι η μικρότερη τιμή ε για την οποία η *ε-γειτονιά* του σημείου περιέχει τουλάχιστον *MinPts* σημεία.
- **Απόσταση Προσβασιμότητας (Reachability Distance):** Η ελάχιστη ακτίνα που απαιτείται, ώστε το σημείο να θεωρηθεί προσβάσιμο από ένα γειτονικό Σημείο Πυρήνα, λαμβάνοντας υπόψη τόσο τη δική του, όσο και τη γειτονική πυκνότητα.

Η παραπάνω διαδικασία αποτυπώνεται στον παρακάτω Αλγόριθμο 4.

Αλγόριθμος 4 OPTICS

1: Επίσκεψη σε κάθε σημείο του συνόλου δεδομένων
2: Αν το σημείο δεν έχει ήδη επεξεργαστεί, υπολογίζεται η πυκνότητά του και καταχωρείται η απόσταση πυρήνα και προσβασιμότητας.
3: Τα σημεία που είναι άμεσα προσβάσιμα βάσει πυκνότητας εισάγονται σε μία δομή προτεραιότητας (OrderSeeds) για περαιτέρω ανάλυση.
4: Τα σημεία από τη λίστα αυτή ταξινομούνται ανάλογα με την απόσταση προσβασιμότητάς τους και επεξεργάζονται με τη σειρά αυτή.
5: Το αποτέλεσμα είναι μια ταξινομημένη λίστα σημείων, μαζί με τις αντίστοιχες αποστάσεις προσβασιμότητας (Reachability Distances), που αποτυπώνει τη δομή πυκνότητας του συνόλου δεδομένων.

Οι κύριες παράμετροι του OPTICS είναι:

- ***min_samples:*** Ο ελάχιστος αριθμός σημείων που απαιτείται, για να θεωρηθεί μια περιοχή ως πυκνή και να σχηματιστεί ένα cluster (ή να είναι ένα σημείο Πυρήνας - Core Point).
- ***max_eps:*** Η μέγιστη απόσταση μεταξύ δύο σημείων, για να θεωρηθεί ότι το ένα ανήκει στη "γειτονιά" του άλλου. Σε αντίθεση με τον DBSCAN, ο OPTICS δεν χρησιμοποιεί αυτή την παράμετρο για να ορίσει τα clusters, αλλά ως ένα ανώτερο όριο για την αναζήτηση γειτόνων.

- ***xi:*** Αυτή η παράμετρος χρησιμοποιείται από τη μέθοδο **cluster_method='xi'** για την αυτόματη εξαγωγή clusters από το Διάγραμμα Προσβασιμότητας (Reachability Plot). Καθορίζει την ελάχιστη "κλίση" ή "πτώση" (σε ποσοστό), που πρέπει να έχει μια κοιλάδα στο Διάγραμμα Προσβασιμότητας, για να θεωρηθεί ως ένα ξεχωριστό cluster.

Ένα από τα βασικά πλεονεκτήματα του OPTICS είναι η δυνατότητα ανίχνευσης συστάδων διαφορετικής πυκνότητας, χωρίς την ανάγκη καθορισμού σταθερής παραμέτρου *epsilon*. Ο αλγόριθμος προσφέρει μια λεπτομερέστερη κατανόηση της εσωτερικής δομής των δεδομένων και είναι κατάλληλος για σύνθετα ή πολυμορφικά δεδομένα. Επιπλέον, παράγει πληροφορίες που επιτρέπουν την εύκολη επιλογή παραμέτρων, εκ των υστέρων ή την αποφυγή τους εντελώς.

Ωστόσο, όπως αναφέρθηκε παραπάνω, ο OPTICS δεν επιστρέφει απευθείας συστάδες, αλλά μια γραφική αναπαράσταση (Reachability Plot), την οποία ο αναλυτής πρέπει να ερμηνεύσει, για να εξάγει τις συστάδες. Αυτό καθιστά την ανάλυση λιγότερο αυτοματοποιημένη και απαιτεί κάποια εμπειρία. Οι "κοιλάδες" σε αυτό το γράφημα αντιστοιχούν σε πυκνά clusters, ενώ οι "κορυφές" υποδεικνύουν σημεία που είναι θόρυβος, ή βρίσκονται μεταξύ clusters.

Επιπλέον, η υπολογιστική πολυπλοκότητα είναι αυξημένη σε σχέση με τον DBSCAN, ιδιαίτερα σε μεγάλα datasets, και το αποτέλεσμα μπορεί να επηρεαστεί από την επιλογή του *min_samples*. Η υπολογιστική πολυπλοκότητα του OPTICS γενικά είναι *O(nlogn)* για χωρικά ευρετήρια, όπως ball trees ή k-d trees, και O(n^2) στη χειρότερη περίπτωση, χωρίς βελτιστοποιήσεις. Παρόλα αυτά, ο αλγόριθμος παραμένει δημοφιλής για αναλύσεις, που η ποικιλία πυκνότητας είναι κρίσιμο χαρακτηριστικό των δεδομένων.

2.6 Μετρικές Αξιολόγησης

Silhouette Coefficient[6]**:** αποτελεί ένα χρήσιμο εργαλείο για την ερμηνεία και αξιολόγηση της ποιότητας μιας συσταδοποίησης. Ενσωματώνει τις δύο βασικές πτυχές [14]:

- τη συνοχή των σημείων εντός της ίδιας συστάδας (τα σημεία μιας συστάδας πρέπει να είναι κοντά μεταξύ τους) και
- τον διαχωρισμό τους από τις άλλες συστάδες (τα σημεία διαφορετικών συστάδων πρέπει να είναι μακριά μεταξύ τους).

Ο δείκτης υπολογίζεται για κάθε σημείο (Σχέση 2.2), συγκρίνοντας τη μέση απόσταση από τα υπόλοιπα σημεία της δικής του συστάδας (συνοχή *a*). Ακολούθως, υπολογίζεται για κάθε σημείο η μέση απόσταση προς σημεία άλλης συστάδας (διαχωρισμός *b*).

$$s = \frac{b - a}{\max(a, b)} \tag{2.2}$$

Η τιμή του κυμαίνεται μεταξύ -1 και +1. Όταν η τιμή του πλησιάζει το +1, αυτό υποδηλώνει, ότι η συσταδοποίηση είναι ιδανική, με ισχυρή συνοχή εντός της συστάδας και σαφή διαχωρισμό από

[6] https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

τις υπόλοιπες. Μια τιμή κοντά στο μηδέν δείχνει ότι οι συστάδες επικαλύπτονται, καθώς τα σημεία είναι εξίσου κοντά σε διαφορετικές συστάδες. Τέλος, αρνητικές τιμές φανερώνουν ανεπιτυχή συσταδοποίηση, καθώς ένα σημείο βρίσκεται πιο κοντά σε μια άλλη συστάδα από ό,τι στη δική του.

Ο συνολικός μέσος όρος του συντελεστή Silhouette χρησιμοποιείται τόσο για την αξιολόγηση της ποιότητας των συστάδων, όσο και για την επιλογή του κατάλληλου αριθμού *k*, στην περίπτωση του K-Means για παράδειγμα. Παρ' όλα αυτά, εμφανίζει κάποιους περιορισμούς, καθώς αποδίδει καλύτερα σε συστάδες με κυρτό σχήμα και ομοιογενές μέγεθος. Ωστόσο, δεν είναι το ίδιο αξιόπιστος σε περιπτώσεις ακανόνιστων ή άνισων συστάδων, όπως συμβαίνει με συστάδες που προκύπτουν από τον DBSCAN. Θα πρέπει να τονιστεί σε αυτό το σημείο, πως μπορεί να σχεδιαστεί γράφημα και στην περίπτωση του δείκτη Silhouette. Το γράφημα Silhouette Score σε συνάρτηση με τον αριθμό των συστάδων (clusters) επιτρέπει την επιλογή της βέλτιστης τιμής του *k*, η οποία αντιστοιχεί στο μέγιστο του δείκτη. Η κορυφή του γραφήματος δείχνει τον ιδανικό αριθμό clusters, με βάση τη συνοχή και τον διαχωρισμό αυτών [21].

Davies-Bouldin Index - DBI[7]**:** είναι ένας εσωτερικός δείκτης αξιολόγησης της ποιότητας μιας συσταδοποίησης. Ο στόχος του είναι να ποσοτικοποιήσει τον βαθμό διαχωρισμού και συνοχής των συστάδων που έχουν δημιουργηθεί από έναν αλγόριθμο συσταδοποίησης. Η βασική ιδέα πίσω από τον δείκτη είναι ότι μία καλή συσταδοποίηση χαρακτηρίζεται από συστάδες, που είναι όσο το δυνατόν πιο συμπαγείς (δηλαδή τα σημεία τους να είναι κοντά μεταξύ τους) και ταυτόχρονα όσο το δυνατόν πιο απομακρυσμένες μεταξύ τους [22], [23].

Ο δείκτης ορίζεται ως η μέση ομοιότητα μεταξύ κάθε συστάδας C_i (για *i=1,...,k*) και της περισσότερο προς αυτήν όμοιας συστάδας C_j. Η ομοιότητα ορίζεται ως ένα μέτρο R_{ij} που εξισορροπεί:

- Το s_i, τη μέση απόσταση μεταξύ κάθε σημείου της συστάδας *i* και του κεντροειδούς αυτής της συστάδας – γνωστή και ως διάμετρος συστάδας (ή εσωτερική διασπορά/συμπαγής δομή).
- Το d_{ij}, την απόσταση μεταξύ των κεντροειδών των συστάδων *i* και *j* – γνωστή και ως απόσταση μεταξύ συστάδων (ή διαχωρισμός).

Μια απλή επιλογή για την κατασκευή του R_{ij}, ώστε να είναι μη αρνητικό και συμμετρικό, είναι η (Σχέση 2.3):

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{2.3}$$

Ο συνολικός δείκτης προκύπτει από τον μέσο όρο των μέγιστων τιμών R_{ij} για κάθε συστάδα (Σχέση 2.4):

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{2.4}$$

Όσο χαμηλότερη είναι η τιμή του δείκτη, τόσο καλύτερη θεωρείται η συσταδοποίηση, καθώς αυτό υποδηλώνει μεγαλύτερο διαχωρισμό και μικρότερη διασπορά εντός των συστάδων. Ο DBI τείνει

[7] `https://scikit-learn.org/stable/modules/clustering.html#davies-bouldin-index`

να ευνοεί λύσεις συσταδοποίησης, που είναι σφαιρικές και περίπου ίσου μεγέθους. Εάν οι φυσικές συστάδες είναι επιμήκεις, ακανόνιστου σχήματος ή διαφέρουν σημαντικά σε μέγεθος, ο DBI ενδέχεται να μην παρέχει την πιο ακριβή αξιολόγηση. Μπορεί επίσης να είναι ευαίσθητος στην παρουσία ακραίων τιμών (outliers) μέσα στις συστάδες, οι οποίες μπορούν να αυξήσουν την τιμή s_i (διασποράς), κάνοντας τη συστάδα να φαίνεται λιγότερο συμπαγής. Μερικές φορές, ένα υψηλό s_i για μία συστάδα λόγω ενός outlier μπορεί να επηρεάσει δυσανάλογα τη συνάρτηση, οδηγώντας σε μια κατώτερη βαθμολογία, ακόμα κι αν άλλες συστάδες είναι καλά σχηματισμένες.

Calinski-Harabasz Index - CHI[8]**:** είναι μια άλλη δημοφιλής εσωτερική μέτρηση, συχνά αναφερόμενη ως Κριτήριο Αναλογίας Διασποράς (Variance Ratio Criterion). Ο δείκτης Calinski–Harabasz στηρίζεται στη σύγκριση μεταξύ της συνοχής των συστάδων και της διακριτότητάς τους. Ουσιαστικά, ο δείκτης εξετάζει πόσο καλά διαχωρίζονται οι συστάδες μεταξύ τους, σε συνδυασμό με το πόσο συμπαγείς είναι εσωτερικά.

Ο δείκτης υπολογίζεται με βάση την αναλογία της δια-συσταδικής διακύμανσης (inter-cluster dispersion) προς την ενδο-συσταδική διακύμανση (intra-cluster dispersion), πολλαπλασιασμένη κατάλληλα με βάση τον αριθμό των σημείων και των συστάδων (Σχέση 2.5):

$$CH = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \cdot \frac{n-k}{k-1} \tag{2.5}$$

όπου:

- *Tr(B_k)* είναι το ίχνος του πίνακα διακύμανσης μεταξύ των συστάδων (δηλαδή η συνολική διασπορά των κέντρων των συστάδων από το συνολικό κέντρο),
- *Tr(W_k)* είναι το ίχνος του πίνακα διακύμανσης εντός των συστάδων (δηλαδή το άθροισμα της διασποράς κάθε συστάδας),
- *n* είναι ο συνολικός αριθμός των δειγμάτων,
- *k* είναι ο αριθμός των συστάδων.

Όσο υψηλότερη είναι η τιμή του δείκτη Calinski–Harabasz, τόσο καλύτερη θεωρείται η συσταδοποίηση, καθώς αυτό υποδηλώνει μεγαλύτερο διαχωρισμό μεταξύ των συστάδων και μικρότερη διασπορά εντός τους. Ο δείκτης αυτός είναι ιδιαίτερα ευαίσθητος στο πλήθος των συστάδων και αποδίδει καλά, όταν οι συστάδες έχουν παρόμοιο μέγεθος και σχήμα. Παρόμοια και με τις προαναφερθείσες μετρικές, ο CHI μπορεί επίσης να ευνοεί σφαιρικές συστάδες και ενδέχεται να μην αποδίδει βέλτιστα με συστάδες ακανόνιστου σχήματος [24].

[8] `https://scikit-learn.org/stable/modules/clustering.html#calinski-harabasz-index`

Κεφάλαιο 3

Τεχνολογίες

3.1 Python

Το όνομα της γλώσσας δεν προέρχεται από το ομώνυμο ερπετό, αλλά από την αγάπη του δημιουργού της για τη βρετανική κωμική σειρά "Monty Python's Flying Circus". Η Python αναπτύχθηκε στα τέλη του 1980 από τον Guido van Rossum στο Εθνικό Ινστιτούτο Έρευνας Μαθηματικών και Επιστήμης Υπολογιστών στην Ολλανδία. Η πρώτη έκδοση της Python ξεκίνησε επίσημα το 1991 (Python 0.9.0) [25]. Η ανάγκη για μια απλή στη σύνταξη και ισχυρή στη λειτουργικότητα γλώσσα, οδήγησε στη δημιουργία της Python. Πλέον, η Python έχει εξελιχθεί σε μία από τις πιο δημοφιλείς και ευρέως χρησιμοποιούμενες γλώσσες προγραμματισμού παγκοσμίως, ιδιαίτερα στον τομέα της Επιστήμης Δεδομένων, της Τεχνητής Νοημοσύνης και της Ανάλυσης Δεδομένων [26], [27].

Η Python έχει καθαρή σύνταξη, η οποία θυμίζει την αγγλική γλώσσα, καθιστώντας την ιδανική για αρχάριους. Παρέχει πλούσιες βιβλιοθήκες και frameworks, όπως για παράδειγμα: Pandas, NumPy και TensorFlow, για απλές εφαρμογές επεξεργασίας δεδομένων, έως πιο σύνθετες, που αφορούν τη Τεχνητή Νοημοσύνη. Είναι μια γλώσσα υψηλού επιπέδου, Ανοιχτού Λογισμικού (Open Source) και διερμηνευόμενη (interpreted). Αυτό σημαίνει, ότι ο πηγαίος κώδικας μεταφράζεται γραμμή προς γραμμή, καθώς εκτελείται ο κώδικας. Έτσι, ο κύκλος ανάπτυξης γίνεται ταχύτερος, γιατί δεν υπάρχει το ενδιάμεσο βήμα της μεταγλώττισης. Σημαντικό, επίσης, χαρακτηριστικό της Python είναι η επεκτασιμότητα, καθώς μπορεί να συνδυαστεί με άλλες γλώσσες προγραμματισμού, όπως C/C++. Εκτός από τον Αντικειμενοστρεφή προγραμματισμό υποστηρίζει επίσης το Διαδικαστικό και Συναρτησιακό προγραμματισμό. Είναι φορητή (portable), μπορεί δηλαδή να εκτελείται σε λειτουργικά συστήματα Windows, macOS, Linux, αυξάνοντας, έτσι, την ευελιξία της [26].

3.2 Scikit-learn

Η Scikit-learn (γνωστή και ως sklearn) είναι μία από τις κυριότερες βιβλιοθήκες της Python για την υλοποίηση αλγορίθμων Μηχανικής Μάθησης. Είναι δωρεάν και ανοιχτού κώδικα. Είναι σχεδιασμένη, έτσι ώστε να διαλειτουργεί με τις βιβλιοθήκες NumPy και SciPy. Υποστηρίζει ένα ευρύ φάσμα αλγορίθμων Εποπτευόμενης και Μη Εποπτευόμενης Μάθησης. Η δημοφιλία της συγκεκριμένης βιβλιοθήκης είναι μεγάλη, καθώς είναι απλή και προσβάσιμη σε ένα ευρύ κοινό, όχι μόνο των ειδικών της Πληροφορικής. Η Scikit-learn χρησιμοποιείται για την ανάλυση δεδομένων και

την υλοποίηση αλγορίθμων Μηχανικής Μάθησης, σε τομείς της βιομηχανίας, αλλά και σε άλλες επιστήμες, όπως είναι, για παράδειγμα, η βιολογία και η ιατρική [28], [29].

Οι εφαρμογές της Scikit-learn συνοψίζονται παρακάτω :

- **Ταξινόμηση (Classification):** Προσδιορισμός της κατηγορίας στην οποία ανήκει ένα αντικείμενο. Για παράδειγμα, διάγνωση ασθενειών, αναγνώριση εικόνων.
- **Παλινδρόμηση (Regression):** Θεμελιώδης μέθοδος στη στατιστική και στη Μηχανική Μάθηση, που χρησιμοποιείται για να μοντελοποιήσει τη σχέση μεταξύ μιας εξαρτημένης μεταβλητής (target/output) και μίας ή περισσότερων ανεξάρτητων μεταβλητών (predictors/features). Η παλινδρόμηση προσπαθεί να προβλέψει συνεχείς αριθμητικές τιμές, π.χ. πρόβλεψη τιμών ακινήτων, ανάλυση τάσεων στις πωλήσεις.
- **Συσταδοποίηση (Clustering):** Αυτόματη ομαδοποίηση παρόμοιων αντικειμένων σε σύνολα. Για παράδειγμα, εντοπισμός ομάδων πελατών, ανάλυση χωρικών δεδομένων.
- **Μείωση Διαστάσεων (Dimensionality Reduction):** Μείωση του αριθμού των τυχαίων μεταβλητών, που λαμβάνονται υπόψη. Ενδεικτικά αναφέρονται η συμπίεση μεγάλων datasets και η οπτικοποίηση δεδομένων.
- **Επιλογή Μοντέλου (Model selection):** Σύγκριση, επικύρωση και επιλογή παραμέτρων και μοντέλων. Εδώ ανήκουν οι μετρικές αξιολόγησης, το Cross-validation (διασταυρούμενη επικύρωση), η οπτική απεικόνιση βαθμολογιών για την αξιολόγηση μοντέλων κ.α.
- **Προεπεξεργασία (Preprocessing):** Εξόρυξη χαρακτηριστικών, κωδικοποίηση κατηγορικών χαρακτηριστικών, κανονικοποίηση κ.α.

3.3 Folium

Η Folium είναι μια βιβλιοθήκη Python, που χρησιμοποιείται για την οπτικοποίηση γεωγραφικών δεδομένων, μέσω διαδραστικών χαρτών. Αξιοποιεί την ευελιξία της Python στην επεξεργασία δεδομένων, σε συνδυασμό με τις προηγμένες δυνατότητες χαρτογράφησης της βιβλιοθήκης Leaflet.js (δημοφιλής βιβλιοθήκη JavaScript για χαρτογραφία [30].

Η βιβλιοθήκη Folium είναι ιδιαίτερα χρήσιμη για την ανάλυση γεωγραφικών δεδομένων, την απεικόνιση χωρικών μοτίβων και τη δημιουργία χαρτών, που μπορούν εύκολα να ενσωματωθούν σε ιστοσελίδες. Υποστηρίζει πληθώρα διαφορετικών τύπων χαρτών (tiles), όπως για παράδειγμα οι: OpenStreetMap, Mapbox και Stamen (επιλογές όπως δορυφορικός, ασπρόμαυρο φόντο κτλ). Επιπλέον, μπορεί να προστεθεί και ένα Layer Control στον χάρτη, έτσι ώστε να μπορεί να γίνει εναλλαγή του τύπου του χάρτη. Με αυτό τον τρόπο, προσφέρεται ευελιξία στον χρήστη, ως προς την αισθητική και τη λειτουργικότητα της απεικόνισης. Παράλληλα, η Folium δίνει τη δυνατότητα προσθήκης διαφόρων διαδραστικών στοιχείων, όπως για παράδειγμα markers (συνήθως απεικονίζεται ως καρφίτσα και είναι ένα σημείο στον χάρτη, σε συγκεκριμένες γεωγραφικές συντεταγμένες - γεωγραφικό πλάτος και μήκος – που υποδηλώνει κάποιο σημαντικό δεδομένο ή γεγονός) και αναδυόμενα παραθύρα (popups). Οι χάρτες είναι πλήρως δυναμικοί, επιτρέποντας την πλοήγηση και

αλληλεπίδραση με τα δεδομένα. Ένα ακόμη σημαντικό πλεονέκτημα της βιβλιοθήκης Folium είναι, ότι οι χάρτες αποθηκεύονται ως αρχεία HTML ή ενσωματώνονται απευθείας σε διαδικτυακές εφαρμογές. Συνεργάζεται άψογα με βιβλιοθήκες, όπως η Pandas για ανάλυση δεδομένων.

3.4 Άλλες Βιβλιοθήκες

Η βιβλιοθήκη **plotly** αποτελεί ένα ισχυρό εργαλείο για τη δημιουργία διαδραστικών και αισθητικά ελκυστικών γραφημάτων σε Python. Είναι ανοιχτού κώδικα, βασισμένη στη βιβλιοθήκη Plotly.js και χρησιμοποιείται ευρέως στην ανάλυση δεδομένων. Υποστηρίζει μια μεγάλη ποικιλία από γεωγραφικές, στατιστικές, οικονομικές, επιστημονικές και τρισδιάστατες περιπτώσεις χρήσεις. Προσφέρει στον χρήστη τη δυνατότητα να εξερευνήσει τα δεδομένα, μέσω του περιβάλλοντος ενός φυλλομετρητή (browser) ή ενσωματωμένων HTML. Οι απεικονίσεις αυτές μπορούν να αποθηκευτούν σε HTML αρχεία ή να ενσωματωθούν σε διαδικτυακές εφαρμογές. Επιπλέον, με τη βοήθεια του εργαλείου Kaleido υποστηρίζει την εξαγωγή στατικών εικόνων υψηλής ποιότητας (δημιουργία PDF εγγράφων με διανυσματικές (vector) εικόνες υψηλής ανάλυσης) [31].

Η βιβλιοθήκη **colorsys** της Python αποτελεί ένα εύχρηστο εργαλείο για τη μετατροπή χρωμάτων μεταξύ διαφόρων χρωματικών χώρων. Συγκεκριμένα, επιτρέπει τη μετατροπή τιμών χρωμάτων μεταξύ του χρωματικού χώρου RGB (Red-Green-Blue), του χρωματικού χώρου YIQ (Luminance-In-phase-Quadrature, χρησιμοποιείται στην NTSC τηλεόραση), και του χρωματικού χώρου HLS (Hue-Luminance-Saturation). Αυτή η λειτουργικότητα είναι ιδιαίτερα χρήσιμη σε εφαρμογές όπου απαιτείται ο χειρισμός χρωμάτων για λόγους απεικόνισης, όπως στην οπτικοποίηση δεδομένων, επιτρέποντας την προσαρμογή της φωτεινότητας ή του κορεσμού των χρωμάτων για τη βελτίωση της αναγνωσιμότητας και της αισθητικής των γραφημάτων [32].

3.5 Google Colab

Το Google Colab (Collaboratory) είναι ένα δωρεάν, διαδικτυακό εργαλείο που προσφέρεται από την Google. Επιτρέπει στους χρήστες να γράφουν, να εκτελούν και να μοιράζονται κώδικα σε Python. Το εργαλείο αυτό είναι ιδιαίτερα διαδεδομένο στις ερευνητικές και επιστημονικές κοινότητες. Η χρήση του συνοδεύεται από πρόσβαση σε ισχυρούς υπολογιστικούς πόρους, γεγονός, που το καθιστά κατάλληλο στην πειραματική ανάπτυξη αλγορίθμων και την εκτέλεση έργων, που απαιτούν αυξημένη υπολογιστική ισχύ. Ουσιαστικά, παρέχει τη δυνατότητα εκτέλεσης κώδικα σε απομακρυσμένους υπολογιστικούς πόρους, όπως GPU και TPU, δίχως να απαιτείται η εγκατάσταση λογισμικού στον τοπικό υπολογιστή.

Το Colab λειτουργεί πλήρως μέσα από τον φυλλομετρητή και προσφέρει ενσωμάτωση με το Google Drive. Αυτό σημαίνει την εύκολη διαχείριση (πρόσβαση και αποθήκευση) των notebooks (σημειωματαρίων). Υποστηρίζει τη χρήση ευρείας γκάμας βιβλιοθηκών Python (NumPy, Pandas, Matplotlib, TensorFlow κ.α) για την ανάλυση δεδομένων και την ανάπτυξη μοντέλων Μηχανικής Μάθησης. Παράλληλα, προσφέρει δυνατότητες για δημιουργία οπτικοποιήσεων, ενώ όπως τονίστηκε προηγουμένως, διευκολύνει τη συνεργασία μεταξύ χρηστών. Ωστόσο, υπάρχουν ορισμένοι περιορισμοί, όπως ο μέγιστος επιτρεπόμενος χρόνος εκτέλεσης ανά συνεδρία (συνήθως έως 12 ώρες), η ανάγκη

συνεχούς σύνδεσης στο διαδίκτυο, και οι περιορισμοί στη διαχείριση πολύ μεγάλων συνόλων δεδομένων [33], [34].

3.6 WEKA

Το WEKA (Waikato Environment for Knowledge Analysis)[1] είναι ένα δημοφιλές λογισμικό ανοιχτού κώδικα για Εξόρυξη Δεδομένων και Μηχανική Μάθηση, το οποίο έχει αναπτυχθεί από το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία. Παρέχει ένα φιλικό περιβάλλον εργασίας μέσω γραφικής διεπαφής, αλλά υποστηρίζει επίσης scripting και ενσωμάτωση με Java, καθιστώντας το ιδιαίτερα ευέλικτο, τόσο για αρχάριους όσο και για προχωρημένους χρήστες. Το WEKA περιλαμβάνει μια μεγάλη συλλογή αλγορίθμων για ταξινόμηση (classification), παλινδρόμηση (regression), συσταδοποίηση (clustering), επιλογή χαρακτηριστικών και εξαγωγή κανόνων συσχέτισης, ενώ υποστηρίζει παράλληλα διαδικασίες προεπεξεργασίας δεδομένων.

Ένα από τα πλεονεκτήματα του WEKA είναι η δυνατότητά του να εφαρμόζει αλγορίθμους απευθείας σε αρχεία δεδομένων σε μορφή .arff ή .csv, χωρίς να απαιτείται προγραμματισμός. Το λογισμικό παρέχει επίσης εργαλεία οπτικοποίησης δεδομένων και αξιολόγησης μοντέλων, μέσω ποικίλων μετρικών επίδοσης. Επιπλέον, υποστηρίζει επαναληπτικές δοκιμές, διασταυρούμενη επικύρωση (cross-validation) και πειραματικά σενάρια μεγάλης κλίμακας [11], [35].

Μια ξεχωριστή λειτουργία του WEKA είναι το εξειδικευμένο πακέτο "Time Series Forecasting". Αυτό το πακέτο επιτρέπει τη δημιουργία μοντέλων πρόβλεψης βάσει χρονοσειρών, αξιοποιώντας αλγορίθμους παλινδρόμησης σε συνδυασμό με χρονικά μεταβλητά χαρακτηριστικά (όπως χρονικές υστερήσεις (lags) και ημερολογιακές πληροφορίες). Μέσω αυτού του ισχυρού εργαλείου, το WEKA μπορεί να χρησιμοποιηθεί αποτελεσματικά για την πρόβλεψη τάσεων σε χρονικά εξαρτώμενα φαινόμενα.

[1] `https://ml.cms.waikato.ac.nz/weka/index.html`

Κεφάλαιο 4

Στατιστική Ανάλυση Δεδομένων

4.1 Διερευνητική Ανάλυση των Δεδομένων (EDA)

Η Διερευνητική Ανάλυση Δεδομένων (EDA) αποτελεί το αρχικό στάδιο ανάλυσης ενός συνόλου δεδομένων και στοχεύει στην κατανόηση της δομής, των προτύπων και των σχέσεων των μεταβλητών. Περιλαμβάνει στατιστικές περιγραφές, απεικονίσεις με γραφήματα και εντοπισμό πιθανών ανωμαλιών ή ελλιπών τιμών, προετοιμάζοντας τα δεδομένα για περαιτέρω επεξεργασία ή μοντελοποίηση. Ο κώδικας εκτελεί Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA) σε ένα συγκεκριμένο σύνολο δεδομένων σεισμών (Ελλάδα), με σκοπό την κατανόηση των χαρακτηριστικών και των τάσεων της σεισμικής δραστηριότητας. Το συγκεκριμένο σύνολο δεδομένων αντλείται από το Γεωδυναμικό Ινστιτούτο[1]. Το εύρος των ημερομηνιών είναι συγκεκριμένο (1/1/1964 - 31/12/2024). Το μέγεθος είναι ορισμένο από 4.5 έως 8R και το βάθος από 0 έως 200 χλμ.

Το πρώτο τμήμα του κώδικα ασχολείται με τις ρυθμίσεις και την προετοιμασία του περιβάλλοντος. Ξεκινά εισάγοντας τις απαραίτητες βιβλιοθήκες: pandas για χειρισμό δεδομένων, numpy για αριθμητικούς υπολογισμούς, matplotlib.pyplot και seaborn για τη δημιουργία γραφημάτων, os για λειτουργίες συστήματος αρχείων, και folium μαζί με το folium.plugins.HeatMap για τη δημιουργία διαδραστικών χαρτών. Στη συνέχεια, ορίζεται ένας φάκελος εξόδου, όπου θα αποθηκευτούν όλα τα παραγόμενα γραφήματα και αρχεία, και δημιουργείται αυτός ο φάκελος, αν δεν υπάρχει ήδη.

Το επόμενο βήμα είναι η φόρτωση και προεπεξεργασία των δεδομένων. Ο κώδικας διαβάζει το αρχείο CSV 'EarthquakesGr.csv' (ο κατάλογος των σεισμών) σε ένα DataFrame της βιβλιοθήκης pandas. Ακολουθεί ένας έλεγχος για διπλότυπες εγγραφές, εντοπίζοντας και εκτυπώνοντας τυχόν διπλότυπα, καθώς και τον συνολικό τους αριθμό. Στη συνέχεια, η στήλη 'Origin Time (GMT)' μετατρέπεται σε αντικείμενο datetime, επιτρέποντας την εξαγωγή του Έτους, Μήνα και Ώρας ως ξεχωριστές στήλες. Δύο διαφορετικά DataFrames δημιουργούνται για συγκεκριμένες ανάγκες: το df_cleaned_for_numerical, που περιέχει μόνο αριθμητικά χαρακτηριστικά χωρίς NaN τιμές (κατάλληλο για στατιστική ανάλυση) και το df_for_maps, που προορίζεται για τους διαδραστικούς χάρτες Folium, διατηρώντας βασικές πληροφορίες και μορφοποιώντας τον χρόνο για εμφανίσεις σε pop-ups.

[1] https://www.gein.noa.gr/ypiresies-proionta/vasi-anazitisis/

Ακολουθεί η Περιγραφική Στατιστική ανάλυση, όπου υπολογίζονται βασικά στατιστικά μέτρα (όπως μέσος όρος, τυπική απόκλιση, ελάχιστο, μέγιστο, τεταρτημόρια) για τις επιλεγμένες αριθμητικές μεταβλητές. Αυτά τα στατιστικά εκτυπώνονται στην κονσόλα και αποθηκεύονται επιπλέον σε ένα αρχείο CSV για μελλοντική αναφορά, όπως φαίνεται παρακάτω στο συγκεκριμένο μπλοκ κώδικα.

Κώδικας εκτέλεσης Περιγραφικής Στατιστικής Δεδομένων:

```
# --- Περιγραφική Στατιστική ---
print("\n--- Περιγραφική Στατιστική Δεδομένων ---")
descriptive_stats = df_cleaned_for_numerical.describe().T

print("Πίνακας Περιγραφικής Στατιστικής:")
print(descriptive_stats)

descriptive_stats_path = os.path.join(output_dir_eda, "descriptive_statistics_eda.csv")
descriptive_stats.to_csv(descriptive_stats_path)
print(f"Η περιγραφική στατιστική αποθηκεύτηκε στο: {descriptive_stats_path}")
```

Ακολούθως, σχεδιάζεται η συνάρτηση add_stats_to_plot, για να ενσωματώνει αυτόματα τα υπολογισμένα στατιστικά στοιχεία απευθείας πάνω στα γραφήματα. Το μεγαλύτερο μέρος του κώδικα αφιερώνεται στις οπτικοποιήσεις δεδομένων. Ένας χάρτης με μεμονωμένα σημεία, όπου κάθε σεισμός αναπαρίσταται ως κύκλος, το μέγεθος και το χρώμα του οποίου αντικατοπτρίζουν το μέγεθος του σεισμού. Κάθε κύκλος έχει ένα pop-up, που εμφανίζει λεπτομερείς πληροφορίες για το σεισμό. Επίσης, παράγεται ένας θερμικός χάρτης (HeatMap), που απεικονίζει τις περιοχές με τη μεγαλύτερη συγκέντρωση σεισμικής δραστηριότητας, προσφέροντας μια οπτική αναπαράσταση της πυκνότητας των σεισμών. Αμφότεροι οι χάρτες επιτρέπουν την επιλογή διαφορετικών στυλ χάρτη, εμπεριέχουν τις τεκτονικές πλάκες και αποθηκεύονται ως αρχεία HTML.

Στη συνέχεια, ο κώδικας δημιουργεί ιστογράμματα (Histograms) και Box Plots για κάθε μία από τις βασικές αριθμητικές μεταβλητές (γεωγραφικό πλάτος, γεωγραφικό μήκος, βάθος, μέγεθος, έτος). Τα ιστογράμματα δείχνουν την κατανομή συχνοτήτων των τιμών, ενώ τα Box Plots αναδεικνύουν την κεντρική τάση, τη διασπορά και την ύπαρξη ακραίων τιμών (outliers). Ακολουθούν τα Διαγράμματα Διασποράς (Scatter Plots) για την οπτικοποίηση των σχέσεων μεταξύ ζευγών μεταβλητών. Ένα ειδικό scatter plot δείχνει τη γεωγραφική κατανομή των σεισμών, όπου το χρώμα και το μέγεθος των σημείων αντικατοπτρίζουν το μέγεθος του σεισμού, ενώ επισημαίνεται και η θέση του σεισμού με το μεγαλύτερο μέγεθος.

Στη συνέχεια, εξετάζεται η εξέλιξη της σεισμικής δραστηριότητας στο χρόνο. Δημιουργείται φίλτρο για να εισάγεται συγκεκριμένο εύρος ημερομηνιών. Δημιουργούνται γραφήματα για:

- Το πλήθος σεισμών ανά έτος, δείχνοντας τη χρονική εξέλιξη της συνολικής σεισμικής δραστηριότητας.
- Τον μέσο όρο μεγέθους σεισμών ανά έτος, αποκαλύπτοντας τις τάσεις στο μέγεθος των σεισμών.
- Το μεγαλύτερο μέγεθος σεισμού ανά έτος, υποδεικνύοντας τις χρονιές με τους ισχυρότερους σεισμούς.

- Το πλήθος σεισμών ανά μήνα, τόσο για όλα τα διαθέσιμα έτη όσο και ειδικά για τα τελευταία 10 έτη, για να εντοπιστούν πιθανές εποχικές τάσεις.
- Το πλήθος σεισμών ανά ώρα της ημέρας, για να διαπιστωθεί αν υπάρχει κάποια ημερήσια περιοδικότητα.

Τέλος, το τμήμα της ανάλυσης τοποθεσιών περιλαμβάνει την εύρεση της τοποθεσίας με το μεγαλύτερο μέγεθος σεισμού και τη δημιουργία ενός γραφήματος με τις 10 περιοχές με τους περισσότερους σεισμούς. Αυτό το γράφημα οπτικοποιεί τις πιο σεισμογενείς περιοχές, τα τελευταία 61 χρόνια. Ο κώδικας ολοκληρώνεται με τη δημιουργία ενός Heatmap πίνακα συσχετίσεων, ο οποίος απεικονίζει τη γραμμική συσχέτιση μεταξύ όλων των ζευγών των αριθμητικών μεταβλητών.

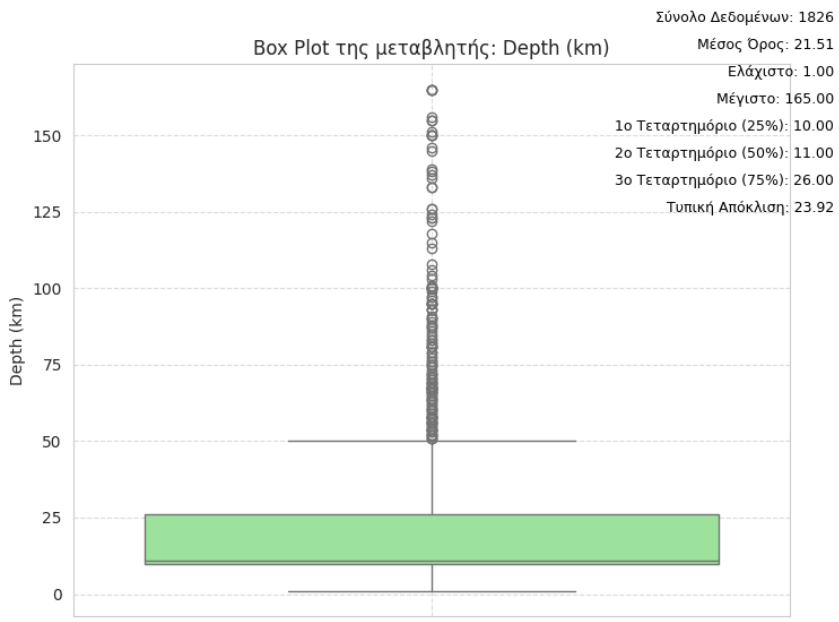
Συνολικά, ο κώδικας εκτελεί μια εκτενή EDA, χρησιμοποιώντας διάφορες οπτικοποιήσεις και ενσωματωμένα στατιστικά στοιχεία, για να παρέχει βαθιά κατανόηση των σεισμολογικών δεδομένων στην Ελλάδα, βοηθώντας στην αναγνώριση προτύπων, τάσεων και ανωμαλιών. Όλα τα παραγόμενα γραφήματα και οι αναφορές αποθηκεύονται συστηματικά σε έναν καθορισμένο φάκελο εξόδου. Ο συνολικός κώδικας παρατίθεται στο Παράρτημα Α (βλ. Διερευνητική Ανάλυση των Δεδομένων (EDA)).

Ερμηνεία Αποτελεσμάτων

Ο συγκεκριμένος κώδικας για την Διερευνητική Ανάλυση Δεδομένων (EDA) είναι σχεδιασμένος να παράγει μια πληθώρα οπτικοποιήσεων, οι οποίες είναι εξαιρετικά χρήσιμες για την αρχική διερεύνηση και την κατανόηση των χαρακτηριστικών του συνόλου δεδομένων. Ωστόσο, αναγνωρίζεται ότι ορισμένα από αυτά τα γραφήματα, όπως τα ιστογράμματα και τα box plots για την ίδια μεταβλητή ή διαφορετικές παραλλαγές διαγραμμάτων διασποράς, ενδέχεται να παρέχουν παρόμοιες ή αλληλοεπικαλυπτόμενες πληροφορίες. Για τον λόγο αυτό, και με γνώμονα τη διασφάλιση μιας συνοπτικής και αποτελεσματικής αναφοράς, πραγματοποιήθηκε προσεκτική επιλογή των πιο αντιπροσωπευτικών οπτικοποιήσεων. Η επιλογή αυτή στοχεύει στην ανάδειξη των σημαντικότερων ευρημάτων και τάσεων της σεισμικής δραστηριότητας, χωρίς περιττές επαναλήψεις, ώστε να διατηρηθεί η σαφήνεια και η ευκρίνεια στην παρουσίαση των αποτελεσμάτων.

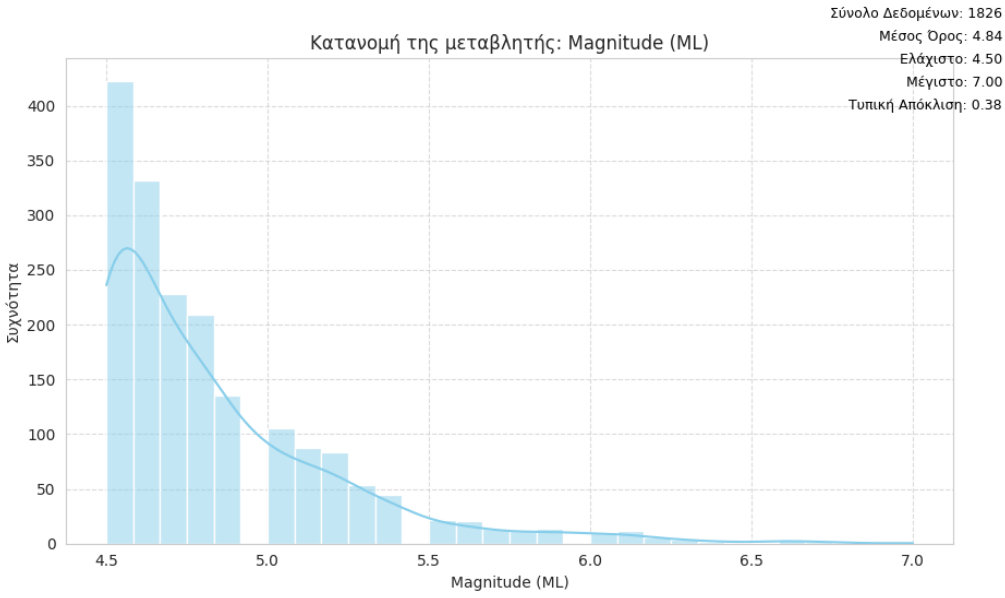
Αναφορικά με το βάθος των σεισμών (Depth) (Σχήμα 4.1), τα δεδομένα παρουσιάζουν ένα εύρος από 1.00 km (ελάχιστο) έως και 165.00 km (μέγιστο), με μέσο όρο τα 21.51 km. Ωστόσο, η διάμεσος τιμή είναι πολύ χαμηλότερη, στα 11.00 km, και το πρώτο τεταρτημόριο βρίσκεται στα 10.00 km, ενώ το τρίτο στα 26.00 km. Αυτά τα στατιστικά στοιχεία αποκαλύπτουν ότι το μεγαλύτερο μέρος των σεισμών, συγκεκριμένα το 75%, εκδηλώνεται σε σχετικά ρηχά βάθη, έως 26 km. Αυτή η συγκέντρωση των σεισμών σε μικρά βάθη είναι ένα χαρακτηριστικό γνώρισμα της σεισμικής δραστηριότητας στον ελλαδικό χώρο, καθώς οι περισσότεροι σεισμοί συνδέονται άμεσα με την επιφανειακή τεκτονική και την αλληλεπίδραση της αφρικανικής με την ευρασιατική πλάκα.

Παρά την κυριαρχία των ρηχών σεισμών, είναι αξιοσημείωτη η παρουσία πολυάριθμων ακραίων τιμών (outliers) στην κατανομή του βάθους, με ορισμένες εστίες να φτάνουν σε βάθη έως και 165.00 km. Αυτές οι βαθύτερες εστίες δεν είναι τυχαίες και υποδηλώνουν την ύπαρξη και τη λειτουργία ζωνών καταβύθισης, όπως είναι το Ελληνικό Τόξο, όπου μια τεκτονική πλάκα βυθίζεται κάτω από την άλλη, προκαλώντας σεισμική δραστηριότητα σε σημαντικά βάθη. Η μεγάλη τυπική απόκλιση του βάθους, που ανέρχεται σε 23.92 km, αποτελεί σαφή αντανάκλαση αυτής της εκτεταμένης δια-



Σχήμα 4.1: Θηκόγραμμα (Boxplot) για Βάθος

σποράς των δεδομένων, η οποία προκύπτει από τη συνύπαρξη τόσο ρηχών, όσο και βαθύτερων σεισμών στην περιοχή. Η κατανομή των δεδομένων, με τη σαφή συγκέντρωση σε ρηχά βάθη (μικρή απόσταση μεταξύ του 1ου και 2ου τεταρτημορίου) και την εκτεταμένη "ουρά" προς τα βαθύτερα βάθη, αποτελεί ένα θεμελιώδες και διακριτό χαρακτηριστικό της σεισμικότητας στην υπό μελέτη περιοχή.

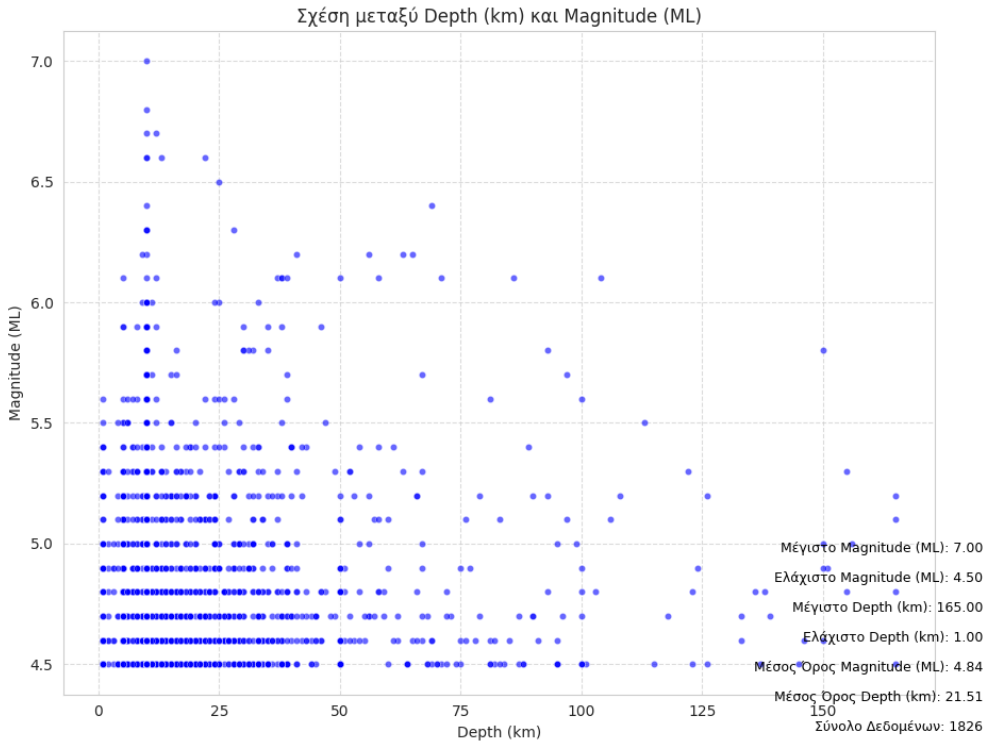


Σχήμα 4.2: Ιστόγραμμα για Μέγεθος

Το ιστόγραμμα του Μεγέθους (Magnitude - ML) (Σχήμα 4.2) παρέχει μια κρίσιμη οπτική αναπαράσταση της κατανομής του μεγέθους των σεισμών στο σύνολο δεδομένων. Με έναν μέσο όρο μεγέθους 4.84 M_L και τυπική απόκλιση 0.38, τα δεδομένα κυμαίνονται από ένα ελάχιστο 4.50 M_L έως ένα μέγιστο 7.00 M_L. Το γράφημα επιβεβαιώνει εμφανώς την έντονη συγκέντρωση της σεισμικής δραστηριότητας σε μικρότερα μεγέθη, γεγονός που αντικατοπτρίζεται σε μια πολύ μεγάλη κορυφή συχνότητας γύρω στο μέγεθος 4.5 - 4.6 M_L.

Η κατανομή αυτή χαρακτηρίζεται από έντονη ασυμμετρία προς τα δεξιά (positive skew), δηλαδή οι τιμές είναι συγκεντρωμένες στην αριστερή πλευρά (μικρότερα μεγέθη) και η "ουρά" της κατανομής εκτείνεται προς τα μεγαλύτερα μεγέθη. Αυτή η ραγδαία μείωση της συχνότητας όσο αυξάνεται το μέγεθος είναι απόλυτα συνεπής και συνάδει με τον νόμο Gutenberg-Richter, ο οποίος περιγράφει την αντίστροφη σχέση μεταξύ μεγέθους και συχνότητας σεισμών. Είναι μια τυπική κατανομή για σεισμολογικά δεδομένα. Παρόλο, που ο αριθμός των σεισμών μειώνεται σημαντικά στα υψηλότερα μεγέθη, η καταγραφή γεγονότων που φτάνουν έως και τα 7.00 M_L είναι ζωτικής σημασίας. Η παρουσία τέτοιων ισχυρών σεισμών υπογραμμίζει την ανάγκη για συνεχή αξιολόγηση του σεισμικού κινδύνου στην περιοχή, καθώς ακόμη και λίγα τέτοια γεγονότα μπορούν να έχουν σοβαρές επιπτώσεις.

Το διάγραμμα διασποράς (Σχήμα 4.3) που παρουσιάζει τη σχέση μεταξύ του βάθους και του μεγέθους των σεισμών προσφέρει κρίσιμες πληροφορίες για τη γεωδυναμική του ελλαδικού χώρου.



Σχήμα 4.3: Σχέση μεταξύ Βάθους και Μεγέθους

Παρατηρείται μια σαφής συγκέντρωση της συντριπτικής πλειονότητας των σεισμών σε ρηχά βάθη, κυρίως κάτω από 50 km, με μια ιδιαίτερα υψηλή συχνότητα σημείων (σεισμών) στην περιοχή των 0 - 25 km. Αυτή η επικράτηση των ρηχών σεισμών, ανεξαρτήτως μεγέθους, είναι ένα χαρακτηριστικό γνώρισμα της ελληνικής σεισμικότητας, όπου οι πιο συχνοί σεισμοί προκαλούνται από ρήγματα κοντά στην επιφάνεια.

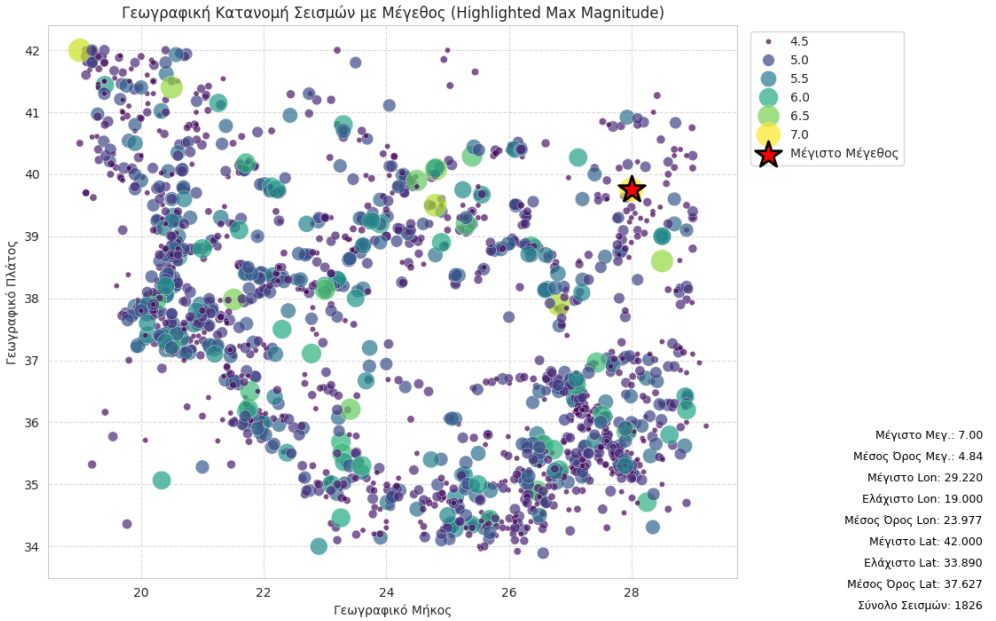
Επιπλέον, το γράφημα καταδεικνύει ότι οι μεγαλύτεροι σεισμοί (άνω των 6.0 M_L), συμπεριλαμβανομένων αυτών που φτάνουν μέχρι και το 7.0 M_L, τείνουν να συμβαίνουν πρωτίστως σε ρηχά βάθη (περίπου 0 - 60 km). Αυτό το εύρημα είναι ζωτικής σημασίας για την εκτίμηση του σεισμικού κινδύνου, καθώς οι ρηχοί σεισμοί, λόγω της εγγύτητάς τους στην επιφάνεια, είναι αυτοί που προκαλούν τις μεγαλύτερες επιφανειακές δονήσεις και, κατά συνέπεια, τις εκτενέστερες ζημιές στις κατασκευές. Αντιθέτως, καθώς το βάθος αυξάνεται, ο αριθμός των σεισμών μειώνεται σημαντικά. Παρατηρούνται σεισμοί σε πολύ μεγάλα βάθη, που φτάνουν έως και τα 150 - 160 km, αλλά αυτοί είναι συγκριτικά λιγότεροι σε αριθμό. Επίσης, λίγο βαθύτεροι σεισμοί (π.χ., άνω των 50 km) τείνουν γενικά να είναι μικρότερου μεγέθους, σπάνια ξεπερνώντας τα 6.0 M_L. Κυμαίνονται κυρίως γύρω στο 4.5 - 5.5 M_L. Αυτή η τάση είναι αναμενόμενη σε ζώνες καταβύθισης, όπως το Ελληνικό Τόξο, όπου η τεκτονική δραστηριότητα εκτείνεται σε μεγάλα βάθη.

Συνολικά, το διάγραμμα διασποράς αποκαλύπτει ότι δεν υπάρχει μια σαφής, ισχυρή γραμμική συσχέτιση μεταξύ βάθους και μεγέθους για το σύνολο των δεδομένων. Ωστόσο, αναδεικνύει δύο διακριτά χαρακτηριστικά: την επικράτηση των ρηχών και δυνητικά καταστροφικών σεισμών, και την παρουσία βαθύτερων γεγονότων. Τα τελευταία είναι συνήθως μικρότερου μεγέθους και λιγότερο επικίνδυνα για τις επιφανειακές κατασκευές, λόγω του βάθους τους, παρ' όλα αυτά είναι ενδεικτικά της σύνθετης τεκτονικής διεργασίας καταβύθισης, που λαμβάνει χώρα στον ελλαδικό χώρο.

Το διάγραμμα διασποράς που αναπαριστά τη γεωγραφική κατανομή των σεισμών στον ελλαδικό χώρο (Σχήμα 4.4), με οπτικοποίηση του μεγέθους και επισήμανση του ισχυρότερου γεγονότος, αποτελεί ένα εξαιρετικά πληροφοριακό εργαλείο για την κατανόηση της χωρικής κατανομής της σεισμικότητας.

Σε αυτό το γράφημα, το χρώμα και το μέγεθος κάθε σημείου (κουκκίδας) αντικατοπτρίζουν απευθείας το μέγεθος (Magnitude) του αντίστοιχου σεισμού, με τις πιο πράσινες αποχρώσεις και τις μεγαλύτερες κουκκίδες να υποδηλώνουν ισχυρότερα σεισμικά γεγονότα. Η οπτική ανάλυση αποκαλύπτει, ότι οι μεγάλοι σεισμοί δεν κατανέμονται ομοιόμορφα σε όλη την ελληνική επικράτεια. Αντιθέτως, παρατηρούνται να συγκεντρώνονται σε συγκεκριμένες, γνωστές ως ιδιαίτερα σεισμογενείς ζώνες. Χαρακτηριστικά παραδείγματα αποτελούν το Ελληνικό Τόξο (ιδιαίτερα η περιοχή νότια της Κρήτης και προς τα Δωδεκάνησα), το Ιόνιο Πέλαγος και η περιοχή του Ανατολικού Αιγαίου.

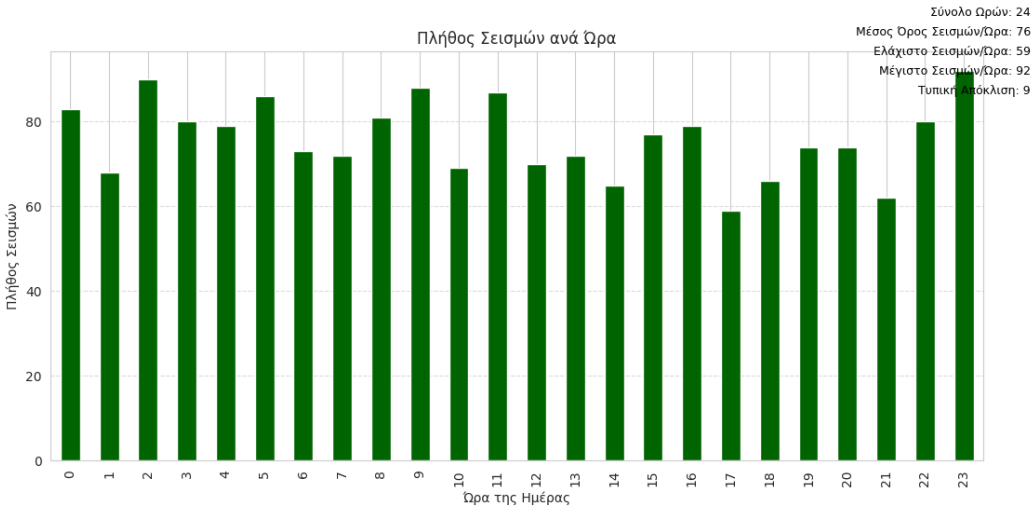
Ένα ιδιαίτερα σημαντικό στοιχείο του γραφήματος είναι η ξεχωριστή επισήμανση (με ένα κόκκινο αστεράκι) της γεωγραφικής θέσης του σεισμού με το μέγιστο μέγεθος, που καταγράφεται στο σύνολο δεδομένων. Αυτό το ισχυρότερο γεγονός εντοπίζεται στο ανατολικό Αιγαίο (Τουρκία). Η επισήμανση αυτή όχι μόνο αναδεικνύει μια περιοχή ιστορικά υψηλού κινδύνου, αλλά χρησιμεύει και ως σημείο αναφοράς για περαιτέρω διερεύνηση των γεωλογικών δομών, που είναι υπεύθυνες για την παραγωγή τόσο μεγάλων σεισμών. Συνολικά, αυτό το διάγραμμα είναι εξαιρετικά πολύτιμο



Σχήμα 4.4: Γεωγραφική Κατανομή Σεισμών με Μέγεθος

καθώς οπτικοποιεί με σαφήνεια όχι μόνο τη γενική χωρική κατανομή των σεισμών, αλλά κυρίως τις περιοχές όπου εκδηλώνονται οι ισχυρότεροι σεισμοί.

Η ανάλυση του πλήθους των σεισμών ανά ώρα της ημέρας (από 0 έως 23) προσφέρει μια ενδιαφέρουσα οπτική γωνία σχετικά με την ημερήσια κατανομή της σεισμικής δραστηριότητας στον ελλαδικό χώρο (Σχήμα 4.5). Το γράφημα αποκαλύπτει ότι η κατανομή των σεισμών ανά ώρα δεν



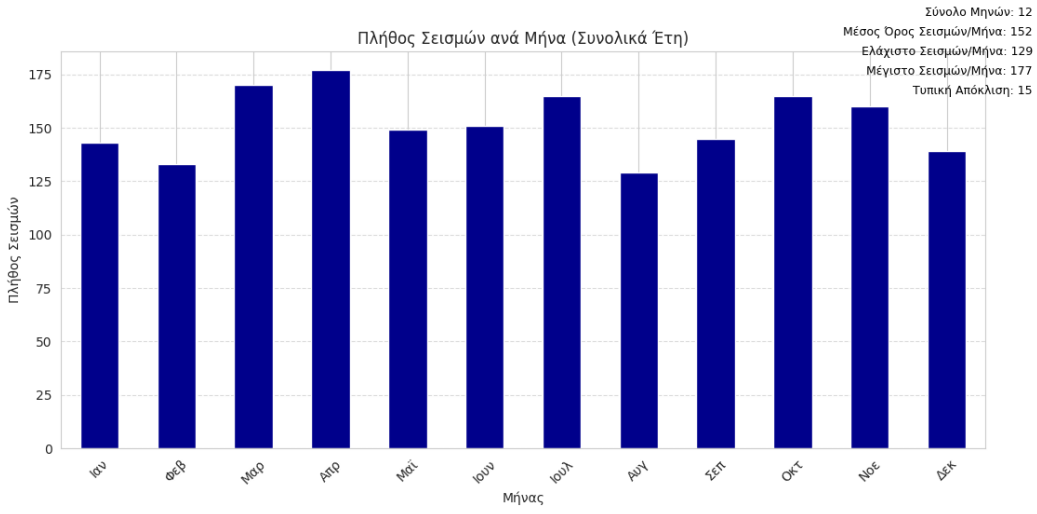
Σχήμα 4.5: Πλήθος Σεισμών ανά Ώρα

είναι πλήρως ομοιόμορφη, υποδηλώνοντας την ύπαρξη ωρών της ημέρας με ελαφρώς αυξημένη ή μειωμένη σεισμική δραστηριότητα.

Πιο συγκεκριμένα, παρατηρούνται κορυφές στη σεισμική δραστηριότητα κατά τις πρώτες πρωινές ώρες, περίπου μεταξύ 02:00 και 05:00, με ιδιαίτερη αύξηση γύρω στις 02:00. Επιπλέον, εμφανίζεται μια άλλη περίοδος αυξημένης δραστηριότητας γύρω στις 09:00-11:00 το πρωί. Μια περαιτέρω αύξηση καταγράφεται αργά το βράδυ, γύρω στις 22:00-23:00. Αντιθέτως, οι λιγότεροι σεισμοί τείνουν να καταγράφονται περίπου στις 17:00, με άλλες ώρες σχετικής ηρεμίας να περιλαμβάνουν την 01:00 και το διάστημα 18:00-21:00. Αυτές οι παρατηρούμενες διακυμάνσεις καθ' όλη τη διάρκεια της ημέρας είναι ενδιαφέρουσες. Οι σεισμοί είναι πρωτίστως φυσικά φαινόμενα, που δεν επηρεάζονται άμεσα από τον κύκλο ημέρας - νύχτας ή τις ανθρώπινες δραστηριότητες με την ίδια έννοια που επηρεάζονται από τους γεωλογικούς παράγοντες.

Το γράφημα που απεικονίζει το συνολικό πλήθος των σεισμών για κάθε μήνα, αθροισμένο για όλα τα διαθέσιμα έτη του dataset, παρέχει μια γενική εικόνα της κατανομής της σεισμικής δραστηριότητας σε μηνιαία βάση (Σχήμα 4.6). Από την ανάλυση, παρατηρούμε κάποιες διακυμάνσεις στο πλήθος των σεισμών ανά μήνα. Συγκεκριμένα, οι μήνες Μάρτιος και Απρίλιος παρουσιάζουν το μεγαλύτερο πλήθος σεισμών, με τον Απρίλιο να καταγράφει τους περισσότερους συνολικά (177). Αντίθετα, ο Αύγουστος εμφανίζει το χαμηλότερο πλήθος σεισμών (129), ακολουθούμενος από τον Φεβρουάριο και τον Ιανουάριο. Η τάση που διαγράφεται είναι μια αύξηση από τον Ιανουάριο προς τον Απρίλιο, μια πτώση προς τον Αύγουστο, και στη συνέχεια μια σχετική σταθεροποίηση ή μικρές διακυμάνσεις τους φθινοπωρινούς και χειμερινούς μήνες.

Ωστόσο, παρά αυτές τις παρατηρούμενες διακυμάνσεις, δεν διακρίνεται μια ισχυρή ή συνεπής εποχική περιοδικότητα στη σεισμική δραστηριότητα, με βάση τον μήνα για το σύνολο των δεδομένων. Οι διαφορές στο πλήθος των σεισμών μεταξύ των μηνών δεν είναι δραματικές και οι μικρές αυτές αυξομειώσεις ενδέχεται να οφείλονται περισσότερο σε τυχαία γεγονότα, ή σε μακροπρόθεσμες τεκ-

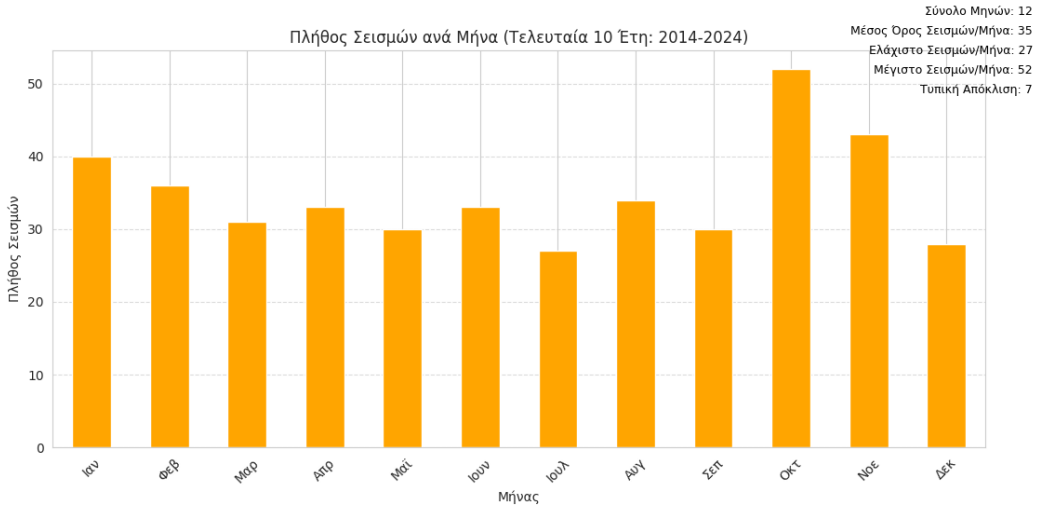


Σχήμα 4.6: Πλήθος Σεισμών ανά Μήνα (Συνολικά Έτη)

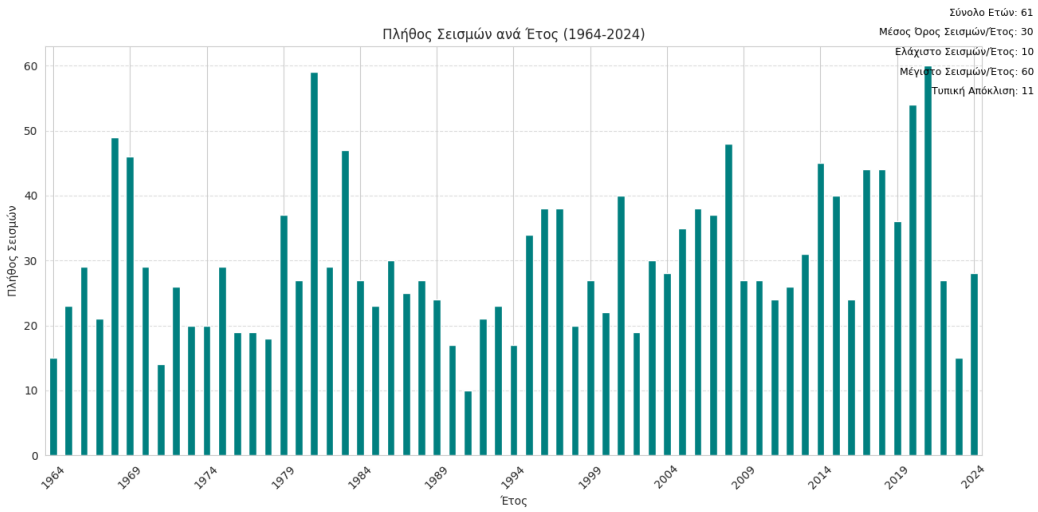
τονικές διεργασίες, που δεν συνδέονται άμεσα με εποχικούς παράγοντες. Επομένως, ενώ αναγνωρίζονται οι περίοδοι με ελαφρώς αυξημένη ή μειωμένη δραστηριότητα, όπως η άνοιξη που εμφανίζει συγκριτικά υψηλότερη συχνότητα σεισμών, αυτές οι διαφορές δεν είναι αρκετά σημαντικές ώστε να υποδηλώνουν ένα σαφές, επαναλαμβανόμενο εποχικό μοτίβο στη σεισμικότητα του ελλαδικού χώρου.

Το γράφημα που εμφανίζει το πλήθος των σεισμών ανά μήνα για την περίοδο των τελευταίων δέκα ετών (2014 - 2024) προσφέρει μια επικαιροποιημένη οπτική γωνία σχετικά με την εποχικότητα ή μη, της σεισμικής δραστηριότητας (Σχήμα 4.7). Από την ανάλυση αυτού του γραφήματος, διαπιστώνεται ότι ο Οκτώβριος εμφανίζει το υψηλότερο πλήθος σεισμών (52 γεγονότα), ακολουθούμενος από τον Νοέμβριο και τον Ιανουάριο. Αντίθετα, οι μήνες Ιούλιος και Δεκέμβριος παρουσιάζουν το χαμηλότερο πλήθος σεισμών, περίπου στα 27 γεγονότα.

Είναι αξιοσημείωτη η σημαντική διαφοροποίηση αυτού του μοτίβου σε σύγκριση με την ανάλυση της εποχικότητας για το σύνολο των διαθέσιμων ετών του σεισμικού καταλόγου. Ενώ σε μεγαλύτερες χρονικές κλίμακες παρατηρούνται διαφορετικές ενεργές περίοδοι (π.χ., η άνοιξη σε ολόκληρο το ιστορικό), τα δεδομένα των τελευταίων 10 ετών αναδεικνύουν τον Οκτώβριο ως τον πλέον ενεργό μήνα. Αυτή η παρατηρούμενη αλλαγή στην εποχική κατανομή ενδεχομένως υποδηλώνει, είτε μια πιθανή μετατόπιση στα χαρακτηριστικά της σεισμικότητας της περιοχής κατά την τελευταία δεκαετία, είτε την επίδραση μεμονωμένων, ισχυρότερων σεισμικών ακολουθιών, που συνέβησαν συγκεκριμένα τον Οκτώβριο εντός αυτής της περιόδου. Το εύρημα αυτό τονίζει με έμφαση τη σημασία της ανάλυσης των σεισμικών δεδομένων σε διαφορετικές χρονικές κλίμακες. Κάτι τέτοιο επιτρέπει την αποτελεσματικότερη ανίχνευση τόσο βραχυπρόθεσμων, όσο και μακροπρόθεσμων τάσεων στη σεισμική δραστηριότητα, οι οποίες είναι κρίσιμες για μια ολοκληρωμένη κατανόηση του φαινομένου.



Σχήμα 4.7: Πλήθος Σεισμών ανά Μήνα (Τελευταία 10 έτη 2014 - 2024)



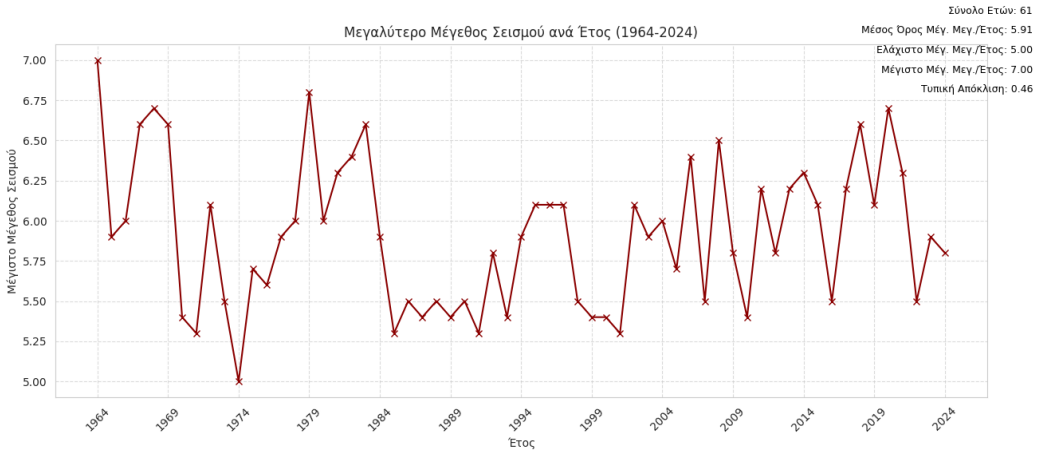
Σχήμα 4.8: Πλήθος Σεισμών ανά Έτος

Το γράφημα που εμφανίζει το πλήθος των σεισμών ανά έτος, καλύπτοντας την περίοδο από το 1964 έως το 2024, παρέχει μια επισκόπηση της συνολικής σεισμικής δραστηριότητας στον ελλαδικό χώρο (Σχήμα 4.8). Από την ανάλυση του γραφήματος, παρατηρούνται σημαντικές διακυμάνσεις στον αριθμό των σεισμών από έτος σε έτος. Υπάρχουν έτη που χαρακτηρίζονται από αυξημένο αριθμό σεισμικών γεγονότων, ενώ άλλα εμφανίζουν αισθητά χαμηλότερες τιμές, υποδηλώνοντας περιόδους αυξημένης ή μειωμένης έντασης στη σεισμική δραστηριότητα.

Η στατιστική σύνοψη, που συνοδεύει το γράφημα επιβεβαιώνει αυτή τη μεταβλητότητα: ο μέσος αριθμός σεισμών ανά έτος ανέρχεται σε 30, με ελάχιστο 10 και μέγιστο 60. Η υψηλή τιμή της τυπικής απόκλισης, που φτάνει το 11, υποδηλώνει αρκετή μεταβλητότητα και την έλλειψη ενός σταθερού ετήσιου μοτίβου. Η παρουσία συγκεκριμένων ετών με ιδιαίτερα υψηλές ή χαμηλές τιμές πλήθους σεισμών μπορεί να υποδηλώνει φυσικές γεωλογικές μεταβολές.

Το γράφημα που αποτυπώνει το μέγιστο μέγεθος σεισμού που καταγράφηκε κάθε έτος, από το 1964 έως το 2024, παρέχει μια κρίσιμη εικόνα της χρονικής εξέλιξης των ισχυρότερων σεισμικών γεγονότων στον ελλαδικό χώρο (Σχήμα 4.9). Η ανάλυση του γραφήματος αποκαλύπτει σημαντικές διακυμάνσεις στο μέγιστο μέγεθος από έτος σε έτος. Παρατηρούνται έτη με πολύ υψηλά μέγιστα μεγέθη, ενδεικτικά έντονης σεισμικής δραστηριότητας. Αυτές οι κορυφώσεις αντικατοπτρίζουν την εκδήλωση μεγάλων σεισμών, που ιστορικά έχουν προκαλέσει σημαντικές επιπτώσεις. Αντίθετα, υπάρχουν και περίοδοι σχετικής ηρεμίας, όπου το μέγιστο καταγραφόμενο μέγεθος είναι αισθητά χαμηλότερο, όπως για παράδειγμα τα μέσα της δεκαετίας του 1970.

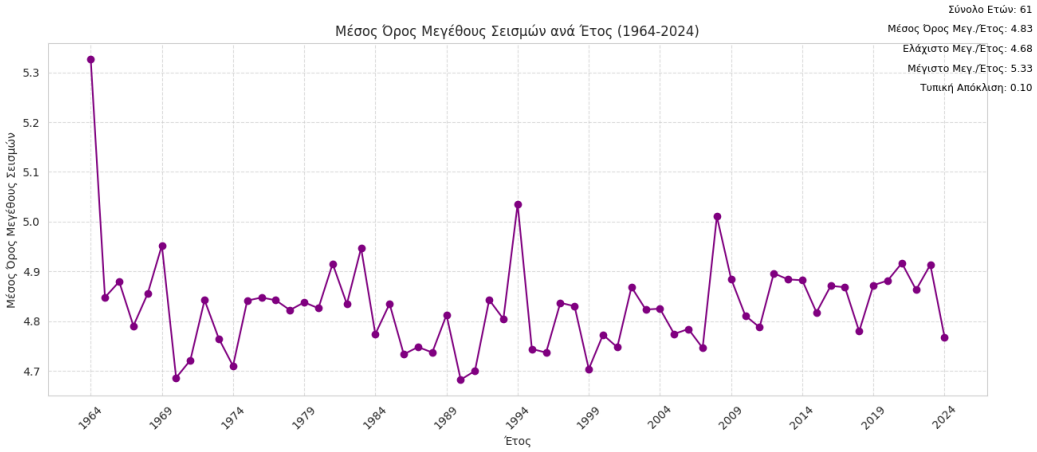
Το απόλυτο μέγιστο μέγεθος που έχει καταγραφεί στο συγκεκριμένο σύνολο δεδομένων είναι 7.0 M_L. Είναι σημαντικό να σημειωθεί ότι, παρά αυτές τις διακυμάνσεις, δεν παρατηρείται μια σαφής ανοδική ή καθοδική τάση στο μέγιστο μέγεθος σεισμού ανά έτος. Αυτό εντείνει την άποψη, ότι η εμφάνιση μεγάλων σεισμών είναι ένα γεγονός, που δεν ακολουθεί προβλέψιμα γραμμικά μοτίβα σε ετήσια βάση. Συμπερασματικά, το γράφημα επιβεβαιώνει ότι η Ελλάδα είναι μια περιοχή, που



Σχήμα 4.9: Μέγιστο Μέγεθος ανά Έτος

βιώνει συχνά ισχυρούς σεισμούς (μεγέθους 6.0 M_L και άνω). Η παρουσία ισχυρών σεισμών είναι ασταθής και απρόβλεπτη σε ετήσια βάση. Το γράφημα δείχνει μια ουσιαστική ιστορική εικόνα των μεγάλων σεισμών, η οποία είναι πολύτιμη για την κατανόηση της σεισμικής δραστηριότητας στον ελλαδικό χώρο.

Το γράφημα που παρουσιάζει τον μέσο όρο μεγέθους των σεισμών για κάθε έτος (1964 - 2024), προσφέρει μια συμπληρωματική οπτική (Σχήμα 4.10). Σε αντίθεση με τις σημαντικές διακυμάνσεις, που παρατηρούνται στο μέγιστο μέγεθος ανά έτος, ο μέσος όρος του μεγέθους ανά έτος διατηρείται σε ένα αξιοσημείωτα σταθερό επίπεδο, παρουσιάζοντας πολύ μικρότερες διακυμάνσεις. Η τιμή του μέσου όρου κυμαίνεται κυρίως μεταξύ 4.7 M_L και 5.0 M_L, με τον συνολικό μέσο όρο μεγέθους για



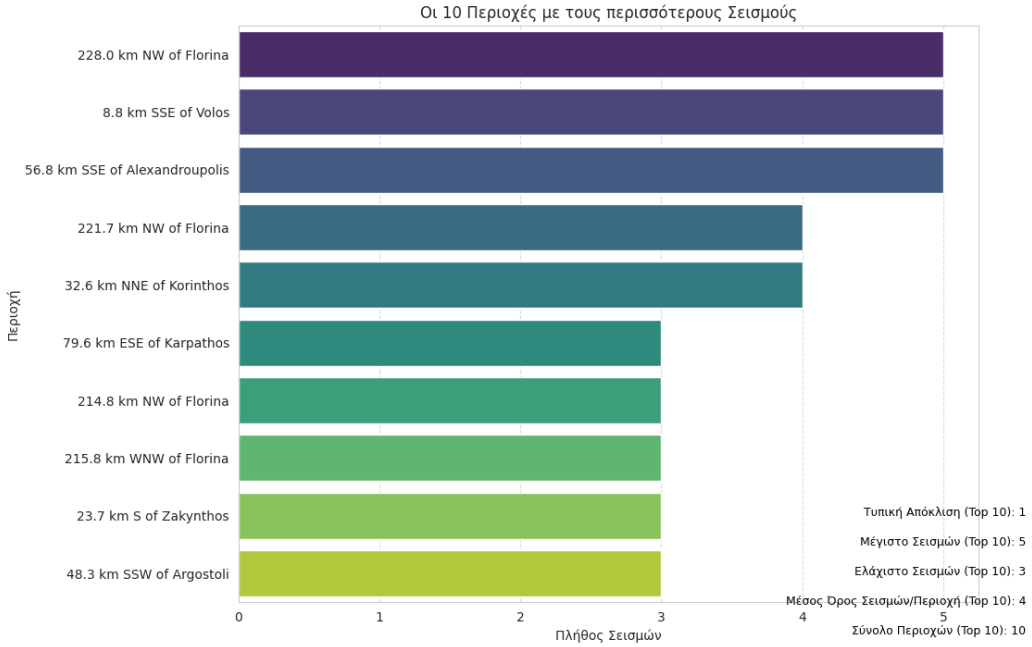
Σχήμα 4.10: Μέσος Όρος Μεγέθους ανά Έτος

όλη την περίοδο να ανέρχεται σε περίπου 4.83 M_L. Παρόλο, που εμφανίζονται κάποιες μικρές κορυφές, αυτές πιθανώς αντιστοιχούν σε έτη με αυξημένο πλήθος σεισμών μεσαίου μεγέθους ή την εμφάνιση ενός ή δύο μεγαλύτερων σεισμών που, μολονότι δεν είναι τα απόλυτα μέγιστα, συμβάλλουν στην αύξηση του ετήσιου μέσου όρου.

Το βασικό συμπέρασμα που προκύπτει από αυτό το γράφημα είναι, ότι οι μεγάλοι σεισμοί είναι απρόβλεπτα και σποραδικά γεγονότα. Η πλειονότητα των καταγεγραμμένων σεισμών στον ελλαδικό χώρο είναι μικρού και μεσαίου μεγέθους. Ο σχετικά σταθερός μέσος όρος του μεγέθους με την πάροδο του χρόνου υποδηλώνει, ότι η συνολική κατανομή των μεγεθών των σεισμών παραμένει σχετικά σταθερή.

Το οριζόντιο ραβδόγραμμα (Σχήμα 4.11) αναδεικνύει τις δέκα γεωγραφικές περιοχές του ελλαδικού χώρου, που έχουν καταγράψει το μεγαλύτερο πλήθος σεισμών στο σύνολο δεδομένων. Αυτό το γράφημα είναι ιδιαίτερα χρήσιμο για τον άμεσο εντοπισμό των "hotspots" της σεισμικής δραστηριότητας. Από την ανάλυση, διαπιστώνεται ότι οι δύο κορυφαίες περιοχές, "228.0 km NW of Florina" και "8.8 km SSE of Volos", παρουσιάζουν σημαντικά υψηλότερο πλήθος σεισμών, σε σύγκριση με τις υπόλοιπες στη δεκάδα. Είναι ιδιαίτερα αξιοσημείωτη η συγκέντρωση πολλών καταχωρήσεων στην πρώτη δεκάδα που αφορούν την ευρύτερη περιοχή της Φλώρινας, όπως "221.7 km NW of Florina", "214.8 km NW of Florina", και "215.8 km WNW of Florina". Αυτό δείχνει ξεκάθαρα, ότι η ευρύτερη περιοχή της Φλώρινας αποτελεί ένα σημαντικό και ενεργό επίκεντρο σεισμικής δραστηριότητας στην Ελλάδα.

Πέραν της Φλώρινας, η λίστα περιλαμβάνει και άλλες περιοχές με αυξημένη σεισμικότητα, οι οποίες είναι γνωστές για τη γεωλογική τους αστάθεια. Ενδεικτικά αναφέρονται οι: "56.8 km SSE of Alexandroupolis", "32.6 km NNE of Korinthos", "79.6 km ESE of Karpathos", "23.7 km S of



Σχήμα 4.11: Οι 10 Περιοχές με τους περισσότερους σεισμούς

Zakynthos" και "48.3 km SSW of Argostoli". Η παρουσία αυτών των περιοχών επιβεβαιώνει τη γενικευμένη και εκτεταμένη σεισμική δραστηριότητα σε όλη την Ελλάδα, συμπεριλαμβανομένων περιοχών, που βρίσκονται κοντά σε κρίσιμες σεισμικές ζώνες όπως ο Κορινθιακός Κόλπος, το Αιγαίο Πέλαγος και το Ιόνιο Πέλαγος.

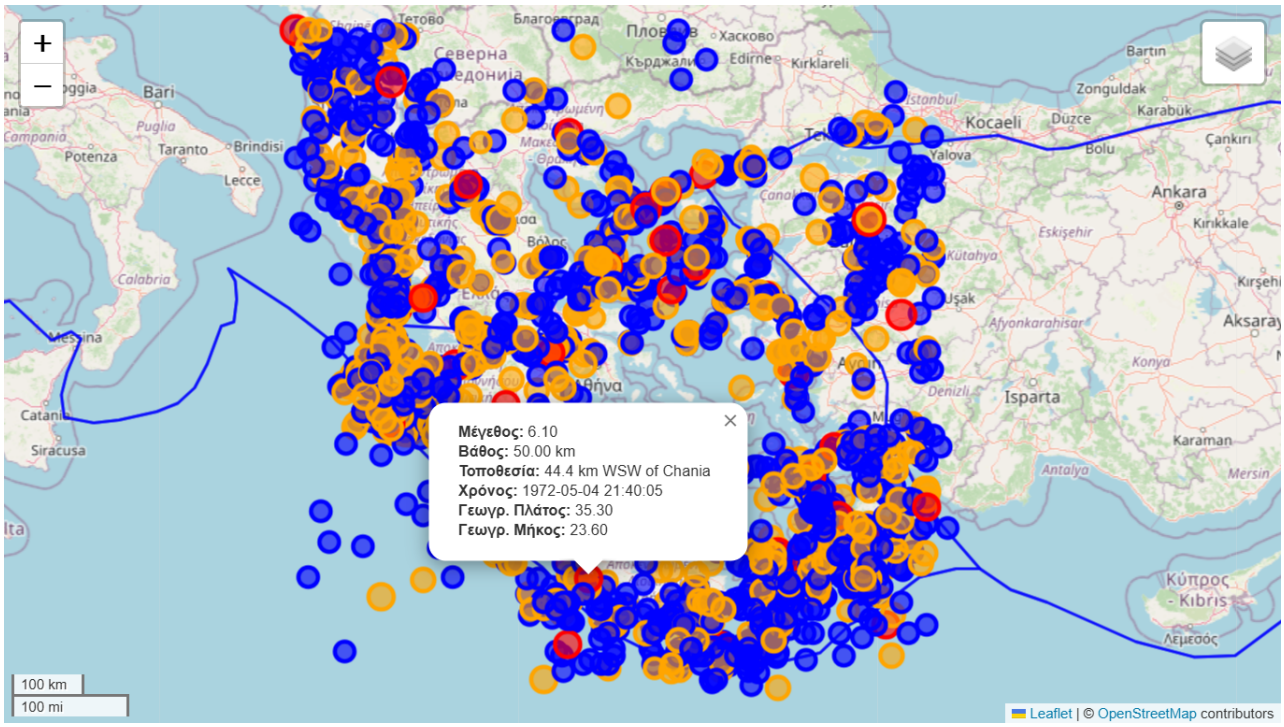
Στο πλαίσιο της ανάλυσης σεισμολογικών δεδομένων, η γεωχωρική οπτικοποίηση μέσω χαρτών αποτελεί ένα θεμελιώδες και αναντικατάστατο εργαλείο. Η δυνατότητα να απεικονιστούν χωρικά τα σεισμικά γεγονότα επιτρέπει στο να αναγνωριστούν άμεσα σημαντικές γεωδυναμικές πληροφορίες, όπως τα επίκεντρα των σεισμών, τις περιοχές συσσωρεύσεων δραστηριότητας, καθώς και τις πιθανές συσχετίσεις αυτών των φαινομένων με γνωστές γεωλογικές δομές, συμπεριλαμβανομένων των ορίων των τεκτονικών πλακών. Ειδικότερα, χρησιμοποιούνται διαφορετικοί τύποι χαρτών για την πλήρη κατανόηση της χωρικής διάστασης:

- Οι διαδραστικοί χάρτες σημείων προσφέρουν μια λεπτομερή επισκόπηση κάθε μεμονωμένου σεισμού. Αυτοί οι χάρτες επιτρέπουν την εξερεύνηση χαρακτηριστικών όπως το μέγεθος, το βάθος και την ακριβή γεωγραφική τοποθεσία για κάθε γεγονός. Η διαδραστικότητά τους βοηθά στην αναγνώριση εστιακών σημείων και την ακριβή εκτίμηση της χωρικής κατανομής της σεισμικής δραστηριότητας σε ένα μεγάλο σύνολο δεδομένων (Σχήμα 4.12).

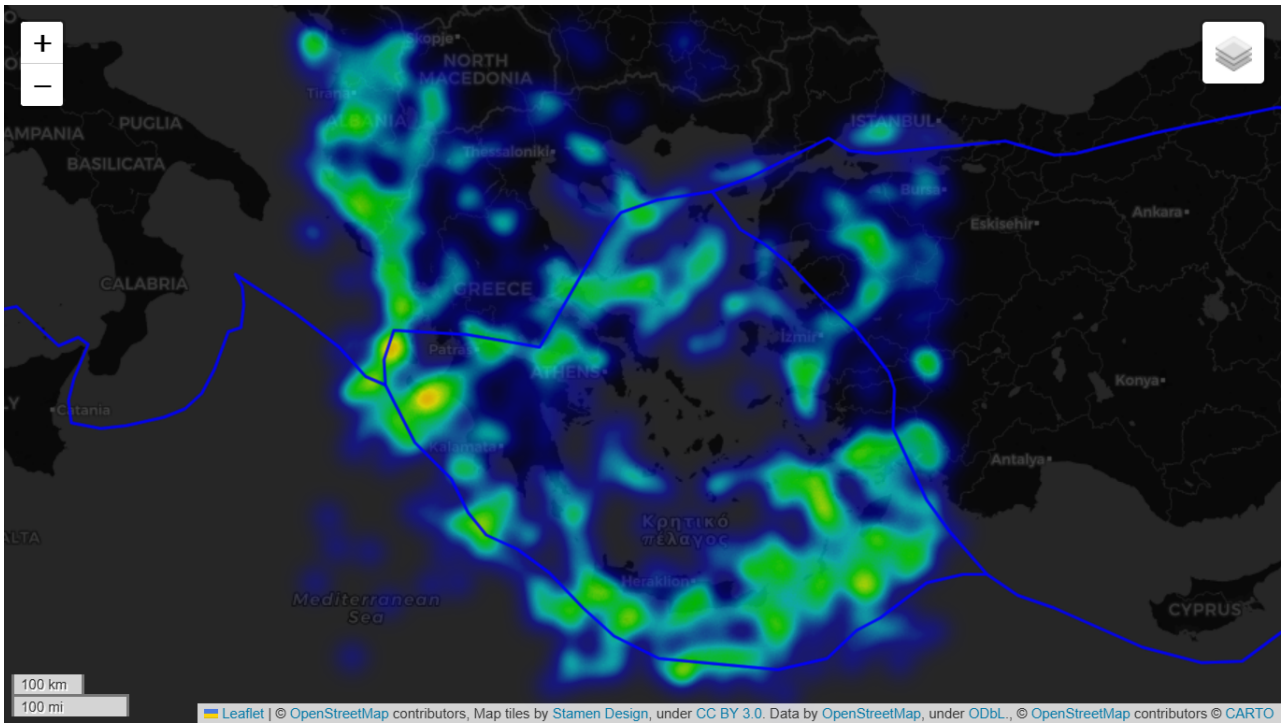
- Οι θερμικοί χάρτες (Heatmaps) συμπληρώνουν την ανάλυση των σημειακών δεδομένων, αναδεικνύοντας οπτικά τις περιοχές υψηλότερης σεισμικής πυκνότητας. Μέσω της απεικόνισης της συγκέντρωσης των σεισμών ως "θερμών" ζωνών, οι θερμικοί χάρτες αποκαλύπτουν με σαφήνεια τις πιο ενεργές σεισμικά περιοχές (Σχήμα 4.13).

Η ενσωμάτωση των ορίων των τεκτονικών πλακών σε αυτούς τους χάρτες είναι ζωτικής σημασίας για την ερμηνεία των σεισμολογικών φαινομένων. Δεδομένου ότι, η συντριπτική πλειονότητα των σεισμών συμβαίνει στα όρια των τεκτονικών πλακών, η ταυτόχρονη απεικόνιση των σεισμών και των ορίων των πλακών ενισχύει με έντονο τρόπο την κατανόηση των υποκείμενων γεωδυναμικών διεργασιών, που τους προκαλούν. Αυτή η συνδυαστική οπτικοποίηση παρέχει ουσιαστικές γνώσεις για την ερμηνεία των σεισμολογικών φαινομένων στο ευρύτερο γεωτεκτονικό πλαίσιο.

Εν κατακλείδι, μέσω αυτής της συστηματικής γεωχωρικής οπτικοποίησης, τα ακατέργαστα σεισμολογικά δεδομένα μετατρέπονται σε αναγνωρίσιμα και ερμηνεύσιμα πρότυπα. Αυτά τα πρότυπα είναι απολύτως απαραίτητα για την εφαρμογή προηγμένων τεχνικών εξόρυξης γνώσης και την εξαγωγή σημαντικών συμπερασμάτων, σχετικά με τη σεισμικότητα μιας περιοχής, συμβάλλοντας στην καλύτερη κατανόηση και διαχείριση του σεισμικού κινδύνου.



Σχήμα 4.12: Χάρτης με τους Σεισμούς 1964 - 2024 με τις Τεκτονικές Πλάκες



Σχήμα 4.13: Θερμικός Χάρτης (Heatmap) με τις Τεκτονικές Πλάκες

4.2 Πρόσφατη Σεισμική Δραστηριότητα Σαντορίνης

Η πρόσφατη σεισμική δραστηριότητα στη Σαντορίνη έχει προκαλέσει έντονο επιστημονικό ενδιαφέρον, καθώς καταγράφονται συνεχείς δονήσεις στην ευρύτερη περιοχή. Οι σεισμοί αυτοί, με μεγέθη που κυμαίνονται από 1 έως 5,3R, επικεντρώνονται κυρίως σε μια ζώνη μεταξύ της Σαντορίνης και της Αμοργού, κοντά στο υποθαλάσσιο ηφαίστειο Κολούμπο.

Στα πλαίσια της παρούσας εργασίας, υλοποιείται κώδικας Python, ο οποίος εκτελεί μια στοχευμένη στατιστική ανάλυση και οπτικοποίηση σεισμικών δεδομένων, συγκεκριμένα για την περιοχή της Σαντορίνης. Ο κατάλογος σεισμών αντλείται από το Γεωδυναμικό Ινστιτούτο[2]. Το εύρος των ημερομηνιών (1/1/2025 - 20/5/2025) είναι συγκεκριμένο, καθώς το φαινόμενο ξεκίνησε μέσα στον Ιανουάριο του 2025. Επίσης, το μέγεθος είναι ορισμένο από 0,1 εως 8R και το βάθος από 0 έως 200 χλμ (είναι οι προκαθορισμένες τιμές). Στη συγκεκριμένη ανάλυση είναι απαραίτητη κάθε λεπτομέρεια, ακόμη και για πολύ μικρούς σεισμούς, για να γίνει πλήρως κατανοητή η συνολική σεισμική δραστηριότητα.

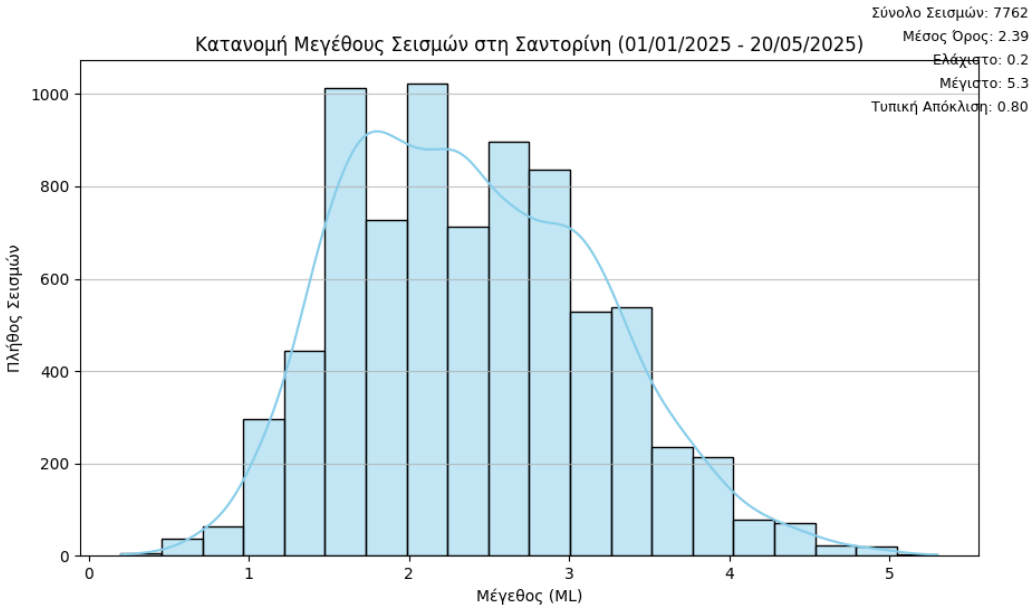
Αρχικά, ο κώδικας (βλ. Παράρτημα Α Πρόσφατη Σεισμική Δραστηριότητα Σαντορίνης) εισάγει τις απαραίτητες βιβλιοθήκες: pandas για την αποτελεσματική διαχείριση και ανάλυση δεδομένων, numpy για αριθμητικούς υπολογισμούς, matplotlib.pyplot και seaborn για τη δημιουργία υψηλής ποιότητας γραφημάτων, και os για λειτουργίες που σχετίζονται με το σύστημα αρχείων. Το πρώτο ουσιαστικό βήμα της ανάλυσης είναι η φόρτωση του συνόλου δεδομένων. Το αρχείο 'EarthquakesSantorini.csv' διαβάζεται σε ένα DataFrame της pandas. Ακολουθεί η προετοιμασία και το φιλτράρισμα των δεδομένων. Η στήλη 'Origin Time (GMT)', η οποία περιέχει την ημερομηνία και ώρα του σεισμού, μετατρέπεται σε τύπο δεδομένων datetime, κάτι που είναι απαραίτητο για χρονολογικές αναλύσεις. Το πιο κρίσιμο βήμα σε αυτή την ενότητα είναι το φιλτράρισμα των σεισμών. Δημιουργείται ένα νέο DataFrame, το df_thira, το οποίο περιέχει μόνο τις εγγραφές όπου η στήλη 'Location' περιέχει τη συμβολοσειρά "Thira" (χωρίς διάκριση πεζών-κεφαλαίων). Αυτό εξασφαλίζει ότι η ανάλυση επικεντρώνεται αποκλειστικά στους σεισμούς που σχετίζονται με τη Σαντορίνη.

Ο κώδικας προχωρά στον υπολογισμό και την εκτύπωση βασικών περιγραφικών στατιστικών. Υπολογίζεται ο συνολικός αριθμός των φιλτραρισμένων σεισμών. Στη συνέχεια, χρησιμοποιείται η μέθοδος .describe() για να ληφθούν περιγραφικά στατιστικά (μέσος όρος, ελάχιστο, μέγιστο, τεταρτημόρια, τυπική απόκλιση κ.ά.) για τις στήλες 'Magnitude (ML)', 'Depth (km)', 'Latitude', και 'Longitude'. Αυτά τα στατιστικά εκτυπώνονται στην κονσόλα, δίνοντας μια γρήγορη αριθμητική σύνοψη των δεδομένων της Σαντορίνης. Επίσης, υπολογίζεται και εκτυπώνεται το χρονικό εύρος των δεδομένων (ελάχιστη και μέγιστη ημερομηνία). Παράκάτω δίνονται τα συγκεκριμένα στατιστικά, που εμφανίζονται στην κονσόλα:

[2] https://www.gein.noa.gr/ypiresies-proionta/vasi-anazitisis/

```
    --- Βασικά Περιγραφικά Στατιστικά για Σαντορίνη (Thira, 01/01/2025 - 20/05/2025) ---

Συνολικός αριθμός σεισμών: 7762

Στατιστικά για Magnitude (ML):
count    7762.000000
mean        2.394821
std         0.795072
min         0.200000
25%         1.800000
50%         2.300000
75%         3.000000
max         5.300000
Name: Magnitude (ML), dtype: float64

Στατιστικά για Depth (km):
count    7762.000000
mean       11.262561
std         4.562934
min         2.000000
25%         9.000000
50%        11.000000
75%        13.000000
max       118.000000
Name: Depth (km), dtype: float64

Στατιστικά για Latitude:
count    7762.000000
mean       36.599615
std         0.113396
min        35.194700
25%        36.579900
50%        36.613300
75%        36.650800
max        36.777600
Name: Latitude, dtype: float64

Στατιστικά για Longitude:
count    7762.000000
mean       25.635306
std         0.351576
min        22.040400
25%        25.628800
50%        25.670000
75%        25.718500
max        26.392800
Name: Longitude, dtype: float64

Εύρος ημερομηνιών/ωρών δεδομένων: Από 2025-01-02 03:07:00 έως 2025-05-20 22:23:41
Η ανάλυση των περιγραφικών στατιστικών ολοκληρώθηκε.
```

Ένα σημαντικό στοιχείο του κώδικα είναι η συνάρτηση που έχει σχεδιαστεί, για να ενσωματώνει αυτόματα και δυναμικά στατιστικές πληροφορίες απευθείας πάνω στα γραφήματα. Το τελευταίο και πιο εκτεταμένο τμήμα του κώδικα αφορά τις οπτικοποιήσεις των δεδομένων. Συνολικά, ο κώδικας παρέχει μια εις βάθος ανάλυση των σεισμών στη Σαντορίνη, συνδυάζοντας την αριθμητική στατιστική με δυναμικές οπτικοποιήσεις, που ενσωματώνουν άμεσα τα βασικά χαρακτηριστικά των δεδομένων, διευκολύνοντας με αυτό τον τρόπο την ερμηνεία των αποτελεσμάτων.

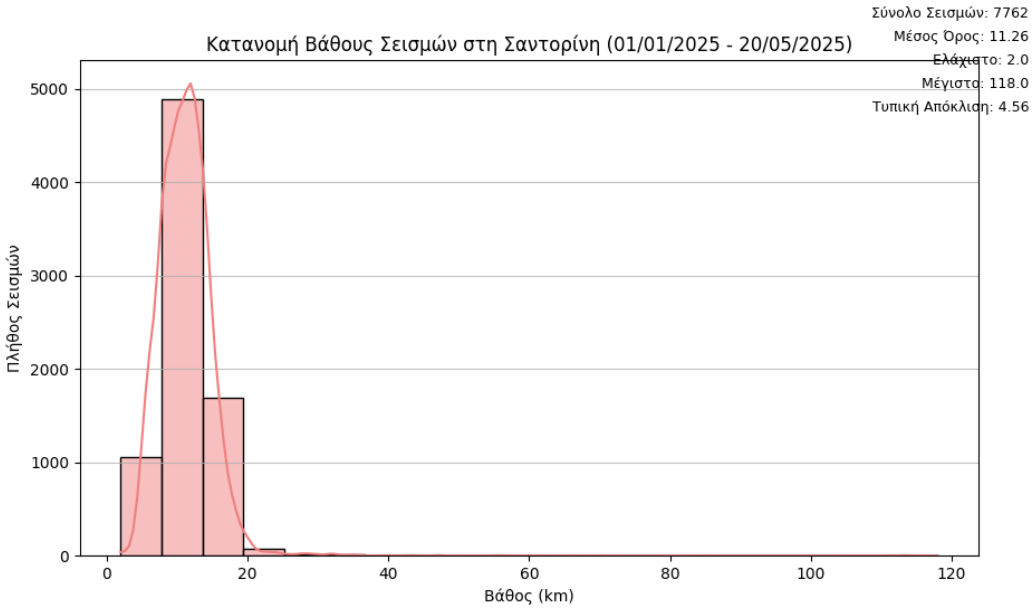


Σχήμα 4.14: Κατανομή Μεγέθους Σεισμών (Magnitude Histogram)

Το ιστόγραμμα που απεικονίζει την κατανομή των μεγεθών (M_L) των σεισμών στην περιοχή της Σαντορίνης για την περίοδο από την 1η Ιανουαρίου έως την 20ή Μαΐου 2025 παρέχει μια σαφή εικόνα της τοπικής σεισμικής δραστηριότητας (Σχήμα 4.14). Η ανάλυση του γραφήματος αποκαλύπτει μια τυπική κατανομή για σεισμική δραστηριότητα, στην οποία παρατηρείται ένα μεγάλο πλήθος μικρών σεισμών και ένας ραγδαία μειούμενος αριθμός, καθώς αυξάνεται το μέγεθος. Συγκεκριμένα, η μεγαλύτερη συγκέντρωση σεισμών εντοπίζεται στα μεγέθη περίπου 1.5 M_L έως 2.5 M_L, ενώ το πλήθος τους μειώνεται σημαντικά για μεγέθη άνω των 2.5 M_L. Ωστόσο, είναι αξιοσημείωτη η παρουσία ενός σημαντικού αριθμού σεισμών μεγέθους 3.0 M_L και 4.0 M_L.

Τα συνολικά στατιστικά στοιχεία για αυτή την περίοδο ενισχύουν αυτά τα ευρήματα: καταγράφηκαν 7.762 σεισμοί με μέσο όρο μεγέθους 2.39 M_L. Το εύρος των μεγεθών εκτείνεται από ένα ελάχιστο 0.2 M_L έως ένα μέγιστο 5.3 M_L, με τυπική απόκλιση 0.80 M_L. Η κατανομή αυτή επιβεβαιώνει τη συνεχή σεισμική συμπεριφορά στην περιοχή, όπου οι μικρότεροι σεισμοί είναι οι πιο συχνοί. Η κυριαρχία μικρών μεγεθών, σε συνδυασμό με την παρουσία σημαντικού αριθμού σεισμών μεγέθους 3.0 - 4.0 M_L (σμηνοσειράς), υποδηλώνει μια συνεχή απελευθέρωση ενέργειας, κάτι που είναι χαρακτηριστικό της σεισμικότητας ενός ενεργού ηφαιστειακού τόξου. Επίσης, ο καταγεγραμμένος μέγιστος σεισμός των 5.3 M_L αποτελεί ένα σημαντικό γεγονός για την εξεταζόμενη περίοδο.

Το ιστόγραμμα που παρουσιάζει την κατανομή των βάθους (km) των σεισμών στην περιοχή της Σαντορίνης αναδεικνύει μια έντονη συγκέντρωση σεισμών σε πολύ ρηχά βάθη, κυρίως μεταξύ 0 και 15 km, με μια σαφή κορύφωση στα 5 - 10 km (Σχήμα 4.15). Αυτό το πρότυπο κατανομής είναι ιδιαίτερα χαρακτηριστικό της επιφανειακής σεισμικότητας, που συχνά σχετίζεται με ηφαιστειακές και τεκτονικές διεργασίες. Η συντριπτική πλειονότητα των σεισμών είναι ρηχοί, κάτι που είναι αναμενόμενο για ηφαιστειακές περιοχές, καθώς οι σεισμοί συχνά συνδέονται με την κίνηση μάγματος ή άλλες διεργασίες.

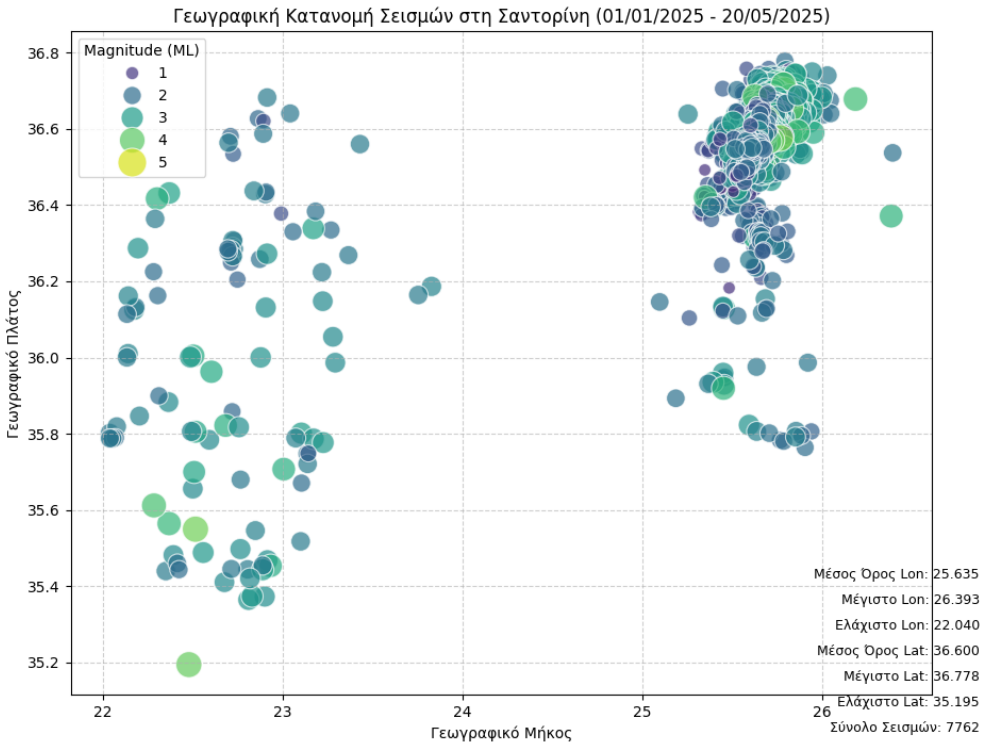


Σχήμα 4.15: Κατανομή Βάθους Σεισμών (Depth Histogram)

Τα συνολικά στατιστικά στοιχεία για αυτή την περίοδο επιβεβαιώνουν την επικράτηση των ρηχών σεισμών: από τους 7.762 καταγεγραμμένους σεισμούς, ο μέσος όρος του βάθους είναι 11.26 km, με ελάχιστο βάθος 2.0 km. Η τυπική απόκλιση ανέρχεται στα 4.56 km, υποδηλώνοντας ότι οι περισσότεροι σεισμοί είναι συγκεντρωμένοι γύρω από αυτόν τον μέσο όρο. Είναι σημαντικό να σημειωθεί ότι, αν και το πλήθος των σεισμών μειώνεται δραματικά καθώς αυξάνεται το βάθος (με πολύ λίγους σεισμούς σε βάθος μεγαλύτερο των 20 - 30 km), το μέγιστο βάθος που καταγράφηκε είναι 118.0 km. Η παρουσία αυτών των εξαιρετικά σπάνιων, βαθύτερων σεισμών μπορεί να υποδηλώνει συγκεκριμένες τεκτονικές διαδικασίες και είναι σημαντικό να αναγνωριστεί η ύπαρξή τους. Συμπερασματικά, οι σεισμοί στη Σαντορίνη είναι κυρίως ρηχοί, υποδηλώνοντας ότι σχετίζονται πρωτίστως με διεργασίες, που λαμβάνουν χώρα στο ανώτερο τμήμα του φλοιού, εντός ή κοντά στο ηφαιστειακό σύστημα.

Το διάγραμμα διασποράς στο (Σχήμα 4.16) δείχνει τα επίκεντρα των σεισμών στην ευρύτερη περιοχή της Σαντορίνης. Το μέγεθος κάθε κύκλου στο διάγραμμα είναι ανάλογο με το μέγεθος του σεισμού, επιτρέποντας την οπτικοποίηση τόσο της γεωγραφικής θέσης, όσο και του σχετικού μεγέθους των γεγονότων. Η ανάλυση του διαγράμματος αποκαλύπτει μια σαφή συγκέντρωση της σεισμικής δραστηριότητας. Η πλειονότητα των σεισμών εντοπίζεται σε μια περιοχή ανατολικά της Σαντορίνης. Οι μεγαλύτεροι σεισμοί της περιόδου (που αναπαρίστανται από τους μεγαλύτερους κύκλους και τα πιο ανοιχτόχρωμα χρώματα στο διάγραμμα) εμφανίζονται επίσης εντός αυτής της συστάδας. Υπάρχει, επίσης, κάποια διάσπαρτη, περιφερειακή δραστηριότητα (δυτικά), αλλά η κύρια συγκέντρωση και το επίκεντρο των ισχυρότερων γεγονότων είναι σαφώς εντοπισμένα.

Η χαρτογράφηση των σεισμών αποκαλύπτει την ύπαρξη μιας ενεργού ηφαιστειακής δομής, η οποία εντοπίζεται με ακρίβεια ανατολικά της νήσου. Η χωρική συγκέντρωση των σεισμών σε αυτή την περιοχή, δείχνει ότι η πρόσφατη σεισμική δραστηριότητα σχετίζεται με διεργασίες εντός του ηφαι-

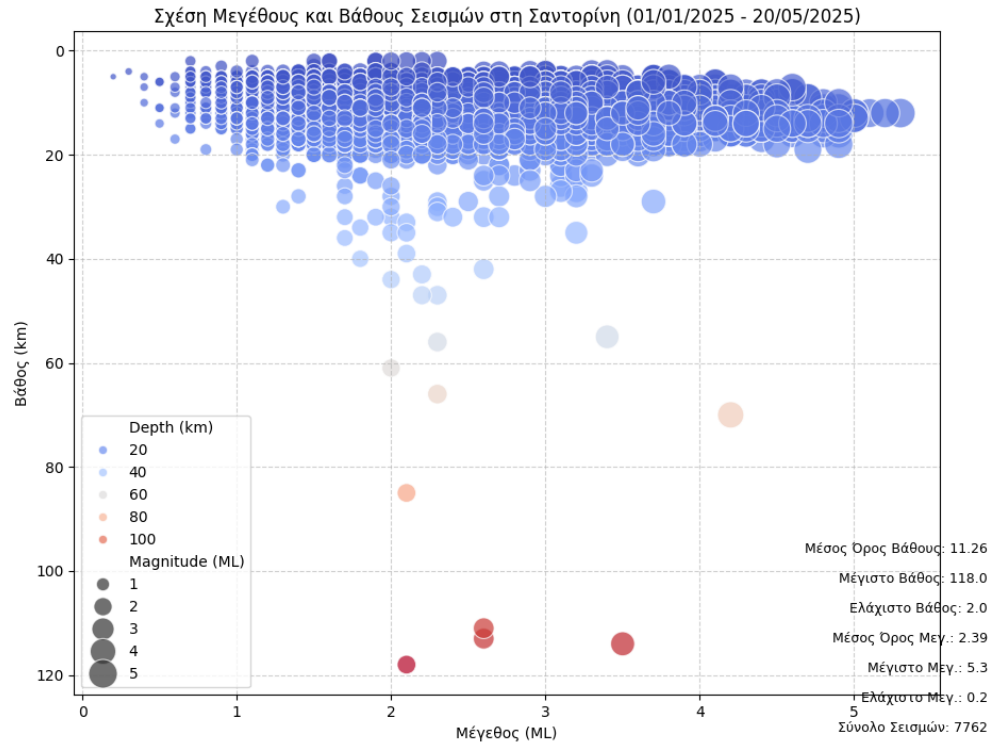


Σχήμα 4.16: Γεωγραφική Κατανομή Σεισμών (Longitude vs Latitude Scatter Plot)

στειακού συστήματος πέριξ της Σαντορίνης. Η ακριβής τοποθέτηση των επίκεντρων είναι κρίσιμη για την κατανόηση της τρέχουσας κατάστασης και της ενδεχόμενης εξέλιξης της ηφαιστειακής και σεισμικής δραστηριότητας στην περιοχή.

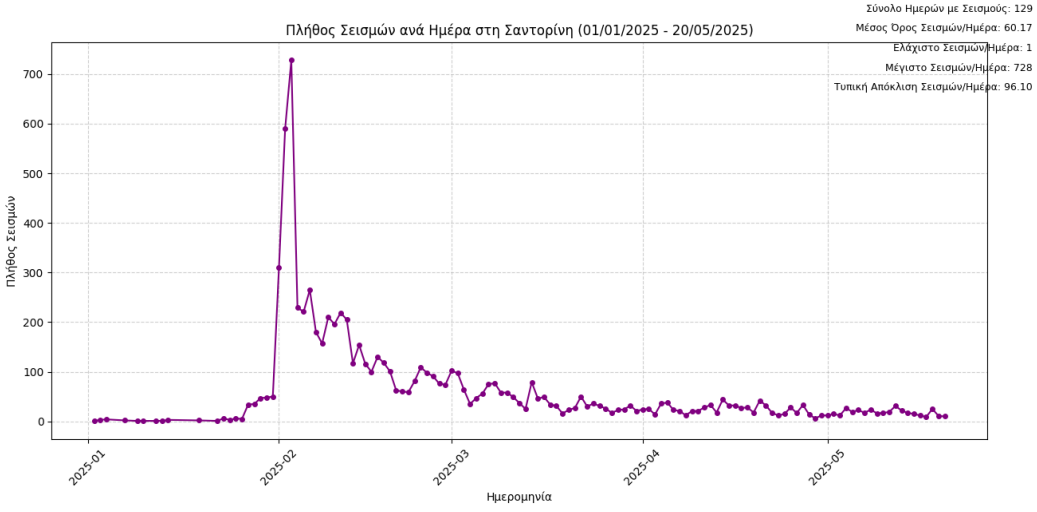
Το διάγραμμα διασποράς που αποτυπώνει τη σχέση μεταξύ του μεγέθους (M_L) και του βάθους (σε χιλιόμετρα) των σεισμών, που καταγράφηκαν στην περιοχή της Σαντορίνης κατά την εξεταζόμενη περίοδο (Σχήμα 4.17). Από την επισκόπηση του διαγράμματος γίνεται άμεσα αντιληπτό, ότι η συντριπτική πλειονότητα των σεισμών είναι συγκεντρωμένη σε μικρά βάθη, κυρίως εντός του εύρους των 0 - 20 χιλιομέτρων. Παράλληλα, οι περισσότεροι σεισμοί χαρακτηρίζονται από μικρά μεγέθη, κυρίως μεταξύ 0.5 M_L και 4.0 M_L.

Είναι ιδιαίτερα σημαντικό να σημειωθεί, ότι οι μεγαλύτεροι σεισμοί που καταγράφηκαν στην περίοδο μελέτης με μέγεθος 4.0 M_L και άνω (φτάνοντας έως τα 5.3 M_L), εμφανίζονται επίσης σε ρηχά βάθη κάτω των 15 - 20 χιλιομέτρων. Αυτό υπερτονίζει την επικινδυνότητα των ρηχών σεισμών στην περιοχή, καθώς η ενέργεια απελευθερώνεται κοντά στην επιφάνεια. Υπάρχουν και ελάχιστοι σεισμοί σε μεγαλύτερα βάθη, που φτάνουν έως και τα 118 χιλιόμετρα. Ωστόσο, αυτοί οι βαθύτεροι σεισμοί είναι γενικά μικρότερου μεγέθους, κυρίως κάτω των 2.0 M_L, και δεν αποτελούν τον κυρίαρχο πληθυσμό της σεισμικής δραστηριότητας στην περιοχή της Σαντορίνης για την εξεταζόμενη περίοδο.



Σχήμα 4.17: Σχέση Μεγέθους και Βάθους Σεισμών (Magnitude vs Depth Scatter Plot)

Το γράφημα που εμφανίζει την ημερήσια καταμέτρηση των σεισμών που καταγράφηκαν στην περιοχή της Σαντορίνης (Σχήμα 4.18). Η σεισμική δραστηριότητα στην αρχή της εξεταζόμενης περιόδου (Ιανουάριος 2025) ήταν εξαιρετικά χαμηλή, με μόλις 0 έως 5 σεισμούς να καταγράφονται

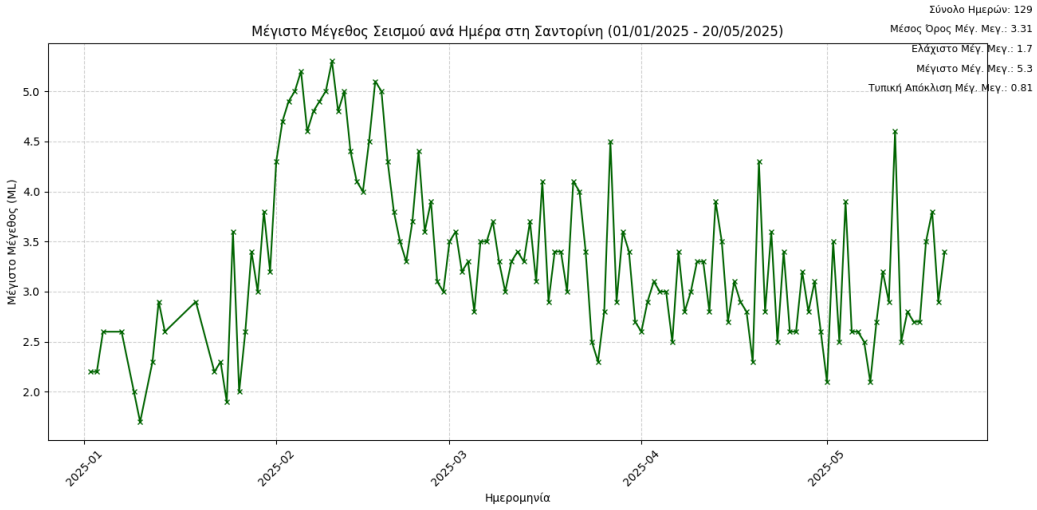


Σχήμα 4.18: Πλήθος Σεισμών ανά Ημέρα

ανά ημέρα. Ωστόσο, προς τα τέλη Ιανουαρίου και στις αρχές Φεβρουαρίου 2025, παρατηρείται μια δραματική και απότομη αύξηση στον αριθμό των ημερήσιων σεισμών. Αυτή η αύξηση οδηγεί σε μια κορύφωση της δραστηριότητας, φτάνοντας στους 728 σεισμούς σε μία μόνο ημέρα, γεγονός που σημαίνει ότι είναι πιθανό ένα "σεισμικό σμήνος" (χωρίς σαφή κύριο σεισμό, αλλά με συνεχή δραστηριότητα), καθώς πρόκειται για ένα σύστημα ηφαιστειακών δομών.

Μετά από αυτή την κορύφωση, το πλήθος των σεισμών αρχίζει να μειώνεται σταδιακά. Η μείωση αυτή είναι αρχικά πιο απότομη και στη συνέχεια επιβραδύνεται. Η ημερήσια συχνότητα διατηρείται σε υψηλά επίπεδα (περίπου 10 - 30 σεισμοί την ημέρα) έως το τέλος της περιόδου μελέτης (Μαΐος 2025) και συγκριτικά με την αρχική περίοδο ηρεμίας. Παρόλο που παρατηρούνται ημερήσιες διακυμάνσεις κατά τη φάση της μείωσης, η γενική τάση παραμένει πτωτική. Αυτά τα φαινόμενα είναι κρίσιμα σε ηφαιστειακές περιοχές, καθώς μπορεί να σχετίζονται με ενδοηφαιστειακές διεργασίες, όπως η κίνηση μάγματος ή η εκτόνωση τεκτονικών τάσεων. Η παρατεταμένη δραστηριότητα, ακόμα και μετά την κορύφωση, καθιστά ζωτικής σημασίας την συνεχή παρακολούθηση του φαινομένου.

Το διάγραμμα γραμμής (Σχήμα 4.19) παρουσιάζει το μέγιστο μέγεθος (M_L) του σεισμού για κάθε ημέρα. Κατά την αρχική περίοδο του Ιανουαρίου 2025, η σεισμική δραστηριότητα χαρακτηρίζεται από πολύ χαμηλά μέγιστα ημερήσια μεγέθη, τα οποία κυμαίνονται κυρίως κάτω από 2.0 M_L. Ωστόσο, παράλληλα με την απότομη αύξηση στο πλήθος των σεισμών, που παρατηρήθηκε στα τέλη Ιανουαρίου και αρχές Φεβρουαρίου 2025, καταγράφηκε και μια αντίστοιχη αύξηση στα μέγιστα ημερήσια μεγέθη. Το μέγιστο μέγεθος που παρατηρήθηκε στην περίοδο είναι 5.3 M_L. Μετά τις αρχικές αυτές κορυφές, τα μέγιστα ημερήσια μεγέθη παρουσιάζουν διακυμάνσεις, αλλά γενικά παραμένουν σε υψηλότερα επίπεδα (μεταξύ 2.5 M_L και 4.0 M_L) σε σύγκριση με την αρχική περίοδο ηρεμίας. Προς τα τέλη της περιόδου μελέτης, τα μέγιστα μεγέθη μειώνονται. Η μείωση δεν είναι τόσο απότομη όσο το πλήθος των σεισμών. Πιθανόν, η συνεχής παρουσία μεσαίου μεγέθους σεισμών να είναι μέρος της εκτόνωσης.



Σχήμα 4.19: Μέγιστο Μέγεθος Σεισμού ανά Ημέρα

Κεφάλαιο 5

Συσταδοποίηση στα Σεισμολογικά Δεδομένα

5.1 Συσταδοποίηση βασιζόμενη στη Διαμέριση (Partition-based Clustering)

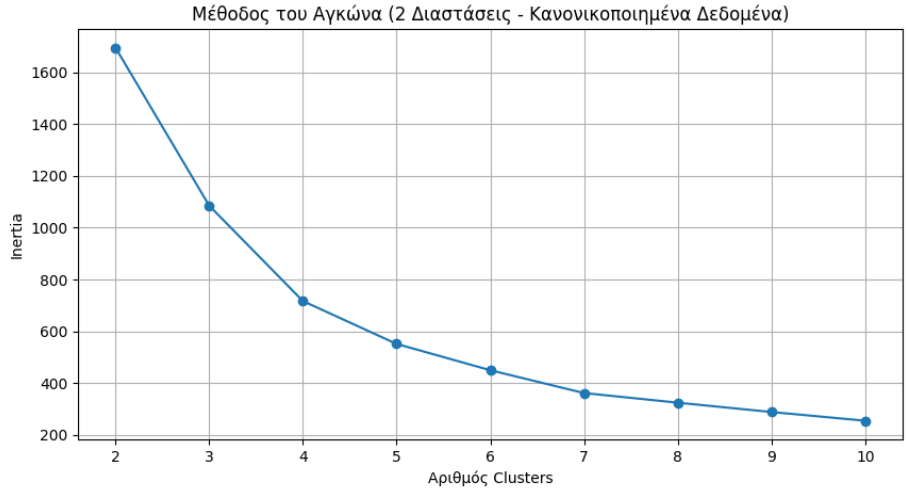
Στο πλαίσιο αυτής της μελέτης αρχικά εξετάζεται ο αλγόριθμος K-Means σε δισδιάστατα δεδομένα, χρησιμοποιώντας ως χαρακτηριστικά τις γεωγραφικές συντεταγμένες (Latitude και Longitude) των σεισμικών γεγονότων. Η παρακάτω ανάλυση παρουσιάζει τα αποτελέσματα του πειράματος. Η διερεύνηση αυτών των χωρικών συσχετίσεων μπορεί να προσφέρει πολύτιμες πληροφορίες για την κατανόηση των σεισμογόνων περιοχών και των γεωλογικών δομών, που τις επηρεάζουν. Σκοπός αυτού του αρχικού πειράματος είναι να θέσει τα θεμέλια για πιο σύνθετη ανάλυση συσταδοποίησης, ενσωματώνοντας στη συνέχεια περισσότερες διαστάσεις των σεισμολογικών δεδομένων (5D).

5.1.1 K-Means 2D

Ο συνολικός κώδικας του αλγορίθμου παρατίθεται στο Παράρτημα Α (βλ. K-Means 2D). Στην αρχή εισάγονται όλες οι απαραίτητες βιβλιοθήκες. Δημιουργείται ένας φάκελος για την αποθήκευση των γραφημάτων και άλλων αρχείων, που παράγονται από τον κώδικα. Τα δεδομένα σεισμών και τεκτονικών πλακών φορτώνονται από CSV αρχεία: EarthquakesGr.csv - είναι ο κατάλογος των σεισμών και TectonicPlates.csv - είναι το αρχείο, που εμπεριέχει τις συντεταγμένες των ορίων των τεκτονικών πλακών.

Η στήλη ώρας μετατρέπεται σε datetime και οι εγγραφές με ελλιπείς συντεταγμένες αφαιρούνται (αν υπάρχουν). Στη συνέχεια, επιλέγονται οι στήλες γεωγραφικού πλάτους και μήκους για την ανάλυση σε δύο διαστάσεις και εμφανίζονται πληροφορίες για τους τύπους δεδομένων του DataFrame earthquakes. Ακολούθως, τα επιλεγμένα χαρακτηριστικά κανονικοποιούνται χρησιμοποιώντας Standard Scaler[1]. Αν και οι μεταβλητές έχουν παρόμοιο εύρος, γίνεται κανονικοποίηση για ακριβέστερη απόσταση. Χρησιμοποιούνται η μέθοδος του αγκώνα (Elbow Method) και το Silhouette Score για την εκτίμηση του βέλτιστου αριθμού συστάδων (clusters). Δημιουργούνται και αποθηκεύονται γραφήματα για τη μέθοδο του αγκώνα και το Silhouette Score.

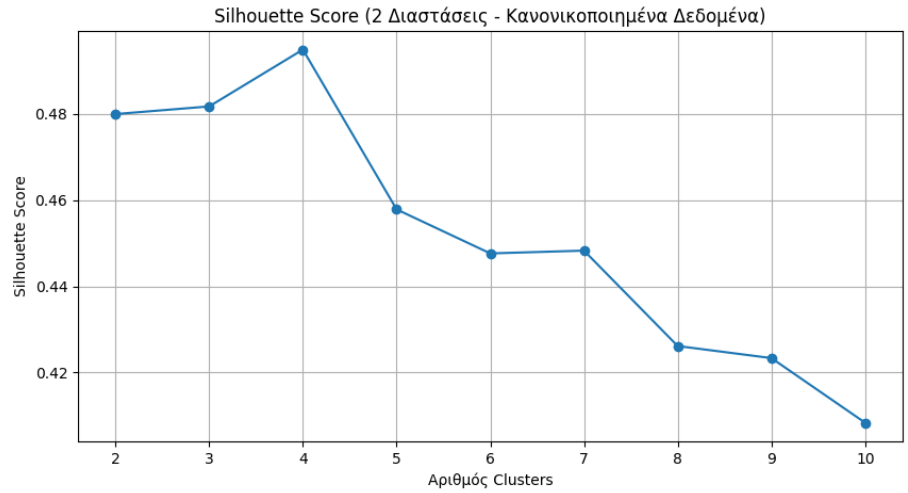
[1] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`



Σχήμα 5.1: Η Μέθοδος του Αγκώνα (Elbow Method) K-Means 2D

Το γράφημα της μεθόδου του αγκώνα (Σχήμα 5.1) απεικονίζει το inertia για διαφορετικούς αριθμούς clusters (από 2 έως 10). Είναι εμφανές, ότι η πτώση του inertia γίνεται πολύ πιο αργή μετά την τιμή *k=4*, κάτι που σημαίνει ότι στα 4 clusters υπάρχει μια καλή ισορροπία μεταξύ συμπαγών ομάδων και διαχωρισμού. Για περισσότερα clusters, το όφελος από τη μείωση του inertia θα ήταν μικρό και πιθανώς περιττό.

Το γράφημα του Silhouette Score (Σχήμα 5.2) εμφανίζει τον μέσο συντελεστή Silhouette για διαφορετικούς αριθμούς clusters (από 2 έως 10). Στο γράφημα παρατηρείται, ότι η υψηλότερη τιμή του Silhouette Score είναι για 4 clusters, η οποία είναι περίπου 0.49. Μετά, το Silhouette Score μειώνεται, υποδεικνύοντας ότι η προσθήκη περισσότερων clusters δεν βελτιώνει τον διαχωρισμό τους. Το γράφημα επιβεβαιώνει το συμπέρασμα της μεθόδου του αγκώνα.



Σχήμα 5.2: Silhouette Score K-Means 2D

Ο αλγόριθμος K-Means εφαρμόζεται με τον επιλεγμένο αριθμό clusters ($optimal\_k$). Ενδεικτικά, δίνεται παρακάτω το τμήμα του κώδικα για την εκτέλεση του αλγορίθμου.

Κώδικας εκτέλεσης K-Means:

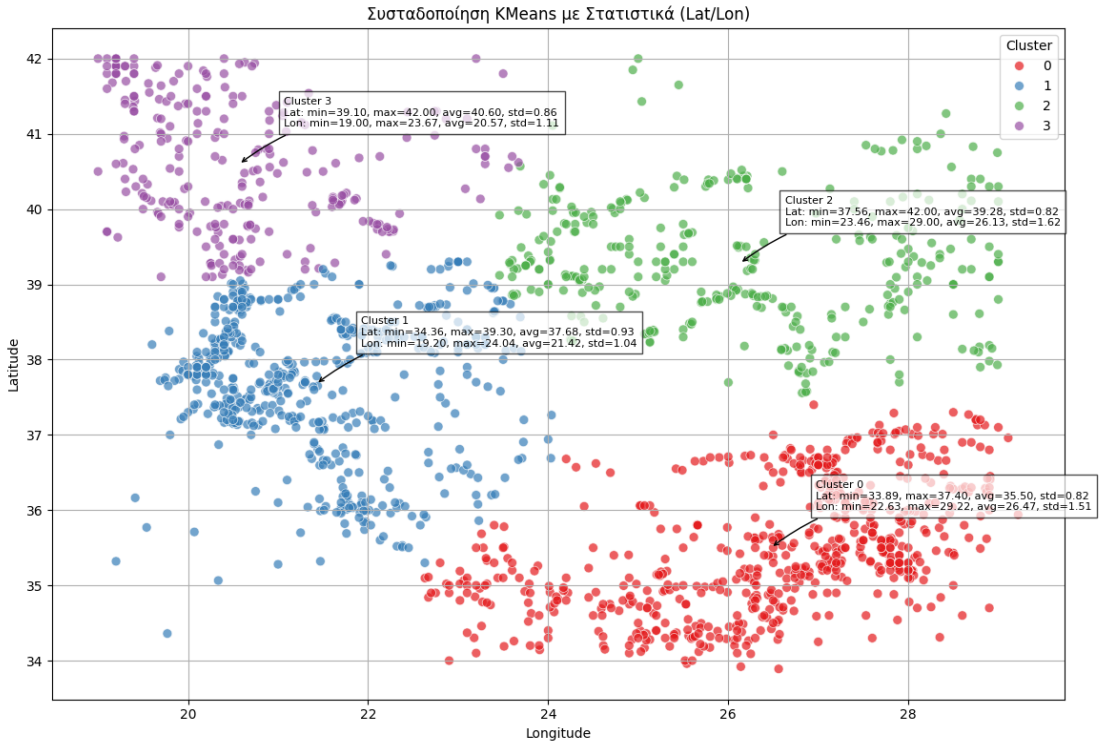
```
# Εκτέλεση KMeans με το επιλεγμένο k
kmeans = KMeans(n_clusters=optimal_k, random_state=0, n_init=10)
earthquakes['Cluster'] = kmeans.fit_predict(scaled_data)
```

Το $n\_init = 10$ καθορίζει τον αριθμό επαναλήψεων με διαφορετικές, τυχαίες αρχικές τιμές για τα κεντρικά σημεία (centroids) στον αλγόριθμο. Ο K-Means είναι ευαίσθητος στον τρόπο με τον οποίο ορίζονται αρχικά τα κέντρα των clusters, και μπορεί να συγκλίνει σε διαφορετικά τοπικά ελάχιστα, ανάλογα με αυτές τις αρχικές τιμές. Ο αλγόριθμος εκτελείται 10 φορές ξεχωριστά και στο τέλος επιλέγεται εκείνο, που δίνει το μικρότερο συνολικό άθροισμα των τετραγώνων των αποστάσεων (inertia). Αυτή η προσέγγιση βελτιώνει τη σταθερότητα και την αξιοπιστία της συσταδοποίησης, εξασφαλίζοντας ότι το αποτέλεσμα είναι όσο το δυνατόν πιο κοντά στο βέλτιστο και δεν εξαρτάται τυχαία από μια μεμονωμένη αρχικοποίηση.

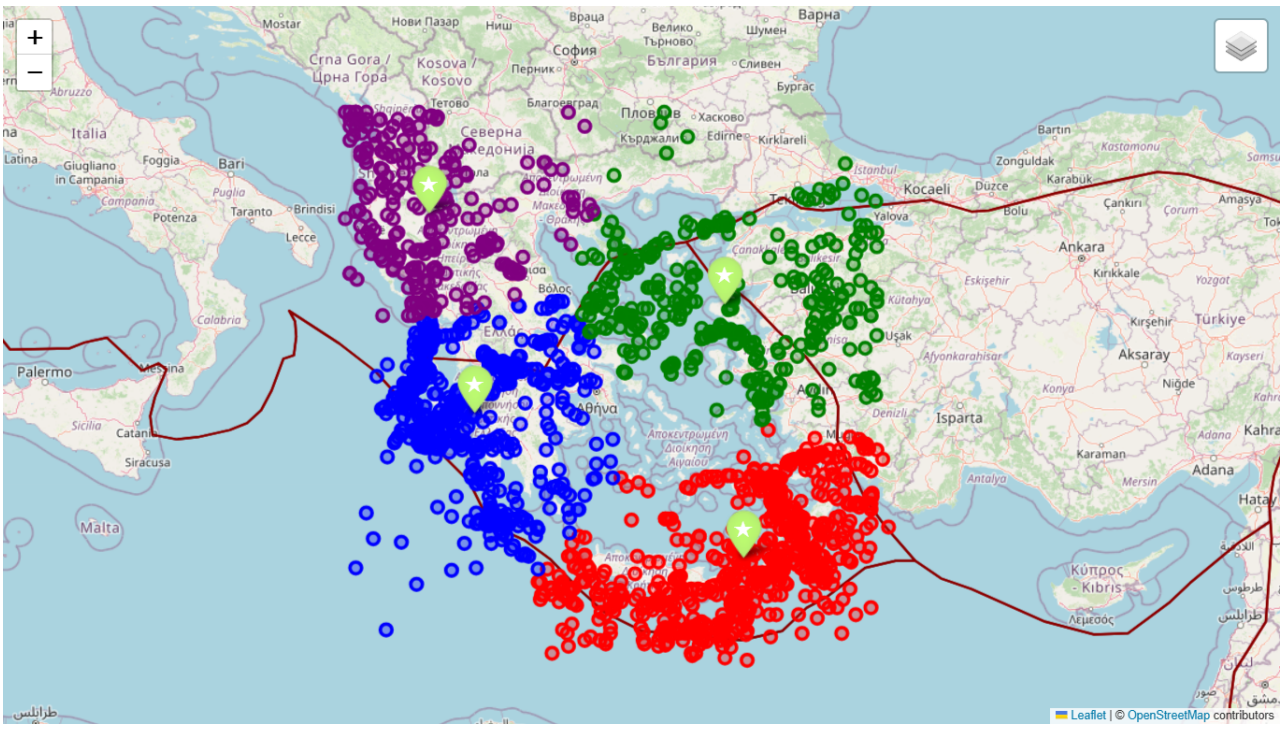
Στη συνέχεια, δημιουργείται και αποθηκεύεται σε ένα αρχείο CSV ένας πίνακας με Περιγραφική Στατιστική, η οποία στοχεύει στην ποσοτική ανάλυση των χαρακτηριστικών κάθε συστάδας (cluster), που προέκυψε από τον αλγόριθμο K-Means. Ουσιαστικά, εξυπηρετεί στην ερμηνεία των clusters που δημιουργήθηκαν, καθώς έτσι, γίνονται κατανοητά τα χαρακτηριστικά κάθε ομάδας σεισμών ποσοτικά.

Σε επόμενο στάδιο, δημιουργείται και αποθηκεύεται ένα Διάγραμμα Διασποράς (Scatter Plot) των σεισμών, χρωματισμένων ανά συστάδα (cluster), και προστίθενται σχόλια με τα βασικά στατιστικά (min, max, avg, std) για το γεωγραφικό πλάτος και μήκος κάθε cluster (Σχήμα 5.3). Επίσης, δημιουργείται και αποθηκεύεται ένα απλό Διάγραμμα Διασποράς (Scatter Plot) των clusters χωρίς τα στατιστικά.

Κατόπιν τούτου, αποθηκεύεται το DataFrame earthquakes με τη στήλη των clusters σε ένα CSV αρχείο. Τέλος, δημιουργείται ένας διαδραστικός χάρτης Folium. Ο διαδραστικός χάρτης απεικονίζει τα clusters των σεισμών (χρωματισμένοι κύκλοι) στην ευρύτερη περιοχή της Ελλάδας, επιτρέποντας την εξερεύνηση των γεωγραφικών τους θέσεων με δυνατότητα zoom και μετακίνησης. Εμφανίζει, επίσης, τα όρια των τεκτονικών πλακών (κόκκινες γραμμές) και τα κέντρα των clusters (πράσινες καρφίτσες). Κάνοντας κλικ σε κάθε σεισμό, εμφανίζεται ένα popup με πληροφορίες για το cluster, το μέγεθος, το βάθος και την ώρα γένεσής του. Επιπρόσθετα, ο χάρτης παρέχει τη δυνατότητα επιλογής διαφορετικών στυλ υποβάθρου (tiles), μέσω ενός ενσωματωμένου μενού, προσφέροντας ποικιλία στην οπτική απεικόνιση των δεδομένων (Σχήμα 5.4). Ο χάρτης αποθηκεύεται ως HTML αρχείο.



Σχήμα 5.3: Διάγραμμα Διασποράς (Scatter Plot) με Στατιστικά K-Means 2D



Σχήμα 5.4: Διαδραστικός Χάρτης K-Means 2D

Η Περιγραφική Στατιστική σε συνδυασμό με τις οπτικοποιήσεις (Διάγραμμα Διασποράς και Διαδραστικός Χάρτης) παρέχουν μια ποσοτική βάση, για να γίνουν κατανοητές οι χωρικές και (σε κάποιο βαθμό) τα άλλα χαρακτηριστικά των σεισμών για κάθε cluster. Η μελέτη και ερμηνεία των αποτελεσμάτων συμβάλλει στην αναγνώριση περιοχών με διαφορετική σεισμική συμπεριφορά. Η συσταδοποίηση με βάση το γεωγραφικό πλάτος και μήκος δημιουργεί 4 clusters, που αντιστοιχούν σε διαφορετικές γεωγραφικές ζώνες με διακριτά χαρακτηριστικά, τόσο σε επίπεδο θέσης (πλάτος και μήκος), όσο και σε επίπεδο σεισμικών παραμέτρων (βάθος και μέγεθος σεισμού).

Ερμηνεία Αποτελεσμάτων K-Means 2D

- **Cluster 0:** περιλαμβάνει κυρίως σεισμούς στο νότιο και ανατολικό τμήμα της περιοχής μελέτης (μέσο πλάτος ≈ 35.5°, μήκος ≈ 26.5°). Πρόκειται για την πολυπληθέστερη ομάδα (644 γεγονότα), με τον υψηλότερο μέσο όρο βάθους (≈ 29.9 km), γεγονός που υποδηλώνει έντονη βαθιά σεισμικότητα, ενδεχομένως υποθαλάσσιας προέλευσης.

- **Cluster 1:** αντιπροσωπεύει σεισμούς στη δυτική - νοτιοκεντρική ενδοχώρα (πλάτος ≈ 37.7°, μήκος ≈ 21.4°), με μικρότερο μέσο βάθος (≈ 18.2 km) και σχετικά υψηλό μέσο μέγεθος σεισμού (≈ 4.86 M_L). Η ομάδα περιλαμβάνει 576 σεισμικά γεγονότα και υποδηλώνει δραστηριότητα σε ηπειρωτικές, κυρίως, περιοχές.

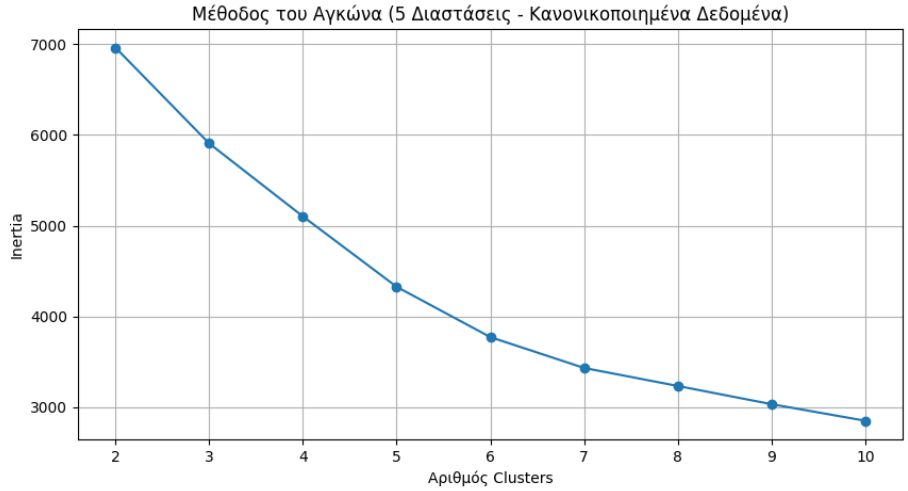
- **Cluster 2:** περιλαμβάνει σεισμούς στο βορειοανατολικό τμήμα (πλάτος ≈ 39.3°, μήκος ≈ 26.1°), με το μεγαλύτερο μέγιστο μέγεθος σεισμού (7.0 M_L). Το μέσο βάθος είναι περίπου 17.1 km και η διακύμανση σχετικά περιορισμένη. Πρόκειται για περιοχή με έντονη σεισμική δραστηριότητα.

- **Cluster 3:** αφορά το βόρειο τμήμα της περιοχής μελέτης (μέσο πλάτος ≈ 40.6°, μήκος ≈ 20.6°), με τα πιο ρηχά σεισμικά γεγονότα (μέσο βάθος ≈ 13.9 km). Παρουσιάζει τη μικρότερη πληθυσμιακή συγκέντρωση σεισμών (259).

5.1.2 K-Means 5D

Ο συνολικός κώδικας του αλγορίθμου παρατίθεται στο Παράρτημα Α (βλ. K-Means 5D). Αρχικά, ο κώδικας προετοιμάζει το περιβάλλον εισάγοντας τις απαραίτητες βιβλιοθήκες, που θα χρησιμοποιηθούν για τη διαχείριση δεδομένων, τις μαθηματικές πράξεις, τη δημιουργία γραφημάτων και τον ίδιο τον αλγόριθμο K-Means. Δημιουργεί έναν ειδικό φάκελο, για να αποθηκεύσει όλα τα αποτελέσματα της ανάλυσης. Στη συνέχεια, φορτώνει τα δεδομένα των σεισμών από το αρχείο CSV (EarthquakesGr.csv). Ένα σημαντικό βήμα εδώ είναι ο καθαρισμός των δεδομένων. Ο χρόνος γένεσης των σεισμών μετατρέπεται σε κατάλληλη μορφή και αφαιρούνται τυχόν σεισμοί, που δεν έχουν πλήρη στοιχεία στις βασικές διαστάσεις. Επιπλέον, το έτος του σεισμού εξάγεται και προστίθεται ως νέα διάσταση για την ανάλυση.

Πριν την εφαρμογή του K-Means, ο κώδικας διασφαλίζει, ότι όλες οι διαστάσεις έχουν την ίδια βαρύτητα στην ανάλυση μέσω της κανονικοποίησης. Αυτό σημαίνει, ότι οι τιμές κάθε διάστασης μετατρέπονται σε μια κοινή κλίμακα, ώστε καμία διάσταση π.χ., το βάθος, να μην επηρεάζει περισσότερο το αποτέλεσμα, λόγω των μεγαλύτερων αριθμητικών της τιμών σε σχέση με το μέγεθος. Όπως και στην περίπτωση του K-Means 2D, για να βρεθεί ο βέλτιστος αριθμός ομάδων (clusters),

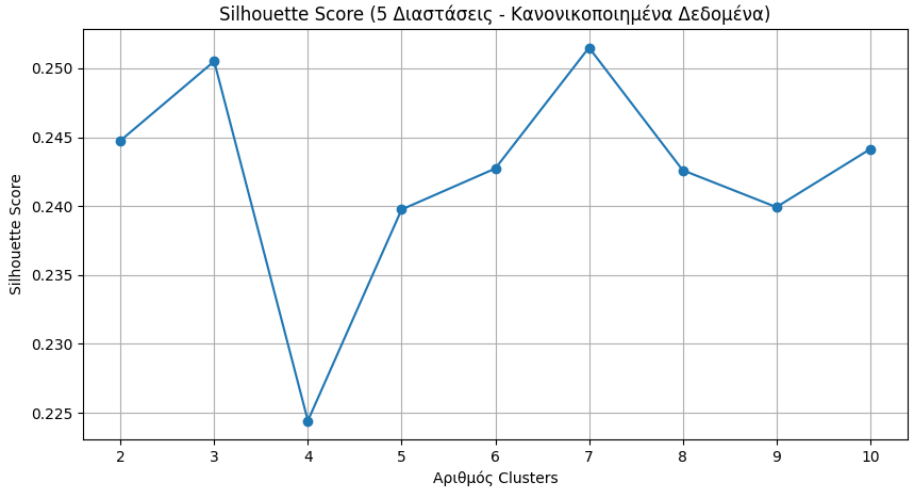


Σχήμα 5.5: Η Μέθοδος του Αγκώνα (Elbow Method) K-Means 5D

στις οποίες θα χωριστούν οι σεισμοί, ο κώδικας χρησιμοποιεί δύο τεχνικές: τη μέθοδο του αγκώνα και το Silhouette Score.

Η μέθοδος του αγκώνα βοηθά στην οπτική αναγνώριση του σημείου, στο οποίο η προσθήκη περισσότερων ομάδων δεν προσφέρει σημαντική βελτίωση στη συνοχή των ομάδων (Σχήμα 5.5). Όπως έχει ήδη τονιστεί, πρόκειται για μία μέθοδο, που χαρακτηρίζεται από υποκειμενικότητα, καθώς πολλές φορές η καμπή δεν είναι πάντα απόλυτα σαφής. Εξετάζοντας το γράφημα της αδράνειας (inertia) έναντι του αριθμού των συστάδων, παρατηρείται μια σημαντική μείωση της αδράνειας κατά την αύξηση του k από 2 έως 5. Πιο συγκεκριμένα, υπάρχει μια αξιοσημείωτη πτώση από 2 σε 3 συστάδες, και μια εμφανής καμπή γύρω στις 3 με 5 συστάδες. Πέρα από αυτό το σημείο η πτώση γίνεται πιο ήπια, πράγμα που σημαίνει, ότι η προσθήκη επιπλέον συστάδων δεν προσφέρει πλέον ανάλογη βελτίωση στη συνοχή των δεδομένων εντός των συστάδων. Η αξιολόγηση του βέλτιστου αριθμού συστάδων *(k)* συμπληρώθηκε με την ανάλυση του Silhouette Score. Το Silhouette Score είναι μια κρίσιμη μετρική, καθώς ποσοτικοποιεί πόσο καλά τα σημεία έχουν ομαδοποιηθεί εντός της δικής τους συστάδας (συνοχή) και πόσο καλά έχουν διαχωριστεί από τις άλλες συστάδες (διαχωρισμός). Μια υψηλότερη τιμή Silhouette Score υποδεικνύει σαφέστερα διαμορφωμένες και διαχωρισμένες συστάδες.

Από το γράφημα του δείκτη Silhouette Score (Σχήμα 5.6), παρατηρούνται κορυφές στις τιμές για $k=3$ (με τιμή περίπου 0.250) και $k=7$ (επίσης με τιμή περίπου 0.250), ενώ για $k=4$ η τιμή ήταν χαμηλότερη (περίπου 0.220). Αυτό, ενδεχομένως, υποδηλώνει ότι τόσο τα 3 όσο και τα 7 clusters αποτελούν ισχυρές υποψήφιες επιλογές, μεγιστοποιώντας την εσωτερική συνοχή και τον εξωτερικό διαχωρισμό. Ωστόσο, παρά τη μέγιστη τιμή στο $k=7$, η διαφορά στις τιμές του δείκτη Silhouette Score ανάμεσα στα 5 (περίπου 0.240) και 7 clusters είναι σχετικά μικρή. Λαμβάνοντας υπόψιν την επιλογή του $k=5$ από τη μέθοδο του αγκώνα – η οποία έδειξε ότι πέρα από αυτό το σημείο η πτώση της αδράνειας γίνεται ηπιότερη – η διατήρηση πέντε συστάδων αποτελεί μια λογική και ισορροπημένη επιλογή. Αυτή η επιλογή συνδυάζει ένα ικανοποιητικό επίπεδο συνοχής εντός των ομάδων με επαρκή διαχωρισιμότητα, χωρίς να εισάγει υπερβολική πολυπλοκότητα, που θα μπορούσε να κατα-



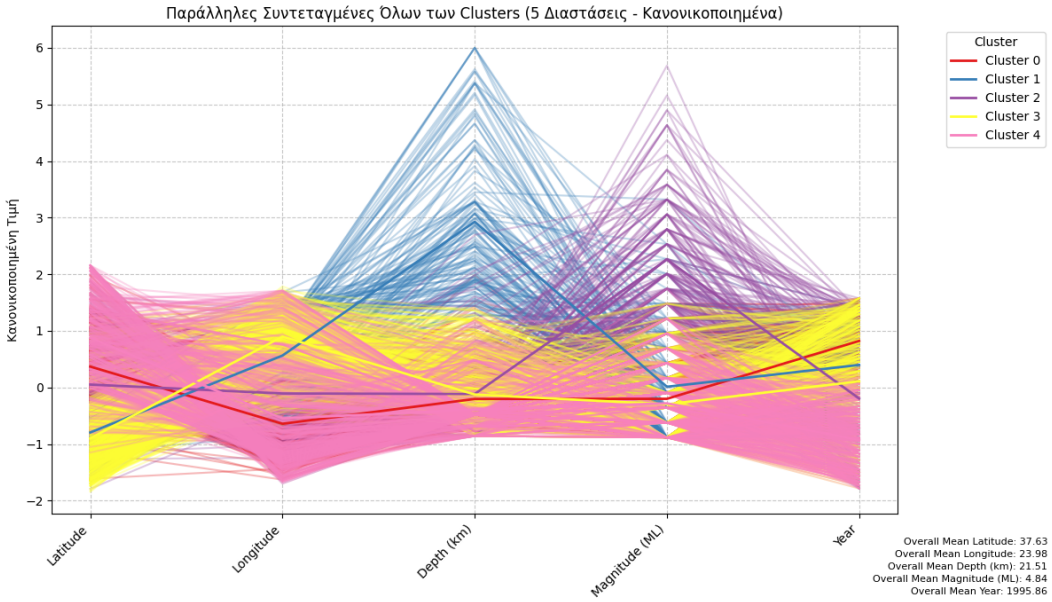
Σχήμα 5.6: Silhouette Score K-Means 5D

στήσει την ερμηνεία των συστάδων πιο δύσκολη.

Με βάση αυτές τις μεθόδους, ο κώδικας προχωρά επιλέγοντας 5 ομάδες ως τον βέλτιστο αριθμό (μετά την επισκόπηση των δύο γραφημάτων). Έπειτα, εφαρμόζει τον αλγόριθμο K-Means για να ομαδοποιήσει τους σεισμούς και προσθέτει μια νέα στήλη στα δεδομένα, που δείχνει σε ποια συστάδα ανήκει κάθε σεισμός. Αφού δημιουργηθούν οι συστάδες (clusters), ο κώδικας υπολογίζει περιγραφικά στατιστικά για κάθε ομάδα, χρησιμοποιώντας τις αρχικές, μη κανονικοποιημένες τιμές των δεδομένων. Αυτό είναι κρίσιμο, για να γίνουν κατανοητά τα χαρακτηριστικά κάθε ομάδας – για παράδειγμα, μια ομάδα μπορεί να περιλαμβάνει κυρίως επιφανειακούς σεισμούς μικρού μεγέθους, ενώ μια άλλη βαθύτερους σεισμούς μεσαίου μεγέθους. Αυτά τα στατιστικά αποθηκεύονται και σε ένα αρχείο CSV για μελλοντική αναφορά.

Τέλος, ο κώδικας δημιουργεί μια σειρά από οπτικοποιήσεις, για να βοηθήσει στην κατανόηση των αποτελεσμάτων της συσταδοποίησης. Περιλαμβάνονται στατικά γραφήματα παράλληλων συντεταγμένων. Τα συγκεκριμένα γραφήματα δείχνουν, πώς οι διάφορες ομάδες κατανέμονται σε όλες τις διαστάσεις. Επίσης, δημιουργεί ένα διαδραστικό 3D Scatter Plot (Plotly) με Animation για την 5η Διάσταση (Ετος). Αυτό το διαδραστικό γράφημα είναι ιδιαίτερα χρήσιμο, καθώς επιτρέπει να εξερευνηθεί η εξέλιξη των σεισμών και των clusters ανά έτος. Όλα τα γραφήματα αποθηκεύονται ως εικόνες ή αρχεία HTML στον προκαθορισμένο φάκελο, μαζί με το τελικό αρχείο CSV που περιέχει τα αρχικά δεδομένα των σεισμών με την προσθήκη των ομάδων στις οποίες ανήκουν.

Ερμηνεία Αποτελεσμάτων K-Means 5D
Τα γραφήματα παράλληλων συντεταγμένων (Parallel Coordinates) αποτελούν ένα εξαιρετικό εργαλείο οπτικοποίησης για την κατανόηση της συμπεριφοράς των δεδομένων σε πολυδιάστατους χώρους και την ανακάλυψη κρυμμένων προτύπων εντός κάθε συστάδας (cluster). Το συγκεκριμένο γράφημα (Σχήμα 5.7) δείχνει ένα μοτίβο σύνδεσης, μεταξύ των πέντε διαστάσεων (Latitude, Longitude, Depth (km), Magnitude (M_L), Year) για τα μέλη του εκάστοτε cluster. Η πυκνότητα των



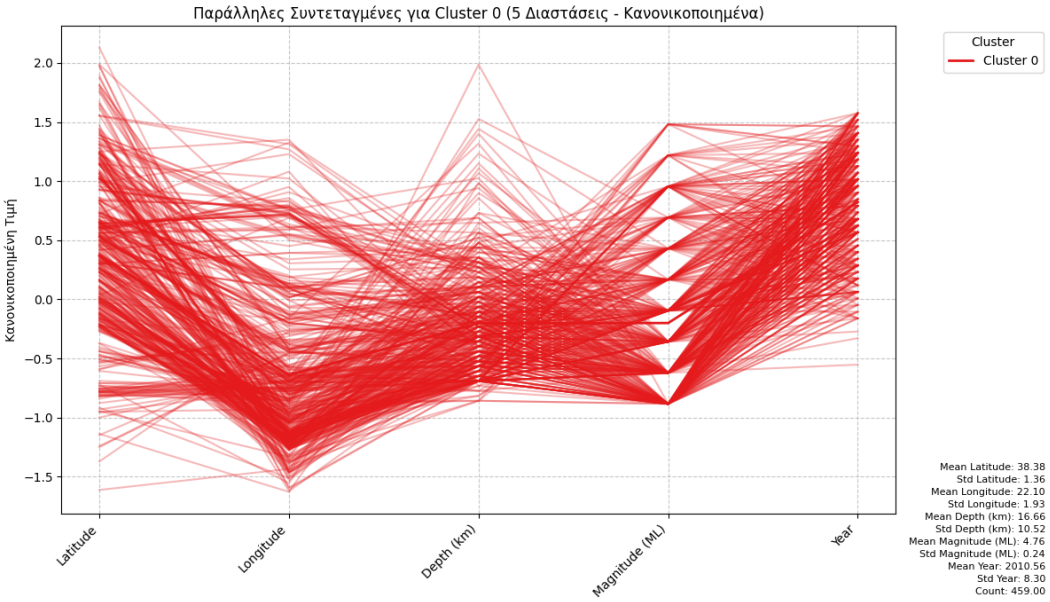
Σχήμα 5.7: Γράφημα Παράλληλων Συντεταγμένων (Parallel Coordinates) K-Means 5D

γραμμών σε συγκεκριμένα εύρη του κάθε άξονα μας δείχνει πού συγκεντρώνονται τα δεδομένα για τη συγκεκριμένη διάσταση εντός του cluster.

Ενώ το συνολικό γράφημα των παράλληλων συντεταγμένων προσφέρει μια γενική επισκόπηση της συμπεριφοράς των δεδομένων σε όλες τις συστάδες, η πολυπλοκότητα των πολλών γραμμών καθιστά την άμεση ερμηνεία των ειδικών χαρακτηριστικών κάθε συστάδας αρκετά δύσκολη. Για να επιτευχθεί μια βαθύτερη αντίληψη των προτύπων, που αναδύθηκαν από την συσταδοποίηση, ο κώδικας παράγει μεμονωμένα γραφήματα παράλληλων συντεταγμένων για κάθε συστάδα ξεχωριστά. Με αυτό τον τρόπο, είναι εφικτή η εστίαση στην πυκνότητα και τη διασπορά των γραμμών εντός της κάθε συστάδας.

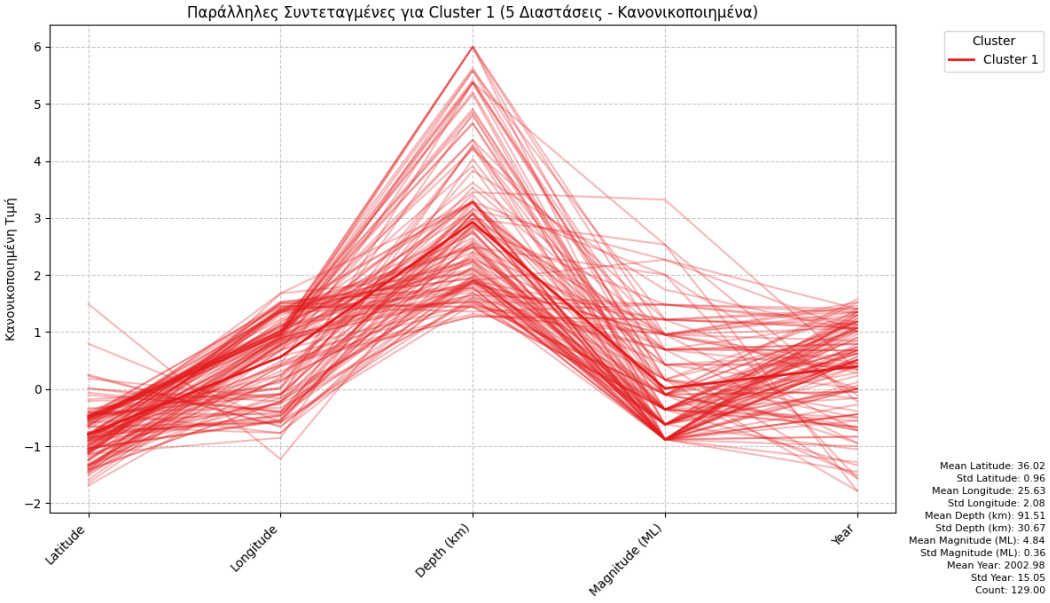
Η ανάλυση των μεμονωμένων γραφημάτων, σε συνδυασμό με τα αποτελέσματα της Περιγραφικής στατιστικής, που υπολογίστηκαν για κάθε συστάδα (μέσες τιμές, τυπικές αποκλίσεις κ.λπ.), παρέχει τη δυνατότητα η ερμηνεία να γίνει με μεγαλύτερη ακρίβεια. Μέσα από αυτή τη διαδικασία, επιχειρείται η περιγραφή των διακριτών όψεων κάθε συστάδας, αναδεικνύοντας τις ομοιότητες μεταξύ των σεισμών εντός της ίδιας ομάδας και τις διαφορές τους από σεισμούς σε άλλες ομάδες, βάσει των πέντε διαστάσεων που εξετάστηκαν.

- Το **Cluster 0** (Σχήμα 5.8) χαρακτηρίζεται από μια σχετικά ευρεία γεωγραφική κατανομή. Οι σεισμοί αυτής της συστάδας δεν περιορίζονται αυστηρά σε μία συγκεκριμένη περιοχή, αν και ενδέχεται να εμφανίζουν κάποιες συγκεντρώσεις, χωρίς ωστόσο σαφή οριοθέτηση. Ως προς το βάθος (Depth km), το συγκεκριμένο cluster περιλαμβάνει σεισμούς με αξιοσημείωτη διακύμανση, καλύπτοντας ένα ευρύ φάσμα από πολύ ρηχούς (αρνητικές κανονικοποιημένες τιμές) έως βαθύτερους σεισμούς (θετικές κανονικοποιημένες τιμές). Αυτό υποδεικνύει μια ποικιλομορφία ως προς το βάθος των σεισμών εντός αυτής της ομάδας. Αναφορικά με το μέγεθος (Magnitude M_L), η πλειοψηφία των σεισμών στο Cluster 0 τείνει να έχει χαμηλές έως μέτριες τιμές, με τις κανονικοποιημένες τιμές τους να συγκεντρώνονται στο κάτω μέρος



Σχήμα 5.8: Γράφημα Παράλληλων Συντεταγμένων Cluster 0 K-Means 5D

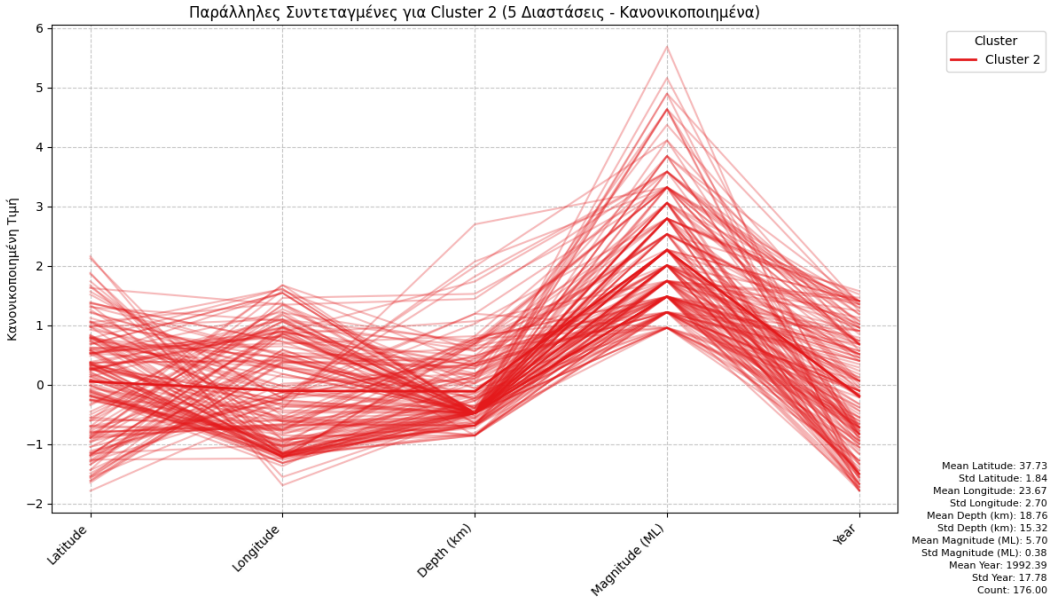
του άξονα, δηλαδή προς τις αρνητικές τιμές. Αυτό υποδηλώνει, ότι πρόκειται κυρίως για μικρότερους σε μέγεθος σεισμούς. Τέλος, η χρονική διάσταση (Year) δείχνει ότι το Cluster 0 καλύπτει ένα ευρύ φάσμα ετών, γεγονός που μαρτυρά, ότι η δραστηριότητα που περιγράφει δεν περιορίζεται σε κάποια συγκεκριμένη χρονική περίοδο, αλλά εμφανίζεται διαχρονικά. Με αρωγό και την Περιγραφική Στατιστική, γίνεται αντιληπτό ότι υπάρχει πρόσφατη (τα τελευταία έτη) δραστηριότητα.



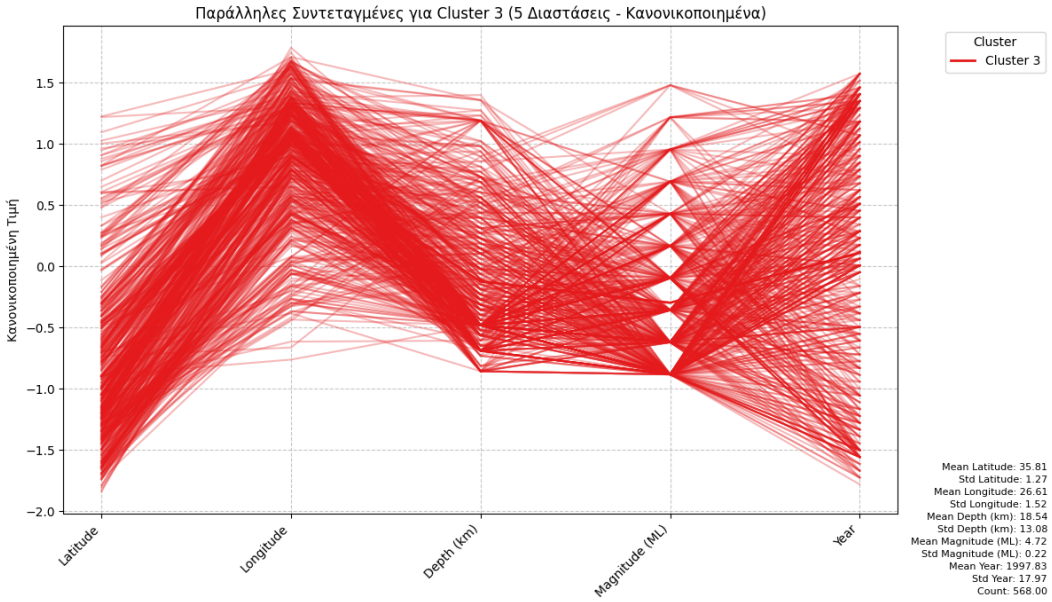
Σχήμα 5.9: Γράφημα Παράλληλων Συντεταγμένων Cluster 1 K-Means 5D

- Το **Cluster 1** (Σχήμα 5.9) παρουσιάζει μια πιο συγκεντρωμένη γεωγραφική κατανομή σε σύγκριση με το Cluster 0. Αυτό γίνεται εμφανές από τις γραμμές των παράλληλων συντεταγμένων, οι οποίες διασχίζουν στενότερα εύρη τιμών στους άξονες του Γεωγραφικού Πλάτους (Latitude) και του Γεωγραφικού Μήκους (Longitude). Ως προς το βάθος (Depth km), οι τιμές των σεισμών του Cluster 1 τείνουν να συγκεντρώνονται προς τα μέτρια έως μεγαλύτερα βάθη, με τις κανονικοποιημένες τιμές να βρίσκονται κυρίως στην θετική πλευρά του άξονα. Στην παράμετρο του μεγέθους (Magnitude M_L), παρατηρείται μια σαφής συγκέντρωση προς τις μέτριες έως υψηλές τιμές (ένα γεγονός 6.1 M_L), γεγονός που διαφοροποιεί το Cluster 1 από το Cluster 0, το οποίο περιελάμβανε κυρίως μικρότερους σεισμούς. Τέλος, και αυτό το cluster φαίνεται να καλύπτει καλύπτει ένα ευρύ χρονικό φάσμα (Year), δείχνοντας ότι η σεισμική δραστηριότητα, που περιγράφει δεν είναι χρονικά περιορισμένη σε κάποια συγκεκριμένη περίοδο.

- Το **Cluster 2** (Σχήμα 5.10) ξεχωρίζει και αυτό για την έντονη γεωγραφική του συγκέντρωση. Είναι σαφές, ότι τόσο οι τιμές του Γεωγραφικού Πλάτους (Latitude), όσο και του Γεωγραφικού Μήκους (Longitude) κινούνται σε ένα πολύ στενό εύρος. Αυτό δείχνει, πως οι σεισμοί αυτής της συστάδας προέρχονται από μια ιδιαίτερα περιορισμένη και καλά καθορισμένη γεωγραφική ζώνη. Ως προς το βάθος (Depth km), το Cluster 2 παρουσιάζει μια ξεκάθαρη τάση προς τα ρηχά βάθη, με τις κανονικοποιημένες τιμές να βρίσκονται κυρίως στην αρνητική πλευρά του άξονα. Αυτό σημαίνει ότι οι σεισμοί σε αυτή την ομάδα εκδηλώνονται κοντά στην επιφάνεια της Γης. Αναφορικά με το μέγεθος (Magnitude M_L), οι σεισμοί του Cluster 2 συγκεντρώνονται σε υψηλότερες τιμές (ένα γεγονός 7.0 M_L), υποδηλώνοντας ότι πρόκειται κατά κύριο λόγο για μέτριους έως μεγάλους σεισμούς. Ο χρόνος (Year) έχει και σε αυτή την περίπτωση ευρύτητα.

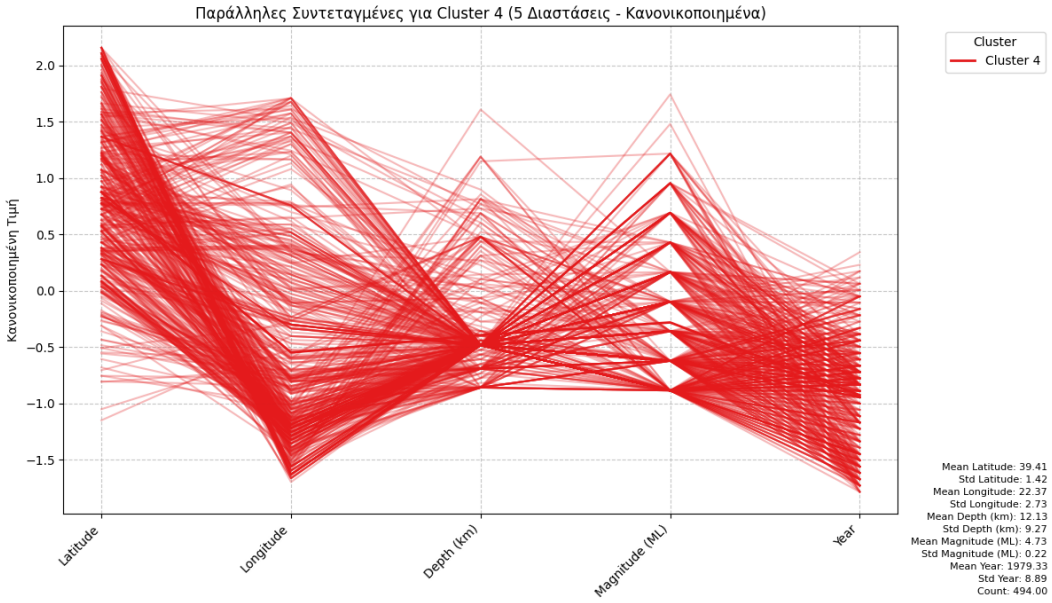


Σχήμα 5.10: Γράφημα Παράλληλων Συντεταγμένων Cluster 2 K-Means 5D



Σχήμα 5.11: Γράφημα Παράλληλων Συντεταγμένων Cluster 3 K-Means 5D

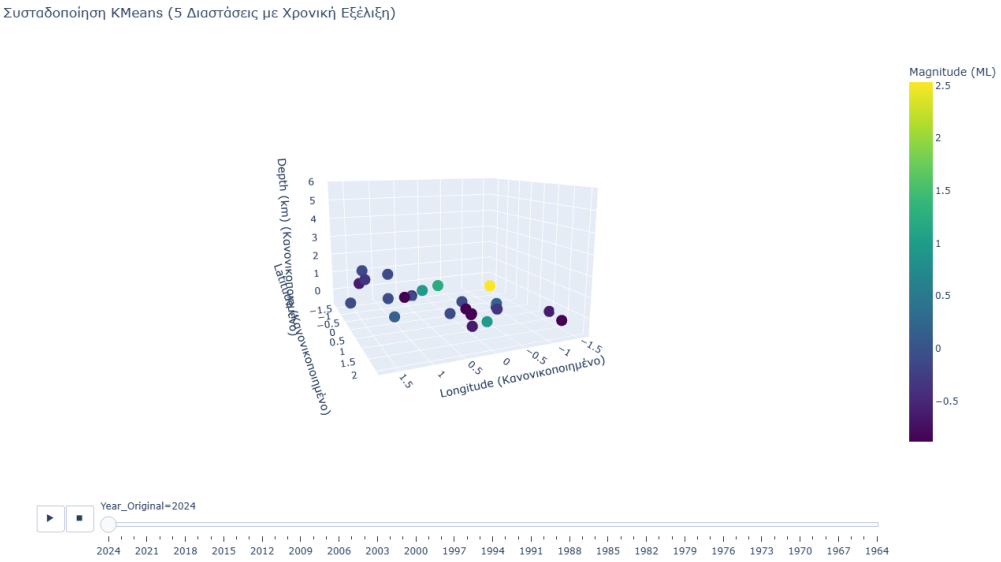
- Το **Cluster 3** (Σχήμα 5.11) χαρακτηρίζεται από ένα ευρύ γεωγραφικό εύρος, παρόμοιο με αυτό που παρατηρήθηκε στο Cluster 0, γεγονός που υποδηλώνει μια διασπορά των σεισμών του σε μια μεγαλύτερη περιοχή. Ως προς το βάθος (Depth km), το Cluster 3 φαίνεται να περιλαμβάνει κυρίως ρηχούς σεισμούς, παρουσιάζοντας μια τάση ανάλογη με αυτή του Cluster 2. Στην παράμετρο του μεγέθους (Magnitude M_L), υπάρχει μια σαφής συγκέντρωση σε χαμηλές έως μέτριες τιμές, υποδεικνύοντας ότι αυτή η ομάδα αποτελείται κυρίως από ασθενείς σεισμούς. Τέλος, και το Cluster 3 δεν περιορίζεται σε συγκεκριμένες χρονικές περιόδους.



Σχήμα 5.12: Γράφημα Παράλληλων Συντεταγμένων Cluster 4 K-Means 5D

- Το **Cluster 4** (Σχήμα 5.12) εμφανίζει μια σχετικά ευρεία γεωγραφική περιοχή. Όσον αφορά το βάθος (Depth km), φαίνεται να περιλαμβάνει κυρίως σεισμούς σε ρηχά βάθη. Σχετικά με το μέγεθος (Magnitude M_L), οι τιμές συγκεντρώνονται χαμηλά προς μέτρια επίπεδα. Αν και το Cluster 4 καλύπτει ένα ευρύ φάσμα ετών (Year), είναι έκδηλη παλαιότερη σεισμική δραστηριότητα στο σύνολο των δεδομένων.

Η διαδραστική 3D οπτικοποίηση (Σχήμα 5.13) επιτρέπει την αποτελεσματική απεικόνιση των πέντε διαστάσεων των σεισμών. Συγκεκριμένα, οι κανονικοποιημένες τιμές του γεωγραφικού μήκους, του γεωγραφικού πλάτους και του βάθους καθορίζουν τη θέση κάθε σεισμού στον τρισδιάστατο χώρο. Το χρώμα των σημείων χρησιμοποιείται για να αναπαραστήσει το κανονικοποιημένο μέγεθος του σεισμού. Η πιο σημαντική προσθήκη είναι η χρήση του αρχικού (μη κανονικοποιημένου) έτους (Year_Original) ως animation frame. Αυτή η λειτουργία προσθέτει μια δυναμική διάσταση στο γράφημα, επιτρέποντας την παρακολούθηση της εξέλιξης των σεισμών και των συστάδων τους στο πέρασμα του χρόνου, μέσω ενός slider. Επιπλέον, εμφανίζονται αναλυτικές πληροφορίες για κάθε σεισμό. Αυτή η προσέγγιση προσφέρει μια πλουσιότερη και πιο εύκολα ερμηνεύσιμη οπτικοποίηση, ειδικά όταν η χρονική διάσταση είναι κεντρική στην ανάλυση.



Σχήμα 5.13: Οπτικοποίηση 5 Διαστάσεων με Χρόνο ως animation K-Means 5D

5.2 Συσταδοποίηση βασιζόμενη στην Πυκνότητα (Density-based Clustering)

Η μεθοδολογία της συσταδοποίησης βασισμένης στην πυκνότητα αποτελεί ένα ιδιαίτερα ισχυρό και ευέλικτο εργαλείο για την ανάδειξη πολύπλοκων, μη γραμμικών δομών, καθώς και για την αποτελεσματική αναγνώριση θορύβου (noise) ή απομονωμένων παρατηρήσεων (outliers), οι οποίες δεν εντάσσονται σε καμία συστάδα, λόγω της αραιής τοπικής πυκνότητάς τους. Στον τομέα της σεισμολογικής ανάλυσης, η εφαρμογή τέτοιου είδους αλγορίθμων - όπως οι DBSCAN, HDBSCAN και OPTICS - καθίσταται εξαιρετικά χρήσιμη, καθώς ανταποκρίνεται με ακρίβεια στις ιδιαιτερότητες των σεισμολογικών δεδομένων, τα οποία συχνά χαρακτηρίζονται από ετερογένεια, αυθαίρετα γεωμετρικά σχήματα και ύπαρξη περιοχών με ποικίλες πυκνότητες.

Σε αντίθεση με άλλες μεθόδους συσταδοποίησης, που βασίζονται σε προκαθορισμένα σχήματα (όπως οι σφαιρικές συστάδες του K-Means), οι αλγόριθμοι που βασίζονται στην πυκνότητα υπερτερούν στην ανακάλυψη συστάδων οποιουδήποτε σχήματος, χωρίς να απαιτείται εκ των προτέρων γνώση του αριθμού των συστάδων. Αυτή η πολύ σημαντική και κρίσιμη λεπτομέρεια θεωρείται ως πρωταρχικής σημασίας για την παρούσα ανάλυση. Επιπλέον, επιτρέπουν την απομόνωση περιοχών χαμηλής πυκνότητας, οι οποίες συχνά ανταποκρίνονται σε τυχαία ή διάσπαρτη σεισμική δραστηριότητα, δίνοντας τη δυνατότητα εστιασμένης μελέτης των πιο σημαντικών γεωδυναμικών δομών.

Ιδιαίτερη έμφαση δίνεται στον αλγόριθμο DBSCAN, ο οποίος χρησιμοποιείται ως ο βασικός αλγόριθμος για την συνολική πειραματική διαδικασία. Η ανάλυση με τον DBSCAN πραγματοποιείται σε ένα βαθύτερο επίπεδο, καθώς πέρα από την απλή εφαρμογή του (βέλτιστη ρύθμιση παραμέτρων), γίνεται επιλογή και λεπτομερή εξέταση συγκεκριμένων συστάδων, που εντοπίζονται στη δισδιάστατη απεικόνιση ορίζοντας αυστηρότερη παραμετροποίηση. Μέσω αυτής της εστιασμένης προσέγγισης, επιχειρείται η ανάδειξη ιδιαίτερων τοπικών σεισμικών μοτίβων, με στόχο την κατανόηση των γεωγραφικών, γεωλογικών ή χρονικών παραμέτρων, που ενδέχεται να καθορίζουν τη μορφολογία των συστάδων.

Η διερεύνηση του αλγόριθμου DBSCAN επεκτείνεται, εξετάζοντας τα δεδομένα τόσο σε δύο (2D) όσο και σε πέντε (5D) διαστάσεις (όπως και στην περίπτωση του K-Means), προκειμένου να αξιολογηθεί η επίδραση της ενσωμάτωσης επιπλέον χαρακτηριστικών (βάθος, μέγεθος, χρόνος) στην ποιότητα, την ακρίβεια και την ερμηνευσιμότητα των παραγόμενων συστάδων. Η πολυδιάστατη ανάλυση ενισχύει την κατανόηση της γεωχωρικής και χρονικής κατανομής της σεισμικής δραστηριότητας, επιτρέποντας μια σφαιρική προσέγγιση στην αναγνώριση σεισμικών μοτίβων. Ωστόσο, θα εξεταστούν και οι HDBSCAN 2D και OPTICS 2D, ώστε να γίνει μια συνολική συγκριτική αξιολόγηση όλων των αλγορίθμων του πειράματος.

5.2.1 DBSCAN 2D

Κατά την αρχική φάση της ανάλυσης, υλοποιείται ο DBSCAN πάνω σε γεωχωρικά σεισμολογικά δεδομένα, χρησιμοποιώντας αποκλειστικά τις δύο γεωγραφικές μεταβλητές, το Γεωγραφικό Πλάτος (Latitude) και το Γεωγραφικό Μήκος (Longitude). Ένα βασικό ερώτημα που προκύπτει, είναι εάν απαιτείται κανονικοποίηση (standardization) των δύο αυτών μεταβλητών, πριν την εφαρμογή

του αλγορίθμου. Στον χώρο της Μηχανικής Μάθησης, η προτυποποίηση των δεδομένων θεωρείται συχνά απαραίτητη για την ισόρροπη συμμετοχή των χαρακτηριστικών στον υπολογισμό των αποστάσεων.

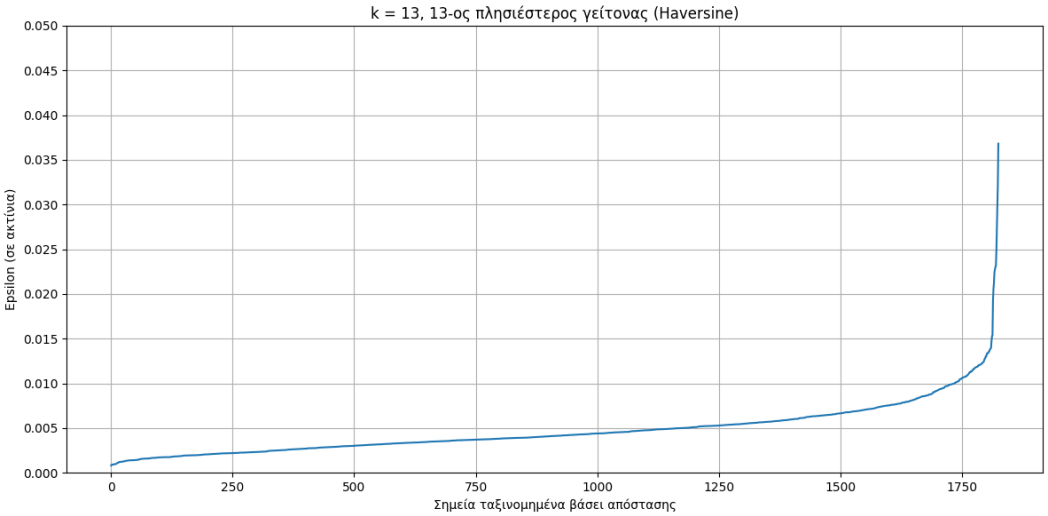
Ωστόσο, λαμβάνοντας υπόψη ότι και οι δύο μεταβλητές εκφράζονται στην ίδια μονάδα (μοίρες) και παρουσιάζουν συγκρίσιμο εύρος τιμών, ενώ παράλληλα η περιοχή μελέτης είναι ο ελλαδικός χώρος - δηλαδή μια γεωγραφικά περιορισμένη περιοχή - κρίνεται ότι η ευκλείδεια απόσταση χωρίς κανονικοποίηση μπορεί να αποδώσει ικανοποιητικά. Παρόλα αυτά, θα πρέπει να τονιστεί ότι η χρήση της ευκλείδειας απόστασης σε συντεταγμένες γεωγραφικού πλάτους και μήκους δεν λαμβάνει υπόψη την καμπυλότητα της Γης, με αποτέλεσμα να μην αντιστοιχεί σε πραγματική γεωδαιτική απόσταση μεταξύ των σημείων.

Για το λόγο αυτό, και με στόχο την καλύτερη γεωγραφική ερμηνεία των αποτελεσμάτων συσταδοποίησης, επιλέγεται τελικώς η χρήση της Haversine απόστασης. Η Haversine αποτελεί μετρική, που υπολογίζει τη μικρότερη απόσταση μεταξύ δύο σημείων πάνω στην επιφάνεια της Γης, λαμβάνοντας υπόψη τη σφαιρική της γεωμετρία. Απαιτεί τη μετατροπή των συντεταγμένων από μοίρες σε ακτίνια και χρησιμοποιείται συχνά σε εφαρμογές χωρικής ανάλυσης. Επιπλέον, είναι συμβατή με τον αλγόριθμο DBSCAN, μέσω του ορισμού της μετρικής απόστασης (metric='haversine'), προσφέροντας μεγαλύτερη ακρίβεια στον εντοπισμό γεωχωρικών συστάδων.

Συνεπώς, η επιλογή της Haversine απόστασης αντί της ευκλείδειας, χωρίς κανονικοποίηση, αποτελεί μια στοχευμένη απόφαση με βάση τη φύση των δεδομένων, τη γεωγραφική περιοχή μελέτης και την επιθυμία για ρεαλιστικότερη ερμηνεία των αποστάσεων μεταξύ σεισμικών γεγονότων.

Η εφαρμογή του αλγορίθμου ομαδοποίησης DBSCAN στα δισδιάστατα δεδομένα (γεωγραφικό πλάτος και μήκος) πραγματοποιείται σε δύο διακριτές φάσεις, με στόχο την ολοκληρωμένη διερεύνηση της σεισμικής δραστηριότητας. Αρχικά, επιχειρείται η εύρεση της βέλτιστης διαμόρφωσης των παραμέτρων ε (ακτίνα) και *MinPts* (ελάχιστος αριθμός σημείων) για την ανάδειξη των πιο φυσικών συστάδων. Αυτή η διαδικασία περιλαμβάνει συστηματικούς πειραματισμούς και την αξιολόγηση δεικτών ποιότητας ομαδοποίησης. Στη συνέχεια, λαμβάνει χώρα μια δεύτερη εκτέλεση (run) του DBSCAN με την εφαρμογή μιας αυστηρότερης συσταδοποίησης.

Α Φάση
Ο κώδικας, ο οποίος παρατίθεται στο Παράρτημα Α (βλ. DBSCAN 2D) υλοποιεί μια ολοκληρωμένη προσέγγιση για την ανάλυση και οπτικοποίηση σεισμικών δεδομένων, εφαρμόζοντας τον συγκεκριμένο αλγόριθμο συσταδοποίησης στις γεωγραφικές συντεταγμένες των σεισμών. Η διαδικασία ξεκινά με τη φόρτωση και προεπεξεργασία των σεισμικών δεδομένων από το EarthquakesGr.csv, καθώς και των γεωγραφικών ορίων των τεκτονικών πλακών από το TectonicPlates.csv. Κατά την προεπεξεργασία, οι χρονικές πληροφορίες των σεισμών μετατρέπονται σε κατάλληλη μορφή, τα δεδομένα καθαρίζονται από τυχόν κενές τιμές, και οι γεωγραφικές συντεταγμένες μετατρέπονται σε ακτίνια για την ορθή εφαρμογή της μετρικής απόστασης Haversine, η οποία είναι κατάλληλη για γεωγραφικά δεδομένα στον αλγόριθμο DBSCAN. Επιπλέον, ο χρόνος καταγραφής ορίζεται ως ευρετήριο του DataFrame για διευκόλυνση στην ανάλυση χρονοσειρών.

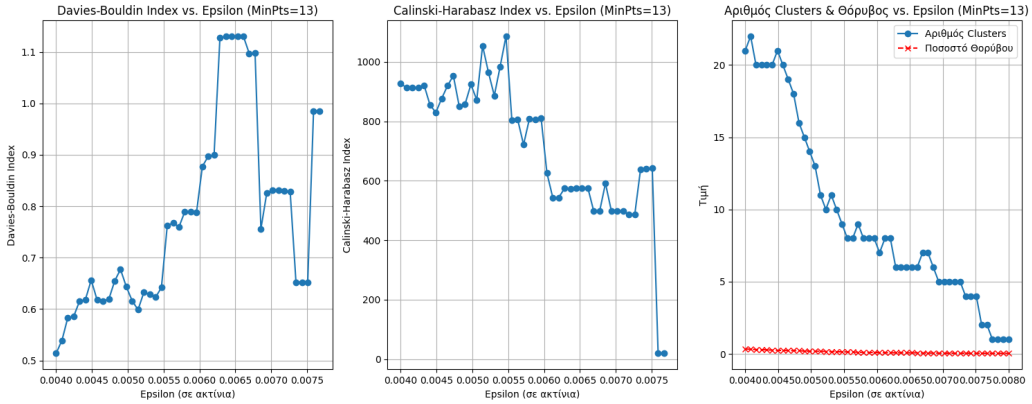


Σχήμα 5.14: Γράφημα 13-ος Πλησιέστερος Γείτονας (Haversine) DBSCAN 2D

Ένα κεντρικό μέρος της ανάλυσης αφορά την εξερεύνηση και επιλογή των βέλτιστων παραμέτρων του DBSCAN, δηλαδή του ε *(epsilon)* και του *MinPt*s. Για το σκοπό αυτό, ο κώδικας παράγει γραφήματα εύρεσης *k-Πλησιέστερων Γειτόνων*, τα οποία παρέχουν οπτική υποστήριξη για την εκτίμηση της βέλτιστης τιμής του ε. Ορίζεται ένα συγκεκριμένο εύρος τιμών k και παράγονται τα αντίστοιχα γραφήματα, τα οποία μελετώνται συναρτήσει των μετρικών, που έχουν επιλεγεί για τον DBSCAN. Στη γραφική παράσταση (Σχήμα 5.14) απεικονίζεται η απόσταση του 13ου πλησιέστερου γείτονα για κάθε σημείο δεδομένων (ταξινομημένα σε αύξουσα σειρά), χρησιμοποιώντας τη μέτρηση Haversine με δεδομένα σε ακτίνια. Η καμπύλη αρχίζει να ανεβαίνει απότομα γύρω από την τιμή 0.005. Ένα ευδιάκριτο σημείο καμπής παρατηρείται στην περιοχή ε=0.005 έως 0.006 ακτίνια. Αυτή η περιοχή υποδηλώνει την απόσταση, στην οποία οι πυκνές συστάδες αρχίζουν να αραιώνουν σημαντικά, καθιστώντας την μια ενδεδειγμένη τιμή για την ακτίνα ε στον DBSCAN.

Συμπληρωματικά, πραγματοποιείται μια συστηματική αξιολόγηση των παραμέτρων, μέσω δεικτών αξιολόγησης συσταδοποίησης, όπως ο Davies-Bouldin Index (DBI) και ο Calinski-Harabasz Index (CHI), οι οποίοι ποσοτικοποιούν την ποιότητα των clusters, ενώ παράλληλα παρακολουθείται ο αριθμός των clusters και το ποσοστό των σημείων, που ταξινομούνται ως θόρυβος. Στην πρώτη απεικόνιση (Σχήμα 5.15) με βάση τον δείκτη Davies-Bouldin, οι καλύτερες συσταδοποιήσεις (δηλαδή, χαμηλότερες τιμές) παρατηρούνται για τιμές ε μεταξύ 0.0040 και 0.0055. Με βάση τον δείκτη Calinski-Harabasz, η βέλτιστη τιμή ε βρίσκεται περίπου στο 0.0050-0.0052, καθώς εκεί έχουμε την υψηλότερη τιμή, υποδεικνύοντας πιο συμπαγείς και καλά διαχωρισμένες συστάδες.

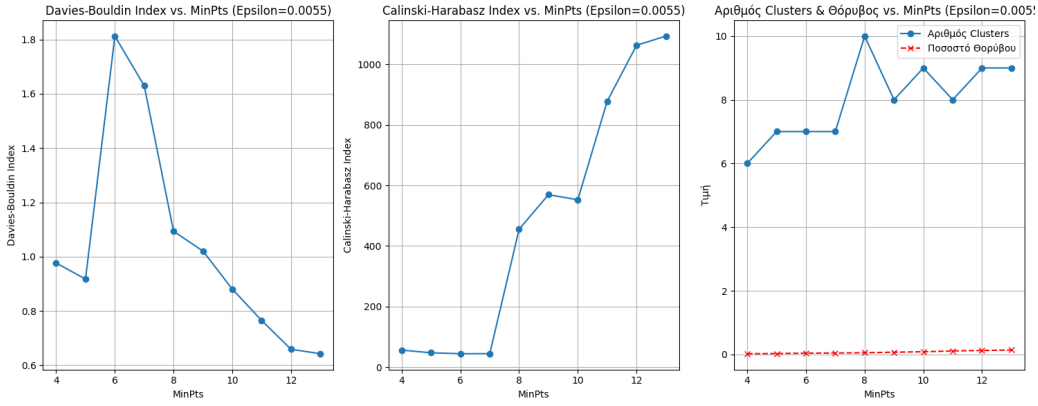
Αναφορικά με τον αριθμό των clusters και τον θόρυβο, παρατηρείται στο γράφημα, ότι ξεκινάει με έναν μεγάλο αριθμό συστάδων (περίπου 22) για μικρές τιμές ε. Αυτό είναι αναμενόμενο, καθώς ένα μικρό ε σημαίνει, πως μόνο σημεία, που είναι πολύ κοντά το ένα στο άλλο ομαδοποιούνται, οδηγώντας σε πολλές μικρές συστάδες. Καθώς το ε αυξάνεται, ο αριθμός των συστάδων μειώνεται σταθερά. Αυτό συμβαίνει επειδή μεγαλύτερο ε επιτρέπει σε περισσότερα σημεία να ομαδοποιηθούν μαζί, συγχωνεύοντας μικρότερες συστάδες σε μεγαλύτερες. Υπάρχουν σημεία, που ο αριθμός των συστάδων παραμένει σταθερός για κάποιο διάστημα και μετά μειώνεται απότομα.



Σχήμα 5.15: Διάγραμμα eps vs metrics για εξερεύνηση eps με σταθερό MinPts DBSCAN 2D

Για πολύ μεγάλες τιμές ε (περίπου 0.0078 και άνω), ο αριθμός των συστάδων πέφτει σε πολύ χαμηλές τιμές (πιθανόν 1 ή 0), υποδηλώνοντας ότι όλα τα σημεία έχουν συγχωνευθεί σε μία ή καμία συστάδα (αν όλα τα σημεία γίνουν θόρυβος). Το ποσοστό θορύβου παραμένει εξαιρετικά χαμηλό (σχεδόν μηδέν) σε όλο το εύρος του ε που εξετάζεται. Αυτό είναι ένα σημαντικό εύρημα, καθώς δείχνει, ότι ο αλγόριθμος καταφέρνει να ταξινομήσει σχεδόν όλα τα σημεία σε συστάδες και ότι δεν υπάρχει σημαντική ποσότητα θορύβου στα δεδομένα με βάση τις επιλεγμένες παραμέτρους.

Με τον ίδιο τρόπο μελετάται και η δεύτερη απεικόνιση (Σχήμα 5.16). Για την επίδραση του *MinPts* (με σταθερό ε=0.0055 ακτίνια) διαπιστώνεται, ότι οι τιμές *MinPts* μεταξύ 11 και 13 αποδίδουν τις καλύτερες συσταδοποιήσεις, με τους δείκτες Davies-Bouldin να είναι στο χαμηλότερο τους και Calinski-Harabasz στο υψηλότερο τους σημείο. Συμπερασματικά, παρατηρείται ότι ένα ε στην περιοχή 0.0050−0.0055 ακτίνια (περίπου 32 - 35 km) οδηγεί σε βέλτιστες τιμές και για τους δύο δείκτες ποιότητας. Ταυτόχρονα, το ποσοστό των σημείων, που ταξινομούνται ως θόρυβος διατηρείται σε αμελητέα επίπεδα. Η επιλογή αυτών των παραμέτρων επιτρέπει τον εντοπισμό γεωγραφικών συστάδων σεισμών εντός λογικών αποστάσεων, κάτι το οποίο είναι εξέχουσας σημασίας για τη σεισμολογική ανάλυση.

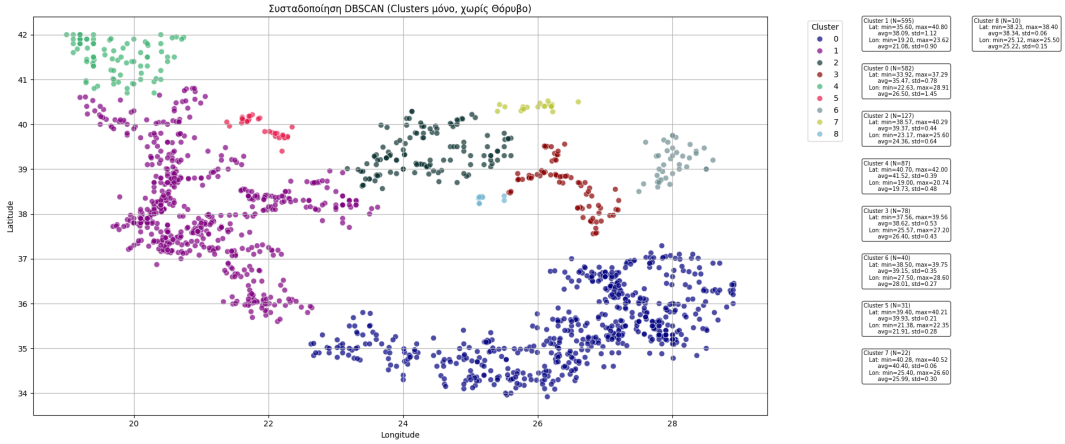


Σχήμα 5.16: Διάγραμμα minpts vs metrics για εξερεύνηση MinPts με σταθερό eps DBSCAN 2D

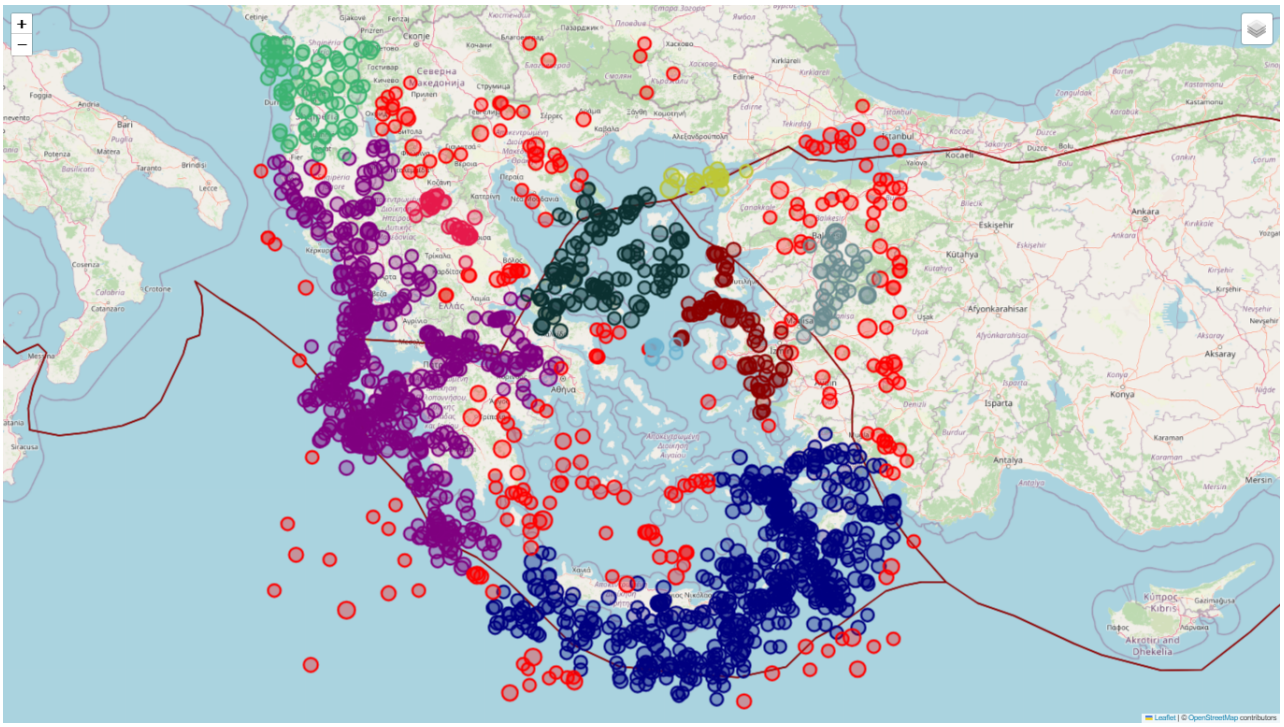
Αφού επιλεγούν οι βέλτιστες παράμετροι (ε=0.0055 ακτίνια / 35 km, *MinPts*=13), ο DBSCAN εφαρμόζεται στα δεδομένα, και οι ετικέτες των clusters (ή -1 για τα σημεία θορύβου - ποσοστό σημείων θορύβου: 13.91%), αποθηκεύονται ως νέα στήλη στο αρχικό DataFrame. Επιπρόσθετα, υπολογίζονται και καταγράφονται αναλυτικά περιγραφικά στατιστικά (μέσοι όροι, ελάχιστα, μέγιστα, τυπικές αποκλίσεις) για κάθε cluster, καλύπτοντας γεωγραφικές συντεταγμένες, βάθος και μέγεθος.

Η οπτικοποίηση των αποτελεσμάτων αποτελεί βασικό κομμάτι του κώδικα. Δημιουργείται Διάγραμμα Διασποράς (Scatter Plot), που απεικονίζει τις χωρικές κατανομές των σεισμών για τα clusters χωρίς τον θόρυβο και παράγεται, επίσης και στην περίπτωση του DBSCAN Περιγραφική Στατιστική Ανάλυση όλων των χαρακτηριστικών, που συνεπικουρούν στην ερμηνεία των αποτελεσμάτων. Η γραφική αναπαράσταση στο (Σχήμα 5.17) παρουσιάζει τα αποτελέσματα της συσταδοποίησης των σεισμολογικών δεδομένων με τον αλγόριθμο DBSCAN (2D), χρησιμοποιώντας τις βέλτιστες παραμέτρους (ε=0.0055 ακτίνια / 35 km, *MinPts*=13), που προσδιορίστηκαν προηγουμένως. Τα σημεία δεδομένων έχουν ομαδοποιηθεί σε 9 διακριτές συστάδες (Clusters 0 έως 8), καθεμία εκ των οποίων αναπαρίσταται με διαφορετικό χρώμα. Σημειώνεται ότι έχουν απεικονιστεί μόνο τα σημεία, που ανήκουν σε κάποια συστάδα, και έχουν παραληφθεί τα σημεία που ταξινομήθηκαν ως θόρυβος. Παρατηρούνται εκτενείς συστάδες (π.χ., Cluster 0, Cluster 1), που υποδηλώνουν μεγάλες και ενεργές τεκτονικές περιοχές, καθώς και μικρότερες, πιο συμπαγείς συστάδες (π.χ., Cluster 5), που ενδέχεται να αντιπροσωπεύουν τοπικές συγκεντρώσεις σεισμικής δραστηριότητας. Συνοδευτικά, ο πίνακας στατιστικών στοιχείων δεξιά παρέχει λεπτομερείς πληροφορίες για κάθε εντοπισμένη συστάδα.

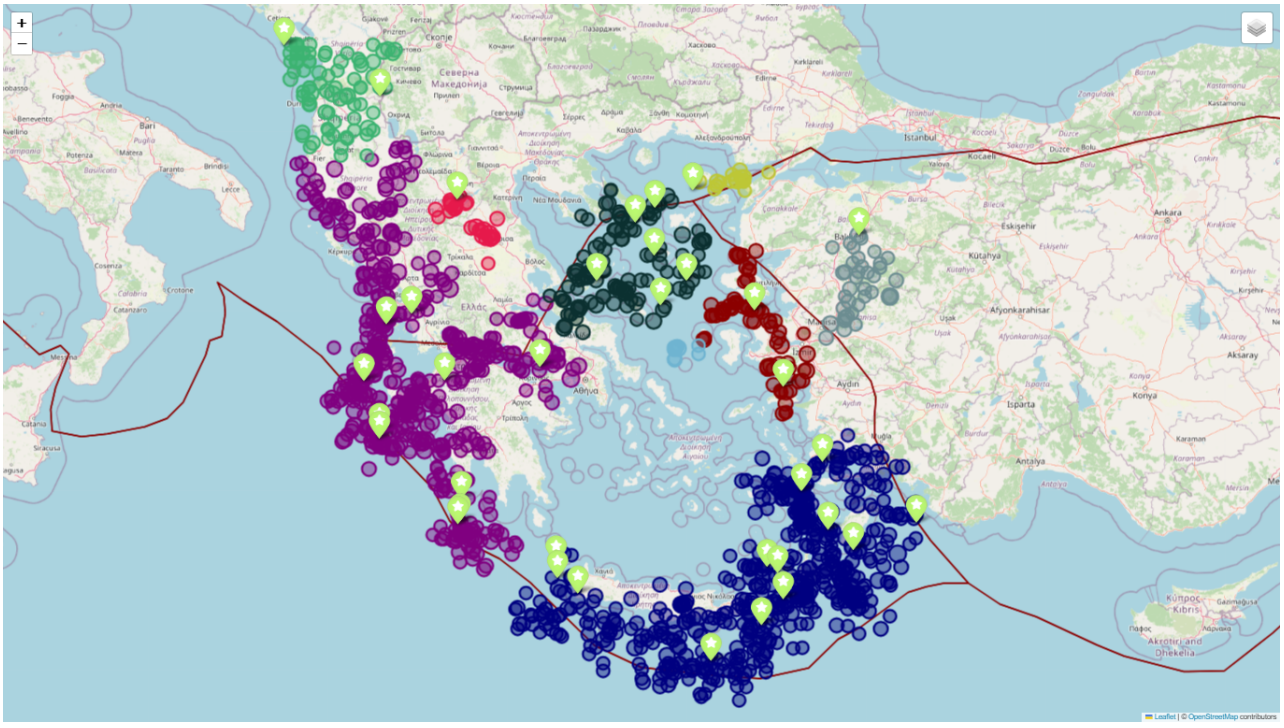
Τέλος, για μια πιο διαδραστική και γεωγραφικά πλούσια παρουσίαση, δημιουργούνται χάρτες Folium. Αυτοί οι χάρτες επιτρέπουν την οπτικοποίηση όλων των clusters (Σχήμα 5.18), των clusters χωρίς θόρυβο (Σχήμα 5.19), των δύο ισχυρότερων clusters βάσει πλήθους σεισμών, καθώς και επιλεγμένων, συγκεκριμένων clusters, με την προσθήκη των τεκτονικών πλακών για γεωλογικό πλαίσιο. Όλα τα παραγόμενα γραφήματα (PNG), οι διαδραστικοί χάρτες (HTML) και τα αρχεία δεδομένων (CSV) αποθηκεύονται συστηματικά σε έναν κεντρικό φάκελο εξόδου για εύκολη πρόσβαση και αναφορά.



Σχήμα 5.17: Διάγραμμα Διασποράς (Scatter Plot) με Στατιστικά DBSCAN 2D



Σχήμα 5.18: Διαδραστικός Χάρτης DBSCAN 2D (με θόρυβο σε κόκκινο χρώμα)



Σχήμα 5.19: Διαδραστικός Χάρτης DBSCAN 2D (χωρίς θόρυβο)

Ερμηνεία Αποτελεσμάτων DBSCAN 2D

- **Cluster 0:** περιλαμβάνει 582 σεισμούς, συγκεντρωμένους γύρω από 35.47° N (Πλάτος) και 26.50° E (Μήκος). Παρουσιάζει μέτρια διασπορά στο πλάτος (0.78) και μεγαλύτερη διασπορά στο μήκος (1.45), υποδηλώνοντας έναν επιμήκη σχηματισμό στον άξονα Ανατολής-Δύσης. Το μέσο βάθος των 29.94 km, με μεγάλη τυπική απόκλιση 31.70 km και εύρος από 1 έως 165 km δηλώνει μια περιοχή με σεισμούς, που εμφανίζονται σε πολύ διαφορετικά βάθη, από επιφανειακά έως ενδιάμεσου βάθους. Το μέσο μέγεθος 4.81 M_L, με τυπική απόκλιση 0.35 M_L και εύρος από 4.5 έως 6.3 M_L μαρτυρά, ότι πρόκειται για μία συστάδα με σχετικά ομοιογενή κατανομή, περιλαμβάνοντας σεισμούς μέσου έως ισχυρού μεγέθους. Ο μεγαλύτερος σεισμός της συγκεκριμένης συστάδας αφορά την περιοχή της Καρπάθου το 2021. Συνολικά, η συστάδα 0 αποτελεί μέρος του Ελληνικού Τόξου (Κρήτη, Κάρπαθος, Ρόδος έως και τη νοτιοδυτική Τουρκία), που είναι το σημείο σύγκλισης της Αφρικανικής με την Ευρασιατική λιθοσφαιρική πλάκα. Επίσης, στη συγκεκριμένη συστάδα τα ενδιάμεσα βάθη των σεισμών εκδηλώνονται στη ζώνη Benioff και φτάνουν περίπου τα 160 km (Νότιο Αιγαίο).

- **Cluster 1:** αντιπροσωπεύει 595 σεισμούς με κέντρο γύρω από 38.09° N (Πλάτος), 21.08° E (Μήκος). Έχει σημαντική διασπορά στο πλάτος (1.12) και μικρότερη στο μήκος (0.89), δείχνοντας μια επιμήκη κατανομή στον άξονα Βορρά-Νότου. Το μέσο βάθος των 14.56 km, έχει μεγάλη τυπική απόκλιση 14.72 km και εύρος από 1 έως 150 km. Όπως και το Cluster 0, περιλαμβάνει σεισμούς σε ένα ευρύ φάσμα βάθους. Το μέσο μέγεθος 4.85 M_L, με τυπική απόκλιση 0.36 M_L και εύρος από 4.5 έως 6.6 M_L - σεισμός Ζάκυνθος 2018 - επισημαίνει ομοιογενή κατανομή μεγέθους. Αφορά, όπως και στην περίπτωση του Cluster 0 το Ελληνικό Τόξο και πιο συγκεκριμένα την αφετηρία του. Σε αυτή την περιοχή, που αφορά το δυτικότερο άκρο του Ελληνικού Τόξου, εντοπίζεται και το λεγόμενο "τρίγωνο του διαβόλου", σημείο με έντονη σεισμική δραστηριότητα, τη μεγαλύτερη σε όλη την Ευρώπη. Η Αφρικανική πλάκα, κινείται προς τα βόρεια και επηρεάζει τη νότια Ελλάδα και η σύγκλιση αυτή προκαλεί τους επιφανειακούς σεισμούς αλλά και τους σεισμούς ενδιάμεσων βαθών, όπως και στο Cluster 0. Στην περιοχή αυτή, όμως, υπάρχει και η πλάκα της Απουλίας, που κινείται με φορά από αριστερά προς τα δεξιά και επηρεάζει τη δυτική και κεντρική Ελλάδα, καθώς συγκρούεται με την Πίνδο.

- **Cluster 2:** περιλαμβάνει 127 σεισμούς, με κέντρο γύρω από 39.36° N (Πλάτος) και 24.35° E (Μήκος). Παρουσιάζει σχετικά χαμηλή διασπορά τόσο στο πλάτος (0.43), όσο και στο μήκος (0.64), υποδηλώνοντας μία πιο συμπαγή και πυκνή συστάδα. Το μέσο βάθος είναι στα 16.51 km, με μέτρια τυπική απόκλιση 8.52 km και εύρος από 1 έως 38 km. Αυτή η συστάδα περιλαμβάνει επιφανειακούς σεισμούς. Το μέσο μέγεθος είναι 4.89 M_L, με τυπική απόκλιση 0.44 M_L και εύρος από 4.5 έως 6.7 M_L. Έχει ελαφρώς υψηλότερο μέσο μέγεθος και το μεγαλύτερο μέγεθος, που εντοπίστηκε σε αυτή την συστάδα είναι 6.7 M_L - ο σεισμός του Αγίου Ευστρατίου το 1968. Η τάφρος του βορείου Αιγαίου διέρχεται νότια της Σαμοθράκης και της ανατολικής Χαλκιδικής και βόρεια της Λήμνου και του Αγίου Ευστρατίου. Η συγκεκριμένη συστάδα αφορά σεισμούς, σχετικούς με το ρήγμα της Βόρειας Ανατολίας. Η Τουρκία κινείται προς τα δυτικά, δηλαδή προς το Αιγαίο. Η πλάκα του Αιγαίου κινείται και αυτή δυτικά, και συγχρόνως επεκτείνεται προς νότο.

- **Cluster 3:** συγκεντρώνει 78 σεισμούς. Το κέντρο βρίσκεται γύρω από 38.61° N (Πλάτος), 26.40° E (Μήκος). Υπάρχει χαμηλή διασπορά τόσο στο πλάτος (0.52), όσο και στο μήκος

(0.43), προσδιορίζοντας μία συμπαγή συστάδα. Το μέσο βάθος είναι στα 19.43 km, με τυπική απόκλιση 10.28 km και εύρος από 3 έως 46 km. Περιλαμβάνει επιφανειακούς σεισμούς. Το μέσο μέγεθος είναι στα 4.89 M_L, με τυπική απόκλιση 0.40 M_L και εύρος από 4.5 έως 6.7 M_L - ο σεισμός της Σάμου το 2020. Είναι σαφές, ότι η συστάδα αυτή αφορά τις σύνθετες τεκτονικές δομές, που αναφέρθηκαν για το Cluster 2. Η πλάκα της Ανατολίας ωθεί την πλάκα του Αιγαίου προς τα νοτιοδυτικά σε σχέση με την Ευρασιατική πλάκα, ενώ ταυτόχρονα η Αφρικανική πλάκα βυθίζεται κάτω από την Αιγιακή πλάκα. Η διαδικασία αυτή της καταβύθισης συμβάλλει και στην τεκτονική επέκταση της ευρύτερης περιοχής του Αιγαίου.

- **Cluster 4:** περιλαμβάνει 87 σεισμούς, με κέντρο γύρω από 41.51° N (Πλάτος), 19.72° E (Μήκος). Υφίσταται χαμηλή διασπορά τόσο στο πλάτος (0.38), όσο και στο μήκος (0.47), υποδηλώνοντας μία πολύ συμπαγή συστάδα στα βορειοδυτικά (περιοχή της Αλβανίας). Το μέσο βάθος είναι στα 14.62 km, με τυπική απόκλιση 10.36 km και εύρος από 1 έως 41 km. Και σε αυτή την περίπτωση, οι σεισμοί είναι επιφανειακοί. Το μέσο μέγεθος είναι 4.77 M_L, με τυπική απόκλιση 0.40 M_L και εύρος από 4.5 έως 6.8 M_L. Στην περίπτωση αυτή, τα περισσότερα επίκεντρα των επιφανειακών σεισμών διατάσσονται κατά μήκος της τοξοειδούς ζώνης στο Ελληνικό Τόξο, το οποίο ξεκινά από την Δυτική Αλβανία. Επιπλέον, ισχύουν και οι συμπιεστικές δυνάμεις της πλάκας της Απουλίας στην Αδριατική.

- **Cluster 5:** εντοπίζει 31 σεισμούς. Το κέντρο βρίσκεται γύρω από 39.93° N (Πλάτος), 21.90° E (Μήκος). Υπάρχει πολύ χαμηλή διασπορά (0.20° N Πλάτος, 0.28° E Μήκος), με αποτέλεσμα μία εξαιρετικά συμπαγή και πυκνή συστάδα. Το μέσο βάθος είναι 10.61 km, με τυπική απόκλιση 7.42 km και εύρος από 5 έως 39 km. Πρόκειται για μία συστάδα με επιφανειακούς σεισμούς. Το μέσο μέγεθος 4.88 M_L, με τυπική απόκλιση 0.42 M_L και εύρος από 4.5 έως 6.1 M_L - σεισμός Κοζάνης και Γρεβενών το 1995. Στην περίπτωση του Cluster 5, ο DBSCAN κατάφερε να απομονώσει μια συστάδα σεισμών, οι οποίοι έχουν κάποιες ιδιαιτερότητες. Αρχικά, η περιοχή όπου εκδηλώθηκε ο σεισμός δεν θεωρούνταν υψηλού σεισμικού κινδύνου, καθώς δεν είχαν εντοπιστεί γνωστά ενεργά ρήγματα. Ωστόσο, έπειτα από εκτεταμένη έρευνα, που διήρκεσε πολλά χρόνια, ο σεισμός συνδέθηκε με ένα σύμπλεγμα ρηγμάτων, που βρίσκεται νοτιοδυτικά της Λίμνης Πολυφύτου.

- **Cluster 6:** αντιπροσωπεύει 40 σεισμούς. Το κέντρο βρίσκεται γύρω από 39.15° N (Πλάτος), 28.01° E (Μήκος). Υπάρχει χαμηλή διασπορά (0.34 Πλάτος, 0.27 Μήκος), προσδιορίζοντας μία συμπαγή συστάδα. Το μέσο βάθος είναι 13.92 km, με τυπική απόκλιση 9.20 km και εύρος από 1 έως 36 km. Οι σεισμοί και σε αυτή την συστάδα είναι επιφανειακοί. Το μέσο μέγεθος 4.82 M_L, με τυπική απόκλιση 0.48 M_L και εύρος από 4.5 έως 7.0M_L - Τουρκία 1964. Αυτή η συστάδα περιλαμβάνει τον ισχυρότερο σεισμό (7.0 M_L), που εντοπίστηκε σε όλες τις συστάδες.

- **Cluster 7:** περιέχει 22 σεισμούς. Το κέντρο είναι γύρω από 40.39° N (Πλάτος), 25.99° E (Μήκος). Υφίσταται πολύ χαμηλή διασπορά (0.06 Πλάτος, 0.29 Μήκος), υποδεικνύοντας μία εξαιρετικά συμπαγή συστάδα. Η διασπορά στο πλάτος είναι ιδιαίτερα μικρή. Το μέσο βάθος είναι στα 19.68 km, με τυπική απόκλιση 11.23 km και εύρος από 5 έως 41 km. Οι σεισμοί είναι επιφανειακοί. Το μέσο μέγεθος 4.91 M_L, με τυπική απόκλιση 0.45 M_L και εύρος από 4.5 έως 6.3 M_L - σεισμός Σαμοθράκης το 2014. Αποτελεί την συστάδα με το υψηλότερο μέσο μέγεθος. Η συστάδα αφορά τη σεισμική δραστηριότητα της τάφρου του βορείου Αιγαίου, η

οποία αποτελεί δυτική προέκταση του ρήγματος της βόρειας Ανατολίας, γνωστού για την παραγωγή ισχυρών σεισμών στο παρελθόν.

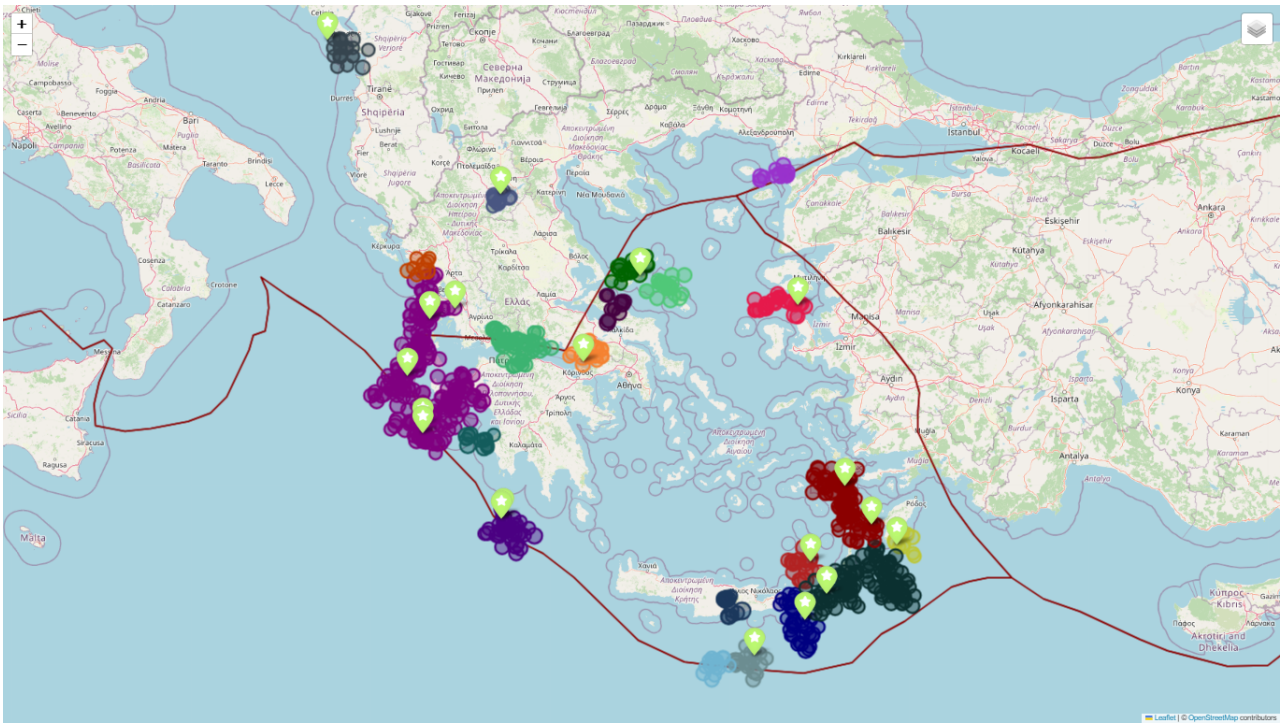
- **Cluster 8:** περιλαμβάνει 10 σεισμούς. Το κέντρο είναι γύρω από 38.34° N (Πλάτος), 25.22° E (Μήκος). Εξαιρετικά χαμηλή διασπορά (0.06 Πλάτος, 0.15 Μήκος), εντοπίζοντας μία πολύ μικρή, αλλά εξαιρετικά πυκνή και τοπική συστάδα. Το μέσο βάθος είναι στα 17.30 km, με τυπική απόκλιση 8.05 km και εύρος από 10 έως 32 km. Αφορά επιφανειακούς σεισμούς. Το μέσο μέγεθος είναι 4.83 M_L, με τυπική απόκλιση 0.24 M_L και εύρος από 4.6 έως 5.3 M_LL. Η συγκεκριμένη συστάδα έχει τη μικρότερη τυπική απόκλιση στο μέγεθος, κάτι που σημαίνει μεγάλη ομοιογένεια, σχετικά με το μέγεθος των σεισμών που περιέχει. Η συγκεκριμένη συστάδα αντανακλά τις γεωτεκτονικές διεργασίες στο Αιγαίο με συμπιεστικές και εφελκυστικές δυνάμεις, που ασκούνται από τις τεκτονικές πλάκες της Ανατολίας, της Αφρικής και της Αιγιακής.

Στο πλαίσιο της δισδιάστατης συσταδοποίησης των σεισμικών δεδομένων (πλάτος, μήκος), ο αλγόριθμος DBSCAN αποδείχθηκε σαφώς ανώτερος έναντι του K-Means, προσφέροντας πιο ρεαλιστικά και γεωφυσικά ερμηνεύσιμα αποτελέσματα. Ενώ ο K-Means βασίζεται σε προκαθορισμένο αριθμό συστάδων (k) και σχηματίζει σφαιρικές συστάδες, ο DBSCAN, ως αλγόριθμος βασισμένος στην πυκνότητα, διαθέτει την εγγενή ικανότητα να εντοπίζει συστάδες ακανόνιστου σχήματος και μεγέθους, καθώς και να αναγνωρίζει σημεία θορύβου. Αυτή η ιδιότητα είναι κρίσιμη στην σεισμολογία, καθώς η κατανομή των σεισμών σπάνια είναι σφαιρική και συχνά ακολουθεί τις πολύπλοκες γεωμετρίες των τεκτονικών δομών.

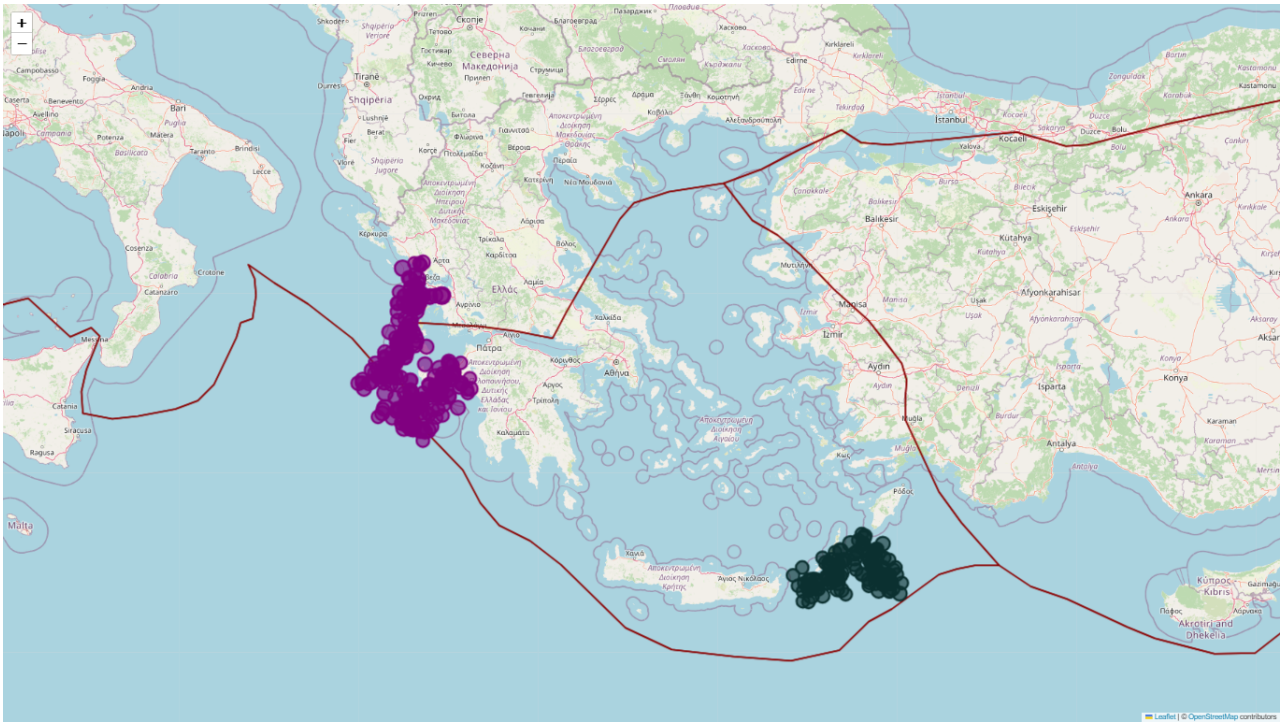
Πιο συγκεκριμένα, ο DBSCAN κατάφερε να εντοπίσει με επιτυχία συστάδες στις παρυφές των λιθοσφαιρικών πλακών, περιοχές που είναι γνωστές για την υψηλή σεισμική τους δραστηριότητα. Οι σχηματιζόμενες συστάδες αντικατοπτρίζουν πυκνές συγκεντρώσεις σεισμών κατά μήκος ενεργών ρηγμάτων και ζωνών σύγκλισης ή απόκλισης πλακών, παρέχοντας πιο ακριβή χωρική αποτύπωση των σεισμοτεκτονικών δομών. Η ικανότητα του DBSCAN να αγνοεί τα μεμονωμένα, απομονωμένα σεισμικά γεγονότα (θόρυβος) και να εστιάζει στις περιοχές υψηλής πυκνότητας, τον καθιστά ένα ισχυρό εργαλείο για την ανάδειξη των κύριων χαρακτηριστικών της σεισμικότητας.

Β Φάση

Μετά από εκτενείς πειραματισμούς και ακολουθώντας ακριβώς την ίδια διαδικασία με την Α Φάση (μελέτη όλων των σχετικών γραφικών αναπαραστάσεων και μετρικών), καθορίζονται οι παράμετροι *MinPts*=14 και ε=0.0033 ακτίνια (περίπου 21 km). Αυτή η αυστηρότερη ρύθμιση έχει ως αποτέλεσμα τον εντοπισμό 21 διακριτών συστάδων (Σχήμα 5.20). Ο κύριος στόχος αυτής της συσταδοποίησης είναι η απομόνωση συγκεκριμένων ομάδων σεισμών με υψηλότερη πυκνότητα, οι οποίες θεωρείται ότι σχετίζονται άμεσα με συγκεκριμένες τεκτονικές διεργασίες και γεωλογικές δομές. Παρά τον αυξημένο αριθμό συστάδων, η προσέγγιση αυτή επιτρέπει την ανάδειξη πιο ομοιογενών και γεωγραφικά συμπυκνωμένων σεισμικών ζωνών, προσφέροντας έτσι μια πιο λεπτομερή εικόνα των ενεργών ρηγμάτων. Ο θόρυβος στην παρούσα φάση δεν αποτελεί σημαντικό στοιχείο και παραβλέπεται, ακόμη και αν αποτελεί μεγάλο ποσοστό. Η ερμηνεία των συγκεκριμένων αποτελεσμάτων δίνεται με συνοπτικό τρόπο και με αρωγό την Περιγραφική Στατιστική, που παράγει ο κώδικας.



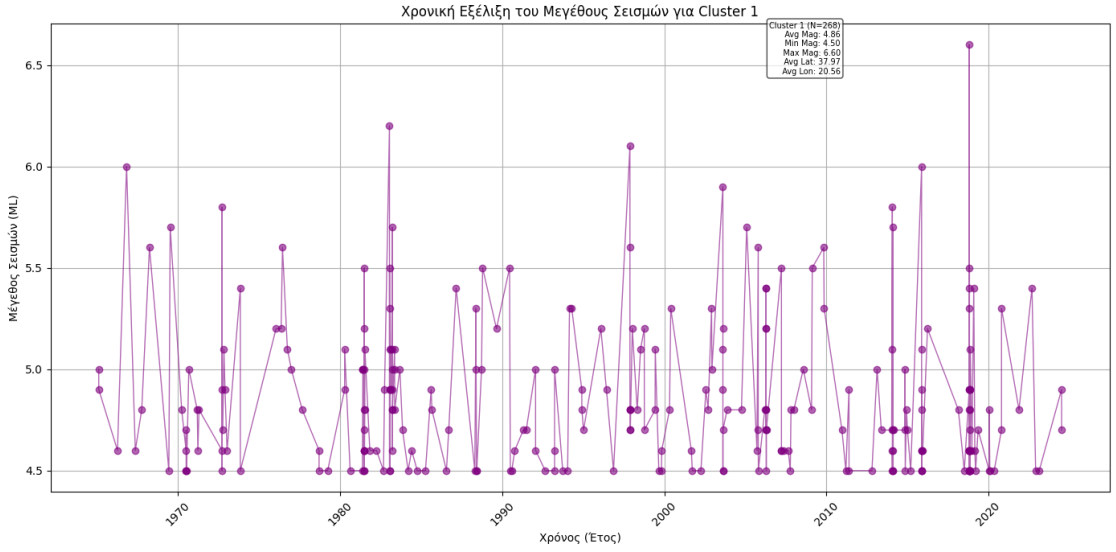
Σχήμα 5.20: Διαδραστικός Χάρτης DBSCAN 2D (χωρίς θόρυβο) - Αυστηροποίηση Παραμέτρων



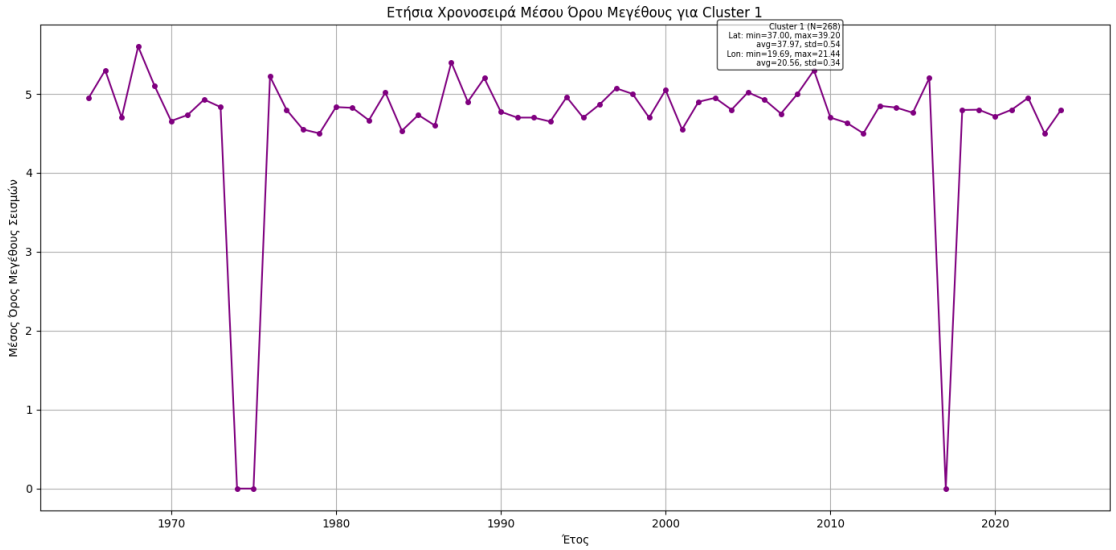
Σχήμα 5.21: Διαδραστικός Χάρτης DBSCAN 2D (χωρίς θόρυβο) - Επιλεγμένες Συστάδες

Η επιλογή έγινε με βάση το πλήθος των σεισμών ανά συστάδα. Στη συγκεκριμένη συσταδοποίηση τα Clusters 1 και 2 αποτελούν τις μεγαλύτερες συστάδες (Σχήμα 5.21). Ο συνολικός κώδικας DBSCAN 2D παράγει δύο τύπους χρονοσειρών για κάθε επιλεγμένη συστάδα: ετήσιες χρονοσειρές, που αναπαριστούν τον μέσο όρο του μεγέθους των σεισμών ανά έτος, και πιο λεπτομερείς χρονοσειρές, που απεικονίζουν το μέγεθος κάθε μεμονωμένου σεισμού σε σχέση με τον χρόνο καταγραφής του, προσφέροντας, έτσι, μια δυναμική εικόνα της σεισμικής δραστηριότητας.

- **Cluster 1:** περιλαμβάνει 268 σεισμούς. Το κέντρο βρίσκεται περίπου 37.97° N (Πλάτος) και 20.55° E (Μήκος), τοποθετώντας το στην περιοχή της Δυτικής Ελλάδας (νησιά Ιονίου) και της Πελοποννήσου (όπως φαίνεται στον χάρτη). Η σχετικά υψηλή τυπική απόκλιση στο πλάτος (0.53) και μέτρια στο μήκος (0.34) υποδηλώνει μια επιμήκη κατανομή των σεισμών στον Βορρά-Νότο άξονα, αντικατοπτρίζοντας τις κύριες τεκτονικές δομές της περιοχής. Η συστάδα χαρακτηρίζεται από μέσο βάθος 11.27 km (ελάχιστο 1 km, μέγιστο 60 km) με σημαντική τυπική απόκλιση 8.62 km, υποδεικνύοντας ότι περιλαμβάνει κυρίως επιφανειακούς σεισμούς. Το μέσο μέγεθος των σεισμών είναι 4.85 M_L (ελάχιστο 4.5 M_L, μέγιστο 6.6 M_L) με τυπική απόκλιση 0.37 M_L. Ο μεγαλύτερος σε μέγεθος σεισμός είναι αυτός της Ζακύνθου το 2018. Είναι το τρίγωνο του διαβόλου, καθώς ακριβώς πάνω από την Κεφαλονιά συναντιούνται η Απουλία, η Αφρικανική και η Αιγιακή πλάκα. Είναι μια περιοχή που δίνει μεγάλους σεισμούς και ιδιαίτερα καταστροφικούς. Στο (Σχήμα 5.22) παρουσιάζονται όλα τα επιμέρους μεγέθη των σεισμών του Cluster 1 ανά έτος και για την εξεταζόμενη περίοδο από 1964 έως 2024, προσφέροντας μια λεπτομερή εικόνα της κατανομής των μεγεθών και της συχνότητας. Η απεικόνιση των μεμονωμένων σημείων επιτρέπει την αναγνώριση όχι μόνο της μέσης τιμής αλλά και της διασποράς των μεγεθών σε κάθε έτος. Παρατηρούνται περίοδοι με αυξημένη συχνότητα μικρότερων μεγεθών, καθώς και μεμονωμένα γεγονότα με μεγαλύτερα μεγέθη (π.χ., κοντά στα 6.0 M_L). Διακρίνονται, επίσης, περίοδοι με έντονη σεισμική δραστηριότητα (πυκνότητα σημείων) και περίοδοι σχετικής ηρεμίας (εξεταζόμενο σύνολο δεδομένων από 4.5 M_L και πάνω).



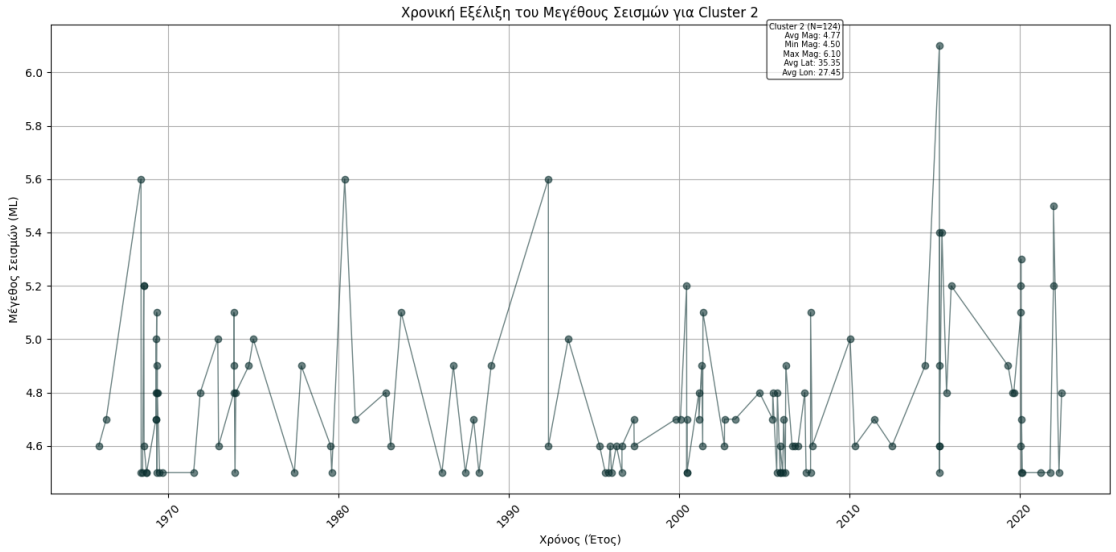
Σχήμα 5.22: Χρονική Εξέλιξη του Μεγέθους Σεισμών (Cluster 1)



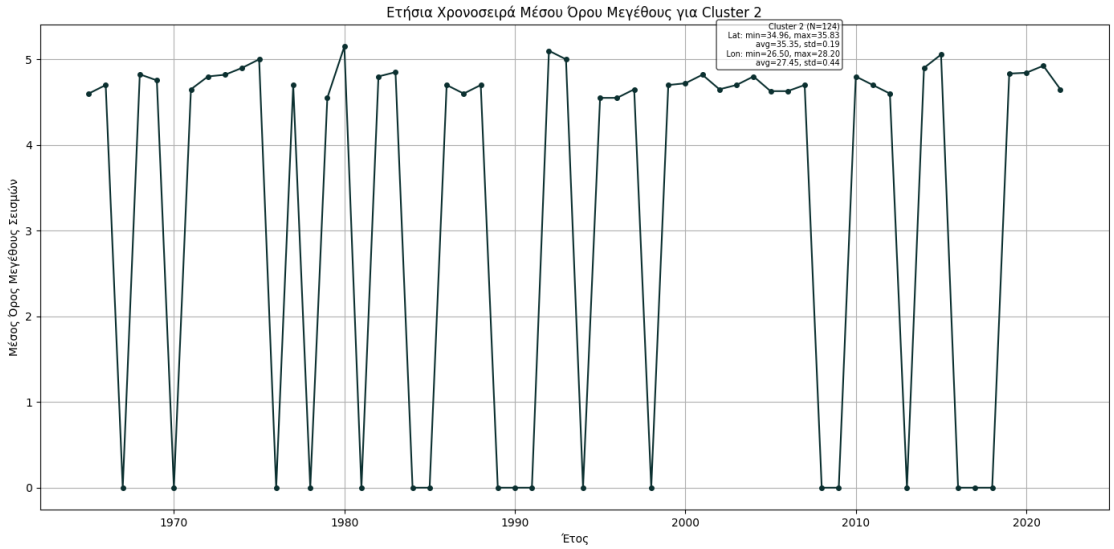
Σχήμα 5.23: Ετήσια Χρονοσειρά Μέσου Όρου Μεγέθους (Cluster 1)

Στο (Σχήμα 5.23) αποτυπώνεται ο ετήσιος μέσος όρος μεγέθους των σεισμών του Cluster 1. Ο μέσος όρος μεγέθους κυμαίνεται μεταξύ 4.5 και 5.5 M_L. Είναι αξιοσημείωτο, ότι υπάρχουν έτη, που η σεισμική δραστηριότητα έπεσε στο μηδέν (αναφορικά με τα μεγέθη που μελετούνται). Η γενική εικόνα είναι μιας συστάδας με συνεχή δραστηριότητα μέτριου έως σημαντικού μέσου μεγέθους, που αντανακλά την ενεργή σεισμικότητα της περιοχής.

- **Cluster 2:** αντιπροσωπεύει 124 σεισμούς. Το Cluster 2 είναι η δεύτερη μεγαλύτερη συστάδα, εντοπιζόμενη στην περιοχή του Νότιου Αιγαίου και της Κρήτης (όπως φαίνεται στον χάρτη). Το κέντρο του βρίσκεται περίπου 35.35° N (Πλάτος) και 27.44° E (Μήκος). Οι χαμηλές τυπικές αποκλίσεις στο πλάτος (0.19) και οι μέτριες στο μήκος (0.44) υποδηλώνουν μια πιο συμπαγή συγκέντρωση σεισμών, με μια τάση επιμήκυνσης στον άξονα Ανατολής-Δύσης.



Σχήμα 5.24: Χρονική Εξέλιξη του Μεγέθους Σεισμών (Cluster 2)



Σχήμα 5.25: Ετήσια Χρονοσειρά Μέσου Όρου Μεγέθους (Cluster 2)

Το μέσο βάθος είναι 24.14 km (ελάχιστο 1 km, μέγιστο 100 km) με μεγάλη τυπική απόκλιση 21.85 km. Αυτό υποδηλώνει ότι το Cluster 2 περιλαμβάνει σεισμούς, που καλύπτουν ένα ευρύ φάσμα βάθους, από πολύ επιφανειακούς έως αρκετά βαθείς. Το μέσο μέγεθος των σεισμών είναι 4.77 M_L (ελάχιστο 4.5 M_L, μέγιστο 6.1 M_L) με τυπική απόκλιση 0.29 M_L. Στο (Σχήμα 5.24) φαίνεται επίσης συνεχής δραστηριότητα, με κορυφώσεις που φτάνουν τα 6.0 M_L, επισημαίνοντας την παρουσία σημαντικών σεισμικών γεγονότων. Εδώ, είναι πιο εμφανή τα κενά στο γράφημα, που αντιστοιχούν σε περιόδους χωρίς σεισμική δραστηριότητα στο εξεταζόμενο εύρος μεγεθών. Ενώ υπάρχουν περίοδοι με πολλαπλά μικρότερα μεγέθη, η συστάδα αυτή χαρακτηρίζεται επίσης από την εμφάνιση κάποιων μεγάλων, μεμονωμένων σεισμών (π.χ., άνω των 6.0 M_L το 2015). Στο (Σχήμα 5.25) είναι εμφανές, ότι ο μέσος όρος μεγέθους διατηρείται σε σχετικά υψηλά επίπεδα, με περιστασιακές αυξήσεις. Τα σημεία που η γραμμή πέφτει στο μηδέν υποδηλώνουν έτη, για τα οποία, δεν υπάρχει καταγεγραμμένη σεισμική δραστηριότητα, εντός των ορίων που μελετούνται (άνω του 4.5 M_L). Η γενική τάση δείχνει μια συστάδα, που όταν είναι ενεργή, παράγει σεισμούς με αξιοσημείωτο μέσο μέγεθος.

5.2.2 DBSCAN 5D

Καθώς η ανάλυση προχωρά, ο αλγόριθμος DBSCAN εφαρμόζεται σε έναν πολυδιάστατο χώρο, που περιλαμβάνει πέντε χαρακτηριστικά: το Γεωγραφικό Πλάτος (Latitude), το Γεωγραφικό Μήκος (Longitude), το Βάθος (Depth km), το Μέγεθος του σεισμού (Magnitude ML) και τον χρόνο γένεσης (Origin Time GMT). Η διεύρυνση αυτή επιβάλλει την αναθεώρηση της προσέγγισης, ως προς τη μετρική απόσταση και την προεπεξεργασία των δεδομένων.

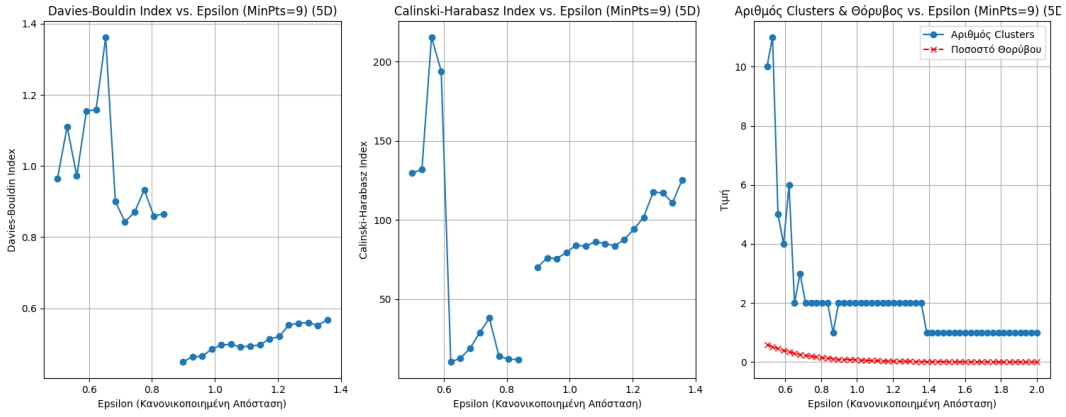
Σε αντίθεση με την περίπτωση των δύο γεωγραφικών διαστάσεων, που η χρήση της Haversine απόστασης προσφέρει γεωδαιτική ακρίβεια, χωρίς την ανάγκη κανονικοποίησης, η εφαρμογή της σε περισσότερες διαστάσεις καθίσταται μη εφικτή ή εννοιολογικά ασαφής, καθώς η Haversine ορίζε-

ται μόνο για σφαιρικές αποστάσεις μεταξύ δύο σημείων με γεωγραφικές συντεταγμένες. Ο αλγόριθμος DBSCAN δεν υποστηρίζει άμεσα συνδυασμό ετερογενών μετρικών (π.χ. Haversine για δύο διαστάσεις και ευκλείδεια για τις υπόλοιπες), γεγονός που καθιστά απαραίτητη την υιοθέτηση της ευκλείδειας απόστασης στον χώρο των πέντε διαστάσεων.

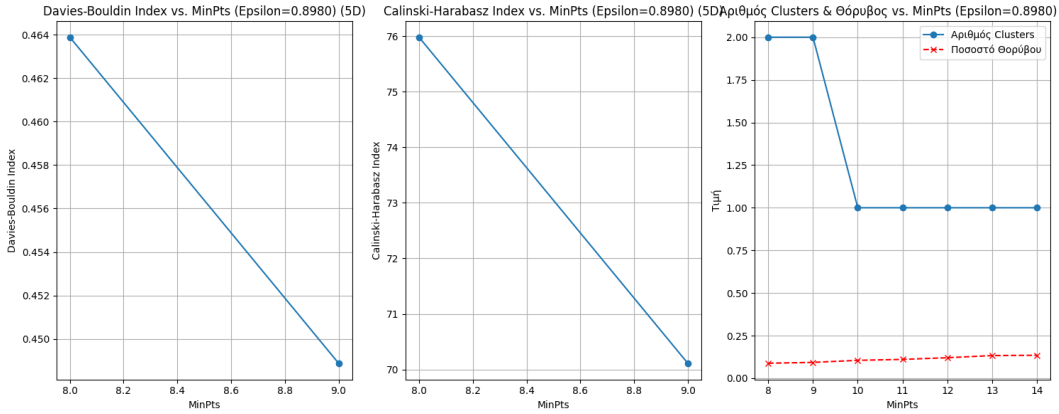
Ωστόσο, τα πέντε χαρακτηριστικά παρουσιάζουν διαφορετικές μονάδες μέτρησης και κλίμακες τιμών, για παράδειγμα το πλάτος και το μήκος εκφράζονται σε μοίρες, το βάθος σε χιλιόμετρα κ.τ.λ. Αν εφαρμοστεί ευκλείδεια απόσταση χωρίς προτυποποίηση των δεδομένων, οι μεταβλητές με μεγαλύτερη διακύμανση θα κυριαρχήσουν στον υπολογισμό των αποστάσεων, οδηγώντας τον DBSCAN σε στρεβλή συσταδοποίηση και υποβάθμιση της συμβολής σημαντικών αλλά αριθμητικά "ασθενέστερων" χαρακτηριστικών.

Για την αντιμετώπιση αυτού του προβλήματος εφαρμόζεται κανονικοποίηση μέσω της μεθόδου StandardScaler, η οποία μετασχηματίζει κάθε μεταβλητή, ώστε να έχει μηδενικό μέσο όρο και μοναδιαία τυπική απόκλιση. Με τον τρόπο αυτό εξασφαλίζεται ότι όλα τα χαρακτηριστικά συμμετέχουν ισότιμα στον υπολογισμό των αποστάσεων, ανεξαρτήτως της αρχικής τους κλίμακας. Η χρήση της κανονικοποίησης είναι απολύτως συμβατή με την ευκλείδεια απόσταση, καθιστώντας τον DBSCAN κατάλληλο για την ανακάλυψη συστάδων σε έναν ομογενοποιημένο πολυδιάστατο χώρο. Ο κώδικας παρατίθεται στο Παράρτημα Α (βλ. DBSCAN 5D).

Από την ενδελεχή μελέτη των οπτικοποιήσεων των μετρικών (Σχήμα 5.26) και (Σχήμα 5.27), διαπιστώνεται ότι η επιλογή των βέλτιστων παραμέτρων οδηγεί σε χαμηλό δείκτη Davies-Bouldin (0.449) και υψηλό δείκτη Calinski-Harabasz (70.113), με περιορισμένο ποσοστό θορύβου (9%), επιβεβαιώνοντας την εσωτερική συνοχή και τον καλό διαχωρισμό των συστάδων. Μετά από σχολαστική πειραματική διερεύνηση, επιλέγεται το ζεύγος των παραμέτρων: *MinPts*=9 και ε=0.898 (παρακάτω δίνεται το αποτέλεσμα στην κονσόλα). Ο θόρυβος (170 σημεία) αποτελεί μικρό ποσοστό (9.31%) και προκύπτουν 2 συστάδες.



Σχήμα 5.26: Διάγραμμα eps vs metrics για εξερεύνηση eps με σταθερό MinPts DBSCAN 5D

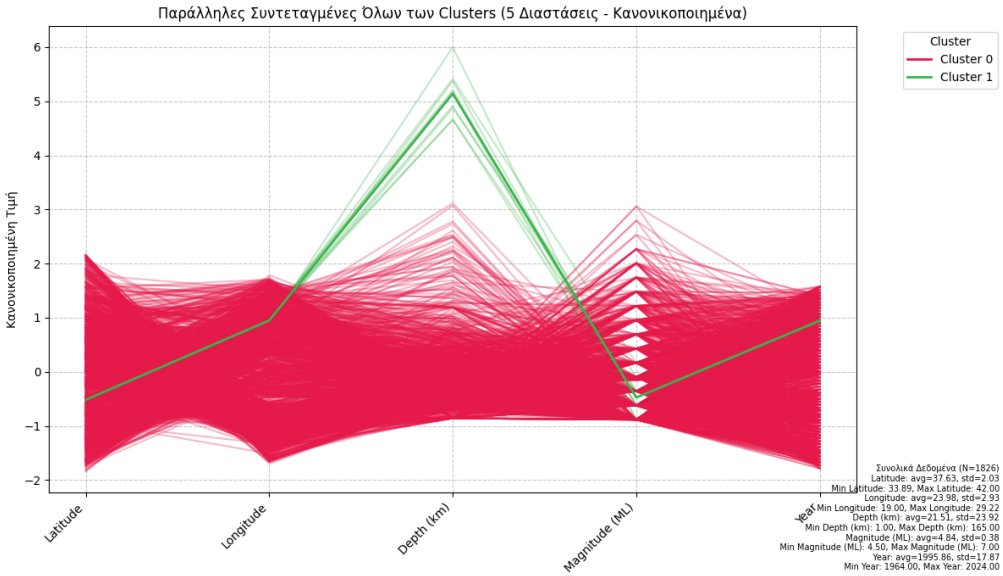


Σχήμα 5.27: Διάγραμμα minpts vs metrics για εξερεύνηση MinPts με σταθερό eps DBSCAN 5D

Αποτελέσματα Εκτέλεσης του DBSCAN (5D):

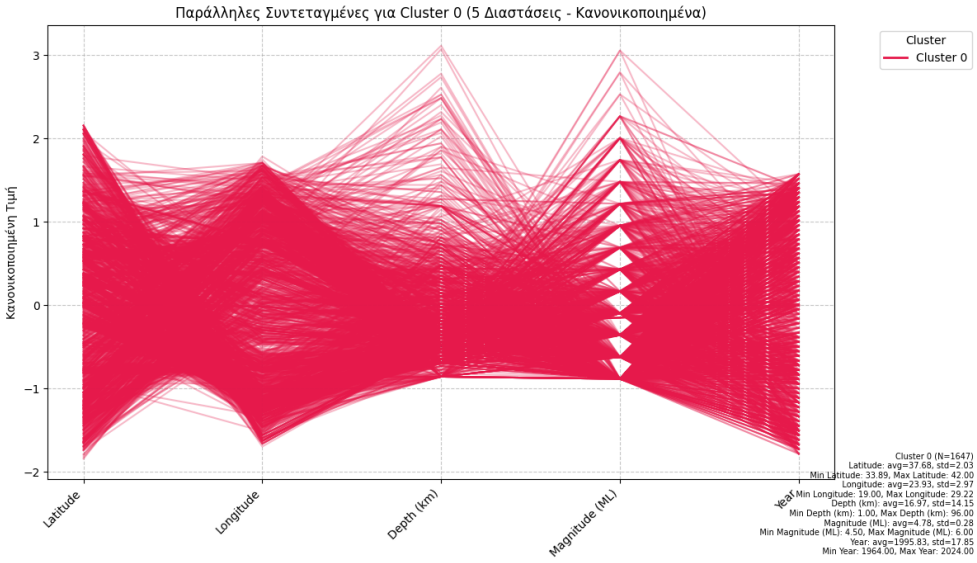
```
Επιλέχθηκαν βέλτιστες παράμετροι για DBSCAN (5D):
   optimal_eps_5d: 0.898 (σε κανονικοποιημένη απόσταση)
   optimal_min_samples_5d: 9

Αξιολόγηση DBSCAN (5D) με τελικές παραμέτρους:
   Αριθμός Clusters: 2
   Davies-Bouldin Index: 0.449 (Χαμηλότερο είναι καλύτερο)
   Calinski-Harabasz Index: 70.113 (Υψηλότερο είναι καλύτερο)
Αριθμός Σημείων Θορύβου (5D): 170
Ποσοστό Σημείων Θορύβου (5D): 9.31%
```

Το (Σχήμα 5.28) επιτρέπει τη σύγκριση προτύπων μεταξύ των συστάδων και την αναγνώριση διακριτών διαφορών σε συγκεκριμένες διαστάσεις, όπως για παράδειγμα το εστιακό βάθος και η χρονολογία. Όπως και στην περίπτωση του K-Means 5D, η ανάλυση γίνεται στοχευμένα σε κάθε συστάδα (μεμονωμένα γραφήματα) και με τη βοήθεια της Περιγραφικής Στατιστικής.



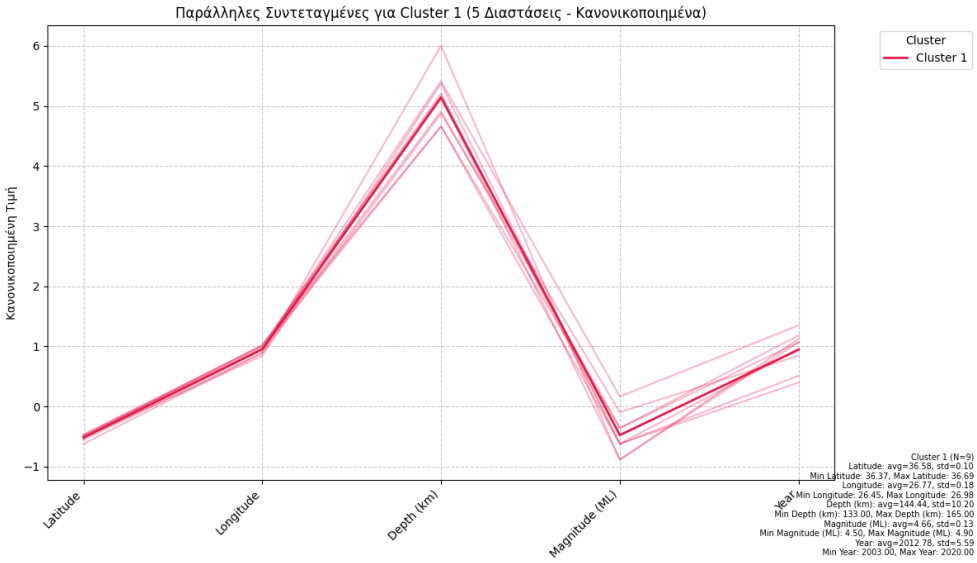
Σχήμα 5.28: Γράφημα Παράλληλων Συντεταγμένων (Parallel Coordinates) DBSCAN 5D

Ερμηνεία Αποτελεσμάτων K-Means 5D



Σχήμα 5.29: Γράφημα Παράλληλων Συντεταγμένων Cluster 0 DBSCAN 5D

- Το **Cluster 0** (Σχήμα 5.29) κυριαρχεί αριθμητικά με 1647 σεισμούς. Χαρακτηρίζεται από ευρεία κατανομή σε πλάτος (33.89°− 42.00°) και μήκος (19.00°− 29.22°), υποδηλώνοντας μια γεωγραφικά εκτεταμένη περιοχή. Το μέσο βάθος μόλις 17 km, δηλώνει κυρίως επιφανειακούς σεισμούς. Οι σεισμοί έχουν σχετικά υψηλό μέσο μέγεθος 4.78 M_L που φτάνει έως 6.00 M_L. Καλύπτει την συνολική, εξεταζόμενη χρονική περίοδο (1964−2024), με μεγάλη διασπορά 17.6, η οποία δείχνει συνεχή δραστηριότητα τις τελευταίες δεκαετίες. Γενικά, το Cluster 0 αντιπροσωπεύει το βασικό σεισμικό πληθυσμό της εξεταζόμενης περιοχής, επισημαίνοντας κυρίως τους επιφανειακούς σεισμούς, αλλά συγχρόνως και τους πιο ισχυρούς.



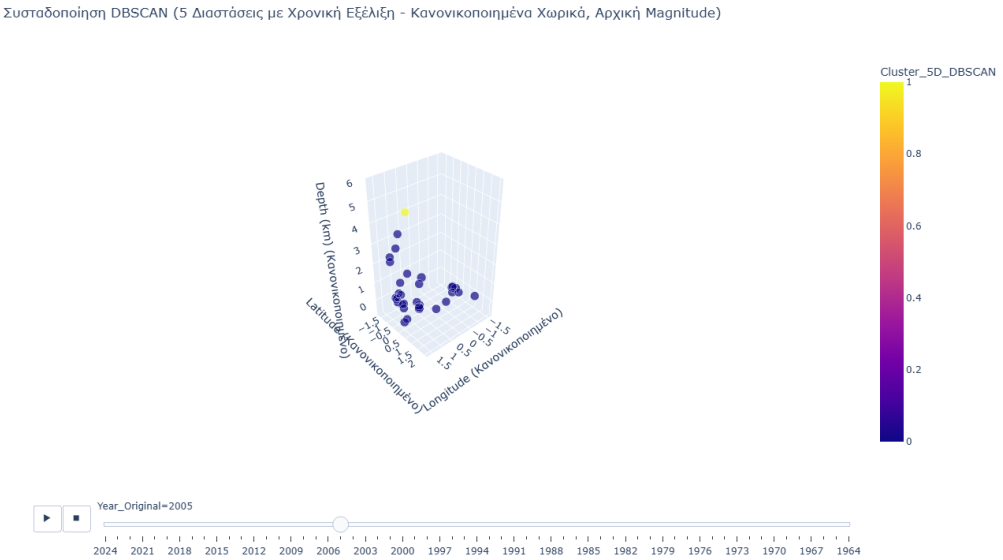
Σχήμα 5.30: Γράφημα Παράλληλων Συντεταγμένων Cluster 1 DBSCAN 5D

- Το **Cluster 1** (Σχήμα 5.30) περιέχει 9 γεγονότα. Εμφανίζει εξαιρετικά μικρή διασπορά σε πλάτος, μήκος και χρόνο. Αφορά πολύ περιορισμένη γεωγραφική περιοχή, καθώς οι γεωγραφικές συντεταγμένες δείχνουν πολύ στενό εύρος – όλα κοντά στην περιοχή (36.5° , 26.7°), που αντιστοιχεί στα Δωδεκάνησα (Ελληνικό Τόξο). Χαρακτηρίζεται από μεγάλα εστιακά βάθη (133–165 km), γεγονός, που την καθιστά σαφώς διακριτή από την κύρια ομάδα σεισμών. Το μέσο μέγεθος στα 4.66 M_L και μικρή τυπική απόκλιση, πράγμα που σημαίνει ότι εκδηλώνονται μεσαίου μεγέθους σεισμοί. Χρονικά, καλύπτει την περίοδο 2003−2020 και πιθανόν σχετίζεται με ιδιαίτερες γεωτεκτονικές διεργασίες.

Τα γραφήματα παράλληλων συντεταγμένων αναδεικνύουν οπτικά τις διαστάσεις, στις οποίες εντοπίζεται ο ουσιαστικός διαχωρισμός των συστάδων. Ειδικότερα:

- Η μεταβλητή "Depth (km)" φαίνεται να έχει τον μεγαλύτερο διαχωριστικό ρόλο, καθώς το Cluster 1 διαχωρίζεται καθαρά από το Cluster 0 μόνο σε αυτόν τον άξονα.
- Η μεταβλητή "Year" προσφέρει πρόσθετη διαχωριστική πληροφορία, καθώς οι σεισμοί του Cluster 1 εντοπίζονται σε πρόσφατα έτη.
- Οι υπόλοιπες μεταβλητές (Latitude, Longitude, Magnitude (ML)) διακρίνονται κυρίως ως συνεχείς και διάσπαρτες στο κύριο Cluster 0.

Τέλος, δημιουργήθηκε ένα δυναμικό, διαδραστικό Scatter Plot (Σχήμα 5.31), που επιτρέπει την οπτικοποίηση της σύνθετης σχέσης των σεισμικών γεγονότων σε πέντε διαστάσεις. Αυτή η προηγμένη απεικόνιση μετατρέπει τα αφηρημένα δεδομένα σε έναν κατανοητό χάρτη και έτσι, διευκολύνεται η αντίληψη όχι μόνο, πού συμβαίνουν οι σεισμοί και σε ποιες ομάδες ανήκουν, αλλά και πώς η κατανομή και η ομαδοποίησή τους μεταβάλλεται μέσα στον χρόνο, παρέχοντας μια ενδελεχή άποψη των αποτελεσμάτων της συσταδοποίησης.



Σχήμα 5.31: Οπτικοποίηση 5 Διαστάσεων με Χρόνο ως animation DBSCAN 5D

5.2.3 HDBSCAN 2D

Ο κώδικας για τον HDBSCAN πραγματοποιεί μια προηγμένη συσταδοποίηση βασισμένη στην πυκνότητα των δεδομένων των σεισμών, αξιοποιώντας την ικανότητά του να εντοπίζει συστάδες διαφορετικών πυκνοτήτων και σχημάτων. Για την εφαρμογή του HDBSCAN σε γεωγραφικά δεδομένα, δεν ήταν δυνατή η απευθείας χρήση της απόστασης Haversine, καθώς ο αλγόριθμος δεν την υποστηρίζει εγγενώς. Επομένως, επιλέχθηκε η κανονικοποίηση των συντεταγμένων (γεωγραφικού πλάτους και μήκους) και η χρήση της ευκλείδειας απόστασης, η οποία αποτελεί κοινή προσέγγιση σε παρόμοιες περιπτώσεις.

Η διαδικασία ξεκινά με την αρχικοποίηση του αλγορίθμου HDBSCAN. Σε αυτό το στάδιο, ορίζεται η βασική παράμετρος *min_cluster_size*. Αυτή η παράμετρος είναι κρίσιμη, καθώς καθορίζει τον ελάχιστο αριθμό σημείων, που απαιτούνται για να σχηματιστεί μια έγκυρη συστάδα. Επιλέγεται προσεκτικά, για να εξισορροπήσει την ανακάλυψη λεπτομερών συστάδων. Οποιαδήποτε ομάδα σημείων μικρότερη από αυτή την τιμή θα θεωρηθεί θόρυβος ή μέρος μιας μεγαλύτερης ασταθούς συστάδας. Εν συνεχεία, ορίζεται η παράμετρος *min_samples*. Αυτή η παράμετρος επηρεάζει το πόσο "σφιχτά" ή "πυκνά" πρέπει να είναι τα σημεία, για να θεωρηθούν μέρος μιας συστάδας. Είναι λιγότερο διαισθητική από το *min_cluster_size*, αλλά παίζει ρόλο στον τρόπο που ο HDBSCAN υπολογίζει την πυκνότητα πυρήνα (core distance). Αφού αρχικοποιηθεί ο αλγόριθμος, εφαρμόζεται η μέθοδος fit() στα προεπεξεργασμένα δισδιάστατα χαρακτηριστικά των σεισμών. Αυτή η διεργασία περιλαμβάνει την κατασκευή μιας ιεραρχίας συσταδοποίησης, με βάση την πυκνότητα και στη συνέχεια, την εξαγωγή των πιο σταθερών συστάδων από αυτή την ιεραρχία, χωρίς την ανάγκη ορισμού μιας καθολικής ακτίνας *ε (epsilon)*.

Δεδομένου ότι, η συσταδοποίηση σεισμών αναφέρεται στη Μη Εποπτευόμενη Μάθηση χρησιμοποιούνται οι εσωτερικές (internal) μετρικές, οι οποίες αξιολογούν την ποιότητα της συσταδοποίησης, με βάση τα ίδια τα δεδομένα και τα αποτελέσματα της συσταδοποίησης, χωρίς να απαιτούν γνώση των αληθινών ετικετών (ground truth). Για την ποσοτική αξιολόγηση της ποιότητας της συσταδοποίησης, υπολογίζονται τρεις εσωτερικές μετρικές: ο Silhouette Coefficient, ο Davies-Bouldin Index (DBI) και ο Calinski-Harabasz Index (CHI). Αυτοί οι δείκτες χρησιμοποιούνται, για να εκτιμήσουν τη συμπαγή δομή και τον διαχωρισμό των αναγνωρισμένων συστάδων. Είναι σημαντικό να σημειωθεί ότι, κατά τον υπολογισμό αυτών των μετρικών, τα σημεία που χαρακτηρίστηκαν ως θόρυβος (-1) αποκλείονται, καθώς δεν αποτελούν μέρος καμίας συστάδας και η συμπερίληψή τους θα παραμόρφωνε την αξιολόγηση της ποιότητας των πραγματικών συστάδων. Στη συνέχεια ο κώδικας, όπως και στους προηγούμενους αλγόριθμους παράγει την Περιγραφική στατιστική, χρονοσειρές, τις οπτικοποιήσεις και αρχεία για μελλοντική αναφορά. Ο συνολικός κώδικας του αλγορίθμου παρατίθεται στο Παράρτημα Α (βλ. HDBSCAN 2D).

Η εύρεση των βέλτιστων παραμέτρων για τον HDBSCAN, όπως και για τον DBSCAN, είναι μια σημαντική πρόκληση και συχνά απαιτεί μια συνδυασμένη προσέγγιση εμπειρίας, δοκιμών και αξιολόγησης. Ο HDBSCAN είναι πιο ευέλικτος από τον DBSCAN, επειδή δεν χρειάζεται το *ε (eps)*, αλλά εξακολουθεί να έχει παραμέτρους, που πρέπει να ρυθμιστούν. Ο καλύτερος τρόπος για να επιλεγεί το *min_cluster_size* και το *min_samples* είναι:

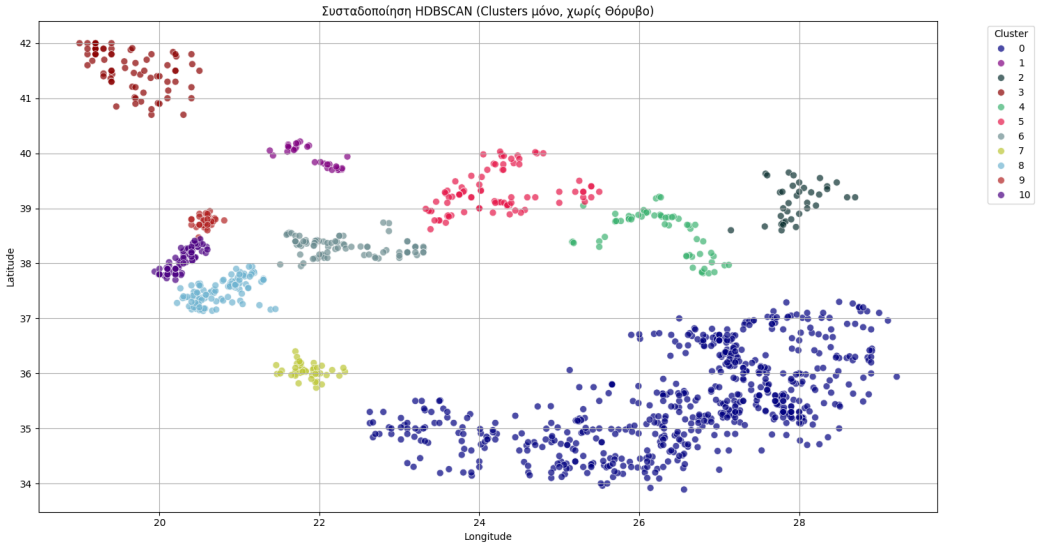
- **Δοκιμή και Σφάλμα (Trial and Error):** Εκτέλεση του αλγόριθμου με διαφορετικές τιμές και παρατήρηση των αποτελεσμάτων.

- **Οπτικοποίηση:** Αναζήτηση clusters, που έχουν οπτικό νόημα (γεωφυσικά).
- **Μετρικές Αξιολόγησης:** Χρήση εσωτερικών μετρικών όπως το Silhouette Score, Davies-Bouldin Index και Calinski-Harabasz Index, για να συγκριθούν οι διαφορετικές εκτελέσεις.
- **Γνώση Πεδίου (Domain Knowledge):** Η κατανόηση της σεισμολογίας μπορεί να προσφέρει καθοδήγηση στο ποιες συστάδες είναι λογικές ή σημαντικές.
- **Στόχος της ανάλυσης:** Η επιλογή παραμέτρων στο clustering είναι συχνά υποκειμενική και έχει άμεση εξάρτηση από αυτόν τον στόχο. Δεν υπάρχει πάντα μία σωστή απάντηση. Η κύρια επιδίωξη είναι να αποκαλυφθούν οι πιο σημαντικές και ερμηνεύσιμες δομές.

Πίνακας 5.1: Αποτελέσματα Εκτελέσεων του HDBSCAN με διαφορετικές παραμέτρους

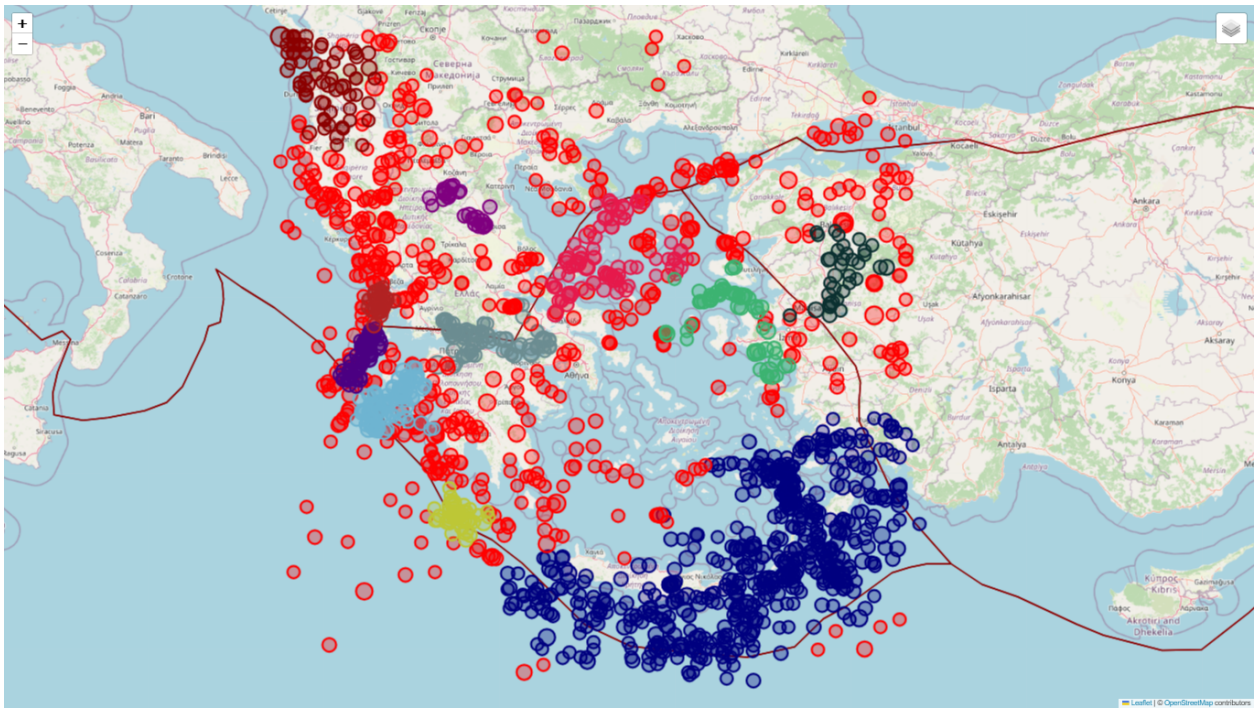
min_cluster_size	min_samples	Clusters	Noise %	DBI (↓)	CHI (↑)	Silhouette (↑)
15	14	12	25,79%	0,569	858,100	0,388
15	15	11	23,60%	0,574	937,578	0,374
16	14	12	25,79%	0,569	858,100	0,388
16	15	11	23,60%	0,574	937,578	0,374
17	15	11	23,60%	0,574	937,578	0,374
17	17	12	28,92%	0,557	862,688	0,395
18	14	11	24,92%	0,580	950,412	0,380
18	16	12	27,05%	0,568	827,267	0,381
18	17	11	27,82%	0,563	960,734	0,417
19	11	11	24,64%	0,612	961,363	0,367
20	10	11	23,99%	0,641	951,893	0,364
20	14	9	25,14%	0,604	1152,916	0,404
20	16	9	27,05%	0,586	1101,075	0,400
25	13	11	31,43%	0,543	930,254	0,474
29	14	9	25,14%	0,604	1152,916	0,404
34	10	9	24,21%	0,689	1145,873	0,357
39	4	10	27,27%	0,600	1006,257	0,456

Στον πίνακα (5.1) παρουσιάζονται ενδεικτικά αποτελέσματα από τις εκτελέσεις του αλγορίθμου HDBSCAN με διαφορετικούς συνδυασμούς παραμέτρων. Έχουν επιλεγεί χαρακτηριστικά παραδείγματα για λόγους συνοπτικής παρουσίασης, καθώς πραγματοποιήθηκαν περισσότερες δοκιμές κατά την πειραματική διαδικασία. Η τελική επιλογή αφορά τις παραμέτρους: *min_cluster_size: 25, min_samples: 13*, καθώς παρουσιάζει το καλύτερο Davies-Bouldin Index (0,543) και το καλύτερο Silhouette Score (0,474), γεγονός που υποδηλώνει εξαιρετική ποιότητα για τις 11 συστάδες που αναγνωρίζονται. Τα σημεία που εντάσσονται σε συστάδες είναι πολύ καλά καθορισμένα και διαχωρισμένα. Η οπτική επιβεβαίωση στο διάγραμμα διασποράς (Σχήμα 5.32) και στον χάρτη είναι το πιο κρίσιμο στοιχείο στην αξιολόγηση μιας συσταδοποίησης, ειδικά σε γεωχωρικά δεδομένα, όπως οι σεισμοί, που εξετάζονται στην παρούσα εργασία.

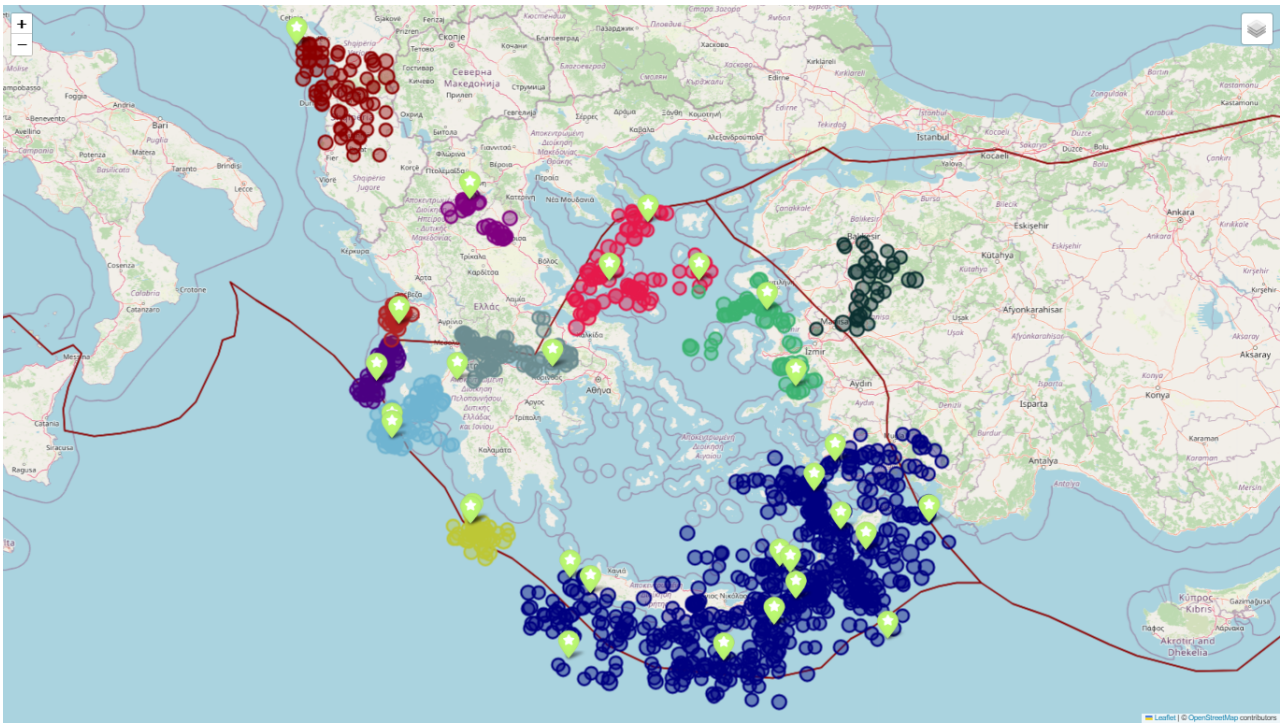


Σχήμα 5.32: Διάγραμμα Διασποράς (Scatter Plot) HDBSCAN 2D

Το ποσοστό θορύβου (31.43%) είναι ένα αναμενόμενο "τίμημα" για την επίτευξη ιδιαίτερα καθαρών και συμπαγών συστάδων (Σχήμα 5.33). Αυτό σημαίνει, ότι ο αλγόριθμος είναι πιο αυστηρός στην αναγνώριση των πυκνών περιοχών, και τα σημεία που δεν πληρούν αυτά τα αυστηρά κριτήρια χαρακτηρίζονται σωστά ως θόρυβος, αντί να "σπρώχνονται" αναγκαστικά σε μια συστάδα, στην οποία δεν ανήκουν πραγματικά. Στο πλαίσιο των σεισμών, αυτά τα σημεία θορύβου μπορεί να αντιπροσωπεύουν μεμονωμένα γεγονότα ή γεγονότα σε περιοχές με πολύ αραιή σεισμική δραστηριότητα.



Σχήμα 5.33: Διαδραστικός Χάρτης HDBSCAN 2D (με θόρυβο σε κόκκινο χρώμα)



Σχήμα 5.34: Διαδραστικός Χάρτης HDBSCAN 2D (χωρίς θόρυβο)

Ερμηνεία Αποτελεσμάτων HDBSCAN 2D
Με βάση, λοιπόν το (Σχήμα 5.34) και την Περιγραφική Στατιστική, που παράγει ο κώδικας, δίδεται η παρακάτω ανάλυση των συστάδων:

- **Cluster 0:** Αυτή είναι η μεγαλύτερη συστάδα σε αριθμό σημείων (619), αντιπροσωπεύοντας μια πολύ ευρεία γεωγραφική περιοχή, που εκτείνεται κυρίως σε χαμηλότερα γεωγραφικά πλάτη (μέση τιμή πλάτους 35.5) και ανατολικά γεωγραφικά μήκη (μέση τιμή μήκους 26.5). Παρουσιάζει τη μεγαλύτερη διασπορά στο βάθος (έως 165 km) αλλά και στο μέγεθος, υποδηλώνοντας πιθανώς μια γενική ή διάχυτη σεισμικότητα σε ένα ευρύ γεωτεκτονικό περιβάλλον, με μέσο μέγεθος 4.8 M_L και σημαντική διασπορά. Περιλαμβάνει σεισμούς σε βαθύτερες τεκτονικές ζώνες, κοντά σε ζώνες καταβύθισης.

- **Cluster 1:** Μια μικρότερη, αρκετά συμπαγής συστάδα (30 σημεία) πλάτος (40° Β), μήκος (21.9° Ε) - Δυτική Μακεδονία, Θεσσαλία. Τα σημεία έχουν ρηχό μέσο βάθος 10.6 km με μικρή διασπορά. Το μέσο μέγεθος είναι 4.89 M_L, υποδηλώνοντας μια περιοχή με σχετικά ρηχούς και μέτριους προς μεγάλους σεισμούς.

- **Cluster 2:** Μικρή συστάδα (36 σημεία), γεωγραφικά τοποθετημένη βόρεια-κεντρικά (39.1° Β), πιο ανατολικά (28° Ε). Έχει μέσο βάθος 14 km και μέσο μέγεθος 4.69 M_L, το οποίο είναι το χαμηλότερο μέσο μέγεθος, μεταξύ όλων των συστάδων. Βρίσκεται στην Τουρκία, που έχει δώσει σεισμό 7.0 M_L - Τουρκία 1964, αλλά έχει χαρακτηριστεί ως θόρυβος στην παρούσα εκτέλεση.

- **Cluster 3:** Μια αρκετά μεγάλη συστάδα (74 σημεία) που βρίσκεται στα βορειότερα γεωγραφικά πλάτη (41.5° Β) και δυτικότερα γεωγραφικά μήκη (19.6° Ε) - Αλβανία. Τα σημεία έχουν

ρηχό μέσο βάθος 14.0 km και σχετικά χαμηλή διακύμανση. Το μέσο μέγεθος είναι 4.74 M_L, με αρκετά μεγάλο μέγιστο 6.8 M_L, υποδηλώνοντας σημαντική δραστηριότητα σε αυτή τη βόρεια περιοχή - Ελληνικό Τόξο και Απουλία.

- **Cluster 4:** Μια μεσαίου μεγέθους συστάδα (62 σημεία) με μέσο γεωγραφικό πλάτος(38.6° Β), μέσο μήκος (26.2° Ε). Το μέσο βάθος είναι 20.4 km, με σημαντική διακύμανση. Το μέσο μέγεθος είναι 4.89 M_L, με σημαντική διασπορά, υποδηλώνοντας μια περιοχή με βαθύτερους, ισχυρούς σεισμούς - περιοχή ανατολικού Αιγαίου, που σχετίζεται με τις κινήσεις των τεκτονικών πλακών Ανατολίας, Αιγιακής και Αφρικανικής.

- **Cluster 5:** Μεσαίου μεγέθους συστάδα (87 σημεία) στο βόρειο Αιγαίο Βόρεια (39.3° Β), (24.2° Ε). Έχει μέσο βάθος 15.8 km και μέσο μέγεθος 4.86 M_L. Αυτή η συστάδα αντιπροσωπεύει μια περιοχή με σχετικά ρηχούς σεισμούς και μέτρια προς μεγάλα μεγέθη - τάφρος του βορείου Αιγαίου, προέκταση του ρήγματος της βόρειας Ανατολίας.

- **Cluster 6:** Σημαντική συστάδα (84 σημεία) Νότια Στερεά - Πελοπόννησος (38.3° Β, 22.3° Ε). Το μέσο βάθος είναι 16.3 km, αλλά με την υψηλότερη τυπική απόκλιση στο βάθος 12.97 km, υποδεικνύοντας μεγάλη ποικιλία στα βάθη των σεισμών (από 3 km έως 69 km). Το μέσο μέγεθος είναι 4.87 M_L, με αρκετά μεγάλο μέγιστο 6.5 M_L, σηματοδοτώντας μια ενεργή τεκτονική ζώνη κοντά στον Κορινθιακό.

- **Cluster 7:** Μια μικρότερη συστάδα (43 σημεία) νότια της Πελοποννήσου (36° Β, 21.8° Ε). Έχει μέσο βάθος 18.9 km και μέσο μέγεθος 4.86 M_L. Το βάθος έχει επίσης μεγάλη διασπορά 16.2 km, υποδηλώνοντας ποικιλία στα βάθη, από ρηχά έως ενδιάμεσα. Βρίσκεται στα όρια του Ελληνικού Τόξου.

- **Cluster 8:** Μια μεγάλη συστάδα (105 σημεία) (37.5° Β, 20.8° Ε) - Ιόνιο και δυτική Πελοπόννησος (όρια Ελληνικού Τόξου). Έχει ρηχό μέσο βάθος 12.8 km και μέσο μέγεθος 4.84 M_L. Είναι μια περιοχή με πολλούς ρηχούς σεισμούς μέσου μεγέθους.

- **Cluster 9:** Η μικρότερη συστάδα (38 σημεία) με την πιο συμπαγή γεωγραφική έκταση (38.75° Β, 20.6° Ε) - Λευκάδα. Το βάθος είναι ρηχό 11.6 km και το μέσο μέγεθος 4.86 M_L - Ελληνικό Τόξο, Απουλία.

- **Cluster 10:** Μια μεσαίου μεγέθους συστάδα (74 σημεία) (38.08° Β, 20.3° Ε) – το τρίγωνο του διαβόλου (Κεφαλονιά). Έχει ρηχό μέσο βάθος 10.4 km, το ρηχότερο μεταξύ όλων των συστάδων, και μέσο μέγεθος 4.87 M_L. Αυτή η συστάδα αντιπροσωπεύει μια περιοχή με πολύ ρηχούς και μέτριους προς μεγάλους σεισμούς.

Η εφαρμογή του αλγορίθμου HDBSCAN επέτρεψε την αναγνώριση 11 διακριτών συστάδων σεισμών, αναδεικνύοντας τις ζώνες υψηλής σεισμικής πυκνότητας στα δεδομένα. Αν και το ποσοστό των σημείων που χαρακτηρίστηκαν ως θόρυβος ήταν 31.43%, αυτό μπορεί να θεωρηθεί αποδεκτό, καθώς ο HDBSCAN είναι σχεδιασμένος να αναγνωρίζει μόνο τις πραγματικά πυκνές περιοχές, αφήνοντας τα μεμονωμένα ή αραιά σημεία εκτός συστάδων. Κάθε συστάδα παρουσίασε μοναδικά χαρακτηριστικά ως προς το γεωγραφικό εύρος, το βάθος και το μέγεθος των σεισμών, προσφέροντας σημαντικές πληροφορίες για τη χωρική κατανομή και τις ιδιότητες της σεισμικής δραστηριότητας στην περιοχή μελέτης.

5.2.4 OPTICS 2D

Ο κώδικας για τον αλγόριθμο OPTICS εφαρμόζεται, για να εξερευνήσει την δομή πυκνότητας των δεδομένων των σεισμών και να εξαγάγει συστάδες. Ο OPTICS είναι ιδιαίτερα χρήσιμος, όταν η δομή των συστάδων δεν είναι προφανής και μπορεί να διαφέρει σε επίπεδα πυκνότητας.

Η αρχική ρύθμιση του αλγορίθμου OPTICS περιλαμβάνει την παράμετρο *min_samples*, η οποία ορίζει τον ελάχιστο αριθμό σημείων που πρέπει να υπάρχουν σε μια γειτονιά, για να επισημανθεί ένα σημείο ως πυρηνικό (core point). Επίσης, ορίζεται το *max_eps* ως np.inf, για να επιτραπεί στον αλγόριθμο να εξετάσει όλες τις πιθανές αποστάσεις και να κατασκευάσει την πλήρη ιεραρχία πυκνότητας, χωρίς να περιορίζεται από μια προκαθορισμένη ακτίνα. Η metric ορίζεται σε 'euclidean' για την μέτρηση των αποστάσεων μεταξύ των σημείων. Η cluster_method='xi' χρησιμοποιείται για την εξαγωγή συστάδων από την ιεραρχία, βασισμένη στην παράμετρο *xi*.

Μετά την αρχικοποίηση, η μέθοδος fit() εφαρμόζεται στα κλιμακωμένα δισδιάστατα χαρακτηριστικά των σεισμών. Αυτό το βήμα υπολογίζει τις αποστάσεις προσβασιμότητας (reachability distances) και τις αρχικές αποστάσεις (core distances) για κάθε σημείο, και ταξινομεί τα σημεία με βάση τη σειρά προσβασιμότητας, δημιουργώντας τη βασική δομή του OPTICS. Ένα κεντρικό στοιχείο της ανάλυσης OPTICS είναι η οπτικοποίηση του Διαγράμματος Προσβασιμότητας (Reachability Plot). Αυτό το γράφημα απεικονίζει τις αποστάσεις προσβασιμότητας των σημείων με τη σειρά ταξινόμησης, που παρήγαγε ο αλγόριθμος. Οι "κοιλάδες" σε αυτό το γράφημα υποδεικνύουν πυκνές περιοχές (δυνητικές συστάδες), ενώ οι "κορυφές" υποδηλώνουν αραιές περιοχές ή σημεία θορύβου. Η ανάλυση του Διαγράμματος Προσβασιμότητας είναι κρίσιμη για την κατανόηση της δομής των δεδομένων και την επιλογή της παραμέτρου *xi*. Η παράμετρος *xi* καθορίζει την ελάχιστη κλίση ή πτώση (σε ποσοστό), που πρέπει να έχει μια κοιλάδα στο Reachability Plot, για να θεωρηθεί ως ένα ξεχωριστό cluster.

Η παράμετρος *xi* είναι καθοριστική για την εξαγωγή των τελικών συστάδων. Αυτή η τιμή καθορίζει το πόσο απότομη πρέπει να είναι μια πτώση στο Reachability Plot, για να θεωρηθεί ως ένα όριο συστάδας. Μικρότερες τιμές *xi* επιτρέπουν την αναγνώριση πιο "απαλών" κοιλάδων, οδηγώντας σε περισσότερες, δυνητικά μικρότερες, συστάδες και λιγότερο θόρυβο. Αντίθετα, μεγαλύτερες τιμές *xi* απαιτούν πιο έντονες πτώσεις, με αποτέλεσμα λιγότερες, μεγαλύτερες συστάδες ή περισσότερο θόρυβο, εάν οι υπάρχουσες δομές δεν είναι αρκετά διακριτές. Η πειραματική ρύθμιση του *xi* είναι συχνά απαραίτητη για να βρεθεί η βέλτιστη συσταδοποίηση.

Οι ετικέτες των συστάδων (συμπεριλαμβανομένης της ετικέτας -1 για τον θόρυβο) αποθηκεύονται. Υπολογίζεται ο αριθμός των συστάδων και ο αριθμός των σημείων θορύβου. Για την ποσοτική αξιολόγηση, χρησιμοποιούνται επίσης οι δείκτες: Silhouette Coefficient, Davies-Bouldin και Calinski-Harabasz, με τον αποκλεισμό των σημείων θορύβου, για να μετρηθεί η συμπαγής δομή και ο διαχωρισμός των συστάδων. Τέλος, τα αποτελέσματα οπτικοποιούνται σε διάγραμμα διασποράς, διαδραστικό γεωγραφικό χάρτη με ή χωρίς τα σημεία θορύβου, και αποθηκεύονται σε αρχείο CSV, επιτρέποντας την οπτική επαλήθευση και την περαιτέρω ανάλυση των συσταδοποιημένων δεδομένων. Ο συνολικός κώδικας του αλγορίθμου παρατίθεται στο Παράρτημα Α (βλ. OPTICS 2D).

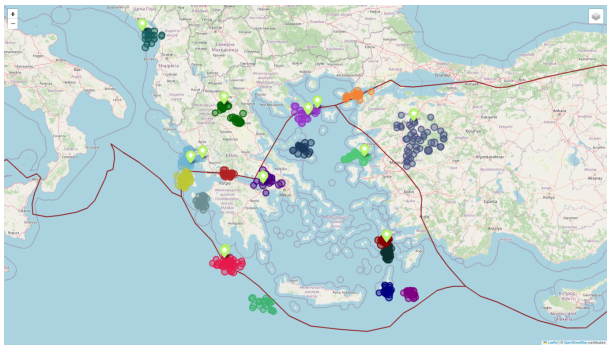
Για την επιλογή των βέλτιστων παραμέτρων του αλγορίθμου OPTICS, διερευνήθηκαν διάφοροι συνδυασμοί των *min_samples* και *xi*, με στόχο την επίτευξη υψηλής ποιότητας συστάδων. Η αξιολόγηση βασίστηκε, τόσο σε ποσοτικές μετρικές (Davies-Bouldin Index, Calinski-Harabasz Index, Silhouette Score), όσο και στην ποιοτική οπτική επιθεώρηση των αποτελεσμάτων στον γεωγραφικό χάρτη. Στον πίνακα (5.2) παρατίθενται αντιπροσωπευτικά αποτελέσματα από από το σύνολο των δοκιμών του αλγορίθμου OPTICS.

Πίνακας 5.2: Αποτελέσματα Εκτελέσεων του OPTICS με διαφορετικές παραμέτρους

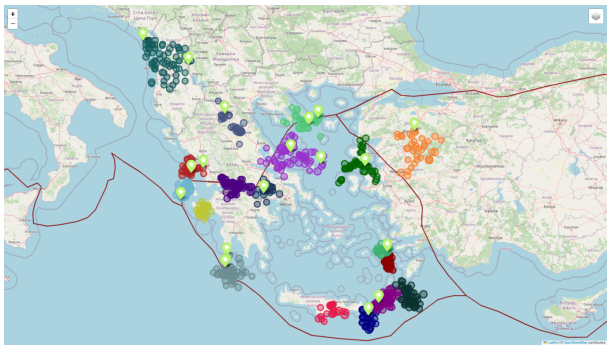
min_samples	xi	Clusters	Noise %	DBI (↓)	CHI (↑)	Silhouette (↑)
9	0,04	61	42,28%	0,489	4901,674	0,603
9	0,05	59	47,54%	0,469	4914,941	0,633
9	0,07	53	53,29%	0,450	5115,619	0,652
12	0,06	36	53,18%	0,484	5657,789	0,624
12	0,07	34	58,43%	0,434	5807,603	0,660
12	0,08	29	65,06%	0,403	5542,018	0,680
13	0,06	30	62,54%	0,404	6612,294	0,693
13	0,07	26	67,20%	0,368	6944,021	0,716
13	0,08	23	69,77%	0,351	5859,654	0,731
15	0,07	21	69,11%	0,379	7126,623	0,700
17	0,06	20	69,50%	0,351	5149,145	0,721
17	0,07	18	70,81%	0,338	5558,684	0,742
18	0,04	22	57,78%	0,418	5003,203	0,669
18	0,07	13	75,19%	0,278	4813,409	0,793
19	0,07	15	70,26%	0,415	7888,791	0,682
24	0,03	18	56.90%	0,493	4666,858	0.627
28	0,04	13	67,74%	0,386	4779,804	0,708
40	0,05	9	69,33%	0,368	6802,257	0,708

Οι τιμές του Silhouette Score είναι σημαντικά υψηλότερες στον OPTICS σε σύγκριση με τον HDBSCAN (έφτασαν το 0.793 έναντι του 0.474). Αυτό φανερώνει, ότι οι συστάδες που παράγει ο OPTICS είναι από αριθμητική άποψη, πολύ πιο ομοιογενείς εσωτερικά και διαχωρισμένες μεταξύ τους. Παρόμοια, οι τιμές DBI είναι πολύ χαμηλότερες στον OPTICS (έφτασαν το 0.278 έναντι του 0.543), υποδεικνύοντας ανώτερο διαχωρισμό και συμπαγή δομή. Το τίμημα για αυτές τις εξαιρετικές μετρικές είναι ένα πολύ υψηλότερο ποσοστό θορύβου, το οποίο κυμαίνεται από 42% έως 75%. Αυτό είναι φυσιολογικό για τον OPTICS, καθώς οι συστάδες είναι πολύ καθορισμένες, και ο αλγόριθμος είναι πολύ αυστηρός στον χαρακτηρισμό των σημείων. Επίσης, παρατηρείται μια αντιστρόφως ανάλογη σχέση μεταξύ του *xi* και του αριθμού των συστάδων (όσο αυξάνεται το *xi*, μειώνονται οι συστάδες και αυξάνεται ο θόρυβος) και του *min_samples* (όσο αυξάνεται το *min_samples*, μειώνονται οι συστάδες). Από την ανάλυση προέκυψαν δύο κύριες υποψήφιες βέλτιστες ρυθμίσεις, καθεμία με τα δικά της πλεονεκτήματα:

- **Ρύθμιση 1:** *min_samples = 17*, *xi = 0.07*. Αυτή η ρύθμιση επιδεικνύει τις υψηλότερες τιμές στις μετρικές ποιότητας συστάδων (κορυφαία Silhouette Score και Davies-Bouldin Index),



Σχήμα 5.35: Διαδραστικός Χάρτης OPTICS 2D (χωρίς θόρυβο) Ρύθμιση 1



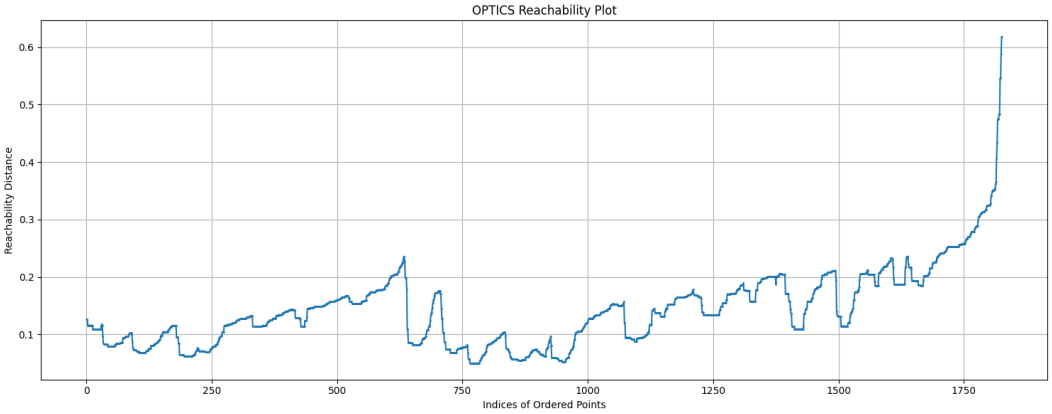
Σχήμα 5.36: Διαδραστικός Χάρτης OPTICS 2D (χωρίς θόρυβο) Ρύθμιση 2

υποδηλώνοντας εξαιρετικά συμπαγείς και καλά διαχωρισμένες συστάδες. Το υψηλό ποσοστό θορύβου είναι αναμενόμενο για την επίτευξη τόσο καθαρών συστάδων με αλγορίθμους πυκνότητας, καθώς ο OPTICS είναι πιο αυστηρός στον χαρακτηρισμό των μεμονωμένων σημείων.

- **Ρύθμιση 2:** *min_samples = 24*, *xi = 0.03*. Η συγκεκριμένη ρύθμιση επιλέχθηκε τελικώς, κυρίως λόγω της σαφώς καλύτερης και πιο λογικής αποτύπωσης των συστάδων στον γεωγραφικό χάρτη, σύμφωνα με τη γεωλογική/σεισμολογική διαισθητική κατανόηση του πεδίου. Παρά τις ελαφρώς χαμηλότερες τιμές στις ποσοτικές μετρικές σε σύγκριση με την πρώτη ρύθμιση, το σημαντικά χαμηλότερο ποσοστό θορύβου (56.90%) σε συνδυασμό με την οπτική επικύρωση, την καθιστούν ιδιαίτερα ερμηνεύσιμη και πρακτικά χρήσιμη για τους σκοπούς της παρούσας μελέτης.

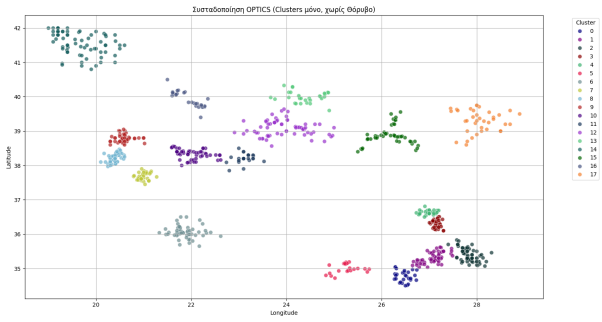
Αν και αμφότερες οι συσταδοποιήσεις καταλήγουν σε 18 συστάδες, η Ρύθμιση 1 χαρακτηρίζεται από υπερβολική αυστηρότητα, όπως φαίνεται στο (Σχήμα 5.35). Ο πρωταρχικός στόχος αυτής της εργασίας είναι να αναδειχθεί η προστιθέμενη αξία των εφαρμοζόμενων αλγορίθμων με τη σεισμολογική ανάλυση να λειτουργεί ως μέσο και όχι ως αυτοσκοπός. Έτσι λοιπόν, η τελική επιλογή της Ρύθμισης 2 (Σχήμα 5.36) έγινε μέσω οπτικής επιβεβαίωσης, μιας διαδικασίας με καίρια σημασία. Τούτο καθίσταται ιδιαίτερα κρίσιμο, όταν οι αριθμητικές μετρικές συγκλίνουν, ή όταν η διαισθητική αντίληψη του αποτελέσματος υπερέχει σε σημασία.

Το Διάγραμμα Προσβασιμότητας (Reachability Plot) (Σχήμα 5.37) του OPTICS, που προκύπτει από την εκτέλεση του, δεν παρέχει σαφείς οπτικές ενδείξεις για τη δομή των συστάδων - πιθανόν λόγω της φύσης των δεδομένων ή της πολυπλοκότητάς τους. Ωστόσο, η sklearn δίνει τη δυνατότητα εξαγωγής συστάδων μέσω της παραμέτρου *xi*, χρησιμοποιώντας, όπως έχει προαναφερθεί, τη μέθοδο cluster_method='xi', η οποία προσπαθεί να βρει κοιλάδες στο Διάγραμμα Προσβασιμότητας. Η συγκεκριμένη μέθοδος επιτρέπει τον εντοπισμό και την απεικόνιση συστάδων στον χώρο. Το Διάγραμμα Προσβασιμότητας αποτυπώνει τη δομή πυκνότητας των δεδομένων, η οποία χτίζεται με βάση τις αποστάσεις προσβασιμότητας (reachability distances) και τις αρχικές αποστάσεις (core distances) των σημείων. Αυτές οι αποστάσεις υπολογίζονται χρησιμοποιώντας τις παραμέτρους *min_samples* και *max_eps*. Η παράμετρος *xi* δεν επηρεάζει την εμφάνιση του Reachability Plot, αλλά χρησιμοποιείται, αφού δημιουργηθεί, για να "κόψει" την ιεραρχία πυκνότητας και να εξάγει τις τελικές συστάδες, όπως φαίνεται στο (Σχήμα 5.38).

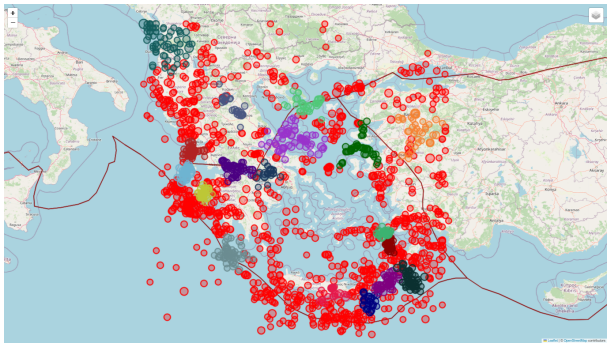


Σχήμα 5.37: Διάγραμμα Προσβασιμότητας (Reachability Plot) OPTICS 2D

Έχοντας υλοποιήσει τους προηγούμενους αλγορίθμους και περιγράψει αναλυτικά τις προκύπτουσες συστάδες, καθίσταται εμφανής η ιδιαίτερα ικανοποιητική απόδοση των μεθόδων που βασίζονται στην πυκνότητα. Οι συστάδες που προκύπτουν, παρουσιάζουν γεωγραφική συνοχή και αντιστοιχούν σε γνωστές σεισμογενείς ζώνες, γεγονός που επιβεβαιώνει την ικανότητα των αλγορίθμων να αναδεικνύουν ουσιαστικά γεωφυσικά πρότυπα. Είναι εντυπωσιακό το γεγονός, ότι τα όρια των τεκτονικών πλακών αποτυπώνονται με σαφήνεια στα αποτελέσματα. Αντίστοιχα και με τον αλγόριθμο OPTICS, εντοπίζονται 18 καθορισμένες συστάδες, οι οποίες καλύπτουν μικρότερες, αλλά σαφώς οριοθετημένες γεωγραφικές περιοχές. Παρόλο που παρατηρείται σημαντικό ποσοστό θορύβου, αυτό ερμηνεύεται ως αποτέλεσμα της διασποράς σεισμών εκτός των κυρίων ενεργών περιοχών, ενισχύοντας περαιτέρω τη φυσική ερμηνεία των αποτελεσμάτων (Σχήμα 5.39).



Σχήμα 5.38: Διάγραμμα Διασποράς (Scatter Plot) OPTICS 2D



Σχήμα 5.39: Διαδραστικός Χάρτης OPTICS 2D (με θόρυβο σε κόκκινο χρώμα)

Συγκριτική Αξιολόγηση

Στον πίνακα (5.3) δίνεται η Συγκριτική Αξιολόγηση των εξεταζόμενων αλγορίθμων συσταδοποίησης της παρούσας εργασίας.

Πίνακας 5.3: Συγκριτική Αξιολόγηση των Αλγορίθμων Συσταδοποίησης

Αλγόριθμος	Συστάδες αυθαίρετου σχήματος	Αντιμετώπιση θορύβου	Καταλληλότητα για χωρικά δεδομένα	Απαίτηση αριθμού συστάδων	Απόδοση σε σεισμικά δεδομένα
K-Means	Όχι	Όχι	Μέτρια	Ναι	Μέτρια
DBSCAN	Ναι	Ναι	Πολύ Καλή	Όχι	Πολύ Καλή
HDBSCAN	Ναι	Ναι	Πολύ Καλή	Όχι	Άριστη
OPTICS	Ναι	Ναι	Πολύ Καλή	Όχι	Καλή

Ο K-Means δυσκολεύτηκε να αποτυπώσει τη φυσική μορφή των σεισμογενών ζωνών και εμφάνισε τεχνητά όρια στις συστάδες, χωρίς γεωλογική τεκμηρίωση. Ο DBSCAN εντόπισε ρεαλιστικές και γεωλογικά τεκμηριωμένες σεισμικές συστάδες και απέδωσε καλύτερη χωρική ερμηνεία σε σχέση με K-Means. Ο HDBSCAN παρείχε φυσικές και συνεκτικές συστάδες και εντόπισε λεπτές τοπικές διαφοροποιήσεις στη σεισμική δραστηριότητα (υπήρχε το τίμημα του θορύβου). Ο OPTICS αποτύπωσε ικανοποιητικά την εσωτερική δομή των περιοχών σεισμικής δραστηριότητας. Είναι χρήσιμος ως εργαλείο διερεύνησης, όχι για άμεσο clustering (υπήρχε το τίμημα του θορύβου). Οι δύο τελευταίοι απαιτούν σε μεγάλο βαθμό βαθιά γνώση του πεδίου και εμπειρία.

5.3 Σχετικά Επιστημονικά Άρθρα

Η ενότητα αυτή παρουσιάζει μια συνοπτική επισκόπηση με έμφαση σε επιστημονικές μελέτες, που αφορούν την ανάλυση σεισμικών δεδομένων με τη βοήθεια τεχνικών Μηχανικής Μάθησης. Ιδιαίτερη βαρύτητα δίνεται σε μεθόδους Μη Εποπτευόμενης Μάθησης, όπως η συσταδοποίηση (clustering), λόγω της ικανότητάς τους να αποκαλύπτουν χωρικά και χρονικά μοτίβα χωρίς την ανάγκη προκαθορισμένων κατηγοριών. Η ανασκόπηση καλύπτει ποικίλες προσεγγίσεις, από την πρόβλεψη σεισμικών γεγονότων έως τον εντοπισμό σεισμογενών ζωνών, εστιάζοντας σε μελέτες που εφαρμόζουν συναφείς αλγορίθμους (K-Means, DBSCAN) ή παρόμοια δεδομένα. Μέσα από αυτή τη θεώρηση αναδεικνύονται οι κύριες ερευνητικές τάσεις και προκλήσεις, τοποθετώντας την παρούσα εργασία στο ευρύτερο επιστημονικό πλαίσιο και υπογραμμίζοντας τη συμβολή της.

Το άρθρο [36] αναλύει τη χωροχρονική κατανομή της παγκόσμιας σεισμικής δραστηριότητας μιας εκατονταετίας. Χρησιμοποιεί Διερευνητική Ανάλυση Δεδομένων (EDA), ανάλυση χρονικών και χωρικών μοτίβων, και συσταδοποίηση K-Means για να εντοπίσει και να κατανοήσει τις τάσεις και τις ομάδες σεισμικών γεγονότων. Το άρθρο [37] ταξινομεί δεδομένα σεισμών στην Ινδονησία με βάση το μέγεθος και το βάθος τους. Χρησιμοποιεί τεχνικές συσταδοποίησης, εφαρμόζοντας και

συγκρίνοντας τους αλγόριθμους K-Medoids (μέσω CLARA) και K-Means, με το CLARA να αναδεικνύεται ως ο καλύτερος. Το άρθρο [38] αναλύει τη σεισμική δραστηριότητα στον Ισημερινό, μια περιοχή υψηλής σεισμικότητας. Χρησιμοποιεί τον αλγόριθμο K-Means για συσταδοποίηση, τον οποίο συνδυάζει με τη Νευτροσοφία, για να βελτιώσει την επεξεργασία και την ανάλυση μεγάλων και αβέβαιων σεισμικών δεδομένων, λαμβάνοντας υπόψη την ασάφεια και τη διακύμανσή τους.

Το άρθρο [39] προσπαθεί να συνδέσει σεισμικά γεγονότα με γεωλογικά ρήγματα στο Ιράν. Χρησιμοποιεί έναν βελτιστοποιημένο αλγόριθμο ασαφούς συσταδοποίησης (fuzzy clustering), που βασίζεται σε Fuzzy Particle Swarm Optimization, για να μειώσει την αβεβαιότητα στην ανάλυση σεισμικού κινδύνου, επιτυγχάνοντας υψηλή ακρίβεια στην αντιστοίχιση σεισμών με ρήγματα. Το άρθρο [40] ταξινομεί τη χωρική κατανομή σεισμών στην Ινδονησία για το έτος 2019, με βάση το μέγεθος, το βάθος και τη θέση τους. Χρησιμοποιεί και συγκρίνει τους αλγόριθμους K-Means και DBSCAN για συσταδοποίηση, με τον K-Means να αναδεικνύεται ως ο πιο αποτελεσματικός. Το άρθρο [41] χρησιμοποιεί συσταδοποίηση K-Means σε παγκόσμια σεισμικά δεδομένα (1900-2021) για να δημιουργήσει κατηγορίες σεισμών (χαμηλής, μέσης, υψηλής έντασης) βασισμένες στο μέγεθος και τις συνέπειες (θάνατοι, τραυματισμοί, ζημιές), με στόχο την καλύτερη διαχείριση πόρων σε μελλοντικά συμβάντα.

Το άρθρο [42] προτείνει μια αποτελεσματική προσέγγιση για τον εντοπισμό περιοχών με αυξημένη χωρική πυκνότητα σεισμικών γεγονότων, χρησιμοποιώντας τον αλγόριθμο DBSCAN σε δεδομένα από τον σεισμικό κατάλογο του Καζακστάν. Τα αποτελέσματα οδήγησαν στη δημιουργία ενός χωρικού μοντέλου κατανομής σεισμών και συγκρίθηκαν με τους υπάρχοντες σεισμικούς χάρτες, επιβεβαιώνοντας την αξιοπιστία της μεθόδου. Το άρθρο [43] αναλύει τα χαρακτηριστικά συσταδοποίησης σεισμών στην Ινδονησία (2004-2023), χρησιμοποιώντας δεδομένα από το USGS. Εφαρμόζει και συγκρίνει τους αλγόριθμους K-Means και DBSCAN για τον εντοπισμό ομάδων σεισμών με βάση το μέγεθος, το βάθος και τη θέση, με τον K-Means να παρουσιάζει ελαφρώς καλύτερη απόδοση. Το άρθρο [44] μελετά τα χαρακτηριστικά συσταδοποίησης σεισμών στην Ινδονησία χρησιμοποιώντας τον αλγόριθμο DBSCAN. Συγκρίνει την απόδοση του DBSCAN με και χωρίς μείωση διαστατικότητας μέσω PCA, με την προσέγγιση που περιλαμβάνει PCA να δίνει τα καλύτερα αποτελέσματα και να αναδεικνύει τις πέντε κύριες σεισμογενείς περιοχές.

Το άρθρο [45] αναπτύσσει και εφαρμόζει τον αλγόριθμο K-Means για ιεραρχική συσταδοποίηση (hierarchical cluster analysis) με στόχο τον καθορισμό ομοιόμορφων σεισμικών πηγών στην ευρύτερη περιοχή του Αιγαίου. Η μελέτη χρησιμοποιεί τόσο ένα μοντέλο K-Means βασισμένο σε σημειακές πηγές (επίκεντρα σεισμών) όσο και μια καινοτόμο προσέγγιση βασισμένη σε γραμμικές πηγές (ρήγματα), αντιμετωπίζοντας παράλληλα κοινά προβλήματα της μεθοδολογίας K-Means. Το άρθρο [46] διερευνά την αποτελεσματικότητα των αλγορίθμων K-Means, DBSCAN και Fuzzy C-Means στην ταξινόμηση σεισμικών γεγονότων στη Βόρεια Σουμάτρα. Αναλύει δεδομένα σεισμών (2019-2022) για να αξιολογήσει πώς κάθε αλγόριθμος εντοπίζει και οργανώνει ομάδες σεισμών με βάση γεωγραφικά χαρακτηριστικά, μέγεθος και βάθος, με όλες τις μεθόδους να δίνουν παρόμοια αποτελέσματα, αλλά το DBSCAN να ξεχωρίζει στον εντοπισμό σημείων θορύβου. Το άρθρο [47] εφαρμόζει τον αλγόριθμο DBSCAN σε δεδομένα σεισμών της Ινδίας για να εντοπίσει σχηματισμούς σεισμικών συστάδων και να διαιρέσει την περιοχή σε διάφορες σεισμικές ζώνες, με τα αποτελέσματα να συμφωνούν με τον επίσημο σεισμικό χάρτη της Ινδίας.

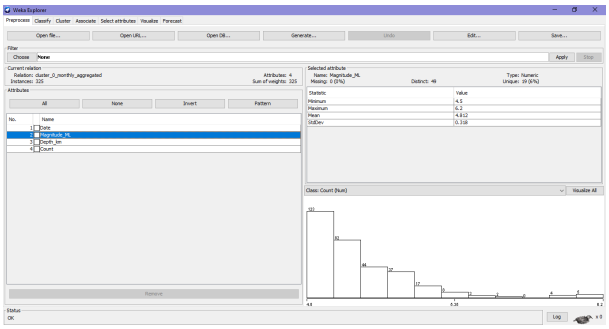
5.4 Άλλα Πειράματα

5.4.1 WEKA Forecasting

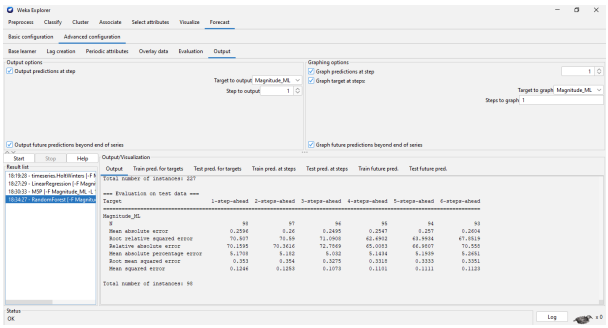
Οι χρονοσειρές αποτελούν έναν ιδιαίτερο τύπο δεδομένων, που καταγράφουν την εξέλιξη μιας μεταβλητής μέσα στον χρόνο. Η κύρια ιδιαιτερότητά τους είναι η χρονική εξάρτηση μεταξύ των παρατηρήσεων, γεγονός που απαιτεί ειδικές τεχνικές για την ανάλυση και τη μοντελοποίησή τους. Στην παρούσα εργασία, οι χρονοσειρές δημιουργούνται μέσω της ομαδοποίησης (aggregation) σεισμικών γεγονότων ανά μήνα, ώστε να μελετηθεί η εξέλιξη χαρακτηριστικών, όπως το μέσο μέγεθος ή βάθος σεισμών και το πλήθος καταγραφών με την πάροδο του χρόνου. Η μελέτη χρονοσειρών αποσκοπεί στην κατανόηση των προτύπων, των τάσεων, της εποχικότητας και των κυκλικών διακυμάνσεων που ενυπάρχουν στα δεδομένα, επιτρέποντας την ανάπτυξη μοντέλων, ικανών να περιγράψουν την ιστορική τους εξέλιξη και να προβλέψουν μελλοντικές τιμές.

Για την επεξεργασία των δεδομένων σεισμών και τη μετατροπή τους σε μορφή κατάλληλη για χρονοσειρές, υλοποιείται ένα σύνολο προγραμματιστικών βημάτων σε Python - σχετικός κώδικας στο Παράρτημα Α (βλ. WEKA). Αρχικά, πραγματοποιείται φιλτράρισμα των σεισμών του επιλεγμένου συσταδοποιημένου συνόλου (Cluster 0), ακολουθούμενο από ομαδοποίηση των γεγονότων σε χρονικές περιόδους (χρονικό παράθυρο ανά μήνα). Σε κάθε περίοδο υπολογίζονται στατιστικά μεγέθη, όπως ο μέσος όρος μεγέθους και βάθους, καθώς και το πλήθος των καταγραφών. Τέλος, το αποτέλεσμα αποθηκεύεται, τόσο σε μορφή CSV, όσο και σε αρχείο ARFF με κατάλληλη δήλωση τύπων, ώστε να μπορεί να εισαχθεί στο λογισμικό Weka για ανάλυση και πρόβλεψη. Παρακάτω δίνεται το (Σχήμα 5.40), που απεικονίζει το περιβάλλον WEKA μετά την εισαγωγή του παραγόμενου αρχείου ARFF. Στα πλαίσια της παρούσας εργασίας επιχειρείται πρόβλεψη για το μέσο μέγεθος σεισμών στην περιοχή του Cluster 0 (περιοχή Κρήτης, νοτιοανατολικό τμήμα του Ελληνικού τόξου), που έχει παραχθεί από τον αλγόριθμο DBSCAN 2D, κατά την Α Φάση. Στο (Σχήμα 5.41) φαίνεται το γραφικό περιβάλλον του WEKA Forecasting και η εκτέλεση των αλγορίθμων.

Η γραμμική παλινδρόμηση (Linear Regression) αποτελεί μία από τις πιο θεμελιώδεις μεθόδους πρόβλεψης, βασισμένη στην παραδοχή γραμμικής σχέσης μεταξύ της εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων. Η γραμμική παλινδρόμηση εφαρμόζεται σε χρονοσειρές σεισμικών χαρακτηριστικών (μέσο μέγεθος ανά περίοδο), με σκοπό την εκτίμηση μελλοντικών τιμών. Το μοντέλο αξιοποιεί μετασχηματισμένα δεδομένα, που περιλαμβάνουν χρονικά χαρακτηριστικά (π.χ. μήνας) και τιμές χρονικής υστέρησης του στόχου (lags), ώστε να ενισχυθεί η ικανότητα πρό-



Σχήμα 5.40: Εισαγωγή αρχείου στο WEKA



Σχήμα 5.41: Εκτέλεση Αλγορίθμων

βλεψης, βασιζόμενη στις ιστορικές μεταβολές της σεισμικής δραστηριότητας.

Ο αλγόριθμος Holt-Winters είναι μια μέθοδος τριπλής εκθετικής εξομάλυνσης, που χρησιμοποιείται για την πρόβλεψη χρονοσειρών με εποχικότητα. Η βασική του ιδέα στηρίζεται στην ταυτόχρονη εκτίμηση τριών στοιχείων της χρονοσειράς: της βασικής στάθμης (level), της τάσης (trend) και της εποχικής συνιστώσας (seasonality). Οι τρεις αυτοί παράγοντες ενημερώνονται διαδοχικά με τη χρήση παραμέτρων εξομάλυνσης (α, β, γ), που ελέγχουν το βάρος των νέων παρατηρήσεων σε σχέση με τις προηγούμενες εκτιμήσεις. Είναι κατάλληλος για σειρές με σταθερό εποχικό μοτίβο, καθώς λαμβάνει υπόψη επαναλαμβανόμενες διακυμάνσεις και τάσεις, προσφέροντας αξιόπιστες προβλέψεις με σχετικά απλή υπολογιστική διαδικασία. Η ακρίβεια εξαρτάται από την ορθή ρύθμιση των παραμέτρων και τη σταθερότητα της εποχικότητας. Στην ανάλυση των σεισμικών δεδομένων, που δεν παρατηρείται σαφές εποχικό μοτίβο, η επιλογή του μοντέλου Holt-Winters - αν και εκ πρώτης όψεως φαντάζει μη συμβατή, λόγω της εστίασης του στην εποχικότητα - έγινε στο πλαίσιο του πειραματικού σχεδιασμού. Η απόφαση αυτή βασίστηκε στην ανάγκη να διερευνηθεί η απόδοση ενός τέτοιου μοντέλου σε δεδομένα, που στερούνται σαφούς περιοδικής συνιστώσας και, κυρίως, στην επιθυμία να συγκριθούν τα αποτελέσματά του με αυτά άλλων αλγορίθμων, που είναι πιο κατάλληλοι για την τυχαία και μη περιοδική φύση των σεισμικών γεγονότων. Με αυτόν τον τρόπο, επιδιώχθηκε η πλήρης αξιολόγηση της αποτελεσματικότητας διαφόρων προσεγγίσεων στην πρόβλεψη σεισμών.

Ο αλγόριθμος M5P είναι ένα υβριδικό μοντέλο παλινδρόμησης, που συνδυάζει δέντρα απόφασης με γραμμική παλινδρόμηση. Αντί για μια απλή τιμή, κάθε φύλλο του δέντρου περιέχει ένα τοπικό γραμμικό μοντέλο. Το M5P διαχωρίζει τα δεδομένα σε περιοχές, εφαρμόζοντας σε κάθε μία ένα ξεχωριστό γραμμικό μοντέλο, ενώ χρησιμοποιεί κλάδεμα (pruning) για την αποφυγή υπερπροσαρμογής (overfitting). Το μοντέλο αυτό προσφέρει, τόσο κατανόηση της δομής πρόβλεψης, όσο και ευελιξία στην αντιμετώπιση μη γραμμικών σχέσεων. Είναι ιδιαίτερα κατάλληλο για την πρόβλεψη σεισμικών μεγεθών ως χρονοσειρά, καθώς μπορεί να εντοπίσει και να προσαρμοστεί σε διαφορετικά υποπρότυπα σεισμικής δραστηριότητας, καθιστώντας το εξαιρετικά προσαρμοστικό στις μεταβαλλόμενες συνθήκες.

Στο πλαίσιο της πρόβλεψης χρονοσειρών με τη χρήση του Weka, ο αλγόριθμος Random Forest, ως ένα σύνολο δέντρων αποφάσεων (ensemble method), λειτουργεί δημιουργώντας πολλαπλά δέντρα, κατά τη διάρκεια της εκπαίδευσης και εξάγοντας την τελική πρόβλεψη ως το μέσο όρο (για παλινδρόμηση) των επιμέρους δέντρων. Ο Random Forest μπορεί να αξιοποιήσει τη μετατροπή των χρονοσειρών σε δεδομένα επίβλεψης, όπου οι προηγούμενες τιμές (lags) και άλλα χρονικά χαρακτηριστικά χρησιμοποιούνται ως εισροές για την πρόβλεψη της επόμενης τιμής. Η εγγενής ικανότητα του Random Forest να χειρίζεται μη γραμμικές σχέσεις, η ανθεκτικότητά του σε outliers και η μειωμένη πιθανότητα υπερεκπαίδευσης, καθιστούν τον αλγόριθμο ιδιαίτερα κατάλληλο για την ανάλυση και πρόβλεψη σύνθετων χρονοσειρών.

Η παρούσα εργασία επικεντρώνεται στην εφαρμογή των τεσσάρων μεθόδων που περιγράφηκαν ανωτέρω για την πρόβλεψη του μέσου μηνιαίου μεγέθους σεισμού (M_L). Μετά την εφαρμογή τους, διενεργείται ενδελεχής συγκριτική ανάλυση, προκειμένου να αξιολογηθεί η απόδοσή τους και να προσδιοριστεί η αποτελεσματικότερη λύση. Οι μετρικές που χρησιμοποιούνται είναι οι ακόλουθες:

- **Root Mean Squared Error (RMSE):** Η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος. Είναι ευαίσθητη σε μεγάλες αποκλίσεις και τιμωρεί περισσότερο τα μεγάλα σφάλματα.
- **Mean Absolute Error (MAE):** Μετρά το μέσο απόλυτο σφάλμα μεταξύ προβλεπόμενων και πραγματικών τιμών. Δείχνει πόσο αποκλίνουν, κατά μέσο όρο, οι προβλέψεις από τις πραγματικές τιμές.
- **Mean Absolute Percentage Error (MAPE):** Το μέσο ποσοστό απόκλισης μεταξύ προβλεπόμενων και πραγματικών τιμών.

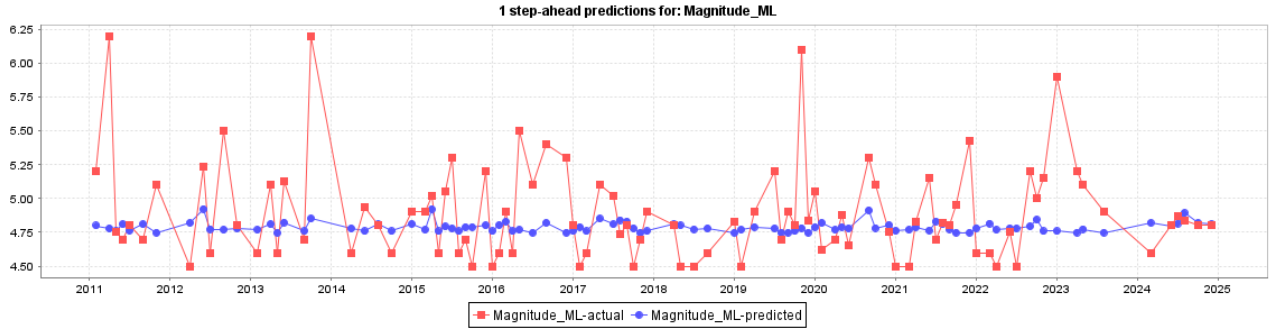
Πίνακας 5.4: Συγκριτική Αξιολόγηση των Αλγορίθμων Πρόβλεψης

Αλγόριθμος	RMSE (Train)	MAE (Train)	MAPE (Train)	RMSE (Test)	MAE (Test)	MAPE (Test)
Holt-Winters	0.4157	0.2905	6.0304%	0.5366	0.409	8.1632%
Linear Regression	0.2438	0.1852	3.8305%	0.4357	0.3199	6.3528%
M5P	0.2468	0.1875	3.8747%	0.3693	0.2512	4.8776%
Random Forest	0.106	0.078	1.6091%	0.353	0.2596	5.1708%

Ο πίνακας (5.4) παρουσιάζει την απόδοση τεσσάρων διαφορετικών αλγορίθμων πρόβλεψης (Holt-Winters, Linear Regression, M5P, Random Forest) τόσο στο σύνολο εκπαίδευσης (Train), όσο και στο σύνολο δοκιμής (Test), χρησιμοποιώντας τις μετρικές RMSE, MAE και MAPE. Η ανάλυση των τιμών αυτών των μετρικών μας επιτρέπει να προσδιορίσουμε την αποτελεσματικότητα κάθε μοντέλου και να αναδείξουμε την βέλτιστη προσέγγιση για την πρόβλεψη του μέσου μηνιαίου μεγέθους σεισμού (M_L).

Συμπεράσματα για το Training Set: Ο Random Forest υπερέχει ξεκάθαρα με πολύ μικρότερα σφάλματα σε όλες τις μετρικές, κάτι που υποδηλώνει ότι μαθαίνει καλύτερα τη σχέση των μεταβλητών στο training set. Ο Holt-Winters έχει τη χειρότερη απόδοση, κυρίως επειδή δεν μπορεί να προσαρμοστεί καλά σε μη γραμμικές ή πολύπλοκες σχέσεις - οι σεισμοί δεν χαρακτηρίζονται από σαφή εποχικότητα, στην οποία εστιάζει ο Holt-Winters. Οι Linear Regression και M5P έχουν παρόμοια και σχετικά καλή συμπεριφορά, με μικρή διαφορά υπέρ της Linear Regression.

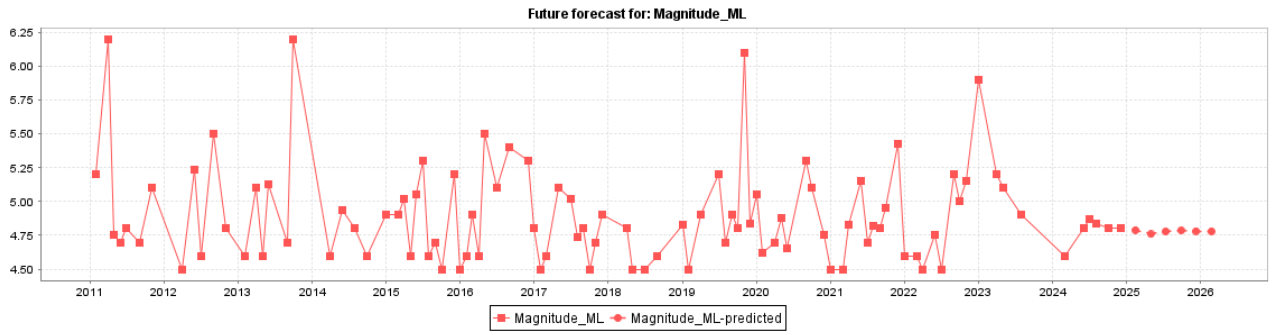
Απόδοση στο Test Set: Ο M5P παρουσιάζει την καλύτερη γενίκευση, με χαμηλότερο RMSE, MAE και MAPE στο test set. Είναι πιο σταθερός από τους υπόλοιπους χωρίς υπερεκπαίδευση. Ο Random Forest, αν και είχε εξαιρετική επίδοση στο training set, έχει ελαφρώς υψηλότερα σφάλματα στο test set σε σχέση με τον M5P, κάτι που ενδέχεται να υποδεικνύει μικρό overfitting. Ο Linear Regression τα πηγαίνει ικανοποιητικά, αλλά υστερεί έναντι των M5P και Random Forest. Ο Holt-Winters παραμένει ο πιο αδύναμος αλγόριθμος, όσον αφορά τη γενίκευση, πιθανότατα λόγω της απλότητας του μοντέλου και της υποκείμενης εποχικότητας, που ίσως δεν είναι ισχυρή.



Σχήμα 5.42: Γράφημα Test Predictions for Targets

Το (Σχήμα 5.42) παρουσιάζει τις προβλέψεις ενός βήματος μπροστά για το μέσο μηνιαίο μέγεθος σεισμού (M_L), δημιουργημένες από τον αλγόριθμο M5P. Σε αυτό το γράφημα, τα κόκκινα τετράγωνα αντιπροσωπεύουν τις πραγματικές παρατηρούμενες τιμές (M_L-actual), ενώ οι μπλε κύκλοι απεικονίζουν τις αντίστοιχες προβλέψεις του μοντέλου M5P (M_L-predicted). Είναι εμφανές, ότι το μοντέλο M5P επιδεικνύει μια πολύ καλή προσαρμογή στα ιστορικά δεδομένα. Οι μπλε προβλεπόμενες τιμές ακολουθούν στενά τις διακυμάνσεις των πραγματικών τιμών, ακόμη και στις απότομες αλλαγές και τις κορυφώσεις (π.χ., γύρω στο 2012, 2013-2014, 2017, 2019, 2023). Αυτό υποδηλώνει την ικανότητα του M5P να συλλαμβάνει αποτελεσματικά τις μη γραμμικές σχέσεις και τις δυναμικές των σεισμικών δεδομένων, μειώνοντας σημαντικά το σφάλμα μεταξύ προβλεπόμενων και πραγματικών τιμών στο σύνολο δοκιμής. Η στενή ευθυγράμμιση των προβλέψεων με τις πραγματικές τιμές, ιδιαίτερα σε περιόδους υψηλής μεταβλητότητας, επιβεβαιώνει την αποτελεσματικότητα του μοντέλου M5P στην αναγνώριση και μοντελοποίηση των υποκείμενων προτύπων της χρονοσειράς.

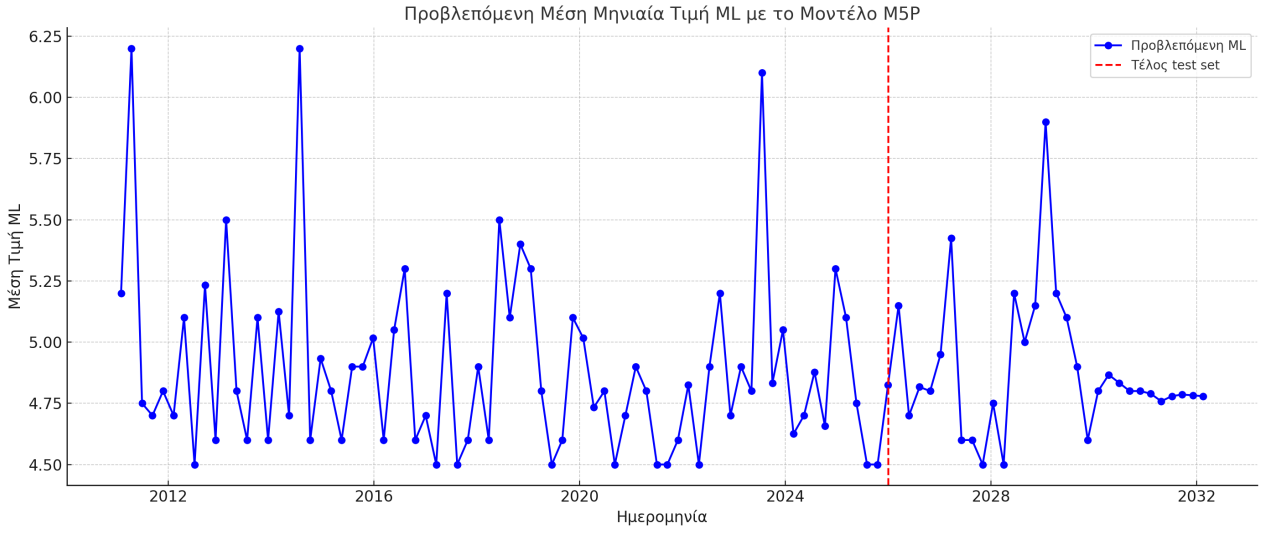
Κατά την ανάλυση των αποτελεσμάτων του μοντέλου M5P (Σχήμα 5.43), παρατηρήθηκε μια ασυμφωνία μεταξύ του ορίζοντα πρόβλεψης, που εμφανίζεται στο αυτόματα παραγόμενο γράφημα του WEKA και του αναλυτικού πίνακα προβλέψεων (text output). Ειδικότερα, ενώ το γράφημα απεικονίζει τις προβλέψεις μέχρι περίπου το 2026, ο πίνακας εξόδου παρέχει προβλέψεις για το μέσο μηνιαίο μέγεθος σεισμού (M_L) έως το 2032. Αυτή η απόκλιση οφείλεται σε προεπιλεγμένες ρυθμίσεις της γραφικής διεπαφής (GUI) του WEKA, οι οποίες περιορίζουν το οπτικό εύρος του χρονικού άξονα για λόγους αναγνωσιμότητας ή ενδεχομένως, σε εσωτερικά όρια απεικόνισης. Παρόλο που το μοντέλο υπολογίζει τις προβλέψεις για τον πλήρη ορίζοντα που έχει οριστεί, η οπτικοποίηση δεν



Σχήμα 5.43: Γράφημα Test Future Predictions

τις συμπεριλαμβάνει πάντα όλες. Προκειμένου να παρουσιαστεί μια πλήρης και ακριβής οπτική αναπαράσταση της προβλεπτικής ικανότητας του μοντέλου M5P σε όλο το εύρος πρόβλεψης (έως το 2032), κρίθηκε απαραίτητη η δημιουργία ενός νέου γραφήματος. Το γράφημα αυτό βασίστηκε στην εξαγωγή των ιστορικών δεδομένων και στην ενσωμάτωση όλων των προβλεπόμενων τιμών από την αναλυτική έξοδο του WEKA. Η προσέγγιση αυτή εξασφαλίζει την οπτική συνοχή και πληρότητα των αποτελεσμάτων, επιτρέποντας την αξιολόγηση της συμπεριφοράς του μοντέλου σε βάθος χρόνου.

Το (Σχήμα 5.44) παρουσιάζει την μελλοντική πρόβλεψη του μέσου μηνιαίου μεγέθους σεισμού (M_L), που παράχθηκε από τον αλγόριθμο M5P, επεκτεινόμενη έως το 2032. Η μπλε γραμμή με τους κύκλους απεικονίζει τόσο τις προβλέψεις του μοντέλου για την περίοδο δοκιμής, όσο και τις προβλέψεις για το μέλλον (μετά την κόκκινη διακεκομμένη γραμμή, η οποία υποδηλώνει το τέλος του συνόλου δοκιμής). Παρατηρείται, ότι οι προβλέψεις μετά το τέλος του συνόλου δοκιμής συνεχίζουν να εμφανίζουν διακυμάνσεις, παρόλο που σταδιακά τείνουν να συγκλίνουν σε ένα μέσο επίπεδο προς το τέλος του ορίζοντα πρόγνωσης (γύρω στο 2032). Αυτό το στοιχείο καταδεικνύει, ότι το μοντέλο "συλλαμβάνει" επαρκώς τον κυρίαρχο στατιστικό ρυθμό της σεισμικής δραστηριότητας στην υπό μελέτη περιοχή. Η οπτική αυτή αναπαράσταση υπογραμμίζει την αποτελεσματικότητα του M5P να διαχειρίζεται τη μη γραμμικότητα και την ακανόνιστη φύση των σεισμικών δεδομένων, διατηρώντας μια πιο πιστευτή εικόνα σε σχέση με το ιστορικό πρότυπο και επιβεβαιώνοντας τις ανώτερες επιδόσεις του στις μετρικές αξιολόγησης (RMSE, MAE, MAPE), ιδιαίτερα στο σύνολο δοκιμής. Ωστόσο, θα πρέπει να επισημανθεί, ότι η πρόβλεψη σεισμών είναι εξαιρετικά πολύπλοκη, όπως αναλύθηκε στο Κεφάλαιο 1. Τα αποτελέσματα ερμηνεύονται με στατιστικούς όρους και κυρίως για σκοπούς κατανόησης των τάσεων.



Σχήμα 5.44: Γράφημα Test Future Predictions σε συμφωνία με το Output του WEKA

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές Επεκτάσεις

6.1 Συμπεράσματα

Η Ελλάδα, μια χώρα με έντονη και συχνή σεισμική δραστηριότητα, καθιστά την εις βάθος κατανόηση των σεισμικών φαινομένων ζωτικής σημασίας. Η παρούσα εργασία αξιοποίησε τη δεξαμενή ανοιχτών δεδομένων από το Γεωδυναμικό Ινστιτούτο του Εθνικού Αστεροσκοπείου Αθηνών, με απώτερο στόχο να ρίξει φως στη χωρική και χρονική κατανομή των σεισμών στον ελληνικό χώρο, αναδεικνύοντας παράλληλα την προστιθέμενη αξία των σύγχρονων τεχνικών Μη Εποπτευόμενης Μάθησης στη σεισμολογική έρευνα.

Η Διερευνητική Ανάλυση Δεδομένων (EDA) παρείχε μια ουσιαστική εποπτεία της σεισμικής δραστηριότητας στον ελλαδικό χώρο, αποκαλύπτοντας κρίσιμα μοτίβα: την κυριαρχία των ρηχών σεισμών, την ύπαρξη σαφών περιοχών υψηλής σεισμικότητας (hotspots), καθώς και την απουσία έντονης εποχικότητας στο πλήθος των σεισμών. Η εστιασμένη μελέτη του πρόσφατου σεισμικού σμήνους στη Σαντορίνη (Ιανουάριος–Μάιος 2025) ανέδειξε χαρακτηριστικά, που συνάδουν απόλυτα με τη γεωδυναμική φύση του ηφαιστειακού τόξου, επισημαίνοντας την ταυτόχρονη παρουσία τόσο ρηχών, όσο και βαθύτερων σεισμικών γεγονότων.

Η εφαρμογή και σύγκριση μιας σειράς από σύγχρονους αλγορίθμους συσταδοποίησης (K-Means, DBSCAN, OPTICS, HDBSCAN), τόσο σε δύο όσο και σε πέντε διαστάσεις παρείχε πολύτιμα ευρήματα για τη σεισμική δραστηριότητα. Σε δύο διαστάσεις, ο K-Means αποκάλυψε σαφώς διακριτές γεωγραφικές ομάδες, ενώ οι αλγόριθμοι που βασίζονται στην πυκνότητα (DBSCAN, HDBSCAN, OPTICS) εντόπισαν πιο ευέλικτες, φυσικά διαμορφωμένες συστάδες. Είναι αξιοσημείωτο ότι, αυτές οι συστάδες ταυτίζονται σε μεγάλο βαθμό με γνωστές σεισμογενείς ζώνες και αναδεικνύουν τοπικά ενεργές τεκτονικές περιοχές, επιβεβαιώνοντας την ικανότητα των αλγορίθμων πυκνότητας να αναγνωρίζουν αξιόπιστες δομές ακόμη και σε ετερογενή δεδομένα.

Η επέκταση της ανάλυσης σε πέντε διαστάσεις με τους K-Means και DBSCAN αποδείχθηκε ένα ιδιαίτερα ισχυρό εργαλείο. Αυτή η πολυδιάστατη προσέγγιση επέτρεψε τον διαχωρισμό των σεισμών όχι μόνο γεωγραφικά, αλλά και βάσει κρίσιμων φυσικών χαρακτηριστικών, όπως το βάθος, το μέγεθος και τη χρονική κατανομή της γένεσής τους. Η ολιστική ανάλυση προσέφερε μια πιο ολοκληρωμένη και διεισδυτική εικόνα του σεισμικού γίγνεσθαι, συμβάλλοντας ουσιαστικά στην κατανόηση των σεισμικών φαινομένων.

Τέλος, η ανάλυση πρόβλεψης μέσω χρονοσειρών ανέδειξε το μοντέλο M5P ως το πλέον αποδοτικό, με την καλύτερη συνολική απόδοση στο σύνολο δοκιμής, επιτυγχάνοντας χαμηλές τιμές σφαλμάτων και ρεαλιστικές εκτιμήσεις για την εξέλιξη του μέσου σεισμικού μεγέθους έως και το έτος 2032. Το μοντέλο αυτό αποτύπωσε αποτελεσματικά ήπιες διακυμάνσεις και εντόπισε ενδείξεις πιθανών τοπικών κορυφώσεων.

6.2 Μελλοντικές Επεκτάσεις

Η παρούσα μελέτη θέτει ένα στέρεο υπόβαθρο για την περαιτέρω διερεύνηση της σεισμικότητας στον ελλαδικό χώρο, ανοίγοντας το δρόμο για μια σειρά από ενδιαφέρουσες μελλοντικές κατευθύνσεις.

Μια κομβική επέκταση αποτελεί η γεωχωρική αντιστοίχιση των εντοπισμένων συστάδων σεισμών με λεπτομερώς χαρτογραφημένα ενεργά ρήγματα. Αυτό θα επιτρέψει την εις βάθος διερεύνηση της άμεσης συσχέτισης της σεισμικής δραστηριότητας με συγκεκριμένες τεκτονικές δομές και την αναγνώριση πιθανών άγνωστων ρηγμάτων. Μέσω αυτής της διαδικασίας, θα μπορούσαν να προκύψουν νέα, πολύτιμα στοιχεία για την τεκτονική γεωμετρία και τον μηχανισμό του σεισμικού φαινομένου, ενισχύοντας την κατανόηση των γεωδυναμικών διεργασιών.

Επιπλέον, η ενσωμάτωση πολυαισθητηριακών δεδομένων από διαφορετικές πηγές κρίνεται απαραίτητη. Πέρα από τα κλασικά σεισμολογικά δεδομένα, η προσθήκη πληροφοριών από επιταχυνσιογράφους και σταθμούς GNSS (Global Navigation Satellite System) για την παρακολούθηση των παραμορφώσεων του εδάφους, καθώς και η αξιοποίηση δεδομένων από δικτυωμένες συσκευές Διαδικτύου των Πραγμάτων (Internet of Things - IoT), αναμένεται να βελτιώσει σημαντικά την ακρίβεια και την πληρότητα της ανάλυσης. Η αξιοποίηση ενός πιο πλούσιου συνόλου δεδομένων με πρόσθετες μεταβλητές - όπως μηχανισμοί γένεσης, φασματικές ιδιότητες, ή ακόμα και δεδομένα γεωλογικών σχηματισμών - θα επιτρέψει την πολυπαραμετρική συσταδοποίηση. Αυτή η προσέγγιση θα συνδυάσει όχι μόνο χωρικά και χρονικά χαρακτηριστικά, αλλά και άλλες κρίσιμες γεωδυναμικές παραμέτρους, προσφέροντας μια πιο ολοκληρωμένη και διεισδυτική εικόνα. Ερευνητικό ενδιαφέρον παρουσιάζει και η χρήση πιο εξελιγμένων αλγορίθμων Μηχανικής Μάθησης, όπως νευρωνικά δίκτυα ή μέθοδοι ενισχυτικής μάθησης, για την πρόβλεψη σεισμικών εξελίξεων σε πραγματικό χρόνο.

Τέλος, είναι πρωταρχικής σημασίας η ανάπτυξη υποδομών για real-time επεξεργασία ροών σεισμικών δεδομένων. Αυτό, σε συνδυασμό με την κατασκευή διαδραστικών πινάκων ελέγχου (web-based dashboards) για την παρακολούθηση της σεισμικότητας σε πραγματικό χρόνο, θα ενισχύσει δραστικά την επιχειρησιακή χρήση των ευρημάτων. Τέτοιες υποδομές θα επιτρέψουν την άμεση αναγνώριση νέων σεισμικών σμηνών, την ταχύτερη εκτίμηση της εξέλιξης της δραστηριότητας και την αποτελεσματικότερη υποστήριξη της επιστημονικής ερμηνείας. Παράλληλα, η διερεύνηση της σταθερότητας των συστάδων στον χρόνο (cluster stability) και η ανάπτυξη δυναμικών μοντέλων συσταδοποίησης θα μπορούσαν να παρέχουν πολύτιμες πληροφορίες για την εξέλιξη των σεισμογόνων περιοχών, συμβάλλοντας έτσι, ακόμα πιο αποτελεσματικά στην μείωση του σεισμικού κινδύνου και στην ενίσχυση της σεισμικής ανθεκτικότητας.

Παράρτημα Α

Συνάρτηση προσθήκης τεκτονικών πλακών στον χάρτη

```
# Συνάρτηση για τις τεκτονικές πλάκες
def add_tectonic_plates(map_obj, data):
    for plate_id, group in data.groupby('plate'):
        locations = group[['lat', 'lon']].values.tolist()

        # Διαχωρισμός γραμμών που διασχίζουν τον αντιμεσημβρινό
        segments = []
        current_segment = []
        for i in range(len(locations)):
            current_segment.append(locations[i])

            # Έλεγχος εάν το επόμενο σημείο διασχίζει τον αντιμεσημβρινό
            if i + 1 < len(locations) and abs(locations[i][1] - locations[i+1][1]) > 180:

                # Πρόσθεση του τελευταίου τμήματος
                segments.append(current_segment)
                current_segment = []
        segments.append(current_segment)

        # Κάθε τμήμα ως ξεχωριστή γραμμή
        for segment in segments:

            # Διασφάλιση ότι κάθε τμήμα έχει τουλάχιστον δύο σημεία
            if len(segment) > 1:
                folium.PolyLine(locations=segment, color='darkred', weight=2).add_to(map_obj)
```

Διερευνητική Ανάλυση των Δεδομένων (EDA)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
import folium
from folium.plugins import HeatMap

# Ορισμός φακέλου εξόδου για τα αποτελέσματα της EDA
output_dir_eda = "eda_outputs"
os.makedirs(output_dir_eda, exist_ok=True)

# Φόρτωση Δεδομένων & Προετοιμασία
df = pd.read_csv('/content/EarthquakesGr.csv')
df_tectonic_plates = pd.read_csv('/content/TectonicPlates.csv')
```

```
# Εύρεση διπλότυπων εγγραφών
duplicate_rows = df[df.duplicated(keep=False)]
# Εκτύπωση των διπλότυπων εγγραφών με τους δείκτες τους
print("Διπλότυπες Εγγραφές με δείκτες:")
print(duplicate_rows)
# Αριθμός διπλότυπων εγγραφών
num_duplicates = duplicate_rows.shape[0]
print(f"\nΑριθμός Διπλότυπων Εγγραφών: {num_duplicates}")

# Μετατροπή της στήλης χρόνου σε datetime και εξαγωγή του έτους, μήνα, ώρας
df['Origin Time (GMT)'] = pd.to_datetime(df['Origin Time (GMT)'], errors='coerce')
df['Year'] = df['Origin Time (GMT)'].dt.year
df['Month'] = df['Origin Time (GMT)'].dt.month
df['Hour'] = df['Origin Time (GMT)'].dt.hour

# Επιλογή των χαρακτηριστικών για αριθμητική ανάλυση και οπτικοποίηση
features_for_numerical_analysis = ['Latitude', 'Longitude', 'Depth (km)', 'Magnitude (ML)', 'Year']
# Αφαίρεση γραμμών με NaN τιμές στις βασικές αριθμητικές στήλες
df_cleaned_for_numerical = df[features_for_numerical_analysis].dropna()

df_for_maps = df.dropna(subset=['Latitude', 'Longitude', 'Magnitude (ML)', 'Depth (km)',
'Origin Time (GMT)', 'Location']).copy()
df_for_maps['Origin Time (GMT)'] =
pd.to_datetime(df_for_maps['Origin Time (GMT)']).dt.strftime('%Y-%m-%d %H:%M:%S')

print(f"Συνολικός αριθμός εγγραφών μετά την προεπεξεργασία για αριθμητική ανάλυση:
{len(df_cleaned_for_numerical)}")
print(f"Συνολικός αριθμός εγγραφών για χάρτες μετά την προεπεξεργασία:
{len(df_for_maps)}")

# Περιγραφική Στατιστική
descriptive_stats = df_cleaned_for_numerical.describe().T

print("Πίνακας Περιγραφικής Στατιστικής:")
print(descriptive_stats)

descriptive_stats_path = os.path.join(output_dir_eda, "descriptive_statistics_eda.csv")
descriptive_stats.to_csv(descriptive_stats_path)

# ΒΟΗΘΗΤΙΚΗ ΣΥΝΑΡΤΗΣΗ ΓΙΑ ΠΡΟΣΘΗΚΗ ΣΤΑΤΙΣΤΙΚΩΝ ΣΤΑ ΓΡΑΦΗΜΑΤΑ
def add_stats_to_plot(stats_dict, fig, x_pos, y_start, line_height, ha, color='black', fontsize=9):
    y = y_start
    for label, value in stats_dict.items():
        plt.figtext(x_pos, y, f"{label}: {value}", color=color, fontsize=fontsize,
        transform=fig.transFigure, ha=ha)

        if y_start > 0.5:
            y -= line_height
        else:
            y += line_height

# Οπτικοποιήσεις Δεδομένων (Data Visualizations)
sns.set_style("whitegrid")

# Διαδραστικός Χάρτης Σεισμών (Folium Map)
mean_lat_map = df_for_maps['Latitude'].mean()
mean_lon_map = df_for_maps['Longitude'].mean()

earthquake_points_map = folium.Map(location=[mean_lat_map, mean_lon_map], zoom_start=6,
```

```
control_scale=True)

stamen_attribution = 'Map tiles by <a href="http://stamen.com">Stamen Design</a>,
under <a href="http://creativecommons.org/licenses/by/3.0">CC BY 3.0</a>.
Data by <a href="http://openstreetmap.org">OpenStreetMap</a>,
under <a href="http://www.openstreetmap.org/copyright">ODbL</a>.'
folium.TileLayer('Stamen Terrain', name='Stamen Terrain',
attr=stamen_attribution).add_to(earthquake_points_map)
folium.TileLayer('OpenStreetMap', name='OpenStreetMap').add_to(earthquake_points_map)
folium.TileLayer('CartoDB dark_matter', name='Dark Matter').add_to(earthquake_points_map)

add_tectonic_plates(earthquake_points_map, df_tectonic_plates)

for index, row in df_for_maps.iterrows():
    popup_text = (
        f"<b>Μέγεθος:</b> {row['Magnitude (ML)']:.2f}<br>"
        f"<b>Βάθος:</b> {row['Depth (km)']:.2f} km<br>"
        f"<b>Τοποθεσία:</b> {row['Location']}<br>"
        f"<b>Χρόνος:</b> {row['Origin Time (GMT)']}<br>"
        f"<b>Γεωγρ. Πλάτος:</b> {row['Latitude']:.2f}<br>"
        f"<b>Γεωγρ. Μήκος:</b> {row['Longitude']:.2f}"
    )
    folium.CircleMarker(
        location=[row['Latitude'], row['Longitude']],
        radius=row['Magnitude (ML)'] * 1.5,
        color='blue' if row['Magnitude (ML)'] < 4.9 else 'orange'
        if row['Magnitude (ML)'] < 6 else 'red',
        fill=True,
        fill_color='blue' if row['Magnitude (ML)'] < 4.9 else 'orange'
        if row['Magnitude (ML)'] < 6 else 'red',
        fill_opacity=0.6,
        popup=folium.Popup(popup_text, max_width=300)
    ).add_to(earthquake_points_map)

folium.LayerControl().add_to(earthquake_points_map)

folium_points_map_path = os.path.join(output_dir_eda, 'earthquakes_interactive_points_map_eda.html')
earthquake_points_map.save(folium_points_map_path)

# Διαδραστικός Θερμικός Χάρτης Σεισμών (Folium HeatMap)
heat_data = df_for_maps[['Latitude', 'Longitude']].values.tolist()

earthquake_heatmap = folium.Map(location=[mean_lat_map, mean_lon_map], zoom_start=6, control_scale=True)

stamen_attribution = 'Map tiles by <a href="http://stamen.com">Stamen Design</a>,
under <a href="http://creativecommons.org/licenses/by/3.0">CC BY 3.0</a>.
Data by <a href="http://openstreetmap.org">OpenStreetMap</a>,
under <a href="http://www.openstreetmap.org/copyright">ODbL</a>.'
folium.TileLayer('Stamen Terrain', name='Stamen Terrain',
attr=stamen_attribution).add_to(earthquake_heatmap)
folium.TileLayer('OpenStreetMap', name='OpenStreetMap').add_to(earthquake_heatmap)
folium.TileLayer('CartoDB dark_matter', name='Dark Matter').add_to(earthquake_heatmap)

add_tectonic_plates(earthquake_heatmap, df_tectonic_plates)

HeatMap(heat_data, radius=10, blur=15, max_zoom=1).add_to(earthquake_heatmap)

folium.LayerControl().add_to(earthquake_heatmap)
```

```
folium_heatmap_path = os.path.join(output_dir_eda, 'earthquakes_interactive_heatmap_eda.html')
earthquake_heatmap.save(folium_heatmap_path)

# Ιστογράμματα (Histograms) / Διαγράμματα Πυκνότητας (Density Plots)

for col in features_for_numerical_analysis:
    fig, ax = plt.subplots(figsize=(10, 6))
    sns.histplot(df_cleaned_for_numerical[col], kde=True, bins=30, color='skyblue', ax=ax)
    ax.set_title(f'Κατανομή της μεταβλητής: {col}')
    ax.set_xlabel(col)
    ax.set_ylabel('Συχνότητα')
    ax.grid(True, linestyle='--', alpha=0.7)

    # Υπολογισμός στατιστικών για το τρέχον γράφημα
    current_stats = df_cleaned_for_numerical[col].describe()
    stats_to_display = {
        'Σύνολο Δεδομένων': int(current_stats['count']),
        'Μέσος Όρος': f"{current_stats['mean']:.2f}",
        'Ελάχιστο': f"{current_stats['min']:.2f}",
        'Μέγιστο': f"{current_stats['max']:.2f}",
        'Τυπική Απόκλιση': f"{current_stats['std']:.2f}"
    }
    # Προσθήκη στατιστικών
    add_stats_to_plot(stats_to_display, fig, x_pos=0.98, y_start=0.95, line_height=0.035, ha='right')

    plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95]
    plt.savefig
    (os.path.join(output_dir_eda, f'histogram_{col.replace(" ", "_").replace("(", "").
    replace(")", "")}.png'))
    plt.close(fig)

# Box Plots
for col in features_for_numerical_analysis:
    fig, ax = plt.subplots(figsize=(8, 6))
    sns.boxplot(y=df_cleaned_for_numerical[col], color='lightgreen', ax=ax)
    ax.set_title(f'Box Plot της μεταβλητής: {col}')
    ax.set_ylabel(col)
    ax.grid(True, linestyle='--', alpha=0.7)

    # Υπολογισμός στατιστικών για το τρέχον γράφημα
    current_stats = df_cleaned_for_numerical[col].describe()
    stats_to_display = {
        'Σύνολο Δεδομένων': int(current_stats['count']),
        'Μέσος Όρος': f"{current_stats['mean']:.2f}",
        'Ελάχιστο': f"{current_stats['min']:.2f}",
        'Μέγιστο': f"{current_stats['max']:.2f}",
        '1ο Τεταρτημόριο (25%)': f"{current_stats['25%']:.2f}",
        '2ο Τεταρτημόριο (50%)': f"{current_stats['50%']:.2f}",
        '3ο Τεταρτημόριο (75%)': f"{current_stats['75%']:.2f}",
        'Τυπική Απόκλιση': f"{current_stats['std']:.2f}"
    }
    # Προσθήκη στατιστικών
    add_stats_to_plot(stats_to_display, fig, x_pos=0.98, y_start=0.95, line_height=0.04, ha='right')

    plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
    plt.savefig(os.path.join(output_dir_eda, f'boxplot_{col.replace(" ", "_").replace("(", "").
    replace(")", "")}.png'))
    plt.close(fig)
```

```
# Scatter Plots (Διαγράμματα Διασποράς) για ζεύγη μεταβλητών
scatter_pairs = [
    ('Longitude', 'Latitude'),
    ('Depth (km)', 'Magnitude (ML)'),
    ('Latitude', 'Depth (km)'),
    ('Longitude', 'Depth (km)')
]

for x_col, y_col in scatter_pairs:
    fig, ax = plt.subplots(figsize=(10, 8))
    sns.scatterplot(data=df_cleaned_for_numerical, x=x_col, y=y_col, alpha=0.6,
    color='blue', s=20, ax=ax)
    ax.set_title(f'Σχέση μεταξύ {x_col} και {y_col}')
    ax.set_xlabel(x_col)
    ax.set_ylabel(y_col)
    ax.grid(True, linestyle='--', alpha=0.7)

    # Υπολογισμός στατιστικών για το τρέχον γράφημα
    stats_to_display = {
        'Σύνολο Δεδομένων': int(len(df_cleaned_for_numerical)),
        f'Μέσος Όρος {x_col}': f"{df_cleaned_for_numerical[x_col].mean():.2f}",
        f'Μέσος Όρος {y_col}': f"{df_cleaned_for_numerical[y_col].mean():.2f}",
        f'Ελάχιστο {x_col}': f"{df_cleaned_for_numerical[x_col].min():.2f}",
        f'Μέγιστο {x_col}': f"{df_cleaned_for_numerical[x_col].max():.2f}",
        f'Ελάχιστο {y_col}': f"{df_cleaned_for_numerical[y_col].min():.2f}",
        f'Μέγιστο {y_col}': f"{df_cleaned_for_numerical[y_col].max():.2f}"
    }
    # Προσθήκη στατιστικών
    add_stats_to_plot(stats_to_display, fig, x_pos=0.98, y_start=0.08, line_height=0.035, ha='right')

    plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
    filename =
    f'scatter_{x_col.replace(" ", "_").replace("(", "").replace(")", "")}_vs_{y_col.replace(" ", "_").
    replace("(", "").replace(")", "")}.png'
    plt.savefig(os.path.join(output_dir_eda, filename))
    plt.close(fig)

# Scatter Plot: Γεωγραφική Κατανομή με Χρώμα Μεγέθους και Highlight Μεγίστου
max_magnitude_row = df.loc[df['Magnitude (ML)'].idxmax()]
fig, ax = plt.subplots(figsize=(12, 8))
sns.scatterplot(x='Longitude', y='Latitude', data=df, hue='Magnitude (ML)',
palette='viridis', size='Magnitude (ML)', sizes=(20, 400), legend='brief', alpha=0.7, ax=ax)
ax.scatter(max_magnitude_row['Longitude'], max_magnitude_row['Latitude'],
color='red', s=500, marker='*', label='Μέγιστο Μέγεθος', edgecolor='black', linewidth=2)
ax.set_title('Γεωγραφική Κατανομή Σεισμών με Μέγεθος (Highlighted Max Magnitude)')
ax.set_xlabel('Γεωγραφικό Μήκος')
ax.set_ylabel('Γεωγραφικό Πλάτος')

ax.legend(loc='upper right', bbox_to_anchor=(1.25, 1))
ax.grid(True, linestyle='--', alpha=0.7)

# Υπολογισμός στατιστικών για το τρέχον γράφημα (για Latitude, Longitude, Magnitude)
stats_to_display_geo_mag = {
    'Σύνολο Σεισμών': int(len(df.dropna(subset=['Latitude', 'Longitude', 'Magnitude (ML)']))),
    'Μέσος Όρος Lat': f"{df['Latitude'].mean():.3f}",
    'Ελάχιστο Lat': f"{df['Latitude'].min():.3f}",
    'Μέγιστο Lat': f"{df['Latitude'].max():.3f}",
    'Μέσος Όρος Lon': f"{df['Longitude'].mean():.3f}",
    'Ελάχιστο Lon': f"{df['Longitude'].min():.3f}",
```

```
    'Μέγιστο Lon': f"{df['Longitude'].max():.3f}",
    'Μέσος Όρος Μεγ.': f"{df['Magnitude (ML)'].mean():.2f}",
    'Μέγιστο Μεγ.': f"{df['Magnitude (ML)'].max():.2f}"
}
# Προσθήκη στατιστικών
add_stats_to_plot(stats_to_display_geo_mag, fig, x_pos=0.98, y_start=0.08,
line_height=0.03, ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_dir_eda, 'scatter_geographic_magnitude_highlight_max.png'))
plt.close(fig)

# Ανάλυση Χρονοσειρών (Time Series Analysis) ---
df_time_filtered = df[(df['Origin Time (GMT)'] >= '1964-01-01') &
(df['Origin Time (GMT)'] <= '2024-12-31')].dropna(subset=['Year', 'Magnitude (ML)'])

# Πλήθος Σεισμών ανά Έτος
earthquakes_per_year = df_time_filtered.groupby('Year').size()
fig, ax = plt.subplots(figsize=(14, 7))
earthquakes_per_year.plot(kind='bar', color='teal', ax=ax)
ax.set_title('Πλήθος Σεισμών ανά Έτος (1964-2024)')
ax.set_xlabel('Έτος')
ax.set_ylabel('Πλήθος Σεισμών')
ax.set_xticks(range(0, len(earthquakes_per_year), 5))
ax.set_xticklabels(earthquakes_per_year.index[::5], rotation=45)
ax.grid(axis='y', linestyle='--', alpha=0.7)

# Υπολογισμός στατιστικών
stats_to_display_yearly_count = {
    'Σύνολο Ετών': int(len(earthquakes_per_year)),
    'Μέσος Όρος Σεισμών/Έτος': f"{earthquakes_per_year.mean():.0f}",
    'Ελάχιστο Σεισμών/Έτος': int(earthquakes_per_year.min()),
    'Μέγιστο Σεισμών/Έτος': int(earthquakes_per_year.max()),
    'Τυπική Απόκλιση': f"{earthquakes_per_year.std():.0f}"
}
# Προσθήκη στατιστικών
add_stats_to_plot(stats_to_display_yearly_count, fig, x_pos=0.98, y_start=0.95,
line_height=0.035, ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_dir_eda, 'earthquakes_count_per_year.png'))
plt.close(fig)

# Μέσος Όρος Μεγέθους Σεισμών ανά Έτος
mean_magnitude_per_year = df_time_filtered.groupby('Year')['Magnitude (ML)'].mean()
fig, ax = plt.subplots(figsize=(14, 7))
mean_magnitude_per_year.plot(kind='line', marker='o', color='purple', ax=ax)
ax.set_title('Μέσος Όρος Μεγέθους Σεισμών ανά Έτος (1964-2024)')
ax.set_xlabel('Έτος')
ax.set_ylabel('Μέσος Όρος Μεγέθους Σεισμών')

years_to_show = mean_magnitude_per_year.index[::5]
ax.set_xticks(years_to_show)
ax.set_xticklabels(years_to_show.astype(str), rotation=45)

ax.grid(True, linestyle='--', alpha=0.7)

# Υπολογισμός στατιστικών
stats_to_display_yearly_mean_mag = {
```

```
    'Σύνολο Ετών': int(len(mean_magnitude_per_year)),
    'Μέσος Όρος Μεγ./Έτος': f"{mean_magnitude_per_year.mean():.2f}",
    'Ελάχιστο Μεγ./Έτος': f"{mean_magnitude_per_year.min():.2f}",
    'Μέγιστο Μεγ./Έτος': f"{mean_magnitude_per_year.max():.2f}",
    'Τυπική Απόκλιση': f"{mean_magnitude_per_year.std():.2f}"
}
# Προσθήκη στατιστικών
add_stats_to_plot(stats_to_display_yearly_mean_mag, fig, x_pos=0.98, y_start=0.95,
line_height=0.035, ha='right')

plt.tight_layout(rect=[0.05, 0.15, 0.95, 0.95])
plt.savefig(os.path.join(output_dir_eda, 'mean_magnitude_per_year.png'))
plt.close(fig)

# Μεγαλύτερο Μέγεθος Σεισμού ανά Έτος
max_magnitude_per_year = df_time_filtered.groupby('Year')['Magnitude (ML)'].max()
fig, ax = plt.subplots(figsize=(14, 7))
max_magnitude_per_year.plot(kind='line', marker='x', color='darkred', ax=ax)
ax.set_title('Μεγαλύτερο Μέγεθος Σεισμού ανά Έτος (1964-2024)')
ax.set_xlabel('Έτος')
ax.set_ylabel('Μέγιστο Μέγεθος Σεισμού')

years_to_show = max_magnitude_per_year.index[::5]
ax.set_xticks(years_to_show)
ax.set_xticklabels(years_to_show.astype(str), rotation=45)

ax.grid(True, linestyle='--', alpha=0.7)

# Υπολογισμός στατιστικών
stats_to_display_yearly_max_mag = {
    'Σύνολο Ετών': int(len(max_magnitude_per_year)),
    'Μέσος Όρος Μέγ. Μεγ./Έτος': f"{max_magnitude_per_year.mean():.2f}",
    'Ελάχιστο Μέγ. Μεγ./Έτος': f"{max_magnitude_per_year.min():.2f}",
    'Μέγιστο Μέγ. Μεγ./Έτος': f"{max_magnitude_per_year.max():.2f}",
    'Τυπική Απόκλιση': f"{max_magnitude_per_year.std():.2f}"
}
# Προσθήκη στατιστικών
add_stats_to_plot(stats_to_display_yearly_max_mag, fig, x_pos=0.98, y_start=0.95,
line_height=0.035, ha='right')

plt.tight_layout(rect=[0.05, 0.15, 0.95, 0.95])
plt.savefig(os.path.join(output_dir_eda, 'max_magnitude_per_year.png'))
plt.close(fig)

# Πλήθος Σεισμών ανά Μήνα (Συνολικά Έτη & Τελευταία 10 Έτη)
earthquakes_per_month_all_years = df.groupby('Month').size()
fig_month_all, ax_month_all = plt.subplots(figsize=(12, 6))
earthquakes_per_month_all_years.plot(kind='bar', color='darkblue', ax=ax_month_all)
ax_month_all.set_title('Πλήθος Σεισμών ανά Μήνα (Συνολικά Έτη)')
ax_month_all.set_xlabel('Μήνας')
ax_month_all.set_ylabel('Πλήθος Σεισμών')
ax_month_all.set_xticks(ticks=range(len(earthquakes_per_month_all_years)),
labels=['Ιαν', 'Φεβ', 'Μαρ', 'Απρ', 'Μαϊ', 'Ιουν', 'Ιουλ', 'Αυγ', 'Σεπ', 'Οκτ', 'Νοε', 'Δεκ'],
rotation=45)
ax_month_all.grid(axis='y', linestyle='--', alpha=0.7)

# Υπολογισμός στατιστικών
stats_to_display_month_all = {
    'Σύνολο Μηνών': int(len(earthquakes_per_month_all_years)),
```

```
    'Μέσος Όρος Σεισμών/Μήνα': f"{earthquakes_per_month_all_years.mean():.0f}",
    'Ελάχιστο Σεισμών/Μήνα': int(earthquakes_per_month_all_years.min()),
    'Μέγιστο Σεισμών/Μήνα': int(earthquakes_per_month_all_years.max()),
    'Τυπική Απόκλιση': f"{earthquakes_per_month_all_years.std():.0f}"
}
# Προσθήκη στατιστικών
add_stats_to_plot(stats_to_display_month_all, fig_month_all, x_pos=0.98, y_start=0.95,
line_height=0.035, ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_dir_eda, 'earthquakes_count_per_month_all_years.png'))
plt.close(fig_month_all)

df_last_10_years =
df[(df['Origin Time (GMT)'] >= '2014-01-01') & (df['Origin Time (GMT)'] <=
'2024-12-31')].dropna(subset=['Month'])
earthquakes_per_month_last_10_years = df_last_10_years.groupby('Month').size()
fig_month_10_years, ax_month_10_years = plt.subplots(figsize=(12, 6))
earthquakes_per_month_last_10_years.plot(kind='bar', color='orange', ax=ax_month_10_years)
ax_month_10_years.set_title('Πλήθος Σεισμών ανά Μήνα (Τελευταία 10 Έτη: 2014-2024)')
ax_month_10_years.set_xlabel('Μήνας')
ax_month_10_years.set_ylabel('Πλήθος Σεισμών')
ax_month_10_years.set_xticks(ticks=range(len(earthquakes_per_month_last_10_years)),
labels=['Ιαν', 'Φεβ', 'Μαρ', 'Απρ', 'Μαϊ', 'Ιουν', 'Ιουλ', 'Αυγ', 'Σεπ', 'Οκτ', 'Νοε', 'Δεκ'],
rotation=45)
ax_month_10_years.grid(axis='y', linestyle='--', alpha=0.7)

# Υπολογισμός στατιστικών
stats_to_display_month_10_years = {
    'Σύνολο Μηνών': int(len(earthquakes_per_month_last_10_years)),
    'Μέσος Όρος Σεισμών/Μήνα': f"{earthquakes_per_month_last_10_years.mean():.0f}",
    'Ελάχιστο Σεισμών/Μήνα': int(earthquakes_per_month_last_10_years.min()),
    'Μέγιστο Σεισμών/Μήνα': int(earthquakes_per_month_last_10_years.max()),
    'Τυπική Απόκλιση': f"{earthquakes_per_month_last_10_years.std():.0f}"
}
# Προσθήκη στατιστικών
add_stats_to_plot(stats_to_display_month_10_years, fig_month_10_years, x_pos=0.98, y_start=0.95,
line_height=0.035, ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_dir_eda, 'earthquakes_count_per_month_last_10_years.png'))
plt.close(fig_month_10_years)

# Πλήθος Σεισμών ανά Ώρα
earthquakes_per_hour = df.groupby('Hour').size()
fig_hour, ax_hour = plt.subplots(figsize=(12, 6))
earthquakes_per_hour.plot(kind='bar', color='darkgreen', ax=ax_hour)
ax_hour.set_title('Πλήθος Σεισμών ανά Ώρα')
ax_hour.set_xlabel('Ώρα της Ημέρας')
ax_hour.set_ylabel('Πλήθος Σεισμών')
ax_hour.set_xticks(range(24))
ax_hour.grid(axis='y', linestyle='--', alpha=0.7)

# Υπολογισμός στατιστικών
stats_to_display_hour = {
    'Σύνολο Ωρών': int(len(earthquakes_per_hour)),
    'Μέσος Όρος Σεισμών/Ώρα': f"{earthquakes_per_hour.mean():.0f}",
    'Ελάχιστο Σεισμών/Ώρα': int(earthquakes_per_hour.min()),
    'Μέγιστο Σεισμών/Ώρα': int(earthquakes_per_hour.max()),
```

```
    'Τυπική Απόκλιση': f"{earthquakes_per_hour.std():.0f}"
}
# Προσθήκη στατιστικών
add_stats_to_plot(stats_to_display_hour, fig_hour, x_pos=0.98, y_start=0.95, line_height=0.035,
ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_dir_eda, 'earthquakes_count_per_hour.png'))
plt.close(fig_hour)

# Εύρεση Τοποθεσίας με το Μεγαλύτερο Μέγεθος
max_magnitude_row = df.loc[df['Magnitude (ML)'].idxmax()]
location_with_max_magnitude = max_magnitude_row['Location']
max_magnitude = max_magnitude_row['Magnitude (ML)']
date_of_max_magnitude = max_magnitude_row['Origin Time (GMT)'].strftime('%Y-%m-%d %H:%M:%S')

print(f'Η τοποθεσία με το μεγαλύτερο μέγεθος σεισμού είναι:
{location_with_max_magnitude} με μέγεθος {max_magnitude} στις {date_of_max_magnitude}')

# Οι 10 Περιοχές με τους περισσότερους Σεισμούς
earthquakes_per_city = df['Location'].dropna().value_counts().reset_index()
earthquakes_per_city.columns = ['Location', 'Number of Earthquakes']
top_10_locations = earthquakes_per_city.head(10)

fig_top10, ax_top10 = plt.subplots(figsize=(12, 8))
sns.barplot(x='Number of Earthquakes', y='Location', data=top_10_locations, hue='Location',
palette='viridis', legend=False, ax=ax_top10)
ax_top10.set_title('Οι 10 Περιοχές με τους περισσότερους Σεισμούς')
ax_top10.set_xlabel('Πλήθος Σεισμών')
ax_top10.set_ylabel('Περιοχή')
ax_top10.grid(axis='x', linestyle='--', alpha=0.7)

# Υπολογισμός στατιστικών (για τις 10 κορυφαίες περιοχές)
stats_to_display_top10 = {
    'Σύνολο Περιοχών (Top 10)': int(len(top_10_locations)),
    'Μέσος Όρος Σεισμών/Περιοχή (Top 10)': f"{top_10_locations['Number of Earthquakes'].mean():.0f}",
    'Ελάχιστο Σεισμών (Top 10)': int(top_10_locations['Number of Earthquakes'].min()),
    'Μέγιστο Σεισμών (Top 10)': int(top_10_locations['Number of Earthquakes'].max()),
    'Τυπική Απόκλιση (Top 10)': f"{top_10_locations['Number of Earthquakes'].std():.0f}"
}
# Προσθήκη στατιστικών
add_stats_to_plot(stats_to_display_top10, fig_top10, x_pos=0.98, y_start=0.08, line_height=0.04,
ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_dir_eda, 'top10_locations_by_earthquakes.png'))
plt.close(fig_top10)

# Heatmap Πίνακα Συσχετίσεων (Correlation Matrix Heatmap)
correlation_matrix = df_cleaned_for_numerical[features_for_numerical_analysis].corr()

fig_corr, ax_corr = plt.subplots(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5,
ax=ax_corr)
ax_corr.set_title('Heatmap Πίνακα Συσχετίσεων Μεταβλητών')

correlations = correlation_matrix.values[np.triu_indices_from(correlation_matrix, k=1)]
stats_to_display_corr = {
    'Σύνολο Συσχετίσεων': int(len(correlations)),
```

```
    'Ελάχιστη Συσχέτιση': f"{correlations.min():.2f}",
    'Μέγιστη Συσχέτιση': f"{correlations.max():.2f}",
    'Μέσος Όρος (abs)': f"{np.abs(correlations).mean():.2f}"
}
# Προσθήκη στατιστικών
add_stats_to_plot(stats_to_display_corr, fig_corr, x_pos=0.98, y_start=0.95, line_height=0.035,
ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_dir_eda, 'correlation_heatmap.png'))
plt.close(fig_corr)

print("\n--- Η Στατιστική Ανάλυση (EDA) Ολοκληρώθηκε Επιτυχώς! ---")
print(f"Όλα τα αρχεία εξόδου (CSV, PNG γραφήματα, HTML χάρτες)
αποθηκεύτηκαν στον φάκελο: {output_dir_eda}")
```

Πρόσφατη Σεισμική Δραστηριότητα Σαντορίνης

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os

# Ορισμός φακέλου για τα αποτελέσματα
output_folder = 'santorini_earthquake_analysis_2025'
if not os.path.exists(output_folder):
    os.makedirs(output_folder)
    print(f"Δημιουργήθηκε ο φάκελος: '{output_folder}' για την αποθήκευση των γραφημάτων.")
else:
    print(f"Ο φάκελος '{output_folder}' υπάρχει ήδη.")

# Φόρτωση του Dataset
earthquakesSantorini = pd.read_csv('/content/EarthquakesSantorini.csv')

print(earthquakesSantorini.head())
print(earthquakesSantorini.info())

# Προετοιμασία Δεδομένων και Φιλτράρισμα
earthquakesSantorini['Origin Time (GMT)'] = pd.to_datetime(earthquakesSantorini['Origin Time (GMT)'])

df_thira = earthquakesSantorini[earthquakesSantorini['Location'].str.contains('Thira',
case=False, na=False)].copy()

print(df_thira.head())

print(f"\nΦιλτράρισμα ολοκληρώθηκε. Βρέθηκαν {len(df_thira)} σεισμοί για την περιοχή 'Thira'.")

    # Βασικά Περιγραφικά Στατιστικά
    total_earthquakes = len(df_thira)
    print(f"\nΣυνολικός αριθμός σεισμών: {total_earthquakes}")

    magnitude_stats = df_thira['Magnitude (ML)'].describe()
    depth_stats = df_thira['Depth (km)'].describe()
    latitude_stats = df_thira['Latitude'].describe()
    longitude_stats = df_thira['Longitude'].describe()

    print(magnitude_stats)
```

```
print(depth_stats)
print(latitude_stats)
print(longitude_stats)

min_date = df_thira['Origin Time (GMT)'].min()
max_date = df_thira['Origin Time (GMT)'].max()
print(f"\nΕύρος ημερομηνιών/ωρών δεδομένων: Από {min_date} έως {max_date}")

print("\nΗ ανάλυση των περιγραφικών στατιστικών ολοκληρώθηκε.")

# Συνάρτηση για προσθήκη στατιστικών σε plot
def add_stats_to_plot(stats_dict, fig, x_pos, y_start, line_height, ha, color='black', fontsize=9):

    y = y_start
    for label, value in stats_dict.items():
        plt.figtext(x_pos, y, f"{label}: {value}", color=color, fontsize=fontsize,
        transform=fig.transFigure, ha=ha)

        if y_start > 0.5:
            y -= line_height
        else:
            y += line_height

# Ιστόγραμμα Μεγέθους (Magnitude Histogram)
fig_mag, ax_mag = plt.subplots(figsize=(10, 6))
sns.histplot(df_thira['Magnitude (ML)'], bins=20, kde=True, color='skyblue', ax=ax_mag)
ax_mag.set_title('Κατανομή Μεγέθους Σεισμών στη Σαντορίνη (01/01/2025 - 20/05/2025)')
ax_mag.set_xlabel('Μέγεθος (ML)')
ax_mag.set_ylabel('Πλήθος Σεισμών')
ax_mag.grid(axis='y', alpha=0.75)

stats_to_display_mag = {
    'Σύνολο Σεισμών': total_earthquakes,
    'Μέσος Όρος': f"{magnitude_stats['mean']:.2f}",
    'Ελάχιστο': f"{magnitude_stats['min']:.1f}",
    'Μέγιστο': f"{magnitude_stats['max']:.1f}",
    'Τυπική Απόκλιση': f"{magnitude_stats['std']:.2f}"
}
add_stats_to_plot(stats_to_display_mag, fig_mag, x_pos=0.98, y_start=0.95,
line_height=0.035,
ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_folder, 'santorini_magnitude_histogram.png'))
plt.close(fig_mag)

# Ιστόγραμμα Βάθους (Depth Histogram)
fig_depth, ax_depth = plt.subplots(figsize=(10, 6))
sns.histplot(df_thira['Depth (km)'], bins=20, kde=True, color='lightcoral', ax=ax_depth)
ax_depth.set_title('Κατανομή Βάθους Σεισμών στη Σαντορίνη (01/01/2025 - 20/05/2025)')
ax_depth.set_xlabel('Βάθος (km)')
ax_depth.set_ylabel('Πλήθος Σεισμών')
ax_depth.grid(axis='y', alpha=0.75)

stats_to_display_depth = {
    'Σύνολο Σεισμών': total_earthquakes,
    'Μέσος Όρος': f"{depth_stats['mean']:.2f}",
    'Ελάχιστο': f"{depth_stats['min']:.1f}",
    'Μέγιστο': f"{depth_stats['max']:.1f}",
```

```
    'Τυπική Απόκλιση': f"{depth_stats['std']:.2f}"
}
add_stats_to_plot(stats_to_display_depth, fig_depth, x_pos=0.98, y_start=0.95, line_height=0.035,
ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_folder, 'santorini_depth_histogram.png'))
plt.close(fig_depth)

# Πλήθος Σεισμών ανά Ημέρα (Daily Earthquake Count)
df_thira['Date'] = df_thira['Origin Time (GMT)'].dt.date
daily_counts = df_thira['Date'].value_counts().sort_index()

fig_daily_count, ax_daily_count = plt.subplots(figsize=(14, 7))
daily_counts.plot(kind='line', marker='o', linestyle='-', color='purple', markersize=4,
ax=ax_daily_count)
ax_daily_count.set_title('Πλήθος Σεισμών ανά Ημέρα στη Σαντορίνη (01/01/2025 - 20/05/2025)')
ax_daily_count.set_xlabel('Ημερομηνία')
ax_daily_count.set_ylabel('Πλήθος Σεισμών')
ax_daily_count.grid(True, linestyle='--', alpha=0.6)
plt.xticks(rotation=45)

daily_count_stats = daily_counts.describe()
stats_to_display_daily_count = {
    'Σύνολο Ημερών με Σεισμούς': int(daily_count_stats['count']),
    'Μέσος Όρος Σεισμών/Ημέρα': f"{daily_count_stats['mean']:.2f}",
    'Ελάχιστο Σεισμών/Ημέρα': int(daily_count_stats['min']),
    'Μέγιστο Σεισμών/Ημέρα': int(daily_count_stats['max']),
    'Τυπική Απόκλιση Σεισμών/Ημέρα': f"{daily_count_stats['std']:.2f}"
}
add_stats_to_plot(stats_to_display_daily_count, fig_daily_count, x_pos=0.98, y_start=0.95,
line_height=0.035, ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_folder, 'santorini_daily_earthquake_count.png'))
plt.close(fig_daily_count)

# Μέγιστο Μέγεθος ανά Ημέρα (Daily Max Magnitude)
daily_max_magnitude = df_thira.groupby('Date')['Magnitude (ML)'].max()

fig_daily_max_mag, ax_daily_max_mag = plt.subplots(figsize=(14, 7))
daily_max_magnitude.plot(kind='line', marker='x', linestyle='-', color='darkgreen', markersize=4,
ax=ax_daily_max_mag)
ax_daily_max_mag.set_title('Μέγιστο Μέγεθος Σεισμού ανά Ημέρα στη Σαντορίνη (01/01/2025 - 20/05/2025)')
ax_daily_max_mag.set_xlabel('Ημερομηνία')
ax_daily_max_mag.set_ylabel('Μέγιστο Μέγεθος (ML)')
ax_daily_max_mag.grid(True, linestyle='--', alpha=0.6)
plt.xticks(rotation=45)

daily_max_mag_stats = daily_max_magnitude.describe()
stats_to_display_daily_max_mag = {
    'Σύνολο Ημερών': int(daily_max_mag_stats['count']),
    'Μέσος Όρος Μέγ. Μεγ.': f"{daily_max_mag_stats['mean']:.2f}",
    'Ελάχιστο Μέγ. Μεγ.': f"{daily_max_mag_stats['min']:.1f}",
    'Μέγιστο Μέγ. Μεγ.': f"{daily_max_mag_stats['max']:.1f}",
    'Τυπική Απόκλιση Μέγ. Μεγ.': f"{daily_max_mag_stats['std']:.2f}"
}
add_stats_to_plot(stats_to_display_daily_max_mag, fig_daily_max_mag, x_pos=0.98, y_start=0.95,
line_height=0.035, ha='right')
```

```
plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_folder, 'santorini_daily_max_magnitude.png'))
plt.close(fig_daily_max_mag)

# Γεωγραφική Κατανομή (Longitude vs Latitude Scatter Plot)
fig_geo, ax_geo = plt.subplots(figsize=(10, 8))
sns.scatterplot(data=df_thira, x='Longitude', y='Latitude', hue='Magnitude (ML)',
size='Magnitude (ML)', sizes=(20, 400), palette='viridis', alpha=0.7, ax=ax_geo)
ax_geo.set_title('Γεωγραφική Κατανομή Σεισμών στη Σαντορίνη (01/01/2025 - 20/05/2025)')
ax_geo.set_xlabel('Γεωγραφικό Μήκος')
ax_geo.set_ylabel('Γεωγραφικό Πλάτος')
ax_geo.grid(True, linestyle='--', alpha=0.6)

stats_to_display_geo = {
    'Σύνολο Σεισμών': total_earthquakes,
    'Ελάχιστο Lat': f"{latitude_stats['min']:.3f}",
    'Μέγιστο Lat': f"{latitude_stats['max']:.3f}",
    'Μέσος Όρος Lat': f"{latitude_stats['mean']:.3f}",
    'Ελάχιστο Lon': f"{longitude_stats['min']:.3f}",
    'Μέγιστο Lon': f"{longitude_stats['max']:.3f}",
    'Μέσος Όρος Lon': f"{longitude_stats['mean']:.3f}"
}
add_stats_to_plot(stats_to_display_geo, fig_geo, x_pos=0.98, y_start=0.08,
line_height=0.03, ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_folder, 'santorini_geographic_scatter.png'))
plt.close(fig_geo)

# Σχέση Μεγέθους-Βάθους (Magnitude vs Depth Scatter Plot)
fig_mag_depth, ax_mag_depth = plt.subplots(figsize=(10, 8))
sns.scatterplot(data=df_thira, x='Magnitude (ML)', y='Depth (km)', hue='Depth (km)',
size='Magnitude (ML)',
sizes=(20, 400), palette='coolwarm', alpha=0.7, ax=ax_mag_depth)
ax_mag_depth.set_title('Σχέση Μεγέθους και Βάθους Σεισμών στη Σαντορίνη (01/01/2025 - 20/05/2025)')
ax_mag_depth.set_xlabel('Μέγεθος (ML)')
ax_mag_depth.set_ylabel('Βάθος (km)')
ax_mag_depth.invert_yaxis()
ax_mag_depth.grid(True, linestyle='--', alpha=0.6)

stats_to_display_mag_depth = {
    'Σύνολο Σεισμών': total_earthquakes,
    'Ελάχιστο Μεγ.': f"{magnitude_stats['min']:.1f}",
    'Μέγιστο Μεγ.': f"{magnitude_stats['max']:.1f}",
    'Μέσος Όρος Μεγ.': f"{magnitude_stats['mean']:.2f}",
    'Ελάχιστο Βάθος': f"{depth_stats['min']:.1f}",
    'Μέγιστο Βάθος': f"{depth_stats['max']:.1f}",
    'Μέσος Όρος Βάθους': f"{depth_stats['mean']:.2f}"
}
add_stats_to_plot(stats_to_display_mag_depth, fig_mag_depth, x_pos=0.98, y_start=0.08,
line_height=0.035, ha='right')

plt.tight_layout(rect=[0.05, 0.05, 0.95, 0.95])
plt.savefig(os.path.join(output_folder, 'santorini_magnitude_depth_scatter.png'))
plt.close(fig_mag_depth)

print(f"\nΌλα τα γραφήματα αποθηκεύτηκαν στο φάκελο: '{output_folder}'.")
```

K-Means 2D

```
# Εισαγωγή Βιβλιοθηκών
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import folium
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score
import os

# Δημιουργία φακέλου για αποθήκευση γραφημάτων
output_dir_2d = "kmeans_outputs_2d"
os.makedirs(output_dir_2d, exist_ok=True)

# Φόρτωση & Προεπεξεργασία Δεδομένων
earthquakes = pd.read_csv('/content/EarthquakesGr.csv')
tectonic_plates = pd.read_csv('/content/TectonicPlates.csv')

# Μετατροπή ώρας & καθαρισμός
earthquakes['Origin Time (GMT)'] = pd.to_datetime(earthquakes['Origin Time (GMT)'], errors='coerce')
earthquakes.dropna(subset=['Latitude', 'Longitude'], inplace=True)

# Επιλογή Χαρακτηριστικών για 2D (Latitude, Longitude)
features_2d_cols = ['Latitude', 'Longitude']
features_2d = earthquakes[features_2d_cols].copy()

# Έλεγχος του τύπου των δεδομένων
print("Πληροφορίες DataFrame για Earthquakes:")
earthquakes.info()
print("\n")

# Κανονικοποίηση
scaler = StandardScaler()
scaled_data = scaler.fit_transform(features_2d)

# Μέθοδος του Αγκώνα & Silhouette
inertia = []
silhouette_scores = []

range_clusters = range(2, 11)
for k in range_clusters:
    kmeans = KMeans(n_clusters=k, random_state=0, n_init=10)
    kmeans.fit(scaled_data)
    inertia.append(kmeans.inertia_)
    if k > 1:
        silhouette_scores.append(silhouette_score(scaled_data, kmeans.labels_))
    else:
        silhouette_scores.append(np.nan)

# Οπτικοποίηση μεθόδου του αγκώνα
plt.figure(figsize=(10, 5))
plt.plot(range_clusters, inertia, marker='o')
plt.title('Μέθοδος του Αγκώνα (2 Διαστάσεις - Κανονικοποιημένα Δεδομένα)')
plt.xlabel('Αριθμός Clusters')
plt.ylabel('Inertia')
plt.grid(True)
```

```
plt.savefig(f"{output_dir_2d}/elbow_method_2d.png")
plt.close()

# Οπτικοποίηση Silhouette Score
plt.figure(figsize=(10, 5))
plt.plot(range_clusters, silhouette_scores, marker='o')
plt.title('Silhouette Score (2 Διαστάσεις - Κανονικοποιημένα Δεδομένα)')
plt.xlabel('Αριθμός Clusters')
plt.ylabel('Silhouette Score')
plt.grid(True)
plt.savefig(f"{output_dir_2d}/silhouette_scores_2d.png")
plt.close()

# Επιλογή του βέλτιστου k
optimal_k = 4

# Εκτέλεση KMeans με το επιλεγμένο k
kmeans = KMeans(n_clusters=optimal_k, random_state=0, n_init=10)
earthquakes['Cluster'] = kmeans.fit_predict(scaled_data)

# Περιγραφική Στατιστική
print("\nΠεριγραφική στατιστική ανά cluster (2 Διαστάσεις - Αρχικά Δεδομένα):")
numeric_cols_for_agg = ['Latitude', 'Longitude', 'Depth (km)', 'Magnitude (ML)']
cluster_stats = earthquakes.groupby('Cluster')[numeric_cols_for_agg].agg(['mean', 'std', 'count', 'min',
'max'])
print(cluster_stats)

# Αποθήκευση περιγραφικής στατιστικής σε CSV
cluster_stats.to_csv(f"{output_dir_2d}/cluster_statistics_2d.csv")

plt.figure(figsize=(12, 8))
sns.scatterplot(data=earthquakes, x='Longitude', y='Latitude', hue='Cluster', palette='Set1', s=50,
alpha=0.7)

# Προσθήκη Στατιστικών στο Scatter Plot (Αρχικά Δεδομένα)
for cluster_id in earthquakes['Cluster'].unique():
    cluster_data = earthquakes[earthquakes['Cluster'] == cluster_id]
    mean_lat = cluster_stats.loc[cluster_id, ('Latitude', 'mean')]
    mean_lon = cluster_stats.loc[cluster_id, ('Longitude', 'mean')]
    min_lat = cluster_stats.loc[cluster_id, ('Latitude', 'min')]
    max_lat = cluster_stats.loc[cluster_id, ('Latitude', 'max')]
    std_lat = cluster_stats.loc[cluster_id, ('Latitude', 'std')]
    min_lon = cluster_stats.loc[cluster_id, ('Longitude', 'min')]
    max_lon = cluster_stats.loc[cluster_id, ('Longitude', 'max')]
    std_lon = cluster_stats.loc[cluster_id, ('Longitude', 'std')]

    text = (f"Cluster {cluster_id}\n"
            f"Lat: min={min_lat:.2f}, max={max_lat:.2f}, avg={mean_lat:.2f}, std={std_lat:.2f}\n"
            f"Lon: min={min_lon:.2f}, max={max_lon:.2f}, avg={mean_lon:.2f}, std={std_lon:.2f}")

    plt.annotate(text,
            xy=(mean_lon, mean_lat),
            xytext=(mean_lon + 0.5, mean_lat + 0.5),
            bbox=dict(facecolor='white', alpha=0.7, edgecolor='black'),
            fontsize=8,
            arrowprops=dict(arrowstyle="->", connectionstyle="arc3,rad=.2"))

plt.title('Συσταδοποίηση KMeans με Στατιστικά (Lat/Lon)')
plt.xlabel('Longitude')
```

```
plt.ylabel('Latitude')
plt.legend(title='Cluster')
plt.grid(True)
plt.tight_layout()
plt.savefig(f"{output_dir_2d}/2D_kmeans_clusters_with_detailed_stats.png")
plt.close()

# Scatter Plot των Clusters ΧΡΗΣΙΜΟΠΟΙΕΙ ΑΡΧΙΚΑ ΔΕΔΟΜΕΝΑ
plt.figure(figsize=(12, 6))
sns.scatterplot(data=earthquakes, x='Longitude', y='Latitude', hue='Cluster', palette='Set1', s=50,
alpha=0.7)
plt.title('Συσταδοποίηση KMeans (Longitude vs Latitude - Αρχικά Δεδομένα)')
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.legend(title='Cluster')
plt.grid(True)
plt.savefig(f"{output_dir_2d}/2D_kmeans_clusters_scatter.png")
plt.close()

# Αποθήκευση του DataFrame με τις ετικέτες των clusters σε CSV
earthquakes.to_csv(f"{output_dir_2d}/earthquakes_with_clusters_2d.csv", index=False)

# Δημιουργία Διαδραστικού χάρτη
map_clusters = folium.Map(location=[39.0742, 21.8243], zoom_start=6)
colors = ['red', 'blue', 'green', 'purple', 'orange']

# Κλήση Συνάρτησης τεκτονικών πλακών
add_tectonic_plates(map_clusters, tectonic_plates)

for cluster in range(optimal_k):
    cluster_data = earthquakes[earthquakes['Cluster'] == cluster]
    for idx, row in cluster_data.iterrows():
        folium.CircleMarker(
            location=[row['Latitude'], row['Longitude']],
            radius=5,
            color=colors[cluster % len(colors)],
            fill=True,
            fill_color=colors[cluster % len(colors)],
            fill_opacity=0.3,
            popup=folium.Popup(
                f"Cluster: {cluster}<br>"
                f"Magnitude: {row['Magnitude (ML)']}<br>"
                f"Depth: {row['Depth (km)']}<br>"
                f"Time: {row['Origin Time (GMT)']}",
                max_width=300
            )
        ).add_to(map_clusters)

# Προσθήκη Centroids στον Χάρτη  (απο-κανονικοποίηση για να ψανούν στο χάρτη)
centroids = kmeans.cluster_centers_

# Υπολογισμός μέσου όρου και τυπικής απόκλισης μία φορά για απο-κανονικοποίηση
lat_mean = earthquakes['Latitude'].mean()
lat_std = earthquakes['Latitude'].std()
lon_mean = earthquakes['Longitude'].mean()
lon_std = earthquakes['Longitude'].std()

for centroid in centroids:
    folium.Marker(
```

```
        location=[
            centroid[0] * lat_std + lat_mean,
            centroid[1] * lon_std + lon_mean
        ],
        icon=folium.Icon(color='lightgreen', icon='star'),
        popup=folium.Popup(f"Centroid: Latitude: {centroid[0]:.2f}, Longitude: {centroid[1]:.2f}",
        max_width=200)
    ).add_to(map_clusters)

base_maps = {
    "OpenStreetMap": folium.TileLayer("OpenStreetMap").add_to(map_clusters),
    "CartoDB Positron": folium.TileLayer("CartoDB Positron").add_to(map_clusters),
    "CartoDB Dark_Matter": folium.TileLayer("CartoDB Dark_Matter").add_to(map_clusters)
}

folium.LayerControl().add_to(map_clusters)

map_clusters.save(f"{output_dir_2d}/2D_kmeans_clusters_map.html")

print("\nΟλοκληρώθηκε η ανάλυση και οπτικοποίηση K-means σε 2 διαστάσεις.")
```

K-Means 5D

```
# Εισαγωγή Βιβλιοθηκών
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score
import os
import plotly.express as px
import plotly.graph_objects as go
import folium

# Ορισμός φακέλου εξόδου για 5D αποτελέσματα
output_dir_5d = "kmeans_outputs_5d"
os.makedirs(output_dir_5d, exist_ok=True)

# Φόρτωση & Προεπεξεργασία Δεδομένων
earthquakes = pd.read_csv('/content/EarthquakesGr.csv')

# Μετατροπή ώρας & καθαρισμός
earthquakes['Origin Time (GMT)'] = pd.to_datetime(earthquakes['Origin Time (GMT)'],
errors='coerce')
earthquakes.dropna(subset=['Latitude', 'Longitude', 'Depth (km)', 'Magnitude (ML)',
'Origin Time (GMT)'], inplace=True)

# Εξαγωγή του έτους ως 5η διάσταση
earthquakes['Year'] = earthquakes['Origin Time (GMT)'].dt.year

# Επιλογή Χαρακτηριστικών για 5D
features_5d_cols = ['Latitude', 'Longitude', 'Depth (km)', 'Magnitude (ML)', 'Year']
features_5d = earthquakes[features_5d_cols].copy()

# Έλεγχος του τύπου των δεδομένων
print("Πληροφορίες DataFrame για 5D Features:")
```

```
features_5d.info()
print("\n")

# Κανονικοποίηση
scaler_5d = StandardScaler()
scaled_data_5d = scaler_5d.fit_transform(features_5d)

scaled_df_5d = pd.DataFrame(scaled_data_5d, columns=features_5d_cols)

# Μέθοδος του Αγκώνα & Silhouette για 5D
inertia_5d = []
silhouette_scores_5d = []

range_clusters_5d = range(2, 11)

for k in range_clusters_5d:
    kmeans_5d = KMeans(n_clusters=k, random_state=0, n_init=10)
    kmeans_5d.fit(scaled_data_5d)
    inertia_5d.append(kmeans_5d.inertia_)

    if k > 1:
        silhouette_scores_5d.append(silhouette_score(scaled_data_5d, kmeans_5d.labels_))
    else:
        silhouette_scores_5d.append(np.nan)

# Οπτικοποίηση μεθόδου του Αγκώνα για 5D
plt.figure(figsize=(10, 5))
plt.plot(range_clusters_5d, inertia_5d, marker='o')
plt.title('Μέθοδος του Αγκώνα (5 Διαστάσεις - Κανονικοποιημένα Δεδομένα)')
plt.xlabel('Αριθμός Clusters')
plt.ylabel('Inertia')
plt.grid(True)
plt.savefig(f"{output_dir_5d}/elbow_method_5d.png")
plt.close()

# Οπτικοποίηση Silhouette Score για 5D
plt.figure(figsize=(10, 5))
plt.plot(range_clusters_5d, silhouette_scores_5d, marker='o')
plt.title('Silhouette Score (5 Διαστάσεις - Κανονικοποιημένα Δεδομένα)')
plt.xlabel('Αριθμός Clusters')
plt.ylabel('Silhouette Score')
plt.grid(True)
plt.savefig(f"{output_dir_5d}/silhouette_scores_5d.png")
plt.close()

# KMeans με το επιλεγμένο K (για 5D)
optimal_k_5d = 5
kmeans_5d = KMeans(n_clusters=optimal_k_5d, random_state=0, n_init=10)

scaled_df_5d['Cluster_5D'] = kmeans_5d.fit_predict(scaled_data_5d)

earthquakes['Cluster_5D'] = scaled_df_5d['Cluster_5D']

# Περιγραφική Στατιστική για 5D Clusters (ΑΡΧΙΚΑ Δεδομένα)
print("\nΠεριγραφική στατιστική ανά cluster (5 Διαστάσεις - Αρχικά Δεδομένα):")
numeric_cols_for_agg_5d = ['Latitude', 'Longitude', 'Depth (km)', 'Magnitude (ML)', 'Year']
cluster_stats_5d =
earthquakes.groupby('Cluster_5D')[numeric_cols_for_agg_5d].agg(['mean', 'std', 'count', 'min', 'max'])
print(cluster_stats_5d)
```

```
# Αποθήκευση περιγραφικής στατιστικής σε CSV
cluster_stats_5d.to_csv(f"{output_dir_5d}/cluster_statistics_5d.csv")

def add_stats_to_plot(stats_dict, fig, x_pos, y_start, line_height, ha, color='black', fontsize=9):

    y = y_start
    for label, value in stats_dict.items():
        formatted_value = f"{value:.2f}" if isinstance(value, (int, float, np.number)) else str(value)
        plt.figtext(x_pos, y, f"{label}: {formatted_value}", color=color, fontsize=fontsize,
        transform=fig.transFigure, ha=ha)

        if y_start > 0.5:
            y -= line_height
        else:
            y += line_height

# Parallel Coordinates Plot (Με Matplotlib - ΧΡΗΣΗ ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ)
def parallel_coordinates_plot_static(df, dimensions, color_col, title, filename, output_dir,
cluster_stats=None, original_earthquakes=None):
    fig = plt.figure(figsize=(12, 7))
    num_dimensions = len(dimensions)

    # Χρωματισμός ανά cluster
    unique_clusters = sorted(df[color_col].unique())
    colors = plt.colormaps['Set1']

    for i, cluster_id in enumerate(unique_clusters):
        cluster_data = df[df[color_col] == cluster_id]
        for _, row in cluster_data.iterrows():
            plt.plot(range(num_dimensions), row[dimensions], color=colors(i / len(unique_clusters)),
            alpha=0.3)

    for i, cluster_id in enumerate(unique_clusters):
        cluster_mean = df[df[color_col] == cluster_id][dimensions].mean()
        plt.plot(range(num_dimensions), cluster_mean, color=colors(i / len(unique_clusters)),
        linewidth=2, label=f'Cluster {cluster_id}')

    plt.xticks(range(num_dimensions), dimensions, rotation=45, ha='right')
    plt.title(title)
    plt.ylabel('Κανονικοποιημένη Τιμή')
    plt.grid(True, linestyle='--', alpha=0.7)
    plt.legend(title='Cluster', bbox_to_anchor=(1.05, 1), loc='upper left')
    plt.tight_layout()

    # Προσθήκη στατιστικών
    if cluster_stats is not None:
        stats_to_display = {}

        if isinstance(cluster_stats, pd.DataFrame):

            if original_earthquakes is not None:
                overall_means = original_earthquakes[dimensions].mean().to_dict()
                overall_stats_dict = {f'Overall Mean {dim}': val for dim, val in overall_means.items()}
                add_stats_to_plot(overall_stats_dict, fig, x_pos=0.99, y_start=0.01 +
                (len(overall_stats_dict) * 0.02) , line_height=-0.02, ha='right', fontsize=8)
        elif isinstance(cluster_stats, pd.Series):
            for dim in dimensions:
                if (dim, 'mean') in cluster_stats.index:
```

```
                    stats_to_display[f'Mean {dim}'] = cluster_stats[(dim, 'mean')]
                if (dim, 'std') in cluster_stats.index:
                    stats_to_display[f'Std {dim}'] = cluster_stats[(dim, 'std')]

            if (dimensions[0], 'count') in cluster_stats.index:
                stats_to_display['Count'] = cluster_stats[(dimensions[0], 'count')]

            add_stats_to_plot(stats_to_display, fig, x_pos=0.99, y_start=0.01 +
            (len(stats_to_display) * 0.02), line_height=-0.02, ha='right', fontsize=8)

    plt.savefig(f"{output_dir}/{filename}.png")
    plt.close()

parallel_coordinates_plot_static(
    scaled_df_5d,
    dimensions=['Latitude', 'Longitude', 'Depth (km)', 'Magnitude (ML)', 'Year'],
    color_col="Cluster_5D",
    title="Παράλληλες Συντεταγμένες Όλων των Clusters (5 Διαστάσεις - Κανονικοποιημένα)",
    filename="parallel_coordinates_all_clusters_5d_scaled_static",
    output_dir=output_dir_5d,
    cluster_stats=cluster_stats_5d,
    original_earthquakes=earthquakes
)

# Μεμονωμένα Parallel Coordinates Plots για κάθε Cluster (με ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΑ ΔΕΔΟΜΕΝΑ)
for cluster_id in range(optimal_k_5d):
    cluster_data_scaled = scaled_df_5d[scaled_df_5d['Cluster_5D'] == cluster_id]

    current_cluster_stats = cluster_stats_5d.loc[cluster_id]

    if not cluster_data_scaled.empty:
        parallel_coordinates_plot_static(
            cluster_data_scaled,
            dimensions=['Latitude', 'Longitude', 'Depth (km)', 'Magnitude (ML)', 'Year'],
            color_col="Cluster_5D",
            title=f"Παράλληλες Συντεταγμένες για Cluster {cluster_id} (5 Διαστάσεις - Κανονικοποιημένα)",
            filename=f"parallel_coordinates_cluster_{cluster_id}_5d_scaled_static",
            output_dir=output_dir_5d,
            cluster_stats=current_cluster_stats
        )

# Διαδραστικό 3D Scatter Plot (Plotly) με Animation για την 5η Διάσταση (Έτος)
plot_df_5d_animated = scaled_df_5d.copy()
plot_df_5d_animated['Year_Original'] = earthquakes['Year']
plot_df_5d_animated['Cluster_5D'] = earthquakes['Cluster_5D']

fig_3d_5d_animated = px.scatter_3d(
    plot_df_5d_animated,
    x='Longitude',
    y='Latitude',
    z='Depth (km)',
    color='Magnitude (ML)',
    animation_frame='Year_Original',
    color_continuous_scale=px.colors.sequential.Viridis,
    title='Συσταδοποίηση KMeans (5 Διαστάσεις με Χρονική Εξέλιξη)',
    hover_name="Cluster_5D",
    hover_data={
        'Longitude': ':.2f',
        'Latitude': ':.2f',
```

```
        'Depth (km)': ':.2f',
        'Magnitude (ML)': ':.2f',
        'Year_Original': True,
        'Cluster_5D': False
    },

    range_z=[plot_df_5d_animated['Depth (km)'].min(), plot_df_5d_animated['Depth (km)'].max()],
    range_x=[plot_df_5d_animated['Longitude'].min(), plot_df_5d_animated['Longitude'].max()],
    range_y=[plot_df_5d_animated['Latitude'].min(), plot_df_5d_animated['Latitude'].max()],

)

fig_3d_5d_animated.update_layout(
    scene = dict(
        xaxis_title="Longitude (Κανονικοποιημένο)",
        yaxis_title="Latitude (Κανονικοποιημένο)",
        zaxis_title="Depth (km) (Κανονικοποιημένο)"
    )
)

fig_3d_5d_animated.write_html(f"{output_dir_5d}/plotly_scatter_3d_5d_animated_scaled.html")

# Αποθήκευση του DataFrame με τις ετικέτες των clusters σε CSV (αρχικά δεδομένα + Cluster_5D)
earthquakes.to_csv(f"{output_dir_5d}/earthquakes_with_clusters_5d.csv", index=False)

print("\nΟλοκληρώθηκε η ανάλυση και οπτικοποίηση K-means σε 5 διαστάσεις.")
```

DBSCAN 2D

```
# Εισαγωγή Βιβλιοθηκών
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import folium
from sklearn.cluster import DBSCAN
from sklearn.neighbors import NearestNeighbors
from sklearn.metrics import davies_bouldin_score, calinski_harabasz_score
import os
from math import radians
import colorsys

# Δημιουργία φακέλου για αποθήκευση γραφημάτων και δεδομένων DBSCAN
output_dir_dbscan_2d = "dbscan_outputs_2d_FINAL"
os.makedirs(output_dir_dbscan_2d, exist_ok=True)

# Συνάρτηση για τις τεκτονικές πλάκες

# Συνάρτηση για κείμενο (στατιστικά στο plot)
def add_stats_to_plot(fig, x_pos, y_pos, ha, text_content, color='black', fontsize=9):

    plt.figtext(x_pos, y_pos, text_content, color=color, fontsize=fontsize,
                transform=fig.transFigure, ha=ha, va='top', bbox=dict(facecolor='white', alpha=0.7,
                edgecolor='black', boxstyle='round,pad=0.3'))

# Φόρτωση & Προεπεργασία Δεδομένων
earthquakes = pd.read_csv('/content/EarthquakesGr.csv')
tectonic_plates = pd.read_csv('/content/TectonicPlates.csv')
```

```
# Μετατροπή ώρας & καθαρισμός
earthquakes['Origin Time (GMT)'] = pd.to_datetime(earthquakes['Origin Time (GMT)'],
errors='coerce')
earthquakes.dropna(subset=['Latitude', 'Longitude'], inplace=True)

earthquakes.set_index('Origin Time (GMT)', inplace=True)

# Επιλογή Χαρακτηριστικών για 2D (Latitude, Longitude)
features_2d_cols = ['Latitude', 'Longitude']
features_2d_dbscan = earthquakes[features_2d_cols].copy()

earthquakes.info()
print("\n")

# ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΓΙΑ DBSCAN (ΜΕΤΑΤΡΟΠΗ ΣΕ ΑΚΤΙΝΙΑ)
features_2d_dbscan_rad = features_2d_dbscan.map(radians)

print(features_2d_dbscan_rad.head())
print("\n")

# ΚΩΔΙΚΑΣ ΓΙΑ ΕΥΡΕΣΗ k-ΕΓΓΥΤΕΡΩΝ ΓΕΙΤΟΝΩΝ
for k in range(4, 14):
    neighb = NearestNeighbors(n_neighbors=k, metric='haversine')
    nbrs = neighb.fit(features_2d_dbscan_rad)
    distances, indices = nbrs.kneighbors(features_2d_dbscan_rad)

    distances_for_neighbor = np.sort(distances[:, k-1], axis=0)

    plt.figure(figsize=(12, 6))
    plt.plot(distances_for_neighbor)
    plt.title(f"k = {k}, {k}-ος πλησιέστερος γείτονας (Haversine)")
    plt.ylim(0, 0.05)
    plt.xlabel('Σημεία ταξινομημένα βάσει απόστασης')
    plt.ylabel('Epsilon (σε ακτίνια)')
    plt.yticks(np.arange(0, 0.051, 0.005))
    plt.grid(True)
    plt.tight_layout()
    plt.savefig(f"{output_dir_dbscan_2d}/k_dist_plot_k_{k}.png")
    plt.close()

# ΕΞΕΡΕΥΝΗΣΗ ΠΑΡΑΜΕΤΡΩΝ DBSCAN ΜΕ ΔΕΙΚΤΕΣ ΑΞΙΟΛΟΓΗΣΗΣ ---
eps_values = np.linspace(0.003, 0.006, 20)
min_samples_values = np.arange(10, 14, 1)

db_scores_eps = []
ch_scores_eps = []
noise_ratios_eps = []
num_clusters_eps = []
best_eps_db = None
best_db_score = float('inf')
best_eps_ch = None
best_ch_score = -float('inf')

fixed_min_samples = 13

for eps in eps_values:
    dbscan = DBSCAN(eps=eps, min_samples=fixed_min_samples, metric='haversine')
    clusters = dbscan.fit_predict(features_2d_dbscan_rad)
```

```
    core_samples_mask = (clusters != -1)
    n_clusters = len(set(clusters[core_samples_mask]))
    noise_ratio = (clusters == -1).sum() / len(clusters)

    db_score = np.nan
    ch_score = np.nan

    if n_clusters >= 2 and len(clusters[core_samples_mask]) >= 2:
        try:
            db_score = davies_bouldin_score(features_2d_dbscan_rad[core_samples_mask],
            clusters[core_samples_mask])
            ch_score = calinski_harabasz_score(features_2d_dbscan_rad[core_samples_mask],
            clusters[core_samples_mask])
        except Exception as e:
            pass

    db_scores_eps.append(db_score)
    ch_scores_eps.append(ch_score)
    noise_ratios_eps.append(noise_ratio)
    num_clusters_eps.append(n_clusters)

    if not np.isnan(db_score) and db_score < best_db_score:
        best_db_score = db_score
        best_eps_db = eps
    if not np.isnan(ch_score) and ch_score > best_ch_score:
        best_ch_score = ch_score
        best_eps_ch = eps

plt.figure(figsize=(15, 6))

plt.subplot(1, 3, 1)
plt.plot(eps_values, db_scores_eps, marker='o', linestyle='-')
plt.title(f'Davies-Bouldin Index vs. Epsilon (MinPts={fixed_min_samples})')
plt.xlabel('Epsilon (σε ακτίνια)')
plt.ylabel('Davies-Bouldin Index')
plt.grid(True)

plt.subplot(1, 3, 2)
plt.plot(eps_values, ch_scores_eps, marker='o', linestyle='-')
plt.title(f'Calinski-Harabasz Index vs. Epsilon (MinPts={fixed_min_samples})')
plt.xlabel('Epsilon (σε ακτίνια)')
plt.ylabel('Calinski-Harabasz Index')
plt.grid(True)

plt.subplot(1, 3, 3)
plt.plot(eps_values, num_clusters_eps, marker='o', linestyle='-',
label='Αριθμός Clusters')
plt.plot(eps_values, noise_ratios_eps, marker='x', linestyle='--',
color='red', label='Ποσοστό Θορύβου')
plt.title(f'Αριθμός Clusters & Θόρυβος vs. Epsilon (MinPts={fixed_min_samples})')
plt.xlabel('Epsilon (σε ακτίνια)')
plt.ylabel('Τιμή')
plt.grid(True)
plt.legend()

plt.tight_layout()
plt.savefig(f"{output_dir_dbscan_2d}/eps_vs_metrics_minpts_{fixed_min_samples}.png")
plt.close()
```

```
print(f"Βέλτιστο eps για Davies-Bouldin Index:
{best_eps_db:.4f} (DBI: {best_db_score:.3f})")
print(f"Βέλτιστο eps για Calinski-Harabasz Index:
{best_eps_ch:.4f} (CH: {best_ch_score:.3f})")

fixed_eps = 0.0055

db_scores_minpts = []
ch_scores_minpts = []
noise_ratios_minpts = []
num_clusters_minpts = []
best_min_samples_db = None
best_db_score_minpts = float('inf')
best_min_samples_ch = None
best_ch_score_minpts = -float('inf')

for min_samples in min_samples_values:
    dbscan = DBSCAN(eps=fixed_eps, min_samples=min_samples, metric='haversine')
    clusters = dbscan.fit_predict(features_2d_dbscan_rad)

    core_samples_mask = (clusters != -1)
    n_clusters = len(set(clusters[core_samples_mask]))
    noise_ratio = (clusters == -1).sum() / len(clusters)

    db_score = np.nan
    ch_score = np.nan

    if n_clusters >= 2 and len(clusters[core_samples_mask]) >= 2:
        try:
            db_score = davies_bouldin_score(features_2d_dbscan_rad[core_samples_mask],
            clusters[core_samples_mask])
            ch_score = calinski_harabasz_score(features_2d_dbscan_rad[core_samples_mask],
            clusters[core_samples_mask])
        except Exception as e:
            pass

    db_scores_minpts.append(db_score)
    ch_scores_minpts.append(ch_score)
    noise_ratios_minpts.append(noise_ratio)
    num_clusters_minpts.append(n_clusters)

    if not np.isnan(db_score) and db_score < best_db_score_minpts:
        best_db_score_minpts = db_score
        best_min_samples_db = min_samples
    if not np.isnan(ch_score) and ch_score > best_ch_score_minpts:
        best_ch_score_minpts = ch_score
        best_min_samples_ch = min_samples

plt.figure(figsize=(15, 6))

plt.subplot(1, 3, 1)
plt.plot(min_samples_values, db_scores_minpts, marker='o', linestyle='-')
plt.title(f'Davies-Bouldin Index vs. MinPts (Epsilon={fixed_eps:.4f})')
plt.xlabel('MinPts')
plt.ylabel('Davies-Bouldin Index')
plt.grid(True)

plt.subplot(1, 3, 2)
```

```
plt.plot(min_samples_values, ch_scores_minpts, marker='o', linestyle='-')
plt.title(f'Calinski-Harabasz Index vs. MinPts (Epsilon={fixed_eps:.4f})')
plt.xlabel('MinPts')
plt.ylabel('Calinski-Harabasz Index')
plt.grid(True)

plt.subplot(1, 3, 3)
plt.plot(min_samples_values, num_clusters_minpts, marker='o', linestyle='-',
label='Αριθμός Clusters')
plt.plot(min_samples_values, noise_ratios_minpts, marker='x', linestyle='--',
color='red', label='Ποσοστό Θορύβου')
plt.title(f'Αριθμός Clusters & Θόρυβος vs. MinPts (Epsilon={fixed_eps:.4f})')
plt.xlabel('MinPts')
plt.ylabel('Τιμή')
plt.grid(True)
plt.legend()

plt.tight_layout()
plt.savefig(f"{output_dir_dbscan_2d}/minpts_vs_metrics_eps_{fixed_eps:.4f}.png")

print(f"Βέλτιστο MinPts για Davies-Bouldin Index:
{best_min_samples_db} (DBI: {best_db_score_minpts:.3f})")
print(f"Βέλτιστο MinPts για Calinski-Harabasz Index:
{best_min_samples_ch} (CH: {best_ch_score_minpts:.3f})")

# ΕΠΙΛΟΓΗ ΒΕΛΤΙΣΤΩΝ ΠΑΡΑΜΕΤΡΩΝ DBSCAN
optimal_eps = 0.0055
optimal_min_samples = 13

print(f"   optimal_eps: {optimal_eps} (σε ακτίνια)")
print(f"   optimal_min_samples: {optimal_min_samples}")

# --- ΕΦΑΡΜΟΓΗ ΤΟΥ DBSCAN ---
dbscan = DBSCAN(eps=optimal_eps, min_samples=optimal_min_samples, metric='haversine')
earthquakes['Cluster'] = dbscan.fit_predict(features_2d_dbscan_rad)

# --- ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ ---
core_samples_mask = (earthquakes['Cluster'] != -1)
unique_clusters = sorted(set(earthquakes['Cluster'][core_samples_mask]))
n_clusters_ = len(unique_clusters)

if n_clusters_ >= 2:
    db_score = davies_bouldin_score(features_2d_dbscan_rad[core_samples_mask],
    earthquakes['Cluster'][core_samples_mask])
    ch_score = calinski_harabasz_score(features_2d_dbscan_rad[core_samples_mask],
    earthquakes['Cluster'][core_samples_mask])

    print(f"   Davies-Bouldin Index: {db_score:.3f} (Χαμηλότερο είναι καλύτερο)")
    print(f"   Calinski-Harabasz Index: {ch_score:.3f} (Υψηλότερο είναι καλύτερο)")
else:
    print("\nΔεν υπάρχουν αρκετές συστάδες (>=2) για τον υπολογισμό του
    Davies-Bouldin Index και του Calinski-Harabasz Index.")

print(f"\nΑριθμός Βρεθέντων Clusters: {n_clusters_}")
print(f"Αριθμός Σημείων Θορύβου: {(earthquakes['Cluster'] == -1).sum()}")
print(f"Ποσοστό Σημείων Θορύβου:
{((earthquakes['Cluster'] == -1).sum() / len(earthquakes)) * 100:.2f}%")

# --- ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑ CLUSTER ---
```

```
print("\nΠεριγραφική στατιστική ανά cluster (2 Διαστάσεις - Αρχικά Δεδομένα):")
cluster_stats_df = earthquakes[earthquakes['Cluster'] != -1].groupby('Cluster')
[['Latitude', 'Longitude', 'Depth (km)', 'Magnitude (ML)']].agg(['mean', 'std',
'count', 'min', 'max'])
print(cluster_stats_df)
cluster_stats_df.to_csv(f"{output_dir_dbscan_2d}/cluster_statistics_dbscan_2d.csv")

# --- CUSTOM ΧΡΩΜΑΤΙΚΗ ΠΑΛΕΤΑ ΓΙΑ ΕΝΤΟΝΗ ΔΙΑΚΡΙΣΗ ---
distinct_colors = [
    '#000080', '#800080', '#0B3030', '#8B0000', '#3CB371',
    '#E6194B', '#6C8D91', '#BDC736', '#6FB4D1', '#B22222',
    '#4B0082', '#1B385C', '#9932CC', '#50C878', '#135C5C',
    '#006400', '#495682', '#F58231', '#450141', '#36454F',
    '#C04000'

]

def get_extended_distinct_colors(num_colors, base_colors):
    if num_colors <= len(base_colors):
        return base_colors[:num_colors]

    extended_colors = list(base_colors)
    rgb_colors = [tuple(int(h.lstrip('#')[i:i+2], 16) / 255
    for i in (0, 2, 4)) for h in base_colors]
    hsl_colors = [colorsys.rgb_to_hls(*rgb) for rgb in rgb_colors]

    current_idx = 0
    while len(extended_colors) < num_colors:
        h, l, s = hsl_colors[current_idx % len(hsl_colors)]
        new_l = max(0.1, min(0.9, l + (0.1 if current_idx % 2 == 0 else -0.1)))
        new_rgb = colorsys.hls_to_rgb(h, new_l, s)
        new_hex = '#%02x%02x%02x' % tuple(int(x * 255) for x in new_rgb)

        if new_hex not in extended_colors:
            extended_colors.append(new_hex)
        current_idx += 1
        if current_idx > num_colors * 5 and len(extended_colors) < num_colors:
            print("Προσοχή: Δεν μπόρεσα να δημιουργήσω αρκετά μοναδικά χρώματα.")
            break
    return extended_colors[:num_colors]

final_palette_colors = get_extended_distinct_colors(n_clusters_ if n_clusters_ > 0
else 1, distinct_colors)

cluster_id_to_map_color = {}
if n_clusters_ > 0:
    for i, cluster_id in enumerate(unique_clusters):
        cluster_id_to_map_color[cluster_id] =
        final_palette_colors[i % len(final_palette_colors)]

# Scatter Plot για τα Clusters χωρίς τους Outliers (Θόρυβο)
fig, ax = plt.subplots(figsize=(15, 8))
clusters_only_df = earthquakes[earthquakes['Cluster'] != -1].copy()

if not clusters_only_df.empty:
    sns.scatterplot(data=clusters_only_df, x='Longitude', y='Latitude',
                    hue='Cluster', palette=final_palette_colors,
                    s=50, alpha=0.7, ax=ax)
```

```
    x_positions = [1.02, 1.15, 1.28]
    y_start_top = 0.95

    current_x_idx = 0
    current_y = y_start_top

    cluster_counts = clusters_only_df['Cluster'].value_counts()
    sorted_unique_clusters = cluster_counts.index.tolist()

    for cluster_id in sorted_unique_clusters:
        mean_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'mean')]
        mean_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'mean')]
        min_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'min')]
        max_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'max')]
        std_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'std')]
        min_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'min')]
        max_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'max')]
        std_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'std')]
        count = cluster_stats_df.loc[cluster_id, ('Latitude', 'count')]

        text_block = (f"Cluster {cluster_id} (N={int(count)})\n"
                      f"   Lat: min={min_lat:.2f}, max={max_lat:.2f}\n"
                      f"        avg={mean_lat:.2f}, std={std_lat:.2f}\n"
                      f"   Lon: min={min_lon:.2f}, max={max_lon:.2f}\n"
                      f"        avg={mean_lon:.2f}, std={std_lon:.2f}")

        num_lines_in_block = text_block.count('\n') + 1
        estimated_line_height_fig = 0.02
        block_height = num_lines_in_block * estimated_line_height_fig + 0.005

        if (current_y - block_height) < 0.05 and (current_x_idx + 1) < len(x_positions):
            current_x_idx += 1
            current_y = y_start_top
        elif (current_y - block_height) < 0.05 and (current_x_idx + 1) >= len(x_positions):
            print(f"Προειδοποίηση: Δεν υπάρχει αρκετός χώρος για όλα τα clusters.
            Σταματάει στο Cluster {cluster_id}.")
            break

        add_stats_to_plot(fig, x_positions[current_x_idx], current_y, ha='left',
        text_content=text_block, fontsize=7)

        current_y -= block_height

else:
    ax.scatter(x=earthquakes['Longitude'], y='Latitude',
               c='gray', s=50, alpha=0.7, label='No Clusters Found')
    ax.legend()

ax.set_title('Συσταδοποίηση DBSCAN (Clusters μόνο, χωρίς Θόρυβο)')
ax.set_xlabel('Longitude')
ax.set_ylabel('Latitude')
ax.legend(title='Cluster', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.grid(True)
plt.tight_layout()
plt.savefig(f"{output_dir_dbscan_2d}/2D_dbscan_clusters_only.png",
bbox_inches='tight')
plt.close(fig)

# ΧΡΟΝΟΣΕΙΡΕΣ ΜΕΓΕΘΟΥΣ ΣΕΙΣΜΩΝ ΑΝΑ CLUSTER (ΕΤΗΣΙΑ ΒΑΣΗ)
```

```
#selected_clusters_for_timeseries = [c_id for c_id in unique_clusters if c_id != -1]
selected_clusters_for_timeseries = [1, 2]

for cluster_id in selected_clusters_for_timeseries:
    cluster_data = earthquakes[earthquakes['Cluster'] == cluster_id]

    if cluster_data.empty:
        print(f"Δεν βρέθηκαν δεδομένα για Cluster {cluster_id}. Παράλειψη.")
        continue

    # Χρονοσειρά: Μέσος όρος μεγέθους σεισμών ανά έτος
    timeseries_magnitude = cluster_data['Magnitude (ML)'].resample('YE').mean().fillna(0)

    # Υπολογισμός στατιστικών για εκτύπωση στο γράφημα
    mean_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'mean')]
    mean_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'mean')]
    min_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'min')]
    max_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'max')]
    std_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'std')]
    min_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'min')]
    max_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'max')]
    std_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'std')]
    count = cluster_stats_df.loc[cluster_id, ('Latitude', 'count')]

    stats_text_block = (f"Cluster {cluster_id} (N={int(count)})\n"
                        f"   Lat: min={min_lat:.2f}, max={max_lat:.2f}\n"
                        f"        avg={mean_lat:.2f}, std={std_lat:.2f}\n"
                        f"   Lon: min={min_lon:.2f}, max={max_lon:.2f}\n"
                        f"        avg={mean_lon:.2f}, std={std_lon:.2f}")

    fig_ts_mag, ax_ts_mag = plt.subplots(figsize=(14, 7))

    ax_ts_mag.plot(timeseries_magnitude.index.year, timeseries_magnitude.values,
                   marker='o', linestyle='-', markersize=4,
                   color=final_palette_colors[cluster_id])

    ax_ts_mag.set_title(f'Ετήσια Χρονοσειρά Μέσου Όρου Μεγέθους για Cluster {cluster_id}')
    ax_ts_mag.set_xlabel('Έτος')
    ax_ts_mag.set_ylabel('Μέσος Όρος Μεγέθους Σεισμών')
    ax_ts_mag.grid(True)

    add_stats_to_plot(fig_ts_mag, 0.75, 0.95, ha='right', text_content=stats_text_block,
    fontsize=7)

    plt.tight_layout()
    plt.savefig(f"{output_dir_dbscan_2d}/timeseries_magnitude_cluster_{cluster_id}.png",
    bbox_inches='tight')
    plt.close(fig_ts_mag)

# selected_clusters_for_all_magnitudes = [c_id for c_id in unique_clusters if c_id != -1]
selected_clusters_for_all_magnitudes = [1, 2]

# Βρόχος για κάθε επιλεγμένο Cluster
for cluster_id in selected_clusters_for_all_magnitudes:
    cluster_data = earthquakes[earthquakes['Cluster'] == cluster_id]

    if cluster_data.empty:
        print(f"Δεν βρέθηκαν δεδομένα για Cluster {cluster_id}. Παράλειψη.")
```

```
        continue

    fig_all_mag, ax_all_mag = plt.subplots(figsize=(14, 7))

    ax_all_mag.plot(cluster_data.index, cluster_data['Magnitude (ML)'],
      linestyle='-',
      linewidth=1,
      marker='o',
      alpha=0.6,
      color=final_palette_colors[cluster_id])

    mean_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'mean')]
    mean_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'mean')]
    count = cluster_stats_df.loc[cluster_id, ('Latitude', 'count')]
    min_mag = cluster_stats_df.loc[cluster_id, ('Magnitude (ML)', 'min')]
    max_mag = cluster_stats_df.loc[cluster_id, ('Magnitude (ML)', 'max')]
    avg_mag = cluster_stats_df.loc[cluster_id, ('Magnitude (ML)', 'mean')]

    stats_text_block = (f"Cluster {cluster_id} (N={int(count)})\n"
                        f"   Avg Mag: {avg_mag:.2f}\n"
                        f"   Min Mag: {min_mag:.2f}\n"
                        f"   Max Mag: {max_mag:.2f}\n"
                        f"   Avg Lat: {mean_lat:.2f}\n"
                        f"   Avg Lon: {mean_lon:.2f}")

    ax_all_mag.set_title(f'Χρονική Εξέλιξη του Μεγέθους Σεισμών για Cluster {cluster_id}')
    ax_all_mag.set_xlabel('Χρόνος (Έτος)')
    ax_all_mag.set_ylabel('Μέγεθος Σεισμών (ML)')
    ax_all_mag.grid(True)
    ax_all_mag.tick_params(axis='x', rotation=45)

    # Προσθήκη Στατιστικών στο Γράφημα
    add_stats_to_plot(fig_all_mag, 0.75, 0.95, ha='right',
    text_content=stats_text_block, fontsize=7)

    plt.tight_layout()
    plt.savefig(f"{output_dir_dbscan_2d}/timeseries_all_magnitudes_cluster_{cluster_id}.png",
    bbox_inches='tight')
    plt.close(fig_all_mag)

# ΧΡΟΝΟΣΕΙΡΕΣ ΜΕΓΕΘΟΥΣ ΣΕΙΣΜΩΝ ΑΝΑ CLUSTER (ΕΤΗΣΙΑ ΒΑΣΗ)
#selected_clusters_for_timeseries = [c_id for c_id in unique_clusters if c_id != -1]
selected_clusters_for_timeseries = [1, 2]

for cluster_id in selected_clusters_for_timeseries:
    print(f"Δημιουργία χρονοσειράς μέγεθους για Cluster {cluster_id} (Ετήσια)...")

    cluster_data = earthquakes[earthquakes['Cluster'] == cluster_id]

    if cluster_data.empty:
        print(f"Δεν βρέθηκαν δεδομένα για Cluster {cluster_id}. Παράλειψη.")
        continue

    # Χρονοσειρά: Μέσος όρος μεγέθους σεισμών ανά έτος
    timeseries_magnitude = cluster_data['Magnitude (ML)'].resample('YE').mean().fillna(0)

    # Υπολογισμός στατιστικών για εκτύπωση στο γράφημα
    mean_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'mean')]
    mean_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'mean')]
```

```
    min_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'min')]
    max_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'max')]
    std_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'std')]
    min_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'min')]
    max_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'max')]
    std_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'std')]
    count = cluster_stats_df.loc[cluster_id, ('Latitude', 'count')]

    stats_text_block = (f"Cluster {cluster_id} (N={int(count)})\n"
                        f"   Lat: min={min_lat:.2f}, max={max_lat:.2f}\n"
                        f"        avg={mean_lat:.2f}, std={std_lat:.2f}\n"
                        f"   Lon: min={min_lon:.2f}, max={max_lon:.2f}\n"
                        f"        avg={mean_lon:.2f}, std={std_lon:.2f}")

    fig_ts_mag, ax_ts_mag = plt.subplots(figsize=(14, 7))

    ax_ts_mag.plot(timeseries_magnitude.index.year, timeseries_magnitude.values,
                   marker='o', linestyle='-', markersize=4,
                   color=final_palette_colors[cluster_id])

    ax_ts_mag.set_title(f'Ετήσια Χρονοσειρά Μέσου Όρου Μεγέθους για Cluster {cluster_id}')
    ax_ts_mag.set_xlabel('Έτος')
    ax_ts_mag.set_ylabel('Μέσος Όρος Μεγέθους Σεισμών')
    ax_ts_mag.grid(True)

    add_stats_to_plot(fig_ts_mag, 0.75, 0.95, ha='right',
    text_content=stats_text_block, fontsize=7)

    plt.tight_layout()
    plt.savefig(f"{output_dir_dbscan_2d}/timeseries_magnitude_cluster_{cluster_id}.png",
    bbox_inches='tight')
    plt.close(fig_ts_mag)

# selected_clusters_for_all_magnitudes = [c_id for c_id in unique_clusters if c_id != -1]
selected_clusters_for_all_magnitudes = [1, 2]

# Βρόχος για κάθε επιλεγμένο Cluster
for cluster_id in selected_clusters_for_all_magnitudes:
    print(f"Δημιουργία χρονοσειράς μεγεθών για Cluster {cluster_id}...")

    cluster_data = earthquakes[earthquakes['Cluster'] == cluster_id]

    if cluster_data.empty:
        print(f"Δεν βρέθηκαν δεδομένα για Cluster {cluster_id}. Παράλειψη.")
        continue

    fig_all_mag, ax_all_mag = plt.subplots(figsize=(14, 7))

    ax_all_mag.plot(cluster_data.index, cluster_data['Magnitude (ML)'],
      linestyle='-',
      linewidth=1,
      marker='o',
      alpha=0.6,
      color=final_palette_colors[cluster_id])

    mean_lat = cluster_stats_df.loc[cluster_id, ('Latitude', 'mean')]
    mean_lon = cluster_stats_df.loc[cluster_id, ('Longitude', 'mean')]
    count = cluster_stats_df.loc[cluster_id, ('Latitude', 'count')]
    min_mag = cluster_stats_df.loc[cluster_id, ('Magnitude (ML)', 'min')]
```

```
    max_mag = cluster_stats_df.loc[cluster_id, ('Magnitude (ML)', 'max')]
    avg_mag = cluster_stats_df.loc[cluster_id, ('Magnitude (ML)', 'mean')]

    stats_text_block = (f"Cluster {cluster_id} (N={int(count)})\n"
                        f"   Avg Mag: {avg_mag:.2f}\n"
                        f"   Min Mag: {min_mag:.2f}\n"
                        f"   Max Mag: {max_mag:.2f}\n"
                        f"   Avg Lat: {mean_lat:.2f}\n"
                        f"   Avg Lon: {mean_lon:.2f}")

    ax_all_mag.set_title(f'Χρονική Εξέλιξη του Μεγέθους Σεισμών
    για Cluster {cluster_id}')
    ax_all_mag.set_xlabel('Χρόνος (Έτος)')
    ax_all_mag.set_ylabel('Μέγεθος Σεισμών (ML)')
    ax_all_mag.grid(True)
    ax_all_mag.tick_params(axis='x', rotation=45)

    # Προσθήκη Στατιστικών στο Γράφημα
    add_stats_to_plot(fig_all_mag, 0.75, 0.95, ha='right',
    text_content=stats_text_block, fontsize=7)

    plt.tight_layout()
    plt.savefig(f"{output_dir_dbscan_2d}/timeseries_all_magnitudes_cluster_{cluster_id}.png",
    bbox_inches='tight')
    plt.close(fig_all_mag)

earthquakes_to_save = earthquakes.reset_index()
earthquakes_to_save.to_csv(f"{output_dir_dbscan_2d}/earthquakes_with_clusters_dbscan_2d.csv",
index=False)
print("\nΟλοκληρώθηκε η ανάλυση και οπτικοποίηση DBSCAN σε 2 διαστάσεις.")
```

DBSCAN 5D

```
# Εισαγωγή Βιβλιοθηκών --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# ΟΡΙΣΜΟΣ ΦΑΚΕΛΟΥ ΕΞΟΔΟΥ ΓΙΑ 5D ΑΠΟΤΕΛΕΣΜΑΤΑ DBSCAN --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# ΕΝΗΜΕΡΩΜΕΝΗ ΣΥΝΑΡΤΗΣΗ add_stats_to_plot
def add_stats_to_plot(fig, x_pos, y_pos, ha, va, text_content, color='black', fontsize=9):
    plt.figtext(x_pos, y_pos, text_content, color=color, fontsize=fontsize,
                transform=fig.transFigure, ha=ha, va=va)

# Φόρτωση & Προεπεξεργασία Δεδομένων --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# Μετατροπή ώρας & καθαρισμός
earthquakes['Origin Time (GMT)'] =
pd.to_datetime(earthquakes['Origin Time (GMT)'], errors='coerce')

earthquakes.dropna(subset=['Latitude', 'Longitude', 'Depth (km)',
'Magnitude (ML)', 'Origin Time (GMT)'], inplace=True)

earthquakes.set_index('Origin Time (GMT)', inplace=True)

# Εξαγωγή του έτους ως 5η διάσταση
earthquakes['Year'] = earthquakes.index.year

# Επιλογή Χαρακτηριστικών για 5D
features_5d_cols = ['Latitude', 'Longitude', 'Depth (km)',
'Magnitude (ML)', 'Year']
features_5d_dbscan = earthquakes[features_5d_cols].copy()
```

```
# Έλεγχος του τύπου των δεδομένων
features_5d_dbscan.info()
print("\n")

# Κανονικοποίηση για DBSCAN 5D (με standard scaler)
scaler_5d = StandardScaler()
scaled_data_5d_dbscan = scaler_5d.fit_transform(features_5d_dbscan)

# ΔΗΜΙΟΥΡΓΙΑ DATAFRAME ΜΕ ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΑ ΔΕΔΟΜΕΝΑ
scaled_df_5d_dbscan = pd.DataFrame(scaled_data_5d_dbscan,
columns=features_5d_cols, index=features_5d_dbscan.index)

print(scaled_df_5d_dbscan.head())
print("\n")

# ΕΞΕΡΕΥΝΗΣΗ ΠΑΡΑΜΕΤΡΩΝ DBSCAN ΜΕ ΔΕΙΚΤΕΣ ΑΞΙΟΛΟΓΗΣΗΣ (5D) --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# ΕΠΙΛΟΓΗ ΒΕΛΤΙΣΤΩΝ ΠΑΡΑΜΕΤΡΩΝ DBSCAN
optimal_eps_5d = 0.8980
optimal_min_samples_5d = 9

print(f"   optimal_eps_5d: {optimal_eps_5d} (σε κανονικοποιημένη απόσταση)")
print(f"   optimal_min_samples_5d: {optimal_min_samples_5d}")

dbscan_5d = DBSCAN(eps=optimal_eps_5d, min_samples=optimal_min_samples_5d, metric='euclidean')
# Προσθήκη των ετικετών των clusters στο αρχικό DataFrame
earthquakes['Cluster_5D_DBSCAN'] = dbscan_5d.fit_predict(scaled_data_5d_dbscan)
# Προσθήκη των ετικετών των clusters και στο scaled_df_5d_dbscan (για plots)
scaled_df_5d_dbscan['Cluster_5D_DBSCAN'] = earthquakes['Cluster_5D_DBSCAN']

# ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ (5D) --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑ CLUSTER (5D - Πάντα με ΑΡΧΙΚΑ Δεδομένα για ερμηνεία) --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# CUSTOM ΧΡΩΜΑΤΙΚΗ ΠΑΛΕΤΑ ΓΙΑ ΕΝΤΟΝΗ ΔΙΑΚΡΙΣΗ --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
def get_extended_distinct_colors(num_colors, base_colors):.....
# Οπτικοποιήσεις 5D (ΟΛΕΣ ΜΕ ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΑ ΔΕΔΟΜΕΝΑ)
# ΕΝΗΜΕΡΩΜΕΝΗ ΣΥΝΑΡΤΗΣΗ parallel_coordinates_plot_static_dbscan
def parallel_coordinates_plot_static_dbscan(df, dimensions, color_col, title, filename,
                                            output_dir, cluster_stats=None, original_df=None,
                                            scaler_5d=None):

    if scaler_5d is None:
        raise ValueError("Ο 'scaler_5d' πρέπει να παρασχεθεί στη συνάρτηση
        'parallel_coordinates_plot_static_dbscan'.")

    fig = plt.figure(figsize=(12, 7))
    num_dimensions = len(dimensions)

    clusters_only_df = df[df[color_col] != -1]
    unique_clusters_in_plot = sorted(clusters_only_df[color_col].unique())

    plot_colors = get_extended_distinct_colors(len(unique_clusters_in_plot),
                                               distinct_colors)
    cluster_id_to_specific_plot_color = {cid: plot_colors[i] for i,
                                         cid in enumerate(unique_clusters_in_plot)}


    for _, row in clusters_only_df.iterrows():
        cluster_id = row[color_col]
        color = cluster_id_to_specific_plot_color.get(cluster_id, 'gray')
        plt.plot(range(num_dimensions), row[dimensions], color=color, alpha=0.3)
```

```
if cluster_stats is not None:
    if isinstance(cluster_stats, pd.DataFrame):
        for cluster_id_to_plot in unique_clusters_in_plot:
            if cluster_id_to_plot in cluster_stats.index:
                cluster_mean = cluster_stats.loc[cluster_id_to_plot].xs('mean', level=1)
                mean_values = [cluster_mean[dim] for dim in dimensions]
                mean_df_for_scaler = pd.DataFrame([mean_values], columns=dimensions)
                scaled_mean_values = scaler_5d.transform(mean_df_for_scaler)[0]

                color = cluster_id_to_specific_plot_color.get(cluster_id_to_plot, 'gray')
                plt.plot(range(num_dimensions), scaled_mean_values, color=color,
                         linewidth=2,
                         label=f'Cluster {cluster_id_to_plot}')
    elif isinstance(cluster_stats, pd.Series):
        cluster_id_for_series = cluster_stats.name
        if cluster_id_for_series in unique_clusters_in_plot:
            cluster_mean_series = cluster_stats.xs('mean', level=1)
            mean_values_series = [cluster_mean_series[dim] for dim in dimensions]
            mean_df_for_scaler_series = pd.DataFrame([mean_values_series],
                                                     columns=dimensions)
            scaled_mean_values_series = scaler_5d.transform(mean_df_for_scaler_series)[0]

            color = cluster_id_to_specific_plot_color.get(cluster_id_for_series, 'gray')
            plt.plot(range(num_dimensions), scaled_mean_values_series, color=color,
                     linewidth=2,
                     label=f'Cluster {cluster_id_for_series}')
        else:
             print(f"Προσοχή: Το Cluster {cluster_id_for_series} δεν
             βρέθηκε στα δεδομένα για σχεδίαση.")


plt.xticks(range(num_dimensions), dimensions, rotation=45, ha='right')
plt.title(title)
plt.ylabel('Κανονικοποιημένη Τιμή')
plt.grid(True, linestyle='--', alpha=0.7)

handles, labels = plt.gca().get_legend_handles_labels()
if handles:
    plt.legend(title='Cluster', bbox_to_anchor=(1.05, 1), loc='upper left')

plt.tight_layout()

# Προσθήκη στατιστικών ως κείμενο στο διάγραμμα
if cluster_stats is not None:
    stats_text_blocks = []
    if original_df is not None:
        overall_means = original_df[dimensions].mean().to_dict()
        overall_std = original_df[dimensions].std().to_dict()
        overall_count = len(original_df)
        overall_min = original_df[dimensions].min().to_dict()
        overall_max = original_df[dimensions].max().to_dict()

        overall_text = f"Συνολικά Δεδομένα (N={overall_count})\n"
        for dim in dimensions:
            overall_text += (f"   {dim}: avg={overall_means[dim]:.2f},
            std={overall_std[dim]:.2f}\n"
                             f"   Min {dim}: {overall_min[dim]:.2f},
                             Max {dim}: {overall_max[dim]:.2f}\n")
```

```
            stats_text_blocks.append(overall_text)

        if isinstance(cluster_stats, pd.Series):
            cluster_id_for_stats = cluster_stats.name
            if (dimensions[0], 'count') in cluster_stats.index:
                count = cluster_stats[(dimensions[0], 'count')]
            else:
                count = "N/A"

            text_block = (f"Cluster {cluster_id_for_stats} (N={int(count)})\n")
            for dim in dimensions:
                if (dim, 'mean') in cluster_stats.index and \
                   (dim, 'std') in cluster_stats.index and \
                   (dim, 'min') in cluster_stats.index and \
                   (dim, 'max') in cluster_stats.index:
                    text_block += (f"   {dim}: avg={cluster_stats[(dim, 'mean')]:.2f},
                    std={cluster_stats[(dim, 'std')]:.2f}\n"
                                   f"   Min {dim}: {cluster_stats[(dim, 'min')]:.2f},
                                   Max {dim}: {cluster_stats[(dim, 'max')]:.2f}\n")
                else:
                    text_block += f"   {dim}: Στατιστικά μη διαθέσιμα.\n"
            stats_text_blocks.append(text_block)

        stats_text_blocks.reverse()

        y_start_bottom = 0.01
        line_height_factor = 0.02
        current_y_pos = y_start_bottom

        for block_text in stats_text_blocks:
            num_lines = block_text.count('\n') + 1

            block_height = num_lines * line_height_factor + 0.005

            add_stats_to_plot(fig, 0.99, current_y_pos, ha='right',
                              va='bottom',
                              text_content=block_text, fontsize=7)

            current_y_pos += block_height

    plt.savefig(f"{output_dir}/{filename}.png", bbox_inches='tight')
    plt.close()

# Γράφημα όλων των clusters (όλα τα σημεία εκτός θορύβου)
parallel_coordinates_plot_static_dbscan(
    scaled_df_5d_dbscan,
    dimensions=features_5d_cols,
    color_col="Cluster_5D_DBSCAN",
    title="Παράλληλες Συντεταγμένες Όλων των Clusters (5 Διαστάσεις -
    Κανονικοποιημένα)",
    filename="parallel_coordinates_all_clusters_dbscan_5d_scaled_static",
    output_dir=output_dir_dbscan_5d,
    cluster_stats=cluster_stats_5d_dbscan,
    original_df=earthquakes,
    scaler_5d=scaler_5d
)

# Μεμονωμένα Parallel Coordinates Plots για κάθε Cluster
selected_clusters_for_pcp = [c_id for c_id in unique_clusters_5d if c_id != -1]
```

```
for cluster_id in selected_clusters_for_pcp:
    cluster_data_scaled_5d = scaled_df_5d_dbscan[scaled_df_5d_dbscan['Cluster_5D_DBSCAN']
                                                 == cluster_id]

    if cluster_id in cluster_stats_5d_dbscan.index:
        current_cluster_stats = cluster_stats_5d_dbscan.loc[cluster_id]
    else:
        print(f"Δεν βρέθηκαν στατιστικά για Cluster {cluster_id}. Παράλειψη.")
        continue

    if not cluster_data_scaled_5d.empty:
        parallel_coordinates_plot_static_dbscan(
            cluster_data_scaled_5d,
            dimensions=features_5d_cols,
            color_col="Cluster_5D_DBSCAN",
            title=f"Παράλληλες Συντεταγμένες για Cluster {cluster_id} (5 Διαστάσεις -
            Κανονικοποιημένα)",
            filename=f"parallel_coordinates_cluster_{cluster_id}_dbscan_5d_scaled_static",
            output_dir=output_dir_dbscan_5d,
            cluster_stats=current_cluster_stats,
            scaler_5d=scaler_5d
        )

plot_df_5d_animated = scaled_df_5d_dbscan.copy()
plot_df_5d_animated = plot_df_5d_animated[['Latitude', 'Longitude', 'Depth (km)']]

plot_df_5d_animated['Magnitude (ML)'] = earthquakes['Magnitude (ML)']
plot_df_5d_animated['Year_Original'] = earthquakes['Year']
plot_df_5d_animated['Cluster_5D_DBSCAN'] = earthquakes['Cluster_5D_DBSCAN']

plot_df_5d_animated = plot_df_5d_animated[plot_df_5d_animated['Cluster_5D_DBSCAN'] != -1].copy()

fig_3d_5d_animated = px.scatter_3d(
    plot_df_5d_animated,
    x='Longitude',
    y='Latitude',
    z='Depth (km)',
    color='Cluster_5D_DBSCAN',
    size='Magnitude (ML)',
    animation_frame='Year_Original',
    color_continuous_scale=px.colors.sequential.Viridis if
    n_clusters_final_5d == 0 else None,
    color_discrete_sequence=px.colors.qualitative.Set1 if
    n_clusters_final_5d > 0 else None,
    title='Συσταδοποίηση DBSCAN (5 Διαστάσεις με Χρονική Εξέλιξη -
    Κανονικοποιημένα Χωρικά, Αρχική Magnitude)',
    hover_name="Cluster_5D_DBSCAN",
    hover_data={
        'Longitude': ':.2f',
        'Latitude': ':.2f',
        'Depth (km)': ':.2f',
        'Magnitude (ML)': ':.2f',
        'Year_Original': True,
        'Cluster_5D_DBSCAN': False
    },

    range_z=[plot_df_5d_animated['Depth (km)'].min(),
             plot_df_5d_animated['Depth (km)'].max()]
```

```
                        if not plot_df_5d_animated.empty
                        else [-1,1],
    range_x=[plot_df_5d_animated['Longitude'].min(),
             plot_df_5d_animated['Longitude'].max()]
                        if not plot_df_5d_animated.empty
                        else [-1,1],
    range_y=[plot_df_5d_animated['Latitude'].min(),
             plot_df_5d_animated['Latitude'].max()]
                        if not plot_df_5d_animated.empty
                        else [-1,1],
)

fig_3d_5d_animated.update_layout(
    scene = dict(
        xaxis_title="Longitude (Κανονικοποιημένο)",
        yaxis_title="Latitude (Κανονικοποιημένο)",
        zaxis_title="Depth (km) (Κανονικοποιημένο)"
    )
)
# Αποθήκευση του DataFrame με τις ετικέτες των clusters σε CSV --- OMOIA ME DBSCAN 2D
```

HDBSCAN 2D

```
# Εισαγωγή Βιβλιοθηκών --- OMOIA ΜΕ DBSCAN 2D
import hdbscan # Για τον HDBSCAN
from sklearn.preprocessing import StandardScaler # Για την κανονικοποίηση (scaling)
from sklearn.metrics import davies_bouldin_score, calinski_harabasz_score, silhouette_score

# Δημιουργία φακέλου για αποθήκευση γραφημάτων και δεδομένων HDBSCAN --- OMOIA ΜΕ DBSCAN 2D
# Συνάρτηση για τις τεκτονικές πλάκες --- OMOIA ΜΕ DBSCAN 2D
# Συνάρτηση για κείμενο (στατιστικά στο plot) --- OMOIA ΜΕ DBSCAN 2D
# Φόρτωση & Προεπεργασία Δεδομένων --- OMOIA ΜΕ DBSCAN 2D
# Μετατροπή ώρας & καθαρισμός --- OMOIA ΜΕ DBSCAN 2D
# Ορισμός 'Origin Time (GMT)' ως index για τις χρονοσειρές
# Επιλογή Χαρακτηριστικών για 2D (Latitude, Longitude) --- OMOIA ΜΕ DBSCAN 2D
scaler_hdbscan_2d = StandardScaler()
features_2d_hdbscan_scaled = scaler_hdbscan_2d.fit_transform(features_2d_hdbscan)

print(pd.DataFrame(features_2d_hdbscan_scaled, columns=features_2d_cols).head())
print("\n")

optimal_min_cluster_size = 25
optimal_min_samples = 13

print(f"\nΕπιλέχθηκαν παράμετροι για HDBSCAN:")
print(f"   optimal_min_cluster_size: {optimal_min_cluster_size}")
print(f"   optimal_min_samples: {optimal_min_samples}")

# ΕΦΑΡΜΟΓΗ ΤΟΥ HDBSCAN ---
hdbscan_clusterer = hdbscan.HDBSCAN(min_cluster_size=optimal_min_cluster_size,
                                    min_samples=optimal_min_samples,
                                    metric='euclidean')
earthquakes['Cluster'] = hdbscan_clusterer.fit_predict(features_2d_hdbscan_scaled)

# ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ ---
clustered_data_indices = (earthquakes['Cluster'] != -1)
features_for_metrics = features_2d_hdbscan_scaled[clustered_data_indices]
```

```
labels_for_metrics = earthquakes['Cluster'][clustered_data_indices]

unique_clusters = sorted(set(labels_for_metrics))
n_clusters_ = len(unique_clusters)

if n_clusters_ >= 2:
    db_score = davies_bouldin_score(features_for_metrics, labels_for_metrics)
    ch_score = calinski_harabasz_score(features_for_metrics, labels_for_metrics)
    silhouette_avg_score = silhouette_score(features_for_metrics, labels_for_metrics)

    print(f"   Davies-Bouldin Index: {db_score:.3f} (Χαμηλότερο είναι καλύτερο)")
    print(f"   Calinski-Harabasz Index: {ch_score:.3f} (Υψηλότερο είναι καλύτερο)")
    print(f"   Silhouette Score: {silhouette_avg_score:.3f} (Υψηλότερο είναι καλύτερο,
    εξαιρουμένου του θορύβου)")
else:
    print("\nΔεν υπάρχουν αρκετές συστάδες (>=2) για τον υπολογισμό των δεικτών
    Davies-Bouldin, Calinski-Harabasz και Silhouette.")

# ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑ CLUSTER --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# CUSTOM ΧΡΩΜΑΤΙΚΗ ΠΑΛΕΤΑ ΓΙΑ ΕΝΤΟΝΗ ΔΙΑΚΡΙΣΗ ---ΟΜΟΙΑ ΜΕ DBSCAN 2D
# Scatter Plot για τα Clusters χωρίς τους Outliers (Θόρυβο) ---ΟΜΟΙΑ ΜΕ DBSCAN 2D
# ΧΡΟΝΟΣΕΙΡΕΣ ... ---ΟΜΟΙΑ ΜΕ DBSCAN 2D
# ΔΗΜΙΟΥΡΓΙΑ ΔΙΑΔΡΑΣΤΙΚΩΝ ΧΑΡΤΩΝ ---ΟΜΟΙΑ ΜΕ DBSCAN 2D
Αποθήκευση του DataFrame με τις ετικέτες των clusters σε CSV --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
```

OPTICS 2D

```
# Εισαγωγή Βιβλιοθηκών --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
from sklearn.cluster import OPTICS
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import davies_bouldin_score, calinski_harabasz_score, silhouette_score

# Δημιουργία φακέλου για αποθήκευση γραφημάτων και δεδομένων OPTICS --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# Συνάρτηση για τις τεκτονικές πλάκες --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# Συνάρτηση για κείμενο (στατιστικά στο plot) --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# Φόρτωση & Προεπεργασία Δεδομένων --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# Μετατροπή ώρας & καθαρισμός --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# Ορισμός 'Origin Time (GMT)' ως index για τις χρονοσειρές
# Επιλογή Χαρακτηριστικών για 2D (Latitude, Longitude) --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
scaler_optics_2d = StandardScaler()
features_2d_optics_scaled = scaler_optics_2d.fit_transform(features_2d_optics)

print(pd.DataFrame(features_2d_optics_scaled, columns=features_2d_cols).head())
print("\n")

optimal_min_samples = 24
optimal_max_eps = np.inf
optimal_xi = 0.03

print(f"   optimal_min_samples: {optimal_min_samples}")
print(f"   optimal_max_eps: {optimal_max_eps}")
print(f"   optimal_xi (για εξαγωγή clusters): {optimal_xi}")

# ΕΦΑΡΜΟΓΗ ΤΟΥ OPTICS ---
optics_clusterer = OPTICS(min_samples=optimal_min_samples,
                          max_eps=optimal_max_eps,
                          metric='euclidean',
```

```
                          cluster_method='xi',
                          xi=optimal_xi)

optics_clusterer.fit(features_2d_optics_scaled)


earthquakes['Cluster'] = optics_clusterer.labels_

plt.figure(figsize=(15, 6))
ordered_reachability_distances = optics_clusterer.reachability_[optics_clusterer.ordering_]

non_inf_indices = np.where(np.isfinite(ordered_reachability_distances))[0]
plt.plot(non_inf_indices, ordered_reachability_distances[non_inf_indices], marker='.',
         linestyle='-', markersize=2)

plt.title('OPTICS Reachability Plot')
plt.xlabel('Indices of Ordered Points')
plt.ylabel('Reachability Distance')
plt.grid(True)
plt.tight_layout()
plt.savefig(f"{output_dir_optics_2d}/optics_reachability_plot.png")
plt.close()

# ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑ CLUSTER --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
# CUSTOM ΧΡΩΜΑΤΙΚΗ ΠΑΛΕΤΑ ΓΙΑ ΕΝΤΟΝΗ ΔΙΑΚΡΙΣΗ ---ΟΜΟΙΑ ΜΕ DBSCAN 2D
# Scatter Plot για τα Clusters χωρίς τους Outliers (Θόρυβο) ---ΟΜΟΙΑ ΜΕ DBSCAN 2D
# ΧΡΟΝΟΣΕΙΡΕΣ ... ---ΟΜΟΙΑ ΜΕ DBSCAN 2D
# ΔΗΜΙΟΥΡΓΙΑ ΔΙΑΔΡΑΣΤΙΚΩΝ ΧΑΡΤΩΝ ---ΟΜΟΙΑ ΜΕ DBSCAN 2D
Αποθήκευση του DataFrame με τις ετικέτες των clusters σε CSV --- ΟΜΟΙΑ ΜΕ DBSCAN 2D
```

WEKA

Aggregation ανά μήνα

```
import pandas as pd

df = pd.read_csv("/content/earthquakes_with_clusters_dbscan_2d.csv")

# Μετατροπή της στήλης χρόνου σε datetime
df['Origin Time (GMT)'] = pd.to_datetime(df['Origin Time (GMT)'])

# Προαιρετικό χρονικό φίλτρο
df = df[df['Origin Time (GMT)'].dt.year >= 1964]

# Επιλογή cluster 0
df_cluster = df[df['Cluster'] == 0].copy()

# Δημιουργία νέας στήλης Date = 1η ημέρα κάθε μήνα (για forecasting)
df_cluster['Date'] = df_cluster['Origin Time (GMT)'].dt.to_period('M').dt.to_timestamp()

# Aggregation ανά μήνα
df_monthly = df_cluster.groupby('Date').agg({
    'Magnitude (ML)': 'mean',
    'Depth (km)': 'mean',
    'Origin Time (GMT)': 'count'
}).rename(columns={'Origin Time (GMT)': 'Count'}).reset_index()
```

```
# Έλεγχος για missing values
if df_monthly.isnull().sum().sum() > 0:
    print("Missing values detected στο μηνιαίο aggregated dataframe:")
    print(df_monthly.isnull().sum())
else:
    print("Δεν βρέθηκαν missing values στο aggregated dataframe.")

# Αποθήκευση αρχείων
df_cluster.to_csv("cluster_0_full.csv", index=False)
df_monthly.to_csv("cluster_0_monthly_aggregated.csv", index=False)

print("Τα αρχεία αποθηκεύτηκαν σωστά για μηνιαίο forecasting.")
```

Δημιουργία .arff αρχείου

```
import pandas as pd

df = pd.read_csv("/content/earthquakes_with_clusters_dbscan_2d.csv")
df['Origin Time (GMT)'] = pd.to_datetime(df['Origin Time (GMT)'])

# Φιλτράρισμα για cluster 0
df_cluster = df[df['Cluster'] == 0].copy()

# Απόρριψη missing values (αν και δεν υπάρχουν)
df_cluster.dropna(subset=['Magnitude (ML)', 'Depth (km)', 'Origin Time (GMT)'], inplace=True)

# Ομαδοποίηση ανά μήνα
df_cluster['YearMonth'] = df_cluster['Origin Time (GMT)'].dt.to_period('M')
df_monthly = df_cluster.groupby('YearMonth').agg({
    'Magnitude (ML)': 'mean',
    'Depth (km)': 'mean',
    'Origin Time (GMT)': 'count'
}).rename(columns={'Origin Time (GMT)': 'Count'}).reset_index()

# Δημιουργία στήλης ημερομηνίας (1η του μήνα) για Weka
df_monthly['Date'] = df_monthly['YearMonth'].dt.to_timestamp()
df_monthly = df_monthly[['Date', 'Magnitude (ML)', 'Depth (km)', 'Count']]

# Εγγραφή σε .arff
with open('cluster_0_monthly_aggregated.arff', 'w', encoding='utf-8') as f:
    f.write('@relation cluster_0_monthly_aggregated\n\n')
    f.write('@attribute Date DATE "yyyy-MM-dd"\n')
    f.write('@attribute Magnitude_ML NUMERIC\n')
    f.write('@attribute Depth_km NUMERIC\n')
    f.write('@attribute Count NUMERIC\n\n')
    f.write('@data\n')

    for _, row in df_monthly.iterrows():
        date_str = row['Date'].strftime('%Y-%m-%d')
        f.write(f"{date_str},{row['Magnitude (ML)']:.4f},{row['Depth (km)']:.4f},{int(row['Count'])}\n")

print("Το αρχείο .arff αποθηκεύτηκε σωστά με ημερομηνία ως τύπο DATE.")
```

Βιβλιογραφία

[1] Ο.Α.Σ.Π, ``Σεισμός: Η γνώση είναι προστασία," *ΟΑΣΠ, Β΄ Έκδοση, Αθήνα*, pp. 9--31, 2007.

[2] Ο.Α.Σ.Π, ``Περί Σεισμών." `https://oasp.gr/peri-seismon`. Accessed: 22-04-2025.

[3] P. M. Shearer, *Introduction to seismology*. Cambridge university press, 2019.

[4] A. D'Alessandro, S. Scudero, and G. Vitale, ``A review of the capacitive mems for seismology," *Sensors*, vol. 19, no. 14, 2019.

[5] Y. Xie, M. E. Sichani, J. E. Padgett, and R. DesRoches, ``The promise of implementing machine learning in earthquake engineering: A state-of-the-art review," *Earthquake Spectra*, vol. 36, no. 4, pp. 1769--1801, 2020.

[6] Ε. Σπυράκος, Κ. Τουτουδάκη, ``Βασικές έννοιες σεισμολογίας," *Εργαστήριο Αντισεισμικής Τεχνολογίας." Εθνικού Μετσόβιου Πολυτεχνείου. Διαθέσιμο στο: http://lee. civil. ntua. gr/pdf/mathimata/antiseismikes_kataskeves/simeioseis/simeioseis_ak.pdf*, 2010.

[7] Ε. Λέκκας, ``Φυσικές και τεχνολογικές καταστροφές.," *Εκδόσεις Access, Αθήνα*, 2000.

[8] Γεωδυναμικό Ινστιτούτο, ``Ιστορία και Στόχοι." `https://www.gein.noa.gr/stoxoi/`. Accessed: 22-04-2025.

[9] J. Oyelade, I. Isewon, O. Oladipupo, O. Emebo, Z. Omogbadegun, O. Aromolaran, E. Uwoghiren, D. Olaniyan, and O. Olawole, ``Data clustering: Algorithms and its applications," in *2019 19th International Conference on Computational Science and Its Applications (ICCSA)*, pp. 71--81, 2019.

[10] O. Bousquet, U. von Luxburg, and G. Rätsch, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, vol. 3176. Springer, 2011.

[11] O. Maimon and L. Rokach, *Data mining and knowledge discovery handbook*, vol. 2. Springer, 2005.

[12] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2012.

[13] J. Han, K. Micheline, and J. Pei, ``Cluster analysis in data mining,'' *Coursera Course. Disponível em: https://www. coursera. org/course/clusteranalysis.[Último acesso: 25 de Maio de 2015]*, 2015.

[14] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.

[15] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, ``Dbscan revisited, revisited: why and how you should (still) use dbscan,'' *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1--21, 2017.

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, ``A density-based algorithm for discovering clusters in large spatial databases with noise,'' in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, p. 226–231, AAAI Press, 1996.

[17] J. Ravi, S. Kulkarni, *et al.*, ``A critical review on density-based clustering algorithms and their performance in data mining,'' *Int. J. Res. Anal. Rev.(IJRAR)*, vol. 9, no. 1, pp. 73--82, 2022.

[18] R. J. Campello, D. Moulavi, and J. Sander, ``Density-based clustering based on hierarchical density estimates,'' in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160--172, Springer, 2013.

[19] L. McInnes, J. Healy, and S. Astels, ``How hdbscan works.'' `https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html#how-hdbscan-works`. Accessed: 30-04-2025.

[20] L. McInnes, J. Healy, and S. Astels, ``hdbscan: Hierarchical density based clustering,'' *The Journal of Open Source Software*, vol. 2, mar 2017.

[21] P. J. Rousseeuw, ``Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,'' *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53--65, 1987.

[22] D. L. Davies and D. W. Bouldin, ``A cluster separation measure,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224--227, 1979.

[23] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, ``On clustering validation techniques,'' *Journal of intelligent information systems*, vol. 17, pp. 107--145, 2001.

[24] T. Caliński and J. H. and, ``A dendrite method for cluster analysis,'' *Communications in Statistics*, vol. 3, no. 1, pp. 1--27, 1974.

[25] Python Software Foundation, ``History and license.'' `https://docs.python.org/3/license.html`. Accessed: 25-04-2025.

[26] Python Software Foundation, ``Documentation.'' `https://docs.python.org/release/3.13.3/tutorial/index.html`. Accessed: 25-04-2025.

[27] D. Sarkar, R. Bali, and T. Sharma, ``Practical machine learning with python,'' *Book'' Practical Machine Learning with Python*, pp. 25--30, 2018.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, ``Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825--2830, 2011.

[29] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, ``API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108--122, 2013.

[30] Folium developers, ``Concepts." `https://python-visualization.github.io/folium/latest/`. Accessed: 27-04-2025.

[31] Plotly, Inc, ``Overview." `https://plotly.com/python/getting-started/`. Accessed: 27-04-2025.

[32] Python Software Foundation, ``Documentation." `https://docs.python.org/3/library/colorsys.html`. Accessed: 30-05-2025.

[33] Google, ``Frequently asked questions." `https://research.google.com/colaboratory/faq.html`. Accessed: 30-04-2025.

[34] DataScientest, ``Google colab: the power of the cloud for machine learning." `https://datascientest.com/en/google-colab-the-power-of-the-cloud-for-machine-learning`. Accessed: 30-04-2025.

[35] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for" Data Mining: Practical Machine Learning Tools and Techniques"*. The University of Waikato, 2016.

[36] M. Ibrahim and B. Al-Bander, ``An integrated approach for understanding global earthquake patterns and enhancing seismic risk assessment," *International Journal of Information Technology*, vol. 16, no. 4, pp. 2001--2014, 2024.

[37] I. H. Rifa, H. Pratiwi, and R. Respatiwulan, ``Clustering of earthquake risk in indonesia using k-medoids and k-means algorithms," *Media statistika*, vol. 13, no. 2, pp. 194--205, 2020.

[38] J. E. Ricardo, J. J. Domínguez Menéndez, I. F. Barcos Arias, J. M. Macías Bermúdez, and N. M. Lemus, ``Neutrosophic k-means for the analysis of earthquake data in ecuador.," *Neutrosophic Sets & Systems*, vol. 44, 2021.

[39] R. G. Nejad, R. A. Abbaspour, and M. Mojarab, ``Associating earthquakes with faults using cluster analysis optimized by a fuzzy particle swarm optimization algorithm for iranian provinces," *Soil Dynamics and Earthquake Engineering*, vol. 140, p. 106433, 2021.

[40] A. Jufriansah, Y. Pramudya, A. Khusnani, and S. Saputra, ``Analysis of earthquake activity in indonesia by clustering method," *Journal of Physics: Theories and Applications*, vol. 5, no. 2, p. 92, 2021.

[41] D. D. Atsa'am, T. Gbaden, and R. Wario, ``A machine learning approach to formation of earthquake categories using hierarchies of magnitude and consequence to guide emergency management," *Data Science and Management*, vol. 6, no. 4, pp. 208--213, 2023.

[42] M. Karmenova, A. Tlebaldinova, I. Krak, N. Denissova, G. Popova, Z. Zhantassova, and G. Györök, ``An approach for clustering of seismic events using unsupervised machine learning," *Acta Polytechnica Hungarica*, vol. 19, no. 5, pp. 7--22, 2022.

[43] J. S. Mufti and A. Dhini, ``Comparative analysis of centroid-based and density-based clustering for indonesian earthquake data," in *2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pp. 464--467, IEEE, 2024.

[44] Mustakim, E. Rahmi, M. R. Mundzir, S. T. Rizaldi, Okfalisa, and I. Maita, ``Comparison of dbscan and pca-dbscan algorithm for grouping earthquake area," in *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, pp. 1--5, 2021.

[45] G. Weatherill and P. W. Burton, ``Delineation of shallow seismic source zones using k-means cluster analysis, with application to the aegean region," *Geophysical Journal International*, vol. 176, no. 2, pp. 565--588, 2009.

[46] M. Sinambela, H. H. Arrizal, E. Darnila, Munawar, M. R. Nugraha, and A. Widodo, ``Clustering analysis of earthquake based on k-means, dbscan, and fuzzy c-means in north sumatra," in *2024 International Conference on Information Technology and Computing (ICITCOM)*, pp. 242--247, 2024.

[47] N. A. Karri, M. Yousuf Ansari, and A. Pathak, ``Identification of seismic zones of india using dbscan," in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, pp. 65--69, 2018.