ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΜΣ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ

ΕΙΔΙΚΕΥΣΗ ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΔΕΙΞΗ HOT SPOT ΚΥΚΛΟΦΟΡΙΑΚΗΣ ΣΥΜΦΟΡΗΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΤΡΟΧΙΑΣ ΟΧΗΜΑΤΩΝ

ΕΚΠΟΝΗΣΗ: ΠΑΝΑΓΙΩΤΑ ΚΕΖΙΟΥ

ΕΠΙΒΛΕΨΗ : ΧΡΉΣΤΟΣ ΔΟΥΛΚΕΡΙΔΗΣ

ΙΟΥΝΙΟΣ 2021, ΑΘΗΝΑ

UNIVERSITY OF PIRAEUS

DEPARTMENT OF DIGITAL SYSTEMS

POSTGRADUATE PROGRAMME INFORMATION SYSTEMS & SERVICES

SPECIALIZATION BIG DATA & ANALYTICS

THESIS

TRAFFIC CONGESTION HOT SPOT ANALYSIS OVER VEHICLE TRAJECTORY DATA

BY: PANAGIOTA KEZIOU

SUPERVISOR: CHRISTOS DOULKERIDIS

JUNE 2021, ATHENS

ΕΥΧΑΡΙΣΤΙΕΣ

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω όλους αυτούς τους ανθρώπους που χωρίς αυτούς η διπλωματική αυτή εργασία δεν θα είχε ολοκληρωθεί.

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Χρήστο Δουλκερίδη για όλη τη βοήθεια, στήριξη και καθοδήγηση που μου παρείχε, καθώς και όλη την εμπιστοσύνη που μου έδειξε κατά την περάτωση της διπλωματικής μου εργασίας. Οι στοχευμένες παρατηρήσεις και συμβουλές του συνέβαλαν στην ολοκλήρωση στη διπλωματικής αυτής εργασίας.

Δεν θα μπορούσα να παραλείψω να ευχαριστήσω την οικογένεια μου για την υποστήριξη τους και την υπομονή και επιμονή που με εφοδίαζαν σε κάθε δυσκολία.

Τέλος, θα ήθελα να ευχαριστήσω τους αγαπημένους μου φίλους για όλη τους τη στήριξη και την καλή τους διάθεση. Η παρουσία τους ήταν καθοριστική δίνοντας μου κίνητρο για ένα καλύτερο αποτέλεσμα.

ΠΕΡΙΛΗΨΗ

Η ανάλυση hotspot αποτελεί ένα πρόβλημα αναγνώρισης στατιστικά σημαντικών συστάδων. Στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η ανάδειξη οδικών τμήματων με στατιστικά σημαντικές τιμές κυκλοφοριακής συμφόρησης για διαφορετικά χρονικά παράθυρα ανάλυσης, κάνοντας χρήση μεγάλου όγκου δεδομένων τροχιάς οχημάτων.

Για την αναγνώριση hotspots, χρησιμοποιήθηκε ο τροποποιημένος στατιστικός δείκτης Getis Ord Gi*, καθώς μπορεί να εφαρμοστεί σε τρισδιάστατα δεδομένα ακμών γράφου. Η σχέση γειτνίασης που επιλέχθηκε για τον υπολογισμό του Gi* z-score είναι το Gaussian Kernel.

Για την αναγνώριση hotspots κυκλοφοριακής συμφόρησης, προτείνονται δυο αλγόριθμοι παράλληλοι με δυνατότητα κλιμάκωσης. Στον πρώτο αλγόριθμο, ο δείκτης χωρικής αλληλεξάρτησης υπολογίζεται για κάθε χωρική ενότητα λαμβάνοντας υπόψη το σύνολο της πληροφορίας των μεταγενέστερων ακμών του 3D γράφου που αναπαριστά το οδικό δίκτυο. Αντίθετα, στο δεύτερο αλγόριθμο ο δείκτης χωρικής αλληλεξάρτησης υπολογίζεται για κάθε χωρική ενότητα λαμβάνοντας υπόψη την πληροφορία μόνο των γειτονικών ακμών εντός μιας οριζόμενης από το χρήστη απόστασης. Στόχος είναι η εύρεση της τομής μεταξύ του βέλτιστου χρόνου εκτέλεσης του αλγορίθμου και του αποτελέσματος από άποψη ποιότητας. Οι αλγόριθμοι υλοποιήθηκαν σε Apache Spark και το deploy του κώδικα πραγματοποιήθηκε μέσω του Google Cloud Platform. Κατά την πειραματική αξιολόγηση του αλγορίθμου, εξάγονται συμπεράσματα για την απόδοση του αλγορίθμου για διαφορετικές παραμέτρους ανάλυσης και το αποτέλεσμα της ανάλυσης για ιστορικά δεδομένα τροχιάς οχημάτων οπτικοποιείται.

ABSTRACT

Hot spot analysis is the problem of identifying statistically significant spatial clusters. The aim of this thesis is to identify spatio-temporal road segments with statistically significant amount of traffic congestion for massive trajectory data of moving objects.

Hotspots are identified using the modified statistical index Getis Ord Gi*, which is appropriately tailored in order to be implemented on a spatio-temporal graph. The spatial-temporal weight function applied to index Getis Ord Gi* is the Gaussian Kernel.

Two parallel and scalable algorithms are proposed for the identification of traffic congestion hotspots. In the first algorithm, the spatial autocorrelation index is calculated per graph edge, taking into consideration all the successor edges of the spatial unit under examination. In the second algorithm, the index is calculated per graph edge taking into consideration the successor edges within a user-defined distance. The goal is to determine the point where the algorithm's optimal execution time intersects with the quality of the outcome. The algorithms are developed in Apache Spark and deployed using Google Cloud Platform. The performance of the algorithms is experimentally evaluated for different analysis parameters and the quality of the outcome both analytically and via visualization.

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ 13
1.1 ΕΙΣΑΓΩΓΙΚΑ ΣΤΟΙΧΕΙΑ 13
1.2 ΑΝΤΙΚΕΙΜΕΝΟ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑ 14
1.3 ΔΟΜΗ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ 15
2 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ 17
2.1 ΕΙΣΑΓΩΓΗ 17
2.2 ΤΕΧΝΙΚΕΣ ΑΝΙΧΝΕΥΣΗΣ HOTSPOTS 19
2.2.1 Point-based Hotspot Analysis 19
2.2.1 Trajectory-based Hotspot Analysis 26
2.3 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΤΡΟΧΙΩΝ 30
2.4 ΣΥΜΠΕΡΑΣΜΑΤΑ 31
3 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ 35
3.1 APACHE SPARK 35
3.2 ΑΝΑΠΑΡΑΣΤΑΣΗ ΟΔΙΚΟΥ ΔΙΚΤΥΟΥ ΣΕ ΜΟΡΦΗ ΓΡΑΦΟΥ 37
3.3 ΧΩΡΙΚΗ ΑΝΑΛΥΣΗ 39
3.3.1 Χωρική Αυτοσυσχέτιση 40
3.3.2 Gi* Spatial Statistics 40
4 ΠΕΡΙΓΡΑΦΗ ΠΡΟΒΛΗΜΑΤΟΣ 43
4.1 ΔΕΔΟΜΕΝΑ 43
4.2 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ 44
4.2.1 Map-matching 44
4.2.2 Δημιουργία αναγνωριστικού τροχιάς 46
4.3 ΔΕΙΚΤΗΣ ΚΥΚΛΟΦΟΡΙΑΚΗΣ ΣΥΜΦΟΡΗΣΗΣ 47

4.4 ΧΩΡΟ-ΧΡΟΝΙΚΗ HOTSPOT ΑΝΑΛΥΣΗ 50

5 ΥΛΟΠΟΙΗΣΗ ΑΛΓΟΡΙΘΜΟΥ 57
5.1 ΣΥΝΟΨΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ 57

5.2 ΥΠΟΛΟΓΙΣΜΟΣ ΔΕΙΚΤΗ ΚΥΚΛΟΦΟΡΙΑΚΗΣ ΣΥΜΦΟΡΗΣΗΣ 58

5.2.1 Ροή εργασίας στο Spark 59

5.2.2 Συνάρτηση temporal_parameters 62

5.2.3 Συνάρτηση calc_attribute_value 62

5.3 ΥΠΟΛΟΓΙΣΜΟΣ BROADCAST ΜΕΤΑΒΛΗΤΩΝ 66

5.4 ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΔΕΙΚΤΗ Gi* 68

5.4.1 Ροή εργασιών για τον υπολογισμό του δείκτη Gi* 68

5.4.2 Συνάρτηση GetisOrdCalculations χωρίς εφαρμογή σημείου αποκοπής 71

5.4.3 Εφαρμογή χωρικού και χρονικού cut-off στη συνάρτηση *GetisOrdCalculations* 75

6 ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ 79
6.1 ΡΥΘΜΙΣΗ ΠΑΡΑΜΕΤΡΩΝ ΠΕΙΡΑΜΑΤΟΣ 79

6.2 ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΙΚΗΣ ΑΞΙΟΛΟΓΗΣΗΣ 83

6.2.1 Αναφορικά με το χρόνο εκτέλεσης 83

6.2.2 Οπτικοποίηση και ανάλυση των αποτελεσμάτων της Hotspot ανάλυσης 92

7 ΣΥΜΠΕΡΑΣΜΑΤΑ 99
7.1 ΣΥΝΟΨΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ 99

7.2 ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΠΕΡΑΙΤΕΡΟ ΕΡΕΥΝΑ 101

8 ΒΙΒΛΙΟΓΡΑΦΙΑ 103

ΠΑΡΑΡΤΗΜΑ ΕΙΚΟΝΩΝ

Εικόνα 1. Σύγκριση αποτελεσμάτων αλγορίθμου DBSCAN και GRHD σε συνθετικά δεδομένα [2] 18

Εικόνα 2. Σύγκριση αποτελεσμάτων αλγορίθμου DBSCAN και Sig-DBSCAN σε πραγματικά δεδομένα επειγόντων περιστατικών ρυμούλκησης οχημάτων σε συνθήκες χιονιού στην Minneapolis, US,2019 [5] 18

Εικόνα 3. Συνιστώσες μιας κατανεμημένης διενέργειας σε Spark..... 36

Εικόνα 4. Μη κατεθυνόμενος γράφος 38

Εικόνα 5. Κατευθυνόμενος γράφος..... 38

Εικόνα 6.Σταθμισμένος κατευθυνόμενος γράφος..... 38

Εικόνα 7. Τροχία οχημάτων όπως προκύπτει από δείκτη GPS σε σχέση με το οδικό δίκτυο..... 44

Εικόνα 8. Διαγραμματική απεικόνισης της αντιστοίχισης ενός σημείο στην εγγύτερη ακμή 46

Εικόνα 9. Απεικόνιση των παραμέτρων που απαιτούνται για εκτίμηση της χρονικής στιγμής εισόδου στην κορυφή..... 48

Εικόνα 10. Απεικόνιση γράφου για την άναδειξη της σχέσης γειτνίασης μεταξύ ακμών..... 52

Εικόνα 11. Απεικόνιση του τρόπου που βαρύνονται τα δεδομένα με εφαρμογή του Gaussian Kernel..... 53

Εικόνα 12. Αναπαράσταση 3D γράφου για την κατανόηση του υπολογισμού της σχέσης γειτνίασης..... 56

Εικόνα 13. Απεικόνιση των παραμέτρων που απαιτούνται για εκτίμηση της χρονικής στιγμής εισόδου στην κορυφή..... 63

ΠΑΡΑΡΤΗΜΑ ΠΙΝΑΚΩΝ

Πίνακας 1. Συνοπτικός πίνακας βιβλιογραφίας για Point-based Hotspot Analysis 23

Πίνακας 2. Συνοπτικός πίνακας βιβλιογραφίας για Point-based Hotspot Analysis ενσωματώνοντας τον περιορισμό του οδικού δικτύου 25

Πίνακας 3. Συνοπτικός πίνακας βιβλιογραφίας για Trajectory-based Hotspot Analysis 29

Πίνακας 4. Συνοτπικός πίνανας παρουσίασης μεταβλητών 43

Πίνακας 5. Παράμετροι πειράματος 83

ΠΑΡΑΡΤΗΜΑ ΔΙΑΓΡΑΜΜΑΤΩΝ

Διαγράμματα 1. Βασικά βήματα αλγορίθμου 58

Διαγράμματα 2. Μετατροπή των δεδομένων σε pair RDD 58

Διαγράμματα 3. Μετασχηματισμοί των δεδομένων κατά τον υπολογισμό του δείκτη συμφόρησης 61

Διαγράμματα 4. Δημιουργία των broadcast μεταβλητών τοπικά στον driver και διανομή τους στους worker nodes ώστε να γίνουν cached στη μνήμη 67

Διαγράμματα 5. Μετασχηματισμοί των δεδομένων κατά τον υπολογισμό του Gi* z-score 69

Διαγράμματα 6. Διαφορετικές υλοποιήσεις της συνάρτησης GetisOrdCalculations 70

Διαγράμματα 7. Απεικόνιση της σχέσης γειτνίασης της εξεταζόμενης ακμής και των ακμών του υπογράφου nn 73

Διαγράμματα 8. Παράδειγμα λογικής για την υλοποίηση της συνάρτησης GetisOrdCalculations 74

Διαγράμματα 9. Πλήθος ακμών που θα συμπεριληφθούν στην ανάλυση για n_neigh=2 77

Διαγράμματα 10. Μετρικές απόδοσης αλγορίθμου στο Stage#1 σε συνάρτηση με το μέγεθος του dataset 84

Διαγράμματα 11. Μετρικές απόδοσης αλγορίθμου σε συνάρτηση με το μέγεθος του dataset 85

Διαγράμματα 12. Μετρική att_edges για διαφορετικά χρονικά παράθυρα ανάλυσης..... 86

Διαγράμματα 13. Μετρικές απόδοσης αλγορίθμου ως προς το χρονικό παράθυρο της ανάλυσης 87

Διαγράμματα 14. Χρόνος εκτέλεσης για διαφορετικές τιμές της παραμέτρου n-neigh 89

Διαγράμματα 15. Πλήθος ακμών γράφου σε συνάρτηση της παραμέτρου temporal partition..... 90

Διαγράμματα 16. Μετρική Gi_edges για διαφορετικές τιμές της παραμέτρου n-neigh 90

Διαγράμματα 17. Απόδοση αλγορίθμου για διαφορετικά επίπεδα παραλληλίας (Dataset 5M)..... 91

Διαγράμματα 18. Απόδοση αλγορίθμου για διαφορετικά επίπεδα παραλληλίας (Dataset 7.4M)..... 92

ΠΑΡΑΡΤΗΜΑ ΧΑΡΤΩΝ

Χάρτης 1. Χάρτης περιοχής μελέτης με το πλήθος παρατηρήσεων ανά ακμή 80

Χάρτης 2. Αποτελέσματα hotspot ανάλυσης με συμμετοχή του συνόλου των γειτονικών παρατηρήσεων (All-neigh)..... 95

Χάρτης 3. Αποτελέσματα hotspot ανάλυσης με συμμετοχή 6 γειτονικών παρατηρήσεων (6-neigh)..... 96

Χάρτης 4. Αποτελέσματα hotspot ανάλυσης με συμμετοχή 12 γειτονικών παρατηρήσεων (12-neigh)..... 97

Χάρτης 5. Αποτελέσματα hotspot ανάλυσης με συμμετοχή 24 γειτονικών παρατηρήσεων (24-neigh)..... 98

1 ΕΙΣΑΓΩΓΗ

1.1 ΕΙΣΑΓΩΓΙΚΑ ΣΤΟΙΧΕΙΑ

Τα τελευταία χρόνια η ανάλυση δεδομένων κινούμενων αντικειμένων έχει προσελκύσει το ενδιαφέρον τόσο της ακαδημαϊκής κοινότητας όσο και των επιχειρήσεων. Ιδιαίτερα δημοφιλής έχει καταστεί η ανίχνευση κινούμενων αντικειμένων στο οδικό δίκτυο. Με την ανάπτυξη της τεχνολογίας GPS και την ενσωμάτωση της στις συσκευές κινητής τηλεφωνίας, καθώς και τη γρήγορη ανάπτυξη της τεχνολογίας καταγραφής δεδομένων, καθίσταται δυνατή η συλλογή μεγάλου όγκου χωροχρονικών δεδομένων.

Οι εφαρμογές που κάνουν χρήση αυτών των δεδομένων είναι ποικίλες. Στη συγκεκριμένη εργασία το ενδιαφέρον μας εστιάζει στο φαινόμενο της κυκλοφοριακής συμφόρησης στο οδικό δίκτυο και ειδικότερα στην ανίχνευση hot spots κυκλοφοριακής συμφόρησης από μεγάλου όγκου δεδομένα τροχιάς προερχόμενα από δείκτες GPS.

Το φαινόμενο της κυκλοφοριακής συμφόρησης αποτελεί συνεχιζόμενη πρόκληση στο αστικό περιβάλλον. Τα παρατηρούμενα έντονα φαινόμενα κυκλοφοριακής συμφόρησης συμβάλλουν σημαντικά στην αύξηση της εκπομπής των αερίων του θερμοκηπίου, στη μείωση της κυκλοφοριακής ικανότητας του οδικού δικτύου και στη υποβάθμιση της ποιότητας ζωής στον αστικό ιστό. Αποτελεί ένα σύνθετο φαινόμενο το οποίο διαδίδεται στα οδικά τμήματα και δύναται να προκαλέσει παρατεταμένα κύματα καθυστέρησης. Πολλές κυβερνήσεις και οργανισμοί διαθέτουν σημαντικά χρηματικά ποσά για την παρακολούθηση του φαινομένου και την ανάδειξη των αιτιών που το προκαλούν. Οι πληροφορίες που συλλέγονται είναι πολύ σημαντικές για τη λήψη αποφάσεων σε ατομικό επίπεδο, αλλά και στη φάση του σχεδιασμού και επαναπροσδιορισμού των μεταφορικών υποδομών.

Παραδοσιακοί μέθοδοι ανίχνευσης της κυκλοφοριακής συμφόρησης βασίζονται σε αισθητήρες εγκατεστημένους στο οδικό δίκτυο, όπως radars, και στην ανάλυση βίντεο από κάμερες εγκατεστημένες στο οδικό δίκτυο. Μειονέκτημα

των μεθόδων αυτών είναι ότι η παρακολούθηση της κυκλοφοριακής συμφόρησης γίνεται σε επιλεγμένα σημεία στρατηγικής σημασίας αλλά και το υψηλό κόστος εγκατάστασης και συντήρησης των συστημάτων. Αντίθετα με τη χρήση μετρήσεων GPS, μπορεί θεωρητικά να επιτευχθεί παρακολούθηση του φαινομένου για το σύνολο του οδικού δικτύου. Η χρήση αυτής της μεθόδου είναι οικονομική και δεν απαιτεί την εγκατάσταση ακριβού εξοπλισμού πάνω στο οδικό δίκτυο.

1.2 ΑΝΤΙΚΕΙΜΕΝΟ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑ

Στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η ανάδειξη hot spots κυκλοφοριακής συμφόρησης στο οδικό δίκτυο, κάνοντας χρήση μεγάλων δεδομένων τροχιάς προερχόμενων από δείκτες GPS. Ειδικότερα, μέσω της ανάλυσης Hotspots, στόχος είναι να βρεθούν οδικά τμήματα με στατιστικά σημαντικές τιμές κυκλοφοριακής συμφόρησης για διαφορετικά χρονικά παράθυρα ανάλυσης.

Για την ανάδειξη των hotspots χρησιμοποιήθηκε ο στατιστικός δείκτης Getis Ord Gi*, ο οποίος χρησιμοποιείται κυρίως για δισδιάστατα σημειακά δεδομένα. Για την εφαρμογή του στο οδικό δίκτυο και την ενσωμάτωση του χρονικού παράγοντα, απαραίτητη ήταν η προσαρμογή του. Ο τροποποιημένος αυτός δείκτης δέχεται ως δεδομένα ένα τρισδιάστατο γράφο οποίος αναπαριστά το οδικό δίκτυο και ενσωματώνει τις τιμές της κυκλοφοριακής συμφόρησης σε επίπεδο ακμής (χωρικό partition) και χρονικό παράθυρο (temporal partition).

Τόσο στο δείκτη Getis Ord Gi* όσο και στο σύνολο των δεικτών αναγνώρισης χωρικών και χωρο-χρονικών προτύπων, καθοριστικής σημασίας είναι ο ορισμός της σχέσης χωρικής αλληλεξάρτησης μεταξύ των εξεταζόμενων χωρικών οντοτήτων. Στη συγκεκριμένη εργασία, η σχέση γειτνίασης μεταξύ χωρικών οντοτήτων ορίζεται μέσω του χωρο-χρονικού Gaussian Kernel. Η σχέση γειτνίασης ορίζεται τόσο συναρτήσει της απόστασης των χωρικών οντοτήτων πάνω στο οδικό δίκτυο μετρούμενη σε hops, όσο και της απόστασης των οντοτήτων στη διάσταση του χρόνου, όπως προκύπτει μέσω των temporal partitions.

Στο πλαίσιο της διπλωματικής εργασίας αναπτύχθηκαν δυο διαφορετικοί αλγόριθμοι. Στον πρώτο αλγόριθμο, κατά τον υπολογισμό του δείκτη Gi* σε επίπεδο ακμής, λαμβάνεται υπόψη η τιμής της κυκλοφοριακής συμφόρησης για το σύνολο των ακμών που προηγούνται στης εξεταζόμενης ακμής. Στο δεύτερο αλγόριθμο που αναπτύχθηκε κατά τον υπολογισμό του δείκτη Gi* λαμβάνεται υπόψη μια γειτονιά ακμών γύρω από την εξεταζόμενη ακμή, εφαρμόζοντας ένα σημείο αποκοπής στους υπολογισμούς. Η γειτονιά αυτή περιλαμβάνει το σύνολο των ακμών που απέχουν απόσταση μικρότερη ή ίση με μια οριζόμενη από το χρήστη τιμή (n-neigh) στο 3D γράφο.

Κατά την πειραματική αξιολόγηση, εξάγονται συμπεράσματα για τον τρόπο που η επιλογή των παραμέτρων της hotspot ανάλυσης επηρεάζουν τον χρόνο εκτέλεσης του αλγορίθμου. Τέλος, μέσω της οπτικοποίησης και της ανάλυσης των αποτελεσμάτων της ανάλυσης hotspot, εξάγονται συμπεράσματα για την ποιότητα του αποτελέσματος στην περίπτωση εφαρμογής ή όχι σημείου αποκοπής στους υπολογισμούς.

1.3 ΔΟΜΗ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Παρακάτω παρουσιάζεται η δομή της διπλωματικής εργασίας ως προς τα κεφάλαια που περιλαμβάνει και το περιεχόμενο αυτών. Αναλυτικότερα:

- Στο πρώτο κεφάλαιο γίνεται καταγραφή του αντικειμένου της διπλωματικής εργασίας και ο στόχος αυτής.
- Στο δεύτερο κεφάλαιο εισάγεται η έννοια της Hotspot Ανάλυσης και πραγματοποιείται η ανασκόπηση της διεθνούς βιβλιογραφίας για έρευνες σχετικές με το αντικείμενο της διπλωματικής.
- Στο τρίτο κεφάλαιο πραγματοποιείται μια συνοπτική παρουσίαση των βασικών στοιχείων της τεχνολογίας Apache Spark και της χωρικής ανάλυσης, ενώ παρουσιάζεται η απαραίτητη θεωρία για το Getis-Ord Gi* Spatial Statistic.
- Στο τέταρτο κεφάλαιο γίνεται καταγραφή των δεδομένων που χρησιμοποιήθηκαν, της προεπεξεργασίας που υπέστησαν ώστε να καταλήξουν σε μια αξιοποιήσιμη μορφή, καθώς και θεωρητική

περιγραφή των βημάτων του αλγορίθμου. Επίσης, γίνεται αναφορά στο δείκτη κυκλοφοριακής συμφόρησης και στην προσαρμογή τους δείκτη Gi*, ώστε να επιτευχθεί 3D Hotspot Ανάλυση.

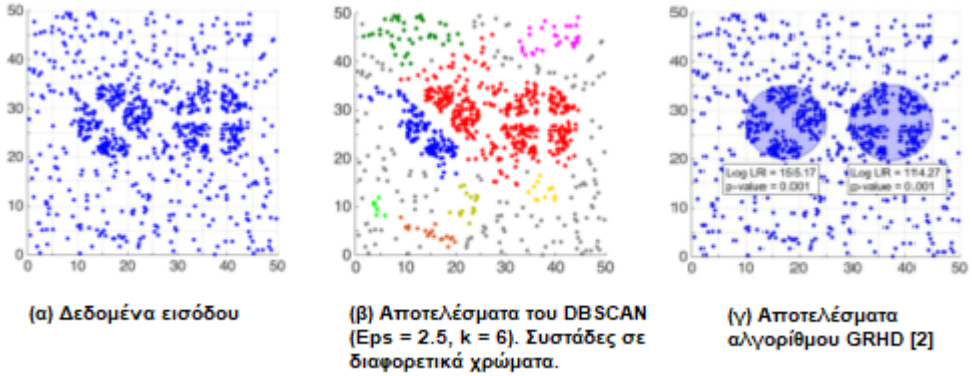
- Στο πέμπτο κεφάλαιο περιγράφονται αναλυτικά τα βήματα υλοποίηση του αλγορίθμου σε Apache Spark.
- Στο έκτο κεφάλαιο παρουσιάζονται τα αποτελέσματα της πειραματική αξιολόγησης του αλγορίθμου και παρουσιάζονται τα αποτελέσματα τόσο ως προς τον χρόνο εκτέλεσης του αλγορίθμου, αλλά και οι ακμές που αναγνωρίστηκαν ως hotspots κυκλοφοριακής συμφόρησης μέσω οπτικοποίησης.
- Στο έβδομο κεφάλαιο παρουσιάζονται τα συμπεράσματα και κάποιες προτάσεις για μελλοντική έρευνα.

2 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

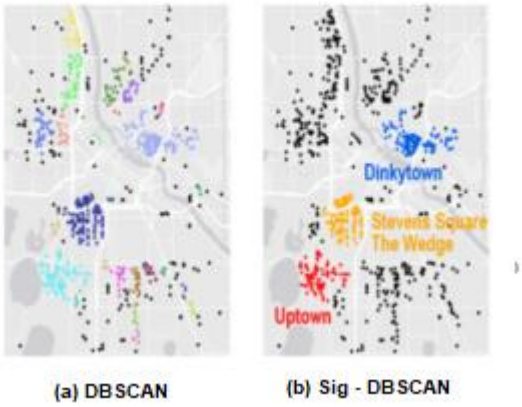
2.1 ΕΙΣΑΓΩΓΗ

Η hotspot ανάλυση αποτελεί μια ειδική περίπτωση ανάλυσης χωρικών και χωρο-χρονικών δεδομένων και εξόρυξης, η οποία έχει αποτελέσει αντικείμενο μελέτης πολλών ερευνών τα τελευταία χρόνια [1]. Ειδικότερα, ως hotspot ανάλυση ορίζεται η διαδικασία αναγνώρισης στατιστικά σημαντικών συστάδων. Δοσμένου ενός συνόλου Ν σημειακών χωρικών δεδομένων, στόχος είναι η ανίχνευση στατιστικά σημαντικών συστάδων, ποικίλων σχημάτων και πυκνοτήτων.

Στην υπάρχουσα βιβλιογραφία, κλασσικοί αλγόριθμοι συσταδοποίησης έχουν χρησιμοποιηθεί για την ανίχνευση περιοχών υψηλής συγκέντρωσης δραστηριοτήτων. Η πρακτική αυτή θεωρείται λανθασμένη στο πλαίσιο της ανάλυσης hotspot, εξαιτίας της παράλειψης ελέγχου των εξαγόμενων από τον αλγόριθμο συστάδων ως προς τη στατιστική τους σημαντικότητα, γεγονός που μπορεί να οδηγήσει στην ανίχνευση εσφαλμένων θετικών hotspots. Για παράδειγμα, στην Εικόνα 1 δίνεται ένα σύνολο δεδομένων από 800 σημεία. Εφαρμόζοντας σε αυτά τον αλγόριθμο συσταδοποίησης DBSCAN που να βασίζεται στην πυκνότητα προκύπτουν 8 διαφορετικές συστάδες, Εικόνα 1(α), αντίθετα κάνοντας χρήση του αλγορίθμου GHRD [2], που εξαλείφει τις εσφαλμένες θετικές συστάδες, προκύπτουν δυο κυκλικά hotspots 200 σημείων το καθένα. Ομοίως, στο παράδειγμα της Εικόνας 2 παρατηρούμε πως το πλήθος των στατιστικά σημαντικών συστάδων είναι μικρότερο από το πλήθος των συστάδων που εντοπίζονται.



Εικόνα 1. Σύγκριση αποτελεσμάτων αλγορίθμου DBSCAN και GRHD σε συνθετικά δεδομένα [2]



Εικόνα 2. Σύγκριση αποτελεσμάτων αλγορίθμου DBSCAN και Sig-DBSCAN σε πραγματικά δεδομένα επειγόντων περιστατικών ρυμούλκησης οχημάτων σε συνθήκες χιονιού στην Minneapolis, US,2019 [5]

Σε κάποιες κρίσιμες εφαρμογές, η αναγνώριση τυχαίων χωρικών προτύπων ως hotspots μπορεί να οδηγήσει σε κοινωνική και οικονομική κρίση. Η ανοχή σε εσφαλμένες θετικές συστάδες είναι ιδιαίτερα χαμηλή στην λήψη αποφάσεων που αφορούν κοινωνικά ζητήματα. Για παράδειγμα, ο ψευδής συναγερμός, αναφορικά με το ξέσπασμα μιας νόσου, μπορεί να οδηγήσει σε σπατάλη των δημόσιων πόρων και να προκαλέσει πανικό και περιττό άγχος στον πληθυσμό. Ομοίως, ο εσφαλμένος χαρακτηρισμός μιας περιοχής ως hotspot εγκληματικής ενέργειας μπορεί να μειώσει την επισκεψιμότητα, τις αξίες των ακίνητων και να πλήξει την τοπική οικονομία. Στην περίπτωση του σχεδιασμού στις μεταφορές η ανάδειξη αληθών θετικών hotspots για ατυχήματα μπορεί να προσανατολίσει σωστά τις επενδύσεις με στόχο τη βελτίωση των υποδομών. Τέλος, στην περίπτωση πυρκαγιάς σε αστική περιοχή, η ορθή ανάδειξη ακμών του οδικού δικτύου ως hotspots υψηλού κυκλοφορικού φόρτου μπορεί να συμβάλλει

σημαντικά στην αποδοτική καθοδήγηση και έγκυρη απομάκρυνση του πληθυσμού από την περιοχή κινδύνου.

2.2 ΤΕΧΝΙΚΕΣ ΑΝΙΧΝΕΥΣΗΣ HOTSPOTS

Η διεθνή βιβλιογραφία, αναφορικά με την ανάλυση hotspots, χωρίζεται σε δυο μεγάλες κατηγορίες. Στην πρώτη κατηγορία, υλοποιούνται τεχνικές βασιζόμενες σε σημειακά δεδομένα, ενώ στην δεύτερη κατηγορία τεχνικές βασιζόμενες σε δεδομένα τροχιών.

2.2.1 Point-based Hotspot Analysis

Στην πρώτη κατηγορία ανήκουν μελέτες στις οποίες η ανίχνευση στατιστικά σημαντικών hotspots στοχεύει στην αναγνώριση περιοχών και γραμμικών στοιχείων (υπο-γράφους ενός οδικού δικτύου), των οποίων η συγκέντρωση σημειακών δεδομένων είναι μεγαλύτερη από αυτή της γύρω περιοχής.

Στην πρώτη κατηγορία ανήκει η γεωμετρική μεθοδολογία που αναπτύσσουν στην μελέτη τους οι E. Eftelioglu et al. [2]. Διερευνούν το πρόβλημα της ανίχνευσης γεωγραφικά ισχυρών hotspots (Geographically Robust Hotspot Detection), που μοντελοποιούνται ως κύκλοι στον δισδιάστατο χώρο. Ο προτεινόμενος αλγόριθμος CGC (Cubic Grid Circle) αποτελεί βελτίωση του αλγορίθμου SaTScan, καθώς εντοπίζει hotspots ακόμα και αν παρεμβάλλονται από γεωγραφικά εμπόδια ή το κέντρο τους εντοπίζεται σε αραιές περιοχές, καθώς και εξαλείφει πολύ μικρά hotspots. Ο αλγόριθμος CGC αποτελείται από τρεις φάσης. Στην πρώτη φάση πραγματοποιείται κατάτμηση του χώρου σε κάνναβο και το σχήμα του κύκλου προσομοιώνεται από ένα σύνολο από κελιά ($cell_{circle}$). Για καθένα $cell_{circle}$ υπολογίζεται το log likelihood ratio, που αποτελεί μια μετρική ενδιαφέροντος της υπό εξέτασης περιοχής. Εφόσον η τιμή του δείκτη υπερβαίνει ένα κατώφλι, εφαρμόζεται μια διαδικασία βελτίωσης του αποτελέσματος για την εύρεση του ακριβούς κύκλου με το μέγιστο log likelihood ratio. Τέλος, καθένα υποψήφιο hotspot ελέγχεται ως προς τη στατιστική σημαντικότητα του με τη μέθοδο Monte Carlo.

Η βασισμένη στη γεωμετρία μεθοδολογία που προτείνεται μέσω του αλγορίθμου CGC μπορεί να εφαρμοστεί για την ανίχνευση περιοχών hotspots οποιουδήποτε σχήματος εφόσον αυτό μπορεί να οριστεί με χρήση παραμέτρων. Ειδικότερα, οι E. Eftelioglu et al. [3] και X. Tang et al. [4] στη μελέτη τους προσαρμόζουν αυτή τη μεθοδολογία για τη ανίχνευση hotspots σχήματος δακτυλίου και έλλειψης, αντίστοιχα.

Η εργασία των Y. Xie et al. [5] αντιμετωπίζει το πρόβλημα της εύρεσης στατιστικά σημαντικών συστάδων διαφορετικών σχημάτων και πυκνοτήτων με τη χρήση του αλγορίθμου DBSCAN. Ειδικότερα, κάνοντας χρήση της μεθόδου Monte Carlo, εντοπίζουν στατιστικά σημαντικές συστάδες για ένα συγκεκριμένο ζεύγος παραμέτρων (ϵ, minPts) και επιταχύνουν την επαναληπτική διαδικασία με χρήση του αλγορίθμου Dual-Convergence. Η ανίχνευση στατιστικά σημαντικών συστάδων διαφορετικών πυκνοτήτων πραγματοποιείται με επιλογή διαφορετικών παραμέτρων. Μέσω της πειραματικής τεκμηρίωσης της προτεινόμενης μεθοδολογίας προκύπτει ότι ο αλγόριθμος sig-DBSCAN βελτιώνει την ποιότητα των αποτελεσμάτων, απορρίπτοντας τυχαία πρότυπα και ότι ο Dual-Convergence μειώνει σημαντικά το υπολογιστικό κόστος.

Μια λύση στο προβλήματος της ανάλυσης hotspot σε μεγάλου όγκου χωρο-χρονικά δεδομένα με παράλληλο και κατανεμημένο τρόπο, προσφέρουν οι P. Nikitopoulos et al. [1] με τον αλγόριθμο BigCAB. Σύμφωνα με την προτεινόμενη προσέγγιση, η περιοχή μελέτης χωρίζεται σε κελιά και στόχος είναι η εύρεση των k κορυφαίων κελιών που έχουν χαρακτηριστεί ως hotspots, δηλαδή στατιστικά σημαντικές χωρικές συστάδες σύμφωνα με το στατιστικό μέγεθος Getis-Ord Gi*. Από την πειραματική ανάλυση σε πραγματικά σημειακά δεδομένα αποβίβασης από ταξί, προκύπτει ότι ο αλγόριθμος διαθέτει ιδιότητες που τον καθιστούν αποδοτικό σε κατανεμημένα συστήματα.

Οι Tang et al. [7], [8] αναπτύσσουν μεθοδολογίες βασιζόμενοι στο υποκείμενο δίκτυο. Αναλυτικότερα, οι Tang et al. [7] επικεντρώνεται στην ανακάλυψη στατιστικά σημαντικών γραμμικών hotspots με υψηλή συγκέντρωση δραστηριοτήτων, όπως ατυχήματα πεζών στο οδικό δίκτυο. Παρουσιάζουν ένα δυναμικό μοντέλο κατάτμησης του δικτύου, που επιτρέπει την αναγνώριση

στατιστικά σημαντικών hotspots σε τμήματα ακμών του υπάρχοντος δικτύου, που σε άλλη περίπτωση δεν θα είχαν αναγνωριστεί. Παράλληλα, η μεθοδολογία που προτείνουν (SRM_THB) ενσωματώνει ένα σύνολο αλγοριθμικών βελτιώσεων με στόχο την αύξηση της δυνατότητας κλιμάκωσης, όπως ο 'neighbor node filter ', ο 'shortest path pruning' και ο 'Monte Carlo speedup' αλγόριθμος. Σύμφωνα με την προτεινόμενη μεθοδολογία, εντοπίζονται τα συντομότερα μονοπάτια μεταξύ δραστηριοτήτων και υπολογίζεται ο δείκτης πυκνότητας για καθένα από αυτά. Εφόσον, ο δείκτης πυκνότητας υπερβαίνει ένα προκαθορισμένο κατώφλι, η τιμή ελέγχεται για τη στατιστική σημαντικότητας της με εφαρμογή της μεθόδου Monte Carlo. Ο αλγόριθμος επιστρέφει όλα τα στατιστικά σημαντικά μονοπάτια που δεν αποτελούν τμήματα άλλων hotspots.

Στη μελέτη τους [8] διαμορφώνουν το πρόβλημα της ανίχνευσης 'Network Isodistance Hotspots', που διαχέονται ισότροπα (προς όλες τις διευθύνσεις με συγκεκριμένη απόσταση) γύρω από κέντρα δραστηριότητας κατά μήκος του δικτύου, με σκοπό τον εντοπισμό περιοχών ενδιαφέροντος (υπο-γράφων) με υψηλή συγκέντρωση δραστηριοτήτων. Η απλοϊκή εκδοχή του αλγορίθμου περιλαμβάνει τον υπολογισμό του log likelihood ratio, που μετράει πόσο πιθανή είναι η απόρριψη της μηδενικής υπόθεσης σύμφωνα με την οποία η συγκέντρωση της δραστηριότητας μέσα στον ισότροπο υπο-γράφο είναι όμοια με αυτή έξω από αυτόν. Η στατιστική σημαντικότητα της τιμής του δείκτη ελέγχεται με τη χρήση της μεθόδου Monte Carlo. Με πειραματική ανάλυση, καταλήγουν ότι η βελτιωμένη εκδοχή του προτεινόμενου αλγόριθμου μπορεί να αντιμετωπίσει τις υπολογιστικές προκλήσεις της εφαρμογής, που βασίζεται στο partitioning του δικτύου και στο κλάδεμα της διαδικασίας με χρήση ενός κριτηρίου γρήγορου τερματισμού.

Τέλος, οι P. Kuo, X. Zeng και D. Lord [6] πραγματοποίησαν μια μελέτη για την εύρεση κατάλληλων αναλυτικών εργαλείων εντοπισμού περιοχών υψηλού κινδύνου για hotspot οδικών ατυχημάτων και εγκληματικών γεγονότων. Η μελέτη τους επικεντρώνεται γύρω από δυο βασικούς άξονες, την ενσωμάτωση περιορισμών του οδικού δικτύου στην διαδικασία αναγνώριση hotspots και την ενσωμάτωση του χωρο-χρονικού παράγοντα. Ο καθορισμός των hotspots

γίνεται εφαρμόζοντας δυο μεθόδους, τη Getis-Ord Gi* (Gi*) και την Kernel Density Estimation (KDE), με χρήση αποστάσεων πάνω στο επίπεδο και πάνω στο οδικό δίκτυο. Η ενσωμάτωση του χωρο-χρονικού παράγοντα στην ανάλυση γίνεται με χρήση οπτικών μεθόδων όπως spider graphs και co-maps για διαφορετικά χρονικά επίπεδα ανάλυσης. Τα αποτελέσματα της μελέτης δείχνουν ότι η μετρική Gi* είναι πολύ ευαίσθητη σε συσταδοποιημένα δεδομένα, ωστόσο ο συνδυασμός των δύο μεθόδων μπορεί να αποκαλύψει προβληματικές περιοχές. Η χρήση του οδικού δικτύου στον υπολογισμό των αποστάσεων αποδεικνύεται ωφέλιμη στην περίπτωσή που η περιοχή μελέτης παρουσιάζει έντονες κλίσεις. Τέλος από την χρονική ανάλυση προκύπτει ότι οι κατανομές των εγκληματικών ενεργειών και των ατυχημάτων είναι παρόμοιες σε ωριαίο και εβδομαδιαίο επίπεδο ανάλυσης.

Πίνακας 1. Συνοπτικός πίνακας βιβλιογραφίας για Point-based Hotspot Analysis

Μελέτη	**Αντικείμενο**	**Δεδομένα**	**Τύπος Hotspot**	**Αλγόριθμος**	**Μετρική ενδιαφέροντος (hotness)**	**Παράμετροι**	**Αρχιτεκτονική Συστήματος**	**Υπολογιστική πολυπλοκότητα**
[2]	Ανίχνευση γεωγραφικά ισχυρών hotspots εγκληματικών γεγονότων	Σημειακά χωρικά δεδομένα δύο διαστάσεων	Κυκλική περιοχή	Baseline: SaTScan Proposed: CGC (Cubic Grid Circle)	Μετρική πυκνότητας - Log likelihood ratio	- Ελάχιστη ακτίνα (r_{min}) - Κατώφλι log likelihood ratio (θ) - Κατώφλι p-value (a_p) - Πλήθος επαναλήψεων Monte Carlo (m)	Centralized (δυνατότητα κλιμάκωσης)	Μέγιστη: O (m x (N^3 + |P|log|P|)) Ελάχιστη: Ω (m x N^3) Όπου N^2 το πλήθος των κελίων, P των σημείων
[3]	Ανίχνευση hotspots σχήματος δακτυλιδιού για δεδομένα ασθενειών	Σημειακά χωρικά δεδομένα δύο διαστάσεων	Περιοχή σχήματος δακτυλιδιού	Baseline: DPG Proposed: DGPLMR (Dual Grid based Prune with Local maxima elimination, Multi cell Length and Refine)	Μετρική πυκνότητας - Log likelihood ratio	- Πλάτος δακτυλιδιού ($(r_o-r_i)_{min}$) - Ελάχιστη εσωτερική ακτίνα (r_{min}) - Κατώφλι log likelihood ratio (θ) - Κατώφλι p-value (a_p) - Πλήθος επαναλήψεων Monte Carlo (m)	Centralized (δυνατότητα κλιμάκωσης)	Μέγιστη: O (m x (N_c^2 x N_w^2 + $\|A\|^2$ log|A|)) Ελάχιστη: Ω (m x N_c^2 x N_w^2) Όπου N^2 το πλήθος των κελίων, P των σημείων
[4]	Ανίχνευση hotspots ελλειπτικού σχήματος εγκληματικών γεγονότων	Σημειακά χωρικά δεδομένα δύο διαστάσεων	Περιοχή σχήματος έλλειψης	Baseline: NaïveEHD Proposed: FastEHD (Fast Elliptical Hotspot Detection)	Μετρική πυκνότητας - Log likelihood ratio	- Ακρίβεια (δ) - Κατώφλι log likelihood ratio (θ) - Κατώφλι p-value (a_p) - Πλήθος επαναλήψεων MoDente Carlo (m)	Centralized (δυνατότητα κλιμάκωσης)	O (m x f x N^4 x n) Όπου $0 < \varphi < 1$

[5]	Ανίχνευση hotspots αυθαίρετων σχημάτων και πυκνοτήτων επειγόντων περιστατικών ρυμούλκησης οχημάτων σε συνθήκες χιονιού	Σημειακά χωρικά δεδομένα δύο διαστάσεων	Περιοχή αυθαίρετων σχημάτων και πυκνοτήτων	Baseline: DBSCAN Proposed: Sig-DBSCAN	-	- Παράμετροι DBSCAN (ε, minPts) - Κατώφλι p-value (a_p) - Πλήθος επαναλήψεων Monte Carlo (m)	Centralized	O (ρMN log N + (1 - ρ) M · max(|G|, N)), όπου ρ ∈ (0, 1] είναι το ποσοστό των δοκιμών που απαιτούνται για τον ακριβή DBSCAN
[1]	Ανίχνευση hotspots από χωρο-χρονικά δεδομένα αποβίβασης ταξί	Σημειακά χωρο-χρονικά δεδομένα (3D)	Περιοχή σχήματος κύβου (3D cells)	Baseline: Getis-Ord statistic Gi* Proposed: BigCAB	-	- Διαστάσεις κελιών και ως προς τις τρεις διαστάσεις (lon, lat, time window) - k-top στατιστικά σημαντικά κελιά	Distributed	-

Πίνακας 2. Συνοπτικός πίνακας βιβλιογραφίας για Point-based Hotspot Analysis ενσωματώνοντας τον περιορισμό του οδικού δικτύου

Μελέτη	**Αντικείμενο**	**Δεδομένα**	**Τύπος Hotspot**	**Baseline Αλγόριθμος**	**Μετρική ενδιαφέροντος (hotness)**	**Παράμετροι**	**Αρχιτεκτονική Συστήματος**	**Υπολογιστική πολυπλοκότητα**
[6]	Εντοπισμός περιοχών υψηλού κινδύνου για hotspots οδικών ατυχημάτων και εγκληματικών γεγονότων	- Σημειακά χωρο-χρονικά δεδομένα - Οδικό δίκτυο	- Gi*: αναδεικνύει σημειακά δεδομένα ως hotspots - KDE: επιφάνειες χωρίς επιβολή περιορισμών δικτύου και ακμές γράφων με επιβολή	Proposed: - Gi*, KDE χωρίς περιορισμούς - Gi* & KDE με περιορισμούς δικτύου	-	Gi*: - Βάρος (w_{ij}) KDE: - Bandwidth (τ)	Centralized	-
[7]	Ανακάλυψη στατιστικά σημαντικών γραμμικών hotspots (SLHD - Significant Linear Hotspot Detection)	- Σημειακά χωρικά δεδομένα - Οδικό δίκτυο	Δυναμικά κατατμημένοι υπογράφοι	Baseline: SRM_Naïve Proposed: SRM_TBD (Significant Route Miner Neighbor Node Filter, Shortest Path Tree Pruning, and Monte Carlo Simulation Speedup)	Μετρική πυκνότητας - Density-based ratio	- Κατώφλι density ratio (θ_λ) - Πλήθος επαναλήψεων Monte Carlo (m)	Centralized	$O\ (\|N_S\|^2 \log\|N_S\| + f_{MC} \times m \times (f_{DFS} \times \|N_D\|\ (\|N_D\| + \|N_S\|) + f_{SP} \times \|N_D\|^2))$
[8]	'Network isodistance hotspot detection' για εγκληματικά γεγονότα	- Σημειακά χωρικά δεδομένα - Οδικό δίκτυο	Υπογράφοι	Baseline: BaseNIHD (Network Isodistance Hotspot Detection) Proposed: NPP (Network Partitioning and Upper-Bound Pruning)	Μετρική πυκνότητας - Log likelihood ratio	- Κατώφλι log likelihood ratio (θ_p) - Κατώφλι p-value (a_p) - Πλήθος επαναλήψεων Monte Carlo (m)	Centralized (δυνατότητα κλιμάκωσης)	$O\ (m \times \|A\|^2 \log\ (\|A\|) + f_1 \times f_2 \times \|A\|^2\|N\|))$, where f_1 and f_2 αντιστοιχούν στην επιτάχυνση του 'upper-bound logLR pruning' and τη μείωση του αριθμού των εξεταζόμενων ακμών αντίστοιχα.

2.2.1 Trajectory-based Hotspot Analysis

Στην δεύτερη κατηγορία ανήκουν μελέτες που στοχεύουν στον εντοπισμό hotspots από κινούμενα αντικείμενα, ειδικότερα από τα δεδομένα τροχιάς που εκπέμπουν. Στη διεθνή βιβλιογραφία έχουν παρουσιαστεί ποικίλες μέθοδοι, που έχουν ως στόχο την αναγνώριση στατιστικά σημαντικών μοτίβων τροχιών.

Θεωρώντας πως η κίνηση των αντικειμένων περιορίζεται από το υποκείμενο οδικό δίκτυο ([9], [10]), οι X. Li et al. [9] αντιμετωπίζει το πρόβλημα της ανακάλυψης μοτίβων στην ροή της κυκλοφορίας, αναφερόμενων ως ‘hot routes’. Ειδικότερα, προτείνουν τον αλγόριθμο FlowScan που λαμβάνει υπόψη την πυκνότητα και δημιουργεί συστάδες από οδικά τμήματα, βασιζόμενος στο ποσοστό της κοινής τους κυκλοφορίας. Ο προτεινόμενος αλγόριθμος είναι επαναληπτικός και απαιτεί τον ορισμό δυο παραμέτρων της ελάχιστης κοινής κυκλοφορίας (MinTraffic) και το πλήθος των ακμών που διαμορφώνουν μια γειτονιά (Eps-neighborhoods). Σε πρώτο στάδιο, με χρήση των παραμέτρων εντοπίζεται η αρχή των hot routes και στη συνέχεια κάθε αρχή επεκτείνεται ώστε να σχηματίζει μια διαδρομή που μοιράζεται κοινό κυκλοφοριακό φόρτο. Για την αύξηση της αποδοτικότητας του αλγορίθμου γίνεται χρήση μεθόδων ευρετηρίασης, όπως R-tree και clustering indexes. Τέλος, μέσω πειραματικής ανάλυσης προκύπτει ότι ο αλγόριθμος είναι ισχυρός στην αναγνώριση περίπλοκων χωρικών προτύπων, που αναπαριστούν με ρεαλισμό την ροή της κυκλοφορίας.

Οι M. Häsner at al. [10] στη μελέτη τους παρουσίασαν την μεθοδολογία OPS (Online Prediction of Hot Spots) για την online πρόβλεψη hotspots που βασίζεται στη πρόσφατη θέση κινούμενων αντικειμένων στο οδικό δίκτυο. Σε πρώτη φάση, πραγματοποιείται πρόβλεψη της θέσης των κινούμενων οχημάτων για μια μελλοντική χρονική στιγμή. Μέσω μιας ευριστικής τεχνικής εντοπίζονται όλοι οι κόμβοι από τους οποίους πιθανώς θα διέλθει το κινούμενο αντικείμενο. Λαμβάνοντας υπόψη το σύνολο των προβλέψεων, πραγματοποιείται ανάθεση βαρών στους κόμβους του δικτύου, που αντιπροσωπεύουν την ένταση της κυκλοφορίας σε αυτούς. Μέσω μιας μεθόδου ανάδειξης ακραίων τιμών, χαρακτηρίζονται κάποιοι από τους κόμβους ως

hotspots. Τέλος εφαρμόζοντας τον αλγόριθμο DBSCAN στους κόμβους που αναδείχθηκαν ως hotspots, είναι δυνατή η εξαγωγή επιφανειών υψηλής συγκέντρωσης hotspots.

Οι D. Sacharidis et al. [11] προτείνουν μια μεθοδολογική προσέγγιση για online εξαγωγή και διατήρηση μοτίβων κίνησης (hot motion paths), που διανύθηκαν από πολλαπλά κινούμενα αντικείμενα. Το πρόβλημα της ανίχνευσης 'hot motion paths' έχει εξεταστεί και τη μελέτη [9], αλλά με τον περιορισμό της κίνησης των αντικειμένων πάνω στον γνωστό οδικό δίκτυο. Η μελέτη των [11] διαφοροποιείται τόσο στη μη επιβολή περιορισμού στην κίνηση των αντικειμένων στον ευκλείδειο χώρο, όσο και στο γεγονός ότι η ένταση της συχνότητας κίνησης στα μονοπάτια (hotness) υπολογίζεται σε συνάρτηση με το χρονικό διάστημα στο οποίο το αντικείμενο διένυσε το μονοπάτι. Τέλος, κρίνεται σκόπιμο να σημειωθεί ότι λαμβάνεται υπόψη η αβεβαιότητα εντοπισμού της θέσης των κινούμενων αντικειμένων. Οι [11] στη μελέτη τους θεωρούν ένα κατανεμημένο περιβάλλον, που με χρήση χωροχρονικής ευρετηρίασης, το 'hotness' και η γεωμετρία των 'hot motion paths' διατηρείται και ανανεώνεται για σημαντικές αλλαγές στον εντοπισμό τους. Εστιάζουν σε μοτίβα κίνησης στο κοντινό παρελθόν, απορρίπτοντας μονοπάτια που υπερβαίνουν ένα κυλιόμενο χρονικό παράθυρο.

Ομοίως, οι μελέτες των P. Nikitopoulos et al. [12] και Y. Qiao et al. [13] αφορούν την ανίχνευση hotspots, χωρίς να λαμβάνουν υπόψη τον περιορισμό κίνησης των αντικειμένων πάνω στον γνωστό οδικό δίκτυο.

Ειδικότερα, οι P. Nikitopoulos et al. [12] εξετάζουν το πρόβλημα της εύρεσης hotspot τροχιών σε μεγάλου όγκου χωρο-χρονικά δεδομένα, προτείνοντας δυο παράλληλες και κατανεμημένες προσεγγίσεις. Η ιδιαιτερότητα της προτεινόμενης μεθοδολογίας οφείλεται στην τροποποίηση του δείκτη Getis-Ord Gi*, ώστε να εντοπίζει hotspots από δεδομένα τροχιών. Σύμφωνα με την προτεινόμενη μεθοδολογία, η περιοχή μελέτης κατατμίζεται σε 3D κελιά και η τιμή του κάθε κελιού αντιστοιχεί στο χρόνο που διανύθηκε από κάθε κινούμενο αντικείμενο μέσα σε αυτό. Η επίδραση του κάθε κελιού στα γειτονικά ορίζεται ως αντιστρόφως εκθετική της μεταξύ τους απόστασης. Ο πρώτος αλγόριθμος

(THS), που προτείνεται, παρέχει μια ακριβής λύση στο πρόβλημα, που μπορεί να αποδειχθεί υπολογιστικά ακριβή συνδυαστικά με την επιλογή των παραμέτρων. Ο δεύτερος αλγόριθμος (aTHS), είναι προσεγγιστικός και αγνοεί τη συνεισφορά των απομακρυσμένων κελιών με ελεγχόμενο τρόπο, ανταλλάσσοντας ένα μέρος της υπολογιστικής ακρίβειας με τη μείωση του υπολογιστικού κόστους.

Η μελέτη των Y. Qiao et al. [13] προτείνει μια μεθοδολογία αποδοτικής ανάλυσης μεγάλου όγκου κυκλοφοριακών δεδομένων πυκνοκατοικημένων περιοχών προερχόμενα από δίκτυο κινητής τηλεφωνίας δεύτερης, τρίτης και τέταρτης γενιάς σε cloud περιβάλλον. Ειδικότερα για τη διερεύνηση της ανθρώπινης κινητικότητας, εφαρμόστηκε μια μη-παραμετρική μέθοδος ανακάλυψης ‘city hotspots’, τη λεγόμενη «καρδιά της πόλης». Ο χαρακτηρισμός μιας περιοχής ως ‘city hotspots’ πραγματοποιείται εφόσον η επισκεψιμότητα της υπερβαίνει ένα καθορισμένο όριο όπως αυτό προκύπτει από τη χρήση της καμπύλης Lorenz. Η μελέτη λαμβάνει υπόψη τη συνιστώσα του χρόνου υπολογίζοντας τα ‘city hotspots’ ωριαία. Στη συνέχεια, για την ανακάλυψη προτύπων κινητικότητας τόσο για ομάδες χρηστών όσο και μεμονωμένους χρήστες μεταξύ ‘city hotspots’, γίνεται χρήση μιας βελτιωμένης εκδοχής του αλγορίθμου A-priori. Μελετώντας την ομοιότητα των διαδρομών διαφορετικών χρηστών, αναδεικνύονται συνήθειες που αφορούν τις μετακινήσεις, καθώς και πληθυσμιακές ομάδες με κοινά ενδιαφέροντα. Τέλος, με χρήση ιστορικών δεδομένων κίνησης και των αλγορίθμων Intelligent Time Divisions (ITD) και Markov, καθίσταται δυνατή η πρόβλεψη των ανθρώπινων μετακινήσεων λαμβάνοντας υπόψη το χρονικό παράγοντα.

Πίνακας 3. Συνοπτικός πίνακας βιβλιογραφίας για Trajectory-based Hotspot Analysis

Μελέτη	**Εφαρμογή**	**Δεδομένα**	**Τύπος Hotspot**	**Αλγόριθμος Hotspots**	**Παράμετροι**	**Έλεγχος Στατιστική σημαντικότητα**	**Αρχιτεκτονική Συστήματος**	**Υπολογιστική πολυπλοκότητα**
[9]	Ανίχνευση hot routes σε οδικό δίκτυο βασιζόμενη στην πυκνότητα της κυκλοφορίας	- Οδικό δίκτυο - Τροχιές αντικειμένων	Διαδρομές: Σύνολο από ακμές γράφου που μοιράζονται κοινή κυκλοφορία	Baseline: DBSCAN Proposed: FlowScan	- Ελάχιστη κοινή κυκλοφορία (MinTraffic) - Γειτνίαση (Eps-neighborhoods)	Όχι	Centralized (δυνατότητα κλιμάκωσης)	-
[10]	Online πρόβλεψη hotspots κινούμενων αντικειμένων στο οδικό δίκτυο	- Οδικό δίκτυο - Τροχιές αντικειμένων	Σημειακά και επιφανειακά hotspots	Baseline: DBSCAN Proposed: OPS	- Χρονικός ορίζοντας της πρόβλεψης - Παράμετροι DBSCAN (ε, minPts)	Όχι	Centralized μ δυνατότητα υλοποίησης σε κατανεμημένο σύστημα	-
[11]	Online εξαγωγή και διατήρηση μοτίβων κίνησης (hot motion paths) σε δεδομένα τροχιών	Τροχιές αντικειμένων	Ακμές γράφου	RayTrace, Single Path Strategy	- Αρχικό χωρο-χρονικό σημείο ανάλυσης - Αβεβαιότητα εντοπισμού στη θέση των αντικειμένων (ε)	Όχι	Θεωρείται ένα κατανεμημένο σύστημα	-
[12]	Hotspot ανάλυση για μεγάλου όγκου δεδομένα τροχιών στο θαλάσσιο χώρο	Τροχιές αντικειμένων	Περιοχή σχήματος κύβου (3D cells)	THS (Trajectory Hot Spot), aTHS	- Παράμετροι για το 3D partitioning - k-top στατιστικά σημαντικά κελιά	Ναι	Distributed	-
[13]	Ανάλυσης μεγάλου όγκου κυκλοφοριακών δεδομένων για την ανάδειξη ‘city hotspots’	Δεδομένα κινητής τηλεφωνίας	Επιφάνεια	A-priori, Lorenz curve	Μη παραμετρική μέθοδος	Ναι	Distributed	-

2.3 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΤΡΟΧΙΩΝ

Οι Zygouras et al. [14] στη μελέτη τους αντιμετωπίζουν το πρόβλημα ανακάλυψης διαδρόμων από δεδομένα τροχιών προερχόμενα από GPS, με στόχο την εξόρυξη πολυσύχναστων διαδρομών. Στην προσέγγιση τους χωρίζουν το χώρο σε ομοιόμορφα κελιά με στόχο την ανακάλυψη περιοχών που μοιράζονται κοινή κυκλοφορία, κάνοντας χρήση του αλγορίθμου LDA που είναι ανάλογος του NLP. Το στάδιο της εξόρυξης διαδρόμων πραγματοποιείται κατάτμηση των τροχιών σε υπο-τροχιές με βάση τα κελιά μεγάλης συχνότητας και συσταδοποίηση αυτών για την αναγνώριση των διαδρόμων. Οι υπο-τροχίες συγχωνεύονται με χρήση ιεραρχικής συσταδοποίησης. Η διαδικασία συγχώνευσης των συστάδων σταματάει όταν η εσωτερική διακύμανση ξεπεράσει ένα προκαθορισμένο κατώφλι. Η απόσταση μεταξύ των τροχιών προκύπτει από τη μέθοδο Dynamic Time Wraping (DTW), καθώς λαμβάνει υπόψη την κατεύθυνση της κίνησης και αναθέτει μικρότερη απόσταση σε τροχιές με παρόμοιο σχήμα και χωρική διάταξη. Διάδρομοι δημιουργούνται για κελιά με κοινή κυκλοφορία που διαθέτουν συστάδες παρόμοιων τροχιών. Για την αντιμετώπιση της αβεβαιότητας που προκαλεί ο χαμηλός ρυθμός δειγματοληψίας από το GPS, κατασκευάζεται ένας κατευθυνόμενος γράφος από τις συστάδες που δημιουργήθηκαν. Σε κάθε ακμή ανατίθεται ένα βάρος που αφορά την πιθανότητα κίνησης σε αυτή, με βάση την επισκεψιμότητα κάθε κελιού και την πιο κοινή κατεύθυνση των συστάδων του. Η πιο κοινή διαδρομή εξάγεται από την εγγύτερη διαδρομή στο γράφο. Οι διάδρομοι προκύπτουν από τις συστάδες και κάθε πλήρης διαδρομή προστίθεται στο σύνολο των διαδρόμων, αν η απόσταση από τους υφιστάμενους διαδρόμους δεν υπερβαίνει ένα προκαθορισμένο όριο.

Οι J.-G. Lee et al. [16] προτείνουν μια μεθοδολογική προσέγγιση ‘partitioning - and-grouping’ με στόχο τη συσταδοποίηση τροχιών. Σύμφωνα με την προσέγγιση, ο αλγόριθμος TRACLUS που προτείνουν αποτελείται από δυο στάδια. Στο πρώτο πραγματοποιείται κατάτμηση των τροχιών σε υπο-τροχιές σε χαρακτηριστικά σημεία με χρήση της αρχής MDL (Minimum Description Length), δηλαδή σε σημεία που παρατηρείται μεγάλη αλλαγή στη συμπεριφορά

των τροχιών. Σε δεύτερη φάση, πραγματοποιείται η ανακάλυψη υπο-τροχιών με μεγάλη ομοιότητα κάνοντας χρήση μιας τροποποιημένης εκδοχής του DBSCAN για ευθύγραμμα τμήματα. Η μεθοδολογία ολοκληρώνεται με την εξαγωγή της χαρακτηριστικής τροχιάς από κάθε συστάδα.

Η μελέτη των Krogh et al. [16] πραγματεύεται τη χρήση τροχιών προερχόμενων από συσκευές GPS για τον υπολογισμό των μεγεθών της κυκλοφοριακής ροής, όπως η ταχύτητα ελεύθερης ροής, ο χρόνος αναμονής σε διασταύρωση και το μήκος της ουράς αναμονής, το πλήθος των «αποτυχημένων» κύκλων, κ.ά.. Μέσω της χρήσης των τροχιών δύναται η δυνατότητα υπολογισμού μεγεθών και δεικτών αναφορικά με το επίπεδο εξυπηρέτησης μεμονωμένων διασταυρώσεων σε διαφορετικές συνθήκες κυκλοφοριακού φόρτου. Παράλληλα με τη χρήση κανόνων συσχέτισης, δίνεται η δυνατότητα αξιολόγησης ολόκληρων διαδρομών με διαδοχικές διασταυρώσεις αναφορικά με την ποιότητα συγχρονισμού της διαδοχικής φωτεινής σηματοδότησης, ενώ σε επίπεδο πλοήγησης μπορεί να συμβάλλει στην μείωση του χρόνου διαδρομής και κατανάλωσης καυσίμου.

2.4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Συμπερασματικά, η διεθνής βιβλιογραφία μπορεί να χωριστεί σε ποικίλες κατηγορίες ως προς διάφορες παραμέτρους της ανάλυσης. Ο βασικότερος διαχωρισμός που γίνεται στη συγκεκριμένη διπλωματική εργασία αφορά τον τύπο δεδομένων εισόδου. Στην πρώτη κατηγορία, τα δεδομένα είναι σημειακά, αναφερόμενα κυρίως σε εντοπισμούς δραστηριοτήτων στο χώρο, ενώ στην δεύτερη, τα δεδομένα εισόδου αφορούν χωρο-χρονικά δεδομένα και ειδικότερα αναφέρονται σε τροχιές κινούμενων αντικειμένων.

Η κατηγοριοποίηση της βιβλιογραφίας, ωστόσο, μπορεί να πραγματοποιηθεί και σύμφωνα με τον τύπο των hotspots που αναδεικνύονται με το πέρας της ανάλυσης. Ειδικότερα, η ανάλυση οδηγεί άλλοτε στην ανάδειξη σημειακών hotspots, άλλοτε περιοχών ποικίλων γεωμετριών, παραμετρικών γεωμετρικών σχημάτων, αυθαίρετης γεωμετρίας, καθώς και υπογράφων που προκύπτουν

είτε από τμήματα του οδικού δικτύου, είτε με δυναμική κατάτμηση αυτού σε σημεία ενδιαφέροντος.

Ένα ακόμη πιθανό κριτήριο διαφοροποίησης των υφιστάμενων μελετών αποτελεί και ο τρόπος υπολογισμού των αποστάσεων μεταξύ των υπό εξέταση κινούμενων αντικειμένων. Οι αποστάσεις μπορεί να υπολογίζονται στον Ευκλείδειο χώρο και εναλλακτικά λαμβάνοντας υπόψη τις εξαρτήσεις που εισάγει το υποκείμενο οδικό δίκτυο. Η επιλογή της μεθόδου υπολογισμού των αποστάσεων εξαρτάται από τη φύση της εφαρμογής. Στην περίπτωση μελέτης της κυκλοφορίας και της κινητικότητας, ο υπολογισμός αποστάσεων πάνω στο οδικό δίκτυο μπορεί να θεωρεί πιο αντιπροσωπευτική.

Αναφορικά με την ποσοτικοποίηση του ‘hotness’, στη διεθνή βιβλιογραφία χρησιμοποιούνται μέτρα ενδιαφέροντος που βασίζονται στην πυκνότητα ,όπως το log-likelihood ratio και density-based ratio. Για την τεκμηρίωση της στατιστικής σημαντικότητας των αποτελεσμάτων, υλοποιείται κάποιο στατιστικό τεστ, όπως ‘Monte Carlo Simulation’. Η μέθοδος Getis - Ord Gi*, που χρησιμοποιείται επί το πλείστο σε εφαρμογές χωρικής ανάλυσης, παράγει τιμές z-score και p-values, αθροίζοντας τις τιμές πυκνότητας γειτονικών δεδομένων. Ένα σημείο θεωρείται hotspot όταν σχετίζεται με υψηλή βαθμολογία z-score και χαμηλή βαθμολογία p-value. Η μέθοδος Gi* παράγει στατιστικά σημαντικά hotspots βασιζόμενη στην πυκνότητα, χωρίς να απαιτεί εισαγωγή κατωφλιού από το χρήση ή περαιτέρω στατιστικούς ελέγχους.

Μέσω της βιβλιογραφικής ανασκόπησης που πραγματοποιήθηκε, συμπεραίνεται ότι οι μελέτες με σημειακά δεδομένα εισόδου ενσωματώνουν στην ανάλυση τους τον παράγοντα της στατιστικής σημαντικότητας στο σύνολο τους, αντίθετα οι εξεταζόμενες μελέτες με δεδομένα εισόδου τροχιές δεν τον έχουν λάβει υπόψη τους στο σύνολο τους. Επίσης σκόπιμο κρίνεται να σημειωθεί ότι περιορισμένες σε πλήθος είναι και οι μελέτες που έχουν υλοποιηθεί για κατανεμημένα συστήματα.

Αναγνωρίζοντας αυτές τις ελλείψεις στον τομέα της ανάλυσης hotspots, η συγκεκριμένη διπλωματική εργασία έρχεται να συμβάλλει προς αυτή την

κατεύθυνση. Στοχεύει στην υλοποίηση μιας μεθοδολογίας ανάλυσης hotspots για μεγάλα δεδομένα τροχιών κινούμενων αντικειμένων που θα οδηγεί σε στατιστικά σημαντικά αποτελέσματα, προτείνοντας μια παράλληλη και κλιμακωτή λύση. Τα hotspots που αναδεικνύονται από την ανάλυση αφορούν ακμές γράφου και συγκεκριμένο χρονικό παράθυρο ανάλυσης, καθώς ο δείκτης Getis Ord Gi*, εφαρμόζεται στα δεδομένα ενός 3D γράφου. Τέλος σκόπιμο είναι να σημειωθεί ότι σύμφωνα με τη βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε, είναι πρώτη φορά που χρησιμοποιείται Γκαουσσιανό Kernel για την ανάδειξη της χωρικής αλληλεπίδρασης χωρικών οντοτήτων σε εφαρμογή ανίχνευσης hotspots κυκλοφοριακής συμφόρησης.

3 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

3.1 APACHE SPARK

Το Apache Spark αποτελεί ένα εργαλείο γενικής χρήσης για την επεξεργασία μεγάλου όγκου δεδομένων, που χρησιμοποιείται σε μεγάλο φάσμα εφαρμογών και περιέχει πολλά ενσωματωμένα στοιχεία. Στον πυρήνα του, το Spark είναι μια υπολογιστική μηχανή που είναι υπεύθυνη για τον προγραμματισμό, την κατανομή και την παρακολούθηση της εκτέλεσης εφαρμογών, με σκοπό την παράλληλη εκτέλεση τους στους κόμβους (nodes) μίας συστάδας υπολογιστών (cluster).

Αποτελεί ίσως το πιο γρήγορο εργαλείο ανοικτού κώδικα. Η επεξεργασία των δεδομένων γίνεται στη μνήμη των κόμβων του cluster, καθιστώντας το έως και 100 φορές πιο γρήγορο από το Hadoop. Ταυτόχρονα είναι σχεδιασμένο ώστε να είναι εξαιρετικά φιλικό προς το χρήστη, προσφέροντας APIs σε Python, Java, Scala και SQL, καθώς και πλούσιες ενσωματωμένες βιβλιοθήκες. Μπορεί επίσης να ενσωματώσει και άλλα εργαλεία διαχείρισης μεγάλων δεδομένων. Ειδικότερα, μπορεί να τρέξει πάνω από Hadoop clusters.

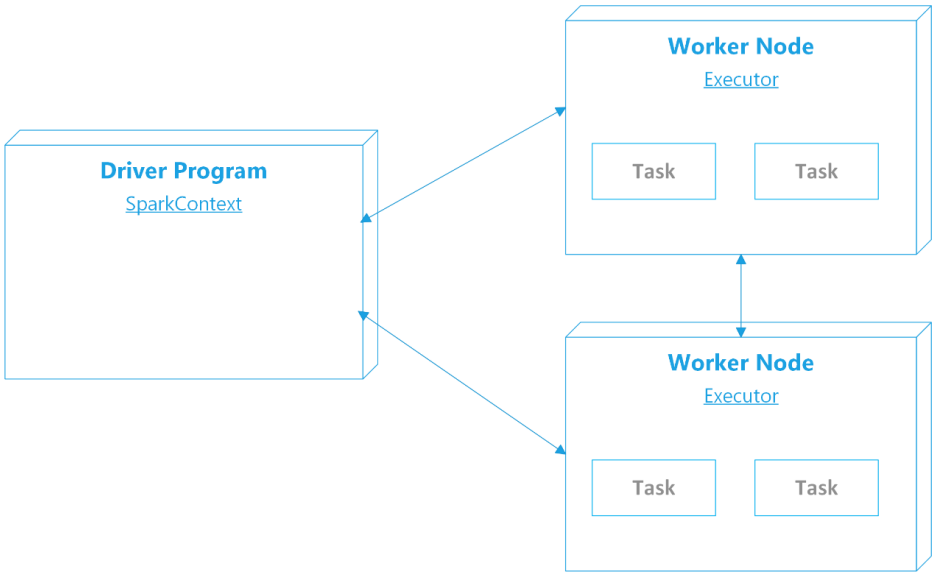
Επεκτείνει τη λογική του MapReduce για αποτελεσματικότερη εκτέλεση παράλληλων εργασιών, εισάγοντας την έννοια των Resilient Distributed Datasets (RDDs). Τα RDDs αποτελούν μια κατανεμημένη συλλογή από Java ή Python αντικείμενα διαμερισμένα σε ένα cluster, με στόχο να επιτρέπουν την επεξεργασία των δεδομένων απευθείας στην μνήμη των κόμβων. Αποτελούνται είτε από τιμές (values) είτε από ζεύγη κλειδιού- τιμής (key-value pairs). Κάθε διεργασία που δίνεται στο Spark μπορεί να είναι είτε η δημιουργία ενός RDD, είτε ένας μετασχηματισμός (transformation) είτε μια ενέργεια (action) πάνω σε ένα RDD. Το σύνολο των διενεργειών δέχονται ως δεδομένα εισόδου RDDs και οδηγούν στη δημιουργία νέων.

Οι μετασχηματισμοί στα RDDs είναι ‘lazy evaluated’, που σημαίνει ότι το Spark δεν εκτελεί τους μετασχηματισμούς αν δεν εντοπίσει πρώτα μια ενέργεια. Το χαρακτηριστικό αυτό του Spark συμβάλλει στην αύξηση της ταχύτητας του. Το

σύνολο των μετασχηματισμών που απαιτούνται καταγράφονται σε ένα λογικό διάγραμμα ροής που ονομάζεται Logical Acyclic Graph (DAG). Το Spark εφαρμόζει σε αυτό διάφορες βελτιστοποιήσεις και μετατρέπει το DAG σε ένα σύνολο από στάδια (stages). Κάθε στάδιο (stage) αποτελείται από πολλές εργασίες (tasks). Το σύνολο των σταδίων εκτελείται όταν ζητηθεί η εκτέλεση μιας ενέργειας πάνω στα δεδομένα.

Επιπρόσθετα, το Spark πέραν των RDDs διαθέτει και δυο επιπλέον τύπους μεταβλητών που διανέμονται στους κόμβους ενός cluster. Οι broadcast μεταβλητές οι οποίες είναι μόνο για ανάγνωση και διανέμονται στο σύνολο τους σε κάθε κόμβο και οι accumulators που είναι μεταβλητές όπως αθροίσματα και μετρητές και μπορούν όλοι οι κόμβοι να προσθέσουν σε αυτές.

Σε ένα υψηλό επίπεδο, κάθε εφαρμογή σε Spark αποτελείται από ένα driver program, που εκκινεί την εκτέλεση διάφορων παράλληλων εργασιών σε ένα cluster. Το πρόγραμμα αυτό περιέχει μια main συνάρτηση που ορίζει τον διαμερισμό του συνόλου δεδομένων στο cluster και τις διενέργειες που θα εκτελεστούν σε αυτό. Το πρόγραμμα αυτό αποκτά πρόσβαση στο Spark με ενός αντικειμένου SparkContent, που ουσιαστικά αναπαριστά τη σύνδεση με ένα σύμπλεγμα υπολογιστών. Για την εκτέλεση των διεργασιών, το driver program τυπικά διαχειρίζεται ένα πλήθος από κόμβους που ονομάζονται executors.



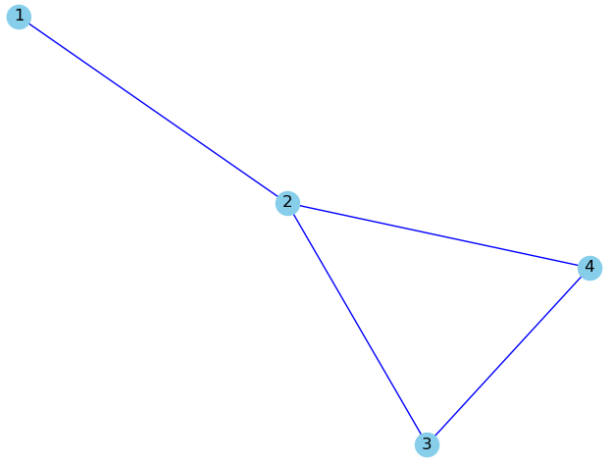
Εικόνα 3. Συνιστώσες μιας κατανεμημένης διενέργειας σε Spark.

Σκόπιμο είναι να σημειωθεί ότι η ανάπτυξη ενός αλγορίθμου σε Spark ακόμα και όταν ο όγκος των διαθέσιμων δεδομένων δεν είναι μεγάλος έχει ως πλεονέκτημα ότι το πρόγραμμα έχει δυνατότητα κλιμάκωσης. Η μείωση του χρόνου εκτέλεσης του προγράμματος γίνεται φανερή σε μεγάλου όγκου δεδομένα.

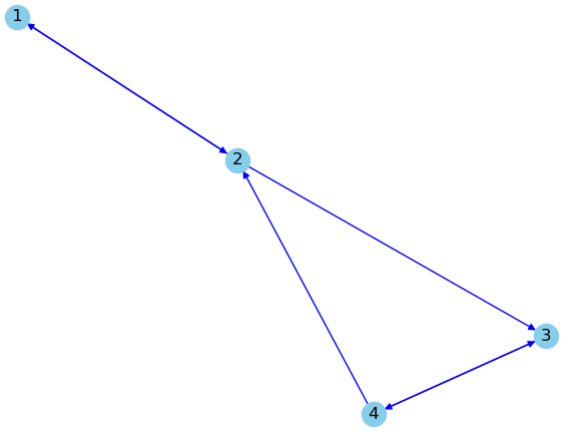
3.2 ΑΝΑΠΑΡΑΣΤΑΣΗ ΟΔΙΚΟΥ ΔΙΚΤΥΟΥ ΣΕ ΜΟΡΦΗ ΓΡΑΦΟΥ

Η μετακίνηση των ανθρώπων από μια δραστηριότητα σε μια άλλη πραγματοποιείται μέσω των υποδομών που προσφέρει το σύστημα μεταφορών. Ένα τέτοιο είδος υποδομής είναι το οδικό δίκτυο. Για τη μελέτη του απαραίτητη είναι η αναπαράσταση του με ακριβή τρόπο. Με μαθηματικούς όρους, τα συστήματα αυτά αναπαρίστανται σε μορφή γράφων ή αλλιώς δικτύων. Ένας γράφος G ορίζεται ως G = (V,E) και αποτελείται από ένα σύνολο ακμών E που αναπαριστούν οδικά τμήματα και ένα σύνολο κόμβων V που αναπαριστούν τις διασταυρώσεις.

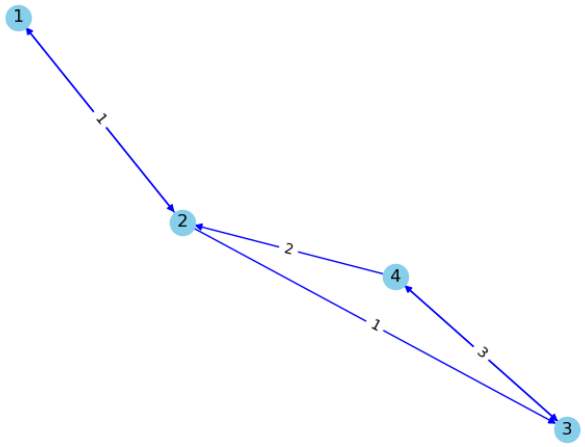
Μια ακμή $e_i \in E$ ορίζεται από δυο κόμβους $v_i, v_j \in V$. Ένας κόμβος συνδέει δυο ή περισσότερες ακμές. Οι ακμές μπορεί να είναι είτε κατευθυνόμενες $e = (v_i, v_j) \in E$, υποδεικνύοντας ότι οι κόμβοι v_i, v_j συνδέονται απευθείας και η κίνηση είναι δυνατή μόνο με κατεύθυνση από το v_i στο v_j ή μη κατευθυνόμενοι. Σημαντικές ιδιότητες των ακμών του οδικού δικτύου που μπορεί να συμπεριλαμβάνονται στο γράφο είναι το μήκος ακμής και το κόστος. Το μήκος της ακμής αντιστοιχεί στο μήκος του οδικού τμήματος που ενώνει δυο κόμβους και η έννοια του κόστους ακμής χρησιμοποιείται για να περιγράψει την ωφέλεια ενός χρήστη όταν αυτός επιλέγει να ταξιδέψει μέσω της συγκεκριμένης ακμής. Ο χρόνος ταξιδιού ή το άμεσο κόστους, για παράδειγμα της κατανάλωσης καυσίμου, μπορεί να χρησιμοποιηθεί για την ποσοτικοποίηση της ωφέλειας. Οι γράφοι που ενσωματώνουν κάποια επιπλέον πληροφορία στις ακμές τους αναφέρονται στη βιβλιογραφία ως σταθμισμένοι (weighted graph).



Εικόνα 4. Μη κατεθυνόμενος γράφος



Εικόνα 5. Κατευθυνόμενος γράφος



Εικόνα 6.Σταθμισμένος κατευθυνόμενος γράφος

Στη συγκεκριμένη διπλωματική εργασία, η αναπαράσταση του οδικού δικτύου γίνεται μέσω ενός κατευθυνόμενου γράφου, καθώς για τη μελέτη της κυκλοφοριακής συμφόρησης και την ανάδειξη χωρικών προτύπων σημαντικό ρόλο έχει η κατεύθυνση της κίνησης των αντικειμένων. Η πληροφορία της κατεύθυνση του γράφου αποτυπώνει την κατεύθυνση με την οποία επιτρέπεται να κινούνται τα οχήματα σε κάθε οδικό τμήμα.

Για την απόκτηση δεδομένων γράφου που αναπαριστούν το οδικό δίκτυο και την επεξεργασία αυτών χρησιμοποιήθηκαν οι βιβλιοθήκες της Python OSMnx και NetworkX.

Η βιβλιοθήκη OSMnx αποτελεί ένα πακέτο της Python που δίνει τη δυνατότητα ανάκτησης, ανάλυσης και οπτικοποίησης του οδικού δικτύου από το Open Street Maps. Το Open Street Map παρέχει ελεύθερη πρόσβαση σε ένα παγκόσμιο χάρτη με δυνατότητα επεξεργασίας, ο οποίος δημιουργήθηκε και ενημερώνεται από μια κοινότητα χαρτογράφων και περιέχει πληροφορία για το οδικό δίκτυο, στάσεις και σταθμούς των αστικών συγκοινωνιών και άλλα σημεία ενδιαφέροντος. Αναφορικά με το οδικό δίκτυο, παρέχεται η πληροφορία της γεωμετρίας, αλλά και η περιγραφική πληροφορία για το όριο ταχύτητας και την κατηγορία της οδού.

Η βιβλιοθήκη NetworkX αποτελεί ένα πακέτο της Python, που διαχειρίζεται γράφους και δίκτυα. Υποστηρίζει τύπους δεδομένων όπως κατευθυνόμενους (directed graphs) και μη κατευθυνόμενους γράφους (undirected graphs), καθώς και πολυγράφους (multigraphs).

3.3 ΧΩΡΙΚΗ ΑΝΑΛΥΣΗ

Η χωρική ανάλυση περιλαμβάνει ένα σύνολο από ποσοτικές τεχνικές που μελετούν οντότητες και φαινόμενα που εξελίσσονται στο χώρο και στο χρόνο, χρησιμοποιώντας τις τοπολογικές, γεωμετρικές ή γεωγραφικές ιδιότητες τους [20].

Ειδικότερα ως χωρική ανάλυση μπορεί να οριστεί η διαδικασία δημιουργίας-εξόρυξης νέων πληροφοριών για ένα σύνολο γεωγραφικών χαρακτηριστικών οντοτήτων μέσα από τη εξέταση, αξιολόγηση και επεξεργασία των στοιχείων μιας γεωγραφικής περιοχής, σύμφωνα με τα προκαθορισμένα χωρικά πρότυπα. Στόχος της είναι η παρακολούθηση, αποτύπωση, προσμέτρηση, πρόβλεψη, ερμηνεία και κατανόηση περίπλοκων χωρικών φαινομένων και κατ΄ επέκταση την τροφοδότηση της διαδικασίας του χωρικού σχεδιασμού. [20]

Η ποσοτική χωρική ανάλυση χρησιμοποιεί χωρικά στατιστικά εργαλεία για την ανάλυση χωρικών κατανομών, προτύπων και σχέσεων. Εστιάζει στο γεωγραφικό παράγοντα και εξαρτάται άμεσα από συγκεκριμένες χωρικές μεταβλητές για την αξιολόγηση ή την ερμηνεία ενός φαινομένου. Βασικό

χαρακτηριστικό της ποσοτικής χωρικής ανάλυσης είναι η ενσωμάτωση της γεωγραφικής πληροφορίας (εγγύτητα, συνδεσιμότητα και άλλες χωρικές σχέσεις) κατευθείαν μέσα στα μαθηματική της σχέση. Σε αντίθεση, η μη χωρική ανάλυση δεν απαιτεί την ενσωμάτωση του χωρικού παράγοντα. Συνεπώς, η θεμελιώδης διαφορά μεταξύ των δυο αυτών μορφών ανάλυσης είναι η παρουσία ή απουσία του χωρικού παράγοντα στη διαδικασία.

3.3.1 Χωρική Αυτοσυσχέτιση

Ως χωρική αυτοσυσχέτιση μπορεί να οριστεί η ύπαρξη ομοιότητας ή αλληλεξάρτησης ενός αντικειμένου με τα γειτονικά του αντικείμενα στο χώρο, δηλαδή χωρική συσχέτιση υπάρχει όταν η τιμή μιας μεταβλητής, που αναφέρεται σε μια συγκεκριμένη χωρική ενότητα, επηρεάζει και επηρεάζεται από τις τιμές της ίδιας μεταβλητής στις γειτονικές χωρικές ενότητες (Kitchin and Tate, 2000).

"The first law of geography: Everything is related to everything else, but near things are more related than distant things." Waldo R. Tobler (Tobler 1970)

Στόχος της χωρικής αυτοσυσχέτισης είναι να εντοπίσει τις οντότητες εκείνες που έχουν ξεχωριστό ρόλο για την ευρύτερη περιοχή τους. Εκτός από την γενική τιμή αυτοσυσχέτισης για το σύνολο της περιοχής μελέτης , η χρήση του τοπικού δείκτη αυτοσυσχέτισης βοηθά στον εντοπισμό οντοτήτων που φέρουν τιμές διαφορετικές από τον περίγυρο τους και μπορούν να αποτελέσουν περιοχές δυναμικές ή αδύναμες για την εκάστοτε μεταβλητή. Οι τεχνικές χωρικής αυτοσυσχέτισης εντοπίζουν κάθε φορά ένα συγκεκριμένο αριθμό οντοτήτων που είναι στατιστικά σημαντικές [18].

3.3.2 Gi* Spatial Statistics

Η οικογένεια των G statistics, που αναπτύχθηκε αρχικά από τους Getis και Ord [18,19], χρησιμοποιείται για την αναγνώριση χωρικών προτύπων. Όπως και οι δείκτες Moran's I και Geary's C, ο γενικός δείκτης G είναι καθολικός, καθώς μελετά το συνολικό βαθμό χωρικής αλληλεξάρτησης και οδηγεί στην εξαγωγή ενός ενιαίου δείκτη για το σύνολο της υπό μελέτη περιοχής. Οι καθολικοί δείκτες

καταλήγουν σε ένα γενικευμένο αποτέλεσμα, που τείνει να εξουδετερώνει τα χωρικά πρότυπα σε τοπικό επίπεδο καθιστώντας τη χωρική αλληλεξάρτηση μη ανιχνεύσιμη. Το γεγονός ότι η χωρική αλληλεξάρτηση μπορεί να διαφέρει τοπικά οδήγησε στην ανάγκη ανάπτυξης τοπικών στατιστικών, που αναδεικνύουν την χωρική ανομοιογένεια. Το τοπικό στατιστικό Gi* αναπτύχθηκε σε συμφωνία με το καθολικό στατιστικό Gi για την αντιστάθμιση του προαναφερθέντος περιορισμού. Διαδεδομένη είναι η χρήση του σε εφαρμογές ανάλυσης hot spots. Πλεονέκτημα του τοπικού δείκτη Gi, σε σχέση με αντίστοιχα στατιστικά (π.χ. τοπικός Moran's I), είναι η ικανότητα του να διακρίνει hot spots και cold spots. Συνεπώς, ο δείκτης Gi* μπορεί να χαρακτηριστεί καταλληλότερος για τη συγκεκριμένη εφαρμογή, καθώς μπορεί να διαχωρίσει να κάνει τη διάκριση μεταξύ ακμών χαμηλής και υψηλής κυκλοφοριακής συμφόρησης.

Μια απλοϊκή μορφή του δείκτη Gi*, έτσι όπως ορίζεται από τους Getis και Ord είναι η εξής:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}\, x_j}{\sum_{j=1}^{n} x_j} \qquad \textbf{(3.1)}$$

Όπου:

G_i^* : στατιστικός δείκτης που περιγράφει τη χωρική αλληλεξάρτηση μεταξύ της παρατήρησης i και του συνόλου των διακριτών χωρικών ενοτήτων
w_{ij} : η σχέση γειτνίασης μεταξύ των χωρικών ενοτήτων i και j
x_j : η τιμή της μεταβλητής για την χωρική ενότητα j
n : ο συνολικός αριθμός των διακριτών χωρικών ενοτήτων

Ο κανονικοποιημένος δείκτης Gi* είναι ουσιαστικά ένα Z score και κατά συνέπεια μπορεί να του αποδοθεί στατιστική σημαντικότητα.

$$Z\left(G_i^*\right) = \frac{\sum_{j=1}^{n} w_{ij}\, x_j \; - \; \bar{X} \sum_{j=1}^{n} w_{ij}^2}{s\sqrt{\frac{n\sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij}\right)^2}{n-1}}} \quad \textbf{(3.2)}$$

Όπου:

G_i^* : στατιστικός δείκτης που περιγράφει τη χωρική αλληλεξάρτηση μεταξύ της παρατήρησης i και του συνόλου των διακριτών χωρικών ενοτήτων

w_{ij} : η σχέση γειτνίασης μεταξύ των χωρικών ενοτήτων i και j

x_j : η τιμή της μεταβλητής για την χωρική ενότητα j

n : ο συνολικός αριθμός των διακριτών χωρικών ενοτήτων

$\bar{X}$: η μέση τιμή της μεταβλητής x

$$\bar{X} = \frac{\sum_{j=1}^{n} x_i}{n} \quad \textbf{(3.3)}$$

s : η τυπική απόκλιση της μεταβλητής x

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{X}^2} \quad \textbf{(3.4)}$$

4 ΠΕΡΙΓΡΑΦΗ ΠΡΟΒΛΗΜΑΤΟΣ

4.1 ΔΕΔΟΜΕΝΑ

Το σύνολο δεδομένων *D* που χρησιμοποιήθηκε για την ανάδειξη της κυκλοφοριακής συμφόρησης αποτελείται από τροχιές οχημάτων $\tau \in D$. Ως τροχιά τ ορίζεται μια αλληλουχία από σημειακά δεδομένα. Κάθε σημείο p είναι της μορφής *(Vehicle_ID, x, y, t)*, δηλαδή περιέχει την πληροφορία της γεωγραφικής θέσης (γεωγραφικό μήκος, γεωγραφικό πλάτος), της χρονικής στιγμής και του αναγνωριστικού του οχήματος στην οποία αναφέρεται. Τα δεδομένα που πρόκειται να χρησιμοποιηθούν για την συγκεκριμένη ανάλυση αποτελούν ιστορικά δεδομένα τροχιάς οχημάτων, όπως αυτά προέκυψαν από δείκτες GPS.

Για τις ανάγκες της εφαρμογής απαραίτητος είναι ο εμπλουτισμός του βασικού συνόλου δεδομένων *D* με χωρική και περιγραφική πληροφορία, που αφορά το οδικό τμήμα πάνω στο οποίο πραγματοποιείται η εκάστοτε κίνηση.

Πίνακας 4. Συνοπτικός πίνανας παρουσίασης μεταβλητών

Συμβολισμός	**Περιγραφή**
D	Σύνολο δεδομένων
τ ∈ D	Τροχιά που ορίζεται ως μια αλληλουχία σημείων
p	Σημειακό δεδομένο / Παρατήρηση
G	Γράφος που αναπαριστά το οδικό δίκτυο
$\mathbf{e_i} \in \mathbf{G}$	Ακμή του γράφου
$\mathbf{x_i}$	Δείκτης κυκλοφοριακής συμφόρησης για μια ακμή e_i
$\mathbf{w_{ij}}$	Σχέση γειτνίασης / Βάρος μεταξύ δυο ακμών e_i και e_j
h	Εύρος ζώνης / Bandwidth του Gaussian kernel
a	Σταθερά που χρησιμοποιείται σε συνάρτηση βάρους
n	Πλήθος ακμών στο γράφο G
$\mathbf{G_i^*}$	Z-score από του στατιστικό Getis-Ord για ακμή e_i
$\mathbf{d_{ij}^T}$	Απόσταση μεταξύ ακμών e_i και e_j στον άξονα του χρόνου
$\mathbf{d_{ij}^S}$	Χωρική απόσταση μεταξύ ακμών e_i και e_j σε hops

4.2 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ

Το στάδιο της προεπεξεργασίας των δεδομένων αποτελείται από δυο βήματα, το βήμα του Map-matching και κατ' επέκταση του εμπλουτισμού των δεδομένων με πληροφορία που αφορά το οδικό δίκτυο και το βήμα της δημιουργίας αναγνωριστικού τροχιάς τ.

4.2.1 Map-matching

Ως τροχιά *τ* ορίζεται μια αλληλουχία από σημεία *p*, που περιγράφουν την κίνηση ενός οχήματος. Τα σημειακά αυτά δεδομένα ενσωματώνουν πληροφορία που αφορά τη γεωγραφική θέση και τη χρονική στιγμή καταγραφής αυτής. Εξαιτίας σφαλμάτων κατά τη συλλογή δεδομένων από τους δείκτες GPS, παρατηρούνται αποκλίσεις ανάμεσα στην καταγραφόμενη και πραγματική θέση των οχημάτων. Αυτό έχει ως αποτέλεσμα να κρίνεται αναγκαία η αντιστοίχιση των συλλεγόμενων δεδομένων με το οδικό τμήμα πάνω στο οποία πραγματοποιείται η κίνηση. Η διαδικασία αυτή αντιστοίχισης ονομάζεται Map-matching και αποτελεί ένα βασικό βήμα προεπεξεργασίας για εφαρμογές που κάνουν χρήση δεδομένων τροχιών.

Εικόνα 7. Τροχία οχημάτων όπως προκύπτει από δείκτη GPS σε σχέση με το οδικό δίκτυο

Το Map-matching αποτελεί ένα διαφορετικό αντικείμενο έρευνας με αποτέλεσμα να έχουν αναπτυχθεί ποικίλοι αλγόριθμοι με στόχο την αύξηση της ακρίβειας του εξαγόμενου αποτελέσματος.

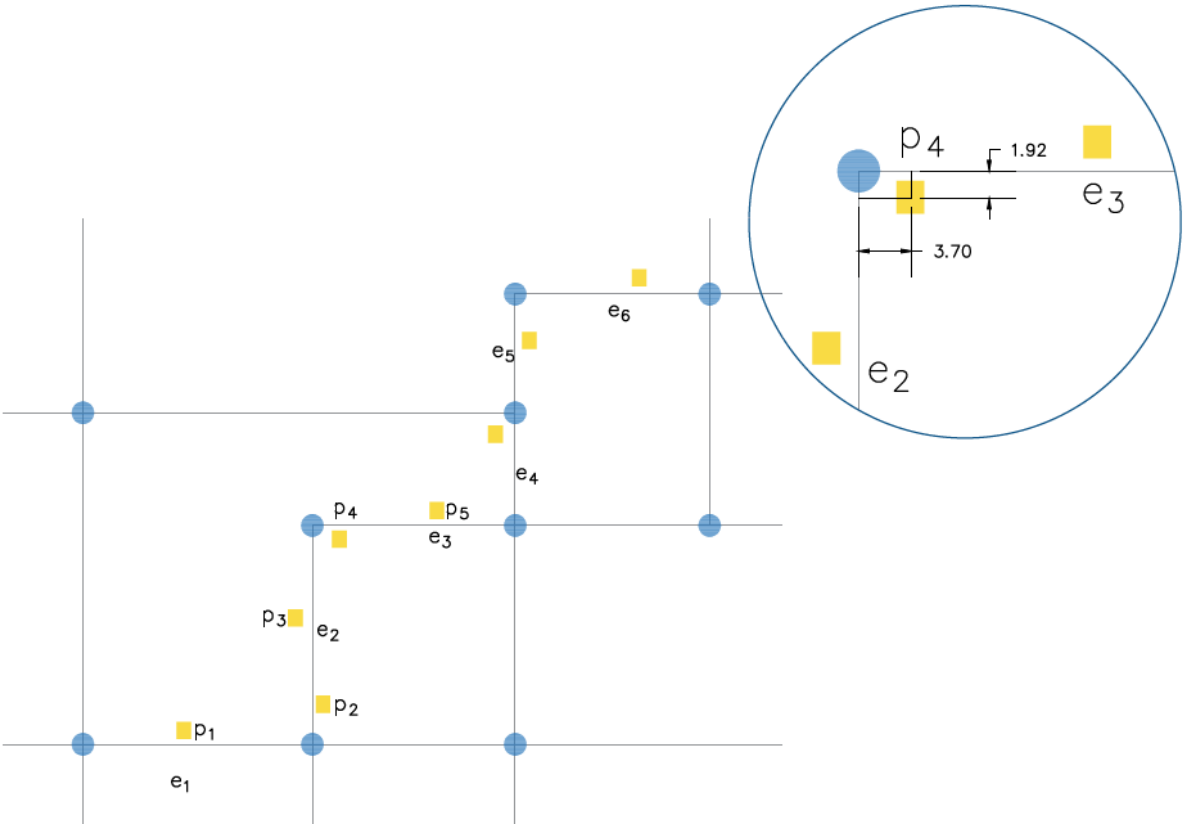
Στη συγκεκριμένη διπλωματική εργασία χρησιμοποιήθηκε μια απλοϊκή τεχνική για την υλοποίηση του Map-matching. Αρχικά, το οδικό δίκτυο της περιοχής

μελέτης αναπαραστάθηκε σε μορφή γράφου. Η περιοχή μελέτης ορίστηκε από ένα παραλληλόγραμμο (box) και το οδικό δίκτυο εντός αυτής ανακτήθηκε μέσω της βιβλιοθήκης OSMnx και της συνάρτησης *graph_from_box*. Από το σύνολο του οδικού δικτύου επιλέχθηκαν μόνο συγκεκριμένες κατηγορίες δρόμων που εξυπηρετούν αποκλειστικά την κίνηση των οχημάτων, με σκοπό τον περιορισμό του όγκου των δεδομένων προς επεξεργασία. Για τον ίδιο σκοπό, πραγματοποιείται ένας έλεγχος ως προς τη γεωμετρία του δικτύου. Η γεωμετρία του δικτύου απλοποιείται, απαλείφοντας αλγοριθμικά τους κόμβους που χωρίζουν ένα ενιαίο τμήμα δρόμου σε μικρότερα χωρίς να εξυπηρετούν τη σύνδεση του με άλλα τμήματα. Με αυτή τη διόρθωση η γεωμετρία της οδού παραμένει αμετάβλητη, ενώ το πλήθος των ακμών μειώνεται στις πλέον απαραίτητες. Ο γράφος που προέκυψε πέραν από την γεωμετρία του οδικού δικτύου περιλαμβάνει πληροφορία για το όριο ταχύτητας και την κατηγορία κάθε οδικού τμήματος που αποτυπώνει. Εξετάζοντας ωστόσο την πληροφορία που είναι ενσωματωμένη στο γράφο, παρατηρείται ότι υφίστανται αγνοούμενες τιμές στο πεδίο που αφορά την ταχύτητα. Οι αγνοούμενες τιμές συμπληρώνονται αποδίδοντας σε κάθε κατηγορία δρόμου μια μέση επιτρεπόμενη ταχύτητα.

Στη συνέχεια, κάθε σημείο p αντιστοιχίζεται στην εγγύτερη ακμή, με χρήση της συνάρτησης *get_nearest_edge*. Η συνάρτηση αυτή δημιουργεί σε κάθε ακμή του γράφου σημεία σε ίση απόσταση μεταξύ τους. Τα σημεία αυτά χρησιμεύουν στη δημιουργία ενός Ball tree, με σκοπό την αναγνώριση των εγγύτερων σημείων σε κάθε μια από τις παρατηρήσεις που διαθέτουμε και την αντιστοίχιση αυτών με την εγγύτερη ακμή κατ’ επέκταση. Το Ball tree αποτελεί μέθοδος κατάτμησης των δεδομένων σε ένα δυαδικό δέντρο. Ειδικότερα, χωρίζει τα δεδομένα σε δυο υπερ-σφαίρες, δημιουργώντας δυο συστάδες. Ένα σημείο ανατίθεται στη σφαίρα από την οποία έχει μικρότερη απόσταση. Κάθε υπερ-σφαίρα χωρίζεται σε μικρότερες, μέχρι ένα συγκεκριμένο βάθος.

Στην Εικόνα 8 γίνεται μια διαγραμματική απεικόνισης της λειτουργία της συνάρτησης *get_nearest_edge* στην περίπτωση του σημείου p_4. Το σημείο p_4 θα μπορούσε να αντιστοιχιστεί είτε στην ακμή e_2 είτε στην ακμή e_3.

Υπολογίζοντας την απόσταση μεταξύ του σημείου και των δυο υποψήφιων ακμών, προκύπτει ότι η εγγύτερη ακμή είναι η e_3.



Εικόνα 8. Διαγραμματική απεικόνισης της αντιστοίχισης ενός σημείο στην εγγύτερη ακμή

Έχοντας αντιστοιχίσει κάθε σημείο στην εγγύτερη ακμή της, γίνεται δυνατός ο εμπλουτισμός των αρχικών δεδομένων με επιπλέον πληροφορία που εμπεριέχεται στο γράφο και αφορά το οδικό δίκτυο. Η πληροφορία που ενσωματώνεται στα αρχικά δεδομένα για τις ανάγκες της ανάλυσης αφορά την αναγνωριστικό της εγγύτερης ακμής, τη γεωγραφική πληροφορία των δυο κορυφών που την ορίζουν, το μήκος της ακμής, καθώς και περιγραφική πληροφορία που αφορά το όριο ταχύτητας.

4.2.2 Δημιουργία αναγνωριστικού τροχιάς

Το δεύτερο βήμα της προεπεξεργασίας αφορά τη δημιουργία ενός αναγνωριστικού τροχιάς, *Trajectory_ID*. Η λογική δημιουργίας του μοναδικού αυτού κωδικού είναι ότι ένα *Trajectory_ID* αντιστοιχεί σε δεδομένα τροχιάς ενός και μόνο οχήματος. Για τη δημιουργία του, τα δεδομένα ταξινομούνται σε

αύξουσα σειρά πρώτα κατά *Vehicle_ID* και στη συνέχεια κατά *timestamp*. Ένας μετρητής τίθεται και η τιμή του αυξάνεται κατά μια μονάδα κάθε φορά που αλλάζει το *Vehicle_ID* ή το χρονικό διάστημα μεταξύ δυο διαδοχικών παρατηρήσεων υπερβαίνει τα πέντε λεπτά. Το χρονικό όριο των πέντε λεπτών αποτελεί μια υπόθεση, σύμφωνα με την οποία θεωρείται ότι στην περίπτωση που το χρονικό διάστημα μεταξύ δυο διαδοχικών παρατηρήσεων από το ίδιο όχημα υπερβαίνει τα πέντε λεπτά οφείλεται είτε σε πιθανή στάθμευση/στάση, είτε σε πιθανή απενεργοποίηση της συσκευής εντοπισμού. Το αναγνωριστικό τροχιάς που ανατίθεται σε κάθε μία από τις παρατηρήσεις αποτελεί ένα σύνθετο αναγνωριστικό, που προκύπτει από συνδυασμό του *Vehicle_ID* και της τιμής του μετρητή.

Με το πέρας του σταδίου της προεπεξεργασίας τα αρχικά δεδομένα της μορφής *(Vehicle_ID, x, y, t)* μετατρέπονται στη μορφή *(Vehicle_ID, Traj_ID, x, y, t, ei, Length, Speed)*.

4.3 ΔΕΙΚΤΗΣ ΚΥΚΛΟΦΟΡΙΑΚΗΣ ΣΥΜΦΟΡΗΣΗΣ

Η κυκλοφοριακή συμφόρηση αποτελεί αντικείμενο της κυκλοφοριακής τεχνικής. Χαρακτηρίζεται από υψηλό χρόνο ταξιδιού και χαμηλή ταχύτητα. Πολλοί αλγόριθμοι ανίχνευσης κυκλοφοριακής συμφόρησης βασίζονται στον υπολογισμό της ταχύτητας των οχημάτων. Ομοίως και στη συγκεκριμένη διπλωματική εργασία, η ανίχνευση των ακμών με κυκλοφοριακή συμφόρηση πραγματοποιήθηκε μέσω μετρήσεων ταχύτητας.

Ειδικότερα, έχοντας αντιστοιχίσει κάθε σημείο δειγματοληψίας στην ακμή που πραγματοποιείται η κίνηση, είναι δυνατός ο υπολογισμός του χρονικού διαστήματος που χρειάστηκε το όχημα για να τη διανύσει. Γνωρίζοντας το χρονικό αυτό διάστημα και το μήκος της ακμής, είναι δυνατός ο υπολογισμός της σταθερής ταχύτητας με την οποία το όχημα διένυσε το οδικό τμήμα.

Γίνεται αντιληπτό, ωστόσο, ότι εφόσον το στίγμα που συλλέγεται από κάθε όχημα είναι δειγματοληπτικό, η χρονική στιγμή εισόδου και εξόδου από την υπό μελέτη ακμή δεν είναι γνωστή εκ των προτέρων. Για την αντιμετώπιση αυτού

του προβλήματος, γίνεται εκτίμηση της χρονικής στιγμής που το όχημα διέσχισε την κορυφή μέσω της μεθόδου της γραμμικής παρεμβολής. Μια αντίστοιχη προσέγγιση χρησιμοποιείται και στη μελέτη [11].

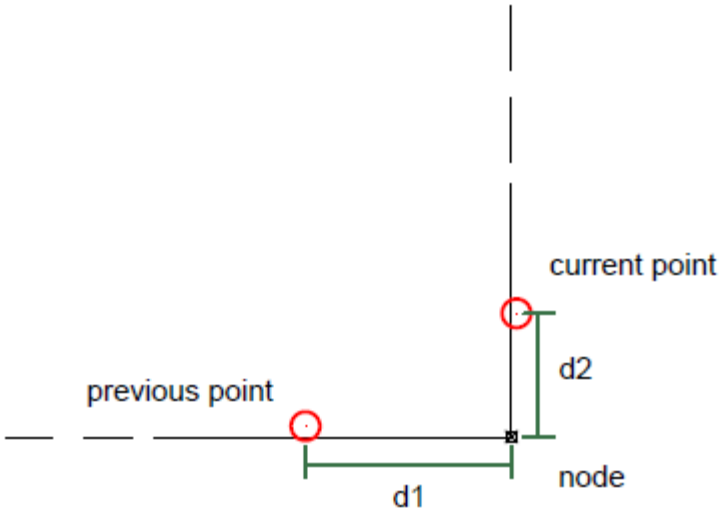
Αναλυτικότερα, για τον υπολογισμό της χρονικής στιγμής εισόδου και εξόδου από μια ακμή, κάθε τροχιά εξετάζεται διαφορετικά. Στόχος είναι η εύρεση διαδοχικών σημείων δειγματοληψίας, που ανήκουν σε διαφορετικές ακμές, με την προϋπόθεση ότι αυτές οι ακμές είναι διαδοχικές, δηλαδή μοιράζονται μια γειτονική κορυφή.

Ο τύπος που χρησιμοποιείται για τον υπολογισμό της χρονικής στιγμής διέλευσης από την κοινή κορυφή είναι ο εξής:

$$t_{node} = t_1 + d_1 \times \frac{t_2 - t_1}{d_1 + d_2} \quad \textbf{(4.1)}$$

Όπου:

t_1: η τελευταία χρονική σήμανση πριν την έξοδο από την ακμή
t_2: η πρώτη χρονική σήμανση αμέσως μετά την είσοδο στην διαδοχική ακμή
d_1, d_2: η γεωδαιτική απόσταση μεταξύ των σημείων των σημείων 1 και 2 με την κοινή κορυφή



Εικόνα 9. Απεικόνιση των παραμέτρων που απαιτούνται για εκτίμηση της χρονικής στιγμής εισόδου στην κορυφή

Διαθέτοντας την πληροφορία της χρονικής στιγμής εισόδου και εξόδου από κάθε ακμή, ο υπολογισμός της ταχύτητας με την οποία το κινούμενο όχημα διέσχισε την υπό εξέταση ακμή είναι εφικτός, μέσω της εφαρμογής του εξής τύπου:

$$V_{obs} = \frac{l}{t_{out} - t_{in}} \quad \textbf{(4.2)}$$

Όπου:

l: το μήκος της υπό εξέταση ακμής
t_{in} , t_{out}: η χρονική στιγμή εισόδου και εξόδου από την ακμή, όπως υπολογίστηκε από τον τύπο (4.1)

Γνωρίζοντας πλέον την ταχύτητα που έχει αναπτυχθεί από όλα τα κινούμενα αντικείμενα σε κάθε ακμή, ο υπολογισμός του δείκτη της κυκλοφοριακής συμφόρησης σε επίπεδο ακμής είναι δυνατός.

Σύμφωνα με τους H.Xiong et al. (2018), ο τύπος που εκφράζει την κυκλοφοριακή συμφόρηση είναι ο εξής:

$$x_i = 1 - \frac{V_{obs}}{V_{ffs}} \quad \textbf{(4.3)}$$

Όπου:

V_{obs}: η παρατηρούμενη ταχύτητα
V_{ffs}: η ταχύτητα ελεύθερης ροής

Ως ταχύτητα ελεύθερης ροής, θεωρείται η ταχύτητα η οποία αναπτύσσεται από τα οχήματα όταν η κίνηση τους γίνεται απρόσκοπτα. Η μέτρηση της ταχύτητας ελεύθερης ροής απαιτεί εργασία πεδίου και αποτελεί μια επίπονη διαδικασία ειδικότερα στην περίπτωση μεγάλων οδικών δικτύων. Για τον λόγο αυτό, κάναμε την παραδοχή ότι το όριο ταχύτητας μίας οδού αποτελεί μια ικανοποιητική προσέγγιση της ταχύτητας ελεύθερης ροής της.

Αναφορικά με την ερμηνεία του δείκτη κυκλοφοριακής συμφόρησης, προκύπτει ότι υψηλές τιμές του δείκτη αντιστοιχούν σε υψηλό επίπεδο συμφόρησης, ενώ χαμηλές τιμές του δείκτη αντιστοιχούν σε χαμηλό επίπεδο συμφόρησης. Οι επιτρεπτές τιμές του δείκτη κυμαίνονται στο κλειστό διάστημα [0,1].

Έχοντας πραγματοποιήσει κατάτμηση της πληροφορίας σε επίπεδο ακμών γράφου και υπολογίσει τις απαραίτητες μεταβλητές, σειρά έχει η κατάτμηση της πληροφορίας σε μια τρίτη διάσταση αυτή του χρόνου. Ο άξονας του χρόνου χωρίζεται σε ίσα μη επικαλυπτόμενα διαστήματα, με στόχο τον υπολογισμό της μέσης τιμής του δείκτη κυκλοφοριακής συμφόρησης ανά αναγνωριστικό ακμής και χρονική περίοδο ανάλυσης.

4.4 ΧΩΡΟ-ΧΡΟΝΙΚΗ HOTSPOT ΑΝΑΛΥΣΗ

Στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η ανάδειξη χωρο-χρονικών hot spot κυκλοφοριακής συμφόρησης σε επίπεδο ακμών γράφου. Για την ανάλυση έχει χρησιμοποιηθεί ο προσαρμοσμένος δείκτης Getis Ord Gi*, έτσι όπως ορίζεται στον τύπο (4.4).

$$Z\left(G_i^*\right) = \frac{\sum_{j=1}^{n} w_{ij}\, x_j \;-\; \bar{X}\, \sum_{j=1}^{n} w_{ij}^2}{s\sqrt{\dfrac{n\, \sum_{j=1}^{n} w_{ij}^2 \;-\; \left(\sum_{j=1}^{n} w_{ij}\right)^2}{n-1}}} \qquad \textbf{(4.4)}$$

Όπου:

G_i^* : η τιμή του στατιστικού δείκτη χωρικής αλληλεξάρτησης για την ακμή i

w_{ij} : η σχέση γειτνίασης μεταξύ των ακμών i και j

x_j : η τιμή της μεταβλητής για την ακμή j

n : ο συνολικός αριθμός των διακριτών ακμών

$\bar{X}$: η μέση τιμή της μεταβλητής x

$$\bar{X} = \frac{\sum_{j=1}^{n} x_i}{n} \quad \textbf{(4.5)}$$

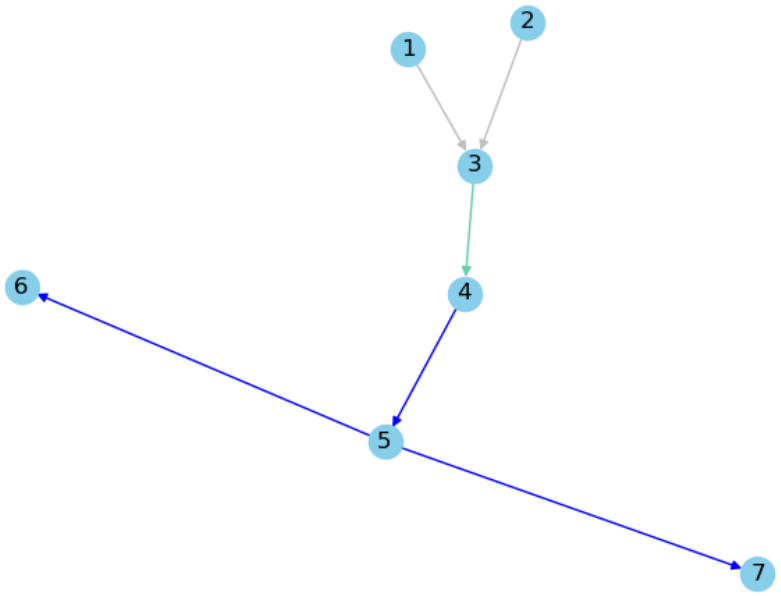
S : η τυπική απόκλιση της μεταβλητής x

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{X}^2} \quad \textbf{(4.6)}$$

Η παράμετρος w_{ij} εκφράζει το χωρικό διαχωρισμό μεταξύ γειτονικών οντοτήτων και συνήθως αποτελεί συνάρτηση της απόστασης d, $w_{ij}(d)$. Μέσω της συνάρτησης του βάρους, ενσωματώνεται στη χωροχρονική ανάλυση ο '1ος Νόμος της Γεωγραφίας' κατά Tobler (1970) : 'Everything is related to everything else, but near things are more related'.

Στην συγκεκριμένη εφαρμογή, η σχέση γειτνίασης w_{ij} αποτελεί συνάρτηση της απόστασης μιας ακμής e_i από τη γειτονική της e_j στο 3D χώρο. Στόχος είναι όσο μικρότερη είναι η απόσταση μεταξύ των γειτονικών ακμών, τόσο μεγαλύτερη να είναι η μεταξύ τους επιρροή.

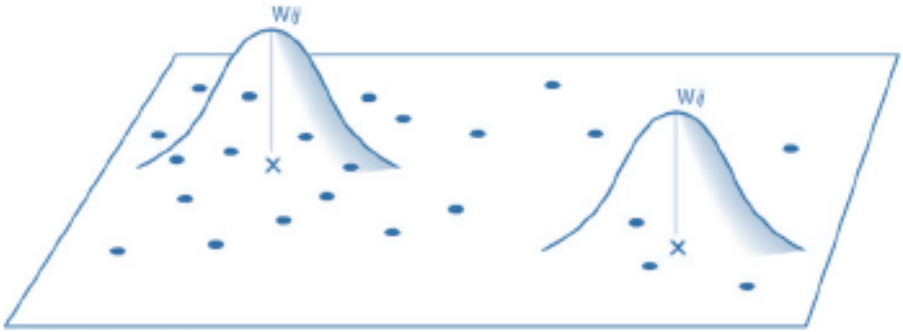
Αναφορικά με τη σχέση γειτνίασης, θεωρούμε ότι για να χαρακτηριστεί μια ακμή ως hotspot κυκλοφοριακής συμφόρησης προϋποθέτει ότι οι ακμές έπειτα από αυτήν έχουν υψηλές τιμές του δείκτη κυκλοφοριακής συμφόρησης. Με άλλα λόγια, θεωρούμε ότι το φαινόμενο της κυκλοφοριακής συμφόρησης διαδίδεται με κατεύθυνση προς τα πίσω στις ακμές ενός γράφου. Εξαιτίας αυτής της θεώρησης, το βάρος w_{ij} υπολογίζεται μόνο για τις ακμές e_j που διαδέχονται την ακμή e_i στον υπό εξέταση κατευθυνόμενο γράφο. Στις υπόλοιπες ακμές αποδίδεται βάρος μηδέν, με αποτέλεσμα να μην επιδρούν στον δείκτη Gi* και συνεπώς να μπορούν να παραλειφθούν από τους υπολογισμούς.



Εικόνα 10. Απεικόνιση γράφου για την άναδειξη της σχέσης γειτνίασης μεταξύ ακμών

Ειδικότερα στο παράδειγμα της Εικόνας 10, εξετάζεται η περίπτωση της ακμής e_i = [3,4], που συμβολίζεται με πράσινο χρώμα. Σύμφωνα με τα παραπάνω, για τον καθορισμό της τιμής $Gi^*_{[3,4]}$, το βάρος w_{ij} υπολογίζεται μόνο για τις ακμές e_j = [4,5], [5,6], [5,7] με μπλε χρώμα που διαδέχονται την ακμή e_i. Οι ακμές [1,3] και [2,3] που είναι προγενέστερες της εξεταζόμενης ακμής δεν συμμετέχουν στους υπολογισμός, καθώς τους αποδίδεται βάρος w_{ij} ίσο με μηδέν.

Η συνάρτηση που επιλέχθηκε για να εκφράσει τη σχέση εγγύτητας w_{ij} μεταξύ δυο ακμών e_i και e_j στη συγκεκριμένη διπλωματική εργασία είναι αυτή του Γκαουσσιανού Kernel (Gaussian Kernel). Η συνάρτηση του Gaussian Kernel βαραίνει τα δεδομένα με συνεχή τρόπο και σταδιακά μειώνει το βάρος, καθώς αυξάνεται η απόσταση από το κέντρο του kernel, χωρίς ωστόσο να παίρνει ποτέ τη μηδενική τιμή. Η τιμή του βάρους για μια ακμή e_i υπολογίζεται για το σύνολο των γειτονικών ακμών ή για το σύνολο των ακμών εντός μια προκαθορισμένης ακτίνας και εξαρτάται τόσο από την απόσταση d_{ij}, όσο και από την επιλογή της τιμής ενός προκαθορισμένου εύρους, h (bandwidth), που προκαλεί μείωση της επίδρασης με την απόσταση.



Εικόνα 11. Απεικόνιση του τρόπου που βαρύνονται τα δεδομένα με εφαρμογή του Gaussian Kernel

Η συνάρτηση του Γκαουσσιανού Kernel είναι η εξής:

$$W_{ij} = \exp\left(-\frac{d_{ij}^2}{h^2}\right) \quad \textbf{(4.7)}$$

Όπου:

d_{ij}: η χωροχρονική απόσταση μεταξύ δυο ακμών. Ο χωρικός διαχωρισμός ορίζεται με βάση το βαθμό διαχωρισμού (degrees of separation). Ο βαθμός διαχωρισμού δείχνει πόσα hops απαιτούνται για να τη σύνδεση της ακμής e_i με την e_j. Η απόσταση υπολογίζεται πάνω στο γράφο. Η χρονική απόσταση προκύπτει μέσω των χρονικών παραθύρων, δηλαδή της απεικόνισης του ίδιου γράφου σε διαφορετικά χρονικά διαστήματα ανάλυσης που ισαπέχουν μεταξύ τους και κατά συνέπεια οδηγούν σε διαφορετικές τιμές του δείκτη κυκλοφοριακής συμφόρησης ανά ακμή.
h : η σταθερή παράμετρος που εκφράζει το εύρος ζώνης, bandwidth.

Σύμφωνα με τους J.Shim et al. [17], θεωρώντας ότι τη χωρική $\left(d_{ij}^S\right)^2$ και τη χρονική $\left(d_{ij}^T\right)^2$ απόσταση μεταξύ δυο χωρικών ενοτήτων, η χρονοχωρική απόσταση $\left(d_{ij}^{ST}\right)^2$ μπορεί να εκφραστεί ως γραμμικός συνδυασμός αυτών των δυο, όπως ορίζεται στη σχέση που ακολουθεί.

$$\left(d_{ij}^{ST}\right)^2 = \mu^S \left(d_{ij}^S\right)^2 + \mu^T \left(d_{ij}^T\right)^2 \quad \textbf{(4.8)}$$

Όπου:

μ^S : συντελεστής κλίμακας χωρικής απόστασης
μ^T : συντελεστής κλίμακας χρονικής απόστασης

Συνεπώς η συνάρτηση του βάρους για τρεις διαστάσεις εκφράζεται ως εξής:

$$W_{ij} = \exp\left(-\left(\frac{\mu^S \left(d_{ij}^S\right)^2 + \mu^T \left(d_{ij}^T\right)^2}{h_{ST}^2}\right)\right) = \exp\left(-\left(\frac{\left(d_{ij}^S\right)^2}{h_S^2} + \frac{\left(d_{ij}^T\right)^2}{h_T^2}\right)\right) \quad \textbf{(4.9)}$$

$$= W_{ij}^S + W_{ij}^T \quad \textbf{(4.10)}$$

Όπου:

h_{ST}^2 : το χωροχρονικό εύρος ζώνης
$h_S^2 = h_{ST}^2 / \mu^S$: το χωρικό εύρος ζώνης
$h_T^2 = h_{ST}^2 / \mu^T$: το χρονικό εύρος ζώνης

Σκόπιμο είναι να σημειωθεί ότι η χρήση kernel στη χωρική ανάλυση είναι συνηθισμένη πρακτική για χωρικά σημειακά δεδομένα, ενώ χρησιμοποιείται κυρίως σε εφαρμογές εκτίμησης αξίας ακινήτων, μέσω μοντέλων γεωγραφικά σταθμισμένης παλινδρόμησης (Geographically Weighted Regression - GWR). Η συγκεκριμένη διπλωματική εργασία διαφοροποιείται από την υπάρχουσα βιβλιογραφία, καθώς σύμφωνα με την έρευνα μας είναι η πρώτη που χρησιμοποιεί Γκαουσσιανό Kernel για να ορίσει τη σχέση γειτνίασης μεταξύ των ακμών ενός κατευθυνόμενου γράφου, έχοντας ως στόχο την ανάδειξη hotspot κυκλοφοριακής συμφόρησης.

Στο παράδειγμα που ακολουθεί με κόκκινο χρώμα απεικονίζονται οι ακμές με χωρική απόσταση 1 (ένα) από την υπό εξέταση ακμή e_i και με πράσινο χρώμα οι ακμές με χωρική απόσταση 2 (δυο) από την εξεταζόμενη ακμή. Με βάση το χρονικό partition, υπολογίζεται η χρονική απόσταση, που αντιστοιχεί στα χρονικά βήματα που χωρίζουν δυο χωρικές ενότητες ακόμα και όταν αυτές βρίσκονται στην ίδια θέση στο χώρο.

Θεωρώντας πως η υπό εξέταση ακμή είναι η e_i στο χρονικό διάστημα Δt=0 και η e_j στο χρονικό διάστημα Δt=1, πρόκειται να υπολογιστεί το βάρος wij κάνοντας χρήση της σχέσης γειτνίασης που ορίστηκε παραπάνω.

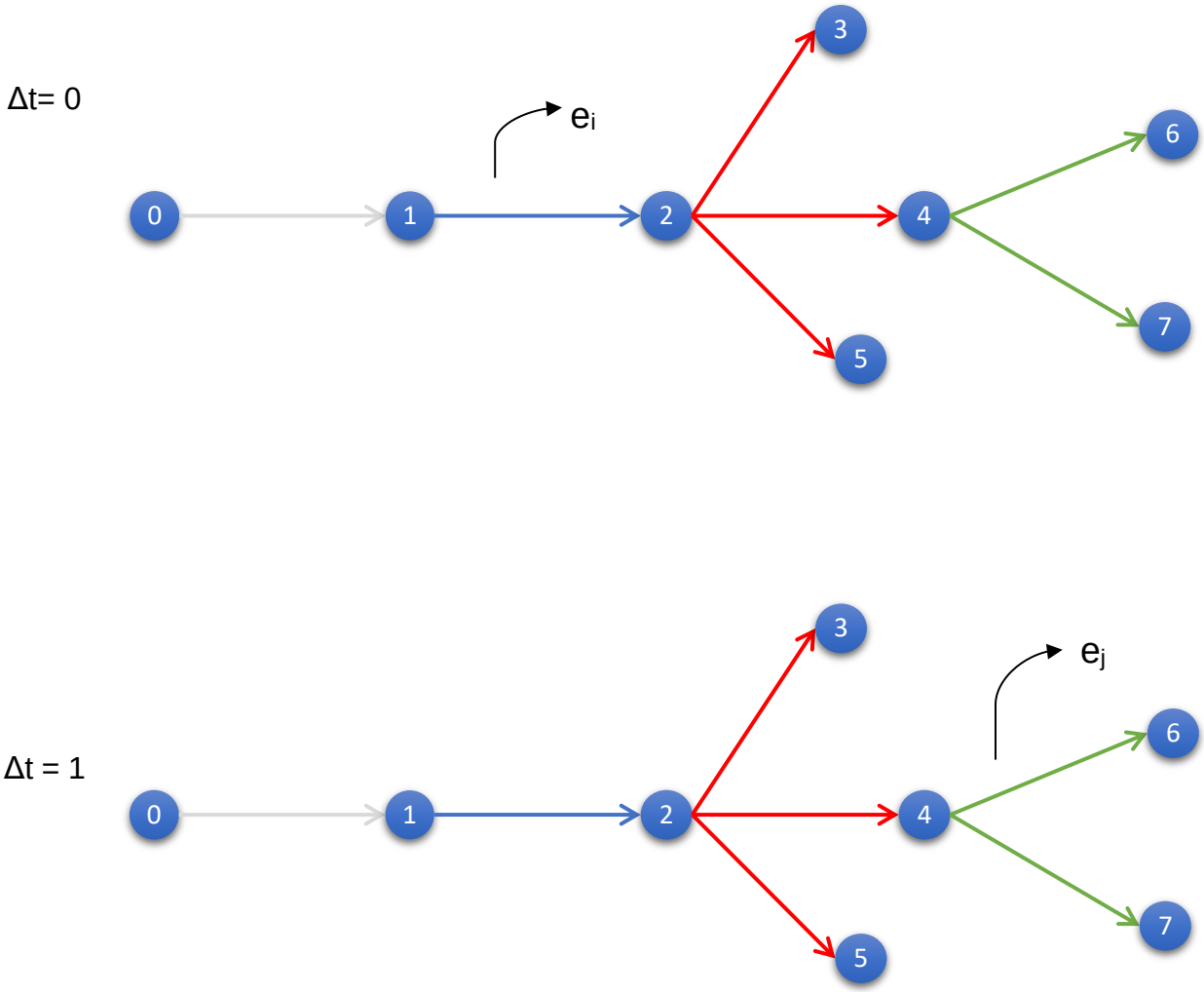
Σύμφωνα με τους τύπους (4.9) και (4.10) και θεωρώντας $h_S = h_T = 2$, το βάρος w_{ij} υπολογίζεται ως εξής:

$$W_{ij} = \exp\left(-\left(\frac{\left(d_{ij}^S\right)^2}{h_S^2} + \frac{\left(d_{ij}^T\right)^2}{h_T^2}\right)\right) = \exp\left(-\left(\frac{(2)^2}{2^2} + \frac{(1)^2}{2^2}\right)\right) = 0.287$$

Αναφορικά με την ερμηνεία του στατιστικού Gi*, ισχύουν τα εξής. Όταν η τιμή του Gi* τείνει στο μηδέν, συνεπάγεται τυχαία κατανομή των παρατηρούμενων χωρικών γεγονότων. Αντίθετα, θετικές και αρνητικές τιμές του δείκτη Gi*, αντιστοιχούν σε συγκεντρώσεις μεταβλητών με υψηλές τιμές και χαμηλές τιμές αντίστοιχα. Ειδικότερα, θετικές τιμές του δείκτη Gi* αντιστοιχούν σε ακμές με υψηλή κυκλοφοριακή συμφόρηση και αρνητικές τιμές Gi* σε ακμές με χαμηλή κυκλοφοριακή συμφόρηση. Συνοπτικά, αν η υπολογισμένη τιμή του δείκτη είναι μεγαλύτερη από ένα όριο που εκφράζει τη στατιστική σημαντικότητα, τότε η ακμή που αντιστοιχεί στην τιμή αυτή χαρακτηρίζεται ως hot spot ή cold spot αντίστοιχα.

Βασιζόμενοι στα παραπάνω, το πρόβλημα της ανάλυσης hot spot κυκλοφοριακής συμφόρησης σε δεδομένα τροχιών εκφράζεται ως εξής:

Δοσμένου ενός συνόλου δεδομένων D με δεδομένα τροχιάς GPS, όπως προκύπτουν από την κίνηση οχημάτων στο υφιστάμενο οδικό δίκτυο, και ενός γράφου G μέσω του οποίου αναπαρίσταται το οδικό δίκτυο, να βρεθούν οι k κορυφαίες στατιστικά σημαντικές ακμές $TOPK = \{e_1, \dots, e_k\} \in G$ για καθορισμένο επίπεδο σημαντικότητας, όπως προκύπτουν από τον δείκτη Gi*, έτσι ώστε: $G_i^* \geq G_j^*, \forall\, e_i \in TOPK, \forall\, e_j \in G - TOPK$.



Εικόνα 12. Αναπαράσταση 3D γράφου για την κατανόηση του υπολογισμού της σχέσης γειτνίασης

5 ΥΛΟΠΟΙΗΣΗ ΑΛΓΟΡΙΘΜΟΥ

Στο πλαίσιο της διπλωματικής εργασίας αναπτύχθηκε ένας αλγόριθμος για τον εντοπισμό χωρο-χρονικών hot spot κυκλοφοριακής συμφόρησης σε επίπεδο ακμών γράφου σε Apache Spark, επιτρέποντας την παράλληλη εκτέλεση του στους κόμβους ενός cluster.

5.1 ΣΥΝΟΨΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ

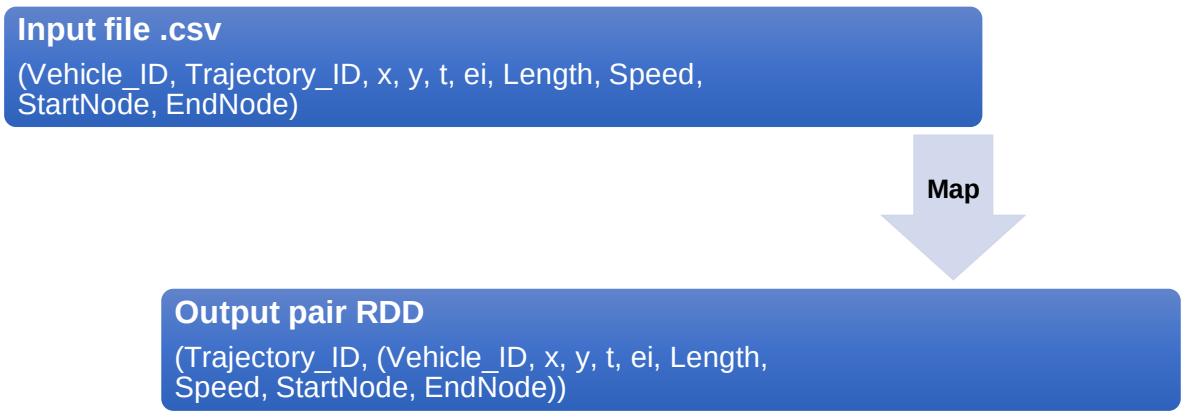
Η υλοποίηση του αλγορίθμου αποτελείται από τέσσερα βασικά βήματα.

- Στο πρώτο βήμα υπολογίζεται η τιμή του δείκτη κυκλοφοριακής συμφόρησης για το σύνολο των ακμών του γράφου για κάθε χώρο-χρονικό partition του οδικού δικτύου.
- Στο δεύτερο βήμα του αλγορίθμου, γίνεται η αναπαράσταση του οδικού δικτύου σε μορφή γράφου, καταγράφοντας με αυτό τον τρόπο τη χωρική σύνδεση των ακμών και κατά συνέπεια των οδικών τμημάτων που αναπαριστούν. Στη συνέχεια, η πληροφορία αυτή θα χρησιμοποιηθεί για τον καθορισμό της σχέσης γειτνίασης μεταξύ των χωρικών ενοτήτων κατά τον υπολογισμό του στατιστικού Gi*. Στο σημείο αυτό του κώδικα, υπολογίζεται η μέση τιμή και η τυπική απόκλιση των τιμών του δείκτη της κυκλοφοριακής συμφόρησης.
- Στο τρίτο βήμα του αλγορίθμου υπολογίζονται τα συστατικά μέλη της συνάρτησης του δείκτη Gi*, καθώς και ο ίδιος ο δείκτης Gi* .
- Τέλος, ο αλγόριθμος στο σύνολο του επιστρέφει μια λίστα με τις k πλέον στατιστικά σημαντικές ακμές σε ένα διάστημα εμπιστοσύνης, δηλαδή τις ακμές με το υψηλότερο z-score.

Διαγράμματα 1. Βασικά βήματα αλγορίθμου

5.2 ΥΠΟΛΟΓΙΣΜΟΣ ΔΕΙΚΤΗ ΚΥΚΛΟΦΟΡΙΑΚΗΣ ΣΥΜΦΟΡΗΣΗΣ

Έχοντας ολοκληρώσει την προεπεξεργασία των δεδομένων, σειρά έχει η εισαγωγή τους στο Spark και η μετατροπή τους σε μορφή pair RDD. Ως κλειδί (key) ορίζεται το πεδίου του Trajectory_ID και ως τιμή (value) οι υπόλοιποι παράμετροι.



Διαγράμματα 2. Μετατροπή των δεδομένων σε pair RDD

5.2.1 Ροή εργασίας στο Spark

Pseudocode 1: Workflow on Spark for calculating the attribute value	
1:	**Input:** rdd, temporal_parameters
2:	**Output:** attr_rdd
3:	**function**
4:	attr_rdd = rdd.groupByKey().mapValues(sort_grouped_values) \
5:	.flatMap(lambda group: calc_attribute_value(group, temp_par)) \
6:	.partitionBy(num_of_partitions, lambda k: hash_partitioner) \
7:	.mapValues(lambda x: (x,1)) \
8:	.reduceByKey(lambda x, y: (x[0] + y[0], x[1] + y[1])) \
9:	.mapValues(lambda x: x[0] / x[1])
10:	**end function**

Για την υλοποίηση του πρώτου βήματος του αλγορίθμου, δηλαδή τον υπολογισμό του δείκτη κυκλοφοριακής συμφόρησης μέσω της συνάρτησης *calc_attribute value*, τα δεδομένα που απαιτούνται είναι το *pair RDD* με τα δεδομένα εισόδου και οι *temporal parameters*, δηλαδή οι παράμετροι του χρονικού διαχωρισμού, όπως ορίζονται στη συνέχεια.

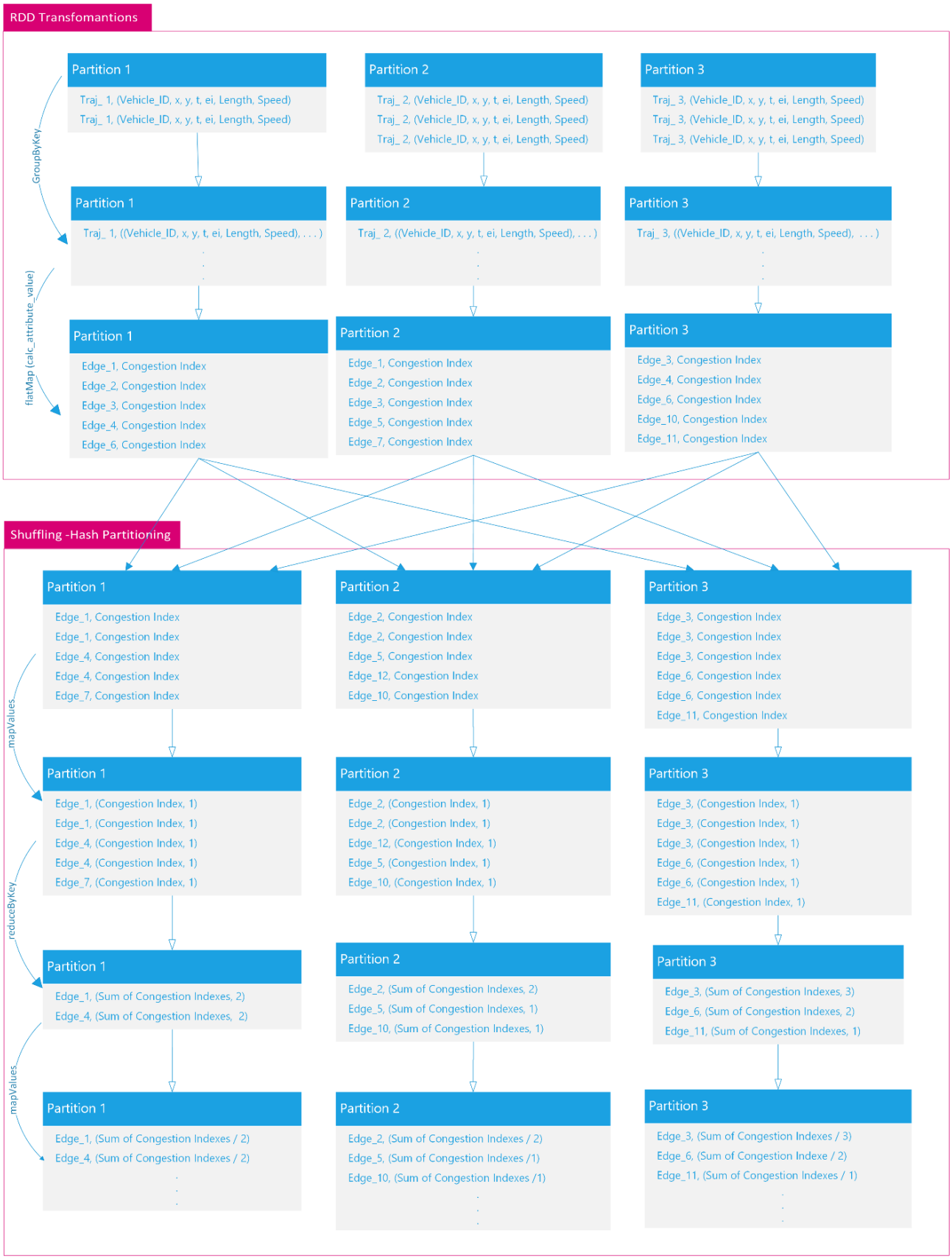
Σε πρώτη φάση, το *pair RDD* ομαδοποιείται ανά κλειδί, δηλαδή ανά *Trajectory_ID* και τα δεδομένα κάθε ομάδας ταξινομούνται σε αύξουσα σειρά σύμφωνα με την τιμή του timestamp, t. Έπειτα, τα ομαδοποιημένα δεδομένα γίνονται *flatMap* και σε κάθε μια από τις παρατηρήσεις της ομάδας εφαρμόζεται η συνάρτηση *calc_attribute_value*, η οποία υπολογίζει το δείκτη κυκλοφοριακής συμφόρησης σε επίπεδο ακμής. Η λειτουργία της περιγράφεται αναλυτικότερα στη συνέχεια.

Το *pair RDD* που προκύπτει από την εφαρμογή της συνάρτησης *calc_attribute_value* έχει ως κλειδί το *Edge_ID* και ως τιμή την υπολογισμένη τιμή του δείκτη κυκλοφοριακής συμφόρησης. Στο σημείο αυτό σκόπιμο είναι να σημειωθεί ότι το κλειδί *Edge_ID ορίζεται ως* (e_i , t_part_i) και είναι σύνθετο, καθώς περιλαμβάνει πληροφορία που αφορά το αναγνωριστικό της ακμής e_i

και το *temporal partition, t_part*$_i$, στο οποίο αναφέρεται ο υπολογισμένος δείκτης κυκλοφοριακής συμφόρησης.

Μια τροχιά αποτελεί ένα σύνολο διαδοχικών σημειακών δεδομένων. Μερικά από τα σημειακά δεδομένα ανήκουν σε διαφορετικές ακμές, διαδοχικές μεταξύ τους. Συνεπώς, εφαρμόζοντας τη συνάρτηση *calc_attribute_value,* κάθε γραμμή του pair RDD που προκύπτει αντιστοιχεί σε μια ακμή του γράφου που αναπαριστά το οδικό δίκτυο. Με τον μετασχηματισμό των δεδομένων το key-value pair διαφοροποιείται και οδηγούμαστε στο να έχουμε σε κάθε partition δεδομένα με διαφορετικό κλειδί, που σε επόμενο βήμα θα οδηγήσει σε μετακίνηση των δεδομένων μέσα στο δίκτυο *(shuffle)*. Για το λόγο αυτό σε αυτό το σημείο επιλέγεται να πραγματοποιηθεί hash partitioning των δεδομένο με βάση το πεδίου του κλειδιού *Edge_ID*.

Τέλος, τα δεδομένα συνοψίζονται και υπολογίζεται η μέση τιμή της κυκλοφοριακής συμφόρησης ανά ακμή. Αναλυτικότερα για τον υπολογισμό της μέσης τιμής, το πεδίο των values εμπλουτίζεται με την τιμή ένα για κάθε παρατήρηση, η πληροφορία συνοψίζεται εφαρμόζοντας *reduceByKey* και αθροίζοντας τις τιμές των values που ανήκουν στο ίδιο *Edge_ID*. Έχοντας στη διάθεση μας το άθροισμα των τιμών του δείκτη κυκλοφοριακής συμφόρησης και το πλήθος των παρατηρήσεων που αντιστοιχεί στο κάθε *Edge_ID*, είναι δυνατός ο υπολογισμός της μέσης τιμής της κυκλοφοριακής συμφόρησης για κάθε *Edge_ID*, διαιρώντας το άθροισμα των τιμών του δείκτη κυκλοφοριακής συμφόρησης με το πλήθος των παρατηρήσεων ανά *Edge_ID*.



Διαγράμματα 3. Μετασχηματισμοί των δεδομένων κατά τον υπολογισμό του δείκτη συμφόρησης

5.2.2 Συνάρτηση temporal_parameters

Pseudocode 2: temporal_parameters	
1:	**Input:** rdd, temp_part_duration
2:	**Output:** minTimestamp, maxTimestamp, t_step, n_steps
3:	**function**
4:	timestamp_RDD= rdd.mapValues(lambda x : x[Datetime]).values()
5:	minTimestamp = timestamp_RDD.reduce(lambda x, y: min(x, y))
6:	maxTimestamp = timestamp_RDD.reduce(lambda x, y: max(x, y))
7:	t_step = 3600* temp_part_duration
8:	n_steps = int((maxTimestamp-minTimestamp).total_seconds()/t_step)+1
9:	return ((minTimestamp, maxTimestamp, t_step, n_steps))
10:	**end function**

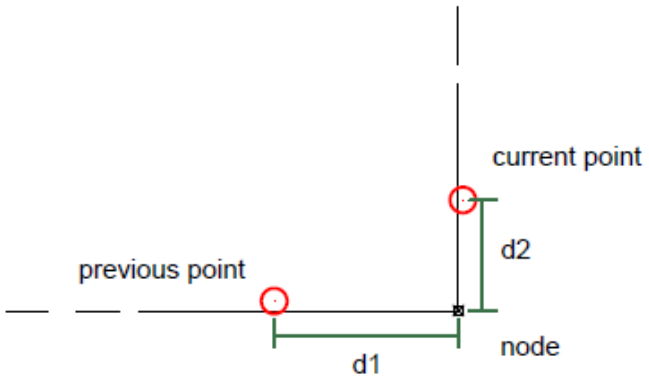
Ο υπολογισμός των παραμέτρων σύμφωνα με τις οποίες θα πραγματοποιηθεί ο χρονικός διαχωρισμός των δεδομένων (*temporal_parameters*) γίνεται μέσω της συνάρτησης *temporal_parameters*. Στη συνάρτηση αυτή εισάγεται το αρχικό *pair RDD* και η χρονική διάρκεια κάθε temporal partition (*temp_part_duration*), όπως ορίζεται από το χρήστη. Μέσω αυτής υπολογίζεται το πλήθος των temporal partitions, δηλαδή το πλήθος των κομματιών που πρόκειται να χωριστεί ο άξονας του χρόνου. Τέλος, η συνάρτηση επιστρέφει 4 παραμέτρους: το μικρότερο *Timestamp* (*minTimestamp*), το μέγιστο *Timestamp* (*maxTimestamp*), τη χρονική διάρκεια κάθε *temporal partition* σε δευτερόλεπτα (*t_step*) και το πλήθος των temporal partitions που προκύπτουν (*n_steps*). Η πληροφορία αυτή πρόκειται να χρησιμοποιηθεί στην συνέχεια για την αντιστοίχιση της πληροφορίας της κυκλοφορίας της κυκλοφοριακής συμφόρησης στο χρονικό παράθυρο που ανήκει.

5.2.3 Συνάρτηση calc_attribute_value

Η συνάρτηση *calc_attribute_value* παίρνει ως όρισμα το σύνολο των ομαδοποιημένων δεδομένων που αντιστοιχούν σε κάθε *Trajectory_ID* και τις χρονικές παραμέτρους, όπως υπολογίζονται από τη συνάρτηση

temporal_parameters, ενώ επιστρέφει ένα *key-value pair*, με κλειδί το *Edge_ID* και τιμή τον υπολογισμένο δείκτη κυκλοφοριακής συμφόρησης.

Ο υπολογισμός γίνεται ακολουθώντας τη μεθοδολογική προσέγγιση που περιεγράφηκε νωρίτερα στο Κεφάλαιο 4.3. Ο αλγόριθμος πραγματοποιεί σύγκριση των διαδοχικών παρατηρήσεων, ελέγχοντας αν έχουν ίδιο κωδικό e_i. Όταν εντοπίζει διαδοχικές παρατηρήσεις με διαφορετικό κωδικό e_i, πραγματοποιεί έναν τοπολογικό έλεγχο για να εντοπίσει αν οι ακμές είναι διαδοχικές, δηλαδή μοιράζονται έναν κοινό κόμβο. Εφόσον η συνθήκη αυτή ικανοποιείται, υπολογίζεται η απόσταση της κάθε παρατήρησης (d1, d2 στην Εικόνα 13) από τον κοινό κόμβο πάνω στο ελλειψοειδές μέσω της συνάρτησης *distance* της βιβλιοθήκης *geopy*. Στη συνέχεια, εφαρμόζοντας τον τύπο της παρεμβολής (4.1), υπολογίζεται η χρονική στιγμή που το κινούμενο όχημα διέρχεται από τον κοινό κόμβο και καταχωρείται στη μεταβλητή t_{node}.



Εικόνα 13. Απεικόνιση των παραμέτρων που απαιτούνται για εκτίμηση της χρονικής στιγμής εισόδου στην κορυφή

Pseudocode 3: calc_atttribute_value

```
1:  Input: traj_group, t_par
2:  Output: tuple (Edge ID, attr)
3:  function
4:    traj, group = traj_group
5:    prev_item = None
6:    edge_dict = dict()
7:    for item in group:
8:        if prev_item != None:
9:          curr_item = item
10:         if not on the same edge:
11              if prev_item[EndNode] == curr_item[StartNode]:
12:               dist_prev_to_node = geopy.distance.distance( prev_item, node)
13:               dist_node_to_curr = geopy.distance.distance( node, curr_item)
14:               node_time = linear interpolation for the calculation of timestamp at the node
15:               t_partition = int( (node_time - minTimestamp).total_seconds()/ t_step )
16:               curr_edge_id = (curr_item[StartNode], curr_item[EndNode])
17:               prev_edge_id = (prev_item[StartNode], prev_item[EndNode])
18:               edge_dict[curr_edge_id] = node_time
19:               if edge_dict.get(prev_edge_id) is not None:
20:                 dt = (node_time - edge_dict.get(prev_edge_id)).total_seconds()/3600
21:                 if dt>0 and prev_item[Road_Speed] > (prev_item[Length] / dt):
22:                   attr = 1 - ( (prev_item[Length] / dt) / prev_item[Road_Speed] )
23:                   yield (prev_edge_id,t_partition),attr
24:                 else:
25:                   attr = 0.0
26:                   yield (prev_edge_id,t_partition),attr
27:             elif prev_item[EndNode] == curr_item[EndNode]:
28:                 Similarly with the correct edge directionality in the calculations
29:             elif prev_item[StartNode] == curr_item[EndNode]:
30:                 Similarly with the correct edge directionality in the calculations
31:             elif prev_item[StartNode] == curr_item[StartNode]:
32:                 Similarly with the correct edge directionality in the calculations
33: end function
```

Το αναγνωριστικό των ακμών που βρίσκονται ανάντη και κατάντη του κοινού κόμβου αποθηκεύονται στις μεταβλητές *curr_edge_id* και *prev_edge_id* αντίστοιχα. Η μεταβλητή t_{node} καταχωρείται σε ένα λεξικό (*dictionary*) με κλειδί τον κωδικό της ακμής στην οποία εισέρχεται, *curr_edge_id*. Στην συνέχεια, αφού ελεγχθεί ότι υπάρχει καταχωρημένη τιμή στο λεξικό για το *prev_edge_id*, εκκινείται η διαδικασία υπολογισμού του δείκτη κυκλοφοριακής συμφόρησης.

Προς διευκόλυνση, αρχικά υπολογίζεται η μεταβλητή dt, που αντιστοιχεί στο χρονικό διάστημα που χρειάστηκε το όχημα για να διανύσει την υπό εξέταση ακμή. Σε κάποιες περιπτώσεις, η τιμή του dt είναι πολύ κοντά ή ίση με το μηδέν, με αποτέλεσμα ο υπολογισμός της ταχύτητας του οχήματος στην ακμή να καθίσταται αδύνατος (διαίρεση με το μηδέν). Για το λόγο αυτό, τίθεται ο περιορισμός η τιμή του dt να είναι μεγαλύτερη του μηδενός.

Μια ακόμα παράμετρος που πρέπει να ληφθεί υπόψη κατά τον υπολογισμό του δείκτη είναι η σχέση μεταξύ της υπολογισμένης ταχύτητας και του ορίου ταχύτητας. Σε κάποιες περιπτώσεις παρατηρείται το φαινόμενο της υπερβολικής ταχύτητας, με την υπολογιζόμενη ταχύτητα να ξεπερνά το όριο ταχύτητας. Σε αυτή την περίπτωση, καθώς και στην περίπτωση που η τιμή του dt ισούται με το μηδέν, θεωρούμε ότι η κίνηση του οχήματος είναι απρόσκοπτη και η τιμή του δείκτη κυκλοφοριακής συμφόρησης ίση με το μηδέν.

Τέλος κατά την εκτέλεση αυτής της συνάρτησης, οι τιμές του δείκτη κυκλοφοριακής συμφόρησης αντιστοιχίζονται στο temporal partition στο οποίο ανήκουν.

Η συνάρτηση αυτή επιστρέφει ένα *key-value pair* με κλειδί το *Edge_ID* και με τιμή τον υπολογισμένο δείκτη κυκλοφοριακής συμφόρησης. Ως *Edge_ID* από εδώ και στο εξής, ορίζεται το tuple που περιέχει τις εξής πληροφορίες: (*Edge_id, t_part*$_i$). Το *Edge_id* αντιστοιχεί και αυτό σε ένα tuple που ορίζει την ακμή στην οποία πραγματοποιείται η κίνηση και την κατεύθυνση αυτής. Για παράδειγμα, όταν (*Start Node, End Node*), η κίνηση να πραγματοποιείται από τον κόμβο *Start Node* στον *End Node*.

5.3 ΥΠΟΛΟΓΙΣΜΟΣ BROADCAST ΜΕΤΑΒΛΗΤΩΝ

Σε δεύτερο βήμα του αλγορίθμου πραγματοποιείται ο υπολογισμός των σταθερών τιμών του δείκτη Gi*, όπως και η μέση τιμή και η τυπική απόκλιση του δείκτη κυκλοφοριακής συμφόρηση, το συνολικό πλήθους ακμών που συμμετέχουν στην ανάλυση, καθώς και η δημιουργία του γράφου G που αναπαριστά το οδικό δίκτυο και καταγράφει τη χωρική σύνδεση των ακμών.

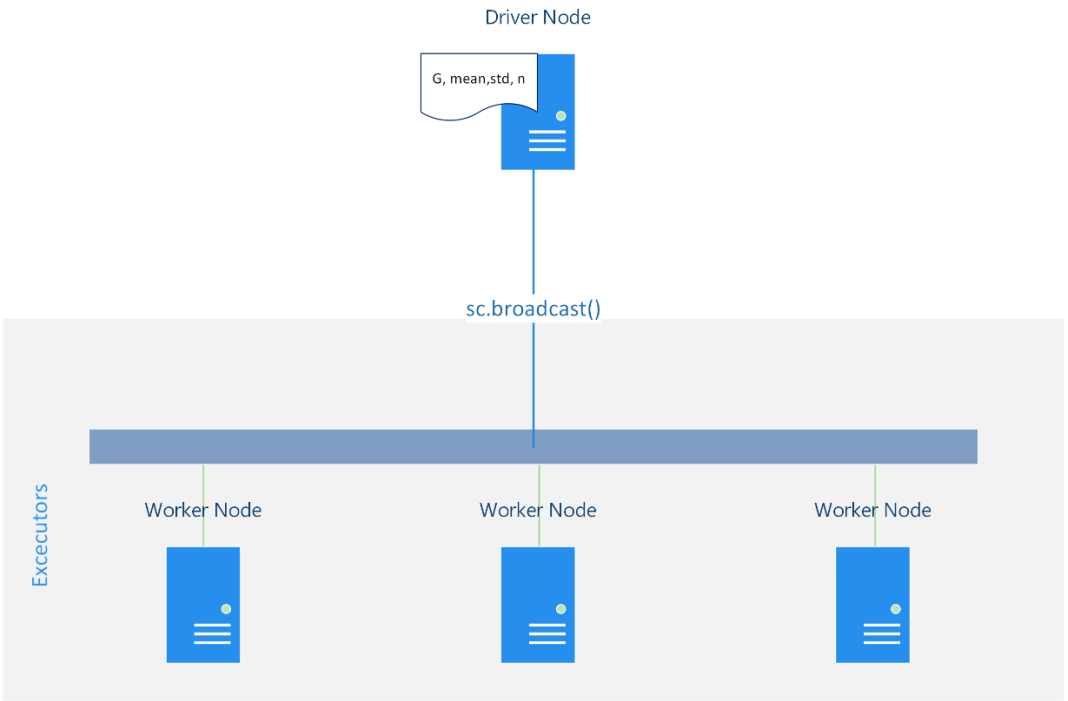
Pseudocode 4: broadcast_variables	
1:	**Input:** att_rdd, temporal_parameters
2:	**Output:** G, meanAttr, stdAttr, n_edges
3:	**function**
4:	meanAttr = att_rdd.values().mean()
5:	stdAttr = att_rdd.values().stdev()
6:	edges = att_rdd.keys().map(lambda x :x[0]).distinct()
7:	len_edges = edges.count()
8:	n_edges = len_edges * n_steps
9:	G = nx.DiGraph()
10:	G.add_edges_from(edges.collect())
11:	G = G.reverse()
12:	**end function**

Η συνάρτηση *broadcast_variables* δέχεται ως όρισμα τα δεδομένα εξόδου της συνάρτησης *calc_attribute_value*. Απομονώνει τα values του pair RDD που δέχεται ως όρισμα (att_rdd) και υπολογίζει τη μέση τιμή (meanAttr) και τη τυπική απόκλιση (stdAttr). Επίσης, απομονώνει από το κλειδί Edge_ID, το πρώτο τμήμα που αφορά τις κορυφές που ορίζουν την ακμή, e_i, και αναθέτει στην μεταβλητή edges το σύνολο των μοναδικών αναγνωριστικών. Υπολογίζει το πλήθος αυτών (len_edges) στη διάσταση του χώρου και στη συνέχεια το πλήθος τους στο σύνολο των temporal partitions (n_edges), ανάγοντας το στη διάσταση του χωρο-χρόνου.

Στη συνέχεια, μέσω της βιβλιοθήκης NetworkX, παράγεται ο κατευθυνόμενος γράφος G, κάνοντας χρήση της πληροφορίας της λίστας edges. Ο γράφος G αντιστρέφεται προς διευκόλυνση των υπολογισμών στη συνέχεια. Τέλος, η

συνάρτηση αυτή εκτελείται και τα δεδομένα εξόδου της γίνεται broadcast στους κόμβους του cluster.

Ειδικότερα, οι broadcast μεταβλητές είναι read-only μεταβλητές, οι οποίες γίνονται τοπικά διαθέσιμες στο driver και στη συνέχεια διανέμονται και αποθηκεύονται στη μνήμη όλων των nodes του cluster. Με αυτό τον τρόπο, η πληροφορία είναι διαθέσιμη για την εκτέλεση των tasks με παράλληλο τρόπο, χωρίς να απαιτείται η επανάληψη της αποστολής των δεδομένων στους nodes κάθε φορά που χρειάζεται μια αναζήτηση σε αυτές για την εκτέλεση ενός task, που θα οδηγήσει στον μετασχηματισμό ενός RDD. Αποφεύγοντας τα περιττά shuffle των δεδομένων, ειδικότερα όταν αυτά είναι μεγάλα σε μέγεθος και απαιτούνται συνεχείς αναζητήσεις σε αυτά, όπως ισχύει στην περίπτωση της μεταβλητής του γράφου G, οδηγούμαστε σε καλύτερη απόδοση του αλγορίθμου αναφορικά με το επικοινωνιακό κόστος.



Διαγράμματα 4. Δημιουργία των broadcast μεταβλητών τοπικά στον driver και διανομή τους στους worker nodes ώστε να γίνουν cached στη μνήμη

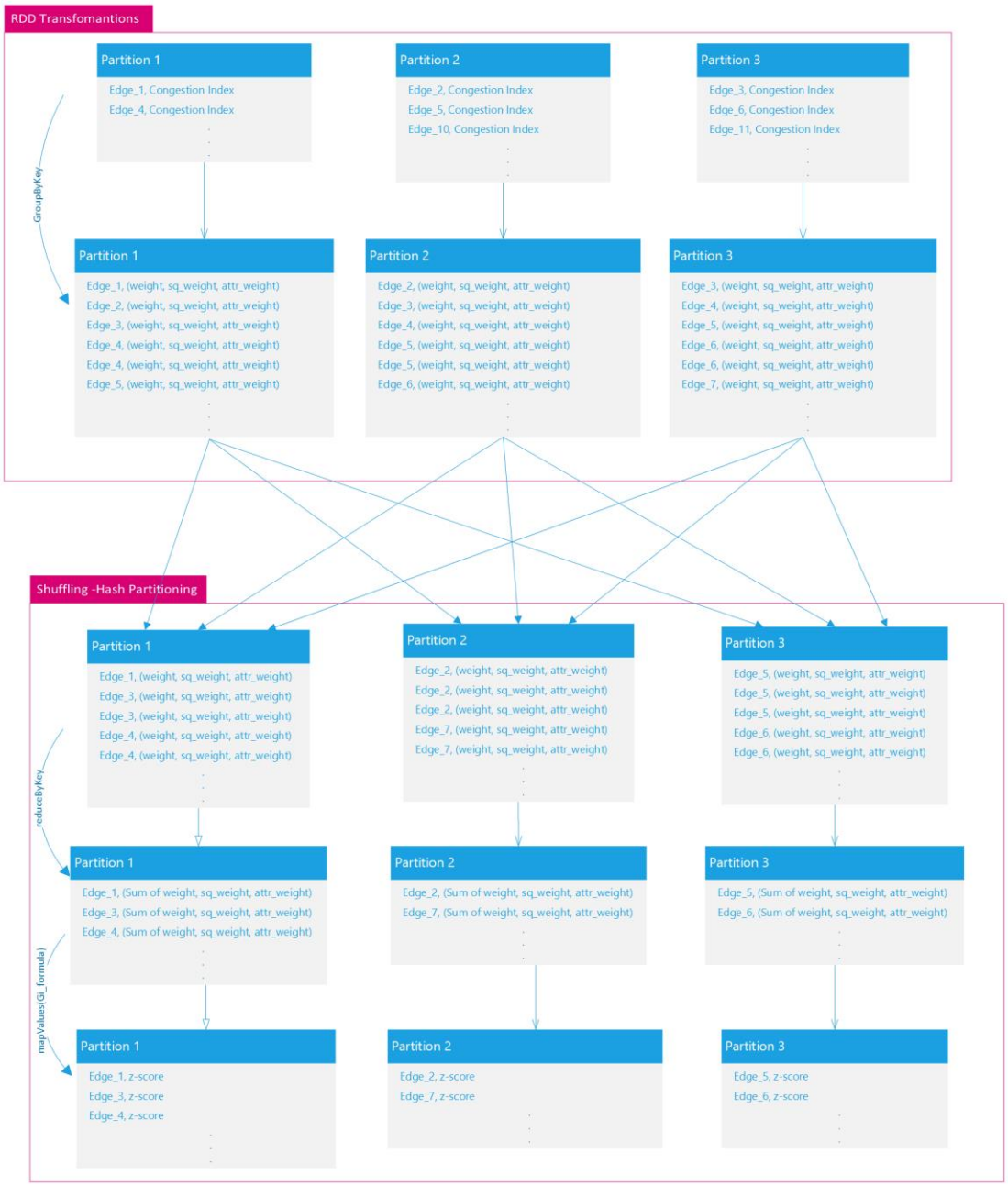
5.4 ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΔΕΙΚΤΗ Gi*

5.4.1 Ροή εργασιών για τον υπολογισμό του δείκτη Gi*

Σε αυτό το βήμα του αλγορίθμου γίνεται ο υπολογισμός του z-score για κάθε ακμή του γράφου G, με χρήση του στατιστικού Getis-Ord Gi*. Τα δεδομένα που απαιτούνται για την υλοποίηση της ροής εργασιών είναι οι χρονικές παράμετροι, οι broadcasted μεταβλητές, που αφορούν το γράφο, και το pair RDD (att_rdd) που προέκυψε από την ολοκλήρωση της πρώτης φάσης επεξεργασίας και περιέχει για κάθε Edge_ID την πληροφορία του δείκτη συμφόρησης.

Αρχικά, το pair RDD γίνεται flatMap και η συνάρτηση *GetisOrdCalculations* εφαρμόζεται σε κάθε ζευγάρι κλειδιού-τιμής. Τα δεδομένα εξόδου της συνάρτησης είναι key-value pairs που έχουν ως κλειδί το Edge_ID και ως τιμή ένα 3-tuple που ορίζεται ως εξής (weight, sq_weight, attr_weight) και αποτελείται από την τιμή του βάρους, το τετράγωνο της τιμής του βάρους και το γινόμενο του βάρους με την τιμή του δείκτη κυκλοφοριακής συμφόρησης.

Με το μετασχηματισμό των δεδομένων, για κάθε γραμμή του pair RDD που εισάγεται στη συνάρτηση *GetisOrdCalculations* επιστρέφεται ένα 3-tuple της μορφής (weight, sq_weight, attr_weight) για τις γειτονικές ακμές Edge_ID στη διάσταση του χώρου και του χρόνου (3D). Αυτό έχει ως αποτέλεσμα τα δεδομένα με το ίδιο Edge_ID να μην εντοπίζονται στο ίδιο partition (στο ίδιο partition ανήκουν πλέον δεδομένα που έχουν χωρική εγγύτητα σε οδικό δίκτυο με τα αρχικά). Σε επόμενο στάδιο της επεξεργασίας απαιτείται η ομαδοποίηση των δεδομένων ανά Edge_ID, οπότε σε αυτό το σημείο είναι απαραίτητο το hash partitioning των δεδομένων.

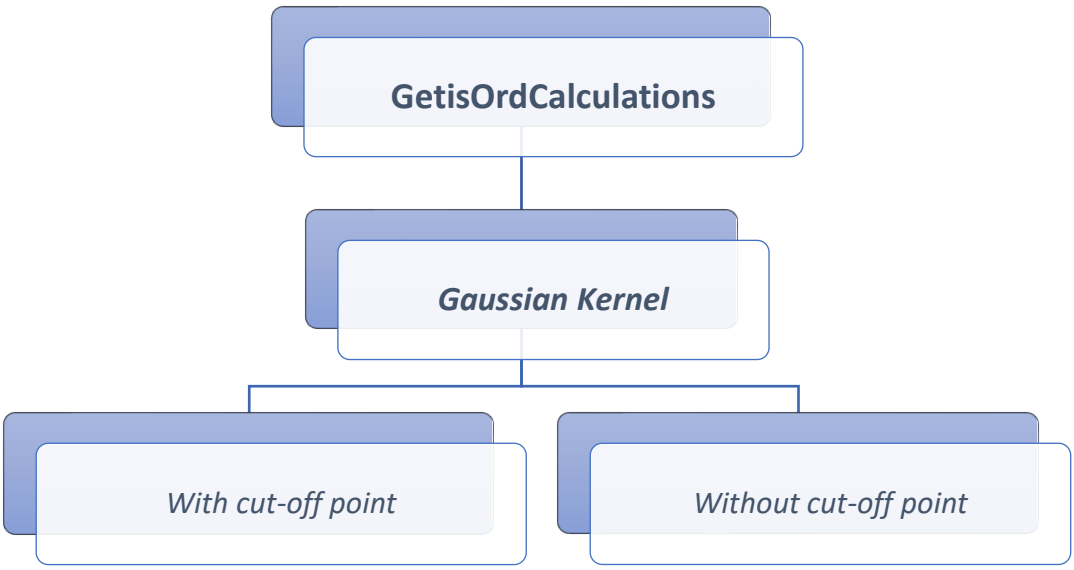


Διαγράμματα 5. Μετασχηματισμοί των δεδομένων κατά τον υπολογισμό του Gi z-score*

Έπειτα από το hash partitioning, το *pair RDD* γίνεται *reduceByKey* και υπολογίζονται τα αθροίσματα για τις τιμές των values με ίδιο *Edge_ID*. Τέλος, σε κάθε γραμμή του *pair RDD*, που προέκυψε από το προηγούμενο βήμα, εφαρμόζεται η φόρμουλα του Getis-Ord Gi*, σύμφωνα με τη σχέση (4.4). Το αποτέλεσμα των παραπάνω διεργασιών είναι ένα *pair RDD* που περιλαμβάνει την πληροφορία του *z-score* για κάθε ακμή του γράφου.

Pseudocode 5: Spark operations for the implementation of Gi* Formula	
1:	**Input:** att_rdd, temporal_parameters, broadcast_variables
2:	**Output:** pair RDD with key the Edge_ID and value the z-scores
3:	**function**
4:	calc_Gi = traj_rdd \
5:	.flatMap(lambda x : GetisOrdCalculations(x, temp_par, graph_parameters)) \
6.	.partitionBy(num_of_partitions, lambda k : : hash_partitioner) \
7:	.reduceByKey(lambda x, y: (x[0] + y[0], x[1] + y[1], x[2] + y[2])) \
8:	.mapValues(lambda x: Gi_formula(x, graph_parameters))
9:	**end function**

Στο πλαίσιο της διπλωματικής εργασίας αναπτύχθηκαν δυο διαφορετικές υλοποιήσεις της συνάρτησης υπολογισμού των απαραίτητων παραμέτρων για τον υπολογισμό του στατιστικού Getis-Ord Gi*, *GetisOrdCalculations*. Στην πρώτη υλοποίηση το σύνολο της πληροφορίας των γειτονικών παρατηρήσεων τόσο στη διάσταση του χώρου, όσο και στη διάσταση του χρόνου λαμβάνονται υπόψιν στο υπολογισμού του Gi* z-score μιας ακμής. Στη δεύτερη υλοποίηση, στους υπολογισμούς συμμετέχουν οι παρατηρήσεις που βρίσκονται εντός μιας καθορισμένης από τον χρήστη απόστασης. Η δεύτερη υλοποίηση αποτελεί προσπάθεια βελτιστοποίησης της πρώτης ως προς το χρόνο εκτέλεσης του αλγορίθμου, εφαρμόζοντας σημείο αποκοπής στους υπολογισμούς.



Διαγράμματα 6. Διαφορετικές υλοποιήσεις της συνάρτησης GetisOrdCalculations

5.4.2 Συνάρτηση GetisOrdCalculations χωρίς εφαρμογή σημείου αποκοπής

Όπως έχει αναφερθεί, οι παράμετροι του δείκτη Gi* υπολογίζονται μέσω της συνάρτησης *GetisOrdCalculations*. Η συνάρτηση δέχεται ως όρισμα μια γραμμή του att_rdd, τις χρονικές παραμέτρους, τις broadcast μεταβλητές του γράφου και τη μεταβλητή του εύρους ζώνης (bandwidth) του Gaussian Kernel.

Πρώτο βήμα της συνάρτησης είναι η ανάθεση της πληροφορίας του γράφου και της υπό εξέταση γραμμής του pair RDD σε μεταβλητές. Παράλληλα, ορίζονται οι μεταβλητές t_max και t_min που αφορούν το πρώτο και το τελευταίο temporal partition που θα ληφθεί υπόψη κατά τον υπολογισμό του Getis-Ord Gi*. Στο συγκεκριμένο αλγόριθμο, λαμβάνεται υπόψη το σύνολο της χρονικής και χωρικής πληροφορίας, ειδικότερα το σύνολο των temporal partitions και το σύνολο των ακμών που διαδέχονται την κάθε ακμή. Έπειτα, μέσω της μεθόδου ego_graph της βιβλιοθήκης NetworkX, παράγεται ένας υπογράφος, nn, που έχει ως αφετηρία την κορυφή u_i της υπό εξέταση ακμής και περιλαμβάνει το σύνολο των προγενέστερων ακμών.

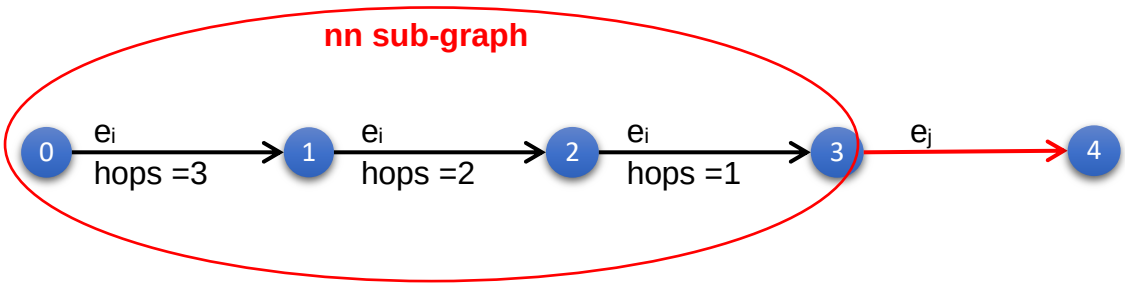
Pseudocode 6: GetisOrdCalculations

```
1:  Input: att_rdd line, temporal_parameters, broadcast_graph parameters, bandwidth
2:  Output: Tuple (Edge_ID, Tuple (weight, sq_weight, attr_weight))
3:  function
4:      G, mean_attr, std_attr, n_edges = graph_par
5:      key,attr = att_rdd line
6:      edge_nodes, t_part = key
7:      ui,vi = edge_nodes
8:      t_min = 0
9:      t_max = maxTimestamp
10:     nn = nx.ego_graph(G, ui)
11:     n = nn.number_of_edges()
12:     if n>=1 :
13:         for (uj,vj) in nn.edges:
14:             hopes = nx.shortest_path_length(nn, ui, vj)
15:             for t in range(t_min, t_max):
16:                 weight = math.exp( - (( hopes**(2) / h**(2)) + ((t - t_part)**(2) / h**(2))))
17:                 sq_weight = weight**(2)
18:                 attr_weight = attr * weight
19:                 yield ( (ui,vi),t ), (weight, sq_weight,attr_weight)
20:     for t in range(t_min, t_max):
21:         weight = math.exp(-((t - t_part)**(2) / h**(2)))
22:         sq_weight = weight**(2)
23:         attr_weight = attr * weight
24:         yield ( (ui,vi),t ), (weight, sq_weight,attr_weight)
```

Στο σημείο αυτό, σκόπιμο είναι να σημειωθεί ότι για τον υπολογισμό του δείκτη Gi* για μια ακμή e_i απαιτείται η γνώση του δείκτη της κυκλοφοριακής συμφόρησης για το σύνολο των ακμών e_j που τη διαδέχονται. Πρακτικά, όμως διαβάζοντας κάθε γραμμή του pair RDD, διαθέσιμη είναι η πληροφορία της κυκλοφοριακής συμφόρησης για μια ακμή, που ουσιαστικά αντιστοιχεί σε ένα από τα e_j. Γνωρίζοντας ότι η πληροφορία του δείκτη συμφόρησης απαιτείται για τον υπολογισμό του δείκτη Gi* για το σύνολο των προγενέστερων ακμών της ακμής e_j, η διαδικασία υπολογισμού των επιμέρους παραμέτρων του

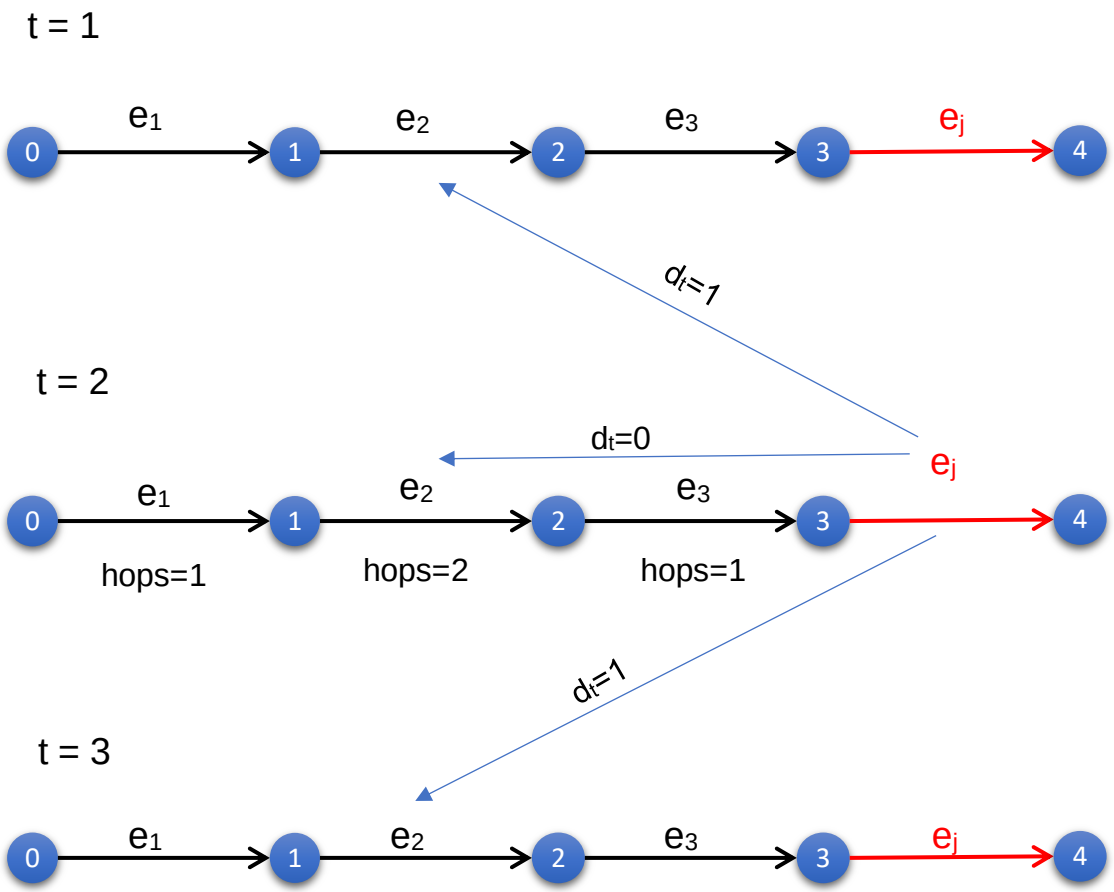
στατιστικού Getis-Ord γίνεται αντίστροφα, διαδίδοντας τη διαθέσιμη πληροφορία προς τα πίσω. Ειδικότερα κάνοντας χρήση του αντίστροφου γράφου G, τον οποίο έχουμε στη διάθεση μας από τη συνάρτηση broadcast_variables, είναι δυνατός ο προσδιορισμός των ακμών που είναι προγενέστερες της υπό εξέταση ακμής e_j και ο καθορισμός της σχέσης γειτνίασης τους με την ακμή e_j. Ο υπολογισμός των hops είναι ανεξάρτητος της κατεύθυνσης με αποτέλεσμα να είναι δυνατός ο υπολογισμός της χωρικής απόστασης και κατά συνέπεια του βάρους w_{ij}.

Τέλος, η πληροφορία του 3-tuple (weight, sq_weight, attr_weight) δεν αποδίδεται στην εξεταζόμενη ακμή, αλλά στις ακμές υπογράφου nn που επιστράφηκαν από τη μέθοδο ego_graph και για τις οποίες υπολογίστηκε η παράμετρος που χρονικού και χωρικού διαχωρισμού.



Διαγράμματα 7. Απεικόνιση της σχέσης γειτνίασης της εξεταζόμενης ακμής και των ακμών του υπογράφου nn

Έχοντας διαθέσιμο τον υπογράφο nn με τις γειτονικές ακμές της e_j (γραμμή 10 του ψευδοκώδικα) και εφόσον το πλήθος των ακμών του υπογράφου nn είναι μεγαλύτερο ή ίσο της μονάδας, για καθεμία από αυτές υπολογίζεται η χωρική απόσταση τους από την ακμή e_j, με βάση τον χωρικό διαχωρισμό τους, hopes. Παράλληλα με την χωρική απόσταση μεταξύ των δυο εκάστοτε ακμών, υπολογίζεται και η χρονική απόσταση τους, πλήθος temporal partitions που μεσολαβούν μεταξύ τους, καθιστώντας εφικτό τον υπολογισμό του βάρους σύμφωνα με τον τύπο (4.9) του *Gaussian Kernel* για κάθε ένα από τα temporal partitions.



Διαγράμματα 8. Παράδειγμα λογικής για την υλοποίηση της συνάρτησης GetisOrdCalculations

Στη συνέχεια πρόκειται να γίνει παρουσίαση ενός παραδείγματος της συνάρτησης *GetisOrdCalculations*. Θεωρούμε ότι το υπό εξέταση *Edge_ID* που εισέρχεται στην συνάρτηση είναι το (e_j, 2), δηλαδή η ακμή e_j που ανήκει στο temporal partition με κωδικό 2. Η συνάρτηση αναθέτει την πληροφορία του εξεταζόμενου Edge_ID σε μεταβλητές. Για τις ανάγκες του παραδείγματος θεωρούμε ότι ο άξονας του χρόνου έχει χωριστεί σε 3 διαστήματα (temporal partitions) και ότι το οδικό δίκτυο αποτελείται από 4 ακμές. Με χρήση της μεθόδου ego_graph, υπολογίζεται ο υπογράφος nn που περιέχει την πληροφορία των προγενέστερων ακμών της ακμής e_j και ειδικότερα τις ακμές, e_1, e_2, e_3. Η χωρική απόσταση της κάθε ακμής του υπογράφου nn είναι από την ακμή e_j είναι ίση με 3,2 και 1 hopes αντίστοιχα. Τα hopes μεταξύ αυτών των ακμών είναι ίδιο ανεξάρτητα του temporal partition που εξετάζεται.

Η επίδραση του δείκτη κυκλοφοριακής συμφόρησης της ακμής (ej,2) στις γειτονικές της ακμές, έστω ακμή e_2, δεν είναι η ίδια για τα διαφορετικά temporal partitions. Η επίδραση είναι συνάρτηση και του χρονικού διαχωρισμού των δυο ακμών. Για τον λόγο αυτό, η συνάρτηση λαμβάνει υπόψη και τον άξονα του χρόνου και υπολογίζεται την χρονική απόσταση μεταξύ των ακμών. Ο χρονικός διαχωρισμός μεταξύ της ακμής (e_j,2) και της ακμής e2 για διαφορετικά χρονικά partitions είναι ίσος με τη μονάδα για το 1 και 3 temporal partition και μηδέν για το 2 temporal partition.

Με βάση την πληροφορία της χρονικής και χωρικής απόστασης που είναι πλέον γνωστή, καθίσταται δυνατός ο υπολογισμός τους χωρο-χρονικού βάρους σύμφωνα με τον τύπο του Gaussian Kernel, σχέση (4.9).

Τελικό βήμα της συνάρτησης, είναι ο υπολογισμός των παραμέτρων που αντιστοιχούν στην υπό εξέταση ακμή για το σύνολο των temporal partitions. Στον υπολογισμό του βάρους αγνοείται το μέλος της συνάρτησης του βάρους που αφορά τη χωρική απόσταση.

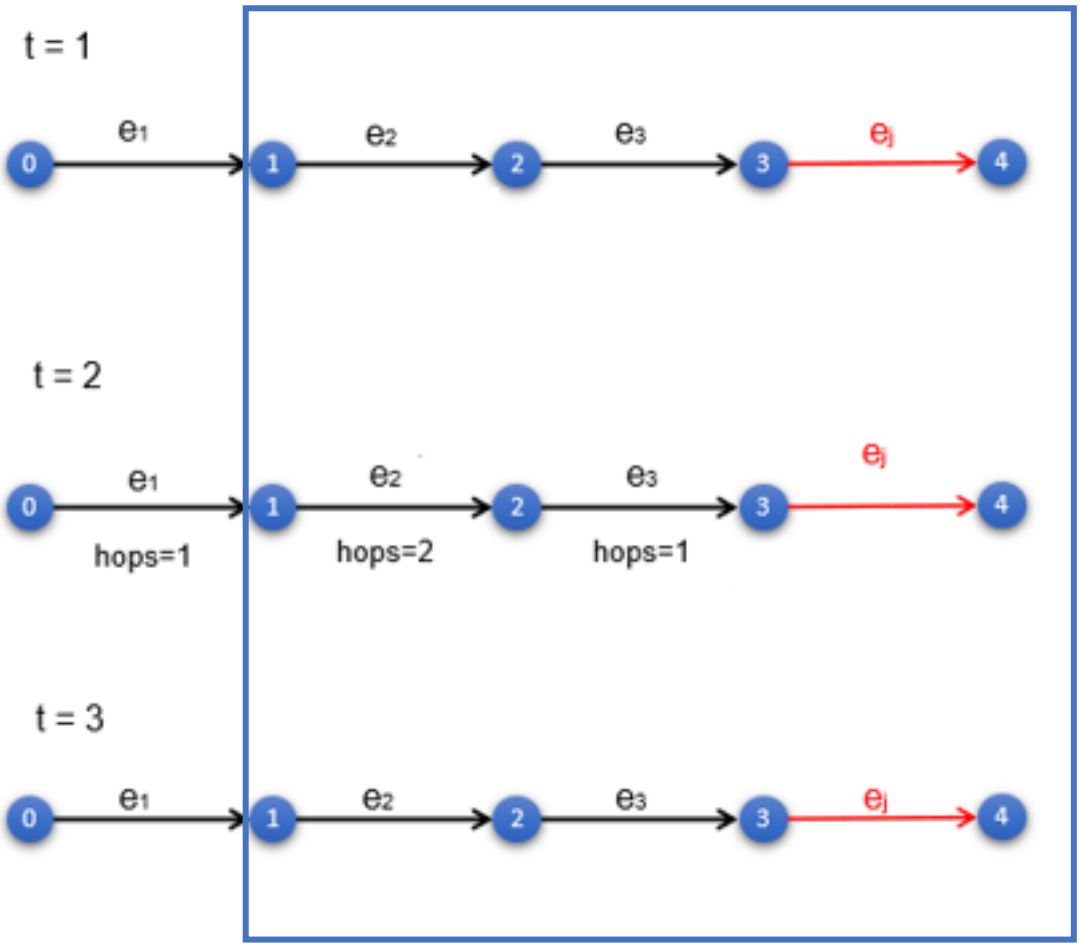
5.4.3 Εφαρμογή χωρικού και χρονικού cut-off στη συνάρτηση *GetisOrdCalculations*

Στην προηγούμενη υλοποίηση της συνάρτησης *GetisOrdCalculations*, πραγματοποιήθηκαν υπολογισμοί για το σύνολο της χωρικής και χρονικής πληροφορίας. Η πραγματοποίηση υπολογισμών για το σύνολο των γειτονικών ακμών τόσο χωρικά όσο και χρονικά, είναι υπολογιστικά ακριβή διαδικασία. Το υπολογιστικό κόστος αυξάνεται ακόμα περισσότερο σε γράφους μεγαλύτερου μεγέθους και για μικρότερο βήμα χρονικής ανάλυσης.

Pseudocode 8: GetisOrdCalculations with cut-off point	
1:	**Input:** att_rdd line, temporal_parameters, broadcast_graph parameters, n_neigh
2:	**Output:** Tuple (Edge_ID, Tuple (weight, sq_weight, attr_weight))
3:	**function**
4:	key,attr = item
5:	edge_nodes, t_part = key
6:	ui,vi = edge_nodes
7:	if (t_part - n_neigh) <= 0:
8:	t_min = 0
9:	else:
10:	t_min = t_part - n_neigh
11:	if (t_part + n_neigh) >= (temp_par[3]-1):
12:	t_max = temp_par[3]
13:	else:
14:	t_max = t_part + n_neigh + 1
15:	nn = nx.ego_graph(G, ui, radius = n_neigh)

Για τη μείωση της πολυπλοκότητας του αλγορίθμου, καθώς και του υπολογιστικού κόστους της διαδικασίας, πραγματοποιείται μια δεύτερη υλοποίηση της συνάρτησης *GetisOrdCalculations* που διαθέτει ένα σημείο αποκοπής για τους υπολογισμούς. Ειδικότερα, ο αλγόριθμος αυτός υπολογίζει το στατιστικό Gi* λαμβάνοντας υπόψη γείτονες που βρίσκονται χρονικά και χωρικά σε μέγιστη απόσταση n_neigh από την εξεταζόμενη ακμή.

Η συνάρτηση *GetisOrdCalculations* διαφοροποιείται στο πρώτο της τμήμα, στο οποίο πραγματοποιείται ο ορισμός του άνω και κάτω ορίου της χρονικής ανάλυσης μέσω των παραμέτρων t_min και t_max, καθώς και του ορισμού της του υπογράφου nn. Ειδικότερα, οι παράμετροι t_min και t_max ορίζονται έτσι ώστε να απέχουν διάστημα r από την εξεταζόμενη ακμή στον άξονα του χρόνου. Αναφορικά με τον υπογράφο nn, γίνεται παραμετροποίηση της μεθόδου ego_graph, ώστε να επιστρέφει της γειτονικές ακμές της εξεταζόμενης ακμής εντός ακτίνας n_neigh. Η τροποποίηση αυτή στη συνάρτηση μπορεί να ενσωματωθεί στην ανάλυση hotspot, ανεξαρτήτως του ορισμού του βάρους στο στατιστικό Gi*.



Διαγράμματα 9. Πλήθος ακμών που θα συμπεριληφθούν στην ανάλυση για n_neigh=2

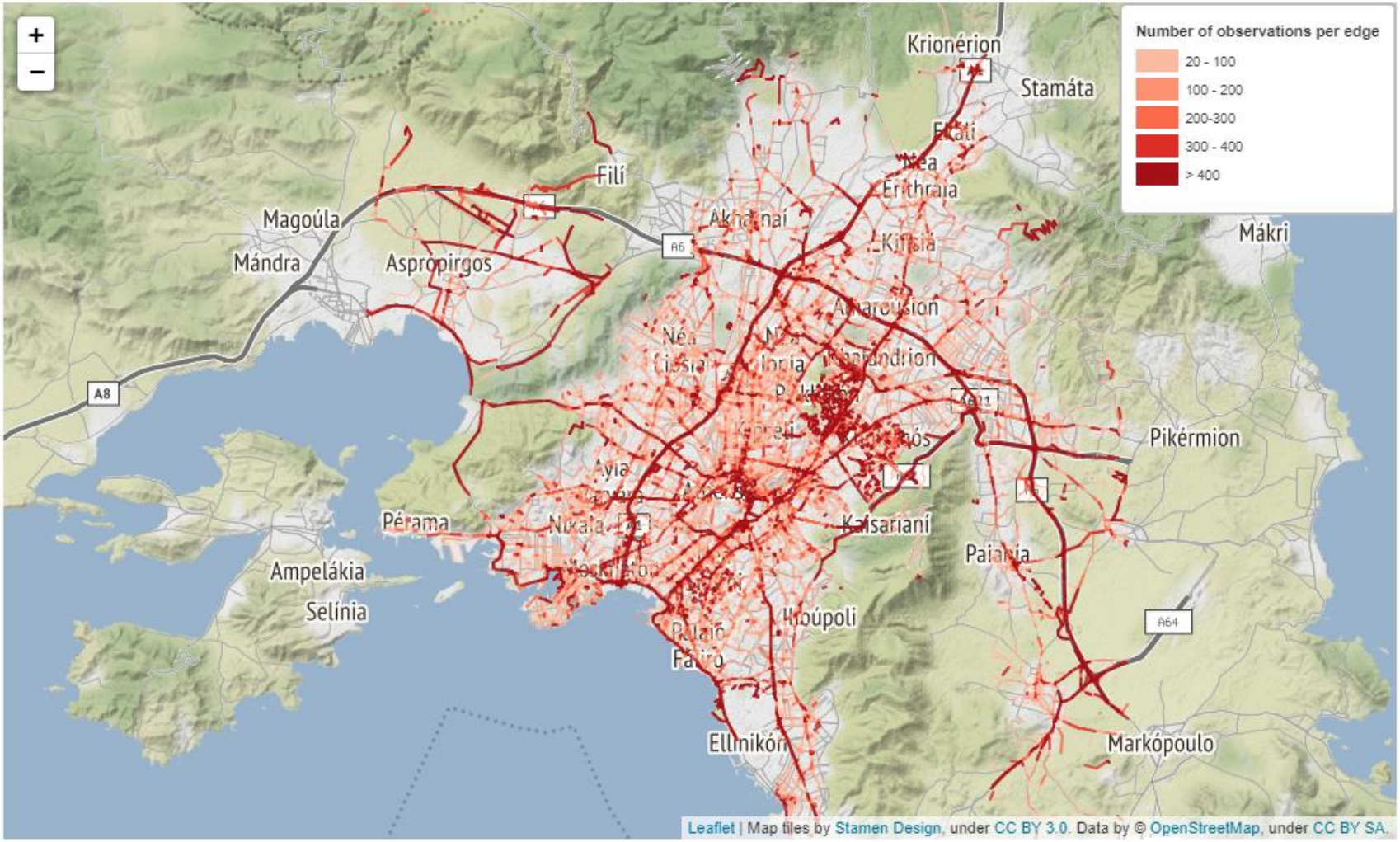
6 ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ

Σε αυτό του κεφάλαιο, παρουσιάζεται η μεθοδολογία αξιολόγησης των αποτελεσμάτων για τις διαφορετικές εκδοχές αλγορίθμων που εφαρμόστηκαν στο πλαίσιο της διπλωματικής εργασίας και τα αποτελέσματα που προέκυψαν. Σε αυτό το σημείο είναι σκόπιμο να σημειωθεί ότι η υλοποίηση των αλγορίθμων έγινε σε PySpark, με χρήση Apache Spark 3.1.1 Core API.

6.1 ΡΥΘΜΙΣΗ ΠΑΡΑΜΕΤΡΩΝ ΠΕΙΡΑΜΑΤΟΣ

Πλατφόρμα. Το deploy του κώδικα πραγματοποιήθηκε μέσω του Google Cloud Platform. Η πλατφόρμα δίνει την δυνατότητα επεξεργασίας μεγάλων δεδομένων σε Apache Spark. Για τη δημιουργία του cluster χρησιμοποιήθηκε το λειτουργικό σύστημα Debian 10, ενώ για την εγκατάσταση του Apache Spark χρησιμοποιήθηκε Spark 3.1.1 σε Hadoop 3.2.2. Το cluster που δημιουργήθηκε αποτελείται από 4 συνολικά μηχανήματα, 1 Master και 3 Workers. Κάθε μηχάνημα διαθέτει 2vCPU μνήμη RAM 13GB, σκληρό δίσκο 500GB και μνήμη SSD 375GB.

Δεδομένα. Για την αξιολόγηση της απόδοσης των αλγορίθμων χρησιμοποιήθηκε ένα σημειακό σύνολο δεδομένων 7.4M παρατηρήσεων, μεγέθους 1.1GB, που ομαδοποιούνται σε 368.977 τροχιές, όπως προέκυψε ύστερα από την προεπεξεργασία των δεδομένων. Το συνολικό dataset αφορά δεδομένα πραγματικής κίνησης οχημάτων συνολικής διάρκειας 5 μηνών από τον Ιούλιο έως τον Νοέμβριο του 2018 στην Περιφέρεια Αττικής. Κάθε παρατήρηση είναι της μορφής: (Vehicle_ID, Trajectory_ID, x, y, t, ei, Length, Speed, StartNode, EndNode).



Χάρτης 1. Χάρτης περιοχής μελέτης με το πλήθος παρατηρήσεων ανά ακμή

Επιλογή Hotspots. Με στόχο την αναγνώριση hotspots κυκλοφοριακής συμφόρησης σε δεδομένα τροχιάς, ως τελικό αποτέλεσμα του αλγορίθμου ορίζονται μόνο εκείνες οι top-k ακμές που έχουν το υψηλότερο Getis-Ord z-score για ένα ορισμένο επίπεδο στατιστικής σημαντικότητας. Για τον λόγο αυτό υλοποιήθηκε και το τέταρτο βήμα του αλγορίθμου, στο οποίο το σύνολο των αποτελεσμάτων του αλγορίθμου φιλτράρονται ως προς ένα διάστημα εμπιστοσύνης και στη συνέχεια επιλέγονται οι top-k κορυφαίες ακμές.

Μετρικές. Η κύρια μετρική αξιολόγησης των πειραματικών δοκιμών ήταν ο χρόνος εκτέλεσης που απαιτείται για την ολοκλήρωση κάθε μεμονωμένου βήματος του αλγορίθμου. Ο χρόνος εκτέλεσης μετράται σε δευτερόλεπτα (seconds).

Επίσης, πέρα από το χρόνο εκτέλεσης για κάθε πείραμα υπολογίζονται και οι εξής μετρικές: (a) το πλήθος των παρατηρήσεων (att_edges) που τους έχει αποδοθεί δείκτης κυκλοφοριακής συμφόρησης για ένα temporal partition κατά το πρώτο βήμα του αλγορίθμου, (b) το πλήθος των ακμών του 3D γράφου, (c) το πλήθος των ακμών (Gi_edges) στις οποίες έχει αποδοθεί Getis-Ord z-score κατά την ολοκλήρωση του τρίτου βήματος του αλγορίθμου. Οι τιμές αυτές επηρεάζουν ουσιαστικά την απόδοση του αλγορίθμου και δίνουν μια ιδέα αναφορικά με τον όγκο των δεδομένων που διαχειρίζεται το σύστημα.

Τέλος, κατά την αξιολόγηση χρησιμοποιείται η μέθοδος της οπτικοποίησης των αποτελεσμάτων. Πρακτικά, η χωρική αναπαράσταση των αποτελεσμάτων έχει ως στόχο τον εντοπισμό διαφορών στα αποτελέσματα με ποιοτικό τρόπο στην περίπτωση εφαρμογής ή όχι σημείου cut-off στους υπολογισμούς. Γίνεται αντιληπτό ότι με την εφαρμογή ή όχι σημείου αποκοπής τους υπολογισμούς, αναμένουμε να παρατηρήσουμε διαφορές όχι μόνο στην μετρική του χρόνου εκτέλεσης του αλγορίθμου, αλλά και πιθανώς στην κατανομή των hotspots στο χώρο.

Διαδικασία Αξιολόγησης. Η απόδοση του αλγορίθμου αξιολογήθηκε ως προς τέσσερεις παραμέτρους (a) το μέγεθος του συνόλου δεδομένων, ώστε να εξεταστεί η συμπεριφορά του αλγορίθμου σε δεδομένα διαφορετικής κλίμακας,

(b) το χρονικό παράθυρο της ανάλυσης, που καθορίζει το πλήθος των temporal partitions, (c) το πλήθος των partitions που θα χωριστεί το εκάστοτε RDD στο cluster, ώστε να παρατηρήσουμε τη διαφορά στην απόδοση όσο αυξάνεται η παραλληλία και τέλος (d) το πλήθος των γειτονικών παρατηρήσεων που λαμβάνονται υπόψη στους υπολογισμούς του Getis-Ord z-score.

Πρακτικά, η χρήση του συνόλου των γειτονικών παρατηρήσεων (Getis-Ord Calculations without cutoff) ή ενός τμήματος αυτών (Getis-Ord Calculations with cutoff) αφορά διαφορετική υλοποίηση του αλγορίθμου, με διαφοροποίηση αποκλειστικά στο τρίτο βήμα του αλγορίθμου. Αναφορικά με τις σταθερές της συνάρτησης βάρους, η παράμετρος του bandwidth (εύρος h) του Gaussian Kernel θεωρήθηκε ίση με την τιμή 8 για την υλοποίηση χωρίς cutoff στους υπολογισμούς και ίση με τα 2/3 του γειτονικών παρατηρήσεων (n-neigh) που λήφθηκαν υπόψη στους υπολογισμούς.

Προς διευκόλυνση του αναγνώστη τα βήματα του αλγορίθμου, συνοψίζονται στην συνέχεια:

- **Stage #1:** Στο αρχικό σύνολο δεδομένων εφαρμόζεται ένα σύνολο από μετασχηματισμούς, ώστε να ανατεθεί μια τιμή κυκλοφοριακής συμφόρησης στο σύνολο των ακμών του 3D γράφου.
- **Stage #2:** Γίνεται αναπαράσταση του οδικού δικτύου του γράφου και υπολογισμός της μέσης τιμής και τυπικής απόκλισης της μεταβλητής της κυκλοφοριακής συμφόρησης. Το σύνολο της πληροφορίας ανατίθεται σε μια broadcast μεταβλητή και γίνεται διαθέσιμη σε όλους τους nodes του cluster.
- **Stage #3:** Υπολογίζονται τα συστατικά μέλη της συνάρτησης του δείκτη Gi*, καθώς και ο ίδιος ο δείκτης Gi* για το σύνολο των ακμών του 3D γράφου.
- **Stage #4:** Τα αποτελέσματα φιλτράρονται ως προς ένα διάστημα εμπιστοσύνης και στη συνέχεια από αυτά επιλέγονται οι top-k κορυφαίες ακμές, δηλαδή οι ακμές με το υψηλότερο z-score.

Το σύνολο των παραμέτρων που χρησιμοποιούνται για την αξιολόγηση του αλγορίθμου συνοψίζονται στον Πίνακα 5.

Πίνακας 5. Παράμετροι πειράματος
** (με σκούρο χρώμα απεικονίζονται οι default παράμετροι)*

Παράμετροι	**Τιμές**
Μέγεθος dataset	2.5M, 5M, **7.4M** παρατηρήσεις
Χρονικό παράθυρο ανάλυσης	**24hrs**, 48hrs, 168hrs
Πλήθος partitions	6, 8, 18, 60, **120**
Γειτονικές παρατηρήσεις	6-neigh, 12-neigh, 24-neigh, All-neigh

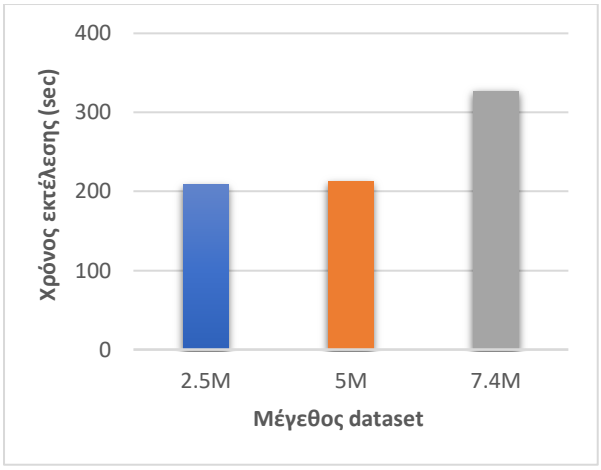
6.2 ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΙΚΗΣ ΑΞΙΟΛΟΓΗΣΗΣ

6.2.1 Αναφορικά με το χρόνο εκτέλεσης

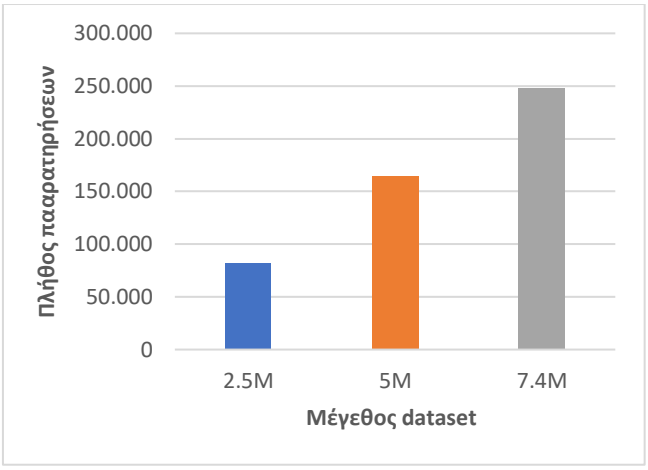
Κατά την παρουσίαση των αποτελεσμάτων της πειραματικής αξιολόγησης, αναφορά θα γίνεται στο χρόνο εκτέλεσης των Stages #1 και #3 του αλγορίθμου, καθώς και στο συνολικό χρόνο εκτέλεσης. Ο χρόνος εκτέλεσης των Stages #2 και #4 είναι αμελητέος σε σύγκριση με αυτόν των υπόλοιπων σταδίων και δεν επηρεάζει ουσιαστικά το αποτέλεσμα. Η συμπεριφορά του αλγορίθμου πρόκειται να εξεταστεί για τις διαφορετικές τιμές των επιμέρους παραμέτρων, τόσο μεμονωμένα όσο και συνδυαστικά. Αναλυτικότερα:

- Μέγεθος συνόλου δεδομένων

Στο Διάγραμμα 10 απεικονίζονται τα αποτελέσματα των πειραμάτων μας, κάνοντας χρήση συνόλων δεδομένων διαφορετικού μεγέθους (2.5M, 5M, 7.4M παρατηρήσεων). Εύκολα γίνεται αντιληπτό ότι καθώς αυξάνεται το πλήθος των παρατηρήσεων προς επεξεργασία, αυξάνεται και ο συνολικός χρόνος εκτέλεσης του αλγορίθμου. Ο χρόνος εκτέλεσης του Stage #1 εξαρτάται άμεσα από το μέγεθος του dataset, καθώς περιλαμβάνει το βήμα του διαβάσματος των δεδομένων από τον δίσκο. Ο χρόνος εκτέλεσης του Stage #1 για το σύνολο δεδομένων των 2.5M και 5M παραμένει σταθερός, ενώ αύξηση του χρόνου εκτέλεσης παρατηρείται για το σύνολο δεδομένων των 7.4M παρατηρήσεων. Σε αντίθεση με τον χρόνο εκτέλεσης, το μέγεθος του attr_rdd (att_edges) αυξάνεται γραμμικά με το μέγεθος του dataset.



(α) Χρόνος εκτέλεσης Stage#1 ως προς το μέγεθος του dataset

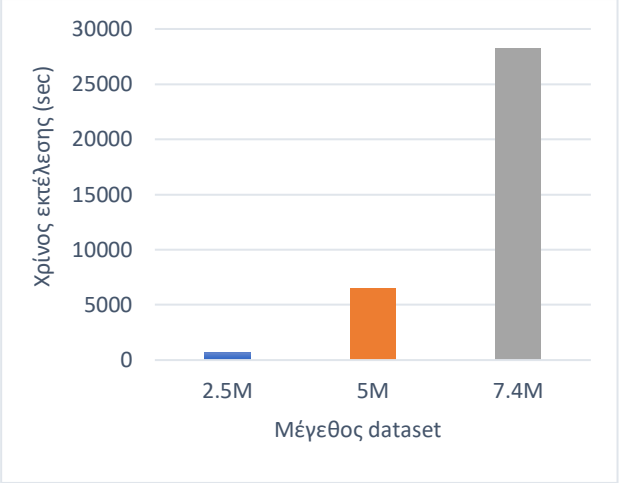


(β) Μέγεθος att_edges RDD που προκύπτει από την εκτέλεση του Stage #1

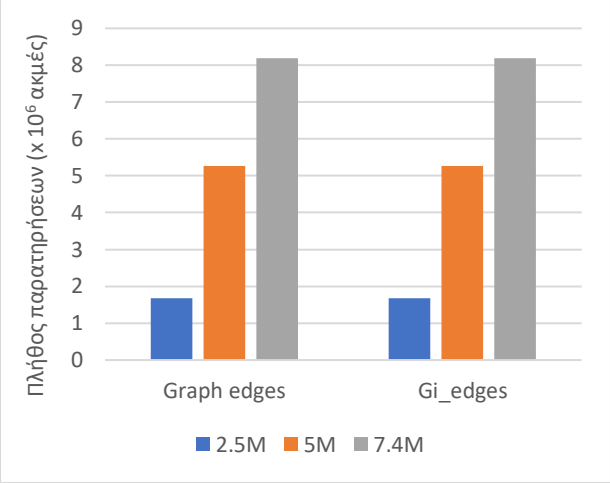
Διαγράμματα 10. Μετρικές απόδοσης αλγορίθμου στο Stage#1 σε συνάρτηση με το μέγεθος του dataset

Παρατηρώντας το Διάγραμμα 11(α), προκύπτει ότι ο χρόνος εκτέλεσης του Stage #3 αυξάνεται εκθετικά σε σχέση με το μεγέθους του dataset. Ο χρόνος εκτέλεσης του Stage #3 είναι συνάρτηση του πλήθος των φορών που χρειάζεται να εκτελεστεί ο αλγόριθμος *GetisOrdCalculations*, που είναι ίσος με την μετρική att_edges. Συνεπώς, με το διπλασιασμό του αρχικού dataset, οδηγούμαστε στο διπλασιασμό της μετρικής att_edges και στο διπλασιασμό των φορών εκτέλεσης του αλγορίθμου.

Παράλληλα, όπως φαίνεται στο Διάγραμμα 11(β) μέσω της μετρική Graph edges, αυξάνοντας το μέγεθος του αρχικού dataset, το μέγεθος του γράφου που αναπαριστά το οδικό δίκτυο αυξάνεται γραμμικά. Αυτό έχει ως αποτέλεσμα να αυξάνεται η πολυπλοκότητα των αναζητήσεων στο γράφο με σκοπό τον προσδιορισμό της σχέσης γειτνίασης μεταξύ δυο ακμών. Τέλος, το πλήθος των αναζητήσεων στο γράφο είναι συνάρτηση και του πλήθους των γειτονικών παρατηρήσεων που επιθυμούμε να λάβουμε υπόψη στους υπολογισμούς του Gi*. Στην περίπτωση του αλγορίθμου χωρίς cut-off, η σχέση γειτνίασης μεταξύ των ακμών πρέπει να προσδιοριστεί για το σύνολο των προγενέστερων ακμών του αντίστροφου γράφου. Από τα παραπάνω γίνεται αντιληπτό ότι καθώς αυξάνεται το μέγεθος του dataset, αυξάνεται όχι μόνο η πολυπλοκότητα του σταδίου αλλά και ο όγκος των δεδομένων που απαιτείται να μετακινηθεί στο δίκτυο (network traffic).



(α) Χρόνος εκτέλεσης Stage #3



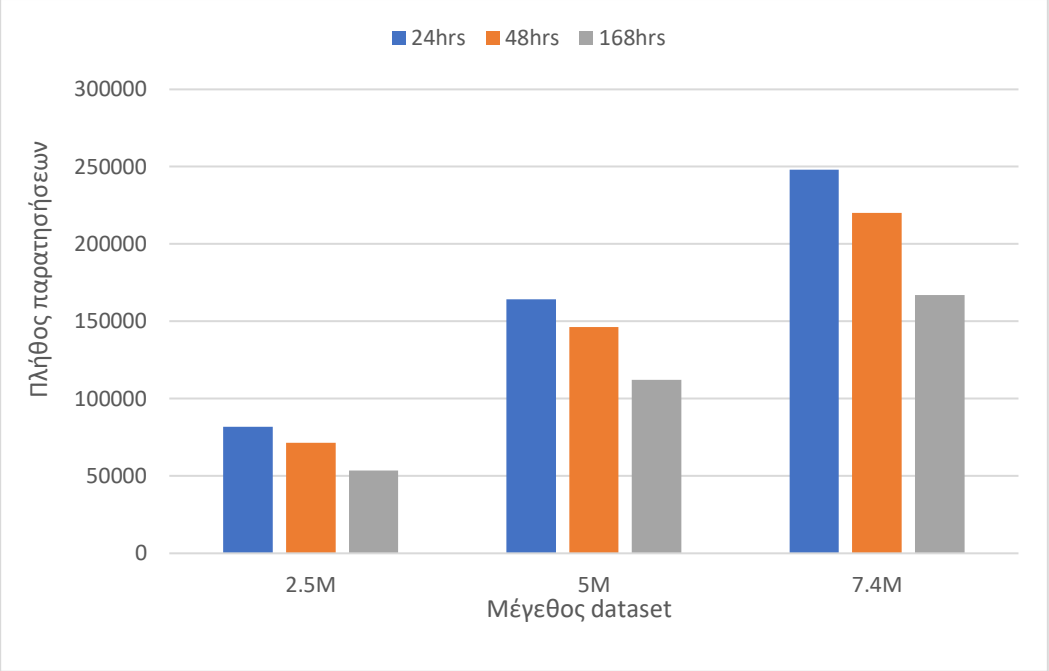
(β) Πλήθος παρατηρήσεων που προκύπτουν από την εκτέλεση του Stage #2 και #3

Διαγράμματα 11. Μετρικές απόδοσης αλγορίθμου σε συνάρτηση με το μέγεθος του dataset

- Χρονικό παράθυρο ανάλυσης

Η σχέση μεταξύ του χρονικού παραθύρου της ανάλυσης και του χρόνου εκτέλεσης του αλγορίθμου είναι αντιστρόφως ανάλογη. Εφαρμόζοντας χρονικά παράθυρα ανάλυσης 168, 48 και 24 ωρών σε ένα σύνολο δεδομένων που αφορά 5 μήνες οδηγούμαστε στη δημιουργία 22, 77 και 153 temporal partitions αντίστοιχα. Όσο λιγότερα είναι τα temporal partitions, τόσο μικρότερος είναι και ο χρόνος εκτέλεσης του αλγορίθμου, καθώς κατά το Stage #1 περισσότερα δεδομένα γίνονται aggregate.

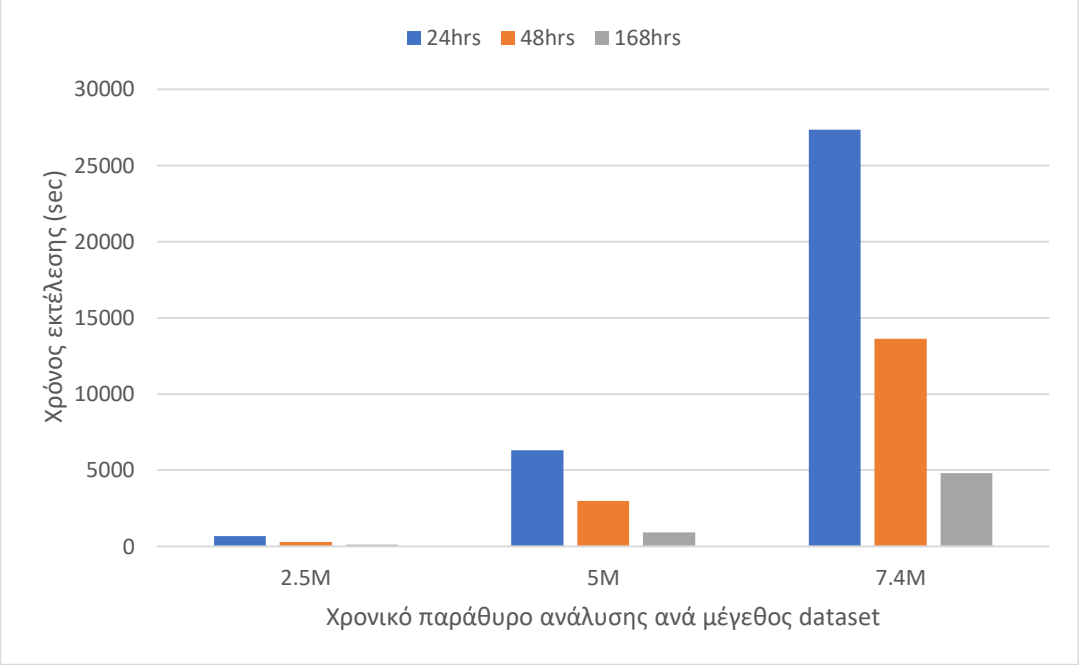
Αναφορικά με το χρόνο εκτέλεσης ανά στάδιο προκύπτει ότι ο χρόνος εκτέλεσης του Stage #1 μένει ανεπηρέαστος από το μέγεθος του χρονικού παραθύρου παρά το γεγονός ότι για μικρότερο χρονικό παράθυρο ανάλυσης, το μέγεθος του RDD που προκύπτει είναι μεγαλύτερου μεγέθους. Η διαπίστωση αυτή οδηγεί στο συμπέρασμα ότι ο χρόνος εκτέλεσης του Stage #1 εξαρτάται κυρίως από την καθυστέρηση της κλίσης των δεδομένων από το δίσκο.



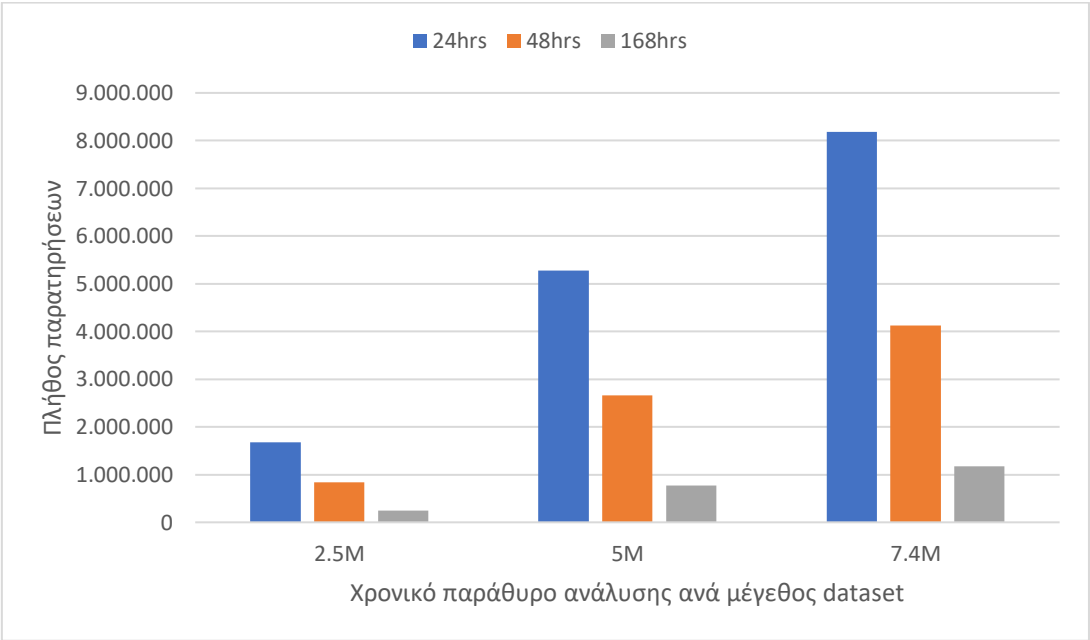
Διαγράμματα 12. Μετρική att_edges για διαφορετικά χρονικά παράθυρα ανάλυσης

Παράλληλα, η αύξηση του μεγέθους του att_rdd, όπως αναφέρθηκε και προηγμένως, οδηγεί σε μεγαλύτερη ανάγκη μετακίνησης δεδομένων μέσα στο cluster και μεγαλύτερη πολυπλοκότητα στους υπολογισμούς κατά την εκτέλεσης του Stage #3, αυξάνοντας το χρόνο εκτέλεσης.

Παρατηρώντας από το Διάγραμμα 13(α) προκύπτει ότι η σχέση μεταξύ του χρονικού παραθύρου στην ανάλυση και του χρόνου εκτέλεσης ποικίλει. Το temporal partitioning των 48 ωρών είναι περίπου δυο φορές πιο αποδοτικό από απόψεως χρόνου εκτέλεσης σε σχέση με αυτό των 24 ωρών και των 168 ωρών περίπου επτά φορές πιο αποδοτική από αυτό των 24 ωρών, υποδεικνύοντας γραμμική σχέση μεταξύ του χρόνου εκτέλεσης και του χρονικού παραθύρου για τα dataset των 2.5M και 5M παρατηρήσεων. Στην περίπτωση του dataset των 7.4M παρατηρήσεων μπορεί να χαρακτηριστεί ως φθίνουσα εκθετική.



(α) Χρόνος εκτέλεσης αλγορίθμου στο Stage#3 για διαφορετικά χρονικά παράθυρα ανάλυσης



(β) Πλήθος Graph edges και Gi_edges για διαφορετικά χρονικά παράθυρα ανάλυσης

Διαγράμματα 13. Μετρικές απόδοσης αλγορίθμου ως προς το χρονικό παράθυρο της ανάλυσης

Το μέγεθος του γράφου στις δυο διαστάσεις καθορίζεται από το πλήθος των μοναδικών ακμών που περιέχει το RDD που εξάγεται κατά την ολοκλήρωση του Stage #1. Η διάσταση του χρόνου οδηγεί ουσιαστικά στη δημιουργία αντιγράφων του γράφου που αναφέρονται σε διαφορετικά χρονικά παράθυρα. Αυξάνοντας το πλήθος των temporal partitions αυξάνεται και το μέγεθος του

3D γράφου, με αποτέλεσμα να αυξάνεται η πολυπλοκότητα υπολογισμού των παραμέτρων της συνάρτησης βάρους.

- Γειτονικές παρατηρήσεις

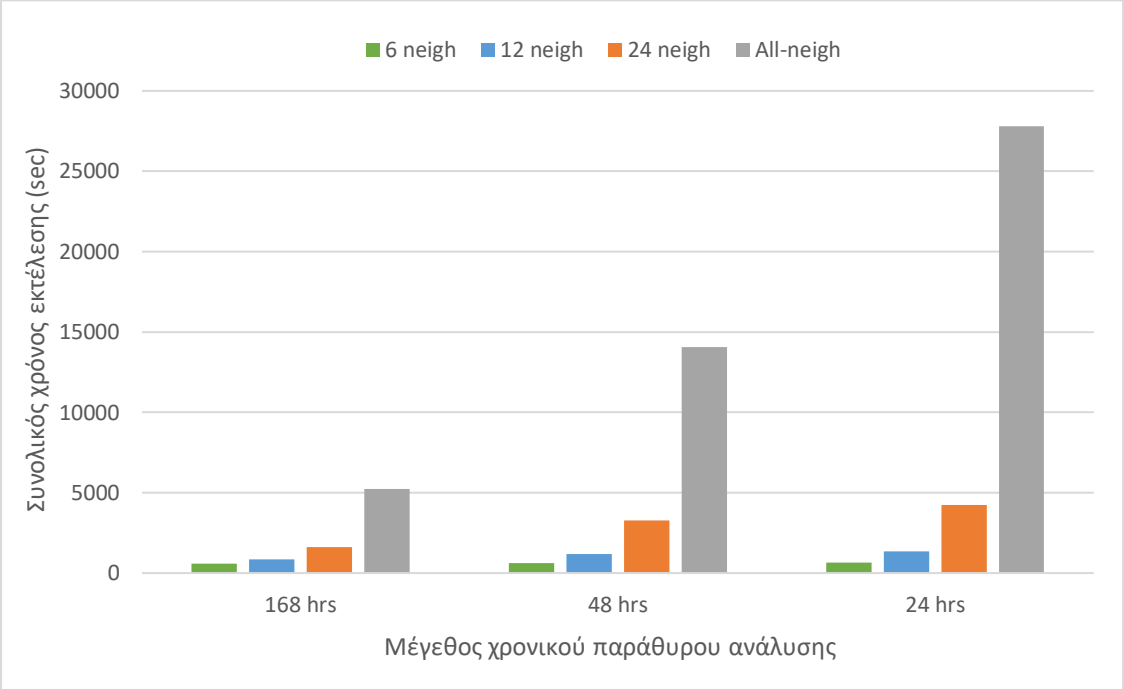
Η παράμετρος της γειτνίασης έχει καθοριστικό ρόλο στη διαμόρφωση του χρόνου εκτέλεσης του αλγορίθμου. Η μείωση του συνολικού χρόνου εκτέλεσης οφείλεται αποκλειστικά στη μείωση του χρόνου εκτέλεσης του Stage #3, καθώς σε αυτό πραγματοποιείται το σύνολο των υπολογισμών που θα καθορίσουν τη σχέση γειτνίασης μεταξύ των ακμών του γράφου. Ο χρόνος εκτέλεσης του Stage #1 δεν εξαρτάται από την παράμετρο n-neigh.

Όπως έχει αναφερθεί και νωρίτερα αναπτύχθηκαν δυο εκδοχές του αλγόριθμου. Στην πρώτη εκδοχή δεν εφαρμόζεται σημείο αποκοπής στους υπολογισμούς, με αποτέλεσμα η παράμετρος της κυκλοφοριακής συμφόρησης του συνόλου των γειτονικών ακμών να συνεισφέρει στο z-score της υπό εξέταση ακμής. Στη δεύτερη εκδοχή εφαρμόζεται σημείο αποκοπής στους υπολογισμούς, με αποτέλεσμα το z-score της κάθε ακμής να επηρεάζεται μόνο από τις τιμές της κυκλοφοριακής συμφόρησης ακμών εντός μια ακτίνας, καθοριζόμενης από το χρήστη.

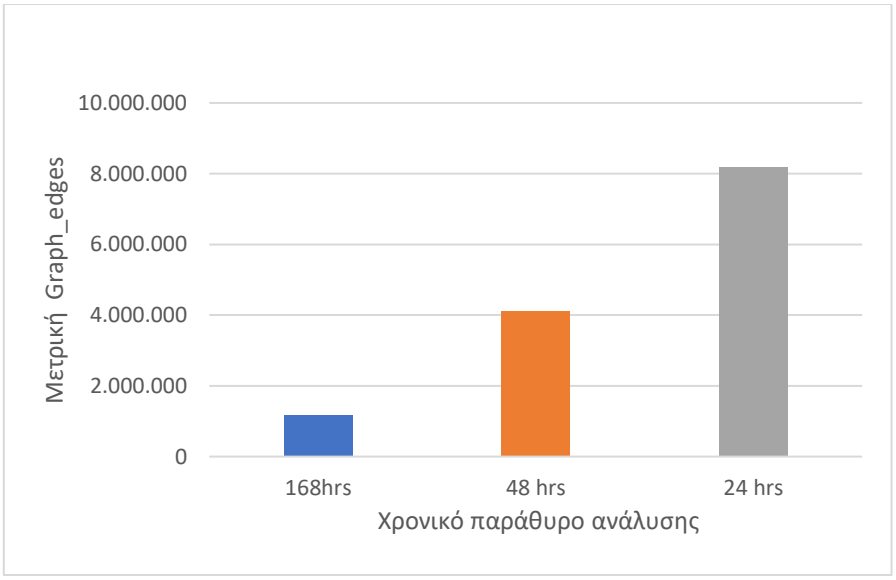
Από το Διάγραμμα 14 προκύπτει ότι ο συνολικός χρόνος εκτέλεσης μειώνεται δραστικά εφαρμόζοντας σημείο αποκοπής στους υπολογισμούς, καθώς η συνάρτηση βάρους υπολογίζεται μόνο για τις ακμές του γράφου που βρίσκονται n hops μακριά από την εξεταζόμενη ακμή τόσο στη διάσταση του χώρου, όσο και στη διάσταση του χρόνου. Αυτό έχει ως αποτέλεσμα τα δεδομένα που επιστρέφονται από την εφαρμογή της συνάρτησης *GetisOrdCalculations* να είναι λιγότερα σε πλήθος, μειώνοντας τον όγκο των δεδομένων που μετακινούνται στο δίκτυο και συνεπώς το συνολικό χρόνο εκτέλεσης του Stage #3.

Μια επιπλέον ένδειξη για την μείωση των απαιτούμενων υπολογισμών στην περίπτωση εφαρμογής σημείου αποκοπής είναι η σχέση μεταξύ του χρόνου εκτέλεσης του αλγορίθμου και του μεγέθους του RDD που προκύπτει κατά την

ολοκλήρωση του σταδίου, όπως περιγράφεται από τη μετρική Gi_edges. Η μετρική Gi_edges μειώνεται κατά 2.8M, 2M και 1.5M παρατηρήσεις για τιμές παραμέτρου 6-neigh, 12-neigh και 24-neigh αντίστοιχα, σε σχέση με την τιμή της μετρικής όταν στους υπολογισμούς συμμετέχουν το σύνολο των ακμών του γράφου (Διάγραμμα 16). Το φαινόμενο οφείλεται στην απόδοση τιμών Gi* z-score σε ακμές που μπορεί να μην περιέχουν πληροφορία σε κάποιο temporal partition, αλλά ο υπολογισμός του Gi* z-score να επιτυγχάνεται εξαιτίας μη κενών γειτονικών τιμών. Καθώς η τιμή της παραμέτρου n-neigh αυξάνεται η πιθανότητα να υπολογιστεί το Gi* z- score για κενές ακμές σε ένα temporal partition αυξάνεται, σε μεγαλύτερο χρόνος εκτέλεσης.

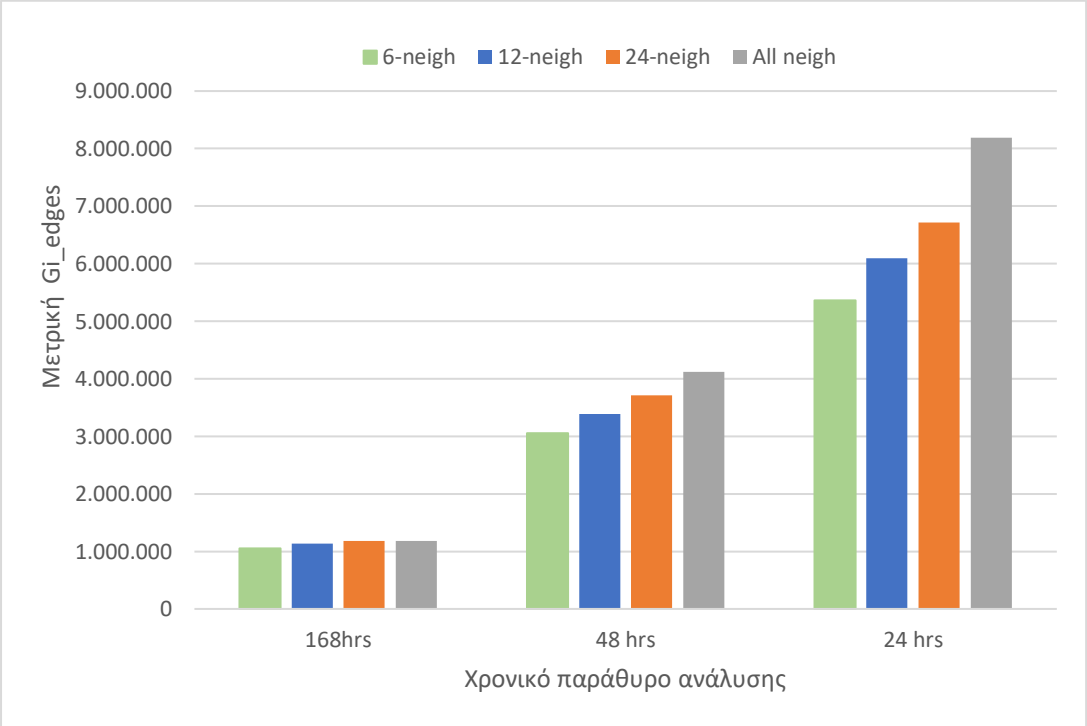


Διαγράμματα 14. Χρόνος εκτέλεσης για διαφορετικές τιμές της παραμέτρου n-neigh



Διαγράμματα 15. Πλήθος ακμών γράφου σε συνάρτηση της παραμέτρου temporal partition

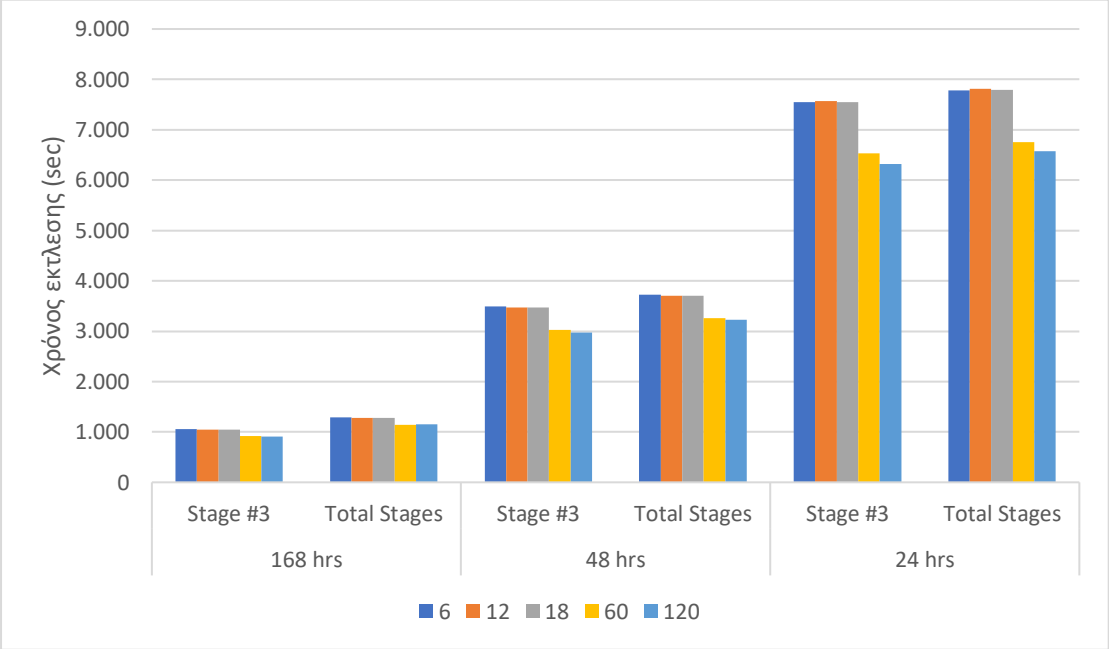
Μελετώντας την επίδραση της παραμέτρου n-neigh στο χρόνο εκτέλεσης, προκύπτει ότι η μείωση του πλήθους των γειτονικών παρατηρήσεων που συμμετέχουν στους υπολογισμούς από 12 σε 6 και από 24 σε 12 οδηγεί σε υποτετραπλασιασμό του χρόνου εκτέλεσης σε κάθε μια από τις δυο περιπτώσεις για οποιοδήποτε μέγεθος 3D γράφου.



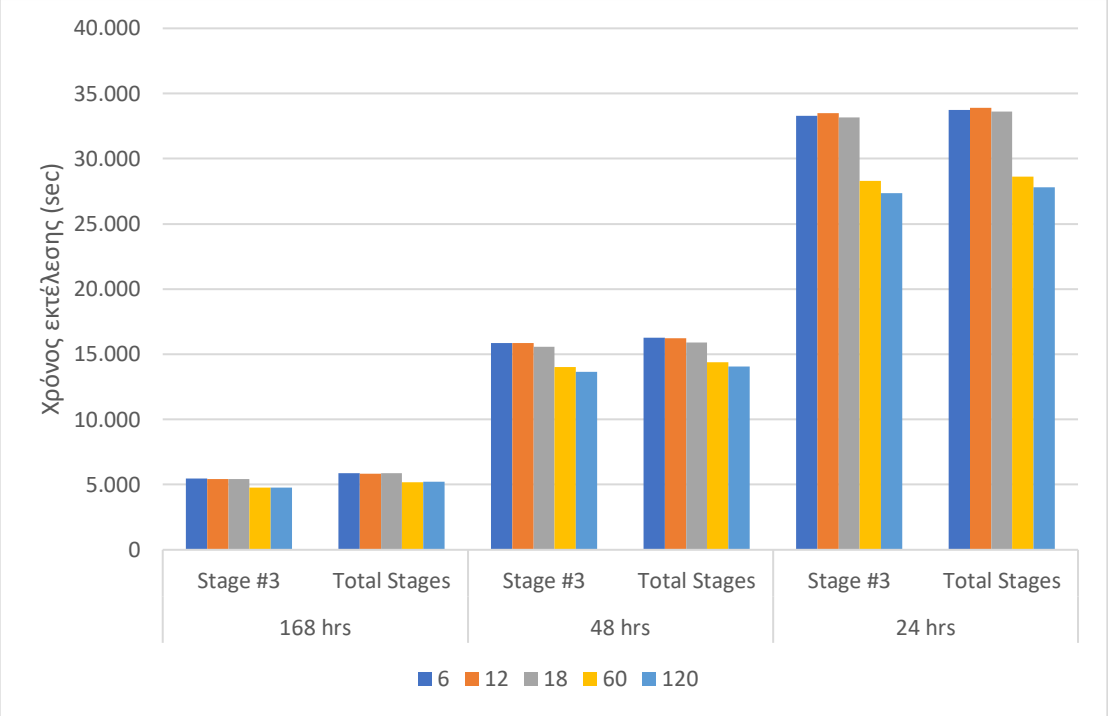
Διαγράμματα 16. Μετρική Gi_edges για διαφορετικές τιμές της παραμέτρου n-neigh

- Πλήθος partitions

Στο Διάγραμμα 17 συνοψίζεται η πληροφορία του χρόνου εκτέλεσης του αλγορίθμου για διαφορετικά επίπεδα παραλληλίας κάνοντας χρήση της εκδοχής του αλγορίθμου χωρίς σημείο αποκοπής στους υπολογισμούς. Για το σύνολο δεδομένων των 7.4M και 5M παρατηρήσεων και για τις διαφορετικές τιμές της παραμέτρου που ορίζει το πλήθος των temporal partitions, παρατηρείται μικρή μείωση του χρόνου εκτέλεσης του αλγορίθμου τόσο στο Stage #3, όσο και στο Stage #4, τάξης του 15% κατά μέσο όρο. Στην περίπτωση του συνόλου δεδομένων των 2.5M παρατηρήσεων, δεν παρατηρείται αλλαγή στον χρόνο εκτέλεσης αυξάνοντας την παραλληλία.



Διαγράμματα 17. Απόδοση αλγορίθμου για διαφορετικά επίπεδα παραλληλίας (Dataset 5M)



Διαγράμματα 18. Απόδοση αλγορίθμου για διαφορετικά επίπεδα παραλληλίας (Dataset 7.4M)

Η συμπεριφορά αυτή του αλγορίθμου μπορεί να ερμηνευτεί αν σκεφτεί κανείς την πολυπλοκότητα του Stage #3, το οποίο είναι και το πιο υπολογιστικά ακριβό. Για τον υπολογισμό του Gi* z -score για μια ακμή πρέπει να καθοριστεί η σχέση γειτνίασης της με το σύνολο προγενέστερων ακμών αυτής στο γράφο. Η σχέση αυτή μεταξύ ενός ζεύγους ακμών υπολογίζεται εφαρμόζοντας τη συνάρτηση *nx.shortest_path_length* (Ψευτοκώδικας 6), με πολυπλοκότητα O (E + V LogV), όπου E το πλήθος των ακμών και V το πλήθος των κόμβων του ego_graph. Το σύνολο των υπολογισμών επαναλαμβάνεται για κάθε μια από τις γραμμές του RDD που γίνεται flatMap.

6.2.2 Οπτικοποίηση και ανάλυση των αποτελεσμάτων της Hotspot ανάλυσης

Κατά την αξιολόγηση των αποτελεσμάτων προκύπτει ότι η παράμετρος n-neigh, που ορίζει το πλήθος των γειτονικών παρατηρήσεων που θα συμμετέχουν στους υπολογισμούς του Gi* z-score, έχει καθοριστικό ρόλο στην διαμόρφωση του χρόνου εκτέλεσης του αλγορίθμου. Μέσω της οπτικοποίησης και της ανάλυσης των αποτελεσμάτων της hotspot ανάλυσης, στόχος είναι να

αξιολογηθεί το ποιοτικό αποτέλεσμα του αλγορίθμου για διαφορετικές τιμές της παραμέτρου n-neigh.

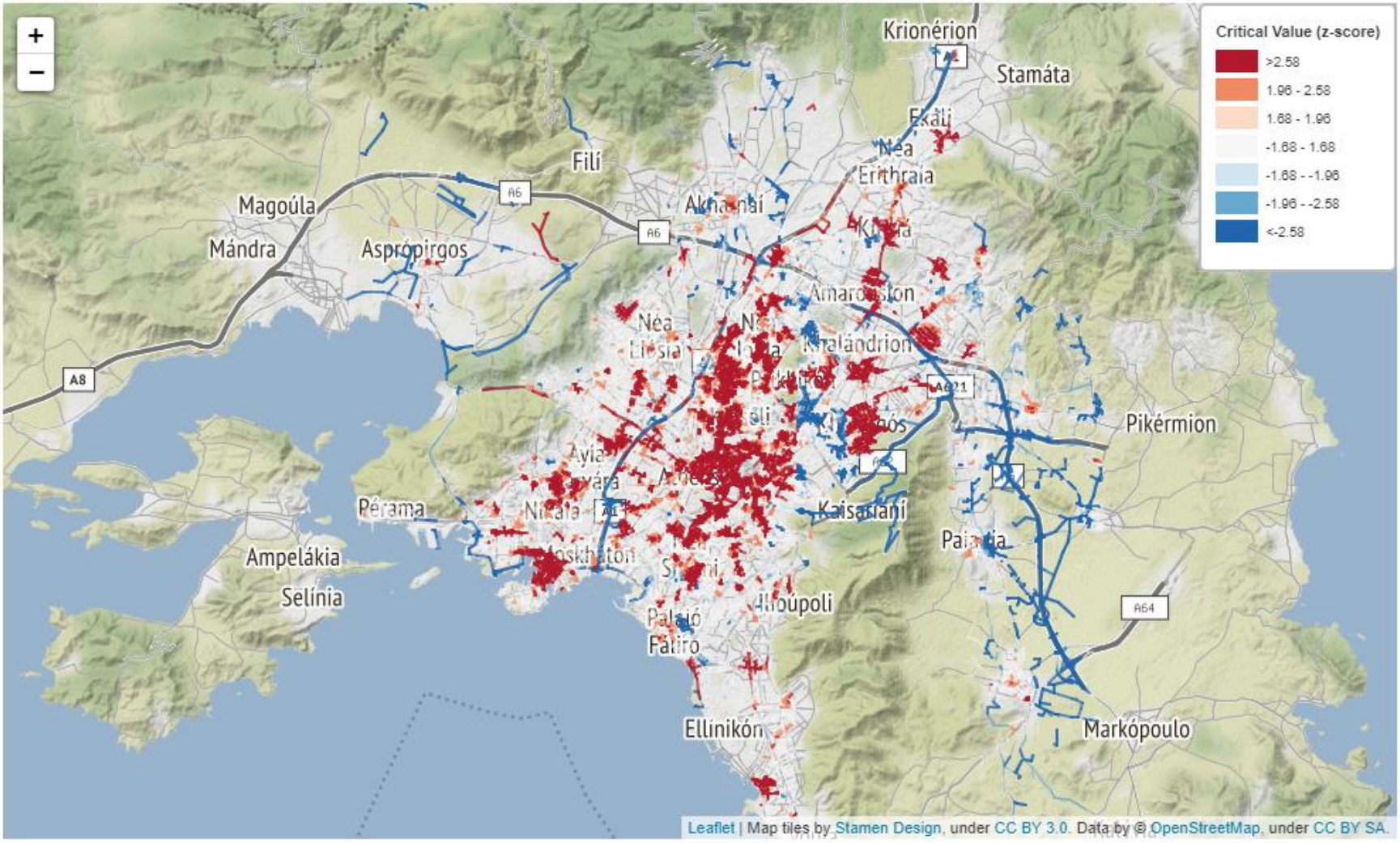
Τα αποτελέσματα της hotspot ανάλυσης, που παρουσιάζονται στη συνέχεια, αφορούν τα πειράματα με τα δυσμενέστερα αποτελέσματα από άποψη υπολογιστικού κόστους. Ειδικότερα, επιλέχθηκαν να παρουσιαστούν τα αποτελέσματα των πειραμάτων με το μεγαλύτερο σύνολο δεδομένων (7.4M παρατηρήσεις) και με χρονικό παράθυρο ανάλυσης αυτό των 24 ωρών. Καθώς το σύνολο των αποτελεσμάτων του αλγορίθμου περιλαμβάνει στατιστικά σημαντικά hot και cold spots κυκλοφοριακής συμφόρησης για χρονικό διάστημα 5 μηνών, η οπτικοποίηση των αποτελεσμάτων έγινε ως προς ένα temporal partition. Οι χάρτες που ακολουθούν παρουσιάζουν τη χωρική κατανομή του φαινομένου της κυκλοφοριακής συμφόρησης σε επίπεδο ημέρας.

Στους Χάρτες 2, 3, 4 και 5 παρουσιάζονται τα οδικά τμήματα με στατιστικά σημαντικές υψηλές τιμές (κόκκινο χρώμα) και χαμηλές τιμές (μπλε χρώμα) κυκλοφοριακής συμφόρησης για διαφορετικές τιμές της παραμέτρου n-neigh. Μέσω της οπτικοποίησης προκύπτει ότι τα αποτελέσματα των πειραμάτων όταν στους υπολογισμούς του δείκτη Gi* συμμετέχουν 12 γειτονικές παρατηρήσεις είναι όμοια με αυτά που προκύπτουν όταν στους υπολογισμούς συμμετέχουν το σύνολο των γειτονικών παρατηρήσεων. Στην περίπτωση των αποτελεσμάτων με χρήση της παραμέτρου 6-neigh, παρατηρείται μικρότερη ευαισθησία στην αναγνώριση ακραίων σημείων, ενώ στην περίπτωση των 24-neigh παρατηρείται υπερευαισθησία με αποτέλεσμα περισσότερες ακμές να αναγνωρίζονται ως hotspots.

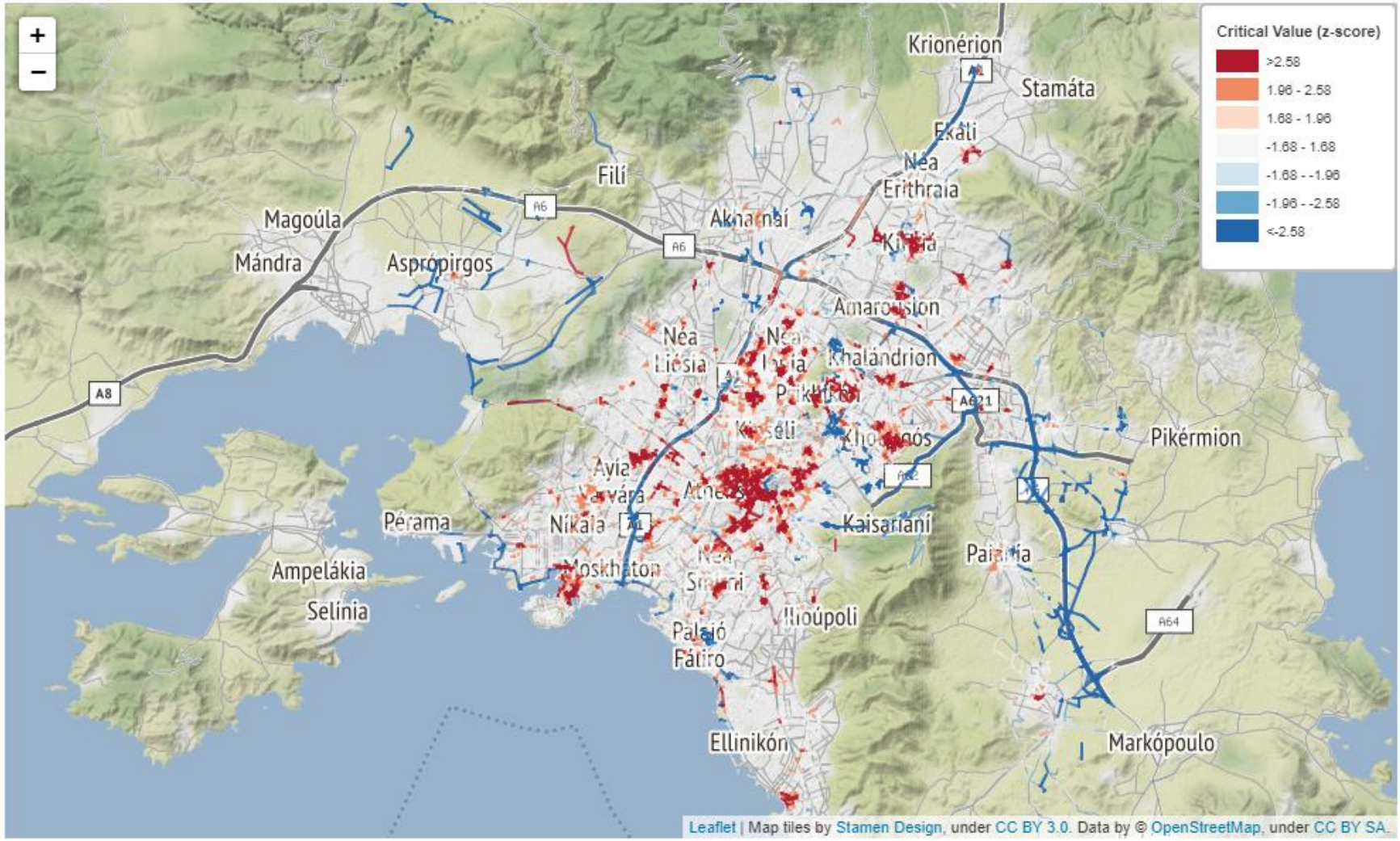
Με στόχο την τεκμηρίωση της παραπάνω διαπίστωσης συγκρίναμε τις 100 κορυφαίες ακμές κυκλοφοριακής συμφόρησης που προκύπτουν όταν στους υπολογισμούς του δείκτη Gi* συμμετέχει το σύνολο των γειτονικών παρατηρήσεων με τις 100 κορυφαίες ακμές όταν εφαρμόζεται σημείο αποκοπής στις 6, 12 και 24 γειτονικές παρατηρήσεις. Εφαρμόζοντας τον τύπο της ομοιότητας Jaccard προέκυψε ομοιότητα 31%, 79% και 53% κατά αντιστοιχία. Μελετώντας τη συσχέτιση των λιστών κατάταξης κατά Spearman, προέκυψε η υψηλή συσχέτιση της τάξης του 82% μεταξύ των αποτελεσμάτων

για παραμέτρους all-neigh και 12 neigh. Με τη θεώρηση ότι τα αποτελέσματα του πειράματος για τον αλγόριθμο που λαμβάνει υπόψιν στους υπολογισμούς το σύνολο των γειτονικών παρατηρήσεων είναι πιο αντιπροσωπευτικά του φαινομένου της κυκλοφοριακής συμφόρησης στην Περιφέρεια Αττικής για ένα temporal partition και εφόσον η ομοιότητα μεταξύ των αποτελεσμάτων είναι υψηλή για τον αλγόριθμο με σημείο αποκοπής ορίζοντας την παράμετρο n-neigh ίση με all-neigh και 12 neigh, προκύπτει ότι η επιλογή της παραμέτρου 12-neigh είναι βέλτιστη, καθώς το υπολογιστικό κόστος είναι μικρότερο.

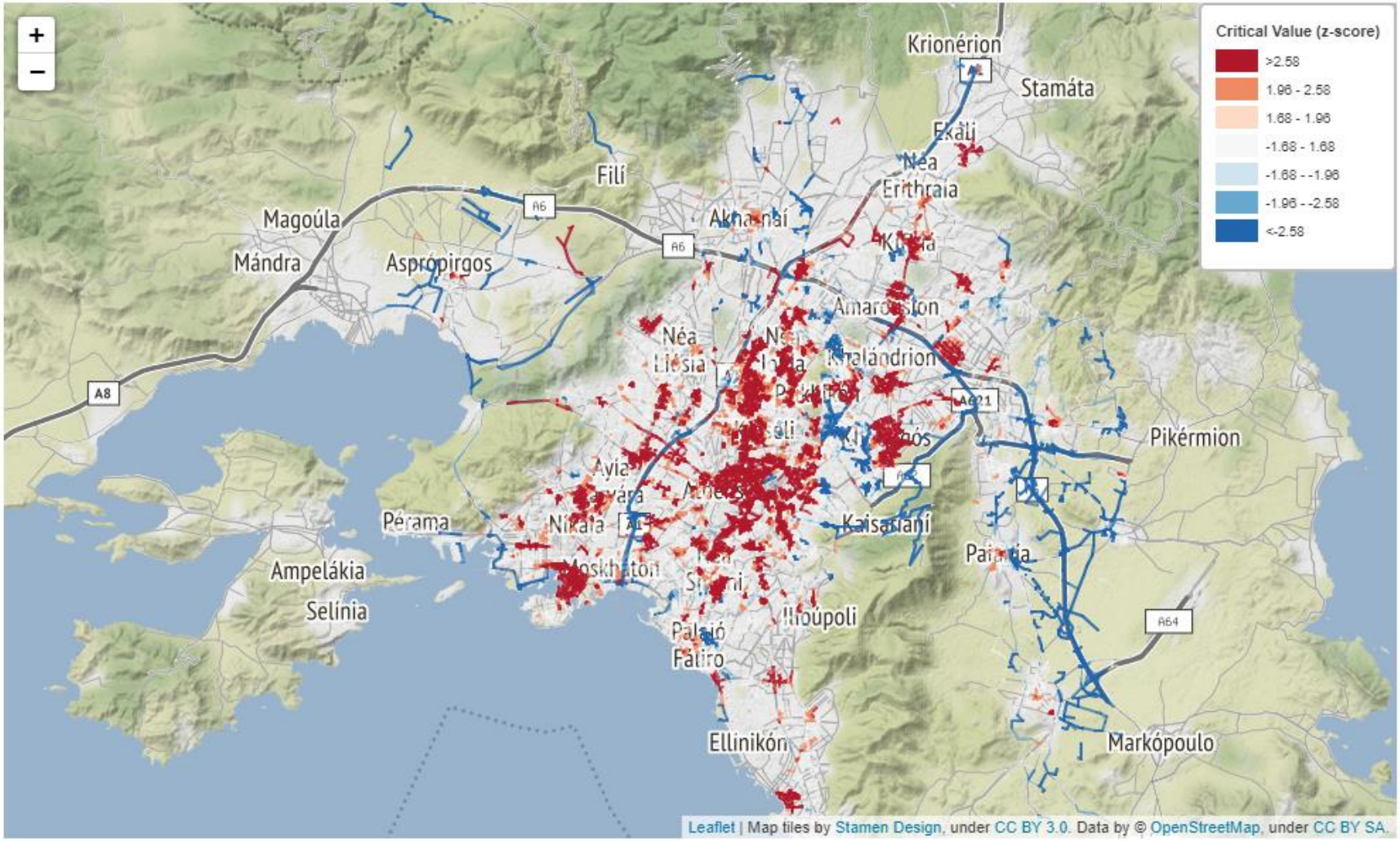
Τέλος, συγκρίνοντας το χάρτη κατηγοριοποίησης των ακμών με βάση το πλήθος παρατηρήσεων ανά ακμή (Χάρτης 1) με τα αποτελέσματα της ανάλυσης hotspot (Χάρτης 2, 3, 4, 5), προκύπτει ότι η υψηλή συγκέντρωση παρατηρήσεων σε ένα οδικό τμήμα δεν συνεπάγεται και υψηλό επίπεδο κυκλοφοριακής συμφόρησης. Χαρακτηριστικό παράδειγμα αποτελεί η περίπτωση της Αττικής Οδού, η οποία παρά την υψηλή συγκέντρωση οχημάτων που παρουσιάζει, τμήματα της χαρακτηρίστηκαν κατά την ανάλυση ως cold spots, δηλαδή ακμές με στατιστικά σημαντικό χαμηλό επίπεδο κυκλοφοριακής συμφόρησης. Γίνεται αντιληπτό ότι η κυκλοφοριακή ικανότητα μιας οδού επηρεάζει την ανοχή της στο φαινόμενο της κυκλοφοριακής συμφόρησης. Στο κέντρο της Αθήνας και σε μεγάλες οδικές αρτηρίες (όπως Λεωφ. Κηφισιάς, Λεωφ. Βουλιαγμένης), ο αλγόριθμος καταφέρνει να αναγνωρίσει οδικά τμήματα και διασταυρώσεις με υψηλή κυκλοφοριακή συμφόρηση, που σύμφωνα με το Χάρτη 1 παρουσιάζουν υψηλή συγκέντρωση παρατηρήσεων.



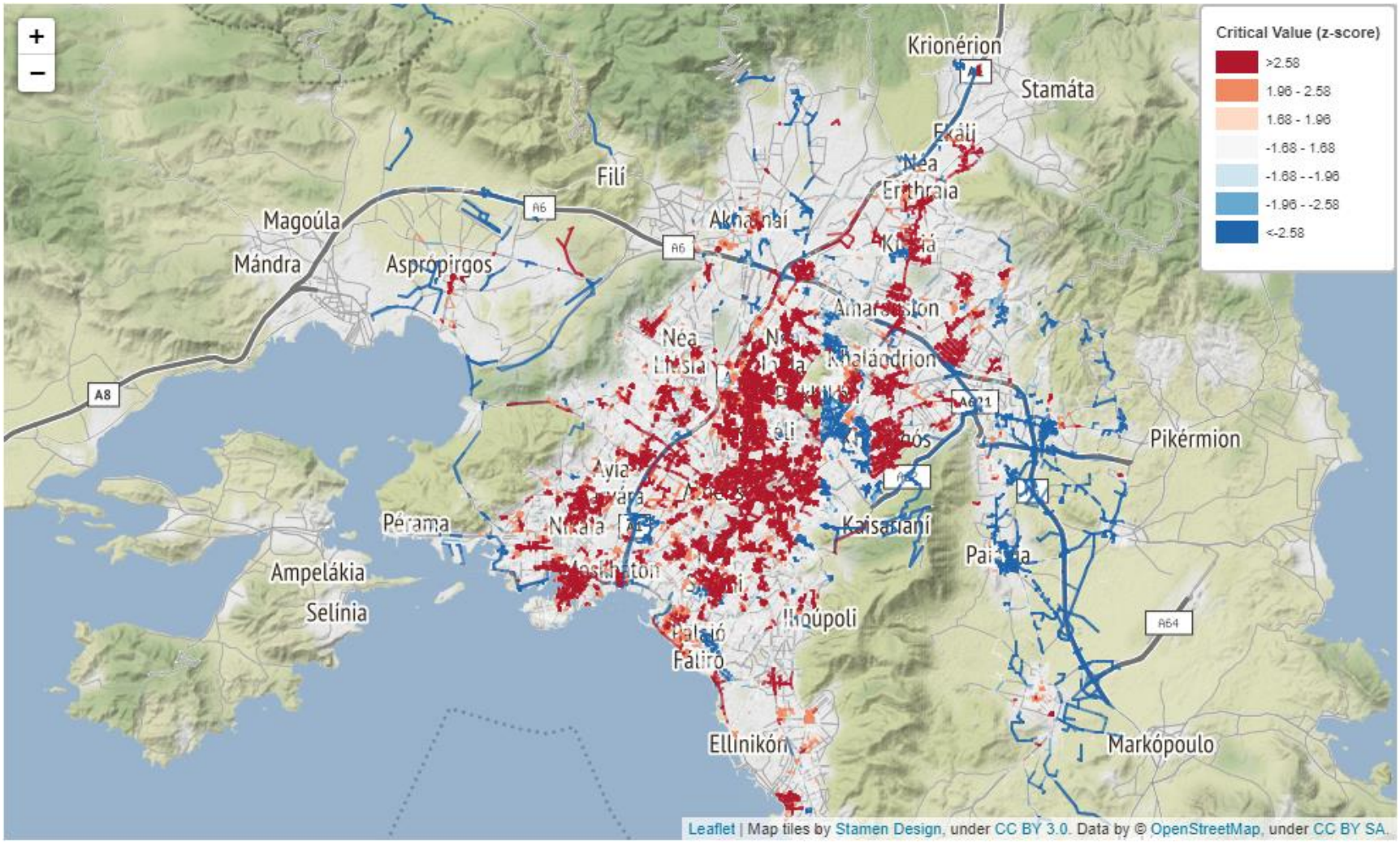
Χάρτης 2. Αποτελέσματα hotspot ανάλυσης με συμμετοχή του συνόλου των γειτονικών παρατηρήσεων (All-neigh)



Χάρτης 3. Αποτελέσματα hotspot ανάλυσης με συμμετοχή 6 γειτονικών παρατηρήσεων (6-neigh)



Χάρτης 4. Αποτελέσματα hotspot ανάλυσης με συμμετοχή 12 γειτονικών παρατηρήσεων (12-neigh)



Χάρτης 5. Αποτελέσματα hotspot ανάλυσης με συμμετοχή 24 γειτονικών παρατηρήσεων (24-neigh)

7 ΣΥΜΠΕΡΑΣΜΑΤΑ

7.1 ΣΥΝΟΨΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η ανάδειξη Hotspots κυκλοφοριακής συμφόρησης στο οδικό δίκτυο, κάνοντας χρήση μεγάλων δεδομένων τροχιάς προερχόμενων από δείκτες GPS. Ειδικότερα, μέσω της ανάλυσης Hotspots, στόχος είναι να βρεθούν οδικά τμήματα με στατιστικά σημαντικές τιμές κυκλοφοριακής συμφόρησης για διαφορετικά χρονικά παράθυρα ανάλυσης.

Με στόχο την ανάδειξη του φαινόμενου της κυκλοφοριακής συμφόρησης, σε πρώτη φάση, επιδιώχθηκε να αποδοθεί μια τιμή κυκλοφοριακής συμφόρησης σε κάθε ακμή του γράφου και η αντιστοίχιση αυτής σε ένα χρονικό παράθυρο ανάλυσης, temporal partition. Ο δείκτης συμφόρησης που χρησιμοποιήθηκε αποτελεί συνάρτηση της παρατηρούμενης ταχύτητας και της ταχύτητας ελεύθερης ροής.

Για ανάδειξη χωρο-χρονικών hotspot κυκλοφοριακής συμφόρησης εφαρμόστηκε ο τροποποιημένος δείκτης Gi*. Ο τροποποιημένος αυτός δείκτης δέχεται ως δεδομένα ένα τρισδιάστατο γράφο οποίος αναπαριστά το οδικό δίκτυο και ενσωματώνει τις τιμές της κυκλοφοριακής συμφόρησης σε επίπεδο ακμής (χωρικό partition) και σε χρονικό παράθυρο (temporal partition). Η σχέση εγγύτητας w_{ij} μεταξύ δυο ακμών e_i και e_j επιλέχθηκε να εκφραστεί μέσω του Gaussian Kernel. Η συνάρτηση του Gaussian Kernel βαραίνει τα δεδομένα με συνεχή τρόπο και σταδιακά μειώνει το βάρος, καθώς αυξάνεται η απόσταση από το κέντρο του kernel, δηλαδή την υπό εξέταση ακμή του γράφου. Εφαρμόζοντας σημείο αποκοπής στους υπολογισμούς του δείκτη Gi*, ουσιαστικά επεμβαίνουμε στη συνάρτηση του Gaussian Kernel και αποδίδουμε μηδενικό βάρος στις ακμές που η απόσταση τους υπερβαίνει ένα threshold, οριζόμενο μέσω της παραμέτρου n-neigh.

Στο πλαίσιο την εργασίας, υλοποιήθηκαν δυο αλγόριθμοι. Στον πρώτο αλγόριθμο, για το χαρακτηρισμό μιας ακμής ως hotspot υπολογίζεται η σχέση

γειτνίασης μεταξύ της εξεταζόμενης ακμής και του συνόλου των γειτονικών της ακμών, πρακτική που αυξάνει την πολυπλοκότητα και το υπολογιστικό κόστος. Ο δεύτερος αλγόριθμος αποτελεί μια προσπάθεια βελτιστοποίησης του πρώτου εφαρμόζοντας σημείο αποκοπής στους υπολογισμούς, με αποτέλεσμα η σχέση γειτνίασης να υπολογίζεται για τις ακμές εντός μια “ακτίνας”.

Η πειραματική αξιολόγηση των αλγορίθμων έγινε ως προς τέσσερα σημεία: το μέγεθος του συνόλου δεδομένων, το μέγεθος του χρονικού παράθυρου της ανάλυσης, ως προς την παραλληλία και το πλήθος των γειτονικών παρατηρήσεων που λαμβάνονται υπόψη.

Αναφορικά με το μέγεθος του αρχικού συνόλου δεδομένων προέκυψε ότι η σχέση μεταξύ του μεγέθους του dataset και του χρόνου εκτέλεσης του αλγορίθμου είναι γραμμική. Ενώ η σχέση του χρόνου εκτέλεσης με το μέγεθος του χρονικού παραθύρου είναι αντιστρόφως ανάλογη. Όσο μικραίνει το χρονικό παράθυρο της ανάλυσης, οδηγούμαστε σε περισσότερα temporal partitions, γεγονός που επηρεάζει το μέγεθος του 3D γράφου και συνεπώς μεγαλύτερο υπολογιστικό κόστος κατά την εκτέλεση του Stage #3. Αυξάνοντας την παραλληλία του αλγορίθμου, παρατηρήθηκε μικρή βελτίωση στον χρόνο εκτέλεσης του αλγορίθμου, της τάξης του 15%. Θεωρήθηκε ότι η μη σημαντική μείωση στο χρόνο εκτέλεσης οφείλεται στο μεγάλο υπολογιστικό κόστος και κατά συνέπεια υψηλή πολυπλοκότητα του Stage #3. Με στόχο την μείωση του όγκου των υπολογισμών και των δεδομένων που πρέπει να μετακινούνται στο γράφο, εφαρμόστηκε σημείο αποκοπής στους υπολογισμούς, οδηγώντας σε σημαντική μείωση του χρόνου εκτέλεσης του αλγορίθμου.

Τέλος, πραγματοποιήθηκε σύγκριση στα αποτελέσματα των δυο αλγορίθμων, με στόχο να διαπιστωθεί πιθανή ποιοτική διαφοροποίηση. Μέσω της οπτικοποίησης και χρήσης αναλυτικών εργαλείων, διαπιστώθηκε μεγάλη ομοιότητα μεταξύ των αποτελεσμάτων του αλγορίθμου που εφαρμόζεται cut-off στα 12 hops και του αλγορίθμου χωρίς εφαρμογή cut-off. Υψηλή ήταν και η τιμή του συντελεστή συσχέτισης κατάταξης για τις top-100 ακμές σε επίπεδο temporal partition. Καταλήγουμε συνεπώς στο συμπέρασμα ότι ο αλγόριθμος με σημείο αποκοπής στα 12 hops οδηγεί στην αναγνώριση των ίδιων ακμών

ως hotspot κυκλοφοριακής συμφόρησης, ενώ παράλληλα έχει μικρότερο υπολογιστικό κόστος.

7.2 ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΠΕΡΑΙΤΕΡΩ ΕΡΕΥΝΑ

Παρόμοιες ή συμπληρωματικές έρευνες είναι δυνατόν να πραγματοποιηθούν με αντικείμενο την ανάδειξη Hotspots κυκλοφοριακής συμφόρησης στο οδικό δίκτυο, κάνοντας χρήση μεγάλων δεδομένων τροχιάς.

Ιδιαίτερο ενδιαφέρον θα είχε η ενσωμάτωση επιπλέον παραμέτρων στην ανάλυση που σχετίζονται άμεσα με το φαινόμενο της κυκλοφοριακής συμφόρησης, όπως η πυκνότητα της εμπορικής δραστηριότητας, η επικρατέστερη χρήση γης ανά οδικό τμήμα, καθώς και μεταβλητές που περιγράφουν το μεταφορικό σύστημα, όπως η ύπαρξη ή όχι φωτεινής σηματοδότησης.

Με στόχο την αύξηση της ακρίβειας των αποτελεσμάτων και την βέλτιστη αξιοποίηση του διαθέσιμου συνόλου δεδομένων, σκόπιμη θα ήταν η υιοθέτηση μιας βέλτιστης μεθόδου map-matching, καθώς και ο ακριβέστερος υπολογισμός της σχέσης γειτνίασης μεταξύ των ακμών. Αντί για την μετρική των hops, ο χωρικός διαχωρισμός μεταξύ δυο ακμών θα μπορούσε να προκύψει με βάση την πραγματική τους απόσταση πάνω στο γράφο.

Ιδιαίτερο ενδιαφέρον θα είχε η υλοποίηση ενός πιο σύνθετου partition στον άξονα του χρόνου, που θα επηρεάσει και τον ορισμό της σχέσης γειτνίασης. Ειδικότερα, θα μπορούσαμε να θεωρήσουμε ότι δυο παρατηρήσεις είναι γειτονικές στον άξονα του χρόνου εφόσον ανήκουν στην ίδια μέρα της βδομάδας και απέχουν μεταξύ τους n-temporal partitions σε επίπεδο ώρας.

Τέλος, με στόχο τη μείωση του υπολογιστικού κόστους του Stage #3, σκόπιμος κρίνεται ο περιορισμός των αναζητήσεων μέσα στο γράφο για τον υπολογισμό των επιμέρους παραμέτρων του δείκτη Gi*. Ένας τρόπος για να επιτευχθεί αυτό θα ήταν η ανάθεση κατά το Stage #1 σε κάθε ακμή πέρα από το temporal partitioning της και μια λίστα με τις n γειτονικές της ακμές και την μεταξύ τους απόσταση. Στην τιμή της μεταβλητής n_neigh θα μπορούσε να ανατεθεί η τιμή

12, καθώς τα αποτελέσματα της hotspot ανάλυσης είναι σχεδόν όμοια, όταν χρησιμοποιείται ολόκληρο το ego graph κάθε ακμής και όταν χρησιμοποιούνται μόνο οι γειτονικές ακμές που απέχουν κατά μέγιστο 12 hops από την εξεταζόμενη ακμή.

8 ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] P. Nikitopoulos, A. Paraskevopoulos, C. Doulkeridis, N. Pelekis and Y. Theodoridis. BigCAB: Distributed hot spot analysis over big spatio-temporal data using Apache Spark. In Proceedings of the 24th ACM International Conference on Advances in Geographic Information Systems. SIGSPATIAL, 2016.

[2] E. Eftelioglu, X. Tang, and S. Shekhar. Geographically robust hotspot detection: A summary of results. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pages 1447-1456. IEEE, 2015.

[3] E. Eftelioglu, S. Shekhar, et al. Ring-shaped hotspot detection: a summary of results. In 2014 IEEE International Conference on Data Mining, pages 815-820. IEEE, 2014.

[4] X. Tang, E. Eftelioglu, S. Shekhar. Elliptical hotspot detection: a summary of results. In Proceedings of the 4th International ACM SIGSPATIAL Workshop on Analytics for Big Geospatial Data, pages. 15-24. ACM, 2015

[5] Y. Xie and S. Shekhar. 2019. Significant DBSCAN towards Statistically Robust Clustering. In Proceedings of the 16th International Symposium on Spatial and Temporal Databases (SSTD ’19). Association for Computing Machinery, New York, NY, USA, 31–40. DOI:https://doi.org/10.1145/3340964.3340968

[6] P. Kuo, X. Zeng, D. Lord. 2011. Guidelines for Choosing Hot-Spot Analysis Tools Based on Data Characteristics, Network Restrictions, and Time Distributions, Submitted for Presentation at the 91st Annual Meeting of the Transportation Research Board, pp. 12-3788

[7] X. Tang, E. Eftelioglu, D. Oliver and S. Shekhar, "Significant Linear Hotspot Discovery," in IEEE Transactions on Big Data, vol. 3, no. 2, pp. 140-153, 1 June 2017.

[8] X. Tang, E. Eftelioglu, S. Shekhar. Detecting Isodistance Hotspots on Spatial Networks: A Summary of Results. Pages 281-299. SSTD, 2017.

[9] X. Li, J. Han, J. Lee, and H. Gonzalez. Traffic density-based discovery of hot routes in road networks. In Advances in Spatial and Temporal Databases, 10th International Symposium. SSTD,2007

[10] M. Häsner, C. Junghans, C. Sengstock, M. Gertz. Online Hot Spot Prediction in Road Networks. pages. 187-206. BTW ,2011.

[11] D. Sacharidis, K. Patroumpas, M. Terrovitis, V. Kantere, M. Potamias, K. Mouratidis and T. K. Sellis. On-line discovery of hot motion paths. In Proceedings of the 11th International Conference on Extending Database Technology. pages. 392-403, EDBT, 2008.

[12] P. Nikitopoulos, A. Paraskevopoulos, C. Doulkeridis, N. Pelekis and Y. Theodoridis. Hot Spot Analysis over Big Trajectory Data. 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 761-770.

[13] Y. Qiao, Y. Cheng, J. Yang, J. Liu, N. Kato (2017). A Mobility Analytical Framework for Big Mobile Data in Densely Populated Area. IEEE Transactions on vehicular technology, vol. 66, No 2, pp. 1443-1455

[14] N. Zygouras and D. Gunopulos. 2017. Discovering Corridors From GPS Trajectories. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '17). Association for Computing Machinery, New York, NY, USA, Article 61, 1–4. DOI:https://doi.org/10.1145/3139958.3139994

[15] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: A partition-and-group framework. In ACM International Conference On Management of Data (SIGMOD), pages 593-604. SIGMOD, 2007.

[16] Benjamin Krogh, Ove Andersen, and Kristian Torp. Trajectories for novel and detailed traffic information. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS '12).

[17] Shim J, Hwang C (2018) Kernel-based geographically and temporally weighted autoregressive model for house price estimation. PLoS ONE 13(10): e0205063. https://doi.org/ 10.1371/journal.pone.0205063

[18] Ord, J. K., and A. Getis. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographic Analysis, Vol. 27, No.4, 1995, pp. 286–306.

[19] Getis, A., and J. K. Ord. The Analysis of Spatial Association by Use of Distance Statistics. Geographic Analysis, Vol. 24, No. 3, 1992, pp. 189–206.

[20] Φώτης Γ. (2009), Ποσοτική Χωρική Ανάλυση, 91-161, Εκδόσεις Γκοβόστης, Αθήνα

[21] Σταθόπουλος Α. και Καρλαύτης Μ. (2016), Τα Συστήματα Μεταφορών - Δραστηριοτήτων, Σχεδιασμός Μεταφορικών Συστημάτων, 14-25, Εκδόσεις Παπασωτηρίου, Αθήνα