Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Ενσωμάτωση Κοινωνικο-περιβαλλοντικών Χωρικών Δεδομένων σε Γράφους Γνώσης

Μελέτη και υλοποίηση

Διπλωματικη Εργασια

της

ΚΑΛΛΙΟΠΗΣ ΕΛΕΥΘΕΡΙΑΣ ΓΙΑΝΝΑΚΟΠΟΥΛΟΥ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής

Αθήνα, Φεβρουάριος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Ενσωμάτωση Κοινωνικο-περιβαλλοντικών Χωρικών Δεδομένων σε Γράφους Γνώσης

Μελέτη και υλοποίηση

Διπλωματικη Εργασια

της

ΚΑΛΛΙΟΠΗΣ ΕΛΕΥΘΕΡΙΑΣ ΓΙΑΝΝΑΚΟΠΟΥΛΟΥ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή τη 19η Φεβρουαρίου του 2024.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Συμεών Παπαβασιλείου	Ιωάννα Ρουσσάκη	Θεοδώρα Βαρβαρίγου
Καθηγητής	Αναπληρώτρια Καθηγήτρια	Καθηγήτρια

Αθήνα, Φεβρουάριος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Καλλιόπη Ελευθερία
Γιαννακοπούλου

Φεβρουάριος 2024

Περίληψη

Στην εποχή μας, όλο και περισσότερα χωρικά δεδομένα γίνονται διαθέσιμα σε περιφερειακό, εθνικό και παγκόσμιο επίπεδο. Η σύνδεσή τους και η εισαγωγή τους σε έναν γράφο γνώσης παρέχει νέες δυνατότητες στην ανάλυση δεδομένων, δεδομένου πως οι γράφοι γνώσης αποτελούν ισχυρά εργαλεία που μπορούν να συγκεντρώσουν δομημένα και μη δομημένα δεδομένα από ποικίλες πηγές και να τα εναρμονίσουν σημασιολογικά. Υπάρχουν ποικίλες προκλήσεις σχετικά με την ενσωμάτωση των χωρικών δεδομένων στους γράφους γνώσης, δεδομένης της ετερογένειας των δεδομένων, την απουσία σημασιολογικής περιγραφής τους, την ανάγκη για διαχείριση μεγάλου όγκου δεδομένων, και σε πολλές περιπτώσεις την απουσία ποιοτικών δεδομένων. Επί του παρόντος, έχει δοθεί μεγαλύτερη προσοχή στην ενσωμάτωση χωρικών δεδομένων σε γράφους RDF, με τους labeled property graphs να αποτελούν έναν ανεξερεύνητο τομέα.

Υπό αυτό το πρίσμα, παρουσιάζεται ένα σύνολο μηχανισμών για την αποδοτική και ποιοτική ενσωμάτωση κοινωνικο-περιβαλλοντικών χωρικών δεδομένων στο SustainGraph, έναν γράφο γνώσης που χρησιμοποιεί το labeled property graph μοντέλο, και περιέχει δεδομένα σχετικά με τους Στόχους Βιώσιμης Ανάπτυξης. Η αρχική επικύρωση των μεθόδων και τα αποτελέσματα των αρχικών αναλύσεων παρέχονται βάσει της ενσωμάτωσης συγκεκριμένων κοινωνικο-περιβαλλοντικών δεδομένων στο SustainGraph, και την πραγματοποίηση μίας χωρικής ανάλυσης για την περιοχή της Αθήνας στην Ελλάδα. Η προσέγγιση αυτή επικυρώνεται με μία ανάλυση στα χωρικά δεδομένα του SustainGraph, η οποία παρήγαγε πολύτιμες πληροφορίες σχετικά με τα χωρικά δεδομένα και τις κλιματικές συνθήκες. Τέτοιου είδους αναλύσεις μπορούν να εφαρμοστούν στα δεδομένα του SustainGraph και να υποστηρίξουν τη λήψη αποφάσεων.

Λέξεις Κλειδιά

χωρικά δεδομένα, γράφος γνώσης, labeled property graph, SustainGraph, κοινωνικο-περιβαλλοντικά δεδομένα, συσταδοποίηση

Abstract

Nowadays, more and more spatial data have become available at regional, national, and global levels. Interlinking and introducing them into a knowledge graph provides new possibilities in data analysis, given that knowledge graphs are powerful tools that can host structured or unstructured data from diverse sources and semantically align them. Various challenges exist for the integration of spatial data into knowledge graphs, considering the data heterogeneity, the lack of semantic description, the need for management of high data volumes, and the absence of qualitative data in many cases. Currently, focus has been given to the population of spatial data into RDF graphs, leaving labeled property graphs an unexplored domain.

Under this perspective, a set of methods is presented for the efficient and qualitative integration of socio-environmental spatial data to the SustainGraph, a labeled property knowledge graph that hosts data related to the Sustainable Development Goals. Initial validation and analysis results are provided based on the population of the SustainGraph with specific socio-environmental data and the realization of spatial analysis for the area of Athens in Greece. The approach is validated with an analysis upon the spatial data of the SustainGraph, which produced valuable insights regarding spatial patterns and climatological conditions. Such analysis processes could be performed over the SustainGraph and support decision making.

Keywords

spatial data, knowledge graph, labeled property graph, SustainGraph, socio-environmental data, clustering

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Συμεών Παπαβασιλείου για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Εργαστήριο Διαχείρισης και Βέλτιστου Σχεδιασμού Δικτύων Τηλεματικής. Επίσης ευχαριστώ ιδιαίτερα τον Δρ. Αναστάσιο Ζαφειρόπουλο, καθώς και τις ερευνήτριες Ελένη Φωτοπούλου και Ιωάννα Μανδηλαρά και την υποψήφια διδακτόρισσα Χριστίνα Μαρία Ανδρωνά, για την καθοδήγησή τους και την εξαιρετική συνεργασία που είχαμε. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου και την αδελφή μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Φεβρουάριος 2024

Καλλιόπη Ελευθερία Γιαννακοπούλου

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Στόχος της εργασίας

Οι γράφοι γνώσης αποτελούν ισχυρά εργαλεία που έχουν τη δυνατότητα να συγκεντρώνουν δομημένα και μη δομημένα δεδομένα προερχόμενα από ποικίλες πηγές, διασφαλίζοντας παράλληλα την κοινή τους σημασιολογία. Σημαντικό τους χαρακτηριστικό είναι πως υποστηρίζουν συνδέσεις μεταξύ των δεδομένων, οι οποίες εκφράζουν πολύτιμες πληροφορίες και συχνά αποκαλύπτουν κρυμμένες σχέσεις. Στην σύγχρονη εποχή, που χαρακτηρίζεται από την συγκέντρωση και την εξερεύνηση δεδομένων, η πληθώρα γεωχωρικών δεδομένων που είναι διαθέσιμα από ποικίλες πηγές όπως στατιστικές υπηρεσίες, περιβαλλοντικές οργανώσεις, και άλλες πλατφόρμες, σε περιφερειακό, εθνικό, και παγκόσμιο επίπεδο, παρέχουν νέες δυνατότητες στην ανάλυση δεδομένων. Οι επιστήμονες διαφορετικών κλάδων μπορούν να αξιοποιήσουν αυτές τις πληροφορίες για να καταλάβουν καλύτερα τα χαρακτηριστικά των χωρικών δεδομένων, να εφαρμόσουν αναλύσεις, και να υποστηρίξουν σε αυτές την λήψη αποφάσεων. Η σύνδεση των γεωχωρικών δεδομένων αποτελεί έναν ανερχόμενο τομέα που μπορεί να αντιμετωπίσει τη σημασιολογική ανομοιομορφία των διαφορετικών πηγών δεδομένων.

Υπάρχουν ποικίλες ερευνητικές εργασίες που έχουν πραγματοποιηθεί σχετικά με την ενσωμάτωση χωρικών δεδομένων στους γράφους γνώσης (Knowledge Graphs (KGs)), παρ' όλα αυτά, υπάρχουν ακόμα προκλήσεις που πρέπει να αντιμετωπιστούν. Σχετικά με την τεχνολογία στην οποία αναπτύσσονται οι γράφοι γνώσης, υπάρχει η δυνατότητα ανάπτυξής τους χρησιμοποιώντας το μοντέλο Resource Description Framework (RDF), είτε το μοντέλο Labeled Property Graph (LPG), ορίζοντας κόμβους και τις μεταξύ τους σχέσεις. Το RDF είναι το πρότυπο μοντέλο για την αντάλλαγή δεδομένων στον Παγκόσμιο Ιστό, συνεισφέροντας στον σημασιολογικό εμπλουτισμό του, ενώ το μοντέλο LPG χρησιμοποείται από βάσεις δεδομένων γράφων για την επίτευξη επαρκή αποθηκευτικού χώρου και γρήγορης διάσχισης των γράφων. Οι υφιστάμενες εργασίες σχετικά με την ενσωμάτωση χωρικών δεδομένων στους γράφους γνώσης αφορούν κυρίως γράφους που χρησιμοποιούν το RDF μοντέλο (π.χ. [1] [2] [3]), ενώ λιγότερες είναι αυτές που αξιοποιούν το LPG μοντέλο ([4] [5]). Ο σχεδιασμός μεθοδολογιών για την ενσωμάτωση χωρικών δεδομένων σε LPG γράφους είναι σημαντικός καθώς θα διευκολύνει την πραγματοποίηση εκτενών χωρικών αναλύσεων, διατηρώντας παράλληλα τα πλεονεκτήματα που προσφέρονται από τους LPG γράφους όσον αφορά την επίδοση των βάσεων δεδομένων γράφων και την ευκολία της χρήσης τους.

Επιπρόσθετα, υπάρχουν ποικίλες προκλήσεις σχετικά με την ανομοιομορφία και τον

όγκο των συνόλων χωρικών δεδομένων. Τα χωρικά δεδομένα μπορεί να περιέχονται σε αρχεία ποικίλων μορφών (π.χ. NetCDF, Shapefiles, CSV), γεγονός που απαιτεί την προσαρμοσμένη μετατροπή τους σε δομικά στοιχεία του γράφου γνώσης, καθώς είναι σημαντικό να αναπαραστηθούν σωστά οι χωρικές σχέσεις, ώστε να υποστηρίζεται η αξιόπιστη ανάλυση και λήψη αποφάσεων. Η ενσωμάτωση των χωρικών δεδομένων συχνά περιλαμβάνει μεγάλα σύνολα δεδομένων, απαιτώντας τη μείωση της χωρικής ανάλυσης για την επίτευξη της βέλτιστης διαχείρισης του μεγάλου όγκου δεδομένων, καθώς και της βάσης δεδομένων γράφων, όσον αφορά τη δυνατότητα κλιμάκωσής της. Επιπλέον, είναι απαραίτητη η διαχείριση της χρονικής διάστασης των δεδομένων μετά την εισαγωγή τους στον γράφο γνώσης, καθώς και η ανάπτυξη μίας διαδικασίας αξιολόγησης της ποιότητάς τους, με σκοπό την εξασφάλιση της ακρίθειας των χωρικών πληροφοριών.

Η παρούσα εργασία έχει ως στόχο την αντιμετώπιση των προκλήσεων που αναφέρθηκαν, εξερευνώντας τεχνικές για την ενσωμάτωση χωρικών δεδομένων χρονοσειράς στο SustainGraph [6], έναν LPG γράφο γνώσης που αναπτύχθηκε για τη διαχείριση κοινωνικοπεριβαλλοντικών πληροφοριών σχετικά με τους Στόχους Βιώσιμης Ανάπτυξης (Sustainable Development Goals (SDGs)). Η ενσωμάτωση χωρικών δεδομένων στον γράφο γνώσης προσφέρει σημασιολογική ομοιογένεια και διαλειτουργικότητα μεταξύ των συνόλων δεδομένων, συνεισφέροντας στη βέλτιστη ανάλυση και αξιοποίηση της γνώσης που προσφέρουν.

1.2 Δομή της εργασίας

Η διπλωματική εργασία αποτελείται από 3 κεφάλαια, τα οποία καλύπτουν πλήρως την ανάπτυξη της, καθώς και το απαιτούμενο θεωρητικό υπόβαθρο και τις τεχνολογίες που χρησιμοποιήθηκαν στα πλαίσια αυτής.

Στο κεφάλαιο 1 παρουσιάζεται το απαραίτητο θεωρητικό υπόβαθρο για την εργασία. Το κεφάλαιο αποτελείται από δύο ενότητες, η πρώτη εκ των οποίων περιέχει μία επισκόπηση της υφιστάμενης βιβλιογραφίας, και αποτελείται από τρεις υπο-ενότητες. Στην πρώτη υποενότητα παρουσιάζονται οι τεχνολογίες γράφων γνώσης με χωρικά δεδομένα, δηλαδή τα μοντέλα RDF και LPG, εστιάζοντας στο μοντέλο της βάσης δεδομένων γράφων του Neo4j. Στη δεύτερη υπο-ενότητα αναφέρονται και περιγράφονται εργαλεία που έχουν αναπτυχθεί για τη διευκόλυνση της δημιουργίας γράφων γνώσης με χωρικά δεδομένα, ενώ στην τρίτη υποενότητα παρουσιάζονται υφιστάμενοι γράφοι γνώσης που ενσωματώνουν γεωχωρική γνώση.

Η δεύτερη ενότητα του κεφαλαίου 1 περιέχει το απαραίτητο θεωρητικό υπόβαθρο για την ενσωμάτωση των χωρικών δεδομένων στο SustainGraph και αποτελείται από πέντε υποενότητες. Στην πρώτη υπο-ενότητα παρουσιάζεται ο γράφος γνώσης SustainGraph, ο στόχος του και η δομή του. Στη δεύτερη υπο-ενότητα αναλύεται ο τρόπος αναπαράστασης των χωρικών δεδομένων στο SustainGraph, τόσο ο τρέχων, όσο και η προτεινόμενη επέκταση για την ενσωμάτωση γεωμετριών. Στις επόμενες τρεις υπο-ενότητες παρουσιάζονται οι μηχανισμοί εισαγωγής χωρικών δεδομένων στο SustainGraph, πιθανοί μηχανισμοί μείωσης του όγκου τους, και μηχανισμοί ανάλυσής τους, αντίστοιχα. Περιέχεται τόσο το θεωρητικό υπόβαθρο που απαιτείται, όσο και η μεθοδολογία που ακολουθήθηκε σε κάθε περίπτωση.

Το κεφάλαιο 2 αποτελεί το πρακτικό μέρος της διπλωματικής εργασίας και περιέχει την

περιγραφή της υλοποίησης των μηχανισμών που παρουσιάστηκαν στο πρώτο κεφάλαιο. Αρχικά παρουσιάζεται το σύνολο χωρικών δεδομένων που χρησιμοποιήθηκε και ο επιθυμητός τρόπος αναπαράστασής τους στο SustainGraph, σύμφωνα με τις τεχνολογίες που παρουσιάστηκαν στο κεφάλαιο 1. Ύστερα, παρουσιάζονται οι μηχανισμοί προεπεξεργασίας και εισαγωγής των χωρικών δεδομένων στο SustainGraph. Έπειτα, περιγράφεται η υλοποίηση και εφαρμογή των μηχανισμών μείωσης όγκου χωρικών δεδομένων που μπορούν να εφαρμοστούν στο στάδιο επεξεργασίας τους, πριν την εισαγωγή τους. Στη συνέχεια, αναφέρονται πιθανοί μηχανισμοί ανάλυσης των χωρικών δεδομένων του SustainGraph, και παρουσιάζεται εκτενώς μία από αυτές. Στην τελευταία υπο-ενότητα του κεφαλαίου, αναφέρονται όλες οι τεχνολογίες που χρησιμοποιήθηκαν για την υλοποίηση του πρακτικού μέρους της εργασίας.

Στο κεφάλαιο 3 περιέχεται ο επίλογος της διπλωματικής εργασίας. Αρχικά παρουσιάζονται τα συμπεράσματα που προέκυψαν από την εργασία, και στη συνέχεια παρουσιάζονται πιθανές μελλοντικές επεκτάσεις του SustainGraph, στο πλαίσιο της ενσωμάτωσης χωρικών δεδομένων σε αυτό.

Στο τέλος της εργασίας δίνεται η βιβλιογραφία και οι πηγές που αξιοποιήθηκαν στο πλαίσιο της διπλωματικής, οι συντομογραφίες και τα αρκτικόλεξα που χρησιμοποιήθηκαν στο κείμενο, καθώς και η απόδοση ξενόγλωσσων όρων.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

Επισκόπηση βιβλιογραφίας

Στο κεφάλαιο αυτό παρουσιάζονται αναλυτικά οι βασικές τεχνολογίες που έχουν σχέση με την εργασία αυτή, καθώς και η χρήση τους σε υφιστάμενες εργασίες. Οι τεχνολογίες που παρουσιάζονται αφορούν την αναπαράσταση και τη διαχείριση γράφων γνώσης, καθώς και τις δυνατότητες αναπαράστασης χωρικών δεδομένων σε αυτούς.

Αρχικά θα παρουσιάσουμε την έννοια του γράφου γνώσης και τις δύο βασικές τεχνολογίες υλοποίησης του, καθώς και τον τρόπο αναπαράστασης χωρικών δεδομένων σε έναν γράφο γνώσης με κάθε μία από τις δύο τεχνολογίες. Ύστερα, θα εξετάσουμε υφιστάμενες μελέτες στις οποίες παρουσιάζονται εργαλεία που διευκολύνουν τη δημιουργία γράφων γνώσης με χωρικά δεδομένα. Τέλος, θα παρουσιαστούν υφιστάμενοι γράφοι γνώσης με χωρικά δεδομένα από ποικίλους τομείς, που έχουν αναπτυχθεί χρησιμοποιώντας τις δύο τεχνολογίες υλοποίησης.

2.1 Τεχνολογίες γράφων γνώσης με χωρικά δεδομένα

Ένας γράφος γνώσης είναι ένα δίκτυο που αποτελείται από οντότητες δεδομένων και τις μεταξύ τους σχέσεις, το οποίο αναπαρίσταται με την μορφή ενός κατευθυνόμενου γράφου. [7] Οι οντότητες, που μπορεί να αφορούν από καθημερινές έννοιες όπως ανθρώπους, εταιρίες και αντικείμενα έως πολύ ειδικές έννοιες επιστημονικών τομέων, αποτελούν τους κόμβους του γράφου, και οι μεταξύ τους σχέσεις αναπαρίστανται με τις ακμές του. [8] Σε έναν γράφο γνώσης με χωρικά δεδομένα εισάγεται και η χωρική ή γεωγραφική ιδιότητα στα δεδομένα, κάτι που γίνεται με την προσθήκη της ιδιότητας της τοποθεσίας σε κόμβους ή ακμές του γράφου.

Δύο βασικά και από τα πιο συχνά χρησιμοποιούμενα μοντέλα [9] για την αναπαράσταση των δεδομένων σε γράφους είναι τα εξής:

- Το Resource Description Framework (RDF)
- Οι Labeled Property Graphs (LPG)

Και τα δύο μοντέλα χρησιμοποιούν τα βασικά στοιχεία των γράφων όπως οι κόμβοι και οι ακμές, αλλά διαφέρουν στον τρόπο μοντελοποίησης τους.

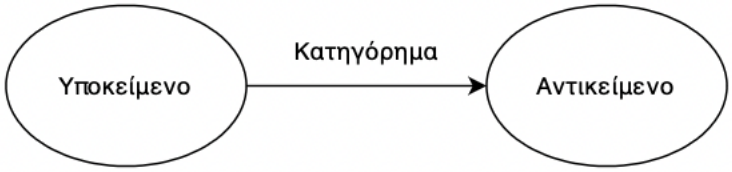
2.1.1 Resource Description Framework

Το Resource Description Framework (RDF) είναι ένα πρότυπο μοντέλο για την ανταλλαγή δεδομένων στον Παγκόσμιο Ιστό. Σχεδιάστηκε από το World Wide Web Consortium (W3C) με σκοπό να επεκτείνει τη δομή του διαδικτύου χρησιμοποιώντας Internationalized Resource Identifiers (IRIs), δηλαδή μοναδικά αναγνωριστικά για να ονομάσει τους πόρους και τις μεταξύ τους σχέσεις, και να συντελέσει στον σημασιολογικό εμπλουτισμό του Παγκόσμιου Ιστού. [10] [11]

Στο μοντέλο του RDF τα δεδομένα αναπαρίστανται και αποθηκεύονται σε μορφή κατευθυνόμενου γράφου. Η μορφή αυτή σχηματίζεται από ένα σύνολο δηλώσεων σχετικά με τους πόρους, κάθε μία εκ των οποίων έχει τη μορφή διατεταγμένης τριάδας [12]:

⟨υποκείμενο⟩ ⟨κατηγόρημα⟩ ⟨αντικείμενο⟩

Το υποκείμενο μπορεί να είναι ένα IRI, ένας κενός κόμβος ή μία άλλη RDF διατεταγμένη τριάδα ενώ το αντικείμενο μπορεί επιπρόσθετα να είναι ένα λεκτικό. Το κατηγόρημα μπορεί να είναι αποκλειστικά ένα IRI. Στην οπτικοποίηση ενός RDF γράφου, το υποκείμενο και το αντικείμενο αναπαριστώνται από κόμβους του γράφου, ενώ το κατηγόρημα από μία ακμή που τους συνδέει. [12] Ένα παράδειγμα φαίνεται στο παρακάτω σχήμα.



Σχήμα 2.1: *Οπτικοποίηση RDF γράφου με δύο κόμβους και μία ακμή*

Το RDF Schema (RDFS) [13] είναι μία σημασιολογική επέκταση του βασικού RDF λεξιλογίου και ορίζει ένα λεξιλόγιο για την μοντελοποίηση των δεδομένων σε έναν RDF γράφο. Είναι γραμμένο στο RDF πρότυπο και προσφέρει τη δυνατότητα της ομαδοποίησης πόρων που είναι σχετικοί μεταξύ τους και της περιγραφής των μεταξύ τους σχέσεων. Οι ομάδες που σχηματίζονται ονομάζονται κλάσεις και οι πόροι που είναι μέλη μίας κλάσης ονομάζονται στιγμιότυπα της κλάσης αυτής. Οι βασικότερες από αυτές είναι οι εξής:

- rdfs:Resource
  Οτιδήποτε περιγράφεται από το RDF, δηλαδή κάθε πόρος, είναι στιγμιόπτυπο της κλάσης αυτής. Κάθε άλλη κλάση είναι υποκλάση αυτής.

- rdfs:Class
  Είναι οι κλάση όλων των πόρων που είναι RDF κλάσεις.

- rdfs:Literal
  Είναι η κλάση όλων των ατόμων, όπως συμβολοσειρών και ακεραίων αριθμών.

- rdfs:Datatype
  Είναι η κλάση όλων των τύπων δεδομένων.

- rdfs:Property
  Είναι η κλάση όλων των RDF ιδιοτήτων.

Οι ιδιότητες στο RDFS είναι οι σχέσεις μεταξύ των υποκειμένων και των αντικειμένων. Οι βασικότερες από αυτές είναι οι εξής:

- rdfs:range
  Δηλώνει ότι μία ιδιότητα μπορεί να έχει ως αντικείμενο ένα στιγμιότυπο μίας ή περισσότερων κλάσεων.

- rdfs:domain
  Δηλώνει ότι το υποκείμενο που έχει μία ιδιότητα είναι στιγμιότυπο μίας ή περισσότερων κλάσεων.

- rdf:type
  Δηλώνει ότι ένας πόρος είναι στιγμιότυπο μίας κλάσης.

- rdfs:subClassOf
  Δηλώνει ότι όλα τα στιγμιότυπα μίας κλάσης είναι στιγμιότυπα και μίας άλλης.

- rdfs:subPropertyOf
  Δηλώνει ότι όλοι οι πόροι που σχετίζονται με μία ιδιότητα σχετίζονται και με μία άλλη.

Με την χρήση του RDFS δίνεται η δυνατότητα ορισμού ταξονομιών και εξαγωγής βασικών συμπερασμάτων για τους πόρους ενός RDF γράφου και συνεπώς η δυνατότητα ορισμού απλών οντολογιών. Για την περιγραφή πιο σύνθετων οντολογιών χρησιμοποιείται η Web Ontology Language OWL 2 [14].

Για την ανάκτηση και διαχείριση δεδομένων που είναι αποθηκευμένα σε γράφους RDF χρησιμοποείται η γλώσσα επερωτήσεων SPARQL [15].

Επεκτάσεις για αναπαράσταση χωρικών δεδομένων σε RDF γράφους

Για την αναπαράσταση χωρικών δεδομένων στους RDF γράφους γνώσης είναι απαραίτητη η επέκταση του βασικού RDF λεξιλογίου ώστε να περιγράφονται τα χωρικά δεδομένα και οι μεταξύ τους τοπολογικές σχέσεις. Παράλληλα, χρειάζεται επέκταση της γλώσσας SPARQL για να δοθεί η δυνατότητα επερωτήσεων επί χωρικών δεδομένων. Κάποιες από τις τεχνολογίες που έχουν προταθεί για αυτόν τον σκοπό και που χρησιμοποιούνται συχνά στους υφιστάμενους γράφους γνώσης με χωρικά δεδομένα σε RDF είναι το πρότυπο GeoSPARQL του Open Geospatial Consortium (OGC) και το μοντέλο stRDF σε συνδυασμό με την γλώσσα επερωτήσεων stSPARQL.

Το πρότυπο GeoSPARQL [16] ορίζει μία οντολογία γραμμένη σε RDFS και OWL για την αναπαράσταση χωρικών αντικειμένων και κάποιες συναρτήσεις για την επέκταση της SPARQL. Οι δύο βασικές κλάσεις που ορίζονται στην οντολογία του GeoSPARQL είναι η geo:SpatialObject, η οποία αναπαριστά οτιδήποτε έχει χωρική αναπαράσταση, και η

geo:Feature, η οποία είναι υποκλάση της geo:SpatialObject, και αναπαριστά κάθε αντικείμενο του πραγματικού κόσμου που έχει κάποια χωρική τοποθεσία, δηλαδή μπορεί να αναπαρασταθεί στον χάρτη. Με την επέκταση γεωμετρίας του GeoSPARQL ορίζεται η κλάση geo:Geometry, η οποία αναπαριστά γεωμετρίες, δηλαδή γεωμετρικά σχήματα όπως σημεία, γραμμές και πολύγωνα, που περιγράφουν την χωρική συνιστώσα των αντικειμένων. Ορίζονται, επίσης, οι ιδιότητες geo:hasGeometry και geo:hasDefaultGeometry, οι οποίες συνδέουν ένα στιγμιότυπο της geo:Feature με ένα στιγμιότυπο της geo:Geometry για την αναπαράσταση της χωρικής έκτασης του πρώτου, δηλαδή της τοποθεσίας του.

Για την κωδικοποίηση της γεωμετρικής πληροφορίας η GeoSPARQL χρησιμοποιεί τις σειριοποιήσεις Well-known Text (WKT) και Geography Markup Language (GML). Η σειριοποίηση WKT ευθυγραμμίζει τους γεωμετρικούς τύπους με το ISO 19125-1 Simple Features και η GML με το ISO 19107 Spatial Schema. Οι σειριοποιήσεις αυτές μπορούν να κωδικοποιήσουν γεωμετρίες όπως σημεία (Point), γραμμές (LineString), πολύγωνα (Polygon) και πιο σύνθετα αντικείμενα όπως MultiPoint, MultiLineString, MultiPolygon. Για τη σύνδεση μίας γεωμετρίας με την WKT σειριοποίηση της αυτή συνδέεται μέσω της ιδιότητας geo:asWKT με ένα λεκτικό που έχει τύπο δεδομένων geo:wktLiteral και για τη σύνδεση με την GML σειριοποίηση της συνδέεται μέσω της ιδιότητας geo:asGML με ένα λεκτικό που έχει τύπο δεδομένων geo:gmlLiteral. Και οι δύο ιδιότητες είναι υπο-ιδιότητες (subproperties) της geo:hasSerialization. Για τα geo:wktLiteral λεκτικά, το προκαθορισμένο σύστημα συντεταγμένων είναι το WGS 84 γεωδαιτικό χωρικό σύστημα αναφοράς γεωγραφικού πλάτους και γεωγραφικού μήκους. Ένα παράδειγμα κωδικοποίησης ενός σημείου στο σύστημα WGS 84 χρησιμοποιώντας σειριοποίηση WKT είναι το εξής:

```
"Point(-83.38 33.95)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>
```

Χρησιμοποιώντας την σειριοποίηση GML το ίδιο σημείο κωδικοποιείται ως εξής:

```
"<gml:Point
     srsName=\"http://www.opengis.net/def/crs/OGC/1.3/CRS84\"
     xmlns:gml=\"http://www.opengis.net/ont/gml\">
   <gml:pos>-83.38 33.95</gml:pos>
</gml:Point>"^^<http://www.opengis.net/ont/geosparql#gmlLiteral>
```

Το GeoSPARQL ορίζει ιδιότητες για την περιγραφή τοπολογικών σχέσεων μεταξύ των χωρικών αντικειμένων, χρησιμοποιώντας τοπολογικές σχέσεις τριών διαφορετικών οικογενειών: των Simple Features, Egenhofer και RCC8. Η οικογένεια Simple Features περιλαμβάνει τις σχέσεις equals, disjoint, intersects, touches, within, contains, overlaps και crosses, οι οποίες εκφράζουν τη σχέση δύο χωρικών αντικειμένων που είναι ισοδύναμα, ξένα, τέμνονται, εφάπτονται, περιέχονται το ένα στο άλλο, έχουν επικάλυψη ή διασχίζει το ένα το άλλο, αντίστοιχα.

Οι συναρτήσεις SPARQL που ορίζονται στο GeoSPARQL για την δυνατότητα εφαρμογής χωρικών μεθόδων σε γεωμετρικά δεδομένα είναι οι geof:distance, geof:buffer, geof:convexHull, geof:intersection, geof:union, geof:difference, geof:symDifference, geof:envelope και geof:boundary. Για παράδειγμα η geof:distance επιστρέφει την απόσταση μεταξύ δύο γεωμετριών, ενώ η

geof:intersection επιστρέφει την τομή των δύο γεωμετριών, δηλαδή ένα γεωμετρικό αντικείμενο που περιέχει όλα τα σημεία που περιέχονται και στις δύο γεωμετρίες.

Το μοντέλο δεδομένων stRDF αποτελεί μία επέκταση του RDF για την αναπαράσταση γεωχωρικών δεδομένων που μπορούν να μεταβληθούν στον χρόνο. Χρησιμοποιεί τα πρότυπα WKT και GML του OGC για τη σειριοποίηση των γεωμετριών, ορίζοντας τους τύπους δεδομένων strdf:WKT και strdf:GML αντίστοιχα. Επιπρόσθετα, ορίζεται ο τύπος δεδομένων strdf:geometry, ο οποίος αναπαριστά την σειριοποίηση μίας γεωμετρίας ανεξαρτήτως του προτύπου που χρησιμοποιείται. Πέρα από τους αναφερθέντες τύπους δεδομένων, το μοντέλο stRDF δεν προσφέρει επιπλέον λεξιλόγιο για την μοντελοποίηση γεωχωρικών δεδομένων, όπως κλάσεις και ιδιότητες για την περιγραφή των χωρικών αντικειμένων και των μεταξύ τους σχέσεων. Ο ορισμός τους αφήνεται στους χρήστες του μοντέλου, κατά τον ορισμό της δικής τους οντολογίας. [17]

Η stSPARQL αποτελεί μία επέκταση της γλώσσας SPARQL. Ορίζει μία συνάρτηση για κάθε τοπολογική σχέση των οικογενειών Simple Features, Egenhofer και RCC8, όπως την strdf:contains και την strdf:overlap. Επιπλέον, ορίζει τις συναθροιστικές συναρτήσεις strdf:union, strdf:intersection και strdf:extent για χωρικά δεδομένα. [17]

2.1.2 Labeled Property Graphs

Το Labeled Property Graph (LPG) είναι ένα μοντέλο για την αναπαράσταση, την αποθήκευση και τη διαχείριση δεδομένων σε μορφή γράφου. Χρησιμοποιεί κόμβους και κατευθυνόμενες ακμές για να αναπαραστήσει τις οντότητες και τις μεταξύ τους σχέσεις, αντίστοιχα. [18] Τόσο οι κόμβοι όσο και οι ακμές έχουν εσωτερική δομή η οποία αποτελείται από το μοναδικό τους αναγνωριστικό, από μία ή περισσότερες ετικέτες που περιγράφουν τον τύπο τους ή την κλάση στην οποία ανήκουν, και τις ιδιότητές τους, μοντελοποιημένες με την μορφή ζευγαριών κλειδιού-τιμής (key-value). Η εσωτερική αυτή δομή προσφέρει πιο συμπαγή και ευκολονόητη αναπαράσταση των δεδομένων όπως επίσης και τη δυνατότητα επαρκή αποθηκευτικού χώρου και γρήγορης διάσχισης του γράφου. [19]

2.1.3 Neo4j

Το Neo4j αποτελεί μία βάση δεδομένων γράφων (graph database) που χρησιμοποιεί το LPG μοντέλο. Κάποια χαρακτηριστικά του Neo4j μοντέλου είναι πως οι κόμβοι μπορούν να έχουν καμία, μία ή περισσότερες ετικέτες για τον προσδιορισμό του είδους τους, ενώ οι σχέσεις έχουν πάντα έναν τύπο, που ορίζει τον τύπο της σχέσης. Ο κόμβος από την οποία ξεκιναεί μία σχέση ονομάζεται κόμβος πηγή (source node), ενώ αυτός στον οποίο καταλήγει ονομάζεται κόμβος στόχος (target node). Οι ιδιότητες περιγράφουν περεταίρω τους κόμβους και τις σχέσεις, και οι τιμές στα ζεύγη κλειδί-τιμή των ιδιοτήτων μπορούν να περιέχουν πολλούς διαφορετικούς τύπους δεδομένων, όπως αριθμούς, συμβολοσειρές και τύπους αληθείας, καθώς και ομογενείς λίστες από τιμές κάποιου τύπου δεδομένων. [20]

Οι κόμβοι και οι σχέσεις μπορούν να θεωρηθούν το μικρότερα δομικά συστατικά ενός LPG, και συνεπώς ενός γράφου στο Neo4j. Με βάση αυτά, δημιουργούνται πρότυπα (patterns) από συνδεδεμένους κόμβους και σχέσεις, που δίνουν τη δυνατότητα κωδικοποίησης πολύπλοκης πληροφορίας, σε αντίθεση με την περιορισμένη πληροφορία που μπορεί να

αναπαρασταθεί σε έναν κόμβο. Στα πρότυπα βασίζεται η δηλωτική γλώσσα επερωτήσεων Cypher, η οποία χρησιμοποιείται για την ανάκτηση δεδομένων από έναν γράφο γνώσης στο Neo4j, καθώς έχει σχεδιαστεί έτσι ώστε να αναγνωρίζει στον γράφο γνώσης τα πρότυπα που γράφονται σε μία επερώτηση. Για την εύρεση των απαντήσεων στις επερωτήσεις, πραγματοποιούνται διασχίσεις του γράφου, ακολουθώντας τις σχέσεις των προτύπων. [21]

Το σχήμα μίας βάσης δεδομένων γράφων στο Neo4j μπορεί να οριστεί χρησιμοποιώντας δείκτες και περιορισμούς. Ο ορισμός ενός σχήματος είναι προαιρετικός, καθώς ένας γράφος γνώσης στο Neo4j μπορεί να δημιουργηθεί ορίζοντας κόμβους και σχέσεις, χωρίς να έχει δηλωθεί εκ των προτέρων το σχήμα του. Οι δείκτες εφαρμόζονται σε ιδιότητες συγκεκριμένων ετικετών κόμβων ή τύπων σχέσεων και χρησιμοποιούνται για να επιταχύνουν τη διάσχιση του γράφου. Οι περιορισμοί χρησιμοποιούνται για να διασφαλίσουν πως τα δεδομένα που περιέχονται στον γράφο γνώσης ακολουθούν κάποιους κανόνες. Οι δείκτες μπορούν να οριστούν μετά τη δημιουργία του γράφου, όπως και οι περιορισμοί, εφόσον τηρούνται ήδη από τα δεδομένα. [20]

Χωρικά δεδομένα στο Neo4j

Για την αναπαράσταση χωρικών δεδομένων στους γράφους γνώσης στο Neo4j, χρησιμοποιούνται χωρικές τιμές που αποθηκεύονται ως ιδιότητες στους κόμβους και στις σχέσεις του γράφου. Οι χωρικές τιμές που υποστηρίζονται από το Neo4j και τη γλώσσα επερωτήσεων Cypher είναι οι τιμές POINT, δηλαδή τιμές που αναπαριστούν ένα σημείο στον χώρο. Οι συντεταγμένες ενός POINT μπορούν να αποτελούνται είτε από δύο αριθμούς, το γεωγραφικό μήκος και το γεωγραφικό πλάτος, είτε από τρεις αριθμούς, συμπεριλαμβάνοντας και το ύψος. Τα γεωγραφικά συστήματα συντεταγμένων που υποστηρίζονται για την μοντελοποίηση των σημείων πάνω στη γη είναι το WGS 84 2D, για τα σημεία με δύο συντεταγμένες, και το WGS 84 3D, για τα σημεία με τρεις συντεταγμένες. Παράλληλα, υποστηρίζονται δύο καρτεσιανά συστήματα συντεταγμένων για την μοντελοποίηση σημείων στον Ευκλείδειο χώρο, το cartesian 2D και το cartesian 3D, για σημεία σε χώρο δύο και τριών διαστάσεων, αντίστοιχα. [22]

Η Cypher υποστηρίζει ένα σύνολο από χωρικές συναρτήσεις προκειμένου να γίνει εφικτή η εφαρμογή επερωτήσεων επί των χωρικών δεδομένων. Συγκεκριμένα, οι συναρτήσεις point χρησιμοποιούνται για να ορίσουν ένα σημείο οποιουδήποτε από τα τέσσερα συστήματα συντεταγμένων που υποστηρίζονται, έχοντας ως ορίσματα τις συντεταγμένες του και το κατάλληλο σύστημα συντεταγμένων. Υποστηρίζεται, επίσης, η συνάρτηση distance, η οποία υπολογίζει την απόσταση μεταξύ δύο σημείων, και η συνάρτηση withinBBox, η οποία υπολογίζει εάν ένα σημείο βρίσκεται εντός ενός ορθογωνίου, και επιστρέφει την κατάλληλη τιμή αληθείας. Το ορθογώνιο ορίζεται από δύο σημεία, το πάνω δεξιά και το κάτω αριστερά του ορθογωνίου. [23]

Η APOC είναι μία βιβλιοθήκη που υποστηρίζεται επίσημα από το Neo4j, η οποία περιέχει διαδικασίες και συναρτήσεις που μπορούν να συμπεριληφθούν στις επερωτήσεις της Cypher, παρέχοντας νέες χρήσιμες λειτουργικότητες. Στη βιβλιοθήκη αυτή συμπεριλαμβάνονται τέσσερις επιπλέον χωρικές συναρτήσεις, οι οποίες δεν χρησιμοποιούν τον τύπο χωρικών τιμών POINT. Η geocodeOnce και η geocode λαμβάνουν ως είσοδο μία διεύθυνση σε μορφή συμ-

βολοσειράς και αναζητούν την τοποθεσία της σε κάποια υπηρεσία γεωκωδικοποίησης, όπως η OpenStreetMap, η οποία είναι και η προκαθορισμένη. Επιστρέφουν μία ή περισσότερες τοποθεσίες, αντίστοιχα, δηλαδή το γεωγραφικό πλάτος και το γεωγραφικό μήκος τους, καθώς και μία περιγραφή κάθε τοποθεσίας και επιπλέον πληροφορίες για αυτή. Η συνάρτηση reverseGeocode, με αντίστροφο τρόπο, λαμβάνει ως είσοδο τις δύο συντεταγμένες μίας τοποθεσίας και επιστρέφει τη διεύθυνσή της. Τέλος, η συνάρτηση sortByDistance λαμβάνει ως είσοδο μία λίστα από μονοπάτια, και τα ταξινομεί σε αύξουσα σειρά απόστασης. Οι κόμβοι που περιέχονται στα μονοπάτια πρέπει να έχουν ως ιδιότητες το γεωγραφικό πλάτος και το γεωγραφικό μήκος της τοποθεσίας που αντιπροσωπεύουν, καθώς με βάση αυτά υπολογίζονται οι αποστάσεις μεταξύ τους. [24]

Χωρική βιβλιοθήκη για το Neo4j

Εκτός από τις δυνατότητες που παρέχονται από τη Cypher και την πρότυπη βιβλιοθήκη APOC, υπάρχει η δυνατότητα διαχείρισης χωρικών δεδομένων στο Neo4j με τη χρήση της βιβλιοθήκης Neo4j Spatial [25]. Πρόκειται για μία ειδική βιβλιοθήκη για χωρικά δεδομένα που δεν υποστηρίζεται επισήμως από το Neo4j, αλλά αναπτύχθηκε από προγραμματιστές και χρήστες του Neo4j σε μορφή λογισμικού ανοιχτού κώδικα, με σκοπό να διευκολύνει την εφαρμογή χωρικών πράξεων στα δεδομένα ενός γράφου Neo4j. Κάποια από τα βασικά της χαρακτηριστικά είναι η δυνατότητα εισαγωγής δεδομένων από αρχεία με χωρικά δεδομένα (π.χ. shapefiles, Open Street Map αρχεία), και η υποστήριξη των βασικών γεωμετριών (π.χ. σημείο, γραμμή, πολύγωνο). Επιπλέον, υποστηρίζει δείκτες που επιταχύνουν τις αναζητήσεις στα χωρικά δεδομένα, καθώς και τοπολογικές πράξεις που εκφράζουν σχέσεις μεταξύ των γεωμετριών όπως οι “περιέχεται”, “επικαλύπτει”, “τεμνει”.

2.2 Εργαλεία δημιουργίας γράφων γνώσης με χωρικά δεδομένα

Ποικίλες ερευνητικές μελέτες έχουν πραγματοποιηθεί σχετικά με την ανάπτυξη γράφων γνώσης με χωρικά δεδομένα, εξερευνώντας μεθοδολογίες για την αναπαράστασή και την ανάλυση τους. Στην παρούσα ενότητα παρουσιάζονται έρευνες που έχουν γίνει για την ανάπτυξη εργαλείων που βοηθούν στην δημιουργία γράφων γνώσης με χωρικά δεδομένα. Όλες οι προσεγγίσεις που παρουσιάζονται πραγματοποιούν την μοντελοποίηση χρησιμοποιώντας το RDF μοντέλο.

Το GeoTriples [26] είναι ένα εργαλείο που αναπτύχθηκε για να διευκολύνει την μετατροπή των γεωχωρικών δεδομένων που προέρχονται από ποικίλες πηγές, σε έναν γράφο γνώσης RDF. Οι πηγές αυτές μπορούν να είναι είτε αρχεία ποικίλων διαφορετικών μορφών (π.χ. αρχεία shapefiles, csv, XML, κ.α.), είτε σχεσιακές βάσεις δεδομένων που έχουν τη δυνατότητα αποθήκευσης χωρικών δεδομένων (π.χ. PostGIS, MonetDB). Τα δεδομένα μετατρέπονται σε γράφο γνώσης στο RDF μοντέλο, χρησιμοποιώντας βιβλιοθήκες για χωρικά δεδομένα όπως η GeoSPARQL και η stRDF/stSPARQL. Για τη μετατροπή χρησιμοποιούνται R2ML και RML απεικονίσεις για γεωχωρικά δεδομένα οι οποίες παράγονται από το GeoTriples, επεκτείνοντας τις αντίστοιχες γλώσσες απεικονίσεων σε RDF γράφους. Το TripleGeo [27] παρουσιάζει μία παρόμοια προσέγγιση, αξιοποιώντας και επεκτείνοντας τη βιβλιοθήκη geometry2rdf για

τη μετατροπή χωρικών δεδομένων ποικίλων μορφών σε RDF διατεταγμένες τριάδες, οι οποίες μπορούν να φορτωθούν σε γράφους γνώσης RDF. Για την αναπαράσταση των γεωμετριών στους RDF γράφους, το TripleGeo υποστηρίζει τη βιβλιοθήκη GeoSPARQL.

Το Theseus [28] αποτελεί ένα πλαίσιο που χρησιμοποιείται για την διαχείριση του κύκλου ζωής δεδομένων που αφορούν τις διαμερίσεις των εδαφών σύμφωνα με την Εδαφική Στατιστική Ονοματολογία (Territorial Statistical Nomenclature (TSN)). Ο κύκλος αυτός ξεκινάει από την συλλογή των χωρικών δεδομένων, συνεχίζεται με την μοντελοποίησή τους, και ολοκληρώνεται με τη δημιουργία και δημοσίευση του γράφου γνώσης. Τα χωρικά δεδομένα που χρησιμοποιούνται ως είσοδος περιέχονται σε αρχεία τύπου shapefile, ένα για κάθε εκδοχή των TSN. Τα αρχεία μετατρέπονται σε μορφή συμβατή με τις οντολογίες του Theseus, οι οποίες είναι δύο: μία για την περιγραφή των TSN και μία για περιγραφή της εξέλιξής τους στον χρόνο. Έπειτα, τα αρχεία αυτά, καθώς και οι διαφορές μεταξύ τους, μετατρέπονται σε έναν γράφο γνώσης που χρησιμοποιεί το μοντέλο RDF και τις δύο οντολογίες του Theseus. Τέλος, μετά τη δημοσίευση του γράφου γνώσης στον Παγκόσμιο Ιστό, το Theseus παρέχει ποικίλες μηχανές αναζήτησης στα δεδομένα του γράφου, παρέχοντας στους χρήστες εργαλεία για την εξερεύνηση του.

2.3 Υφιστάμενοι γράφοι γνώσης με χωρικά δεδομένα

Παράλληλα με τις μελέτες για την ανάπτυξη εργαλείων που διευκολύνουν τη δημιουργία των γράφων, έχουν πραγματοποιηθεί ποικίλες μελέτες σχετικά με την ανάπτυξη και δημοσίευση συγκεκριμένων γράφων γνώσης με χωρικά δεδομένα για ποικίλους τομείς. Στην παρούσα ενότητα παρουσιάζονται υφιστάμενοι γράφοι γνώσης με χωρικά δεδομένα, που βασίζονται είτε στο RDF μοντέλο, είτε στο LPG. Στην περίπτωση του LPG, οι γράφοι γνώσης έχουν αναπτυχθεί στο Neo4j.

Το YAGO2 [29] και το YAGO2geo [1] αποτελούν δύο γράφους που περιέχουν γενική γνώση για τον κόσμο και μοντελοποιούν, μεταξύ άλλων, γεωχωρικά δεδομένα. Στο YAGO2 τα χωρικά δεδομένα προέρχονται από τη Wikipedia και το GeoNames. Το GeoNames αποτελεί ένα λεξικό γεωχωρικής πληροφορίας, καθώς για κάθε όνομα που του δίνεται ως είσοδος (π.χ. όνομα πόλης, όνομα βουνού, κ.α.), δίνει την αντίστοιχη τοποθεσία. Οι οντότητες του YAGO2 που έχουν χωρική διάσταση ονομάζονται geoentities και χαρακτηρίζονται από την ύπαρξη δύο κατηγορημάτων που ορίζουν την τοποθεσία τους. Τα κατηγορήματα αυτά είναι το haslongitude και το haslatitude και συνδέουν το geoentity με το γεωγραφικό μήκος και το γεωγραφικό πλάτος του κέντρου της γεωμετρίας του, αντίστοιχα. Το YAGO2geo αποτελεί μία επέκταση του YAGO2 με ακριβή γεωχωρική γνώση και λεπτομερείς γεωμετρίες (σημεία, γραμμές, πολύγωνα, κ.λπ.). Οι γεωχωρικές πληροφορίες προέρχονται από δύο πηγές: τα διοικητικά δεδομένα της Ελλάδας, του Ηνωμένου Βασιλίου και της Ιρλανδίας, και το σύνολο δεδομένων του Open Street Map. Και οι δύο γράφοι γνώσης, το YAGO2 και το YAGO2geo, χρησιμοποιούν το μοντέλο RDF. Για την αναπαράσταση των γεωμετριών και την έκφραση τοπολογικών σχέσεων στο YAGO2geo χρησιμοποείται, επίσης, η βιβλιοθήκη GeoSPARQL.

Το UrbanKG [2] είναι ένας γράφος γνώσης που αναπτύχθηκε για διευκολύνει την εξερεύνηση του τεράστιου όγκου αστικών δεδομένων που παράγονται καθημερινά σε μία πόλη, ώστε να μπορούν να αξιοποιηθούν για την βελτίωση της ζωής σε αυτή. Ο γράφος γνώσης

είναι μοντελοποιημένος με βάση το RDF μοντέλο. Οι βασικές χωρικές οντότητές του είναι η POI, η οποία αναπαριστά σημαντικά σημεία στην πόλη όπου παρατηρούνται ανθρώπινες δραστηριότητες, η Region, η οποία αναπαριστά χωρικές υποδιαιρέσεις της πόλης, και η Business Area, η οποία αναπαριστά εμπορικά και επιχειρηματικά κέντρα της πόλης. Η χωρική διάσταση αυτών των οντοτήτων περιγράφεται από μία από τις ιδιότητές τους. Επιπλέον, ορίζονται τοπολογικές σχέσεις μεταξύ των χωρικών οντοτήτων, όπως η κοντινή απόσταση μεταξύ δύο περιοχών, η ύπαρξη κοινών συνόρων, και η ύπαρξη ενός σημείου μέσα σε μία περιοχή. Παράλληλα, άλλες οντότητες του UrbanKG, όπως οι User, Brand και Organization, που αναπαριστούν χρήστες, μάρκες και οργανισμούς, αντίστοιχα, συνδέονται μέσω κατηγορημάτων με σημεία και περιοχές της πόλης.

Το KnowWhereGraph [3] είναι ένας πυκνά συνδεδεμένος γράφος γνώσης με δεδομένα ποικίλων τομέων, που περιέχει πληροφορίες για κάθε τοποθεσία στην επιφάνεια της γης, σε υψηλή χωρική και χρονική κλίμακα. Τα δεδομένα που εισάγονται στο KnowWhereGraph αφορούν ακραία γεγονότα, διοικητικά σύνορα, τις καλλιέργειες, το κλίμα, τις μεταφορές, κ.α.. Τα χωρικά δεδομένα αναπαριστώνται ως κελιά ενός ιεραρχικού πλέγματος, του “S2 Grid System” (Discrete Global Grid), προσφέροντας υψηλή χωρική ανάλυση. Στο KnowWhereGraph αναπαριστώνται επίσης περιοχές και οι σχέσεις μεταξύ τους, καθώς επίσης και συνδέσεις μεταξύ γεγονότων και τοποθεσιών. Ο γράφος γνώσης βασίζεται στο RDF μοντέλο και χρησιμοποιεί, μεταξύ άλλων, το πρότυπο GeoSPARQL για την αναπαράσταση των γεωμετριών και των χωρικών σχέσεων.

Όσον αφορά τους υφιστάμενους γράφους που χρησιμοποιούν το LPG μοντέλο, στη διδακτορική διατριβή [4], αναπτύχθηκε μία Neo4j βάση δεδομένων γράφων με δεδομένα από ποικίλες γεωγραφικές βάσεις στο διαδίκτυο, ώστε να συγκριθεί με μία σχεσιακή βάση δεδομένων στην επίδοση των επερωτήσεων. Στη βάση δεδομένων γράφων, οι κόμβοι αναπαριστούν μέρη στον κόσμο και οι σχέσεις αναπαριστούν τοπολογικές ή άλλου είδους σχέσεις μεταξύ τους. Τόσο στους κόμβους όσο και στις σχέσεις περιέχεται ένα σύνολο από ιδιότητες που περιέχουν περισσότερες πληροφορίες για αυτά. Επιπλέον, στο επιστημονικό άρθρο [5], χωρικά και μη χωρικά δεδομένα εισήχθησαν σε ποικίλες αποθήκες σημασιολογικών δεδομένων, συμπεριλαμβανομένου ενός LPG γράφου γνώσης, με σκοπό την εκτίμηση της επίδοσης στον χειρισμό γεωμετριών, τοπολογικών σχέσεων και επερωτήσεων με χωρική σημασιολογία. Για την ανάπτυξη του γράφου γνώσης, χρησιμοποιήθηκε το Neo4j, και για την διευκόλυνση των χωρικών λειτουργιών στα δεδομένα του γράφου, χρησιμοποιήθηκε η Neo4j Spatial Library.

Κεφάλαιο 3

Χωρικά δεδομένα στο SustainGraph

Στο κεφάλαιο αυτό παρουσιάζεται το θεωρητικό υπόβαθρο και η μεθοδολογία που ακολουθήθηκε για την προσθήκη χωρικών δεδομένων στο SustainGraph. Αρχικά περιγράφονται οι στόχοι και η δομή του γράφου γνώσης SustainGraph, ενώ στη συνέχεια περιγράφεται ο τρόπος αναπαράστασης χωρικών δεδομένων σε αυτόν και οι μηχανισμοί εισαγωγής τους, καθώς επίσης και μηχανισμοί προεπεξεργασίας και ανάλυσής τους.

3.1 Εισαγωγή στο SustainGraph

Το SustainGraph [6] είναι ένας γράφος γνώσης που αναπτύχθηκε για την παρακολούθηση και την καταγραφή πληροφοριών σχετικών με την πρόοδο προς την επίτευξη των Στόχων Βιώσιμης Ανάπτυξης (Sustainable Development Goals (SDGs)) που έχουν οριστεί από τον Οργανισμό Ηνωμένων Εθνών. Στόχος του είναι να αποτελέσει μία πηγή γνώσης για πληροφορίες σχετικές με τους SDGs, αξιοποιώντας την τεχνολογία των βάσεων δεδομένων σε μορφή γράφου και τις ποικίλες τεχνικές για τον εμπλουτισμό τους με δεδομένα, την παραγωγή γνώσης από αυτά, και την ανάλυσή τους. Το SustainGraph έχει αναπτυχθεί ως ένας LPG γράφος, χρησιμοποιώντας την τεχνολογία του Neo4j, και είναι διαθέσιμο στο [30].

Τόσο οι SDGs όσο και ποικίλες άλλες πολιτικές, που έχουν οριστεί είτε σε εθνικό είτε σε Ευρωπαϊκό επίπεδο, στοχεύουν στην ανάπτυξη λύσεων σχετικά με τη μείωση των επιπτώσεων της κλιματικής αλλαγής. Σύμφωνα με αυτές τις πολιτικές γίνεται διαθέσιμη μία πληθώρα από δεδομένα που αφορούν την παρακολούθηση της τιμής διαφόρων δεικτών, οι οποίοι δείχνουν την πρόοδο προς την επίτευξη των στόχων που έχουν οριστεί. Τα δεδομένα αυτά συγκεντρώνονται από ποικίλους οργανισμούς σε όλο τον κόσμο, γεγονός που έχει ως αποτέλεσμα αυτά να μην είναι σημασιολογικά συνεπή μεταξύ τους και να μην είναι εύκολη η διαχείρισή τους. Επιπλέον, η πρόσβαση σε αυτά πολλές φορές δεν είναι εύκολη, καθώς επίσης και η ποιότητα των δεδομένων δεν είναι δεδομένη. Ο SustainGraph αναπτύχθηκε, λοιπόν, για να καταγράφει τη σχέση μεταξύ των πολιτικών και των στόχων που έχουν οριστεί σε εθνικό και παγκόσμιο επίπεδο, καθώς και την εξέλιξη των δεικτών στον χρόνο, συγκεντρώνοντας δεδομένα που προέρχονται από πολλές διαφορετικές πηγές και με διαφορετική σημασιολογία, αξιοποιώντας παράλληλα τη δομή γράφου για να αναπαραστήσει γνώση. [6]

Πολιτικές σχετικές με την κλιματική αλλαγή στο SustainGraph

Οι πολιτικές που έχουν οριστεί για την μείωση των επιπτώσεων της κλιματικής αλλαγής και οι οποίες συμπεριλαμβάνονται στο SustainGraph, όπως περιγράφεται στο [6], είναι ποικίλες. Μία από τις πιο βασικές είναι η Συμφωνία του Παρισιού, η οποία υπογράφηκε το 2016 από 196 κράτη και θέτει ως στόχο τη μείωση της εκπομπής αερίων. Τα πλάνα της συμφωνίας περιέχονται σε έγγραφα που ονομάζονται εθνικά καθορισμένες συνεισφορές (nationally determined contributions (NDCs)) και περιέχουν ένα σύνολο στόχων για την προσαρμογή στην κλιματική αλλαγή. Οι SDGs, γύρω από τους οποίους επικεντρώνεται το SustainGraph, είναι 17 στόχοι (goals) που ορίστηκαν από τα Ηνωμένα Έθνη το 2015 στο πλαίσιο της Agenda 2030 για τη Βιώσιμη Ανάπτυξη. Έχουν συνολικά 169 υπο-στόχους (targets), 8 με 12 για κάθε στόχο, και δείκτες οι οποίοι εκφράζουν την πρόοδο προς την επίτευξη κάθε υπο-στόχου. Οι δείκτες που αντιστοιχούν σε κάθε υπο-στόχο είναι ένας με τέσσερις. Για την κατηγοριοποίηση των SDGs έχουν οριστεί από τα Ηνωμένα Έθνη 6 μετασχηματισμοί που είναι απαραίτητοι σε κάθε χώρα για την επίτευξή τους, και σε κάθε μετασχηματισμό έχουν αντιστοιχηθεί συγκεκριμένοι στόχοι.

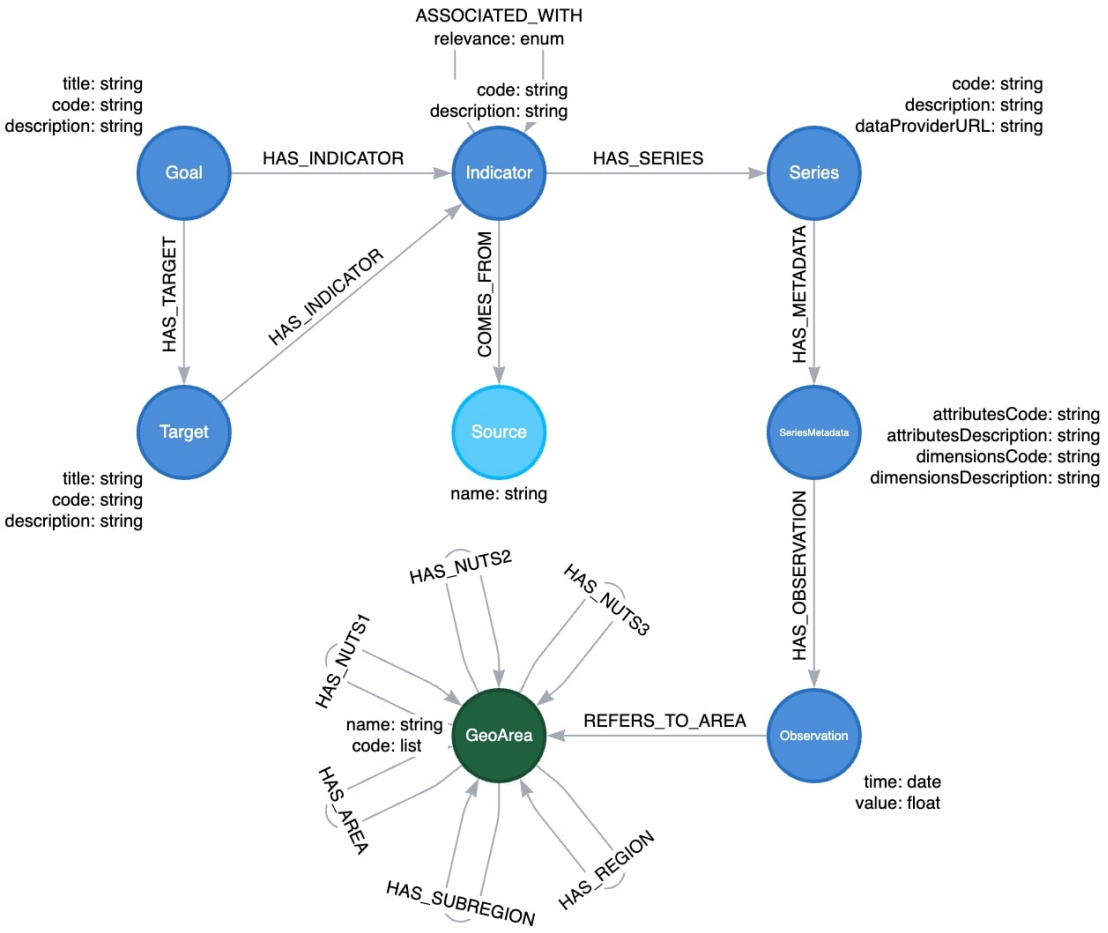
Κάποιες από τις πολιτικές της Ευρωπαικής Ένωσης που είναι σχετικές με την επίτευξη των SDGs, οι οποίες περιέχονται στο SustainGraph, είναι ένα σύνολο πολιτικών σε διάφορους τομείς που ορίστηκαν στην Ευρωπαική Πράσινη Συμφωνία (European Green Deal (EGD). Στόχος της συμφωνίας αυτής αποτελεί η μείωση των εκπομπών αερίου του θερμοκηπίου και η μετάβαση σε μια κλιματική ουδέτερη Ευρώπη έως το 2050. Επιπλέον, έχει οριστεί ένα σύνολο από Ειδικές Συστάσεις ανά Χώρα (Country Specific Recommendations (CSRs)) που έχουν ως στόχο την υιοθέτηση πολιτικών βιώσιμης ανάπτυξης σε εθνικό επίπεδο, καθώς επίσης και η Ταξινομία της Ευρωπαικής Ένωσης για επενδύσεις βιώσιμης ανάπτυξης που προωθούν την πραγματοποίηση της Ευρωπαικής Πράσινης Συμφωνίας. Τέλος, η Ευρωπαική Στατιστική Υπηρεσία (Eurostat) παρακολουθεί 101 δείκτες που αφορούν την πρόοδο προς την επίτευξη των SDGs συγκεκριμένα στις χώρες της Ευρωπαικής Ένωσης.

Δομή του SustainGraph

Το SustainGraph σχεδιάστηκε και υλοποιήθηκε σε μορφή LPG μοντέλου. Στο μοντέλο αυτό, όπως περιγράφηκε στο κεφάλαιο 2, ο γράφος γνώσης αποτελείται από κόμβους και κατευθυνόμενες ακμές που έχουν μία ή περισσότερες ετικέτες καθώς και ένα σύνολο από ιδιότητες που περιγράφουν τα χαρακτηριστικά τους.

Οι βασικές οντότητες στο SustainGraph είναι αυτές που σχετίζονται άμεσα με τους SDGs και αναπαριστώνται από κόμβους του γράφου με τις κατάλληλες ετικέτες. Συγκεκριμένα, κάθε στόχος αναπαρίσταται από έναν κόμβο με ετικέτα *Goal*, οι υπο-στόχοι του από κόμβους *Target*, και οι δείκτες από κόμβους *Indicator*. Οι στόχοι συνδέονται μέσω ακμής με τους υπο-στόχους τους και κάθε υπό-στόχος συνδέεται με έναν ή περισσότερους δείκτες. Κάθε *Indicator* συνδέεται με έναν κόμβο *Series* που αντιπροσωπεύει τη χρονοσειρά με τις τιμές που παίρνει ο δείκτης, ενώ καθε κόμβος *Series* συνδέεται με έναν κόμβο *SeriesMetadata* που περιέχει στις ιδιότητές του επιπλέον πληροφορίες για την αντίστοιχη μετρική. Ο κόμβος αυτός συνδέεται με ένα σύνολο κόμβων τύπου *Observation*. Κάθε *Observation* αναπαριστά μία παρατήρηση της μετρικής και περιέχει δύο ιδιότητες: μία για τη χρονική στιγμή της

παρατήρησης και μία για την τιμή της μετρικής τη συγκεκριμένη χρονική στιγμή. Οι παρατηρήσεις έχουν επίσης και γεωχωρικά χαρακτηριστικά, δηλαδή αφορούν μία συγκεκριμένη γεωγραφική περιοχή. Οι περιοχές αναπαριστούνται στο SustainGraph με τους κόμβους *GeoArea*, με τους οποίους συνδέονται οι κόμβοι *Observation* μέσω ακμής. Οι δείκτες μπορεί να είναι δείκτες που ορίστηκαν από τα Ηνωμένα Έθνη στην Agenda 2030, δείκτες που παρακολουθούνται από την Ευρωπαική Στατιστική Υπηρεσία (Eurostat) σε Ευρωπαικό επίπεδο, καθώς επίσης και δείκτες από εξωτερικές πηγές. Η πηγή των δεικτών ορίζεται στον γράφο γνώσης από την οντότητα *Source*. Τέλος, δείκτες από διαφορετικές πηγές μπορούν να σχετίζονται μεταξύ τους μέσω σχέσεων. [6] Οι οντότητες που περιγράφηκαν καθώς και οι μεταξύ τους σχέσεις φαίνονται στο σχήμα 3.1.



Σχήμα 3.1: *Στόχοι βιώσιμης ανάπτυξης, υπο-στόχοι, δείκτες και παρατηρήσεις μέσα στο SustainGraph*

Ένα άλλο σύνολο οντοτήτων στο SustainGraph αφορά τις πολιτικές που περιγράφηκαν στην υποενότητα *Πολιτικές σχετικές με την κλιματική αλλαγή στο SustainGraph*, και τις μεταξύ τους σχέσεις. Όσον αφορά την Ευρωπαική Πράσινη Συμφωνία, στον γράφο γνώσης συμπεριλαμβάνονται οι οντότητες *EGD* και *Ambitions*, δηλαδή φιλοδοξίες της συμφωνίας

που υλοποιούνται σε συγκεκριμένους τομείς. Οι τομείς αυτοί εκφράζονται από την οντότητα *PolicyArea* και συνδέονται με έναν ή περισσότερους κόμβους *Goal*. Επιπλέον, συμπεριλαμβάνεται η οντότητα *EGDPolicyDocument* που σχετίζεται με τους 6 SDG μετασχηματισμούς. Για τις Ειδικές Συστάσεις ανά Χώρα δημιουργήθηκε η οντότητα *CSRecommendation* η οποία σχετίζεται με κάποιους SDG αλλά και με μία χώρα, δηλαδή έναν κόμβο *GeoArea*. Οι εθνικά καθορισμένες συνεισφορές εκφράζονται μέσω κόμβων με ετικέτα *NDC*, οι οποίοι συνδέονται με συγκεκριμένους υπο-στόχους που πρέπει να επιτευχθούν. Σχετικά με τους έξι μετασχηματισμούς των SDGs, περιέχονται οι οντότητες *Transformation*, *Intervention*, *Ministry* και *IntermediateOutput*, που εκφράζουν τις παρεμβάσεις που πρέπει να πραγματοποιηθούν για κάθε μετασχηματισμό από το αντίστοιχο υπουργείο, καθώς επίσης και τα αναμενόμενα αποτελέσματα από κάθε παρέμβαση. Τα αποτελέσματα αυτά σχετίζονται με κάποιον SDG στόχο. Τέλος, στο SustainGraph περιέχονται οι οντότητες *CaseStudies*, *ClimateHazard*, *Innovation* και *Stakeholder*. Οι συγκεκριμένες οντότητες αναπαριστούν μελέτες που έχουν γίνει στην Ευρώπη με στόχο τη μείωση των επιπτώσεων της κλιματικής αλλαγής σε συγκεκριμένες περιοχές (κόμβοι *GeoArea*), τις καινοτομίες που μπορούν να εφαρμοστούν για κάθε μελέτη, και τους κινδύνους που επιφυλάσσει η κλιματική αλλαγή και καλούνται να αντιμετωπίσουν. [6]

Σχετικά με την αναπαράσταση της γεωχωρικής πληροφορίας, το SustainGraph περιέχει την οντότητα *GeoArea* που σχετίζεται με μία συγκεκριμένη γεωγραφική περιοχή. Η γεωχωρική πληροφορία δηλώνεται με ιεραρχικό τρόπο, με τη χρήση μίας δεύτερης ετικέτας στους κόμβους *GeoArea*. Η ιεραρχία ορίζεται σύμφωνα με το πρότυπο M49 των Ηνωμένων Εθνών, σύμφωνα με το οποίο οι περιοχές χωρίζονται σε ηπείρους (οντότητα *Region*), οι οποίες περιέχουν υπο-περιοχές (οντότητα *SubRegion*), που με τη σειρά τους σχετίζονται με συγκεκριμένες χώρες (οντότητα *Area*). Για τις χώρες της Ευρωπαικής Ένωσης, χρησιμοποιείται η ταξινόμηση σύμφωνα με την κοινή ονοματολογία των εδαφικών στατιστικών μονάδων (Nomenclature of Territorial Units for Statistics (NUTS)) της Ευρωπαικής Στατιστικής Υπηρεσίας, όπου μία χώρα υποδιαιρείται σε μικρότερες περιοχές NUTS επιπέδου 1, 2 και 3. Το επίπεδο 2 είναι υποδιαίρεση του επιπέδου 1 και το επίπεδο 3 υποδιαίρεση του επιπέδου 2. Οι περιοχές αυτές αναπαριστώνται από τις οντότητες *NUTS1*, *NUTS2* και *NUTS3* του SustainGraph. [6]

Εμπλουτισμός του SustainGraph με δεδομένα και δυνατότητες αξιοποίησής του

Για τον εμπλουτισμό του SustainGraph με δεδομένα χρησιμοποιούνται ποικίλοι μηχανισμοί, με τους οποίους δεδομένα που προέρχονται από διαφορετικές πηγές και σε διαφορετική μορφή εισάγονται στον γράφο γνώσης. Η διαδικασία αυτή γίνεται με την ανάπτυξη προσαρμοσμένων script, που αυτοματοποιούν μερικώς ή πλήρως τη διαδικασία της εισαγωγής των δεδομένων που παρέχονται απο κάποιες σημαντικές πηγές δεδομένων. [6] Οι πηγές αυτές αποτελούνται κυρίως από παγκόσμιους οργανισμούς και στατιστικές υπηρεσίες που παρέχουν δεδομένα σε μορφή πίνακα ή μέσω Διεπαφές Προγραμματισμού Εφαρμογών (Application Programming Interfaces (APIs)). Επιπλέον, χρησιμοποιούνται έγγραφα και αναφορές σχετικές με τις πολιτικές που παρουσιάστηκαν στην υπο-ενότητα *Πολιτικές σχετικές με την κλιματική αλλαγή στο SustainGraph*. Αξιοποιώντας τεχνικές μηχανικής μάθησης, και

συγκεκριμένα επεξεργασίας φυσικής γλώσσας (Natural Language Processing (NLP)), εξάγονται πληροφορίες από τα έγγραφα οι οποίες στη συνέχεια εισάγονται στον γράφο γνώσης. Έπειτα, εφαρμόζεται ένα σύνολο μηχανισμών, διαφορετικοί για κάθε πηγή δεδομένων, ώστε να επιτευχθεί η ομοιομορφία των δεδομένων και να διασφαλιστεί η ποιότητά τους. Αυτό γίνεται με ποικίλους τρόπους όπως με την αφαίρεση ακραίων τιμών ή τιμών που δεν πρέπει να συμπεριληφθούν στο SustainGraph, με την συμπλήρωση τιμών που λείπουν από το εκάστοτε dataset όπως και με την προσαρμογή της χρονικής κλίμακας των δεδομένων χρονοσειράς. Παράλληλα, λαμβάνεται υπόψη η πιθανή ύπαρξη προκαταλήψεων στα δεδομένα που αξιοποιούνται, και δίνεται έμφαση στην επαρκή και δίκαιη αντιπροσώπευση όλων των διαφορετικών ομάδων που εμφανίζονται στα dataset. [6]

Όπως περιγράφεται στο [6], η γνώση που περιέχεται στο SustainGraph μπορεί να αξιοποιηθεί με ποικίλους τρόπους. Ένας από αυτούς είναι μέσω της εξερεύνησης των δεδομένων του γράφου με την εφαρμογή επερωτήσεων επί αυτού. Το αποτέλεσμα των επερωτήσεων μπορεί να προσφέρει χρήσιμες πληροφορίες και απαντήσεις σε ερωτήματα των χρηστών καθώς επίσης και να αποτελέσει είσοδο για μηχανισμούς περεταίρω ανάλυσης των δεδομένων. Η εξερεύνηση των δεδομένων μπορεί επίσης να γίνει μέσω της περιήγησης στις οντότητες και τις σχέσεις του SustainGraph, αξιοποιώντας την οπτικοποίησή του σε μορφή γράφου με κόμβους και ακμές. Επιπλέον, υπάρχει η δυνατότητα για οπτικοποίηση των δεδομένων με περιέχονται με διαφορετικούς τρόπους, όπως για παράδειγμα η οπτικοποίηση των δεδομένων χρονοσειράς που μπορεί να δώσει πληροφορίες για την ύπαρξη τάσεων, και η οπτικοποίηση και σύγκριση μετρικών που περιέχονται στον γράφο. Οι αναλύσεις που μπορούν να εφαρμοστούν στα δεδομένα εκτείνονται από εφαρμογή αλγορίθμων σε δεδομένα με μορφή πίνακα, όπως η διερεύνηση συσχέτισης μεταξύ των δεδομένων, ο υπολογισμός ποικίλων στατιστικών μετρικών και η κατηγοριοποίηση των δεδομένων σε ομάδες με κοινά χαρακτηριστικά, μέχρι εφαρμογή αλγορίθμων γράφου, όπως ο υπολογισμός της πυκνότητας του γράφου και ο εντοπισμός κοινοτήτων. Επιπλέον, μπορούν να εφαρμοστούν αλγόριθμοι μηχανικής μάθησης για γράφους για την εισαγωγή νέων σχέσεων και την εφαρμογή αναλύσεων στους κόμβους του γράφου. Η γνώση και οι πληροφορίες που παρέχονται με τους παραπάνω τρόπους στον χρήστη μπορούν να συντελέσουν σε βαθύτερη κατανόηση των δεδομένων καθώς και τη λήψη καλύτερων αποφάσεων σχετικά με την προσπάθεια μείωσης των επιπτώσεων της κλιματικής αλλαγής. [6]

3.2 Αναπαράσταση χωρικών δεδομένων στο SustainGraph

Στην παρούσα ενότητα παρουσιάζεται ο τρόπος αναπαράστασης της χωρικής πληροφορίας σύμφωνα με το υπάρχον σχήμα του SustainGraph, καθώς και η προτεινόμενη επέκτασή του ώστε να δοθεί η δυνατότητα αναπαράστασης γεωμετριών. Παρουσιάζονται, επίσης, το θεωρητικό υπόβαθρο και οι δυνατότητες του Neo4j που αξιοποιούνται για τον σκοπό αυτό.

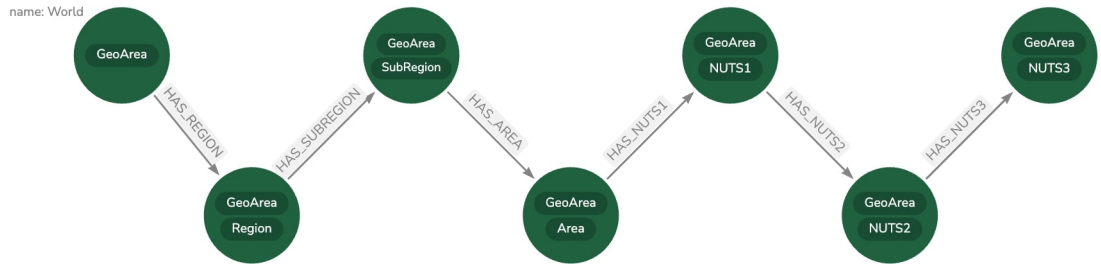
3.2.1 Αναπαράσταση περιοχών μέσω της οντότητας GeoArea

Επί του παρόντος, τα χωρικά χαρακτηριστικά των οντοτήτων δηλώνονται μέσω της συσχέτισης τους με την οντότητα *GeoArea*, η οποία, όπως περιγράφηκε στην ενότητα 3.1,

εκφράζει μία γεωγραφική περιοχή (π.χ. ήπειρο, χώρα, πόλη, περιοχή). Η οντότητα αυτή περιέχει ως ιδιότητες το όνομα της περιοχής που αναπαριστά, και των κωδικό της περιοχής σύμφωνα με το πρότυπο M49 ή την ταξινόμηση NUTS. Παράλληλα, δίνεται έμφαση στην αναπαράσταση της πληροφορίας πως μία περιοχή αποτελεί υποδιαίρεσή μίας άλλης, και συνεπώς περιέχεται μέσα σε αυτή, η οποία εκφράζεται μέσω των σχέσεων που συνδέουν τους διαφορετικούς τύπους *GeoArea* (*Region*, *SubRegion*, *Area*, *NUTS1*, *NUTS2*, *NUTS3*). [6] Οι σχέσεις αυτές είναι οι εξής:

- *HAS_REGION*: Συνδέει τον κόμβο *GeoArea* που αντιπροσωπεύει τον κόσμο με τους κόμβους Region, δηλαδή τις ηπείρους.
- *HAS_SUBREGION*: Συνδέει ένα *Region* με τις υπο-περιοχές του, δηλαδή με κόμβους *SubRegion*.
- *HAS_AREA*: Συνδέει ένα *SubRegion* με κόμβους *Area*, που αναπαριστούν τις χώρες που περιέχονται σε αυτό.
- *HAS_NUTS1*: Συνδέει μία χώρα (*Area*) με τις περιοχές NUTS επιπέδου 1 που περιέχονται σε αυτή.
- *HAS_NUTS2*: Συνδέει μία περιοχή NUTS επιπέδου 1 με τις περιοχές NUTS επιπέδου 2 που περιέχονται σε αυτή.
- *HAS_NUTS3*: Συνδέει μία περιοχή NUTS επιπέδου 2 με τις περιοχές NUTS επιπέδου 3 που περιέχονται σε αυτή.

Οι παραπάνω σχέσεις, οι οποίες συνδέουν κόμβους *GeoArea* μεταξύ τους, φαίνονται στο σχήμα 3.1. Στο 3.2 παρουσιάζονται σχηματικά και οι διαφορετικοί τύποι *GeoArea* που συνδέονται από αυτές τις σχέσεις για την αναπαράσταση εμφωλευμένων περιοχών, ξεκινώντας από τον κόμβο που αναπαριστά τον κόσμο, μέχρι μία περιοχή NUTS επιπέδου 3.



Σχήμα 3.2: *Διαφορετικοί τύποι κόμβων GeoArea στο SustainGraph και οι μεταξύ τους σχέσεις*

Οι οντότητες που έχουν χωρικά χαρακτηριστικά, δηλαδή αφορούν μία τοποθεσία στην επιφάνεια της Γης, όπως οι *NDC*, *CSRecommendation*, *CaseStudy* και *Observation*, συνδέονται με έναν κόμβο *GeoArea* μέσω της σχέσης *REFERS_TO_AREA*. Η σχέση αυτή δεν περιέχει κάποια ιδιότητα στην εσωτερική της δομή.

3.2.2 Αναπαράσταση γεωμετριών

Οι οντότητες του SustainGraph που έχουν χωρικά χαρακτηριστικά μπορούν να χωριστούν σε αυτές που αφορούν πολιτικές και μελέτες (*NDC*, *CSRecommendation*, *CaseStudy*) και σε αυτές που αφορούν παρατηρήσεις από την παρακολούθηση της εξέλιξης των δεικτών (*Observation*). Οι πολιτικές ισχύουν συνήθως σε επίπεδο χώρας ή κάποιας μικρότερης ή ευρύτερης περιοχής η οποία έχει κωδικοποιηθεί σύμφωνα με τα πρότυπα κωδικοποίησης που έχουμε χρησιμοποιήσει, και συνεπώς η χωρική πληροφορία για αυτές αναπαριστάται ικανοποιητικά από την οντότητα *GeoArea*. Το ίδιο ισχύει και για τις σχετικές μελέτες που έχουν πραγματοποιηθεί στην Ευρώπη για την αντιμετώπιση κινδύνων σχετικών με το κλίμα. Ωστόσο οι δείκτες, κυρίως από εξωτερικές πηγές, μπορεί να περιέχουν παρατηρήσεις από μικρότερες περιοχές, οι οποίες δεν αντιστοιχούν σε κάποια από τις κωδικοποιημένες. Προκειμένου να συμπεριληφθούν στον γράφο γνώσης είναι σημαντικό να δοθεί η δυνατότητα αναπαράστασης χωρικών χαρακτηριστικών με τη μορφή γεωμετρίας. Με αυτόν τον τρόπο θα γίνει εφικτή η εισαγωγή δεικτών των οποίων οι παρατηρήσεις αφορούν οποιαδήποτε τοποθεσία, ενισχύοντας τον περεταίρω εμπλουτισμό του SustainGraph με δεδομένα γύρω από την κλιματική αλλαγή. Παράλληλα, το SustainGraph θα υποστηρίζει την αναπαράσταση των παρατηρήσεων με χωρικά χαρακτηριστικά στον χάρτη, και την εφαρμογή χωρικών αναλύσεων σε αυτά.

Δυνατότητες Neo4j για αναπαράσταση γεωμετριών

Η βάση δεδομένων Neo4j, στην οποία έχει υλοποιηθεί το SustainGraph, και η Cypher, η γλώσσα επερωτήσεων του Neo4j, υποστηρίζουν τον τύπο δεδομένων POINT για την αναπαράσταση χωρικών τιμών. Οι τιμές με αυτόν τον τύπο δεδομένων μπορούν να περιέχονται είτε σε κόμβους είτε σε ακμές του γράφου, ως ιδιότητες. [22]

Το POINT εκφράζει την γεωμετρία σημείου. Ένα σημείο περιγράφει μία συγκεκριμένη τοποθεσία στον χώρο που δεν έχει μήκος, πλάτος και ύψος, δηλαδή δεν έχει διαστάσεις. Στο Neo4j ένα σημείο μπορεί να βρίσκεται σε χώρο δύο ή τριών διαστάσεων (2D ή 3D), άρα οι συντεταγμένες του να αποτελούνται από μία διατεταγμένη σειρά από δύο ή τρεις float τιμές. Οι τιμές αυτές συνδέονται με κάποιο σύστημα αναφοράς (Coordinate Reference Systems (CRS)), το οποίο ορίζει τον τρόπο με τον οποίο οι συντεταγμένες του σημείου σχετίζονται με μία τοποθεσία στη Γη [31]. Τα συστήματα αναφοράς που υποστηρίζει το Neo4j είναι δύο γεωγραφικά και δύο καρτεσιανά. Τα γεωγραφικά, τα οποία χρησιμοποιούν μοίρες γεωγραφικού πλάτους και γεωγραφικού μήκους και κάποιες φορές το ύψος για να περιγράψουν μία τοποθεσία [31], είναι το WGS 84 2D και το WGS 84 3D. Τα καρτεσιανά συστήματα αναφοράς, τα οποία ορίζονται από δύο ή τρεις άξονες, είναι το Cartesian 2D και το Cartesian 3D. Μοντελοποιούν σημεία στον ευκλίδειο χώρο δύο και τριών διαστάσεων αντίστοιχα και δεν έχουν προσδιορισμένες μονάδες μέτρησης. [22]

Ο τύπος δεδομένων POINT αποτελείται από δύο τμήματα: τις συντεταγμένες του σημείου και το σύστημα αναφοράς. Οι συντεταγμένες περιέχουν δύο ή τρεις float τιμές, ανάλογα με τις διαστάσεις του χώρου στον οποίο βρίσκεται το σημείο. Για τον ορισμό ενός σημείου χρησιμοποιείται η συνάρτηση point του Neo4j, στην οποία παρέχονται ως ορίσματα οι συντεταγμένες και το σύστημα αναφοράς. Οι συντεταγμένες δηλώνονται χρησιμοποιώντας

τα κλειδιά longitude και latitude για τα σημεία σε γεωγραφικό σύστημα αναφοράς δύο διαστάσεων, και επιπλέον το κλειδί height για τα σημεία σε γεωγραφικό σύστημα αναφοράς τριών διαστάσεων. Χρησιμοποιώντας αυτά τα κλειδιά δεν χρειάζεται να δηλωθεί και το σύστημα αναφοράς, καθώς αυτομάτως θεωρείται πως είναι το WGS 84 2D ή το WGS 84 3D, και δημιουργείται ένα POINT που έχει την τιμή 4326 ή 4979 αντίστοιχα στο πεδίο srid. Εάν όμως χρησιμοποιηθούν τα κλειδιά x, y και z, το γεωγραφικό σύστημα αναφοράς πρέπει να δοθεί ως όρισμα στη συνάρτηση point, είτε με τη συμβολοσειρά 'WGS-84' ή 'WGS-84-3D' στο πεδίο crs, είτε με τον αριθμό 4326 ή 4979 στο πεδίο srid. Για τα σημεία στα καρτεσιανά συστήματα αναφοράς, οι συντεταγμένες δηλώνονται χρησιμοποιώντας τα κλειδιά x, y και z, χωρίς να χρειάζεται να δηλωθεί και το σύστημα αναφοράς, καθώς αυτομάτως θεωρείται πως είναι το Cartesian ή το Cartesian-3D. Έτσι, δημιουργείται ένα POINT που έχει την τιμή 7203 ή 9157 αντίστοιχα στο πεδίο srid. Εάν θέλαμε να δώσουμε και το προαιρετικό όρισμα crs ή srid, τότε αυτό θα είναι η συμβολοσειρά 'cartesian' ή 'cartesian-3D' για το crs, και ο αριθμός 7203 ή 9157 για το srid. Παρακάτω παρουσιάζονται τέσσερα παραδείγματα τιμών με τύπο δεδομένων POINT, ένα για κάθε σύστημα αναφοράς. [22]

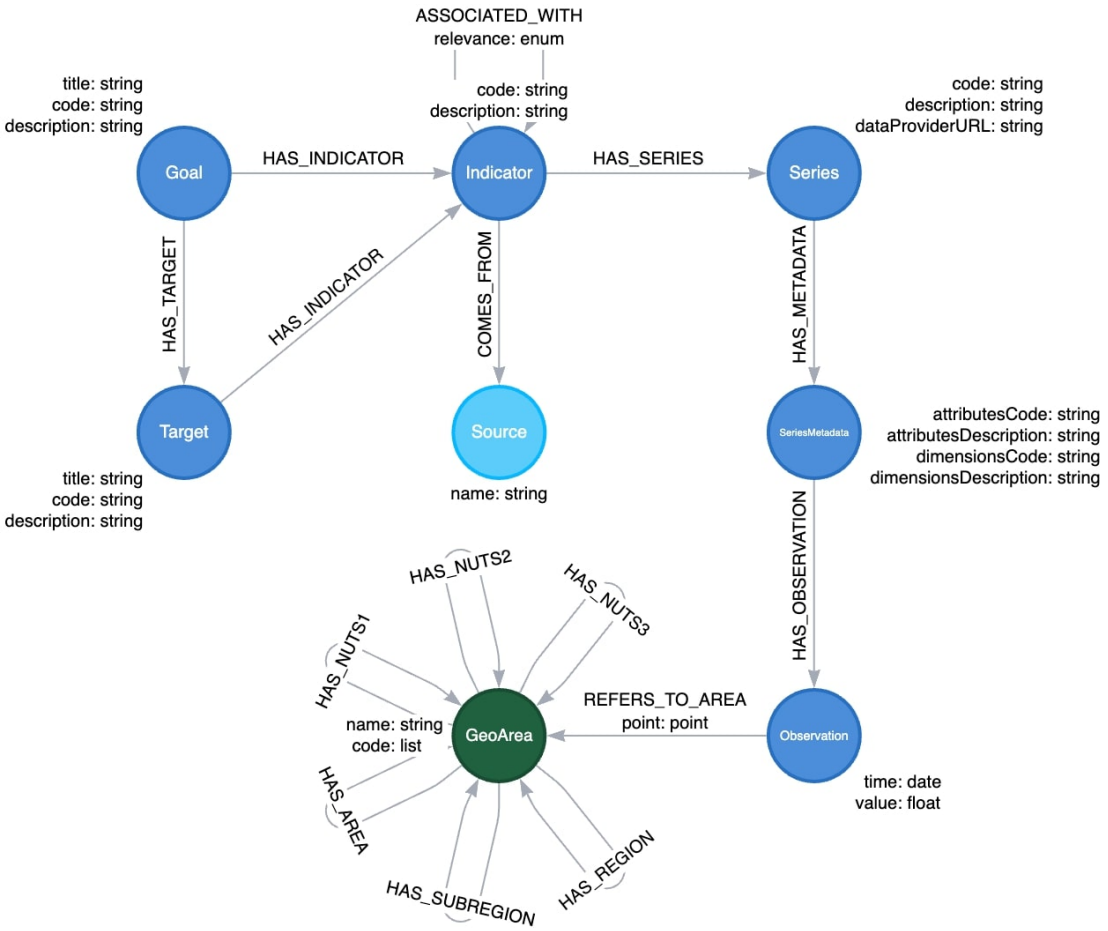
- point(srid:4326, x:53.9, y:14.71) - WGS-84 σύστημα αναφοράς
- point(srid:4979, x:53.9, y:14.71, z:2.0) - WGS-84-3D σύστημα αναφοράς
- point(srid:7203, x:3.6, y:5.2) - Cartesian σύστημα αναφοράς
- point(srid:9157, x:3.6, y:5.2, z:7.0) - Cartesian-3D σύστημα αναφοράς

Το Neo4j παρέχει δύο επιπλέον χωρικές συναρτήσεις, τη distance και τη withinBBox. Η distance υπολογίζει την απόσταση μεταξύ δύο σημείων που βρίσκονται στο ίδιο σύστημα αναφοράς, και η withinBBox υπολογίζει εάν ένα σημείο βρίσκεται χωρικά εντός του ορθογωνίου που ορίζεται από δύο σημεία: το πάνω αριστερά (βορειοδυτικό) και το κάτω δεξιά (νοτιοανατολικό). Το σύστημα αναφοράς με το οποίο συνδέονται τα σημεία πρέπει, και σε αυτή τη συνάρτηση, να είναι το ίδιο. [22]

Ανανεωμένο σχήμα του SustainGraph

Ο τύπος δεδομένων POINT του Neo4j που περιγράφηκε παραπάνω θα αξιοποιηθεί για την αναπαράσταση δεδομένων με γεωμετρία σημείου στο SustainGraph. Οι παρατηρήσεις οι οποίες αφορούν μία τοποθεσία που μπορεί να αναπαρασταθεί με γεωμετρία σημείου θα πρέπει να σχετίζονται με κάποιον τρόπο με αυτήν την πληροφορία στον γράφο γνώσης. Προς το παρόν, οι παρατηρήσεις (οντότητα *Observation*) περιέχουν την ιδιότητα time για τη χρονική στιγμή της παρατήρησης και την ιδιότητα value για την τιμή της μετρικής τη χρονική στιγμή αυτή. Συνδέονται με τη χρονοσειρά στην οποία ανήκουν, και με την περιοχή *GeoArea* την οποία αφορά η παρατήρηση, μέσω της σχέσης *REFERS_TO_AREA*. Η προτεινόμενη αξιοποίηση του τύπου POINT για την εισαγωγή της χωρικής πληροφορίας μίας παρατήρησης με τη μορφή γεωμετρίας είναι η προσθήκη μίας ιδιότητας στη σχέση *REFERS_TO_AREA*, η οποία θα περιέχει το σημείο που αφορά η παρατήρηση ως τιμή τύπου POINT. Η σχέση αυτή

θα συνδέει την παρατήρηση με τη μικρότερη κωδικοποιημένη - από τα πρότυπα γεωγραφικής κωδικοποίησης που λαμβάνει υπόψιν το SustainGraph - περιοχή *GeoArea* στην οποία περιέχεται χωρικά το σημείο, όπως φαίνεται στο σχήμα 3.3.



Σχήμα 3.3: *Στόχοι βιώσιμης ανάπτυξης, υπο-στόχοι, δείκτες και παρατηρήσεις μέσα στο SustainGraph με την ιδιότητα point για την αναπαράσταση της τοποθεσίας των παρατηρήσεων*

3.3 Μηχανισμοί εισαγωγής χωρικών δεδομένων στο SustainGraph

Όπως και για τα υπόλοιπα δεδομένα που εισάγονται στο SustainGraph, έτσι και για τα χωρικά δεδομένα, πρέπει να ακολουθείται μία διαδικασία προεπεξεργασίας τους και διασφάλισης της ποιότητάς τους πριν εισαχθούν στον γράφο γνώσης. Τα δεδομένα αυτά προέρχονται από διαφορετικές πηγές, έχουν διαφορετική σημασιολογία και πολλές φορές μπορεί να περιέχουν ακραίες τιμές που πρέπει να αφαιρεθούν. Επίσης, ένα μέρος τους μπορεί να μη μας ενδιαφέρει να εισαχθούν στον γράφο γνώσης. Ειναι σημαντικό να ακολουθήσουμε μια σειρά απο μηχανισμούς ώστε να τα προετοιμάσουμε κατάλληλα και να τα φέρουμε στην

επιθυμητή μορφή για να τα εισάγουμε στον γράφο γνώσης. Κάποιοι από τους μηχανισμούς εφαρμόζονται και σε μη χωρικά δεδομένα, ενώ άλλοι περιλαμβάνουν πράξεις μεταξύ γεωμετριών. Στην παρούσα ενότητα παρουσιάζεται η μεθοδολογία που ακολουθήθηκε και το απαραίτητο θεωρητικό υπόβαθρο για τους μηχανισμούς που υλοποιήθηκαν.

3.3.1 Μηχανισμοί προεπεξεργασίας δεδομένων σε δομή πίνακα

Οι πηγές δεδομένων που θα αξιοποιηθούν για την εισαγωγή παρατηρήσεων στον γράφο γνώσης παρέχουν τα δεδομένα σε αρχεία NetCDF, μία μορφή που χρησιμοποιείται για την αποθήκευση επιστημονικών δεδομένων πολλαπλών διαστάσεων. Προς διευκόλυνση της προεπεξεργασίας τους, τα δεδομένα φορτώνονται σε δομή πίνακα. Τα δεδομένα σε αυτή τη δομή (tabular data) είναι δομημένα σε γραμμές, η κάθε μία εκ των οποίων περιέχει πληροφορίες για μία παρατήρηση. Κάθε γραμμή περιέχει τον ίδιο αριθμό κελιών, τα οποία περιέχουν τις τιμές των ιδιοτήτων της αντίστοιχης παρατήρησης, ενώ αν για κάποια γραμμή δεν είναι διαθέσιμη η τιμή κάποιας ιδιότητας, το αντίστοιχο κελί μπορεί να έχει null τιμή. Τα κελιά κάθε στήλης αφορούν την ίδια ιδιότητα. [32] Συγκεκριμένα, η δομή που χρησιμοποιείται είναι το DataFrame.

Το πρώτο βήμα για την προετοιμασία των δεδομένων των παρατηρήσεων, είναι η κατανόηση των πληροφοριών που περιέχουν. Εξετάζεται το πλήθος των παρατηρήσεων, ποιες ιδιότητες περιέχονται για τις παρατηρήσεις, το πλήθος των κενών κελιών και σε ποιες στήλες παρατηρούνται. Παράλληλα, εξετάζεται η χρονική κλίμακα, δηλαδή ποιο είναι το χρονικό διάστημα μεταξύ δύο διαδοχικών παρατηρήσεων, και η μορφή με την οποία αναπαριστώνται οι συντεταγμένες των παρατηρήσεων, δηλαδή σε πόσες και ποιες στήλες περιέχονται, με ποια μορφή, και σε ποιο σύστημα αναφοράς.

Η διαχείριση των δεδομένων σε δομή πίνακα καθιστά εύκολη την προεπεξεργασία τους και την προετοιμασία τους μέχρι την εισαγωγή τους στον γράφο γνώσης. Δίνεται η δυνατότητα αφαίρεσης στηλών οι οποίες περιέχουν ιδιότητες που δεν είναι χρήσιμες για τον γράφο γνώσης, είτε αφαίρεσης γραμμών που δεν ικανοποιούν κάποια επιθυμητή συνθήκη. Μία περίπτωση στην οποία επιθυμείται η αφαίρεση ορισμένων γραμμών είναι η ύπαρξη ακραίων τιμών (outliers) σε κάποια ιδιότητα των δεδομένων, οι οποίες φιλτράρονται και δεν συμπεριλαμβάνονται στο τελικό dataset. Ένας από τους μηχανισμούς που εφαρμόζονται για την ανίχνευση ακραίων τιμών στα δεδομένα είναι η μέθοδος του ενδοτεταρτημοριακού εύρους (Interquartile Range (IQR)).

Το ενδοτεταρτημοριακό εύρος αποτελεί μέτρο διασποράς των δεδομένων. Τα δεδομένα, αφότου ταξινομηθούν σε αύξουσα σειρά, χωρίζονται σε τέσσερα ίσα μέρη, και η τιμές που χωρίζουν τα δεδομένα σε αυτά ονομάζονται τεταρτημόρια. Το πρώτο τεταρτημόριο συμβολίζεται ως Q_1 και αναπαριστά το 25ο εκατοστημόριο των δεδομένων, δηλαδή το 25% των δεδομένων έχουν τιμή μικρότερη από αυτό. Το δεύτερο τεταρτημόριο συμβολίζεται ως Q_2 και αναπαριστά το 50ο εκατοστημόριο, και το τρίτο τεταρτημόριο συμβολίζεται ως Q_3 και αναπαριστά το 75ο εκατοστημόριο των δεδομένων. Το ενδοτεταρτημοριακό εύρος είναι το εύρος μεταξύ του πρώτου και του τρίτου τεταρτημορίου, δηλαδή ισχύει IQR = Q_3 - Q_1. Τα δεδομένα που θεωρούνται ακραίες τιμές με βάση αυτήν τη μέθοδο είναι όσα έχουν τιμή μικρότερη από Q_1 - 1.5IQR ή μεγαλύτερη από Q_3 + 1.5IQR. [33]

Μία άλλη δυνατότητα που δίνεται για την επεξεργασία των δεδομένων σε δομή πίνακα, είναι η ομαδοποίησή τους με βάση μία ή περισσότερες ιδιότητες. Κατά την ομαδοποίηση, οι γραμμές που έχουν κοινές τιμές στις επιθυμητές στήλες σχηματίζουν μία ομάδα. Οι γραμμές που περιέχονται σε κάθε ομάδα μπορούν να συναθροιστούν σε μία γραμμή που θα περιέχει τις κοινές τιμές, και η τιμή για κάθε μία από τις υπόλοιπες ιδιότητές της μπορεί να υπολογιστεί εφαρμόζοντας κάποια συναθροιστική συνάρτηση επί των τιμών της αντίστοιχης ιδιότητας όλων των γραμμών που συναθροίστηκαν (π.χ. μέση τιμή, άθροισμα, ελάχιστη τιμή, μέγιστη τιμή). Μία περίπτωση στην οποία χρειάζεται να εφαρμοστεί η ομαδοποίηση στα δεδομένα των παρατηρήσεων, είναι η αύξηση της χρονικής κλίμακας, για την οποία απαιτείται η ομαδοποίηση των παρατηρήσεων που έχουν κοινή κάποια χρονική μεταβλητή, όπως ημέρα, εβδομάδα ή μήνα.

Επιπλέον, υπάρχει η δυνατότητα σύζευξης μεταξύ δύο πινάκων. Ένας τύπος σύζευξης που θα χρησιμοποιηθεί είναι η αριστερή εξωτερική σύζευξη, με την οποία συνδυάζονται οι γραμμές των πινάκων που έχουν κοινή τιμή στις επιθυμητές στήλες. Προκύπτει ένας νέος πίνακας που περιέχει όλες τις γραμμές του πρώτου πίνακα, και σε κάθε γραμμή περιέχει επιπλέον τις ιδιότητες των σειρών που αντιστοιχήθηκαν από τον δεύτερο πίνακα. Στις γραμμές στις οποίες δεν βρέθηκε αντιστοίχιση, οι αντίστοιχες ιδιότητες παίρνουν την τιμή null. Η συγκεκριμένη πράξη μεταξύ πινάκων μπορεί να εφαρμοστεί για την σύζευξη του πίνακα των παρατηρήσεων με έναν πίνακα που περιέχει επιπλέον πληροφορίες για κάθε σημείο στο οποίο υπάρχει κάποια παρατήρηση. Η σύζευξη σε αυτήν την περίπτωση γίνεται επί των στηλών που περιέχουν τις συντεταγμένες των σημείων.

3.3.2 Μηχανισμοί προεπεξεργασίας χωρικών δεδομένων

Για την κατάλληλη προεπεξεργασία των χωρικών δεδομένων, είναι απαραίτητο να αναπαρασταθεί η γεωμετρία τους πριν από την εισαγωγή τους στον γράφο γνώσης όπου, όπως περιγράφηκε στην ενότητα 3.2, θα αξιοποιηθεί ο τύπος δεδομένων POINT του Neo4j. Ο λόγος για τον οποίο αυτό είναι σημαντικό είναι διττός. Πρώτον, θα δοθεί η δυνατότητα οπτικοποίησης των δεδομένων στον χάρτη σε όλη την πορεία της προεπεξεργασίας τους, από την άντλησή τους από τις πηγές δεδομένων, έως την εισαγωγή τους στον γράφο γνώσης, διευκολύνοντας την καλύτερη κατανόηση των πληροφοριών που περιέχουν και των ενεργειών που πρέπει να εφαρμοστούν σε αυτά. Δεύτερον, θα γίνει εφικτή η εφαρμογή πράξεων μεταξύ των γεωμετριών, που είναι απαραίτητες για την κατάλληλη προετοιμασία των χωρικών δεδομένων για το SustainGraph.

Χρησιμοποιούμενο μοντέλο χωρικών δεδομένων

Το μοντέλο που θα χρησιμοποιηθεί για την αναπαράσταση των γεωμετριών των δεδομένων και τις πράξεις μεταξύ τους βασίζεται στο μοντέλο Simple Features του Open Geospatial Consortium [34]. Οι γεωμετρίες που αναπαριστώνται είναι οι εξής:

- Point: Τύπος γεωμετρίας που αναπαριστά ένα σημείο, δηλαδή ένα γεωμετρικό αντικείμενο που δεν έχει διαστάσεις και εκφράζει μία συγκεκριμένη τοποθεσία στον χώρο. Οι συντεταγμένες του αποτελούνται από δύο ή τρεις τιμές.

- LineString: Τύπος γεωμετρίας που αποτελείται από ένα ή περισσότερα ευθύγραμμα τμήματα. Κάθε ευθύγραμμο τμήμα ορίζεται από δύο σημεία.
- LinearRing: LineString που κλείνει και τα ευθύγραμμά τμήματά του δεν τέμνονται.
- Polygon: Τύπος γεωμετρίας που αναπαριστά ένα πολύγωνο, δηλαδή μία περιοχή που περιβάλλεται από ένα LinearRing.
- MultiPoint: Συλλογή από ένα ή περισσότερα σημεία (Points).
- MultiLineString: Συλλογή από ένα ή περισσότερα LineStrings.
- MultiPolygon: Συλλογή από ένα ή περισσότερα πολύγωνα (Polygons).
- GeometryCollection: Συλλογή από μία ή περισσότερες γεωμετρίες, ίδιου ή διαφορετικού τύπου.

Το παραπάνω μοντέλο συνοδεύεται από κατηγορήματα, τα οποία εκφράζουν χωρικές σχέσεις μεταξύ δύο γεωμετριών, και από πράξεις μεταξύ των γεωμετριών. Κάποια από τα κατηγορήματα είναι τα intersects, touches, disjoint, crosses, within, contains, overlaps, equals, covers. Το κατηγόρημα που θα χρησιμοποιηθεί περισσότερο είναι το intersects, το οποίο εκφράζει τη σχέση μεταξύ δύο γεωμετρικών αντικειμένων που τέμνονται χωρικά, δηλαδή μοιράζονται ένα μέρος του χώρου που καταλαμβάνουν. Κάποιες από τις πράξεις που ορίζονται είναι οι union, intersection, difference, convex_hull, envelope, buffer, centroid. Δύο πράξεις που θα αξιοποιηθούν είναι η union, η οποία συγχωνεύει δύο γεωμετρίες σε μία, και η centroid, η οποία υπολογίζει το γεωμετρικό κέντρο (κεντροειδές) μίας γεωμετρίας, και επιστρέφει το αντίστοιχο Point.

Χωρικά δεδομένα σε δομή πίνακα

Το παραπάνω μοντέλο χωρικών δεδομένων χρησιμοποιείται για την περιγραφή των χωρικών χαρακτηριστικών των δεδομένων που βρίσκονται σε δομή πίνακα. Αυτό επιτυγχάνεται δημιουργώντας ένα GeoDataFrame, μία νέα δομή πίνακα η οποία περιέχει και χωρικά δεδομένα. Δημιουργείται ορίζοντας μία νέα στήλη στον προϋπάρχων πίνακα, η οποία περιέχει τις γεωμετρίες των γραμμών. Στη στήλη αυτή εφαρμόζονται οι χωρικές μεθόδοι που εφαρμόζονται στα δεδομένα.

Μία από τις μεθόδους που μπορούμε να εφαρμόσουμε είναι η ομαδοποίηση χωρικών δεδομένων. Όπως και στην ομαδοποίηση μη χωρικών δεδομένων που περιγράφηκε στην υποενότητα 3.3.1, η ομαδοποίηση γίνεται με βάση μία ή περισσότερες ιδιότητες, οι οποίες θα έχουν κοινή τιμή σε κάθε ομάδα που δημιουργείται. Παράλληλα όμως με την συνάθροιση των τιμών κάθε στήλης σε κάθε ομάδα σύμφωνα με μία συναθροιστική συνάρτηση, συναθροίζονται και οι γεωμετρίες κάθε ομάδας σε μία. Η συνάθροιση των γεωμετριών γίνεται με την εφαρμογή της πράξης union σε αυτές, κατά την οποία υπολογίζεται η ένωσή τους.

Επιπλέον, αντίστοιχα με τη σύζευξη δύο πινάκων μη χωρικών δεδομένων, ορίζεται η χωρική σύζευξη δύο πινάκων με χωρικά δεδομένα. Στην χωρική σύζευξη, δύο γεωμετρίες συγχωνεύονται βάσει της μεταξύ τους χωρικής σχέσης. Η χωρική σχέση ορίζεται από το κατηγόρημα που επιλέγεται, το οποίο μπορεί να είναι ένα από τα intersects, touches, crosses,

within, contains, overlaps. Για παράδειγμα, στη χωρική σύζευξη με κατηγόρημα intersects, οι γραμμές των δύο πινάκων που θα συγχωνευτούν, είναι αυτές των οποίων οι γεωμετρίες τέμνονται χωρικά. Οι γεωμετρίες που προκύπτουν στον νέο πίνακα είναι αυτές ενός από τους δύο πίνακες που συγχωνεύτηκαν, ανάλογα με το είδος της σύζευξης (αριστερή εξωτερική, δεξιά εξωτερική, εσωτερική).

Η χωρική σύζευξη είναι ιδιαίτερα χρήσιμη για την προεπεξεργασία των χωρικών δεδομένων. Εφαρμόζοντάς τη στον πίνακα με τις παρατηρήσεις, οι οποίες αφορούν σημεία στον χώρο, και σε έναν πίνακα που περιέχει τις γεωμετρίες περιοχών (πολύγωνα), πληροφορούμαστε για την ευρύτερη περιοχή την οποία αφορά κάθε παρατήρηση. Με αυτόν τον τρόπο, γίνεται εφικτή η αντιστοίχιση κάθε παρατήρησης στην κατάλληλη κωδικοποιημένη περιοχή *GeoArea*, όπως σε κάποια από τις περιοχές NUTS.

3.4 Μηχανισμοί μείωσης όγκου χωρικών δεδομένων στο SustainGraph

Στις προηγούμενες ενότητες περιγράφηκε ο τρόπος αναπαράστασης και προεπεξεργασίας των χωρικών δεδομένων με γεωμετρία σημείου. Συχνά, λόγω της υψηλής χωρικής ανάλυσης των παρατηρήσεων που προσφέρονται από μία πηγή δεδομένων, το πλήθος των σημείων για την παρατήρηση ενός δείκτη σε μία ευρεία περιοχή (π.χ. μία πόλη) μπορεί να είναι πολύ μεγάλο. Τα δεδομένα, μάλιστα, δεν έχουν μόνο χωρική διάσταση, αλλά και χρονική, και - ανάλογα με τη συχνότητα και το συνολικό χρονικό διάστημα των παρατηρήσεων - το πλήθος τους αυξάνεται ακόμα περισσότερο. Στην παρούσα ενότητα θα παρουσιάσουμε πιθανούς μηχανισμούς για την μείωση του όγκου των παρατηρήσεων μειώνοντας τα σημεία στα οποία αυτές αναφέρονται, καθώς και το απαραίτητο θεωρητικό υπόβαθρο.

3.4.1 Ομαδοποίηση σημείων σε περιοχές GeoArea

Ένας από τους πιθανούς τρόπους μείωσης του πλήθους των σημείων τα οποία αφορούν οι παρατηρήσεις, είναι η ομαδοποίηση των σημείων ανάλογα με την ευρύτερη περιοχή στην οποία ανήκουν. Οι περιοχές που επιλέγονται είναι αυτές που έχουν κωδικοποιηθεί από τα πρότυπα που χρησιμοποιούνται από το SustainGraph, όπως οι περιοχές NUTS, και οι οποίες αναπαριστώνται σε αυτό από κόμβους *GeoArea*. Η ομαδοποίηση των δεδομένων μπορεί να πραγματοποιηθεί με τη χρήση της μεθόδου χωρικής ομαδοποίησης σε χωρικά δεδομένα που βρίσκονται σε δομή πίνακα, η οποία περιγράφηκε στην υποενότητα 3.3.2. Έχοντας ήδη την πληροφορία της περιοχής στην οποία περιέχονται τα σημεία, έπειτα από την χωρική σύζευξη που εφαρμόστηκε με τον πίνακα που περιέχει τις γεωμετρίες ευρύτερων περιοχών κατά την προεπεξεργασία τους, μπορούμε, με βάση αυτό το χαρακτηριστικό τους, να τα ομαδοποιήσουμε. Η συναθροιστική συνάρτηση που επιλέγεται είναι η μέση τιμή, ώστε να υπολογιστεί η μέση τιμή της μετρικής των παρατηρήσεων, για κάθε ομάδα. Παράλληλα, η γεωμετρία που προκύπτει για κάθε ομάδα υπολογίζεται συναθροίζοντας τις γεωμετρίες της, όπως περιγράφηκε στην υποενότητα 3.3.2. Η συνάθροιση των σημείων (γεωμετρία Point) δίνει ως αποτέλεσμα μία συλλογή σημείων (γεωμετρία MultiPoint), από την οποία μπορεί να υπολογιστεί το κεντροειδές με την εφαρμογή της πράξης centroid.

3.4.2 Ομαδοποίηση με χρήση πλέγματος

Η ομαδοποίηση των σημείων σε ευρύτερες περιοχές, όπως π.χ. σε επίπεδο πόλης, μειώνει σε μεγάλο βαθμό την χωρική ανάλυση της πληροφορίας. Προκειμένου να διατηρηθεί μεγαλύτερο αριθμό σημείων σε κάθε περιοχή, ένας εναλλακτικός μηχανισμός μείωσης τους είναι μέσω της ομαδοποίησής τους με χρήση πλέγματος. Ομαδοποιούνται, δηλαδή, σημεία που περιέχονται στα τετράγωνα κελιά πλέγματος με διαστάσεις της επιλογής μας, και οι ομάδες εκπροσωπούνται από τα κεντροειδή τους, όπως στον προηγούμενο μηχανισμό. Στη συνέχεια περιγράφεται η μεθοδολογία που ακολουθείται.

Αρχικά δημιουργείται ένα σύνολο από πολύγωνα με σχήμα τετραγώνου, τα οποία σχηματίζουν ένα πλέγμα που περιέχει όλα τα σημεία για τα οποία υπάρχουν διαθέσιμες παρατηρήσεις. Αυτό πετυχαίνεται βρίσκοντας, αρχικά, τα όρια όλων των γεωμετριών (δηλαδή των σημείων) που περιέχονται στον πίνακα των παρατηρήσεων. Έπειτα, δημιουργείται μία λίστα από γεωγραφικά μήκη, ξεκινώντας από το μικρότερο όριο και φτάνοντας στο μέγιστο, με βήμα το επιθυμητό μήκος πλευράς των κελιών, καθώς και μία λίστα από γεωγραφικά πλάτη, ακολουθώντας την αντίστοιχη διαδικασία. Στη συνέχεια, χρησιμοποιούνται οι δύο λίστες για να δημιουργηθούν όλα τα δυνατά πολύγωνα που σχηματίζονται από τα τέσσερα ευθύγραμμα τμήματα που έχουν ως άκρα δύο σημεία με κοινό γεωγραφικό πλάτος και διαδοχικά γεωγραφικά μήκη, ή κοινό γεωγραφικό μήκος και διαδοχικά γεωγραφικά πλάτη. Αποθηκεύοντας τα πολύγωνα αυτά σε μία δομή πίνακα, έχει δημιουργηθεί ένα πλέγμα από τετράγωνα κελιά επιθυμητών διαστάσεων, τα οποία περιέχουν όλα τα σημεία των παρατηρήσεων.

Στη συνέχεια χρησιμοποιείται το πλέγμα από τετράγωνα για να χωριστούν τα σημεία των παρατηρήσεων σε ομάδες με βάση το τετράγωνο στο οποίο περιέχονται χωρικά. Πρώτα εφαρμόζεται χωρική σύζευξη μεταξύ του πίνακα με τις παρατηρήσεις, οι οποίες έχουν τύπο γεωμετρίας Point, και του πίνακα με τα τετράγωνα κελιά, τα οποία έχουν τύπο γεωμετρίας Polygon. Το είδος χωρικής σύζευξης που χρησιμοποιείται είναι η αριστερή εξωτερική, και το κατηγόρημα που επιλέγεται είναι το intersects. Από αυτήν την πράξη προκύπτει ένας πίνακας με τις αρχικές παρατηρήσεις και την γεωμετρία τους αλλά και με μία επιπλέον στήλη που περιέχει το αναγνωριστικό του πολυγώνου με το οποίο τέμνεται χωρικά κάθε σημείο, δηλαδή του πολυγώνου στο οποίο περιέχεται. Έπειτα γίνεται η χωρική ομαδοποίηση των σημείων κάθε ομάδας, με παρόμοιο τρόπο που έγινε και στον μηχανισμό της ομαδοποίησης σε περιοχές. Στην συγκεκριμένη περίπτωση, τα χαρακτηριστικά με βάση τα οποία γίνεται η ομαδοποίηση είναι το αναγνωριστικό των πολυγώνων και το αναγνωριστικό της περιοχής *GeoArea* που χρησιμοποιήθηκε και στον προηγούμενο μηχανισμό. Η συναθροιστική συνάρτηση είναι πάλι η μέση τιμή. Από τη συλλογή σημείων που προκύπτει για κάθε ομάδα υπολογίζεται το κεντροειδές της μέσω της πράξης centroid.

Η δυνατότητα επιλογής του επιθυμητού μήκους πλευράς των κελιών, προσφέρει μεγαλύτερο έλεγχο στην επιλογή του πλήθους των σημείων που προκύπτουν, στην χωρική ανάλυση της πληροφορίας και στο πλήθος των κόμβων που θα εισαχθούν τελικά στον γράφο γνώσης. Η μεθοδολογία που θα ακολουθηθεί για την επιλογή του επιθυμητού μήκους πλευράς είναι η εφαρμογή της ομαδοποίησης χρησιμοποιώντας διαφορετικά πλέγματα με αυξανόμενο μέγεθος κελιού, και η σύγκριση των αποτελεσμάτων με ποικίλους τρόπους. Βασικό κριτήριο επιλογής αποτελεί το τελικό πλήθος των σημείων και το εάν μειώνεται σε ικανοποι-

ητικό βαθμό ο όγκος των παρατηρήσεων. Για να ελεγθεί αυτό υπολογίζεται το πλήθος των δεδομένων μετά από κάθε ομαδοποίηση και το ποσοστό μείωσης τους από τα αρχικά. Από την άλλη πλευρά, όσο αυξάνεται το μήκος πλευράς των τετράγωνων κελιών του πλέγματος, τόσο πιο πολλά σημεία ομαδοποιούνται σε κάθε ομάδα, γεγονός που οδηγεί σε απώλεια πληροφορίας για μικρότερες περιοχές. Ο έλεγχος αυτού του φαινομένου για την ομαδοποίηση με κάθε χρησιμοποιούμενο πλέγμα γίνεται με δύο τρόπους. Αρχικά, οπτικοποιούνται σε χάρτη οι παρατηρήσεις τόσο πριν όσο και μετά από κάθε ομαδοποίηση, δίνοντας τη δυνατότητα της ποιοτικής σύγκρισης του κάθε χάρτη που προκύπτει με τον αρχικό. Επειδή οι παρατηρήσεις έχουν και χρονική διάσταση πέρα από χωρική, η οπτικοποίηση γίνεται ενδεικτικά για κάποιες χρονικές στιγμές. Ο δεύτερος είναι μέσω του σχεδιασμού της κατανομής των τιμών στα αρχικά σημεία αλλά και των τιμών μετά από κάθε ομαδοποίηση, για μία συγκεκριμένη χρονική στιγμή. Έπειτα παρατηρείται η διαφορά της κατανομής των ομαδοποιημένων σημείων κάθε πλέγματος από την αρχική κατανομή. Οι κατανομές αυτές σχεδιάζονται για κάθε περιοχή ξεχωριστά. Μπορούν επίσης να υπολογιστούν και να συγκριθούν στατιστικά μεγέθη όπως η μέση τιμή, η τυπική απόκλιση, η ελάχιστη και η μέγιστη τιμή. Τέλος, υπολογίζεται ένας δείκτης χωρικής αυτοσυσχέτισης των χωρικών δεδομένων, τόσο των αρχικών, όσο και των ομαδοποιημένων, και συγκρίνονται τα αποτελέσματα.

Χωρικά βάρη

Για τον υπολογισμό της χωρικής αυτοσυσχέτισης των χωρικών δεδομένων, απαιτείται πρώτα ο υπολογισμός των χωρικών βαρών των σημείων. Τα χωρικά βάρη αποτελούν έναν τρόπο αναπαράστασης γράφων στην επιστήμη γεωχωρικών δεδομένων. Χρησιμοποιούνται για την αναπαράσταση των χωρικών σχέσεων μεταξύ των παρατηρήσεων μέσα σε ένα dataset με χωρικά δεδομένα, εκφράζοντας τις έννοιες της εγγύτητας και της συνδεσιμότητας. [35] Συγκεκριμένα, τα βάρη εκφράζουν τη γειτνίαση μεταξύ των παρατηρήσεων ως έναν πίνακα W μεγέθους n x n, όπου n είναι το πλήθος των παρατηρήσεων και τα στοιχεία w_{ij} του πίνακα είναι τα χωρικά βάρη μεταξύ τους. Ο πίνακας είναι ο εξής:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

Τα χωρικά βάρη w_{ij} είναι μη μηδενικά όταν οι παρατηρήσεις i και j είναι γείτονες, και μηδενικά στην αντίθετη περίπτωση. Στην απλούστερη μορφή του πίνακα χωρικών βαρών, τα μη μηδενικά βάρη έχουν την τιμή 1. Κατά σύμβαση θεωρείται πως μία παρατήρηση δεν συνορεύει με τον εαυτό της, συνεπώς τα στοιχεία στη διαγώνειο του W έχουν μηδενική τιμή. [36]

Τα χωρικά βάρη μεταξύ των παρατηρήσεων μπορούν να υπολογιστούν με πολλούς διαφορετικούς τρόπους. Αυτοί που χρησιμοποιούνται πιο συχνά χωρίζονται σε δύο κατηγορίες: σε αυτούς που λαμβάνουν υπόψη τις σχέσεις γειτνίασης μεταξύ των παρατηρήσεων, και σε αυτούς λαμβάνουν υπόψη τις σχέσεις που βασίζονται στην απόσταση μεταξύ των παρατηρήσεων. Τα βάρη που υπολογίζονται με βάση τις σχέσεις γειτνίασης υπολογίζονται συνήθως

για παρατηρήσεις που έχουν γεωμετρία πολυγώνου. Σε αυτήν την προσέγγιση θεωρείται πως δύο παρατηρήσεις είναι γείτονες, και συνεπώς έχουν μη μηδενικά βάρη, εάν μοιράζονται ένα κοινό σύνορο. Η έννοια του συνόρου ερμηνεύεται διαφορετικά, ανάλογα με το κριτήριο γειτνίασης που επιλέγεται. Σύμφωνα με το κριτήριο του πύργου, δύο πολύγωνα είναι γείτονες εάν έχουν μία κοινή πλευρά, ενώ σύμφωνα με το κριτήριο της βασίλισσας, δύο πολύγωνα είναι γείτονες εάν έχουν μία κοινή πλευρά ή ένα κοινό σημείο. [36]

Από την άλλη πλευρά, τα χωρικά βάρη που υπολογίζονται με βάση την απόσταση μεταξύ των χωρικών παρατηρήσεων χρησιμοποιούνται συνήθως για παρατηρήσεις με γεωμετρία σημείου. Για τον υπολογισμό αυτού του είδους των χωρικών βαρών απαιτείται ο υπολογισμός των αποστάσεων μεταξύ όλων των ζευγαριών παρατηρήσεων, και ο ορισμός των σχέσεων γειτνίασης ως συνάρτηση αυτών των αποστάσεων. Ένας τύπος χωρικών βαρών που βασίζονται στην απόσταση μεταξύ των χωρικών παρατηρήσεων είναι τα βάρη των K-κοντινότερων γειτόνων, όπου κάθε χωρική παρατήρηση έχει ως γείτονες τις K παρατηρήσεις που έχουν τη μικρότερη απόσταση από αυτήν. Ένας άλλος, συχνά χρησιμοποιούμενος τύπος χωρικών βαρών που βασίζονται στην απόσταση, είναι τα βάρη πυρήνα , τα οποία εκφράζουν την χωρική εγγύτητα μεταξύ των χωρικών παρατηρήσεων, η οποία μειώνεται όσο αυξάνεται η απόσταση. Για τον υπολογισμό τους χρησιμοποιείται μία συνάρτηση πυρήνα που ορίζει τον τρόπο με τον οποίο υπολογίζεται κάθε χωρικό βάρος από την αντίστοιχη απόσταση, και το εύρος ζώνης, το οποίο ορίζει από ποια απόσταση και έπειτα δύο παρατηρήσεις θεωρείται πως δεν είναι γείτονες και συνεπώς έχουν μηδενικό χωρικό βάρος μεταξύ τους. [35]

Χωρική αυτοσυσχέτιση και συντελεστής I του Moran

Η έννοια της χωρικής αυτοσυσχέτισης εξετάζει την ύπαρξη συσχέτισης μεταξύ των χωρικών παρατηρήσεων, δηλαδή τον βαθμό στον οποίο οι παρατηρήσεις δεν είναι ανεξάρτητες μεταξύ τους, αλλά η ομοιότητα των τιμών τους σχετίζεται με την ομοιότητα των τοποθεσιών τους στον χώρο. Η χωρική αυτοσυσχέτιση των παρατηρήσεων που περιέχονται σε ένα σύνολο χωρικών δεδομένων υπολογίζεται είτε σε καθολικό επίπεδο είτε σε τοπικό. Η καθολική χωρική αυτοσυσχέτιση εκφράζει την συνολική τάση συσχέτισης μεταξύ των χωρικών παρατηρήσεων και τον βαθμό στον οποίο παρατηρήσεις με κοντινές τιμές είναι κοντινές χωρικά, ενώ η τοπική χωρική αυτοσυσχέτιση εκφράζει την ύπαρξη περιοχών που αποκλίνουν από την συνολική τάση χωρικής αυτοσυσχέτισης, στις οποίες οι παρατηρήσεις εμφανίζουν πιο ισχυρή ή λιγότερο ισχυρή συσχέτιση. [35]

Ο συντελεστής I προτάθηκε από τον Moran ως μέτρο χωρικής αυτοσυσχέτισης χωρικών παρατηρήσεων και ορίζεται ως:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

όπου n είναι ο αριθμός των χωρικών παρατηρήσεων, w_i η τιμή της μεταβλητής που μας ενδιαφέρει στην τοποθεσία i, $\overline{x}$ η μέση τιμή της x και w_{ij} το αντίστοιχο χωρικό βάρος από τον πίνακα W με τα χωρικά βάρη μεταξύ των παρατηρήσεων. [37]

Η αναμενόμενη τιμή του συντελεστή I του Moran εάν δεν υπάρχει χωρική αυτοσυσχέτιση στα χωρικά δεδομένα είναι $E(I) = \frac{-1}{n-1}$. Φαίνεται, λοιπόν, πως όσο το πλήθος των χωρικών

παρατηρήσεων αυξάνεται, η αναμενόμενη τιμή τείνει στο 0. Οι τιμές του I που είναι μεγαλύτερες από την $E(I)$ δείχνουν πως υπάρχει θετική χωρική αυτοσυσχέτιση, δηλαδή κοντινές τιμές τείνουν να βρίσκονται κοντά χωρικά, ενώ διαφορετικές τιμές τείνουν να έχουν μακρινές τοποθεσίες. Αντίθετα, οι τιμές του συντελεστή που είναι μικρότερες από την $E(I)$ υποδηλώνουν αρνητική χωρική αυτοσυσχέτιση, δηλαδή κοντινές τιμές βρίσκονται μακριά μεταξύ τους στον χώρο. Συνήθως ο συντελεστής I παίρνει τιμές από το -1 έως το +1. [37]

Για τον υπολογισμό της χωρικής αυτοσυσχέτισης στις παρατηρήσεις που θα εισαχθούν στο SustainGraph, θα υπολογιστούν πρώτα τα χωρικά βάρη που βασίζονται στην μεταξύ τους απόσταση, αφού πρόκειται για παρατηρήσεις με γεωμετρία σημείου. Έπειτα, θα υπολογιστεί ο συντελεστής χωρικής αυτοσυσχέτισης I του Moran. Εφόσον οι παρατηρήσεις έχουν και χρονική διάσταση, η διαδικασία αυτή θα γίνει ενδεικτικά για συγκεκριμένες χρονικές στιγμές, για κάθε περιοχή, και για κάθε περίπτωση ομαδοποίησης με χρήση πλέγματος, δηλαδή για κάθε μέγεθος κελιού του πλέγματος.

3.4.3 Ομαδοποίηση μέσω δημιουργίας χωρικών συστάδων

Ένας εναλλακτικός τρόπος χωρισμού των χωρικών παρατηρήσεων κάθε περιοχής σε ομάδες, με στόχο την συγχώνευσή τους και τελικά την μείωση τους, είναι μέσω της ομαδοποίησής τους σε συστάδες, χρησιμοποιώντας κάποιον αλγόριθμο συσταδοποίησης με χωρικό περιορισμό. Η προσέγγιση αυτή λαμβάνει υπόψη την τιμή που έχουν οι παρατηρήσεις, δηλαδή την τιμή του αντίστοιχου δείκτη, σε αντίθεση με την ομαδοποίηση που περιγράφηκε στις προηγούμενες υποενότητες, η οποία γινόταν αποκλειστικά με βάση την γεωμετρία τους.

Οι αλγόριθμοι συσταδοποίησης στοχεύουν στην διαίρεση ενός συνόλου αντικειμένων σε ομάδες, οι οποίες ονομάζονται συστάδες, με τέτοιο τρόπο ώστε αντικείμενα της ίδιας ομάδας να είναι περισσότερο όμοια μεταξύ τους από ό,τι με αντικείμενα άλλων ομάδων, όσον αφορά τις τιμές των χαρακτηριστικών τους. [38] Όταν τα αντικείμενα χαρακτηρίζονται από μία τοποθεσία στον χώρο, τότε στην διαδικασία της συσταδοποίησης πρέπει να ληφθούν υπόψη και οι χωρικές τους σχέσεις, και συγκεκριμένα οι σχέσεις γειτνίασης μεταξύ τους. Η διαδικασία αυτή ονομάζεται regionalization ή συσταδοποίηση με χωρικό περιορισμό και οι συστάδες που δημιουργούνται ονομάζονται περιοχές (regions) ή χωρικές συστάδες. [39]

Συσσωρευτική ιεραρχική συσταδοποίηση με χωρικό περιορισμό

Ο αλγόριθμος συσσωρευτικής ιεραρχικής συσταδοποίησης (Agglomerative Hierarchical Clustering (AHC)) είναι ένας αλγόριθμος συσταδοποίησης, ο οποίος στην κλασική του μορφή δεν λαμβάνει υπόψη τα χωρικά χαρακτηριστικά των παρατηρήσεων. Ανήκει στην οικογένεια των ιεραρχικών αλγορίθμων συσταδοποίησης, οι οποίοι παράγουν ένα σύνολο από εμφωλευμένες συστάδες που μπορούν να αναπαρασταθούν σε ένα ιεραρχικό δέντρο. Οι συστάδες στο χαμηλότερο επίπεδο του δέντρου περιέχουν από μία παρατήρηση, ενώ το υψηλότερο επίπεδο αποτελείται από μία συστάδα που περιέχει όλες τις παρατηρήσεις. Ο υπολογισμός των συστάδων μπορεί να γίνει είτε ξεκινώντας από το χαμηλότερο επίπεδο και συγχωνεύοντας αναδρομικά ένα ζευγάρι συστάδων σε κάθε επίπεδο (συσσωρευτικός τρόπος), είτε ξεκινώντας από το υψηλότερο επίπεδο και διαχωρίζοντας αναδρομικά μία συστάδα σε κάθε επίπεδο, δημιουργώντας δύο νέες συστάδες (διαιρετικός τρόπος). Στην περίπτωση του AHC, εφαρμόζεται

ο συσσωρευτικός τρόπος παραγωγής του ιεραρχικού δέντρου. [40]

Οι δύο συστάδες που συγχωνεύονται σε κάθε επίπεδο κατά την εκτέλεση ενός αλγορίθμου AHC είναι αυτές που έχουν την μικρότερη απόσταση μεταξύ τους. Η απόσταση μπορεί να οριστεί με διαφορετικούς τρόπους, ένας εκ των οποίων είναι η μέθοδος του Ward. Σύμφωνα με αυτήν την μέθοδο, σε κάθε επίπεδο επιλέγεται το ζευγάρι συστάδων το οποίο όταν συνδεθεί θα οδηγήσει στην μικρότερη δυνατή αύξηση του αθροίσματος τετραγώνων εντός των συστάδων (Within-Cluster Sum of Square (WCSS)). [41] Ως WCSS ορίζεται το άθροισμα του τετραγώνου της διαφοράς της τιμής κάθε παρατήρησης από την μέση τιμή των τιμών όλων των παρατηρήσεων της συστάδας στην οποία ανήκει, δηλαδή το άθροισμα της διακύμανσης σε κάθε συστάδα. [42]

Ο αλγόριθμος AHC με χωρικό περιορισμό αποτελεί μία προσαρμοσμένη εκδοχή του AHC, η οποία λαμβάνει υπόψη τις σχέσεις γειτνίασης μεταξύ των παρατηρήσεων. Συκεκριμένα, λαμβάνει ως είσοδο μία αναπαράσταση της χωρικής σύνδεσης μεταξύ των παρατηρήσεων με τη μορφή ενός πίνακα με χωρικά βάρη, τα οποία μπορούν να είναι οποιουδήποτε τύπου, όπως περιγράφηκε στην προηγούμενη ενότητα. Ο αλγόριθμος επιτρέπει να συνδεθούν σε κάθε επίπεδο του ιεραρχικού δέντρου μόνο συστάδες που συνδέονται χωρικά, δηλαδή τέτοιες ώστε τουλάχιστον μία παρατήρηση της μίας να συνδέεται με μία παρατήρηση της άλλης, δηλαδή να είναι γείτονες σύμφωνα με τον πίνακα χωρικών βαρών. [43]

Όπως περιγράφηκε παραπάνω, ο αλγόριθμος AHC παράγει μία συσταδοποίηση σε κάθε βήμα του αλγορίθμου. Ένα κριτήριο επιλογής της επιθυμητής συσταδοποίησης, και συνεπώς ένα κριτήριο τερματισμού του αλγορίθμου, είναι το επιθυμητό πλήθος συστάδων.

Μέθοδος του αγκώνα

Η μέθοδος του αγκώνα (elbow method) χρησιμοποιείται για την εύρεση του βέλτιστου αριθμού συστάδων σε ένα σύνολο παρατηρήσεων. Επιλέγει τον αριθμό συστάδων στον οποίο εάν προσθέσουμε ακόμα μία συστάδα δεν θα μειωθεί κατά πολύ το WCSS, ή αλλιώς διακύμανση των παρατηρήσεων μέσα σε αυτές, δηλαδή δεν θα καταλήξουμε με πολύ καλύτερη μοντελοποίηση των δεδομένων. Για την εφαρμογή της μεθόδου αρχικά σχεδιάζεται η καμπύλη του συνολικού WCSS όλων των συστάδων, συναρτήσει του αριθμού τους. Έπειτα επιλέγεται ως βέλτιστος αριθμός αυτός στον οποίο η καμπύλη σχηματίζει “αγκώνα”, δηλαδή έχει έντονη εναλλαγή από μεγάλη αρνητική κλίση σε μικρότερη. [44] [45]

Για τον εντοπισμό του “αγκώνα”, εκτός από την παρατήρηση της καμπύλης μπορεί να χρησιμοποιηθεί και κάποιος αλγόριθμος ανίχνευσης “αγκώνων”, όπως ο Kneedle [46], ο οποίος λαμβάνει ως είσοδο τη συνάρτηση του WCSS με τον αριθμό των συστάδων και επιστρέφει το σημείο στο οποίο βρίσκεται ο “αγκώνας”.

Προσαρμοσμένος δείκτης Rand

Ο προσαρμοσμένος δείκτης Rand είναι ένα μέτρο ομοιότητας μεταξύ δύο συσταδοποιήσεων και αποτελεί μία παραλλαγή του κλασικού δείκτη Rand. [47]

Ο δείκτης Rand παίρνει τιμές στο κλειστό διάστημα $[0,1]$, με το 0 να υποδηλώνει πως οι δύο συσταδοποιήσεις είναι τελείως διαφορετικές, και το 1 ακριβώς ίδιες. Θεωρώντας ως S ένα σύνολο παρατηρήσεων και ως X και Y δύο διαμερίσεις (συσταδοποιήσεις) του S σε υποσύνολα

(συστάδες), αν ορίσουμε ως A τον αριθμό των συμφωνιών μεταξύ των δύο διαμερίσεων, δηλαδή τον αριθμό των ζευγών παρατηρήσεων που κατηγοριοποιήθηκαν είτε στην ίδια συστάδα και στις δύο διαμερίσεις είτε σε διαφορετικές συστάδες και στις δύο διαμερίσεις, και ως B τον συνολικό αριθμό ζευγών, τότε ως δείκτης Rand ορίζεται ο $RI = \frac{A}{B}$.

Για τον προσαρμοσμένο δείκτη Rand θεωρούμε ως S ένα σύνολο n παρατηρήσεων και ως $X = \{X_1, X_2, \cdots, X_r\}$ και $Y = \{Y_1, Y_2, \cdots, Y_s\}$ δύο συσταδοποιήσεις αυτών των παρατηρήσεων. Η επικάλυψη των δύο συσταδοποίησεων μπορεί να περιγραφεί από έναν πίνακα του οποίου κάθε στοιχείο n_{ij} δηλώνει των αριθμό των παρατηρήσεων που είναι κοινές στις δύο συστάδες X_i και Υ_j. Ο πίνακας είναι ο εξής:

$X \setminus Y$	Υ_1	Υ_2	$\cdots$	Υ_s	Αθροίσματα
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
Αθροίσματα	b_1	b_2	$\cdots$	b_s	

Με βάση τον παραπάνω πίνακα, ο προσαρμοσμένος δείκτης Rand ορίζεται ως εξής:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2}\sum_j + \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}} \text{[47]}$$

Μεθοδολογία ομαδοποίησης μέσω δημιουργίας χωρικών συστάδων

Στην περίπτωση των παρατηρήσεων που προετοιμάζονται για να εισαχθούν στο SustainGraph, προτείνεται η εφαρμογή του αλγορίθμου AHC με χωρικό περιορισμό στις παρατηρήσεις κάθε περιοχής. Καθώς οι παρατηρήσεις έχουν γεωμετρία σημείου, τα χωρικά βάρη θα υπολογιστούν με βάση την μεταξύ τους απόσταση, και το πλήθος των παραγόμενων συστάδων θα καθοριστεί με την εφαρμογή της μεθόδου του αγκώνα. Εφόσον τα δεδομένα των παρατηρήσεων είναι και χρονικά πέρα από χωρικά, ο αλγόριθμος AHC θα εκτελεστεί επαναληπτικά για κάθε διαθέσιμη χρονική στιγμή. Με αυτόν τον τρόπο, οι παρατηρήσεις που παίρνει ως είσοδο θα χαρακτηρίζονται μόνο από την τιμή που παίρνει ο αντίστοιχος δείκτης και από μία τοποθεσία στον χώρο. Ως αποτέλεσμα, θα παραχθούν χωρικές συστάδες με διαφορετικές γεωμετρίες για κάθε χρονική στιγμή. Κάθε συστάδα θα αναπαρασταθεί από την μέση τιμή των παρατηρήσεων που ανήκουν σε αυτή και η γεωμετρία της θα αναπαρασταθεί από το κεντροειδές του συνόλου των σημείων της. Ο υπολογισμός θα γίνει εφαρμόζοντας χωρική ομαδοποίηση στις παρατηρήσεις κάθε συστάδας και έπειτα την συνάρτηση centroid.

Χρησιμοποιώντας τον παραπάνω μηχανισμό ομαδοποίησης, για κάθε χρονική στιγμή υπάρχουν ομαδοποιημένες παρατηρήσεις με διαφορετική γεωμετρία. Το γεγονός αυτό καθιστά πιο δύσκολη την μετέπειτα ανάλυση των δεδομένων με τη μορφή χρονοσειράς, καθώς η σύγκριση των παρατηρήσεων στον άξονα του χρόνου απαιτεί σύγκριση παρατηρήσεων σε διαφορετικά σημεία του χάρτη και, πιθανότατα, με διαφορετικό πλήθος. Για τον λόγο αυτό, για την αξιολόγηση της αξιοποιησιμότητας του συγκεκριμένου μηχανισμού στην παρούσα

εφαρμογή, είναι επιτακτικό να συγκριθούν οι συσταδοποιήσεις που παράγονται για κάθε χρονική στιγμή, και να κριθεί εάν η μεταξύ τους επικάλυψη, ή αλλιώς ομοιότητα, είναι επαρκής. Για την σύγκριση θα χρημοποιηθεί η οπτικοποίηση των συσταδοποιήσεων στον χάρτη αλλά και ο προσαρμοσμένος δείκτης Rand.

3.5 Μηχανισμοί ανάλυσης χωρικών δεδομένων στο SustainGraph

Αναπαριστώντας τα χωρικά δεδομένα στο SustainGraph με τον τρόπο που περιγράφηκε στις προηγούμενες ενότητες, καθίσταται εφικτή η εφαρμογή ποικίλων αναλύσεων σε αυτά, τα αποτελέσματα των οποίων μπορούν να παρέχουν χρήσιμες πληροφορίες και συμπεράσματα σχετικά με τις επιπτώσεις της κλιματικής αλλαγές σε διαφορετικές περιοχές. Η διαδικασία που ακολουθείται για την ανάλυσή τους ξεκινάει με την ανάκτηση των χωρικών δεδομένων από το SustainGraph με τη χρήση κατάλληλων επερωτήσεων, και την αναπαράστασή τους σε δομή πίνακα. Τα χωρικά δεδομένα, όπως περιγράφηκε, περιέχονται στους κόμβους *Observation* συνδυαστικά με τις σχέσεις *REFERS_TO_AREA*, οι οποίες τους συνδέουν με τα αντίστοιχα *GeoAreas*, και περιέχουν τις συντεταγμένες του σημείου που αφορά η παρατήρηση. Εφόσον είναι επιθυμητό, η ανάκτηση των δεδομένων μπορεί να γίνει ορίζοντας κάποιον χωρικό περιορισμό στην επερώτηση, είτε με τη χρήση κάποιας από τις χωρικές συναρτήσεις distance και withinBBox που παρέχει το Neo4j, είτε ορίζοντας συγκεκριμένο κόμβο *GeoArea*, δηλαδή την περιοχή στην οποία περιέχονται χωρικά οι παρατηρήσεις του ενδιαφέροντος.

Έπειτα από την ανάκτηση των παρατηρήσεων και την αναπαράστασή τους σε δομή πίνακα, μπορούν να εφαρμοστούν σε αυτές ποικίλοι αλγόριθμοι για την ανάλυσή τους. Ανάμεσα σε αυτούς είναι ο υπολογισμός και η σύγκριση κλασικών στατιστικών μεγεθών σε διαφορετικές περιοχές, και ο υπολογισμός στατιστικών μεγεθών που λαμβάνουν υπόψη τη γεωμετρία των δεδομένων, όπως η χωρική αυτοσυσχέτιση σε καθολικό και σε τοπικό επίπεδο. Μπορούν επίσης να εφαρμοστούν αλγόριθμοι χωρικής παλινδρόμησης, οι οποίοι διαφέρουν από τους κλασικούς αλγόριθμους παλινδρόμησης στο ότι χρησιμοποιούν μεταβλητές που κατασκευάζονται από τα χωρικά χαρακτηριστικά των παρατηρήσεων και τις μεταξύ τους χωρικές σχέσεις, καθώς και αλγόριθμοι συσταδοποίησης με ή χωρίς χωρικό περιορισμό.

3.5.1 Συσταδοποίηση με πολλαπλές μεταβλητές

Μία ιδιαίτερα χρήσιμη ανάλυση η οποία διευκολύνεται από τον τρόπο αναπαράστασης των χωρικών δεδομένων στο SustainGraph είναι η συσταδοποίηση των χωρικών δεδομένων χωρίς χωρικό περιορισμό, λαμβάνοντας υπόψη παρατηρήσεις πολλαπλών δεικτών. Η συσταδοποίηση βοηθά στη δημιουργία ενός μικρού αριθμού κατηγοριών των σημείων με βάση τις διαφορετικές μεταβλητές των παρατηρήσεων σε αυτά, και συνεπώς στην συνοπτική περιγραφή τους. Εφόσον επιλέγεται κάποιος αλγόριθμος συσταδοποίησης χωρίς χωρικό περιορισμό, τα σημεία κάθε συστάδας μπορούν να βρίσκονται σε διαφορετικές περιοχές στον χώρο, χωρίς περιορισμό στις χωρικές τους σχέσεις. Έτσι, κάθε περιοχή *GeoArea* (για παράδειγμα κάθε περιοχή NUTS επιπέδου 3) μπορεί να χαρακτηριστεί από τις κατηγορίες, δηλαδή συστάδες, στις οποίες κατηγοριοποιήθηκαν τα σημεία της, καθώς και από το ποσοστό της περιοχής που

καταλαμβάνουν.

Συσταδοποίηση K-means

Ο K-means είναι ένας αλγόριθμος για την εύρεση K συστάδων και των κεντρικών σημείων τους σε ένα σύνολο δεδομένων. Ο αριθμός των συστάδων K δίνεται ως είσοδος, και ο αλγόριθμος αλλάζει τα κεντρικά σημεία επαναληπτικά ώστε να ελαχιστοποιήσει το WCSS. [40]

Αλγοριθμος 3.1: *Αλγόριθμος K-means*

Είσοδος: K (*ο αριθμός των συστάδων*)
Επιλογή των K αρχικών κεντρικών σημείων
repeat
Για κάθε κεντρικό σημείο εύρεση του υποσυνόλου των σημείων που είναι κοντινότερα σε αυτό από ό,τι σε οποιδήποτε άλλο κεντρικό σημείο.
Για κάθε συστάδα υπολογισμός της μέσης τιμής κάθε μεταβλητής των σημείων της και ορισμός του διανύσματος τους ως το νέο κεντρικό της σημείο.
until τα κεντρικά σημεία να μην αλλάζουν

Συνήθως τα αρχικά κεντρικά σημεία επιλέγονται τυχαία. Όπως φαίνεται παραπάνω, κατά τον τερματισμό του αλγορίθμου, το κεντρικό σημείο κάθε συστάδας είναι το διάνυσμα με τις μέσες τιμές των μεταβλητών των σημείων της, και συνεπώς δεν είναι ένα από τα σημεία εισόδου. [40] Για την επιλογή του K συχνά ακολουθείται η μέθοδος του αγκώνα, η οποία περιγράφηκε στην υπο-ενότητα 3.4.3.

Συντελεστής σκιαγράφησης

Έπειτα από τον χωρισμό των δεδομένων, είναι σημαντικό να αξιολογηθούν οι παραγόμενες συστάδες. Πέρα από το WCSS, το οποίο υπολογίζεται κατά την εφαρμογή της μεθόδου του αγκώνα, ένα άλλο μέτρο αξιολόγησης των συστάδων είναι ο συντελεστής σκιαγράφησης. Το μέτρο αυτό εξετάζει πόσο όμοια είναι μεταξύ τους τα σημεία κάθε συστάδας, δηλαδή τη συνοχή, και πόσο ανόμοια είναι τα σημεία διαφορετικών συστάδων, δηλαδή τον διαχωρισμό. Παίρνει τιμές στο κλειστό διάστημα [-1,1], με τις υψηλές τιμές να υποδεικνύουν πως τα σημεία κάθε συστάδας είναι αρκετά όμοια μεταξύ τους και αρκετά ανόμοια με τα σημεία των υπόλοιπων συστάδων. Δηλαδή, όσο υψηλότερη είναι η τιμή, τόσο καλύτερα έχουν χωριστεί τα σημεία σε συστάδες. [48]

Για τον ορισμό του συντελεστή σκιαγράφησης θεωρούμε ως $a(i)$ την μέση απόσταση ενός σημείου i από όλα τα υπόλοιπα σημεία της συστάδας του (συνοχή), και ως $b(i)$ την μέση απόσταση του σημείου i από όλα τα σημεία μίας άλλης συστάδας, αυτής με την οποία πετυχαίνεται η μικρότερη μέση απόσταση (διαχωρισμός). Τότε, στις συστάδες που περιέχουν περισσότερα από ένα σημεία, ο συντελεστής σκιαγράφησης στο σημείο i ορίζεται ως:

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

ενώ στις συστάδες που περιέχουν ακριβώς ένα σημείο, ο συντελεστής σκιαγράφησης σε αυτό

είναι:

$$s(i) = 0$$

Η αξιολόγηση της συσταδοποίησης συνολικά πραγματοποιείται υπολογίζοντας τη μέση τιμή των συντελεστών συσταδοποίησης όλων των σημείων.

3.5.2 Υπολογισμός νέων παρατηρήσεων στις γεωμετρίες του SustainGraph

Τόσο κατά την εφαρμογή της συσταδοποίησης με πολλαπλές μεταβλητές που περιγράφηκε στην προηγούμενη υπο-ενότητα, όσο και σε άλλες αναλύσεις των χωρικών δεδομένων, χρησιμοποιούνται τιμές πολλαπλών δεικτών, πολλοί από τους οποίους δεν αναπαριστώνται ήδη στον γράφο γνώσης, αλλά παρέχονται από εξωτερικές πηγές. Για την εφαρμογή των αναλύσεων απαιτείται πρώτα ο υπολογισμός των τιμών όλων των νέων δεικτών στις γεωμετρίες για τις οποίες περιέχονται οι παρατηρήσεις στο SustainGraph.

Οι πηγές δεδομένων που αξιοποιούνται, συχνά δεν παρέχουν τα χωρικά δεδομένα με την αναπαράσταση που περιγράφηκε στις προηγούμενες ενότητες, δηλαδή χρησιμοποιώντας γεωμετρίες όπως σημείο, γραμμή και πολύγωνο (διανυσματικά δεδομένα), αλλά αναπαριστώντας τα σε ψηφιδωτή μορφή. Πρόκειται για ένα μοντέλο χωρικών δεδομένων που αναπαριστά τον κόσμο σαν μία επιφάνεια χωρισμένη σε πλέγμα κελιών. [31] Το μοντέλο αυτό χρησιμοποιείται για την αναπαράσταση δεδομένων που μεταβάλλονται με συνεχή τρόπο, όπως αυτά που προέρχονται από αεροφωτογραφίες και δορυφορικές εικόνες. Το δομικό στοιχείο, δηλαδή το κελί ή αλλιώς η ψηφίδα, έχει συνήθως τετράγωνο σχήμα. Για τον υπολογισμό των τιμών σε συγκεκριμένα σημεία, και συγκεκριμένα στα σημεία τα οποία αφορούν οι παρατηρήσεις του SustainGraph που θα αξιοποιηθούν στην αντίστοιχη ανάλυση, για κάθε σημείο εξάγεται η τιμή που έχει η ψηφίδα στην οποία περιέχεται χωρικά.

Είναι σημαντικό να αναφερθεί πως τα αποτελέσματα των αναλύσεων μπορούν να αποτελέσουν είσοδο στον γράφο γνώσης. Δηλαδή, οι παρατηρήσεις με γεωμετρία σημείου που υπολογίστηκαν με βάση τα χωρικά δεδομένα που αντλήθηκαν από πηγές δεδομένων σε ψηφιδωτή μορφή, μπορούν να εισαχθούν στο SustainGraph, ως παρατηρήσεις ενός νέου δείκτη, αναπαριστώντας τα με τον τρόπο που περιγράφηκε στην ενότητα 3.2.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 4

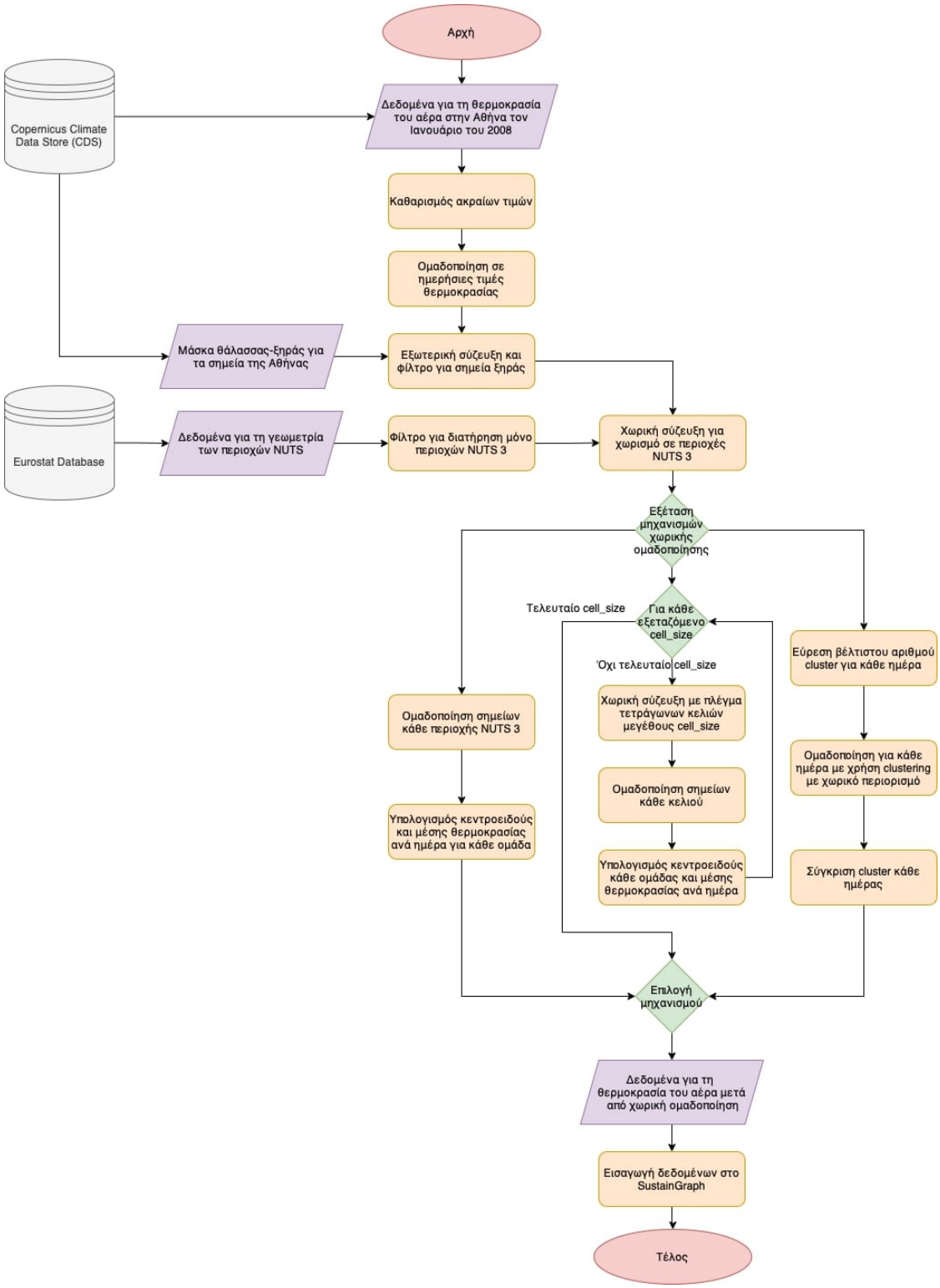
Υλοποίηση

Στο κεφάλαιο αυτό περιγράφεται η υλοποίηση της επέκτασης του SustainGraph για την αναπαράσταση χωρικών δεδομένων, με βάση τις τεχνολογίες και την μεθοδολογία που παρουσιάστηκαν στο προηγούμενο κεφάλαιο. Αρχικά παρουσιάζονται τα δεδομένα που χρησιμοποιήθηκαν και η επιθυμητή αναπαράσταση τους στον γράφο γνώσης. Στη συνέχεια δίνονται οι λεπτομέρειες υλοποίησης των μηχανισμών προεπεξεργασίας τους, εισαγωγής τους και περεταίρω ανάλυσης τους, καθώς επίσης και η δομή του κώδικα, τα ενδιάμεσα αποτελέσματα και οι τεχνολογίες που χρησιμοποιήθηκαν.

Στο σχήμα 4.1 παρουσιάζονται σε μορφή διαγράμματος ροής τα βασικά βήματα που ακολουθήθηκαν, από την επιλογή του dataset με τα χωρικά δεδομένα, έως και την εισαγωγή τους στο SustainGraph. Οι ελλείψεις αναπαριστούν την αρχή και το τέλος της διαδικασίας που ακολουθήθηκε, οι κύλινδροι αναπαριστούν τις πηγές δεδομένων που αξιοποιήθηκαν, και τα πλάγια παραλληλόγραμμα αναπαριστούν τα σύνολα δεδομένων (datasets) που επιλέχθηκαν και χρησιμοποιήθηκαν. Τα κίτρινα παραλληλόγραμμα δείχνουν τις ενέργειες που εφαρμόστηκαν στα δεδομένα, όπως η ομαδοποίηση με βάση κάποιο χαρακτηριστικό, η εφαρμογή φίλτρων και η σύζευξη μεταξύ δύο datasets. Οι ρόμβοι αναπαριστούν βήματα στα οποία πάρθηκε κάποια απόφαση σχετικά με τις επόμενες ενέργειες ή χωρίστηκε η ροή των βημάτων σε περισσότερες από μία, ανεξάρτητες μεταξύ τους, ροές. Τα βέλη δείχνουν τη σειρά με την οποία ακολουθήθηκαν οι ενέργειες επί των δεδομένων, αλλά και την πορεία των δεδομένων από την πηγή τους ως την εισαγωγή τους στο SustainGraph. Τα βήματα που παρουσιάζονται σχηματικά, εξηγούνται αναλυτικά στις ενότητες 4.1-4.3. Στην τελευταία ενότητα του κεφαλαίου, παρουσιάζεται η υλοποίηση και εφαρμογή των μηχανισμών ανάλυσης χωρικών δεδομένων. Τα βήματα που ακολουθήθηκαν δεν συμπεριλαμβάνονται στο σχήμα 4.1.

4.1 Αναπαράσταση χωρικών δεδομένων στο SustainGraph

Στην ενότητα αυτή παρουσιάζεται το σύνολο των δεδομένων με χωρικά χαρακτηριστικά που προστέθηκαν στο SustainGraph και στο οποίο εφαρμόστηκαν οι μηχανισμοί εισαγωγής, προεπεξεργασίας και ανάλυσης που θα περιγραφούν στις επόμενες ενότητες. Έπειτα παρουσιάζεται η αναπαράστασή τους μέσα στον γράφο γνώσης σύμφωνα με το ανανεωμένο σχήμα αναπαράστασης δεδομένων που επιλέχθηκε για την εύκολη αποθήκευση τους και εφαρμογή επερωτήσεων σε αυτά.



Σχήμα 4.1: *Διάγραμμα ροής με τα βήματα που ακολουθήθηκαν απο την αρχική επιλογή του dataset μέχρι την τελική εισαγωγή των δεδομένων στον γράφο γνώσης*

4.1.1 Χρησιμοποιούμενο dataset

Το dataset [49] που χρησιμοποιήθηκε περιέχει δεδομένα που αφορούν τις τιμές τεσσάρων διαφορετικών κλιματικών μεταβλητών για πόλεις της Ευρώπης κατά την χρονική περίοδο από

το 2008 έως το 2017. Οι μεταβλητές αυτές είναι οι εξής: θερμοκρασία αέρα, ειδική υγρασία, σχετική υγρασία και ταχύτητα αέρα.

Το συγκεκριμένο dataset είναι διαθέσιμο από το πρόγραμμα Copernicus, τμήμα του διαστημικού προγράμματος της Ευρωπαϊκής Ένωσης που είναι υπεύθυνο για την παρατήρηση της Γης τόσο μέσω δορυφόρων όσο και μέσω της συλλογής δεδομένων από ποικίλα συστήματα μέτρησης στη Γη. Συγκεκριμένα, τα δεδομένα παρέχονται μέσω της αποθήκης δεδομένων για το κλίμα (CDS) από την υπηρεσία κλιματικής αλλαγής του Copernicus, η οποία αποτελεί μία από τις έξι θεματικές υπηρεσίες πληροφόρησης που παρέχονται, και έχει ως στόχο την προσαρμογή στην κλιματική αλλαγή και τον μετριασμό των επιπτώσεών της, μέσω της τακτικής παροχής έγκυρων πληροφοριών σχετικά με αυτή.

Τα δεδομένα αφορούν 100 Ευρωπαϊκές πόλεις, συμπεριλαμβανομένου της Αθήνας. Τα δεδομένα για τις κλιματικές μεταβλητές στις πόλεις αυτές δίνονται σε μορφή πλέγματος με χωρική ανάλυση 100 μέτρα. Η χρονική περίοδος που καλύπτει το dataset είναι από τον Ιανουάριο του 2008 μέχρι τον Δεκέμβρη του 2017 και περιέχονται οι τιμές των μεταβλητών ανά χρονικά διαστήματα της μίας ώρας. Κάποιες πληροφορίες για τις κλιματικές μεταβλητές που περιέχονται είναι οι εξής:

- Θερμοκρασία του αέρα κοντά στην επιφάνεια:

  Καταγράφεται στη μονάδα μέτρησης Kelvin (K) και αναφέρεται στην θερμοκρασία του αέρα 2 μέτρα πάνω από την επιφάνεια της Γης.

- Ειδική υγρασία κοντά στην επιφάνεια:

  Καταγράφεται σε ποσοστό επί τοις εκατό (%) και αναφέρεται στην ειδική υγρασία 2 μέτρα πάνω από την επιφάνεια της Γης. Αποτυπώνει τη σχέση μεταξύ της πραγματικής υγρασίας και της υγρασίας κορεσμού και παίρνει τιμές στο διάστημα [0, 100]. Η τιμή 0% σημαίνει ότι ο αέρας μέσα στο κελί του πλέγματος είναι εντελώς στεγνός, ενώ η τιμή 100% δείχνει ότι ο αέρας στο κελί είναι κορεσμένος με υδρατμούς.

- Σχετική υγρασία κοντά στην επιφάνεια:

  Αναφέρεται στην μάζα υδρατμών μέσα σε μία μονάδα μάζας αέρα με τη μέγιστη ποσότητα υδρατμών, 2 μέτρα πάνω από την επιφάνεια της Γης. Οι μονάδες μέτρησης της σχετικής υγρασίας είναι $kgkg^{-1}$.

- Ταχύτητα αέρα κοντά στην επιφάνεια:

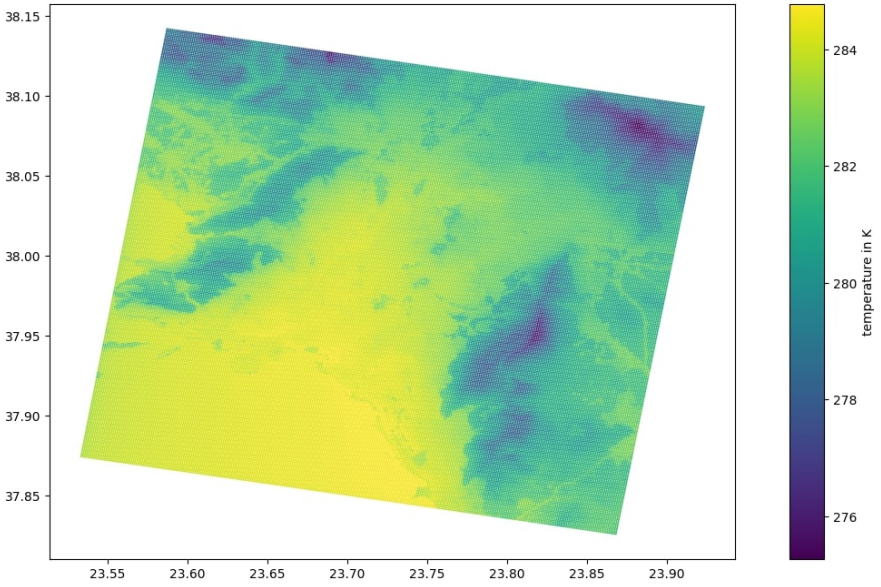
  Καταγράφεται στη μονάδα μέτρησης m/s και αναφέρεται στην ταχύτητα του αέρα 2 μέτρα πάνω από την επιφάνεια της Γης. Υπολογίζεται λαμβάνοντας υπόψη και την ζωνική (u), δηλαδή παράλληλα με τις γραμμές του γεωγραφικού πλάτους, και την μεσημβρινή (v), δηλαδή παράλληλα με τις γραμμές του γεωγραφικού μήκους κυκλοφορία του αέρα, από τον τύπο $sqrt(u^2 + v^2)$.

Για κάθε πόλη η χωρική έκταση που καλύπτεται έχει σχήμα τετραγώνου. Οι τιμές των κλιματικών μεταβλητών δίνονται για κάθε κελί του πλέγματος μαζί με το γεωγραφικό πλάτος και το γεωγραφικό μήκος του κεντροειδούς του. Το πλέγμα που χρησιμοποιήθηκε είναι το

Ευρωπαϊκό Πλέγμα (EPSG:3035), όμως οι συντεταγμένες των σημείων μετατράπηκαν στο σύστημα συντεταγμένων WGS84 (EPSG:4326).

Τα δεδομένα περιέχονται σε αρχεία τύπου NetCDF-4 και η πρόσβαση σε αυτά παρέχεται από το API του CDS. Τα δεδομένα κάθε κλιματικής μεταβλητής για κάθε μήνα και έτος περιέχεται σε ένα διαφορετικό αρχείο NetCDF-4.

Το τμήμα του dataset με το οποίο ασχοληθήκαμε στα πλαίσια αυτής της εργασίας είναι αυτό που αφορά την Αθήνα. Οι μηχανισμοί για την εισαγωγή στο SustainGraph εφαρμόστηκαν ξεχωριστά σε κάθε αρχείο (που περιέχει τα δεδομένα για κάθε κλιματική μεταβλητή ανά έτος και μήνα), αλλά η διαδικασία που ακολουθήθηκε είναι η ίδια. Ως παράδειγμα των δεδομένων για την πόλη της Αθήνας, παρουσιάζεται στο σχήμα 4.2 ο χάρτης με τη μέση θερμοκρασία του αέρα σε όλα τα σημεία που παρέχονται για την Αθήνα, την πρώτη μέρα της χρονικής περιόδου του dataset, δηλαδή την 1η Ιανουαρίου του 2008.

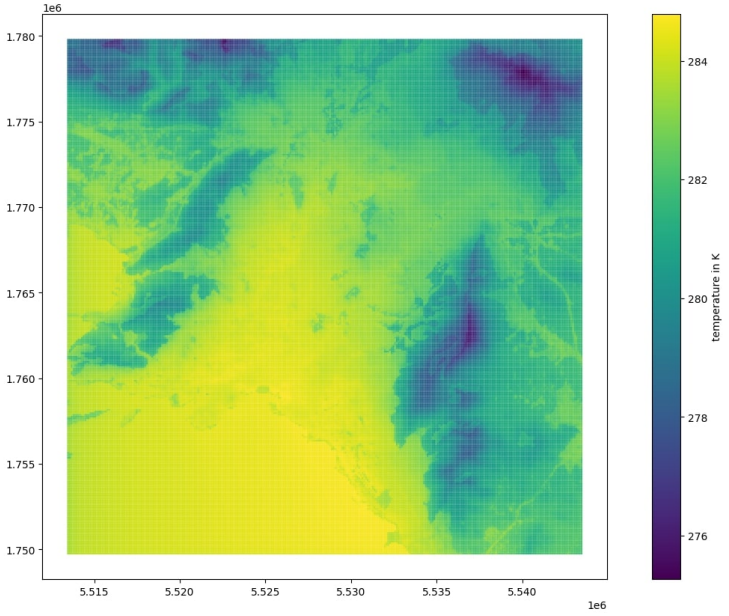


Σχήμα 4.2: *Χάρτης μέσης θερμοκρασίας την 1η Ιανουαρίου του 2008 για την Αθήνα (EPSG:4326)*

Ο λόγος που ο χάρτης παρουσιάζεται λοξός είναι η μετατροπή του σε WGS84 (EPSG:4326) σύστημα συνεταγμένων ενώ χρησιμοποιήθηκε το Ευρωπαϊκό Πλέγμα (EPSG:4326). Για την καλύτερη οπτικοποίησή του αλλά και για την επεξεργασία του dataset που θα ακολουθήσουμε στη συνέχεια, μετατρέπουμε το σύστημα συντεταγμένων των σημείων σε EPSG:3035. Ο χάρτης που προκύπτει φαίνεται στο σχήμα 4.3.

4.1.2 Αναπαράσταση κλιματικών μεταβλητών στο SustainGraph

Όπως περιγράφηκε στο προηγούμενο κεφάλαιο, το SustainGraph έχει ως στόχο την καταγραφή της προόδου προς τους SDG στόχους η οποία αποτυπώνεται στην εξέλιξη ορισμένων σχετικών δεικτών. Οι κλιματικές μεταβλητές της θερμοκρασίας του αέρα, της ειδικής και σχετικής υγρασίας και της ταχύτητας του αέρα που παρέχονται από το πρόγραμμα Copernicus μπορούν να αποτελέσουν δείκτες στον γράφο γνώσης που σχετίζονται με τον αντίκτυπο της κλιματικής αλλαγής, η οποία επηρεάζει την πρόοδο προς την επίτευξη της βιώσιμης ανάπτυ-



Σχήμα 4.3: *Χάρτης μέσης θερμοκρασίας την 1η Ιανουαρίου του 2008 για την Αθήνα (EPSG:3035)*

ξης.

Για την αναπαράσταση των μεταβλητών αυτών στο SustainGraph δημιουργούνται νέοι κόμβοι και σχέσεις μεταξύ τους με τις κατάλληλες ετικέτες και ιδιότητες. Αρχικά δημιουργούνται τέσσερις νέοι κόμβοι στον γράφο με την ετικέτα *Indicator*, ένας για κάθε κλιματική μεταβλητή. Ως ιδιότητες περιέχουν: τον κωδικό του αντίστοιχου δείκτη (π.χ. "air_temperature" για τη θερμοκρασία του αέρα) και μία περιγραφή για την αντίστοιχη κλιματική μεταβλητή. Οι κόμβοι αυτοί συνδέονται μέσω μίας σχέσης με ετικέτα *COMES_FROM* με τον κόμβο τύπου *Source* με όνομα "TPS", καθώς δεν αποτελούν δείκτες των Ηνωμένων Εθνών ή της Ευρωπαϊκής Ένωσης, αλλά δείκτες από εξωτερική πηγή. Προκειμένου να εισάγουμε τη χρονοσειρά με τις τιμές που παίρνει κάθε κλιματική μεταβλητή σε κάθε σημείο, δημιουργούμε έναν κόμβο με ετικέτα *Series*, που περιέχει τον κωδικό της χρονοσειράς (π.χ. "air_temperature_copernicus" για τη θερμοκρασία του αέρα), την περιγραφή της αντίστοιχης μεταβλητής, και το URL του προμηθευτή των δεδομένων, δηλαδή το URL της αποθήκης δεδομένων CDS του Copernicus. Οι κόμβοι αυτοί συνδέονται με τον αντίστοιχό τους *Indicator* με μία σχέση *HAS_SERIES*. Επιπλέον, για κάθε έναν από τους τέσσερις *Series* κόμβους, δημιουργούμε έναν κόμβο με ετικέτα *SeriesMetadata* που περιέχει επιπλέον πληροφορίες για την αντίστοιχη μετρική, όπως τις μονάδες μέτρησης της (π.χ. "Kelvin | Average" για τη θερμοκρασία του αέρα) και τη χρονική διάρκεια στην οποία αναφέρεται κάθε τιμή της (π.χ. "Daily"). Οι κόμβοι αυτοί συνδέονται με τον αντίστοιχο κόμβο με ετικέτα *Series* μέσω μίας σχέσης *HAS_METADATA*. Η οπτικοποίηση των νέων κόμβων που περιγράφηκαν μέσα στον γράφο γνώσης δίνεται στο σχήμα 4.4. Μέσα σε κάθε κόμβο φαίνεται το όνομα της ετικέτας και ο κωδικός του.

Η εισαγωγή των δεδομένων χρονοσειράς για κάθε κλιματική μεταβλητή γίνεται με τη δημιουργία κόμβων με ετικέτα *Observation*, όπως περιγράφηκε στο προηγούμενο κεφάλαιο. Για κάθε παρατήρηση που προκύπτει από την ανάλυση των δεδομένων του dataset που θα

Σχήμα 4.4: *Νέοι κόμβοι Indicator, Series και SeriesMetadata για τις κλιματικές μεταβλητές μέσα στο SustainGraph*

περιγραφεί στις επόμενες ενότητες, δημιουργείται ο αντίστοιχος Observation κόμβος με τα εξής στοιχεία ως ιδιότητες: την τιμή της κλιματικής μεταβλητής στο πεδίο value και την ημερομηνία στην οποία αναφέρεται η τιμή αυτή στο πεδίο time. Για να δοθεί η χωρική τοποθεσία της συγκεκριμένης παρατήρησης, ο κόμβος *Observation* συνδέεται μέσω της σχέσης *REFERS_TO_AREA* με έναν κόμβο *GeoArea*, δηλαδή με μία συγκεκριμένη γεωγραφική περιοχή. Η περιοχή αυτή μπορεί να είναι η περιοχή "Αττική" της ταξινόμησης NUTS επιπέδου 2 ή κάποια από τις περιοχές ταξινόμησης NUTS επιπέδου 3 που περιέχονται σε αυτή.

Για την αναπαράσταση της γεωμετρίας των παρατηρήσεων, αξιοποιούμε, όπως περιγράφηκε στο προηγούμενο κεφάλαιο, τον τύπο δεδομένων POINT του Neo4j. Η γεωμετρία που παίρνουμε από το dataset για κάθε παρατήρηση μίας κλιματικής μεταβλητής είναι ένα σημείο, το οποίο αποτελείται από τις συντεταγμένες του, δηλαδή το γεωγραφικό του πλάτος και το γεωγραφικό του μήκος. Το σημείο αυτό αποθηκεύεται σαν ιδιότητα της σχέσης *REFERS_TO_AREA* μεταξύ του κόμβου *Observation*, με την ημέρα της παρατήρησης και την αντίστοιχη τιμή της κλιματικής μεταβλητής, και του κόμβου *GeoArea* με το όνομα της NUTS περιοχής στην οποία περιέχεται το συγκεκρινένο σημείο. Ένα παράδειγμα ενός τέτοιου *Observation* και της αναπαράστασης των χωρικών του χαρακτηριστικών μέσα στο SustainGraph δίνεται στο σχήμα 4.5. Μέσα σε κάθε κόμβο φαίνονται κάποιες από τις ιδιότητές του. Πάνω στη σχέση *REFERS_TO_AREA* βλέπουμε την τιμή της ιδιότητας point, που είναι μία τιμή τύπου Point.

Σχήμα 4.5: *Χωρική αναπράσταση ενός Observation μέσα στο SustainGraph*

4.2 Μηχανισμοί εισαγωγής χωρικών δεδομένων στο SustainGraph

Για την αναπαράσταση των δεδομένων στο SustainGraph που παρουσιάστηκε στην προηγούμενη ενότητα χρειάζεται να ακολουθηθεί πρώτα μία διαδικασία επεξεργασίας των δεδομένων που λαμβάνουμε από τα διαθέσιμο dataset, καθώς επίσης και η σύνδεση με την βάση δεδομένων με οργάνωση γράφου στο Neo4j και η εφαρμογή των κατάλληλων επερωτήσεων εισαγωγής δεδομένων.

4.2.1 Περιγραφή των δεδομένων

Αρχικά κατεβάζουμε τα αρχεία με τις τιμές των κλιματικών μεταβλητών για την Αθήνα από την αποθήκη δεδομένων για το κλίμα του Copernicus. Πρώτα απαιτείται εγγραφή στην αποθήκη CDS και εγκατάσταση του CDS API κλειδιού. Ύστερα κάνουμε αίτημα στο API ορίζοντας την πόλη της Αθήνας, την επιθυμητή κλιματική μεταβλητή και το έτος και τους μήνες που μας ενδιαφέρουν. Στόχος μας αποτελεί η εισαγωγή όλων των διαθέσιμων δεδομένων για την Αθήνα από το συγκεκριμένο dataset στο SustainGraph. Παρόλα αυτά, πρώτα εφαρμόζουμε τους μηχανισμούς που θα περιγραφούν στη συνέχεια για ένα μέρος του dataset, αυτό που αφορά τη θερμοκρασία του αέρα τον πρώτο διαθέσιμο μήνα, δηλαδή τον Ιανουάριο του 2008.

Φορτώνουμε τα δεδομένα για τη θερμοκρασία του αέρα τον Ιανουάριο του 2008 σε δομή DataFrame και παρατηρούμε τη δομή των δεδομένων. Οι στήλες που αποτελούν το index του πίνακα είναι: δύο στήλες “x” και “y” για τις συντεταγμένες του σημείου στο οποίο αναφέρεται η παρατήρηση, σε EPSG:3035 σύστημα συντεταγμένων, και μία στήλη “time” για την ημερομηνία και ώρα της ημέρας. Οι υπόλοιπες στήλες είναι οι εξής: “latitude” και

“longitude” για το γεωγραφικό μήκος και το γεωγραφικό πλάτος του σημείου αντίστοιχα, σε EPSG:4326 σύστημα συντεταγμένων, και “tas” για τη τιμή της θερμοκρασίας του αέρα στο σημείο αυτό τη συγκεκριμένη ώρα. Το πλήθος των σειρών είναι 66772937 και δεν υπάρχουν null τιμές.

4.2.2 Προεπεξεργασία των δεδομένων

Για να εισάγουμε τα χωρικά δεδομένα στον γράφο γνώσης θα ακολουθήσουμε μία διαδικασία προεπεξεργασίας τους.

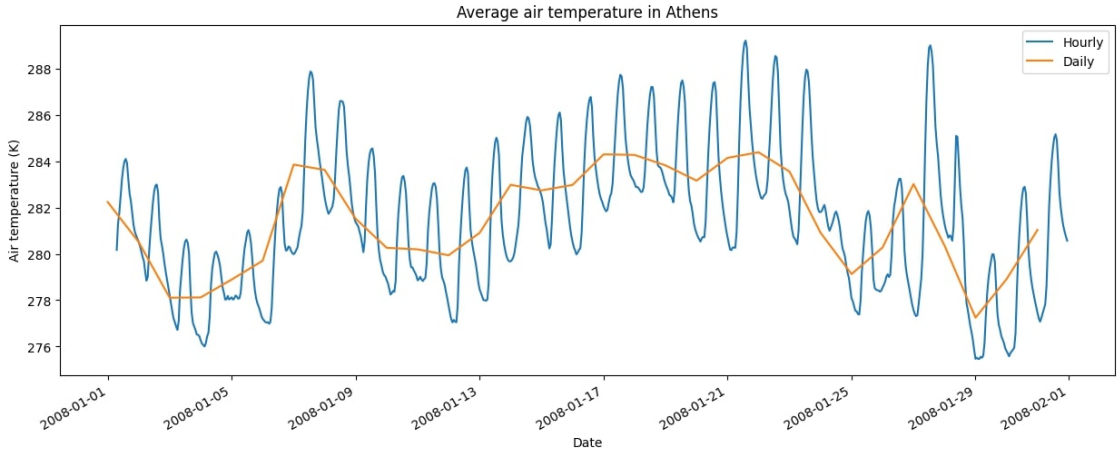
Αρχικά επαναφέρουμε σαν index του DataFrame το προκαθορισμένο, δηλαδή ένα εύρος ακέραιων αριθμών ξεκινώντας από το 0, και καθιστούμε έτσι τις “x”, “y” και “time” ως κανονικές στήλες του πίνακα. Έπειτα κρατάμε μόνο τις στήλες που περιέχουν πληροφορία που θέλουμε να συμπεριλάβουμε στον γράφο γνώσης. Οι στήλες “x” και “y” που αποτελούσαν δύο από τις τρεις στήλες του index δεν μας χρειάζονται καθώς η χωρική πληροφορία που περιέχουν βρίσκεται και στις “latitude” και “longitude”. Τελικά, κάθε σειρά του DataFrame περιέχει τις συντεταγμένες του σημείου, την ώρα της παρατήρησης, και την τιμή της θερμοκρασίας του αέρα.

Ομαδοποίηση σε ημερήσιες τιμές

Όπως αναφέρθηκε προηγουμένως, τα δεδομένα που χρησιμοποιούμε από το dataset περιέχουν τιμές για την θερμοκρασία του αέρα ανά χρονικά διαστήματα μίας ώρας. Η υψηλή αυτή συχνότητα των παρατηρήσεων αυξάνει σε μεγάλο βαθμό τον όγκο των δεδομένων αλλά είναι και περιττή αφού συνήθως η παρακολούθηση των δεικτών γίνεται για την καταγραφή της εξέλιξής τους σε μεγαλύτερο εύρος χρόνου. Επιλέγουμε να ομαδοποιήσουμε τα δεδομένα σε ημερήσια, υπολογίζοντας τον μέσο όρο της θερμοκρασίας για κάθε ημέρα σε κάθε σημείο.

Αρχικά θα εντοπίσουμε ακραίες τιμές της θερμοκρασίας του αέρα, οι οποίες αποκλίνουν σε μεγάλο βαθμό από της υπόλοιπες τιμές των δεδομένων. Αυτές οι τιμές μπορεί να προέρχονται από σφάλμα στην μέτρηση και επιλέγουμε να μην τις συμπεριλάβουμε στον υπολογισμό της μέσης θερμοκρασίας. Για τον εντοπισμό τους εφαρμόζουμε την μέθοδο του ενδοτεταρτημοριακού εύρους στο σύνολο των δεδομένων. Ως άνω όριο του αποδεκτού εύρους τιμών ώστε να μη θεωρούνται ακραίες υπολογίζεται η θερμοκρασία 292.348 K και ως κάτω όριο η θερμοκρασία 270.708 K. Παρατηρούμε ότι δεν υπάρχουν τιμές που ξεπερνούν το πάνω όριο, ενώ το πλήθος των τιμών που είναι μικρότερες από το κάτω όριο είναι 170482. Οι αντίστοιχες σειρές αποτελούν το 0.255% του αρχικού dataset.

Συνεχίζουμε με την ομαδοποίηση των ωριαίων παρατηρήσεων σε ημερήσιες. Από τη στήλη “time” που περιέχει την ημερομηνία και την ώρα δημιουργούμε τη στήλη “date”, η οποία περιέχει μόνο την ημερομηνία. Μπορούμε τώρα να ομαδοποιήσουμε τα δεδομένα που έχουν κοινό “date”, “latitude” και “longitude”, δηλαδή αυτά που αναφέρονται στο ίδιο σημείο και την ίδια ημέρα, και να υπολογίσουμε την μέση τιμή της θερμοκρασίας του αέρα (στήλη “tas”) σε κάθε ομάδα. Καταλήγουμε έτσι με ένα DataFrame που περιέχει τις παρατηρήσεις της μέσης θερμοκρασίας του αέρα για κάθε σημείο του αρχικού dataset, για κάθε ημέρα του Ιανουαρίου του 2008. Στο σχήμα 4.6 βλέπουμε πως μεταβάλλεται η μέση θερμοκρασία όλης της Αθήνας τόσο ανά μία ώρα, όσο και ανά μία ημέρα.



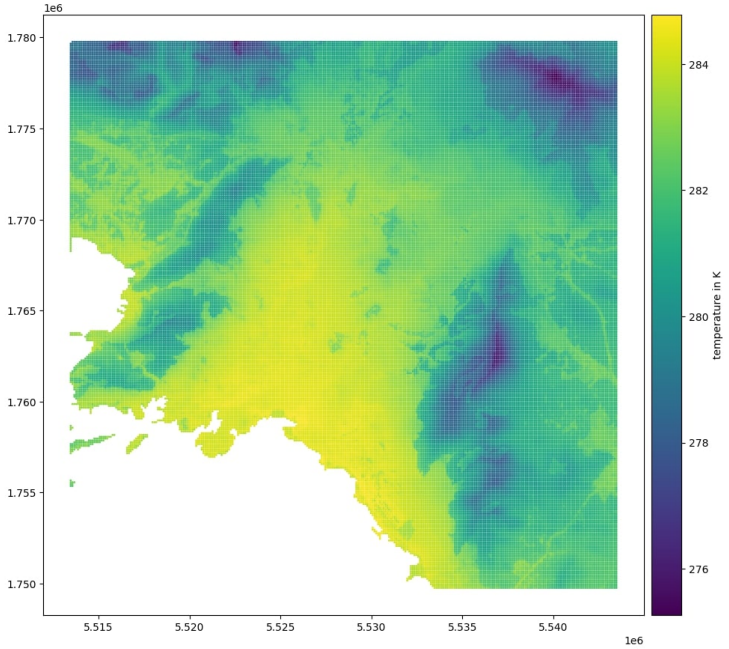
Σχήμα 4.6: *Χρονοσειρά μέσης θερμοκρασίας στην Αθήνα για τον Ιανουάριο του 2008 πριν και μετά την ομαδοποίηση σε ημερήσιες τιμές*

Φιλτράρισμα σημείων στη ξηρά

Έπειτα φιλτράρουμε τα δεδομένα ώστε να κρατήσουμε μόνο σειρές που αφορούν σημεία που βρίσκονται στη ξηρά. Κατεβάζουμε το αρχείο που περιέχει την μάσκα μεταξύ σημείων στη ξηρά και στη θάλασσα από το ίδιο dataset της αποθήκης CDS του Copernicus, ορίζοντας την μεταβλητή landseamask κατά το αίτημα μας στο API. Από αυτό το αρχείο δημιουργούμε μία νέα δομή DataFrame. Οι στήλες του είναι οι εξής: "latitude" και "longitude" για το γεωγραφικό μήκος και το γεωγραφικό πλάτος κάθε σημείου, σε EPSG:4326 σύστημα συντεταγμένων, και "landseamask", που παρουσιάζει την τιμή 1 για σημεία στη ξηρά και null τιμή για σημεία στη θάλασσα. Παρατηρούμε ότι το πλήθος των σειρών είναι 90601, όσο το πλήθος των σημείων της Αθήνας όπου έχουμε παρατηρήσεις για τις κλιματικές μεταβλητές, και ότι υπάρχουν 16195 null τιμές στην στήλη "landseamask", άρα 16195 σημεία βρίσκονται στη θάλασσα και τα υπόλοιπα 74406 στη ξηρά.

Για να κρατήσουμε τις τιμές θερμοκρασίας του αέρα που αναφέρονται σε σημεία ξηράς, συγχωνεύουμε τα δύο DataFrames με την εφαρμογή αριστερής εξωτερικής σύζευξης μεταξύ των δύο. Προκύπτει ένα DataFrame με τις συντεταγμένες των σημείων, τις ημερομηνίες και τις τιμές της θερμοκρασίας του αρχικού, με την επιπλέον στήλη "landseamask", που δείχνει για κάθε σημείο στο οποίο αναφέρεται η παρατήρηση αν είναι ξηράς ή θάλασσας. Τέλος, διαγράφουμε τις σειρές που έχουν null τιμή στην στήλη της μάσκας, καταλήγοντας μόνο με παρατηρήσεις που αφορούν την ξηρά. Διαγράφουμε, επιπλέον, την στήλη της μάσκας καθώς δεν μας χρειάζεται πλέον.

Για την οπτικοποίηση των σημείων στον χάρτη αλλά και για την περεταίρω ανάλυση των χωρικών δεδομένων, δημιουργούμε από το DataFrame μία νέα δομή δεδομένων, ένα GeoDataFrame, ώστε να εκφράσουμε τη γεωμετρία στην οποία αναφέρεται κάθε σειρά, δηλαδή κάθε παρατήρηση. Ορίζουμε ως γεωμετρία κάθε σειράς ένα σημείο, το οποίο προκύπτει από τις συντεταγμένες που βρίσκονται στις στήλες "latitude" και "longitude". Έπειτα αφαιρούμε τις δύο αυτές στήλες. Ορίζουμε, επίσης, πως το σύστημα συντεταγμένων των γεωμετριών είναι το EPSG:4326, το μετατρέπουμε όμως σε EPSG:3035 για την καλύτερη αναπαράστασή τους, όπως αναφέρθηκε στην ενότητα 4.1. Στο σχήμα 4.7 βλέπουμε τον χάρτη με τα σημε-



Σχήμα 4.7: *Χάρτης θερμοκρασίας του αέρα την 1η Ιανουαρίου του 2008 για τα σημεία ξηράς*

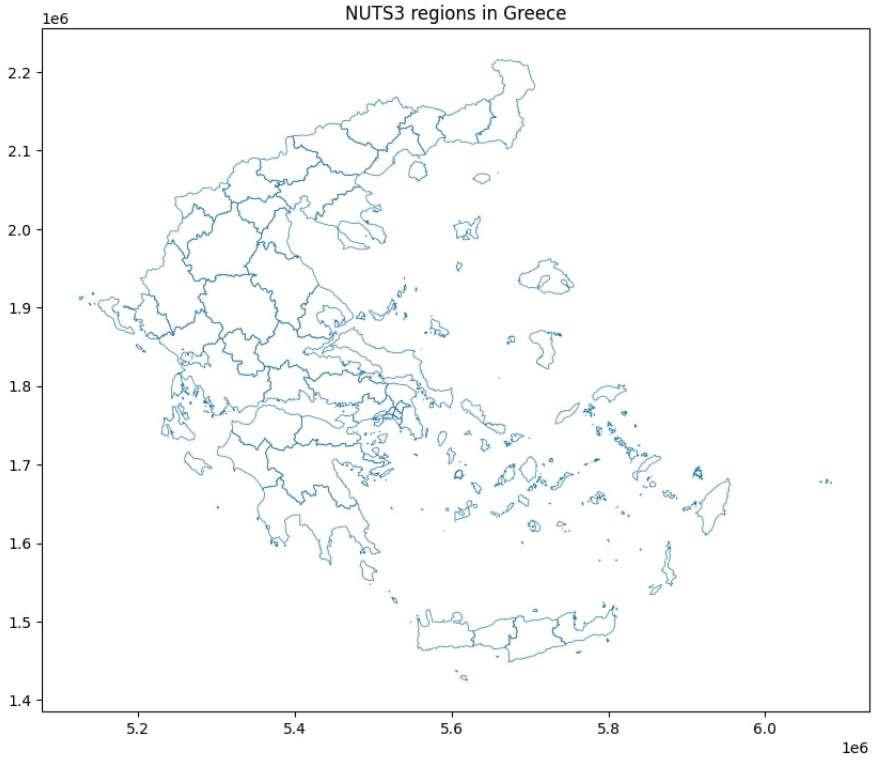
ία που έμειναν μετά το φιλτράρισμα, και τη μέση θερμοκρασία τους την πρώτη ημέρα του Ιανουαρίου.

Χωρισμός σε περιοχές NUTS επιπέδου 3

Κατά την εισαγωγή των δεδομένων στο SustainGraph, είναι προτιμότερο να συνδέσουμε κάθε παρατήρηση με έναν κόμβο GeoArea που περιγράφει τη μικρότερη δυνατή περιοχή, δηλαδή με κόμβο *NUTS3*. Πρέπει, λοιπόν, κατά την προεπεξεργασία των δεδομένων, να προσδιορίσουμε σε ποιά περιοχή NUTS επιπέδου 3 ανήκει κάθε σημείο.

Κατεβάζουμε από την ιστοσελίδα της Ευρωπαϊκής Στατιστικής Υπηρεσίας ένα αρχείο τύπου shapefile που περιέχει τις απαραίτητες πληροφορίες για τις περιοχές NUTS [50]. Επιλέγουμε την πιο πρόσφατη έκδοση του αρχείου, τα πολύγωνα ως τύπο γεωμετρίας, την μικρότερη δυνατή κλίμακα (01 M) και το EPSG:3035 σύστημα συντεταγμένων. Ανοίγουμε το αρχείο αυτό σε μορφή GeoDataFrame και ορίζουμε πως το σύστημα συντεταγμένων των γεωμετριών είναι το EPSG:3035. Κάθε σειρά της δομής αυτής περιέχει ως γεωμετρία ένα πολύγωνο (Polygon) ή μία συλλογή από πολύγωνα (Multipolygon) η οποία εκφράζει την γεωγραφική έκταση που καταλαμβάνει η κάθε περιοχή NUTS. Κρατάμε μόνο τις σειρές που περιέχουν περιοχές NUTS επιπέδου 3 και βρίσκονται στην Ελλάδα. Οι γεωμετρίες τους φαίνονται στον χάρτη που απεικονίζεται στο σχήμα 4.8.

Εφαρμόζουμε χωρική σύζευξη μεταξύ των παρατηρήσεων της θερμοκρασίας που χαρακτηρίζονται χωρικά από ένα σημείο και των NUTS 3 περιοχών που έχουν γεωμετρία πολυγώνων. Χρησιμοποιούμε το κατηγόρημα intersects, το οποίο εξετάζει εάν κάθε σημείο βρίσκεται μέσα (είτε στο εσωτερικό είτε στο περίγραμμα) κάθε πολυγώνου. Το αποτέλεσμα είναι ένα νέο GeoDataFrame που περιέχει τις στήλες (συμπεριλαμβανομένου της γεωμετρίας) του πρώτου, και δύο νέες στήλες με το αναγνωριστικό και το όνομα της NUTS επιπέδου 3 πε-



Σχήμα 4.8: *Γεωμετρίες περιοχών NUTS επιπέδου 3 στην Ελλάδα*

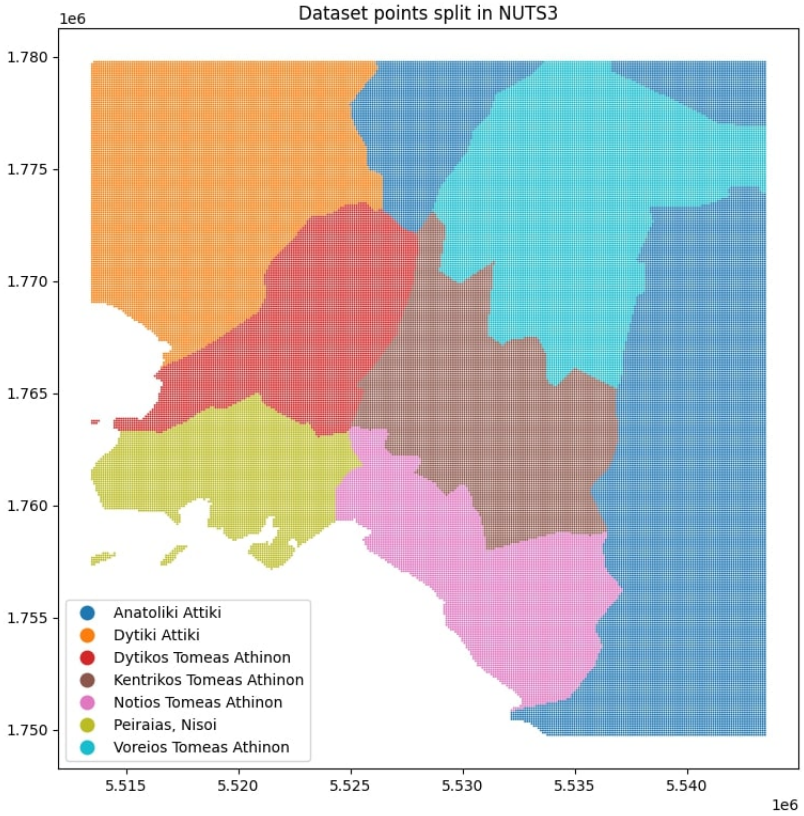
ριοχής στην οποία βρίσκεται το σημείο. Παρατηρούμε πως κάποια από τα σημεία ξηράς για τα οποία είχαμε ημερήσιες τιμές θερμοκρασίας δεν βρίσκονταν μέσα στο πολύγωνο κάποιας περιοχής NUTS επιπέδου 3 και διαγράφηκαν από τη δομή μας. Δεν τα χρειαζόμαστε καθώς μας ενδιαφέρει να εισάγουμε σημεία που συνδέονται με κάποιον από τους κόμβους GeoArea που αναφέραμε. Από τα 74406 σημεία ξηράς που είχαμε, μένουν πλέον 73876. Η οπτικοποίηση των σημείων αυτών, χωρισμένα σε περιοχές NUTS 3 φαίνονται στον χάρτη του σχήματος 4.9

4.2.3 Εισαγωγή στο SustainGraph

Έπειτα από την επεξεργασία των δεδομένων που περιγράφηκε παραπάνω, τα δεδομένα που προέκυψαν μπορούν να εισαχθούν στο SustainGraph. Στην παρούσα ενότητα θα περιγράψουμε τον τρόπο εισαγωγής τους στον γράφο γνώσης, με τη δημιουργία κόμβων με τις κατάλληλες ετικέτες και τιμές ιδιοτήτων. Παρόλα αυτά, τα δεδομένα που τελικά θα εισαχθούν στον γράφο γνώσης αποτελούν αποτέλεσμα ανάλυσης που γίνεται στην ενότητα 4.3.

Εισαγωγή κόμβων τύπου Indicator, Series, SeriesMetadata

Για την εισαγωγή των κόμβων *Indicator*, *Series* και *SeriesMetadata* που αφορούν τη κλιματική μεταβλητή της θερμοκρασία του αέρα, εφαρμόζουμε την κατάλληλη επερώτηση. Η οπτικοποίηση των κόμβων αυτών μέσα στο SustainGraph φαίνεται στο σχήμα 4.4.

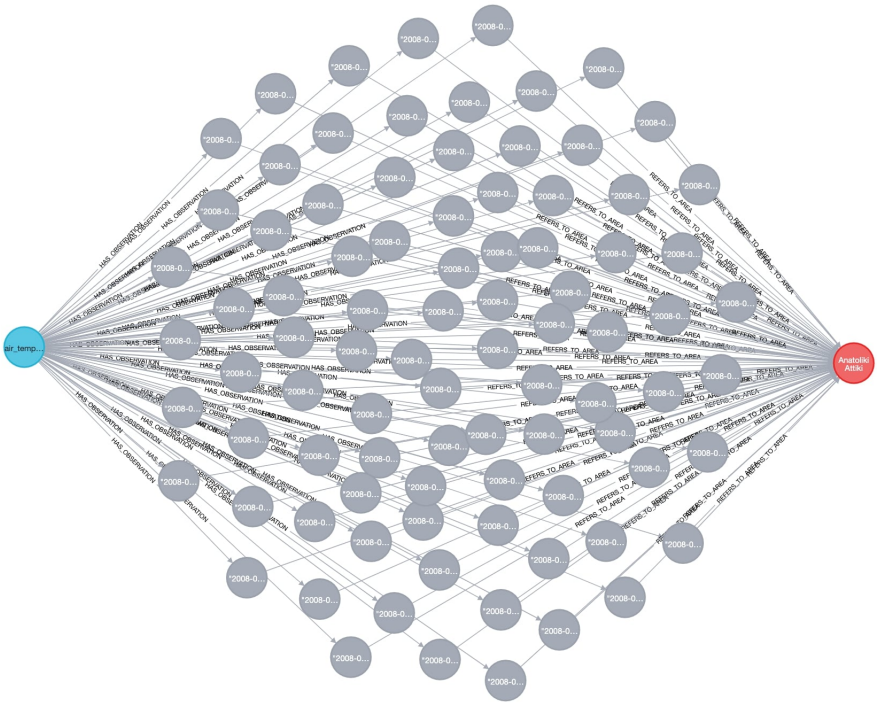


Σχήμα 4.9: *Τα σημεία της Αθήνας χωρισμένα σε περιοχές NUTS επιπέδου 3*

Εισαγωγή κόμβων τύπου Observation

Για την εισαγωγή των κόμβων *Observation* που περιέχουν τις παρατηρήσεις της θερμοκρασία του αέρα, πρώτα μετατρέπουμε τις συντεταγμένες των σημείων τα οποία αφορούν οι παρατηρήσεις στο σύστημα συντεταγμένων WGS84 (EPSG:4326). Ο λόγος για τον οποίο το κάνουμε αυτό είναι διότι, όπως αναφέραμε στο προηγούμενο κεφάλαιο, το Neo4j, στο οποίο έχει υλοποιηθεί το SustainGraph, υποστηρίζει μόνο το WGS84 ως γεωγραφικό σύστημα αναφοράς δύο διαστάσεων. Έπειτα, εφαρμόζουμε την κατάλληλη επερώτηση. Για κάθε γραμμή της δομής δεδομένων που προέκυψε από την ανάλυσή μας, βρίσκουμε το κατάλληλο *SeriesMetadata* σύμφωνα με τον κωδικό της κλιματικής μεταβλητής, και το κατάλληλο *GeoArea* σύμφωνα με την περιοχή NUTS επιπέδου 3 στην οποία ανήκει το σημείο. Δημιουργούμε ένα νέο *Observation* που περιέχει την τιμή της θερμοκρασίας του σημείου και την ημερομηνία της παρατήρησης, και το συνδέουμε μέσω μίας σχέσης *REFERS_TO_AREA*, που έχει ως ιδιότητα το σημείο (με τύπο δεδομένων POINT, με το κατάλληλο *GeoArea*. Επίσης, συνδέουμε το *Observation* με το *SeriesMetadata* μέσω της σχέσης *HAS_OBSERVATION*. Στο σχήμα 4.10 βλέπουμε την οπτικοποίηση κάποιων από του κόμβους *Observation* που εισάγαμε, οι οποίοι προέκυψαν από την ακόλουθη επερώτηση:

```
MATCH (sm:SeriesMetadata{seriesCode:'air_temperature_copernicus'})-
      [:HAS_OBSERVATION]-(o:Observation)-(ga:NUTS3{name:'Anatoliki Attiki'})
WHERE o.time = date({year:  2008, month:  1, day:  31})
RETURN sm,o,ga
```



Σχήμα 4.10: *Κόμβοι Observation για τη θερμοκρασία του αέρα την 31/01/2008 στην Ανατολική Αττική μέσα στο SustainGraph*

Πρόκειται για τους κόμβους *Observation* για τη θερμοκρασία του αέρα που έχουν ως ημερομηνία την 31η Ιανουαρίου και βρίσκονται στην περιοχή της Ανατολικής Αττικής.

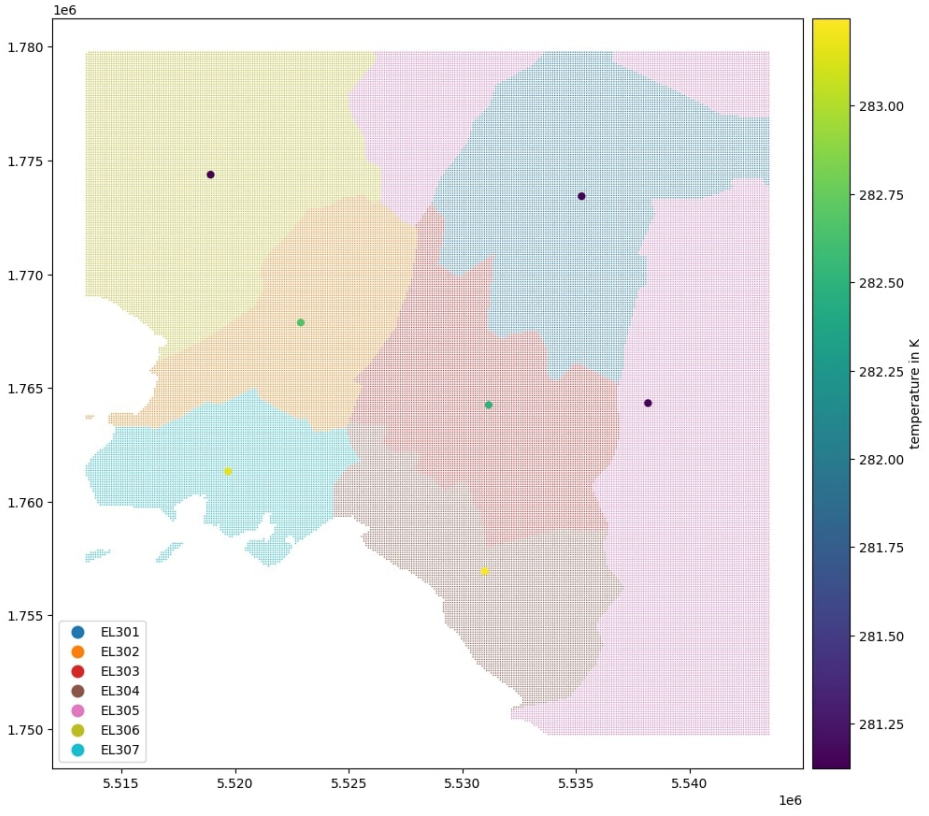
4.3 Μηχανισμοί μείωσης όγκου χωρικών δεδομένων στο SustainGraph

Έπειτα από την προεπεξεργασία των δεδομένων που περιγράφηκε στην προηγούμενη ενότητα, το πλήθος των παρατηρήσεων είναι 2290156. Αποτελούνται από 31 ημερήσιες παρατηρήσεις για κάθε ένα από τα 73876 σημεία ξηράς. Δεδομένου ότι τελικός μας στόχος αποτελεί η εισαγωγή δεδομένων και για τις τέσσερις κλιματικές μεταβλητές του dataset που χρησιμοποιούμε από Copernicus, αλλά και για όλους τους μήνες κάθε έτους από το 2008 έως το 2017, ο όγκος των δεδομένων προς εισαγωγή στο SustainGraph αυξάνεται σε μεγάλο βαθμό. Παράλληλα, έχουμε ήδη μειώσει τις παρατηρήσεις ομαδοποιώντας σε χρονικό επίπεδο. Στην παρούσα ενότητα θα παρουσιάσουμε την υλοποίηση μηχανισμών με τους οποίους μπορούμε να μειώσουμε τον όγκο των παρατηρήσεων, μειώνοντας τα σημεία στα οποία αυτές αναφέρονται.

4.3.1 Διατήρηση κεντροειδούς κάθε NUTS επιπέδου 3 ως εκπρόσωπο της περιοχής

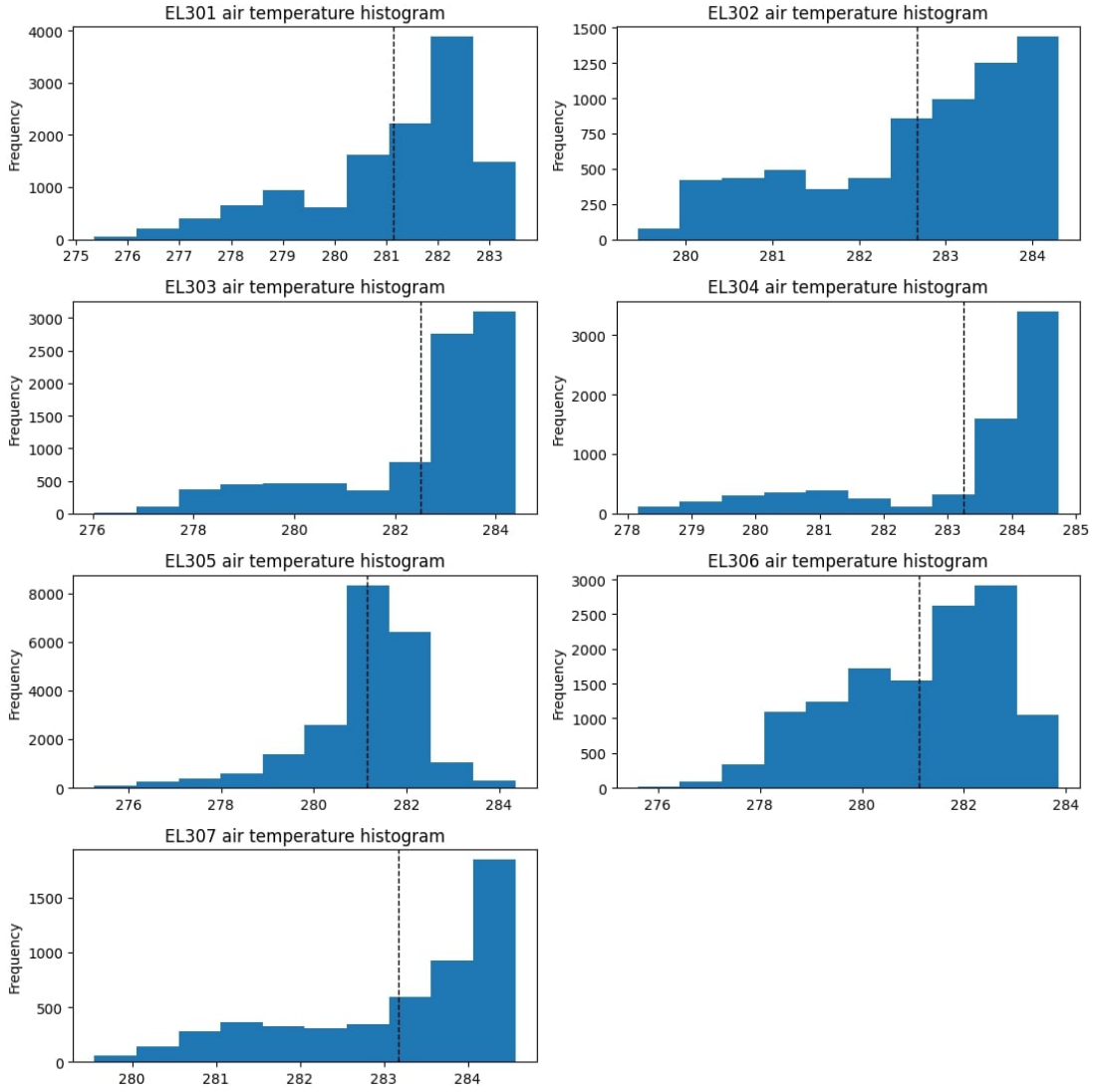
Πρώτα εφαρμόζουμε τον πιο απλό τρόπο μείωσης των σημείων. Επιλέγουμε να κρατήσουμε μόνο ένα σημείο για κάθε NUTS3 ως εκπρόσωπο της περιοχής, το οποίο θα έχει ως θερμοκρασία τη μέση θερμοκρασία όλων των σημείων που περιέχονται σε αυτή. Το σημείο αυτό είναι το κεντροειδές όλων των σημείων της περιοχής. Ομαδοποιούμε τα δεδομένα με κοινή ημερομηνία και αναγνωριστικό NUTS και υπολογίζουμε τη μέση θερμοκρασία τους. Συγχρόνως ομαδοποιούνται και οι γεωμετρίες τους, δηλαδή τα σημεία, και δημιουργείται μία συλλογή σημείων (Multipoint) για κάθε ομάδα. Έπειτα υπολογίζουμε το κεντροειδές κάθε Multipoint και το θέτουμε ως γεωμετρία της γραμμής. Με αυτόν τον τρόπο έχουμε υπολογίσει για κάθε μέρα του μήνα το μέσο όρο της θερμοκρασίας των σημείων που ανήκουν σε κάθε περιοχή NUTS3 και ένα σημείο - το κεντροειδές τους - ως εκπρόσωπο τους.

Στο σχήμα 4.11 φαίνεται στον χάρτη η θερμοκρασία του αέρα την πρώτη μέρα του Ιανουαρίου που έχει υπολογιστεί για κάθε κεντροειδές, καθώς και η περιοχή στην οποία ανήκει, με το αναγνωριστικό της.



Σχήμα 4.11: *Χάρτης θερμοκρασίας του αέρα την 1η Ιανουαρίου 2008 στο κεντροειδές κάθε περιοχής NUTS3*

Θα εξετάσουμε την κατανομή της θερμοκρασίας μέσα σε κάθε περιοχή για μία μέρα του μήνα ώστε να κρίνουμε κατά πόσο η μέση τους τιμή είναι αντιπροσωπευτική. Στο σχήμα 4.12 παρουσιάζεται το ιστόγραμμα της θερμοκρασίας για κάθε περιοχή NUTS 3 την πρώτη μέρα του μήνα. Με διακεκομμένη γραμμή βλέπουμε την μέση τιμή της θερμοκρασίας όλων των σημείων στην αντίστοιχη περιοχή.



Σχήμα 4.12: *Κατανομή των τιμών της θερμοκρασίας του αέρα στα σημεία κάθε περιοχής NUTS επιπέδου 3 για την πρώτη μέρα του μήνα*

Παρατηρούμε πως οι τιμές της θερμοκρασίας του αέρα σε κάθε περιοχή καλύπτουν ένα μεγάλο εύρος τιμών. Επίσης, οι περισσότερες κατανομές εμφανίζουν αρνητική ασυμμετρία, δηλαδή έχουν μεγαλύτερη ουρά προς τα αριστερά. Ομαδοποιώντας όλα τα σημεία μίας ολόκληρης NUTS 3 περιοχής σε ένα με την μέση τιμή της θερμοκρασίας τους, εξομαλύνουμε τιμές της θερμοκρασίας μικρότερων περιοχών μέσα σε αυτήν που μπορεί να είναι σημαντικές για την περεταίρω ανάλυση των δεδομένων σε μικρότερη χωρική κλίμακα. Αντίστοιχα συμπεράσματα προκύπτουν εξετάζοντας τις κατανομές που προκύπτουν από αυτήν τη μέθοδο και για άλλες ημέρες του μήνα. Επίσης, όπως βλέπουμε στον χάρτη του σχήματος 4.11, το κεντροειδές δεν μπορεί να αναπαραστήσει επαρκώς τα χωρικά χαρακτηριστικά των σημείων

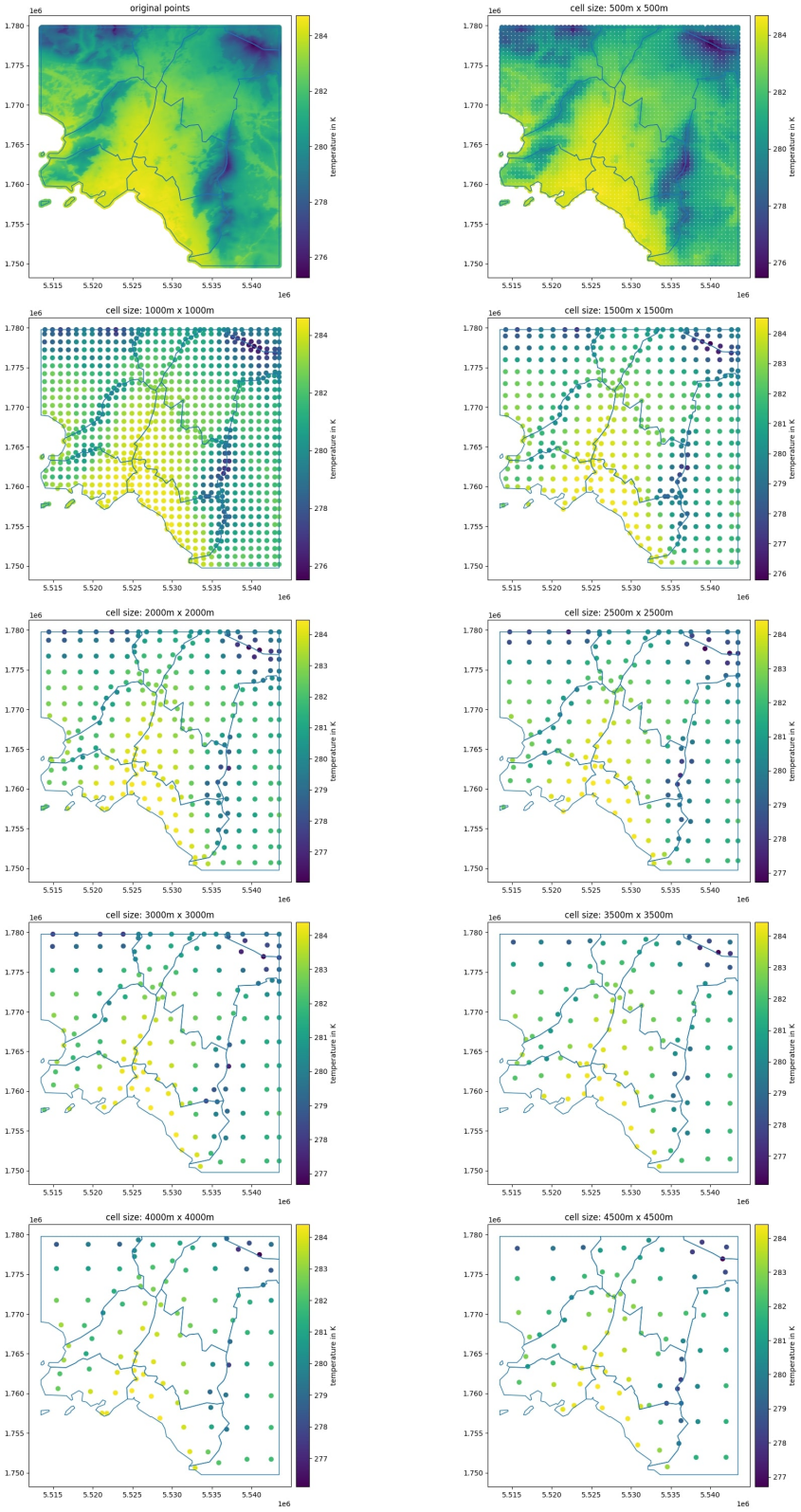
της κάθε περιοχής, καθώς η έκταση που καλύπτουν είναι αρκετά μεγάλη. Μάλιστα, στην EL305 περιοχή η έκταση αυτή δεν είναι συνεχής. Αποτελείται από τρεις περιοχές σημείων EL305 που διαχωρίζονται από μία περιοχή με EL301 σημεία. Ως αποτέλεσμα, το κεντροειδές των σημείων EL305 βρίσκεται μέσα στη μεγαλύτερη από τις τρεις περιοχές και κοντά στα σύνορα με άλλες περιοχές.

4.3.2 Ομαδοποίηση με χρήση πλέγματος

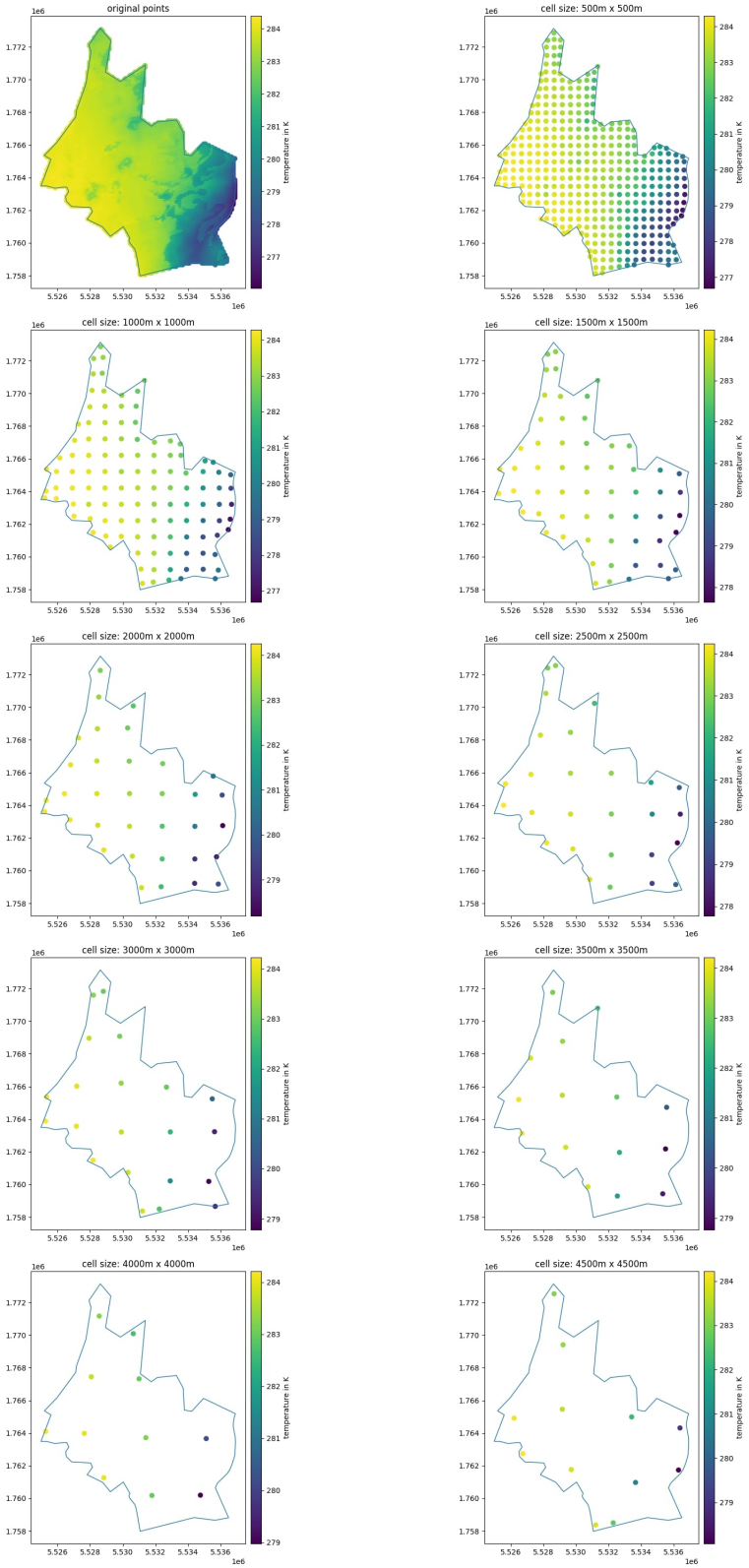
Στη συνέχεια παρουσιάζεται η υλοποίηση του μηχανισμού μείωσης όγκου χωρικών δεδομένων μέσω της ομαδοποίησης με χρήση πλέγματος με τετράγωνα κελιά. Στο τέλος κάθε κελί θα εκπροσωπείται από το κεντροειδές των σημείων που βρίσκονται μέσα στο κελί και θα ανατεθεί σε αυτό η μέση θερμοκρασία τους.

Ορίζουμε μία συνάρτηση που παίρνει σαν παράμετρο όλες τις ημερήσιες παρατηρήσεις για τη θερμοκρασία σε όλα τα σημεία της Αθήνας και το επιθυμητό μέγεθος κελιού. Μέσα στη συνάρτηση αυτή βρίσκουμε τα όρια των σημείων και δημιουργούμε μία λίστα από πολύγωνα που έχουν σχήμα τετραγώνου και μήκος πλευράς αυτό που έχει οριστεί, τα οποία καλύπτουν όλη την έκταση που ορίζεται από τα όρια των σημείων. Έπειτα δημιουργούμε ένα GeoDataFrame που περιέχει τις γεωμετρίες αυτών των πολυγώνων και ορίζουμε πως το σύστημα συντεταγμένων τους είναι το EPSG:3035. Εφαρμόζουμε χωρική σύζευξη μεταξύ των σημείων με τις παρατηρήσεις της θερμοκρασίας και των πολυγώνων του πλέγματος, χρησιμοποιώντας το κατηγόρημα intersects. Πλέον για κάθε παρατήρηση θερμοκρασίας, πέρα από το σημείο στο οποίο αναφέρεται, έχουμε και την πληροφορία του κελιού στο οποίο αυτό περιέχεται. Ομαδοποιούμε τα δεδομένα με κοινή ημερομηνία, αναγνωριστικό κελιού του πλέγματος και αναγνωριστικό περιοχής NUTS 3, και υπολογίζουμε τη μέση θερμοκρασία τους. Συγχρόνως ομαδοποιούνται και τα σημεία και δημιουργείται ένα Multipoint για κάθε ομάδα. Έπειτα υπολογίζουμε το κεντροειδές κάθε Multipoint και το θέτουμε ως γεωμετρία της σειράς. Τέλος, σβήνουμε όσα κεντροειδή βρίσκονται εκτός της αρχικής περιοχής NUTS επιπέδου 3 των σημείων. Με τον τρόπο αυτό μπορούμε να αλλάξουμε τη χωρική ανάλυση των παρατηρήσεων κάθε περιοχής NUTS 3 ώστε αυτές να αφορούν μία έκταση με γεωμετρία τετραγώνου και μέγεθος που εμείς ορίζουμε.

Καλούμε τη συνάρτηση που ορίσαμε για διαφορετικά μεγέθη κελιών και εξετάζουμε τα αποτελέσματα. Συγκεκριμένα, θα εφαρμόσουμε αυτή τη μέθοδο για μήκος πλευράς του κελιού από 500 μέτρα έως 4500 μέτρα με βήμα 500 μέτρων. Στους χάρτες του σχήματος 4.13 βλέπουμε την θερμοκρασία των σημείων που προκύπτουν μετά από κάθε ομαδοποίηση πλέγματος για την πρώτη ημέρα του μήνα. Στους χάρτες του σχήματος 4.14 βλέπουμε την θερμοκρασία των σημείων μετά από κάθε ομαδοποίηση πλέγματος μόνο για τον Κεντρικό Τομέα Αθηνών.

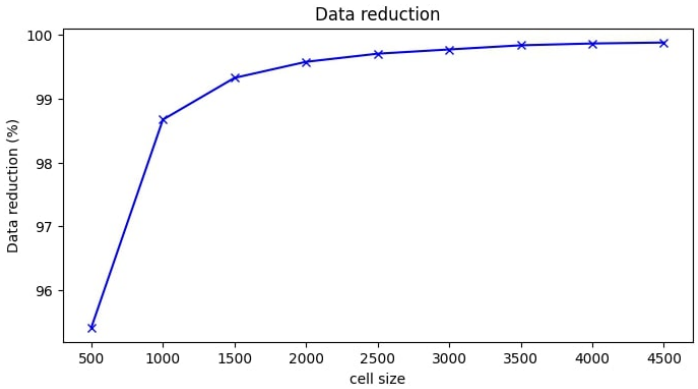


Σχήμα 4.13: *Χάρτες θερμοκρασίας του αέρα την 1η Ιανουαρίου 2008 σε χωρική ανάλυση από 500 έως 4500 μέτρα*



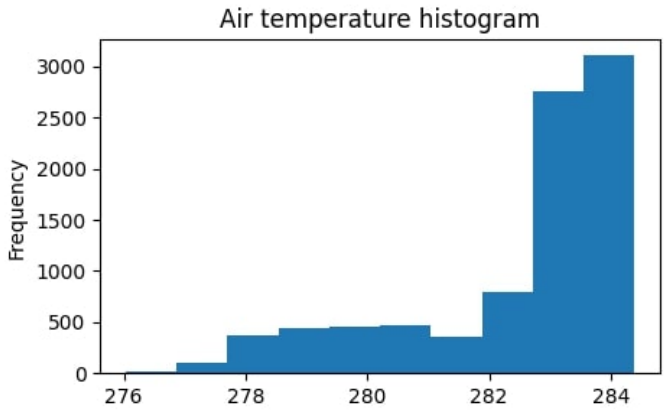
Σχήμα 4.14: *Χάρτες θερμοκρασίας του αέρα την 1η Ιανουαρίου 2008 στον Κεντρικό Τομέα Αθηνών σε χωρική ανάλυση από 500 έως 4500 μέτρα*

Ήδη από την ομαδοποίηση σε κελιά μεγέθους $500m x 500m$ πετυχαίνουμε πολύ μεγάλη μείωση του όγκου των δεδομένων (μείωση κατά 95.419%), ενώ με τη χωρική ανάλυση των 1500 μέτρων πετυχαίνουμε μείωση μεγαλύτερη από 99%. Όσο αυξάνουμε το μέγεθος του κελιού, τόσο μικρότερη βελτίωση παρατηρείται στο ποσοστό μείωσης των δεδομένων. Η μεταβολή του ποσοστού αυτό φαίνεται στο διάγραμμα του σχήματος 4.15.



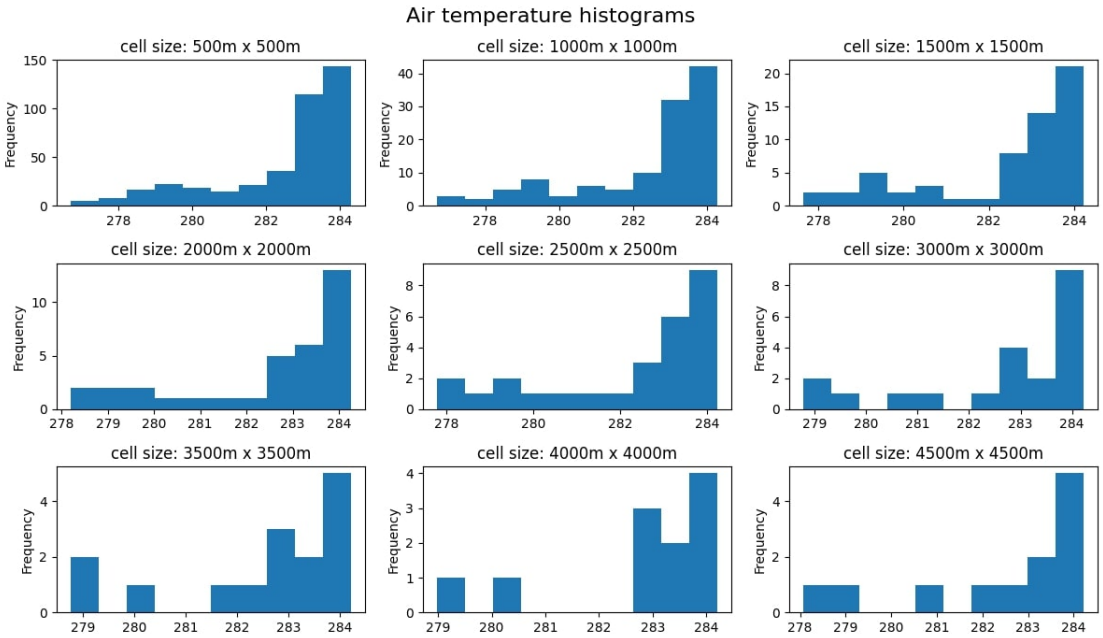
Σχήμα 4.15: *Ποσοστό μείωσης όγκου δεδομένων μετά από ομαδοποίηση των σημείων σε κελιά μεγέθους από 500 έως 4500 μέτρα*

Προκειμένου να εξετάσουμε την απώλεια της πληροφορίας που προκαλείται από την μείωση των δεδομένων, θα συγκρίνουμε την κατανομή της θερμοκρασίας των σημείων που είχαμε αρχικά, με την κατανομή της θερμοκρασίας των σημείων που προκύπτουν μετά την ομαδοποίηση με την χρήση κάθε πλέγματος. Εφόσον κάναμε την ομαδοποίηση των σημείων λαμβάνοντας υπόψη όχι μόνο το κελί του πλέγματος στο οποίο βρίσκονται, αλλά και την περιοχή NUTS 3, τα νέα σημεία δεν είναι πάντα κεντροειδή τετραγώνων. Κοντά στα σύνορα μεταξύ των περιοχών τα σημεία που προκύπτουν είναι πιο πυκνά αφού στο ίδιο κελί περιέχεται το κεντροειδές των σημείων και για τις δύο περιοχές που τέμνονται με αυτό. Για τον λόγο αυτό θα συγκρίνουμε τις κατανομές της θερμοκρασίας των σημείων κάθε περιοχής NUTS 3 ξεχωριστά. Στο σχήμα 4.16 βλέπουμε την κατανομή της θερμοκρασίας των σημείων πριν την ομαδοποίηση, τα οποία βρίσκονται στον Κεντρικό Τομέα Αθηνών, την πρώτη ημέρα του μήνα.



Σχήμα 4.16: *Κατανομή των τιμών της θερμοκρασίας του αέρα στα σημεία του Κεντρικού Τομέα Αθηνών για την 1η Ιανουαρίου του 2008*

Στο σχήμα 4.17 παρουσιάζονται οι κατανομές της θερμοκρασίας στα σημεία του Κεντρικού Τομέα Αθηνών μετά την ομαδοποίηση με κάθε μέγεθος κελιού, για την πρώτη ημέρα του μήνα.



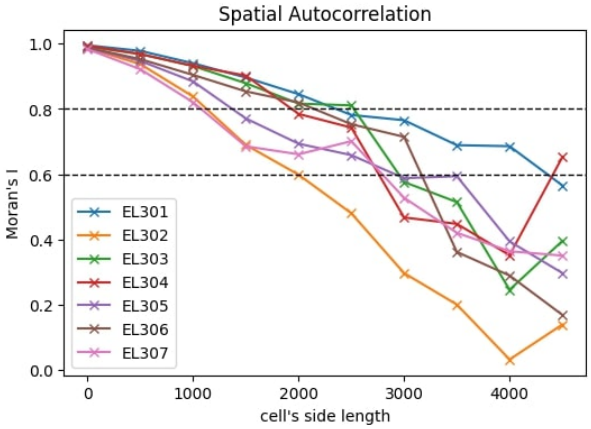
Σχήμα 4.17: *Κατανομή των τιμών της θερμοκρασίας του αέρα στα σημεία του Κεντρικού Τομέα Αθηνών για την 1η Ιανουαρίου του 2008 μετά από κάθε ομαδοποίηση πλέγματος*

Παρατηρούμε ότι οι κατανομές της θερμοκρασίας των σημείων που προέκυψαν σε κάθε περίπτωση ομαδοποίησης με χρήση πλέγματος, μοιάζουν σε μεγάλο βαθμό με αυτή της θερμοκρασίας των αρχικών σημείων. Συγκεκριμένα, μεγαλύτερη συχνότητα παρουσιάζουν τα σημεία με υψηλές θερμοκρασίες, ενώ με χαμηλή συχνότητα παρουσιάζεται ένα μεγάλο εύρος χαμηλών και ενδιάμεσων θερμοκρασιών. Παρατηρούμε, όμως, πως όσο αυξάνεται το μέγεθος κελιού του πλέγματος, τόσο αυξάνεται και η ελάχιστη τιμή της θερμοκρασίας των σημείων και συνεπώς μειώνεται το εύρος των τιμών της θερμοκρασίας. Παράλληλα, στις κατανομές αυξάνεται η συχνότητα των χαμηλών τιμών και μειώνεται αυτή των ενδιάμεσων. Το φαινόμενο αυτό παρατηρούμε οτι γίνεται περισσότερο αισθητό σε μεγαλύτερα μεγέθη κελιών, ειδικά από μέγεθος $2000mx2000m$ και άνω.

Η ίδια ομοιότητα μεταξύ των κατανομών, αλλά και οι μικρές διαφοροποιήσεις από ένα σημείο ομαδοποίησης των σημείων και μετά, παρατηρείται και στις κατανομές των σημείων των υπόλοιπων περιοχών NUTS επιπέδου 3, αλλά και στις κατανομές της θερμοκρασίας των σημείων που προκύπτουν εξετάζοντας και άλλες ημέρες του dataset. Το γεγονός αυτό δείχνει πως δεν υπάρχει μεγάλη απώλεια πληροφορίας κατά την εφαρμογή της συγκεκριμένης μεθόδου, παρόλα αυτά αυξάνεται με την αύξηση του μεγέθους των κελιών.

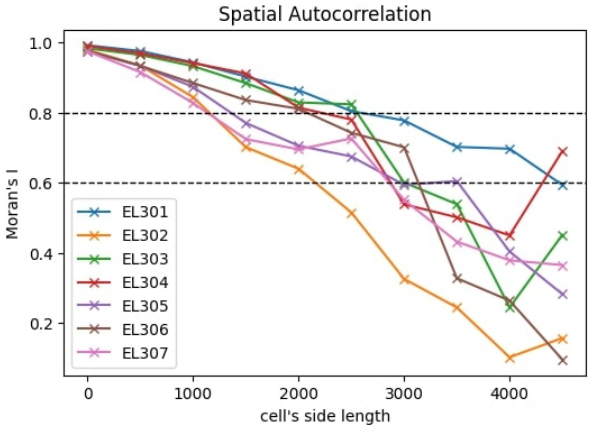
Στη συνέχεια υπολογίζουμε, για μία ημέρα του μήνα και για κάθε περιοχή, την τιμή της χωρικής αυτοσυσχέτισης της θερμοκρασίας των σημείων, τόσο των αρχικών, όσο και αυτών που προέκυψαν μετά από κάθε ομαδοποίηση. Για κάθε περίπτωση χωρικής ανάλυσης, υπολογίζουμε πρώτα τα χωρικά βάρη μεταξύ των σημείων με βάση τις αποστάσεις τους. Συγκεκριμένα, υπολογίζουμε τα βάρη των K κοντινότερων γειτόνων με $K = 4$. Χρησιμοποιώντας

αυτά τα βάρη, υπολογίζουμε την τιμή της μετρικής Moran's I, που μας δείχνει τον βαθμό χωρικής αυτοσυσχέτισης της θερμοκρασίας των σημείων. Στο σχήμα 4.18 παρουσιάζεται το διάγραμμα με τις τιμές της μετρικής Moran's I για την πρώτη ημέρα του Ιανουαρίου. Κάθε γραμμή αντιστοιχεί σε μία περιοχή NUTS 3. Στον οριζόντιο άξονα βρίσκεται το μήκος πλευράς των κελιών του πλέγματος που χρησιμοποιήθηκε σε κάθε ομαδοποίηση, και στον κατακόρυφο άξονα βρίσκεται η τιμή της μετρικής Moran's I. Η πρώτη τιμή κάθε γραμμής (για μηδενικό μήκος πλευράς κελιού) αναφέρεται στα σημεία πριν την ομαδοποίηση.



Σχήμα 4.18: *Χωρική αυτοσυσχέτιση της θερμοκρασίας των σημείων κάθε περιοχής την 1η Ιανουαρίου του 2008 μετά από από ομαδοποίηση των σημείων σε κελιά με μήκος πλευράς από 500 έως 4500 μέτρα*

Παρατηρούμε πως στα αρχικά σημεία υπάρχει πολύ υψηλή θετική χωρική αυτοσυσχέτιση, δηλαδή κοντινά μεταξύ τους σημεία έχουν κοντινές τιμές θερμοκρασίας, ενώ μακρινές τιμές θερμοκρασίες τείνουν να βρίσκονται σε μακρινά σημεία. Όσο χαμηλώνουμε την χωρική ανάλυση μέσω της ομαδοποίησης των σημείων με τη χρήση πλέγματος, η τιμή του Moran's I μειώνεται, δηλαδή η χωρική αυτοσυσχέτιση γίνεται πιο ασθενής. Αυτό εξηγείται από το γεγονός πως τα κοντινά μεταξύ τους σημεία του αρχικού dataset, τα οποία όπως παρατηρήσαμε, έχουν κοντινές τιμές θερμοκρασίας, ομαδοποιούνται σε ένα σημείο, το κεντροειδές τους. Τα



Σχήμα 4.19: *Χωρική αυτοσυσχέτιση της θερμοκρασίας των σημείων κάθε περιοχής την 15η Ιανουαρίου του 2008 μετά από από ομαδοποίηση των σημείων σε κελιά με μήκος πλευράς από 500 έως 4500 μέτρα*

κεντροειδή που προκύπτουν αντιπροσωπεύουν τη μέση θερμοκρασία εκτάσεων με γεωμετρία τετραγώνου και δεν έχουν τόσο ισχυρή χωρική αυτοσυσχέτιση μεταξύ τους. Παρατηρούμε πως ενώ η τιμή της μετρικής Moran's I μειώνεται, διατηρείται σε πολύ υψηλές τιμές (πάνω από 0.8) για μήκος πλευράς κελιού έως 1000 μέτρα, και σε αρκετά υψηλές τιμές (πάνω από 0.6) για μήκος πλευράς κελιού έως 2000 μέτρα.

Αντίστοιχα συμπεράσματα προκύπτουν κι από τον υπολογισμό της χωρικής αυτοσυσχέτισης πριν και μετά την ομαδοποίηση και για άλλες ημέρες του μήνα. Για παράδειγμα, στο σχήμα 4.19 παρουσιάζεται το αντίστοιχο διάγραμμα για την 15η Ιανουαρίου.

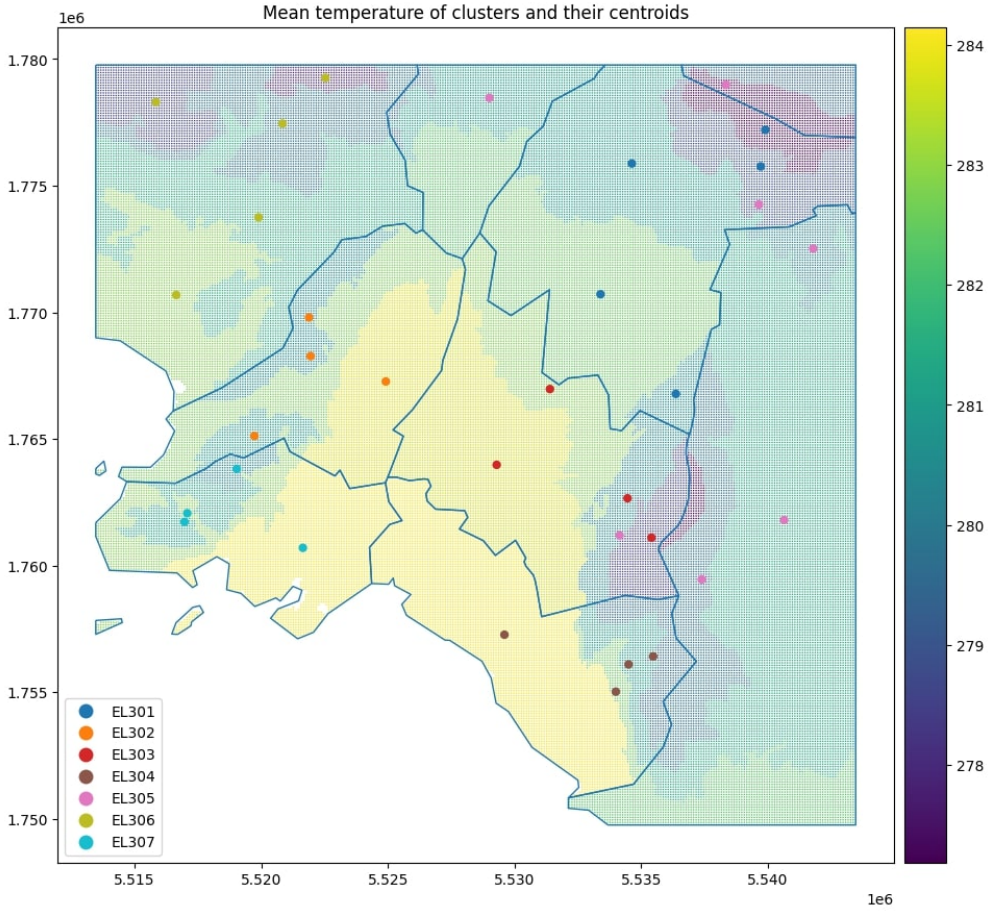
4.3.3 Ομαδοποίηση μέσω δημιουργίας συστάδων με χωρικό περιορισμό

Στη συνέχεια παρουσιάζεται η υλοποίηση του μηχανισμού μείωσης όγκου χωρικών δεδομένων μέσω της δημιουργίας συστάδων με χωρικό περιορισμό.

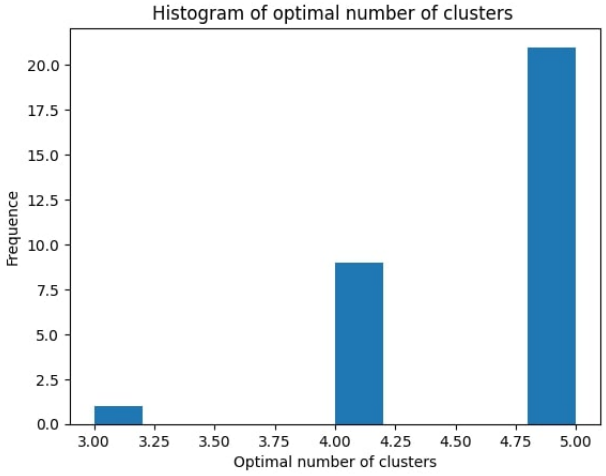
Αρχικά εκτελούμε τον αλγόριθμο συσσωρευτικής ιεραρχικής συσταδοποίησης (AHC) με χωρικό περιορισμό για τις παρατηρήσεις της θερμοκρασίας της πρώτης ημέρας του dataset. Ως χωρικό περιορισμό θέτουμε τη συνεκτικότητα των συστάδων με βάση τα χωρικά βάρη των K κοντινότερων γειτόνων με $K = 4$. Υπολογίζουμε τα βάρη αυτά για τα σημεία κάθε περιοχής NUTS 3. Έπειτα, εφαρμόζουμε την μέθοδο του αγκώνα για να βρούμε τον βέλτιστο αριθμό συστάδων για κάθε περιοχή. Εκτελούμε τον αλγόριθμο AHC και υπολογίζουμε την τιμή του WCSS για κάθε συσταδοποίηση με αριθμό συστάδων στο εύρος από 2 έως 12. Ύστερα, εφαρμόζουμε τον αλγόριθμο Kneedle για να βρούμε το σημείο του αγκώνα κάθε περιοχής στο εύρος αυτό. Εκτελούμε ξανά τον αλγόριθμο AHC για κάθε περιοχή NUTS 3 ορίζοντας ως πλήθος συστάδων για κάθε μία αυτό που βρήκαμε από την μέθοδο του αγκώνα. Τέλος, βρίσκουμε το κεντροειδές σημείο της κάθε συστάδας που προέκυψε και υπολογίζουμε τη μέση θερμοκρασία των σημείων του.

Στο σχήμα 4.20 φαίνεται ο χάρτης των σημείων της Αθήνας. Κάθε σημείο έχει την τιμή της μέσης θερμοκρασίας της συστάδας στο οποίο ανήκει. Οπτικοποιούμε, επίσης, το κεντροειδές κάθε συστάδας με χρώμα που δείχνει σε ποια περιοχή NUTS 3 ανήκει. Παρατηρούμε πως τα κεντροειδή κάποιων συστάδων βρίσκονται έξω από την περιοχή των σημείων που τα αποτελούν. Αυτό συμβαίνει σε μία συστάδα της περιοχής 'EL302', δηλαδή του Δυτικού Τομέα Αθηνών, όπου λόγω της γεωμετρίας του συνόλου των σημείων του, το κεντροειδές του βρίσκεται χωρικά μέσα στα όρια μίας άλλης συστάδας της ίδιας περιοχής. Παρατηρείται, επίσης, ακόμα πιο έντονα, στην περιοχή 'EL305', δηλαδή στην Ανατολική Αττική, όπου λόγω της γεωμετρίας της, που είναι χωρισμένη σε τρία πολύγωνα, κάποιες από τις συστάδες της εκτείνονται σε περισσότερα από ένα πολύγωνα. Ως αποτέλεσμα, τα κεντροειδή τους βρίσκονται χωρικά όχι μόνο εκτός των αντίστοιχων συστάδων, αλλά και εκτός της Ανατολικής Αττικής. Το γεγονός αυτό δείχνει πως η χωρική αναπαράσταση των συστάδων από τα κεντροειδή των σημείων τους δεν είναι αντιπροσωπευτική.

Έπειτα εξετάζουμε κατά πόσο οι συστάδες στις οποία χωρίζεται κάθε περιοχή είναι όμοιες για κάθε ημέρα του dataset. Θα συγκρίνουμε τις συστάδες που προκύπτουν από τον αλγόριθμο AHC με χωρικό περιορισμό για κάθε ημέρα του Ιανουαρίου του 2008 για την περιοχή του Κεντρικού Τομέα Αθηνών. Αρχικά εφαρμόζουμε την μέθοδο του αγκώνα για να βρούμε τον βέλτισμο αριθμό συστάδων για τις παρατηρήσεις της θερμοκρασίας κάθε η-



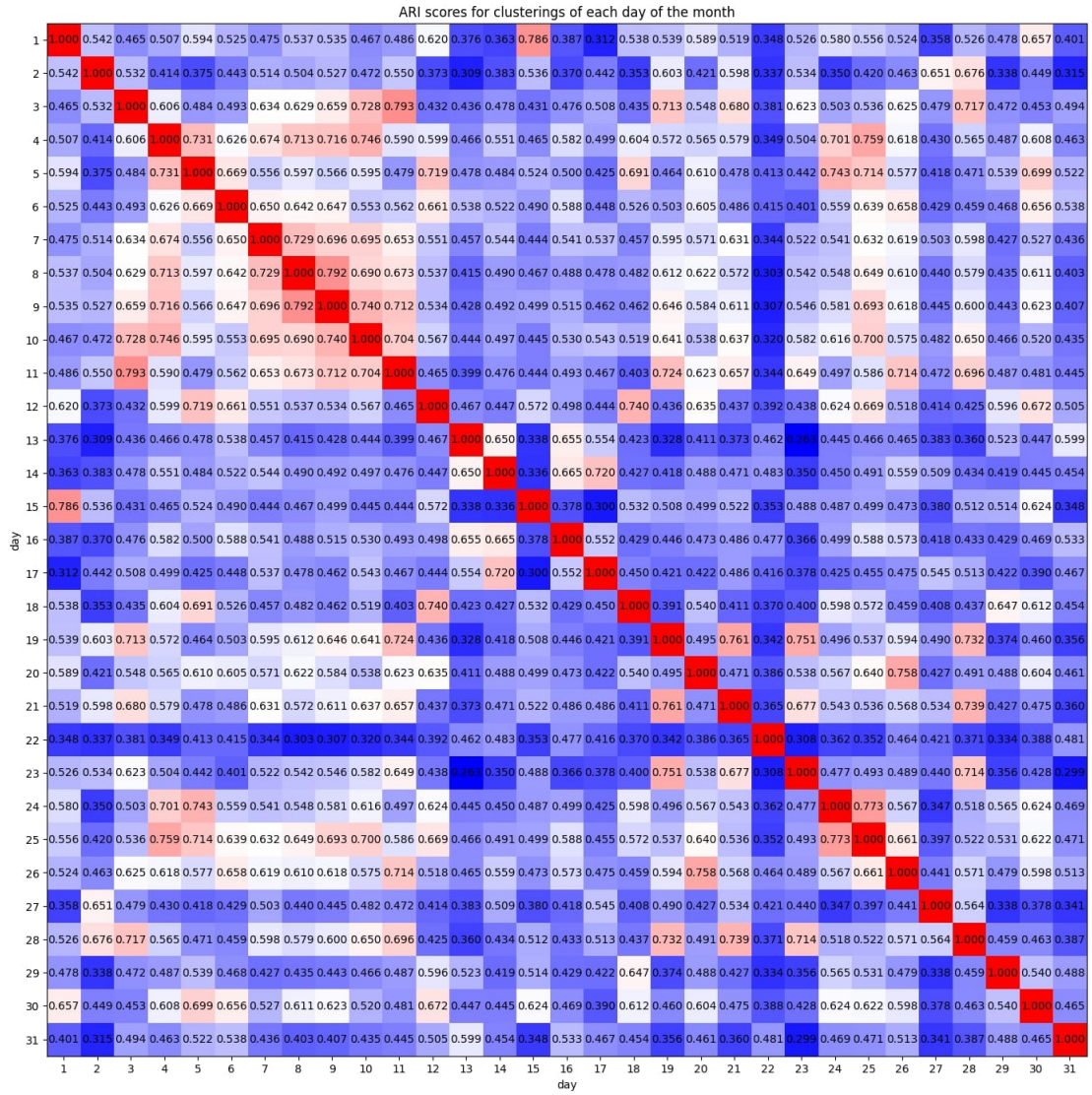
Σχήμα 4.20: *Χάρτης μέσης θερμοκρασίας κάθε συστάδας την 1η Ιανουαρίου του 2008*



Σχήμα 4.21: *Ιστόγραμμα βέλτιστου αριθμού συστάδων για τις ημερήσιες παρατηρήσεις θερμοκρασίας των σημείων του Κεντρικού Τομέα Αθηνών*

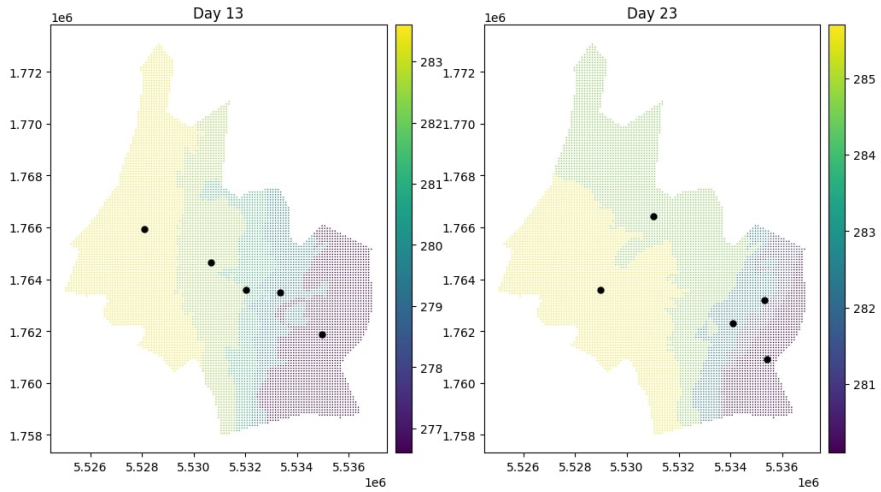
μέρας. Για τις περισσότερες ημέρες, ο βέλτιστος αριθμός είναι 5 συστάδες, ενώ για κάποιες είναι 4 συστάδες, και για μία είναι 3. Η συχνότητα των βέλτιστων αριθμών συστάδων για τις ημερήσιες παρατηρήσεις των σημείων φαίνεται και στο ιστόγραμμα του σχήματος 4.21. Η μέση τιμή του βέλτιστου αριθμού συστάδων είναι 5.

Παρόλο που ο βέλτιστος αριθμός συστάδων που υπολογίστηκε από την μέθοδο του αγκώνα δεν είναι ο ίδιος για όλες τις ημέρες όλων παρατηρήσεων, χρησιμοποιούμε τη μέση τιμή του και χωρίζουμε τις παρατηρήσεις κάθε ημέρας σε 5 συστάδες. Για να συγκρίνουμε την ομοιότητα μεταξύ των συστάδων που δημιουργούνται κάθε ημέρα, υπολογίζουμε την μετρική ARI μεταξύ των συσταδοποιήσεων των ημερών ανά 2. Οι τιμές για κάθε ζευγάρι φαίνονται στον πίνακα του σχήματος 4.22.



Σχήμα 4.22: *Πίνακας μετρικών ARI μεταξύ των συσταδοποιήσεων των παρατηρήσεων της θερμοκρασίας των ημερών του μήνα ανά δύο*

Παρατηρούμε ότι η μετρική ARI παίρνει κυρίως μεσαίες και χαμηλές θετικές τιμές (0.26-0.65), ενώ λίγα ζεύγη ημερών έχουν συσταδοποιήσεις με σχετικά υψηλή τιμή ARI (πάνω από 0.75). Προκύπτει, λοιπόν, το συμπέρασμα πως οι συστάδες που δημιουργούνται από τον αλγόριθμο AHC με χωρικό περιορισμό για κάθε ημέρα διαφέρουν σε μεγάλο βαθμό μεταξύ τους. Στο σχήμα 4.23 παρουσιάζονται οι χάρτες με τις συστάδες για τις δύο ημέρες που είχαν το χαμηλότερο ARI (0.263). Το χρώμα των σημείων δείχνει τη μέση θερμοκρασία της συστάδας στην οποία ανήκουν και τα μαύρα σημεία αναπαριστούν τα κεντροειδή τους.

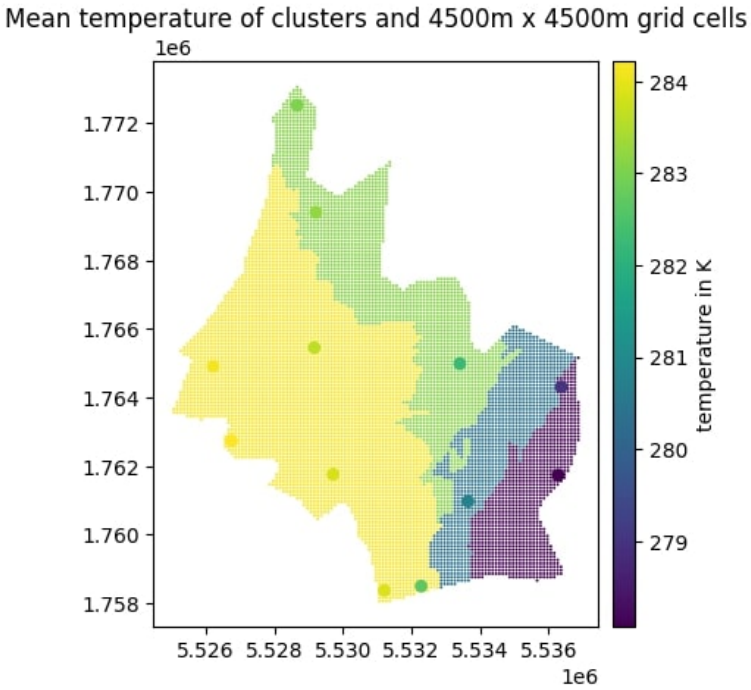


Σχήμα 4.23: *Χάρτες μέσης θερμοκρασίας κάθε συστάδας τις δύο ημέρες του μήνα με τις πιο ανόμοιες συσταδοποιήσεις*

Πράγματι, παρατηρούμε πως οι συστάδες σε κάθε χάρτη αποτελούνται απο πολύ διαφορετικό σύνολο σημείων, άρα έχουν διαφορετική γεωμετρία, και ως αποτέλεσμα τα κεντροειδή τους βρίσκονται σε διαφορετική τοποθεσία.

Τόσο λόγω της ανεπαρκούς χωρικής αναπράστασης των συστάδων από τα κεντροειδή τους, όσο και λόγω των ανόμοιων συστάδων σημείων που δημιουργούνται για κάθε ημέρα του μήνα, η συγκεκριμένη μέθοδος ομαδοποίησης των σημείων δεν είναι κατάλληλη για το συγκεκριμένο dataset.

Προκειμένου να συγκρίνουμε τα αποτελέσματα της ομαδοποίησης με συσταδοποίηση και της ομαδοποίησης των σημείων με την χρήση πλέγματος, συγκρίνουμε τους χάρτες της θερμοκρασίας μεταξύ αυτών. Συγκεκριμένα, θα συγκρίνουμε ενδεικτικά τον χάρτη που προκύπτει για την πρώτη ημέρα του Ιανουαρίου του 2008 στην περιοχή του Κεντρικού Τομέα Αθηνών με κάθε έναν από τους δύο μηχανισμούς. Για τον μηχανισμό με χρήση πλέγματος, επιλέγουμε το μεγαλύτερο μέγεθος κελιών που χρησιμοποιήσαμε, καθώς είναι και αυτό που έχει ως αποτέλεσμα τη μεγαλύτερη απώλεια πληροφορίας. Στον χάρτη του σχήματος 4.24 παρουσιάζονται οι συστάδες με χρώμα που εκφράζει τη μέση θερμοκρασία των σημείων τους τη συγκεκριμένη ημέρα, καθώς επίσης και τα σημεία που προέκυψαν από την ομαδοποίηση με χρήση κελιών με μήκος πλευράς $4500m$. Το χρώμα τους δείχνει τη τιμή της θερμοκρασίας που τα χαρακτηρίζει μετά την ομαδοποίηση, δηλαδή τη μέση θερμοκρασία των σημείων που βρίσκονται στο κελί τους. Στο σχήμα αυτό παρατηρούμε πως τα σημεία της ομαδοποίησης πλέγματος που περιέχονται χωρικά σε κάθε συστάδα που έχει προκύψει από τα αρχικά, μη ομαδοποιημένα, σημεία, έχουν χρώμα (δηλαδή τιμές θερμοκρασίας) όμοιο με αυτό των αντίστοιχων συστάδων. Τα σημεία των οποίων το χρώμα αποκλίνει περισσότερο, βρίσκονται χωρικά κοντά στα σύνορα μεταξύ δύο συστάδων, και το χρώμα τους, δηλαδή η τιμή της θερμοκρασίας τους, είναι μεταξύ αυτών που έχουν οι δύο συστάδες. Σύμφωνα με αυτές τις παρατηρήσεις, και τις αντίστοιχες που προκύπτουν εξετάζοντας άλλες ημέρες του dataset, συμπεραίνουμε πως τα αποτελέσματα των δύο μεθόδων συμφωνούν. Παρόλα αυτά, για τους λόγους που αναφέραμε πριν, η χρήση πλέγματος είναι πιο κατάλληλη μέθοδος για τη μείωση του όγκου των δεδομένων.



Σχήμα 4.24: *Χάρτες μέσης θερμοκρασίας των συστάδων και των σημείων της ομαδοποίησης με κελιά* 4500*m*x4500*m την 1η Ιανουαρίου του 2008*

4.4 Μηχανισμοί ανάλυσης χωρικών δεδομένων στο SustainGraph

Η αναπαράσταση γεωμετριών στο SustainGraph με τον τρόπο που περιγράφηκε στις προηγούμενες ενότητες δίνει τη δυνατότητα σχεδιασμού και υλοποίησης διαδικασιών ανάλυσης των χωρικών δεδομένων που γίνονται διαθέσιμα μέσω του SustainGraph. Σε αυτήν την ενότητα παρουσιάζεται η υλοποίηση των μηχανισμών ανάλυσης χωρικών δεδομένων που αναλύθηκαν στο προηγούμενο κεφάλαιο, και η εφαρμογή τους σε χωρικά δεδομένα με γεωμετρία σημείου, τα οποία υποστηρίζονται από το SustainGraph.

Ανάλυση χωρικών δεδομένων για το κλίμα της Αθήνας

Στόχος μας είναι να εξετάσουμε τη σχέση της θερμοκρασίας του αέρα στα σημεία της Αθήνας με άλλους δείκτες οι οποίοι σχετίζονται με τους SDG στόχους, έχουν χωρικά χαρακτηριστικά και αφορούν την ίδια περιοχή. Οι δείκτες αυτοί μπορούν, όπως και ο δείκτης της θερμοκρασίας, να προέρχονται από το SustainGraph, μετά από τον εμπλουτισμό του με περισσότερα χωρικά δεδομένα. Στο συγκεκριμένο παράδειγμα ανάλυσης, όμως, χρησιμοποιούμε χωρικά δεδομένα που παίρνουμε από εξωτερικές πηγές. Συγκεκριμένα, θα χρησιμοποιήσουμε τους παρακάτω δείκτες:

- Τη σχετική υγρασία στην περιοχή της Αθήνας, που είναι διαθέσιμη από το dataset των κλιματικών μεταβλητών της αποθήκης CDS του Copernicus, το οποίο αξιοποιήσαμε και για τη θερμοκρασία του αέρα.

- Την κάλυψη από δέντρα (Tree Cover Density (TCD)) στην περιοχή της Αθήνας, που

είναι διαθέσιμη από την υπηρεσία παρακολούθησης ξηράς του προγράμματος Copernicus.

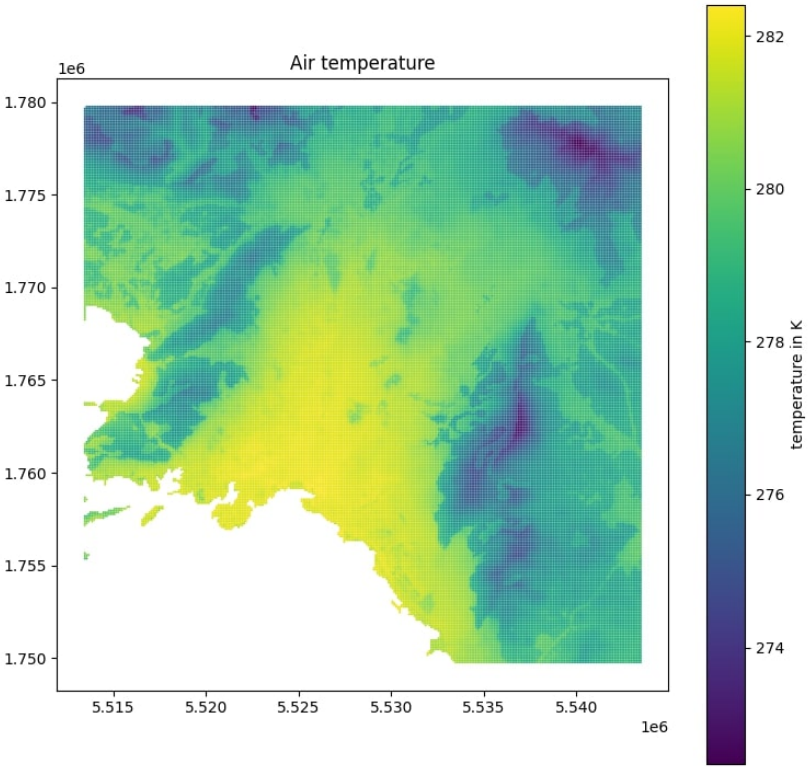
- Την κάλυψη από αδιαπέρατη επιφάνεια (Imperviousness) στην περιοχή της Αθήνας, που είναι επίσης διαθέσιμη από την υπηρεσία παρακολούθησης ξηράς του προγράμματος Copernicus.

- Την πληροφορία για το ύψος των κτιρίων (Building height) στην περιοχή της Αθήνας, που είναι επίσης διαθέσιμη από την υπηρεσία παρακολούθησης ξηράς του προγράμματος Copernicus.

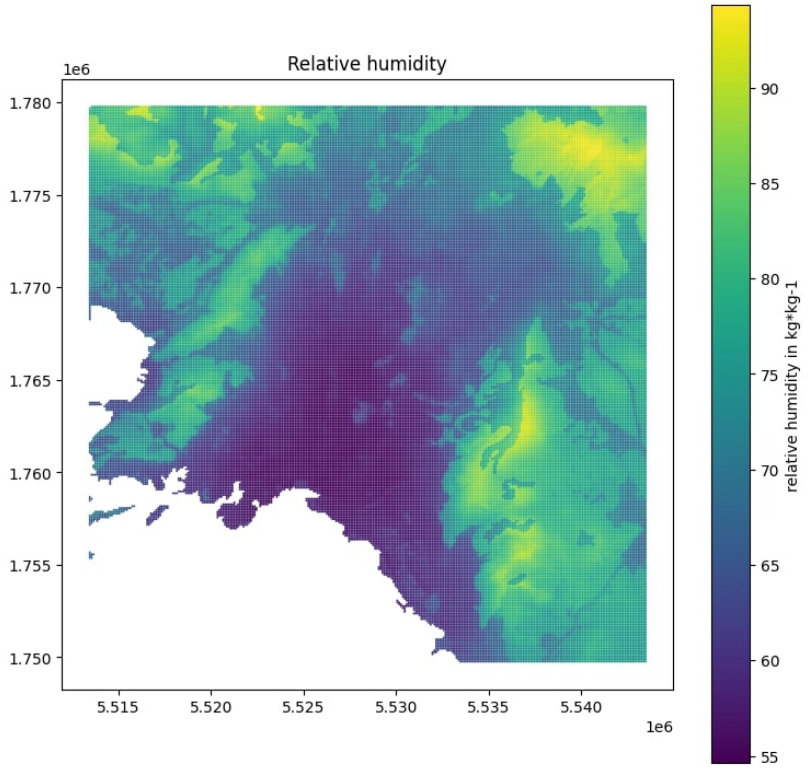
Θα εξετάσουμε τη σχέση τους για τον Ιανουάριο του 2012, έτος για το οποίο έχουμε πληροφορία για όλους τους παραπάνω δείκτες.

Για τις κλιματικές μεταβλητές της θερμοκρασία του αέρα και της σχετικής υγρασίας ακολουθούμε τους μηχανισμούς προεπεξεργασίας των δεδομένων που περιγράφηκαν στις προηγούμενες ενότητες. Συγκεκριμένα, φιλτράρουμε τα διαθέσιμα σημεία ώστε να κρατήσουμε μόνο σημεία ξηράς, και θέτουμε το σημείο που προκύπτει από το γεωγραφικό πλάτος και γεωγραφικό μήκος κάθε παρατήρησης ως τη γεωμετρία της. Ορίζουμε, επίσης, πως το σύστημα συντεταγμένων των γεωμετριών είναι το EPSG:4326, και ύστερα το μετατρέπουμε σε EPSG:3035. Επιπλέον, χωρίζουμε τα σημεία των παρατηρήσεων σε περιοχές NUTS επιπέδου 3 ώστε να έχουμε την πληροφορία της περιοχής στην οποία ανήκουν. Τέλος, για κάθε σημείο ομαδοποιούμε τις παρατηρήσεις όλου του μήνα σε μία και υπολογίζουμε τη μέση θερμοκρασία του αέρα και τη μέση σχετική υγρασία. Αποθηκεύουμε και τις δύο αυτές κλιματικές μεταβλητές σε ένα GeoDataFrame που περιέχει μία στήλη για κάθε κλιματική μεταβλητή, μία στήλη με τη γεωμετρία στην οποία αναφέρεται κάθε παρατήρηση (δηλαδή ένα σημείο ξηράς), και δύο στήλες για το όνομα και το αναγνωριστικό της περιοχής NUTS 3 στην οποία περιέχεται το σημείο. Οι χάρτες με τη μέση θερμοκρασία και τη μέση σχετική υγρασία του Ιανουαρίου του 2012 για κάθε σημείο ξηράς της Αθήνας φαίνονται στα σχήματα 4.25 και 4.26 αντίστοιχα.

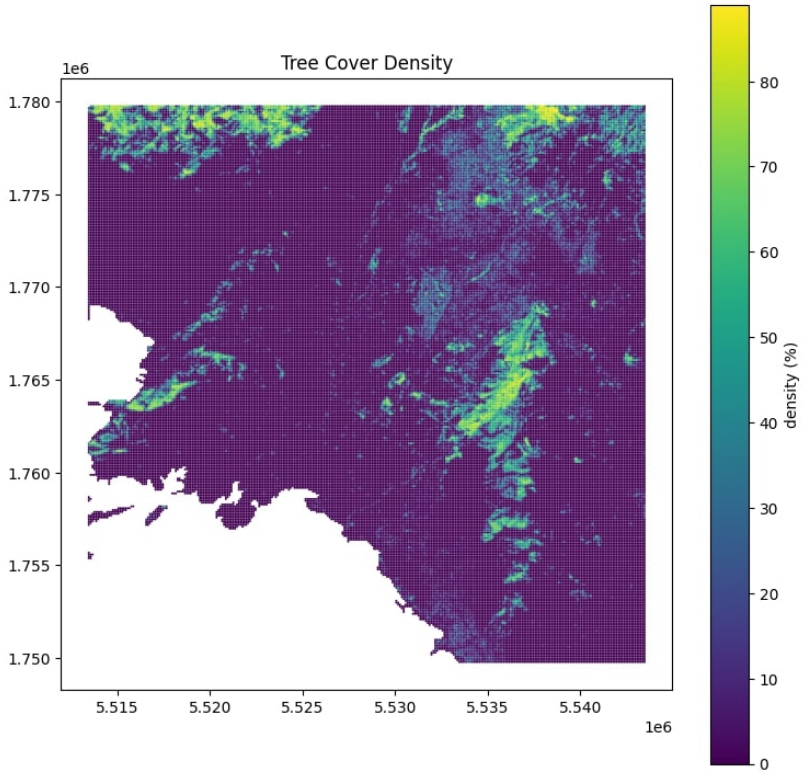
Τα δεδομένα για τους υπόλοιπους τρεις δείκτες παρέχονται από την υπηρεσία παρακολούθησης ξηράς του Copernicus σε κανονικοποιημένη ψηφιδωτή (raster) μορφή. Για κάθε ένα από τα τρία αυτά dataset βρίσκουμε την τιμή που έχει ο αντίστοιχος δείκτης στα σημεία ξηράς που μας ενδιαφέρουν. Αποθηκεύουμε και αυτούς τους δείκτες στο ίδιο GeoDataFrame με τη θερμοκρασία του αέρα και τη σχετική υγρασία, σε τρεις νέες στήλες. Οι χάρτες που προκύπτουν για την κάλυψη από δέντρα, την κάλυψη από αδιαπέρατη επιφάνεια και το ύψος των κτιρίων στα σημεία ξηράς της Αθήνας φαίνονται στα σχήματα 4.27, 4.28 και 4.29 αντίστοιχα.



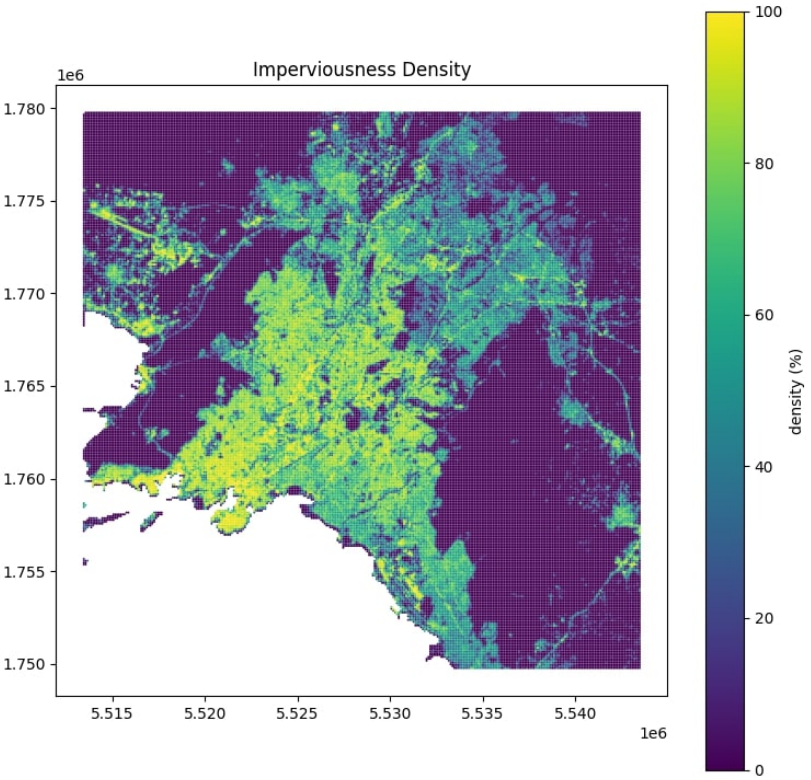
Σχήμα 4.25: *Χάρτης μέσης θερμοκρασίας του αέρα τον Ιανουάριο του 2012 στα σημεία ξηράς της Αθήνας*



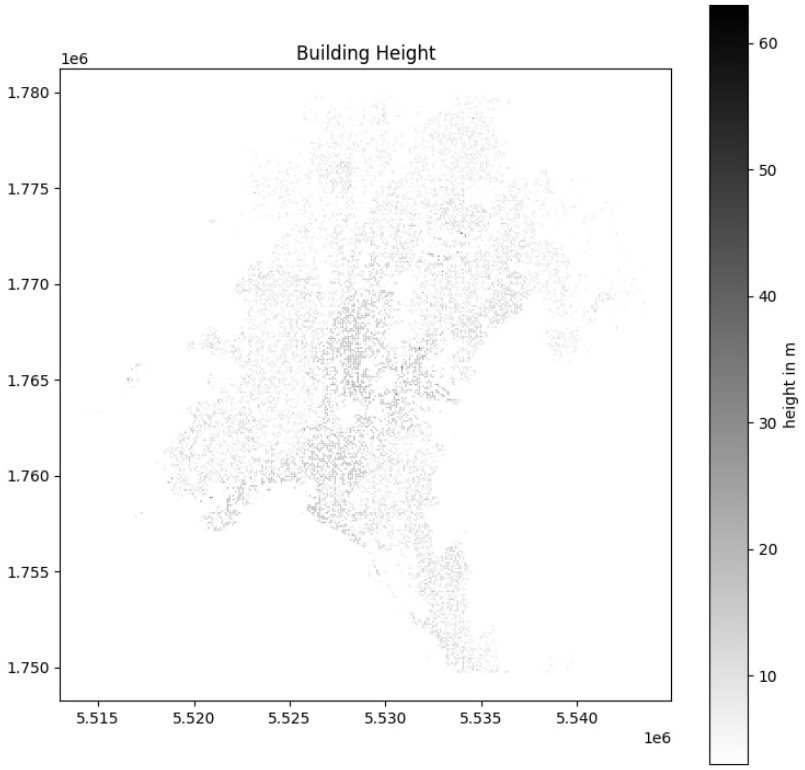
Σχήμα 4.26: *Χάρτης μέσης σχετικής υγρασίας του Ιανουάριο του 2012 στα σημεία ξηράς της Αθήνας*



Σχήμα 4.27: *Χάρτης κάλυψης από δέντρα το 2012 στα σημεία ξηράς της Αθήνας*

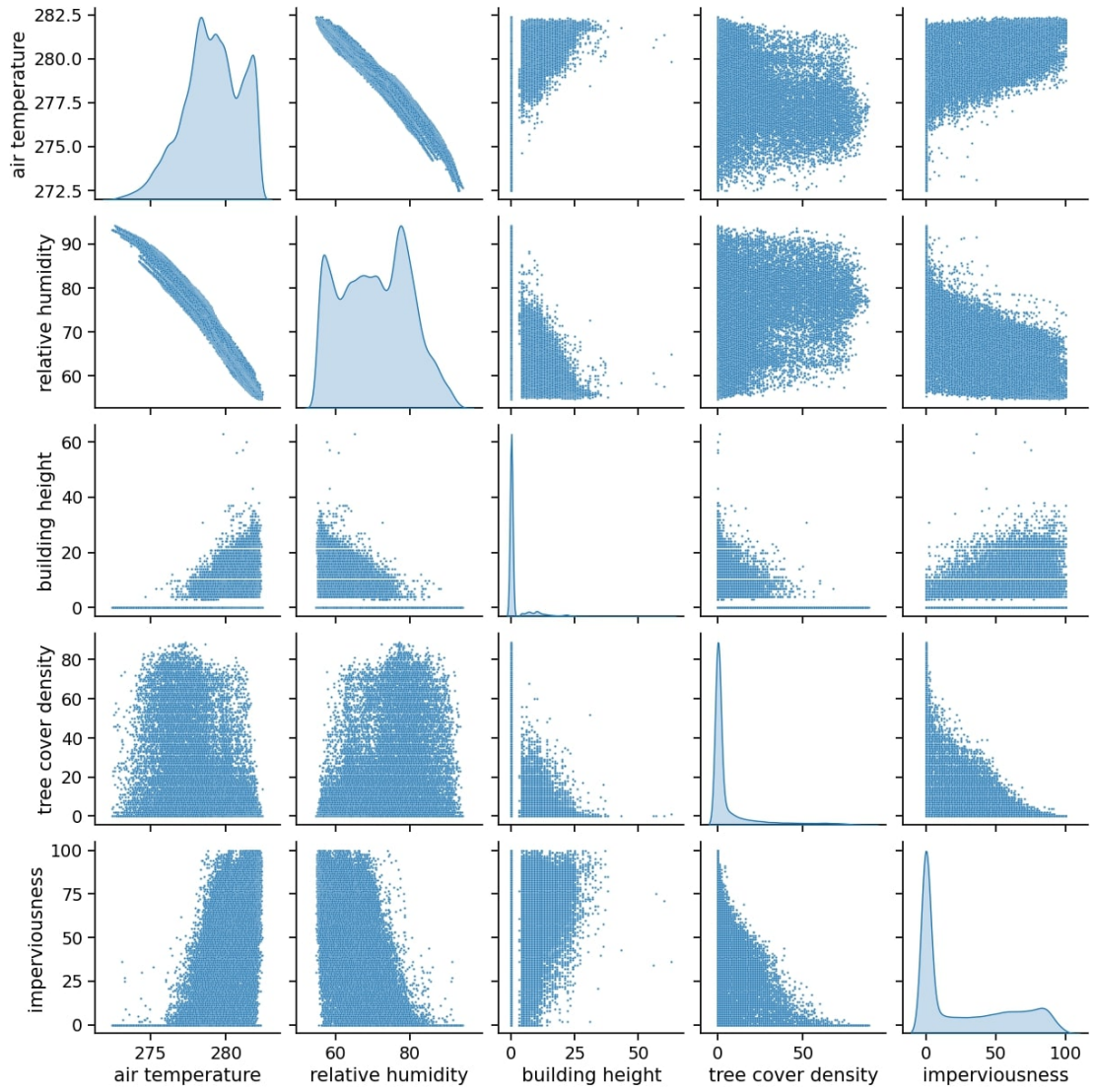


Σχήμα 4.28: *Χάρτης κάλυψης από αδιαπέρατη επιφάνεια το 2012 στα σημεία ξηράς της Αθήνας*



Σχήμα 4.29: *Χάρτης ύψους κτιρίων το 2012 στα σημεία ξηράς της Αθήνας*

Αρχικά θα εξετάσουμε τη συσχέτιση μεταξύ των 5 παραπάνω μεταβλητών (θερμοκρασία αέρα, σχετική υγρασία, κάλυψη από δέντρα, κάλυψη από αδιαπέρατη επιφάνεια, ύψος κτιρίων) ανά δύο. Για τον σκοπό αυτό θα σχεδιάσουμε το διάγραμμα διασποράς των σημείων για κάθε ζευγάρι μεταβλητών. Τα διαγράμματα αυτά φαίνονται στο σχήμα 4.30.

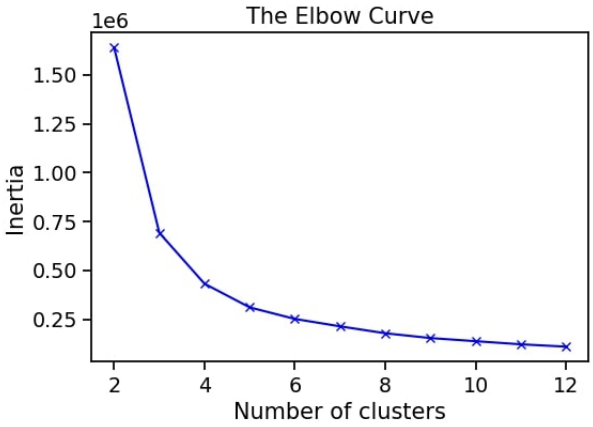


Σχήμα 4.30: *Διαγράμματα διασποράς των 5 μεταβλητών των σημείων ανά δύο*

Από τα διαγράμματα του σχήματος προκύπτουν κάποια συμπεράσματα για τη συσχέτιση κάθε ζευγαριού μεταβλητών. Συγκεκριμένα, παρατηρούμε πως η θερμοκρασία του αέρα και η σχετική υγρασία έχουν ισχυρή αρνητική σχέση, δηλαδή τα σημεία με υψηλή τιμή θερμοκρασίας έχουν χαμηλή τιμή υγρασίας και αντίθετα. Παράλληλα, το διάγραμμα διασποράς της θερμοκρασίας του αέρα με οποιαδήποτε από τις υπόλοιπες τρεις μεταβλητές είναι συμμετρικό με αυτό της σχετικής υγρασίας με την ίδια μεταβλητή. Όσον αφορά το ύψος των κτιρίων, παρατηρούμε από τα διαγράμματα ότι τα σημεία όπου παρατηρείται χαμηλό ύψος κτιρίων μπορούν να έχουν μεγάλο εύρος θερμοκρασίας του αέρα και σχετικής υγρασίας, ενώ τα σημεία όπου παρατηρείται υψηλό ύψος κτιρίων μπορούν να έχουν μόνο υψηλές τιμές θερμοκρασίας και χαμηλές τιμές υγρασίας. Επίσης, πολύ χαμηλές τιμές θερμοκρασίας

και πολύ υψηλές τιμές υγρασίας παρατηρούνται μόνο σε σημεία με μηδενικό ύψος κτιρίων. Σχετικά με την κάλυψη των σημείων από δέντρα, παρατηρούμε πως δεν υπάρχει ισχυρή σχέση μεταξύ αυτής και της θερμοκρασίας του αέρα ή της σχετικής υγρασίας. Παρόλα αυτά υπάρχει μεγαλύτερη συγκέντρωση σημείων στις μεσαίες τιμές θερμοκρασίας (και αντίστοιχα υγρασίας) για υψηλές τιμές κάλυψης πρασίνου και μεγαλύτερη συγκέντρωση στις υψηλές τιμές θερμοκρασίας (αντίστοιχα στις χαμηλές τιμές υγρασίας) για χαμηλή κάλυψη πρασίνου. Όσον αφορά τη σχέση της κάλυψης πρασίνου με το ύψος κτιρίων, παρατηρούμε ότι τα σημεία με υψηλή κάλυψη πρασίνου έχουν μηδενικό ύψος κτιρίων. Για μη μηδενικό ύψος κτιρίων η κάλυψη πρασίνου μπορεί να πάρει ένα εύρος τιμών, από μηδενικές έως κάποια τιμή η οποία μειώνεται όσο αυξάνεται το ύψος. Τέλος, η κάλυψη από αδιαπέρατη επιφάνεια φαίνεται να έχει ασθενή θετική σχέση με την θερμοκρασία και ασθενή αρνητική σχέση με τη σχετική υγρασία, και όσο μεγαλύτερη τιμή έχει σε ένα σημείο, τόσο μικρότερη μπορεί να είναι η κάλυψη πρασίνου που παρατηρείται στην περιοχή που βρίσκεται το σημείο αυτό.

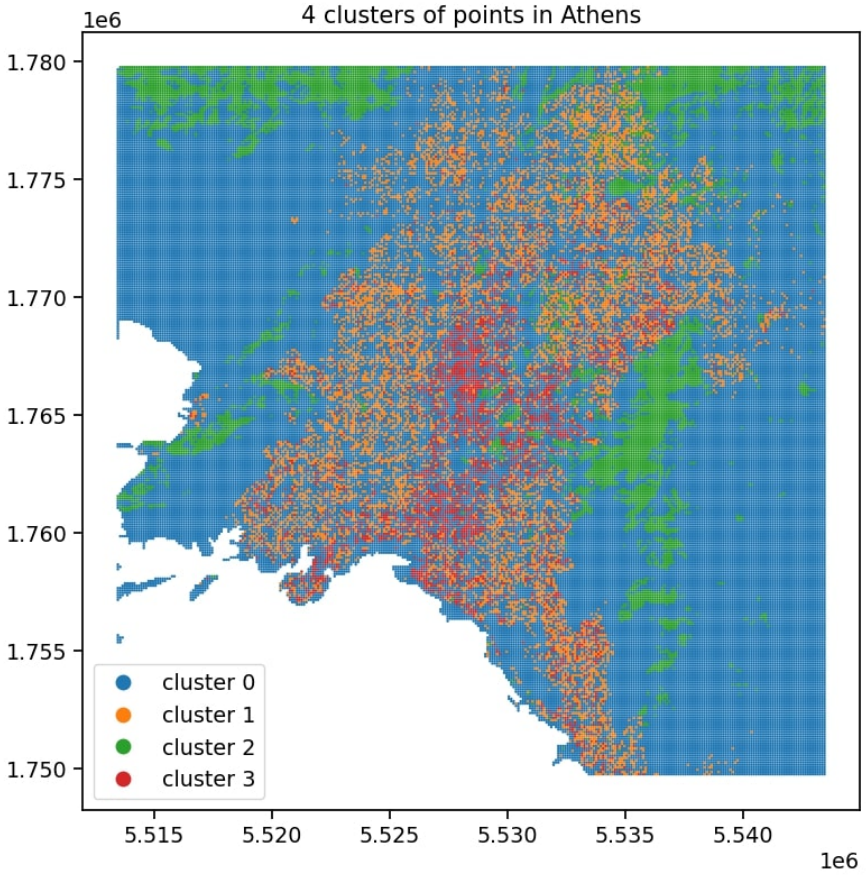
Για να εξετάσουμε τη σχέση μεταξύ όλων των μεταβλητών συνολικά, θα χωρίσουμε τα σημεία σε ομάδες σημείων με βάση αυτές τις μεταβλητές, ώστε τα σημεία κάθε ομάδας να έχουν παρόμοια χαρακτηριστικά, αλλά να διαφέρουν από τα σημεία των υπόλοιπων ομάδων. Θα χρησιμοποιήσουμε τον K-means αλγόριθμο συσταδοποίησης. Πρώτα θα εφαρμόσουμε τη μέθοδο του αγκώνα για να βρούμε τον βέλτιστο αριθμό συστάδων. Στο διάγραμμα του σχήματος 4.31 παρατηρούμε πως ο αριθμός αυτός είναι οι 4 συστάδες.



Σχήμα 4.31: *Καμπύλη μεθόδου elbow για την εύρεση βέλτιστου αριθμού συστάδων*

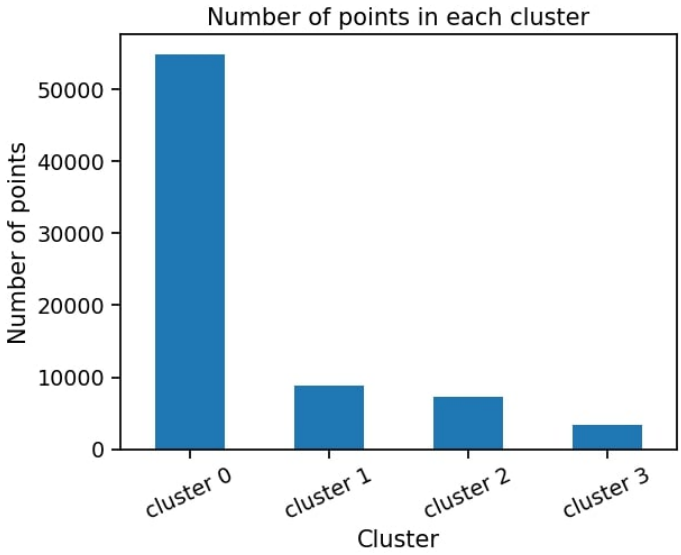
Εκτελούμε τον αλγόριθμο K-means στα δεδομένα ορίζοντας ως αριθμό συστάδων τον βέλτιστο που υπολογίστηκε με τη μέθοδο του αγκώνα, δηλαδή $K = 4$. Ύστερα εξετάζουμε εάν ο χωρισμός των σημείων στις 4 αυτές συστάδες είναι κατάλληλος σύμφωνα και με άλλες μετρικές πέρα από την WCSS. Συγκεκριμένα, υπολογίζουμε την τιμή σκιαγράφησης, δηλαδή τον μέσο συντελεστή σκιαγράφησης όλων των σημείων. Το αποτέλεσμα είναι 0.711, μία αρκετά υψηλή τιμή, που δείχνει πως κατά μέσο όρο τα σημεία έχουν χωριστεί στις συστάδες με κατάλληλο τρόπο. Προσθέτουμε, λοιπόν, τις ετικέτες των συστάδων που υπολογίστηκαν από τον αλγόριθμο K-means ως μία νέα στήλη στο GeoDataFrame των σημείων.

Για να προσδιορίσουμε πού βρίσκονται χωρικά τα σημεία κάθε μίας από τις 4 συστάδες, οπτικοποιούμε τα σημεία ξηράς της Αθήνας με διαφορετικό χρώμα, ανάλογα με τη συστάδα στην οποία ανήκουν, στον χάρτη του σχήματος 4.32.



Σχήμα 4.32: *Χάρτης των σημείων της Αθήνας χωρισμένα σε 4 συστάδες*

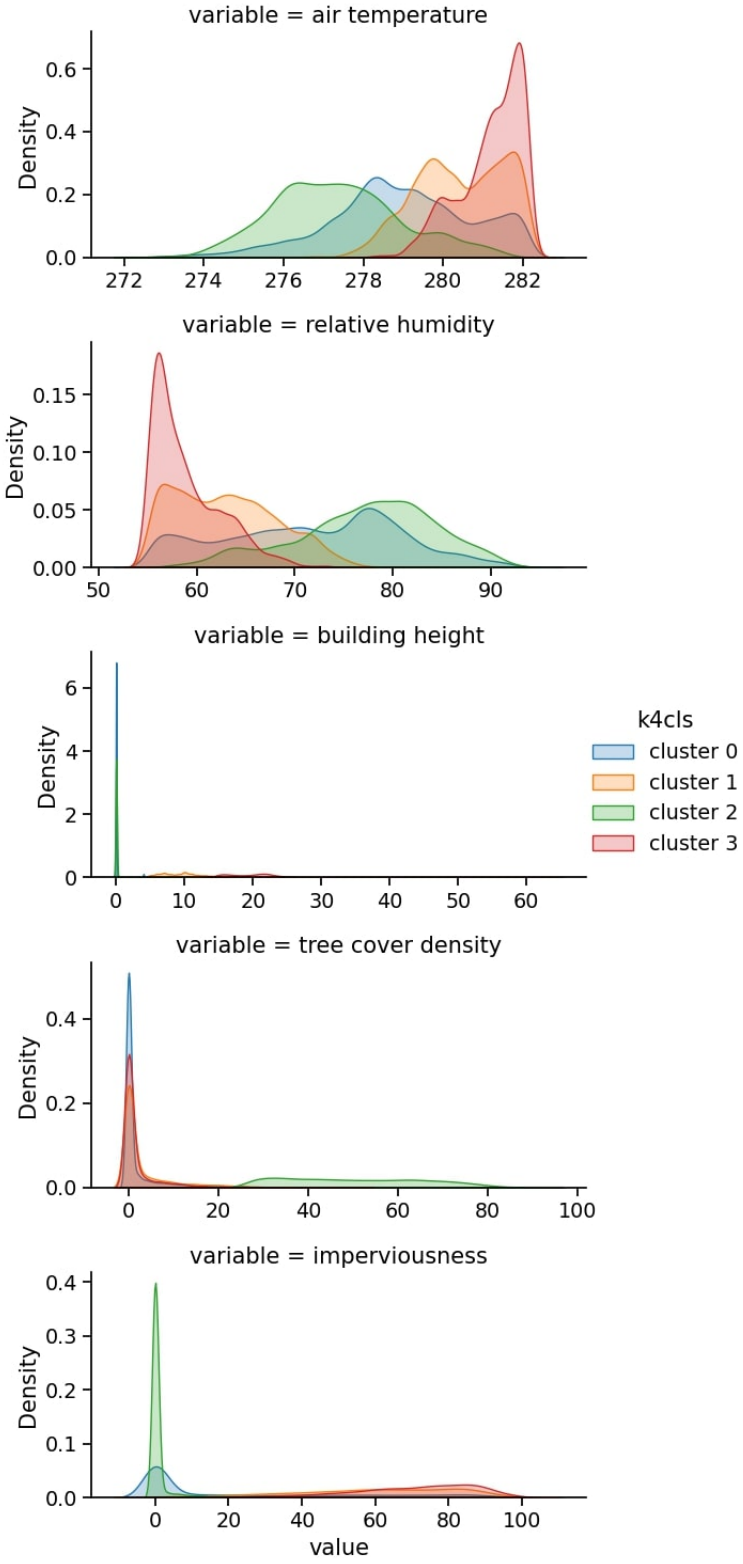
Έπειτα εξετάζουμε τα χαρακτηριστικά κάθε ομάδας. Πρώτα βρίσκουμε από πόσα σημεία αποτελείται κάθε ομάδα. Το αποτέλεσμα φαίνεται στο διάγραμμα του σχήματος 4.33.



Σχήμα 4.33: *Πλήθος σημείων σε κάθε συστάδα*

Παρατηρούμε πως πάνω από 50000 σημεία ανήκουν στη συστάδα 0, ενώ οι υπόλοιπες τρεις συστάδες περιέχουν λιγότερα από 10000 σημεία, με τη συστάδα 3 να έχει τα λιγότερα

σημεία από όλα. Για να περιγράψουμε τα χαρακτηριστικά κάθε συστάδας, οπτικοποιούμε την κατανομή κάθε μεταβλητής στα σημεία κάθε συστάδας, στα διαγράμματα του σχήματος 4.34.



Σχήμα 4.34: *Κατανομή κάθε μεταβλητής ανά συστάδα*

Από τα διαγράμματα αυτά προκύπτουν τα εξής συμπεράσματα για κάθε συστάδα:

- Συστάδα 0

  Τα σημεία που ανήκουν στη συστάδα 0 χαρακτηρίζονται από έλλειψη κτιρίων στην περιοχή γύρω από αυτά κι από πολύ χαμηλή, στα περισσότερα σημεία μηδενική, κάλυψη από δέντρα. Η κάλυψη από αδιαπέρατη επιφάνεια είναι επίσης μηδενική στα περισσότερα σημεία, αλλά μπορεί να πάρει κι ένα μεγάλο εύρος θετικών τιμών. Η θερμοκρασία του αέρα στα σημεία αυτά παίρνει τιμές σε ένα μεγάλο εύρος, από πολύ χαμηλές έως πολύ υψηλές, και η κατανομή της έχει μορφή κανονικής κατανομής με μέση τιμή $278.93K$. Τα περισσότερα σημεία έχουν ενδιάμεσες τιμές θερμοκρασίας. Η σχετική υγρασία παίρνει επίσης από πολύ χαμηλές έως πολύ υψηλές τιμές και η μέση τιμή της είναι 71.69.

- Συστάδα 1

  Τα σημεία που ανήκουν στην ομάδα αυτή χαρακτηρίζονται από παρουσία χαμηλών κτιρίων (χαμηλές μη μηδενικές τιμές ύψους κτιρίων), κυρίως μηδενική ή πολύ χαμηλή κάλυψη από δέντρα, και από αρκετά μεγάλη κάλυψη από αδιαπέρατη επιφάνεια στις περιοχές γύρω από αυτά. Η θερμοκρασία του αέρα στα σημεία αυτά παίρνει από μεσαίες έως πολύ υψηλές τιμές με αρκετά υψηλή μέση τιμή ($280.41K$). Η σχετική υγρασία, αντίθετα, παίρνει από πολύ χαμηλές έως μεσαίες τιμές και η μέση τιμή της είναι αρκετά χαμηλή (63.06).

- Συστάδα 2

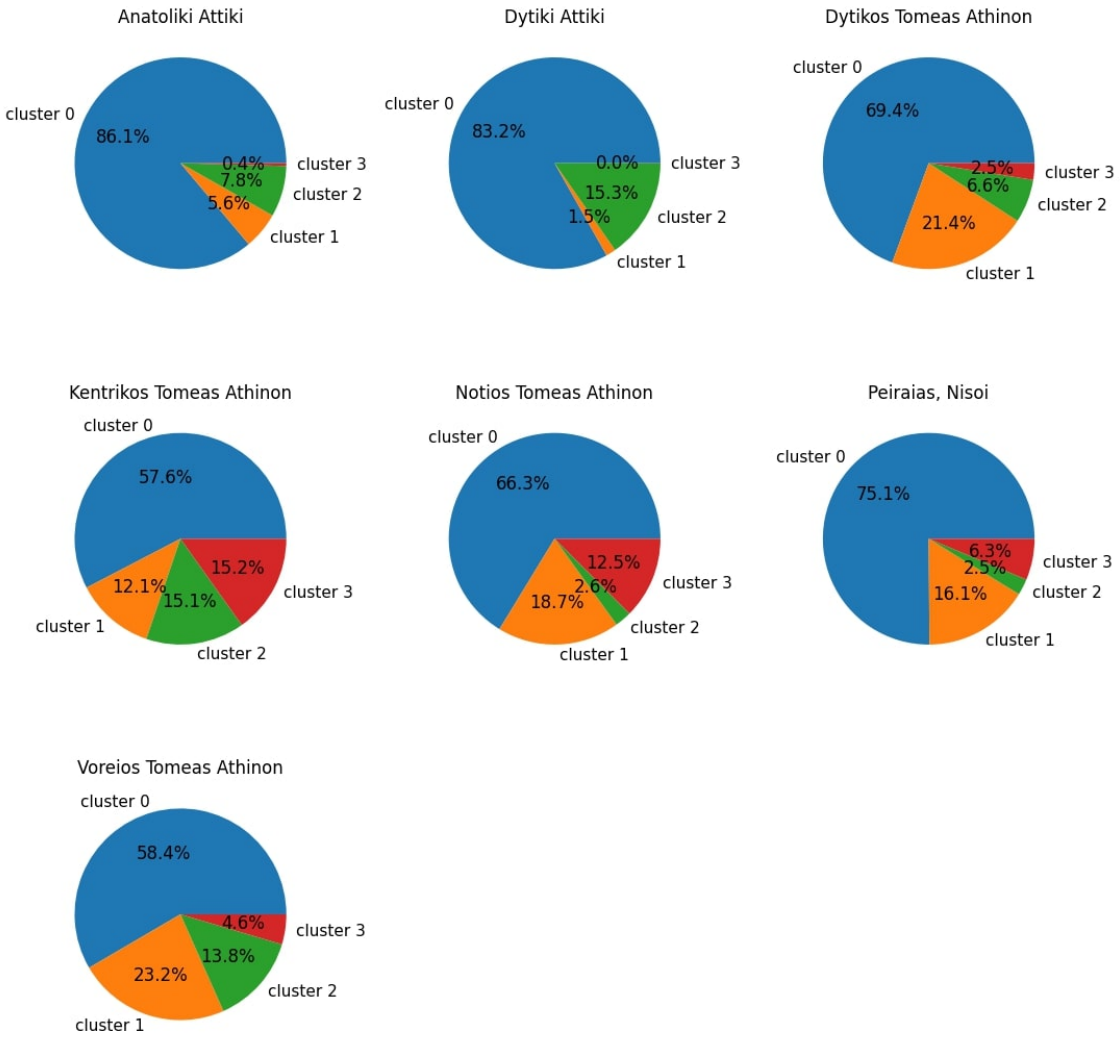
  Τα σημεία αυτής της συστάδας έχουν την μεγαλύτερη κάλυψη από δέντρα. Όλα βρίσκονται σε περιοχές με κάλυψη $26 - 89\%$, σε αντίθεση με τα σημεία των υπόλοιπων συστάδων, όπου η πλειοψηφία τους βρίσκονται σε περιοχή με μηδενική ή πολύ μικρή κάλυψη από δέντρα. Παράλληλα, τα σημεία αυτά έχουν κυρίως μηδενικό ύψος κτιρίων και μηδενική κάλυψη από αδιαπέρατη επιφάνεια. Η θερμοκρασία στα σημεία αυτά παίρνει ένα μεγάλο εύρος τιμών, αλλά η κατανομή της έχει την μορφή κανονικής κατανομής με μέση τιμή $277.32K$, την χαμηλότερη από όλες τις συστάδες, και πάνω από τα μισά σημεία της έχουν χαμηλότερη τιμή θερμοκρασίας από αυτή. Τέλος, τα σημεία της συστάδας 2 έχουν την υψηλότερη κατά μέσο όρο σχετική υγρασία.

- Συστάδα 3

  Η συστάδα 3 χαρακτηρίζεται από τις υψηλές τιμές ύψους κτιρίων σε όλα τα σημεία του, τη μηδενική ή χαμηλή κάλυψη από πράσινο και την υψηλότερη κάλυψη από αδιαπέρατη επιφάνεια από όλες τις υπόλοιπες συστάδες. Τα σημεία αυτά έχουν πολύ υψηλές τιμές μέσης θερμοκρασίας σε σχέση με τα σημεία των υπόλοιπων συστάδων, με μέση τιμή $281.19K$, και πολύ χαμηλές τιμές σχετικής υγρασίας, με μέση τιμή 58.98.

Στη συνέχεια, υπολογίζουμε το ποσοστό των σημείων που ανήκουν σε κάθε συστάδα για κάθε περιοχή NUTS επιπέδου 3. Τα αποτελέσματα φαίνονται στα διαγράμματα του σχήματος 4.35.

Παρατηρούμε πως σε όλες τις περιοχές NUTS επιπέδου 3, το μεγαλύτερο ποσοστό των σημείων τους ανήκουν στη συστάδα 0, στης οποία παρατηρείται έλλειψη κτιίων και πολύ χαμηλή κάλυψη από δέντρα. Οι περιοχές στις οποίες παρατηρείται η μεγαλύτερη παρουσία



Σχήμα 4.35: *Ποσοστό σημείων που ανήκουν σε κάθε συστάδα για κάθε περιοχή NUTS 3*

τέτοιων σημείων είναι η Ανατολική Αττική και η Δυτική Αττική, στις οποίες τα σημεία της συστάδας 0 καταλαμβάνουν πάνω από το 80% των σημείων τους. Αντίθετα, στον Κεντρικό και τον Βόρειο τομέα Αθηνών, τα σημεία αυτά καταλαμβάνουν το 57.7% και το 58.5% της κάθε περιοχής αντίστοιχα.

Τα σημεία που δεν ανήκουν στη συστάδα 0 μοιράζονται διαφορετικά μεταξύ των υπόλοιπων συστάδων, αναλόγως την περιοχή που εξετάζουμε. Στον Βόρειο Τομέα Αθηνών, στον Δυτικό, τον Νότιο, καθώς και στον Πειραιά, τα σημεία που επικρατούν, μετά από τα σημεία της συστάδας 0, είναι αυτά της συστάδας 1. Τα σημεία αυτά, όπως παρατηρήσαμε, χαρακτηρίζονται από μεσαίου ύψους κτίρια, χαμηλή κάλυψη από πράσινο, μεσαίες έως υψηλές τιμές θερμοκρασίας του αέρα και χαμηλές έως ενδιάμεσες τιμές σχετικής υγρασίας. Τα σημεία αυτά εμφανίζονται σε αρκετά μεγάλο ποσοστό και στον Κεντρικό Τομέα Αθηνών, και σε αρκετά μικρό ποσοστό στις περιοχές της Ανατολικής και Δυτικής Αττικής.

Στον Κεντρικό Τομέα Αθηνών, μετά τα σημεία της συστάδας 0, επικρατέστερα είναι αυτά της συστάδας 3, δηλαδή αυτά με υψηλές τιμές ύψους κτιρίων, έλλειψη πρασίνου, υψηλές τιμές θερμοκρασίας και χαμηλές τιμές σχετικής υγρασίας. Τα σημεία αυτά καταλαμβάνουν

επίσης ένα αρκετά μεγάλο ποσοστό του Νότιου Τομέα Αθηνών, και ένα μικρότερο ποσοστό του Πειραιά, του Βόρειου και του Δυτικού Τομέα Αθηνών. Στην Ανατολική και τη Δυτική Αττική υπάρχουν πολύ λίγα τέτοια σημεία. Στις δύο αυτές περιοχές, περισσότερο παρατηρούνται σημεία της συστάδας 2. Τα σημεία αυτά, όπως παρατηρήσαμε, έχουν υψηλή κάλυψη από δέντρα, χαμηλή κάλυψη από αδιαπέρατη επιφάνεια, και τις πιο χαμηλές, κατά μέσο όρο, τιμές θερμοκρασίας του αέρα. Τα σημεία αυτά καλύπτουν επίσης αρκετά μεγάλο ποσοστό του Κεντρικού Τομέα Αθηνών και του Βόρειου Τομέα Αθηνών και συναντώνται πολύ λιγότερο στον Δυτικό Τομέα, στον Νότιο Τομέα, και στον Πειραιά.

Στη συνέχεια επαναλαμβάνουμε όλα τα παραπάνω βήματα χρησιμοποιώντας τα δεδομένα των κλιματικών μεταβλητών (θερμοκρασία του αέρα και σχετική υγρασία) για έναν διαφορετικό μήνα, ώστε να εξετάσουμε κατά πόσο η χρονική διάσταση των δεδομένων επηρεάζει τα αποτελέσματα της ανάλυσης. Συγκεκριμένα, επιλέγουμε τα δεδομένα για τον Αύγουστο του ίδιου έτους. Ομοίως με πριν, υπολογίζουμε τη μέση θερμοκρασία του αέρα και τη μέση σχετική υγρασία κάθε σημείου τον μήνα αυτό. Τα δεδομένα για την κάλυψη από δέντρα, την κάλυψη από αδιαπέρατη επιφάνεια και το ύψος των κτιρίων στα σημεία ξηράς της Αθήνας είναι κοινά και στις δύο εφαρμογές της ανάλυσής μας, αφού αφορούν ολόκληρο το έτος του 2012.

Σε πρώτη φάση παρατηρούμε πως οι σχέσεις των μεταβλητών ανά δύο παρουσιάζουν παρόμοια χαρακτηριστικά με αυτά που παρατηρήσαμε κατά την ανάλυση για τις παρατηρήσεις του Ιανουαρίου. Μία διαφορά που παρατηρούμε είναι πως η αρνητική σχέση μεταξύ των δύο κλιματικών μεταβλητών, δηλαδή της θερμοκρασίας και της υγρασίας, δεν είναι τόσο ισχυρή όσο τον Ιανουάριο. Δηλαδή, δεν παρατηρείται τόσο έντονα το φαινόμενο πως όσο αυξάνεται η σχετική υγρασία που έχει ένα σημείο, τόσο μειώνεται η θερμοκρασία του αέρα σε αυτό και αντίστροφα. Κατά την εφαρμογή του αλγορίθμου K-means για ένα εύρος τιμών αριθμού συστάδων και την εφαρμογή της μεθόδου του αγκώνα, προκύπτει ο ίδιος βέλτιστος αριθμός από συστάδες, δηλαδή 4. Εφαρμόζουμε, λοιπόν, και πάλι τον αλγόριθμο K-means για 4 συστάδες. Προχωρώντας στον υπολογισμό της μετρικής σκιαγράφησης, βρίσκουμε πως έχει και πάλι την τιμή 0.711, η οποία είναι αρκετά υψηλή. Συνεπώς, χωρίζουμε τα σημεία στις 4 ομάδες που προέκυψαν από τον αλγόριθμο. Κατά την οπτικοποίηση των συστάδων στον χάρτη, παρατηρούμε πως ο χάρτης είναι παρόμοιος με αυτόν που προέκυψε χρησιμοποιώντας τα δεδομένα του Ιανουαρίου για τη συσταδοποίηση. Παρατηρούμε, δηλαδή, πως τα δεδομένα χωρίστηκαν στις 4 συστάδες με παρόμοιο τρόπο. Το γεγονός αυτό επιβεβαιώνεται κι από το ποσοστό σημείων που καταλαμβάνει κάθε συστάδα σε κάθε περιοχή NUTS επιπέδου 3, καθώς είναι παρόμοιο και για τους δύο μήνες που εξετάσαμε. Τέλος, οι κατανομές των μεταβλητών σε κάθε συστάδα για τον Αύγουστο έχουν τα ίδια χαρακτηριστικά με αυτά που παρατηρήσαμε στην ανάλυση για τον Ιανουάριο.

Με βάση τις παραπάνω παρατηρήσεις, συμπεραίνουμε πως οι μεταβλητές που εξετάσαμε συσχετίζονται μεταξύ τους με τον ίδιο τρόπο και τους δύο μήνες, και πως οι ομάδες σημείων με κοινά χαρακτηριστικά περιλαμβάνουν κατά κύριο λόγο τα ίδια σημεία και στις δύο περιπτώσεις.

4.5 Τεχνολογίες υλοποίησης

Ο γράφος γνώσης SustainGraph, στον οποίο πραγματοποιήθηκε η εισαγωγή των χωρικών δεδομένων, είναι υλοποιημένος στην πλατφόρμα γράφων δεδομένων Neo4j, σε μορφή μοντέλου LPG. Οι μηχανισμοί εισαγωγής και ανάλυσης των δεδομένων υλοποιήθηκαν μέσω Python χρησιμοποιώντας τη βιβλιοθήκη Py2neo, η οποία υποστηρίζεται από το Neo4j. Για την αναπαράσταση και επεξεργασία γεωμετριών και γεωχωρικών δεδομένων χρησιμοποιήθηκαν επίσης οι βιβλιοθήκες Shapely, Geopandas και Rasterio, και για την ανάλυσή τους και εφαρμογή αλγορίθμων μηχανικής μάθησης, και συγκεκριμένα συσταδοποίησης, σε αυτά, χρησιμοποιήθηκε η βιβλιοθήκη scikit-learn. Για τις χωρικές αναλύσεις χρησιμοποιήθηκαν, επιπρόσθετα, οι βιβλιοθήκες PySal και ESDA. Για τις οπτικοποιήσεις των γεωμετριών χρησιμοποιήθηκε η βιβλιοθήκη Matplotlib, ενώ για τις οπτικοποιήσεις τμημάτων του γράφου γνώσης αξιοποιήθηκαν τα εργαλεία οπτικοποιήσεων Neo4j Bloom και arrows.app, και το διαδραστικό κέλυφος εντολών Neo4j Browser.

Μέρος III

Επίλογος

Κεφάλαιο 5

Επίλογος

5.1 Συμπεράσματα

Στη διπλωματική εργασία παρουσιάστηκε ένα σύνολο μεθόδων και μηχανισμών για την αποδοτική ενσωμάτωση χωρικών δεδομένων σε έναν γράφο γνώσης που χρησιμοποιεί το μοντέλο LPG, τον γράφο SustainGraph, ο οποίος στοχεύει στη διαχείριση κοινωνικών και περιβαλλοντικών πληροφοριών σχετικά με τους SDGs.

Τα γεωχωρικά δεδομένα είναι ένα ισχυρό εργαλείο στην ανάλυση δεδομένων καθώς μπορούν να παρέχουν χρήσιμες πληροφορίες για τα χαρακτηριστικά μίας περιοχής. Παρόλα αυτά, η ετερογένεια των μορφών αρχείων στις οποίες παρέχονται, επιφέρει μία επιπλέον πρόκληση όσον αφορά τη διαχείρισή τους. Παράλληλα, τα γεωχωρικά σύνολα δεδομένων, λόγω του μεγάλου τους όγκου, είναι απαιτητικά όσον αφορά τους πόρους αποθήκευσης. Η συμπερίληψή τους στο SustainGraph διευκολύνει τους επιστήμονες να έχουν πρόσβαση σε αυτά μέσω απλών επερωτήσεων, να τα λαμβάνουν σε κοινή αναπαράσταση, και να εφαρμόζουν σε αυτά ποικίλες αναλύσεις. Οι δυνατότητες κλιμάκωσης των γράφων γνώσης τους καθιστούν κατάλληλους για αυτού του είδους τα δεδομένα.

Παρόλα αυτά, η πλατφόρμα του Neo4j, στην οποία αναπτύχθηκε το SustainGraph, υποστηρίζει αποκλειστικά τη γεωμετρία Point, δηλαδή σημεία, περιορίζοντας την αναπαράσταση άλλων γεωμετριών όπως πολύγωνα και συλλογές πολυγώνων στον γράφο γνώσης. Συνεπώς τα χωρικά δεδομένα πρέπει να εισαχθούν ως ομαδοποιημένες τιμές ανά περιοχή *GeoArea*, είτε από σημεία-εκπροσώπους χρησιμοποιώντας παράλληλα μηχανισμούς μείωσης του όγκου τους και συνεπώς της χωρικής ανάλυσης, ώστε να αποφευχθεί η εκθετική αύξηση του μεγέθους του SustainGraph. Από τις προτεινόμενες μεθόδους, καταλληλότερη κρίθηκε αυτή της ομαδοποίησης μέσω εφαρμογής πλέγματος, όπου οι περιοχές χωρίζονται σε τετράγωνα κελιά, τα οποία αναπαριστώνται από την μέση τιμή τους, και χωρικά από το κεντροειδές τους. Η επιλογή του μεγέθους κελιού εξαρτάται από τον επιθυμητό βαθμό μείωσης της χωρικής ανάλυσης.

Η ενδεικτική ανάλυση που εφαρμόστηκε για την περιοχή της Αθήνας για ένα σύνολο κοινωνικο-περιβαλλοντικών δεικτών, παρείχε σημαντικά αποτελέσματα σχετικά με την συσχέτιση των δεικτών και την κατηγοριοποίηση των περιοχών της Αθήνας βάσει των χωρικών και κλιματικών της χαρακτηριστικών. Το SustainGraph μπορεί πλέον να υποστηρίξει γεωχωρικά δεδομένα, και επιστήμονες διαφορετικών κλάδων μπορούν να εφαρμόσουν αναλύσεις επί των αυτών και να υποστηρίξουν τις διαδικασίες λήψης αποφάσεων.

5.2 Μελλοντικές Επεκτάσεις

Η επέκταση του σχήματος του SustainGraph και οι μηχανισμοί για την ενσωμάτωση των χωρικών δεδομένων σε αυτό, τα οποία εξερευνήθηκαν στα πλαίσια αυτής της διπλωματικής εργασίας, θα μπορούσαν να επεκταθούν περαιτέρω, τουλάχιστον ως προς τρεις κατευθύνσεις. Συγκεκριμένα, αναφέρονται τα ακόλουθα:

- Ενδεχόμενη χρήση της ειδικής βιβλιοθήκης Neo4j Spatial [25], η οποία αναπτύχθηκε από την κοινότητα των χρηστών του Neo4j, σε κάποια μελλοντική σταθερή της έκδοση. Αυτό θα έδινε τη δυνατότητα για αναπαράσταση πιο περίπλοκων γεωμετριών εκτός από σημείο, όπως γραμμή, πολύγωνο και συλλογές γεωμετριών. Χρησιμοποιώντας τέτοιου τύπου γεωμετρίες, θα γινόταν εφικτή η συμπερίληψη των χωρικών χαρακτηριστικών των *GeoAreas*, αλλά και σε μερικές περιπτώσεις η ακριβής αναπαράσταση των περιοχών που αφορούν οι παρατηρήσεις. Θα χρησιμοποιούνταν δηλαδή πολύγωνα για την περιγραφή της περιοχής στην οποία παρατηρούνται οι τιμές των δεικτών, αντί για τα σημεία που αναπαριστούν τα κεντροειδή τους. Θα παρέχονταν, επιπλέον, περισσότερες δυνατότητες εφαρμογής χωρικών πράξεων στα δεδομένα του γράφου.

- Εμπλουτισμός του SustainGraph με κοινωνικο-περιβαλλοντικά χωρικά δεδομένα από περισσότερες πηγές, όπως στατιστικές υπηρεσίες και περιβαλλοντικές οργανώσεις, διασφαλίζοντας την ομοιόμορφη αναπαράστασή τους, όπως περιγράφηκε στην παρούσα εργασία, και αξιοποιώντας τις νέες δυνατότητες του SustainGraph για συμπερίληψη χωρικών δεδομένων.

- Μελέτη και υλοποίηση μεθόδων για την ανάλυση πολλαπλών δεικτών του SustainGraph που αφορούν διαφορετικές περιοχές, Σε αυτές τις περιπτώσεις οι τοποθεσίες των παρατηρήσεων αναπαριστώνται από διαφορετικά σημεία. Για την εισαγωγή των παρατηρήσεων σε αλγορίθμους ανάλυσης χωρικών δεδομένων, όπως σε αλγορίθμους συσταδοποίησης και παλινδρόμησης, χρειάζεται πρώτα να εκτιμηθούν οι τιμές των παρατηρήσεων των διαφορετικών δεικτών σε κοινά σημεία. Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας χωρική παρεμβολή.

Βιβλιογραφία

[1] Nikolaos Karalis, Georgios Mandilaras και Manolis Koubarakis. *Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge*. *The Semantic Web - ISWC 2019*Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois και Fabien Gandon, επιμελητές, σελίδες 181–197, Cham, 2019. Springer International Publishing.

[2] Yu Liu, Jingtao Ding και Yong Li. *Developing Knowledge Graph Based System for Urban Computing*. 2022.

[3] Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby K. Fisher, Ling Cai, Gengchen Mai, Joseph Zalewski, Lu Zhou, Shirly Stephen, Seila Gonzalez, Bryce Mecum, Anna Lopez-Carr, Andrew Schroeder, David Smith, Dawn Wright, Sizhe Wang, Yuanyuan Tian, Zilong Liu, Meilin Shi, Anthony D'Onofrio, Zhining Gu και Kitty Currier. *Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence*. *AI Magazine*, 43(1):30–39, 2022.

[4] Diamantino Romeu Garanito Ferreira. *Using Neo4J geospatial data storage and integration*. Διδακτορική Διατριβή, Universidade da Madeira (Portugal), 2014.

[5] Wenwen Li, Sizhe Wang, Sheng Wu, Zhining Gu και Yuanyuan Tian. *Performance benchmark on semantic web repositories for spatially explicit knowledge graph applications*. *Computers, Environment and Urban Systems*, 98:101884, 2022.

[6] E. Fotopoulou et al. *SustainGraph: A knowledge graph for tracking the progress and the interlinking among the sustainable development goals' targets*. *Frontiers in Environmental Science*, 10, 2022.

[7] Lisa Ehrlinger και Wolfram Wöß. *Towards a definition of knowledge graphs*. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2, 2016.

[8] Vinay K. Chaudhri, Chaitanya Baru, Naren Chittar, Xin Luna Dong, Michael Genesereth, James Hendler, Aditya Kalyanpur, Douglas B. Lenat, Juan Sequeda, Denny Vrandečić και Kuansan Wang. *Knowledge graphs: Introduction, history, and perspectives*. *AI Magazine*, 43(1):17–29, 2022.

[9] Sumit Purohit, Nhuy Van και George Chin. *Semantic Property Graph for Scalable Knowledge Graph Analytics*. *2021 IEEE International Conference on Big Data (Big Data)*, σελίδες 2672–2677, 2021.

[10] *W3C - Resource Description Framework (RDF)*. https://www.w3.org/2001/sw/wiki/RDF.

[11] TIM BERNERS-LEE, JAMES HENDLER και ORA LASSILA. *THE SEMANTIC WEB*. *Scientific American*, 284(5):34–43, 2001.

[12] World Wide Web Consortium και others. *RDF 1.1 concepts and abstract syntax*. 2014.

[13] *W3C - RDF Schema (RDFS)*. https://www.w3.org/TR/rdf-schema.

[14] W3C Owl Working Group και others. *OWL 2 web ontology language document overview*. *http://www. w3. org/TR/owl2-overview/*, 2009.

[15] World Wide Web Consortium και others. *SPARQL 1.1 overview*. 2013.

[16] *OGC GeoSPARQL - A Geographic Query Language for RDF Data*. https://opengeospatial.github.io/ogc-geosparql/geosparql11/spec.html.

[17] Manolis Koubarakis, Manos Karpathiotakis, Kostis Kyzirakos, Charalampos Nikolaou και Michael Sioutis. *Data Models and Query Languages for Linked Geospatial Data*, σελίδες 290–328. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[18] Renzo Angles. *The Property Graph Database Model*. *AMW*, 2018.

[19] Jesús Barrasa. *RDF Triple Stores vs. Labeled Property Graphs: What's the Difference?* https://neo4j.com/blog/rdf-triple-store-vs-labeled-property-graph-difference/, 2017.

[20] *The Neo4j property graph database model*. https://neo4j.com/docs/getting-started/appendix/graphdb-concepts/.

[21] *Cypher query language*. https://neo4j.com/docs/getting-started/cypher-intro/.

[22] *Spatial values in Cypher and Neo4j databases*. https://neo4j.com/docs/cypher-manual/current/values-and-types/spatial/.

[23] *Spatial functions in in the Cypher query language*. https://neo4j.com/docs/cypher-manual/current/functions/spatial/.

[24] *Spatial functions in the APOC library*. https://neo4j.com/labs/apoc/4.4/overview/apoc.spatial/.

[25] *Neo4j Spatial library*. https://neo4j-contrib.github.io/spatial/.

[26] Kostis Kyzirakos, Dimitrianos Savva, Ioannis Vlachopoulos, Alexandros Vasileiou, Nikolaos Karalis, Manolis Koubarakis και Stefan Manegold. *GeoTriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings*. *Journal of Web Semantics*, 52-53:16–32, 2018.

[27] Kostas Patroumpas, Michalis Alexakis, Giorgos Giannopoulos και Spiros Athanasiou. *TripleGeo: an ETL Tool for Transforming Geospatial Data into RDF Triples*. *EDBT/ICDT Workshops*, 2014.

[28] Camille Bernard, Marlène Villanova-Oliver και Jérôme Gensel. *Theseus: A framework for managing knowledge graphs about geographical divisions and their evolution*. *Transactions in GIS*, 26(8):3202–3224, 2022.

[29] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich και Gerhard Weikum. *YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia*. *Artificial Intelligence*, 194:28–61, 2013.

[30] *I. Mandilara, E. Fotopoulou, C.M Androna, A. Zafeiropoulos, SustainGraph Gitlab Repository*. https://gitlab.com/netmode/sustaingraph.

[31] Kang Tsung Chang. *Introduction to geographic information systems*. Mcgraw-hill Boston, 3η έκδοση, 2006.

[32] *W3C - Model for Tabular Data and Metadata on the Web*. https://www.w3.org/TR/tabular-data-model/.

[33] *Interquartile range*. https://en.wikipedia.org/wiki/Interquartile_range.

[34] *OpenGIS Implementation Standard for Geographic information - Simple feature access - Part 1: Common architecture*. Standard OGC 06-103r4, Open Geospatial Consortium, 2011.

[35] Sergio Rey, Dani Arribas-Bel και Levi John Wolf. *Geographic data science with python*. CRC Press, 1η έκδοση, 2023.

[36] Luc Anselin και Sergio J Rey. *Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC, 2014.

[37] *Moran's I*. https://en.wikipedia.org/wiki/Moran%27s_I.

[38] *Cluster analysis*. https://en.wikipedia.org/wiki/Cluster_analysis.

[39] Leonardo Vilela Teixeira, Renato M Assunção και Rosangela Helena Loschi. *Bayesian Space-Time Partitioning by Sampling and Pruning Spanning Trees*. *J. Mach. Learn. Res.*, 20(85):1, 2019.

[40] Trevor Hastie, Robert Tibshirani, Jerome H Friedman και Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, τόμος 2. Springer, 2009.

[41] *Ward's method*. https://en.wikipedia.org/wiki/Ward%27s_method.

[42] *k-means clustering*. https://en.wikipedia.org/wiki/K-means_clustering.

[43] Juan Carlos Duque, Raúl Ramos και Jordi Suriñach. *Supervised Regionalization Methods: A Survey*. *International Regional Science Review*, 30(3):195–220, 2007.

[44] *Elbow method (clustering)*. https://en.wikipedia.org/wiki/Elbow_method_(clustering).

[45] *Knee of a curve*. https://en.wikipedia.org/wiki/Knee_of_a_curve.

[46] Ville Satopaa, Jeannie Albrecht, David Irwin και Barath Raghavan. *Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior*. *2011 31st International Conference on Distributed Computing Systems Workshops*, σελίδες 166–171, 2011.

[47] *Rand index*. https://en.wikipedia.org/wiki/Rand_index.

[48] *Silhouette coefficient*. https://en.wikipedia.org/wiki/Silhouette_(clustering).

[49] *Climate variables for cities in Europe from 2008 to 2017, Copernicus Climate Data Store*. https://cds.climate.copernicus.eu/cdsapp#!/dataset/10.24381/cds.c6459d3a.

[50] *NUTS (Nomenclature of territorial units for statistics), GISCO, Eurostat*. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
κ.α.	και άλλα
KG	Knowledge Graph
RDF	Resource Description Framework
LPG	Labeled Property Graph
IRI	Internationalized Resource Identifier
RDFS	Resource Description Framework Schema
WKT	Well-known Text
GML	Geography Markup Language
ISO	International Organization for Standardization
OGC	Open Geospatial Consortium
WGS	World Geodetic System
RML	RDF Mapping Language
TSN	Territorial Statistical Nomenclature
SDG	Sustainable Development Goals
NDC	Nationally Determined Contribution
EGD	European Green Deal
CSR	Country Specific Recommendation
NUTS	Nomenclature of Territorial Units for Statistics
CRS	Coordinate Reference System
NetCDF	Network Common Data Form
IQR	Interquartile Range
AHC	Agglomerative Hierarchical Clustering
WCSS	Within-Cluster Sum of Square
RI	Rand Index
ARI	Adjusted Rand Index
CDS	Climate Data Store
URL	Uniform Resource Locator
API	Application Programming Interface
TCD	Tree Cover Density

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
ακμή	edge
αλγόριθμος παλινδρόμησης	regression algorithm
αλγόριθμος συσταδοποίησης	clustering algorithm
απόσταση	distance
αριστερή εξωτερική σύζευξη	left outer join
ακραίες τιμές	outliers
βάρη κοντινότερων γειτόνων	nearest neighbor weights
βάρη πυρήνα	kernel weights
βάση δεδομένων γράφων	graph database
γειτνίαση	contiguity
γραμμή (γεωμετρία)	line
γραμμή (του πίνακα)	row
δείκτης	indicator
διανυσματικά δεδομένα	vector data
διαχωρισμός	separation
εγγύτητα	proximity
ένωση	union
επερώτηση	query
ετικέτα	label
εύρος ζώνης	bandwidth
ιδιότητα	property
καθολική χωρική αυτοσυσχέτιση	global spatial autocorrelation
κάλυψη από αδιαπέρατη επιφάνεια	impeviousness
κατηγόρημα	predicate
κελί	cell
κεντροειδές	centroid
κόμβος	node
κριτήριο του πύργου	rook criterion
κριτήριο της βασίλισσας	queen criterion
μέθοδος του αγκώνα	elbow method
ομαδοποίηση	aggregation
παρατήρηση	observation
περιορισμός	constraint
περιοχή	region

πολύγωνο	polygon
πλέγμα	grid
σημείο	point
στήλη	column
σύζευξη	join
συναθροιστική συνάρτηση	aggregate function
συνάρτηση πυρήνα	kernel function
σύνορο	border
συνδεσιμότητα	connectivity
σύνολο δεδομένων	dataset
συνοχή	cohesion
συντελεστής σκιαγράφησης	silhouette coefficient
συντεταγμένες	coordinates
συσσωρευτικός	agglomerative
συστάδα	cluster
συσταδοποίηση	clustering
τομή	intersection
τοπική χωρική αυτοσυσχέτιση	local spatial autocorrelation
χωρικά δεδομένα	spatial data
χωρικά βάρη	spatial weights
χωρική αυτοσυσχέτιση	spatial autocorrelation
χωρική παλινδρόμηση	spatial regression
χωρική παρεμβολή	spatial interpolation
ψηφιδωτή	raster