ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Διαχείριση Δεδομένων Τροχιών Κινούμενων Αντικειμένων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΗΛΙΑ Κ. ΦΡΕΝΤΖΟΥ

Διπλ. Πολιτικού Μηχανικού Ε.Μ.Π. (1997)
ΜΔΕ στη Γεωπληροφορική, Ε.Μ.Π. (2002)

Αθήνα, Ιούλιος 2008

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Συμβουλευτική Επιτροπή:

Επιβλέπων:

Ιωάννης Θεοδωρίδης
Επ. Καθηγητής Πανεπιστημίου Πειραιώς

Μέλη:

Γεώργιος Βασιλακόπουλος
Καθηγητής Πανεπιστημίου Πειραιώς

Τιμολέων Σελλής
Καθηγητής Ε.Μ.Πολυτεχνείου

Πανεπιστήμιο Πειραιώς

Τμήμα Πληροφορικής

Διατριβή

για την απόκτηση
Διδακτορικού Διπλώματος
του Τμήματος Πληροφορικής

ΗΛΙΑ Κ. ΦΡΕΝΤΖΟΥ

"Διαχείριση Δεδομένων Τροχιών Κινούμενων Αντικειμένων"

Εξεταστική επιτροπή:

Νικόλαος Αλεξανδρής
Καθηγητής Πανεπιστημίου Πειραιώς

Θεμιστοκλής Παναγιωτόπουλος
Καθηγητής Πανεπιστημίου Πειραιώς

Εμμανουήλ Στεφανάκης
Επ. Καθηγητής Χαροκόπειου Πανεπιστημίου

Χαράλαμπος Κωνσταντόπουλος
Λέκτορας Πανεπιστημίου Πειραιώς

.....

ΗΛΙΑΣ Κ. ΦΡΕΝΤΖΟΣ

Πολιτικός Μηχανικός Ε.Μ.Π.

Πρόλογος

Ο τομέας των Βάσεων Δεδομένων Κινούμενων Αντικειμένων (Moving Object Databases – MODs) είναι ένας σημαντικός χώρος έρευνας που έχει προσελκύσει ιδιαίτερο ενδιαφέρον την τελευταία δεκαετία. Οι MODs είναι ένας αναδυόμενος τεχνολογικός τομέας λόγω της ανάπτυξης των πανταχού παρόντων συσκευών εντοπισμού θέσης, όπως τα PDAs, τα κινητά τηλέφωνα κλπ. όπως επίσης και λόγω της ποικιλομορφίας των πληροφοριών που μπορούμε να εξάγουμε από αυτές. Από την άλλη πάλι, η ανάπτυξη μηχανισμών που επιτρέπουν στις MODs να υποστηρίζουν αποτελεσματικά τα πολύπλοκα δεδομένα τροχιών που παράγονται από κινούμενα αντικείμενα, εμπλέκει διάφορα φυσικά θέματα της τεχνολογίας βάσεων δεδομένων, όπως για παράδειγμα τη δεικτοδότηση, την προηγμένη επεξεργασία επερωτήσεων καθώς και τεχνικές βελτιστοποίησής τους.

Η πρόκληση λοιπόν που αντιμετωπίζεται σε αυτή τη διατριβή είναι η παροχή αποτελεσματικών μηχανισμών που επιτρέπουν στις MODs να αποθηκεύσουν και να διαχειριστούν δεδομένα τροχιών κινούμενων αντικειμένων με αποτελεσματικό τρόπο. Προκειμένου να επιτευχθεί αυτός ο στόχος, αναπτύσσουμε μια σειρά από μεθόδους προσπέλασης και συγκεκριμένες προηγμένες τεχνικές επεξεργασίας επερωτήσεων που στη συνέχεια υλοποιούνται σε πρωτότυπα συστήματα για να αποδειχθεί η αποτελεσματικότητά τους. Ακολουθώντας τις προτάσεις της διατριβής, υφιστάμενα ευρετήρια τροχιών κινούμενων αντικειμένων μπορούν να απαντήσουν μία ευρεία γκάμα απλών αλλά και προηγμένων επερωτήσεων. Πέρα από αυτό, εφαρμόζοντας υφιστάμενα μοντέλα αναπαράστασης της αβεβαιότητας σε τροχιές κινούμενων αντικειμένων, προτείνουμε ένα μοντέλο πρόβλεψης της επίδρασής της σε χωροχρονικές ερωτήσεις. Τα αποτελέσματα μας σε αυτό το θέμα μπορούν να εφαρμοστούν απευθείας σε βάσεις και αποθήκες χωροχρονικών και χωρικών δεδομένων καθώς και για τη βελτιστοποίηση επερωτήσεων που τίθενται σε κατανεμημένα συστήματα με ασάφεια. Τέλος, παρέχουμε ένα μοντέλο για την πρόβλεψη της επίδρασης των τεχνικών συμπίεσης επάνω στις επερωτήσεις που τίθενται σε συμπιεσμένα δεδομένα τροχιών. Το μοντέλο αποκαλύπτει ενδιαφέρουσες πτυχές της κατανομής του σφάλματος, που μπορούν να οδηγήσουν στην ανάπτυξη πιο αποδοτικών αλγορίθμων συμπίεσης, ενώ μπορεί να χρησιμοποιηθεί και σαν ένα επιπλέον κριτήριο απόφασης για τους χρήστες των MODs σχετικά με την καταλληλότητα των συμπιεσμένων δεδομένων.

Ιούλιος 2008 Ηλίας Φρέντζος

Ευχαριστίες

Θα ήθελα πρώτα απ' όλα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου Γιάννη Θεοδωρίδη για τη συνεχή υποστήριξη που μου παρείχε κατά τη διάρκεια της εκπόνησης της διατριβής στο Πανεπιστήμιο Πειραιά. Οι γνώσεις του, οι συμβουλές του και η συνεχής παρουσία του δίπλα μου ήταν πολύτιμα στοιχεία για την εκπόνηση της διατριβής αυτής. Ευχαριστώ ακόμη τον καθ. Τίμο Σελλή ο οποίος μου κέντρισε αρχικά το ενδιαφέρον για τον επιστημονικό χώρο τον οποίο πραγματεύεται η διατριβή, καθώς και τους καθ. Γεώργιο Βασιλακόπουλο, Νικόλαο Αλεξανδρή, Θεμιστοκλή Παναγιωτόπουλο, Εμμανουήλ Στεφανάκη και Χαράλαμπο Κωνσταντόπουλο που δέχτηκαν ο πρώτος να είναι μέλος της συμβουλευτικής επιτροπής και οι υπόλοιποι να είναι μέλη της επταμελούς εξεταστικής επιτροπής μου.

Επίσης θα ήθελα να ευχαριστήσω όλους τους συναδέλφους μου στο Εργαστήριο Βάσεων Δεδομένων και ιδιαίτερα, δε, τον Κώστα Γρατσία, για την σημαντική συνεργασία και τις συζητήσεις που είχαμε επάνω σε ιδέες που αποτελούν τη βάση της διατριβής αυτής· αρκετές από τις ιδέες αυτές ανήκουν, σε ένα μεγάλο ποσοστό και σε αυτούς.

Κατά τη διάρκεια εκπόνησης της διδακτορικής μου διατριβής συμμετείχα σε δύο ερευνητικά προγράμματα της Ευρωπαϊκή Ένωσης (πρόγραμμα GeoPKDD) και της Γ.Γ.Ε.Τ. (πρόγραμμα ΔΙΑΧΩΡΟΝ), από τα οποία είχα και οικονομική υποστήριξη. Ευχαριστώ τους υπευθύνους των παραπάνω προγραμμάτων.

Θα ήθελα επίσης να ευχαριστήσω τους γονείς μου για την υποστήριξη που μου παρείχαν στα χρόνια των σπουδών μου, καθώς και τον θείο μου Ευθύμιο Σταμπουλόγλου, για τον τρόπο που με δίδαξε να προσεγγίζω την επιστήμη. Τέλος θα ήθελα να ευχαριστήσω τη γυναίκα μου Μαρία και τις δύο μου κόρες στις οποίες και αφιερώνω την εργασία αυτή, για τη συνεχή συμπαράσταση και την ανεξάντλητη υπομονή που έδειξαν όλα αυτά τα χρόνια της εκπόνησής της.

Πίνακας Περιεχομένων

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

1. Εισαγωγή

Το παρόν κεφάλαιο παρουσιάζει το υπόβαθρο της διατριβής και δίνει μία σύντομη περιγραφή της δομής της. Στην Ενότητα 1.1 εισάγονται κάποιες βασικές γνώσεις περί τροχιών και εξηγείται ο σκοπός της διατριβής. Η Ενότητα 1.2 παρουσιάζει την έννοια της τροχιάς, που είναι και το γενικότερο θέμα της εργασίας μας. Στην Ενότητα 1.3 καθορίζονται τα προβλήματα που θα εξετάσουμε, ενώ η Ενότητα 1.4 σκιαγραφεί τη συμβολή της διατριβής στον χώρο των βάσεων δεδομένων. Η Ενότητα 1.5 δίνει μία εικόνα των δεδομένων που χρησιμοποιούνται και η Ενότητα 1.6 κλείνει το κεφάλαιο με το περίγραμμα τις διατριβής.

1.1. Βάσεις Δεδομένων Τροχιών

Ο τομέας των Βάσεων Δεδομένων Κινούμενων Αντικειμένων (Moving Object Databases - MODs) είναι ένας σημαντικός χώρος έρευνας που έχει προσελκύσει ιδιαίτερο ενδιαφέρον την τελευταία δεκαετία. Στόχος των βάσεων δεδομένων κινούμενων αντικειμένων είναι η επέκταση της τεχνολογίας βάσεων δεδομένων προκειμένου να υποστηρίξουν την αναπαράσταση και τις επερωτήσεις κινούμενων αντικείμενων και της τροχιάς τους. Οι MODs είναι πλέον ένας αναδυόμενος τεχνολογικός τομέας λόγω της ανάπτυξης των πανταχού παρόντων συσκευών εντοπισμού θέσης, όπως τα PDAs, τα κινητά τηλέφωνα κλπ. όπως επίσης και της ποικιλομορφίας των πληροφοριών που μπορούμε να εξάγουμε από τέτοιες βάσεις δεδομένων. Επί του παρόντος, την παρουσία των MODs μπορούμε να την εκμεταλλευτούμε σε μια σειρά από διαδικασίες υποστήριξης αποφάσεων όπως για παράδειγμα εκτίμηση και πρόβλεψη κυκλοφοριακής κίνησης, ανάλυση συνθηκών κυκλοφοριακής συμφόρησης, συστήματα διαχείρισης στόλου, πεδίων μάχης και ανάλυση μεταναστευτικών συνηθειών ζώων [GS05].

Από ιστορικής πλευράς, υπάρχει η εξής συστηματική κατάταξη στη βιβλιογραφία των χωροχρονικών βάσεων δεδομένων: (a) εργασίες σχετικά με τις παρούσες και μελλοντικές θέσεις κινούμενων αντικειμένων, όπως οι [SJLL00], [BJKS02], [MXA04] και (b) εργασίες για τις προηγούμενες θέσεις αντικειμένων, που θέτουν ιστορικές επερωτήσεις όπως οι [TVS96], [PJT00]. Η τελευταία κατηγορία, μπορεί επίσης να ταξινομηθεί σε 2 επιπλέον κατηγορίες: (a) προσεγγίσεις που μοντελοποιούν και διαχειρίζονται χωρικά δεδομένα που αλλάζουν σε διακριτές χρονικές στιγμές, με παραδείγματα που περιλαμβάνουν διαχείριση πολυμέσων [TVS96] απλών χωρικών [NST99], και πιο σύνθετων δεδομένων με χωρική αναφορά, όπως εφαρμογές κτηματολογικών δεδομένων [ACNV99], και (b) προσεγγίσεις που ασχολούνται με δεδομένα που αλλάζουν συνεχώς θέση συν τω χρόνω [GBE+00], [PJT00] · στην τελευταία κατηγορία ανήκει και η παρούσα διδακτορική διατριβή.

Τα κινούμενα αντικείμενα είναι γεωμετρίες· μπορεί να είναι σημεία, γραμμές, επιφάνειες ή όγκοι χρονικά μεταβαλλόμενοι, ενώ τροχιά είναι η περιγραφή της κίνησης των εν λόγω αντικειμένων. Καθώς ο γεωγραφικός χώρος αυτός καθ' εαυτός είναι συνεχής, η φυσική κίνηση περιγράφεται από μία συνεχή αλλαγή θέσης, ήτοι μία συνάρτηση από το χρόνο στο γεωγραφικό χώρο. Η κίνηση υποδηλώνει και μία χρονική συνιστώσα δεδομένου ότι δεν μπορούμε να αντιληφθούμε την κίνηση παρά μόνον μέσω σύγκρισης δύο διαφορετικών χρονικών στιγμών. Συνεπώς, μπορούμε αντιστοίχως να ορίσουμε μία τροχιά ως την καταγραφή ενός χρονικά μεταβαλλόμενου χωρικού φαινομένου.

Βάσει της προηγούμενης συζήτησης μία ιστορική τροχιά μπορεί να οριστεί απλώς και μόνον ως μία συνάρτηση από το χρόνο προς ένα γεωγραφικό χώρο· από την άλλη, η περιγραφή, η αναπαράσταση και η αντιμετώπισή της είναι πολύ πιο περίπλοκα θέματα. Πράγματι, από πλευράς εφαρμογής, μία τροχιά είναι η καταγραφή της κίνησης κάποιου αντικειμένου, δηλαδή η καταγραφή των θέσεων του αντικειμένου σε συγκεκριμένες χρονικές στιγμές. Έτσι, ενώ λογικά φανταζόμαστε μία καλοσχηματισμένη καμπύλη ως αναπαράσταση της τροχιάς ενός αντικειμένου, στην πραγματικότητα η τροχιά πρέπει να δημιουργηθεί από μία σειρά σημείων, δηλαδή από τις δειγματοληπτούμενες θέσεις του αντικειμένου· έτσι η καμπύλη της τροχιάς προκύπτει εφαρμόζοντας μεθόδους παρεμβολής (interpolation) στα σημεία δειγματοληψίας. Ωστόσο, οποιαδήποτε μέθοδο παρεμβολής και να χρησιμοποιήσουμε, η καμπύλη που προκύπτει είναι απλώς μία δυνητική αναπαράσταση της πραγματικής τροχιάς· γεγονός που επιδεινώνεται αν λάβουμε υπόψη και τα πιθανά σφάλματα που είναι αναπόφευκτα κατά την αρχική καταγραφή των θέσεων του αντικειμένου. Υπάρχει συνεπώς μία εγγενής αβεβαιότητα στις τροχιές. Στη βιβλιογραφία έχει προταθεί ένα πλήθος μοντέλων [TWHC04], [TWZC02], [PJ99], για τη διαμόρφωση και την κατάλληλη αντιμετώπιση αυτής της αβεβαιότητας.

Επιπλέον, δεδομένου ότι οι τροχιές πρέπει να είναι ένα μοντέλο πρώτης τάξης και όχι δεδομένα που προκύπτουν από υπολογισμούς, η έννοιά τους εισήχθη σε κάποιες αρχικές εργασίες [CR99], [EGSV99], [FGNS00], οι οποίες αντιμετώπιζαν την ανάγκη να αντιληφθούμε και να μοντελοποιήσουμε το πλήρες ιστορικό της κίνησης των αντικειμένων. Αξιολογώντας το γεγονός ότι τα δεδομένα θέσης μπορεί να μεταβάλλονται χρονικά, η αντίστοιχη βάση δεδομένων πρέπει να περιλαμβάνει το όλο ιστορικό αυτής της εξέλιξης· και το Σύστημα Διαχείρισης Βάσης Δεδομένων (Database Management System – DBMS) θα πρέπει να έχει τη δυνατότητα να ανατρέξει στο παρελθόν σε οποιοδήποτε χρονικό αποτύπωμα και να ανακαλέσει τις πληροφορίες της βάσης δεδομένων σ' εκείνη τη χρονική στιγμή. Πιο συγκεκριμένα, βάσει της [GBE+00] τα κινούμενα σημεία (*mpoints*) και οι κινούμενες περιοχές (*mregions*) περιγράφονται ως 3D (2D χώρος + χρόνος) ή περισσότερων διαστάσεων οντότητες η δομή και η συμπεριφορά των οποίων γίνεται αντιληπτή αναπαριστώντας τα ως αφηρημένους τύπους δεδομένων (abstract data types). Τέτοιοι τύποι και οι λειτουργίες τους για χωρικές τιμές χρονικά μεταβαλλόμενες μπορούν να ενσωματωθούν ως βασικοί τύπων δεδομένων (attribute) σε ένα DBMS. Η [GBE+00] εισήγαγε ένα κατασκευαστή τύπου τ που μεταμορφώνει οποιοδήποτε ατομικό τύπο δεδομένων σε ένα τύπο $\tau(a)$ με σημαντική $\tau(a) = time \rightarrow a$. Μ' αυτόν τον τρόπο, οι δύο προαναφερθέντες βασικοί τύποι, ήτοι το *mpoint* και η *mregion*, μπορούν ν' αναπαρασταθούν ως $\tau(point)$ και $\tau(region)$, αντίστοιχα. Η [GBE+00] παρέχει επίσης και μία άλγεβρα με τύπους δεδομένων (όπως ένα κινούμενο σημείο, μία κινούμενη περιοχή, ένα κινούμενο αντικείμενο κλπ.) καθώς και ένα σύνολο λειτουργιών, που υποστηρίζουν μια σειρά από επερωτήσεις για δεδομένα

χωροχρονικών τροχιών. Η υλοποίηση τέτοιων μοντέλων που προτείνονται στη βιβλιογραφία, καθώς και ενσωμάτωσή της αντίστοιχης λειτουργικότητα σε συγκεκριμένες τεχνικές λύσεις έχει ως τελικό αποτέλεσμα τη δημιουργία *μηχανών* βάσεων δεδομένων κινούμενων αντικειμένων. Στη βιβλιογραφία, μπορούμε να βρούμε τουλάχιστον δύο τέτοιες μηχανές MOD οι οποίες έχουν αναπτυχθεί για την υλοποίηση του προτεινόμενου μοντέλου από τους Gutting et al. στη [GBE+00], πιο συγκεκριμένα το πρωτότυπο SECONDO [AGB06] και τη μηχανή HERMES [PT06], [PTVP06].

Από την άλλη πάλι, η ανάπτυξη τέτοιων μηχανών περιλαμβάνει και φυσικές πτυχές της τεχνολογίας βάσεων δεδομένων, όπως για παράδειγμα δεικτοδότηση, προηγμένη επεξεργασία επερωτήσεων και τεχνικές βελτιστοποίησης επερωτήσεων. Η πρόκληση λοιπόν που έχουμε δεχτεί σε αυτή τη διατριβή είναι η ανάπτυξη αποτελεσματικών μηχανισμών που επιτρέπουν στις Μηχανές Βάσεων Δεδομένων Κινούμενων Αντικειμένων (Moving Object Database Engines) να αποθηκεύσουν και να κάνουν επερωτήσεις σε δεδομένα τροχιών με αποτελεσματικό τρόπο. Προκειμένου να επιτύχουμε το στόχο μας στα πλαίσια της διδακτορικής διατριβής, αναπτύσσουμε μια σειρά από μεθόδους προσπέλασης και συγκεκριμένες προηγμένες τεχνικές επεξεργασίας επερωτήσεων που στη συνέχεια εφαρμόζονται για να αποδειχθεί η αποτελεσματικότητά τους. Όλες αυτές οι μέθοδοι αρχικά υλοποιούνται ως πρωτότυπα σε ανεξάρτητα περιβάλλοντα ανάπτυξης, ενώ η μεταφορά τους σε εμπορικά DBMS παραμένει ως μελλοντικός στόχος· ωστόσο ένας αριθμός από τις προτεινόμενες σε αυτή τη διατριβή τεχνικές, έχει ήδη υλοποιηθεί στο Αντικειμενο-Σχεσιακό Σύστημα Διαχείρισης Βάσης Δεδομένων (Object – Relational DBMS) ORACLE και ενσωματωθεί στη μηχανή HERMES [PFGT08], καθώς και στην PostgreSQL [Post08b] σε συνδυασμό με τη χωρική της επέκταση PostGIS [Post08a].

Εν συντομία τα θέματα που θα συζητηθούν στα πλαίσια της διατριβής που είναι φυσικά θέματα μιας μηχανής MOD, περιλαμβάνουν τεχνικές δεικτοδότησης για τροχιές κινούμενων αντικειμένων, τεχνικές προηγμένης; επεξεργασίας επερωτήσεων, μοντέλα για επερωτήσεις παρουσία αβεβαιότητας και τέλος θέματα συμπίεσης τροχιών.

1.2. Βασικές Έννοιες περί Τροχιών

Σε γενικές γραμμές οι χωροχρονικές τροχιές διαιρούνται σε δύο μεγάλες κατηγορίες, ανάλογα με τη φύση του σχετικού χωρικού αντικειμένου: (i) αντικείμενα χωρίς επιφάνεια που αναπαρίστανται ως κινούμενα σημεία, και (ii) αντικείμενα με επιφάνεια, που αναπαρίστανται ως κινούμενες περιοχές· σε αυτή την περίπτωση η έκταση της περιοχής μπορεί επίσης να είναι χρονικά μεταβαλλόμενη. Η πρώτη εκ των δύο παραπάνω κατηγοριών έχει προσελκύσει τη μερίδα του λέοντος του ερευνητικού ενδιαφέροντος, δεδομένου ότι η πλειονότητα των εφαρμογών που περιλαμβάνουν χωροχρονικές τροχιές στον πραγματικό κόσμο θεωρούν αντικείμενα που αναπαρίστανται ως σημεία π.χ. συστήματα διαχείρισης στόλων που παρακολουθούν οχήματα σε οδικά δίκτυα. Γι' αυτό και η διδακτορική διατριβή θα επικεντρωθεί στην πρώτη κατηγορία και από τούδε και στο εξής θα περιοριστούμε σε τροχιές κινουμένων σημείων.

Υπ' αυτό το πρίσμα, μία τροχιά μπορεί να οριστεί ως μία συνάρτηση από το χρονικό $I \subseteq \mathbb{R}$ πεδίο στο γεωγραφικό χώρο $\mathbb{R}^2$, ήτοι στο 2D επίπεδο. Πιο συγκεκριμένα, μία τροχιά *T είναι μία συνεχής αναπαράσταση από το χρονικό* $I \subseteq \mathbb{R}$ στο χωρικό πεδίο ($\mathbb{R}^2$, το 2D επίπεδο):

$$I \subseteq \mathbb{R} \to \mathbb{R}^2 : t \mapsto a(t) = \left(a_x(t), a_y(t)\right), \tag{1.1}$$

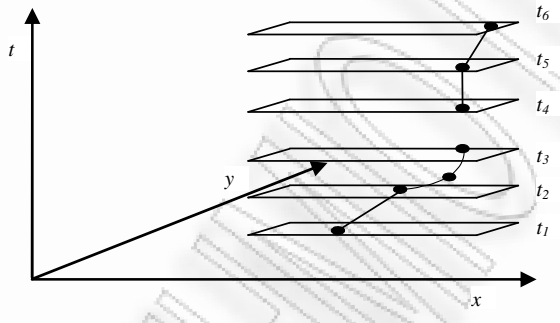
και,

$$T = \left\{\left(a_x(t), a_y(t), t\right) \mid t \in I\right\} \subset \mathbb{R}^2 \times \mathbb{R} \tag{1.2}$$

Αφ'έτερου, από πλευράς εφαρμογής, μία τροχιά είναι η καταγραφή της κίνησης ενός αντικειμένου, ήτοι η καταγραφή των θέσεων ενός αντικειμένου σε συγκεκριμένα χρονικά αποτυπώματα· ενώ η πραγματική τροχιά αποτελείται από μία καμπύλη, οι απαιτήσεις στον πραγματικό κόσμο υποδηλώνουν ότι η τροχιά πρέπει να δημιουργηθεί από μία σειρά από δειγματοληπτούμενα σημεία, δηλαδή τις θέσεις του αντικειμένου σε συγκεκριμένα χρονικά αποτυπώματα. Έτσι, οι τροχιές κινούμενων σημείων συχνά ορίζονται ως ακολουθίες τριάδων (x, y, t):

$$T = \left\{\left(x_1, y_1, t_1\right), \left(x_2, y_2, t_2\right), ..., \left(x_n, y_n, t_n\right)\right\}, \tag{1.3}$$

όπου $x_i, y_i, t_i \in \mathbb{R}$, και $t_1 < t_2 < .. < t_n$, και η καμπύλη της τροχιάς δημιουργείται κατά προσέγγιση εφαρμόζοντας χωροχρονικές μεθόδους παρεμβολής στο σύνολο των τυχαίων σημείων (Σχήμα 1.1).

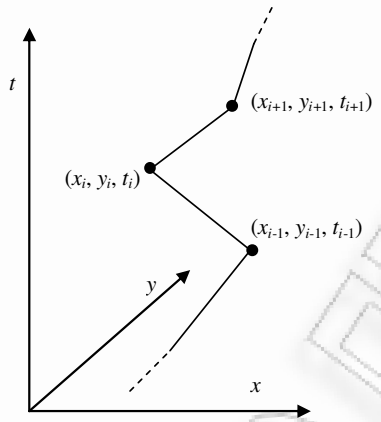


Σχήμα 1.1: Η χωροχρονική τροχιά ενός κινούμενου αντικειμένου: οι τελείες αντιπροσωπεύουν τις τυχαίες θέσεις και οι γραμμές ενδιάμεσα τις εναλλακτικές τεχνικές παρεμβολής (γραμμική σε αντιδιαστολή με την παρεμβολή τόξου). Σε μία τροχιά μπορούμε επίσης να έχουμε και άγνωστη μορφή κίνησης (δες χρονικό διάστημα [t_3, t_4])

Ο πρώτος και σημαντικότερος περιορισμός που θέτουν οι μέθοδοι χωροχρονικής παρεμβολής αυτής της μορφής, είναι ότι μία τροχιά που συνδέεται μ' ένα δείγμα δεδομένων θα πρέπει να περιλαμβάνει τα αρχικά σημεία. Δηλαδή για όλα τα σημεία $\left(x_i, y_i, t_i\right)$ στο δείγμα ισχύει $\left(x_i, y_i, t_i\right) = \left(a_x(t_i), a_y(t_i), t_i\right)$. Δεύτερον, αφ' ης στιγμής δοθεί ένα δείγμα δεδομένων, υπάρχουν άπειρες τροχιές που συνδέονται με το συγκεκριμένο δείγμα δεδομένων γεγονός που υποδηλώνει ότι η τροχιά δεν είναι σε καμία περίπτωση μοναδική. Η εύρεση της κατάλληλης καμπύλης που θα συνδέει τα αρχικά σημεία στα οποία έγινε η δειγματοληψία της θέσης του αντικειμένου ονομάζεται παρεμβολή. Η παρεμβολή φυσικά έχει και τα προβλήματά της. Εμείς θέλουμε να είναι ταχύτατη, εύκολη στη χρήση, ευέλικτη και ακριβής. Δυστυχώς αν βελτιώσουμε μία ιδιότητα δεν βελτιώνονται απαραίτητα όλες. Η γραμμική παρεμβολή είναι η πιο γρήγορη και πιο εύκολη απ' όλες (Σχήμα 1.2). Η ιδέα είναι να ενώσουμε τα αρχικά σημεία με ευθείες γραμμές· η γραμμικότητα εκφράζεται από το γεγονός ότι ίσα χρονικά άλματα (μεταξύ δύο σημείων) οδηγούν σε ίσα άλματα στο χώρο. Επί παραδείγματι, το τμήμα μεταξύ των σημείων $\left(x_i, y_i, t_i\right)$ και $\left(x_{i+1}, y_{i+1}, t_{i+1}\right)$ δίνεται από το

$$\left(x, y, t\right) = \left(x_i, y_i, t_i\right) + \frac{t - t_i}{t_{i+1} - t_i}\left(x_{i+1} - x_i, y_{i+1} - y_i, t_{i+1} - t_i\right), \text{ και } t_i \le t \le t_{i+1}, \tag{1.4}$$

που είναι ένα ευθύγραμμο τμήμα στο $\mathbb{R}^2 \times \mathbb{R}$ που παραμετροποιείται από το $t \in [t_i, t_{i+1}]$. Τέλος, η τροχιά αποτελείται από την αλληλουχία όλων αυτών των τμημάτων. Επομένως, μία τροχιά μπορεί επίσης να αναπαρασταθεί από μία συλλογή *n*-1 ευθύγραμμων τμημάτων $T = \{L_1, L_2, ..., L_{n-1}\}$ με $L_i = \{(x_i, y_i, t_i), (x_{i+1}, y_{i+1}, t_{i+1})\}$



Σχήμα 1.2: Γραμμική παρεμβολή σε τροχιές κινούμενων αντικειμένων

Η γραμμική παρεμβολή αυτής της μορφής δεν είναι και τόσο αθώα· στην πορεία έχουν γίνει κάποιες υποθέσεις. Η πρώτη είναι ότι το κινούμενο αντικείμενο διατηρεί σταθερή ταχύτητα και κατεύθυνση μεταξύ των αρχικών σημείων. Δεύτερον, οι αλλαγές ταχύτητας και κατεύθυνσης συμβαίνουν μόνο στα αρχικά σημεία ενώ είναι συχνά απότομες και ασυνεχείς. Από την άλλη μεριά, η γραμμική παρεμβολή είναι εύκολη στη δημιουργία και στο χειρισμό της και αυτός είναι και ο κύριος λόγος για τον οποίο είναι ευρύτατα διαδεδομένη στη βιβλιογραφία βάσεων δεδομένων τροχιών. Από τούδε και στο εξής, ο όρος τροχιά θα χρησιμοποιείται για την περιγραφή αυτών των τριάδων όπως στην Εξ.(1.3), εφαρμόζοντας γραμμική παρεμβολή μεταξύ τους όπως καθορίζεται στην Εξ. (1.4).

1.3. Ερευνητικά Προβλήματα και Προκλήσεις στις Βάσεις Δεδομένων Τροχιών

Μεταξύ των ποικίλων τεχνολογιών που συμμετέχουν στην ανάπτυξη των MODs για την υποστήριξη ιστορικών τροχιών κινούμενων αντικειμένων, στην παρούσα διατριβή εστιάζουμε σε μία σειρά από φυσικές πτυχές, πιο συγκεκριμένα στη δεικτοδότηση και τη προηγμένη επεξεργασία επερωτήσεων, σε αναλυτικά μοντέλα για την υποστήριξη της αβεβαιότητας και τέλος, στη συμπίεση τροχιών. Στις επόμενες ενότητες παρουσιάζουμε εν συντομία τα κύρια ερευνητικά προβλήματα και τις προκλήσεις που εμφανίζονται τις βάσεις δεδομένων τροχιών και με τις οποίες θα ασχοληθούμε σε αυτή τη διατριβή.

1.3.1. Δεικτοδότηση

Οι επερωτήσεις στις MODs θα μπορούσαν να γίνουν πολύ ακριβές λόγω της φύσης των σχετικών δεδομένων και της πολυπλοκότητα των αλγορίθμων επεξεργασίας των ερωτημάτων. Δεδομένου δε ότι οι συσκευές εύρεσης θέσης είναι πανταχού παρούσες, οι βάσεις δεδομένων τροχιών, αργά ή γρήγορα θα βρεθούν αντιμέτωπες με ένα τεράστιο όγκο δεδομένων. Συνάγουμε συνεπώς ότι η απόδοση παρουσία τεραστίου πλήθους δεδομένων, θα αποτελέσει σημαντικό πρόβλημα για τις βάσεις δεδομένων τροχιών και ο μόνος τρόπος να αντιμετωπίσουμε αυτές τις καταστάσεις είναι η

εκμετάλλευση εξειδικευμένων μεθόδων προσπέλασης που χρησιμοποιούνται για τη χωροχρονική δεικτοδότηση τροχιών.

Στον τομέα της χωροχρονικής δεικτοδότησης κυριαρχεί η παρουσία των R-δέντρων [Gut84], με τις παραλλαγές και τις προεκτάσεις τους· στην πράξη κάτι τέτοιο είναι αναμενόμενο δεδομένου ότι τα R-δέντρα είναι ιδιαίτερα δημοφιλή στις χωρικές βάσεις δεδομένων. Οι παραλλαγές και οι προεκτάσεις του R-δέντρου στο χωροχρονικό πεδίο περιλαμβάνουν, μεταξύ άλλων, 3D R-δέντρα [TVS96], TB-δέντρα και STR-δέντρα [PJT00], PA-δέντρα [NR07], MON-δέντρα [AG05], ενώ το SETI [CEP03] είναι μία υβριδική τεχνική που βασίζεται σε R-δέντρα και διαμέριση του χώρου. Επειδή το ενδιαφέρον μας στην παρούσα διατριβή επικεντρώνεται στις ιστορικές MODs, θα περιορίσουμε τη συζήτησή μας σε τεχνικές δεικτοδότησης που καταγράφουν προηγούμενες θέσεις. Ο αναγνώστης που ενδιαφέρεται για δεικτοδότηση τρεχουσών θέσεων και ανυσμάτων κίνησης μπορεί να βρει αρκετά ενδιαφέροντα πράγματα στα [SJ02], [SJLL00], [TPS03], και [XP03].

Ωστόσο, όπως επισημαίνεται στην [PJT00], η μεγάλη πλειονότητα των προτεινόμενων χωροχρονικών ευρετηρίων παραβλέπουν τις προκλήσεις που τίθενται από τη ίδια τη φύση των δεδομένων τροχιάς και απλά δεικτοδοτούν συλλογές γραμμικών τμημάτων στο χωροχρόνο, ασχολούμενα μόνο με την επεξεργασία παραδοσιακών επερωτήσεων που *βασίζονται στις συντεταγμένες (coordinate-based)* και αγνοώντας ταυτόχρονα άλλους χρήσιμους τύπους, όπως τοπολογικές επερωτήσεις και επερωτήσεις πλοήγησης, οι οποίες είναι *βασισμένες στη τροχιά* (*trajectory-based)*. Επιπλέον τα υπάρχοντα χωροχρονικά ευρετήρια που δεν διατηρούν τις τροχιές των κινούμενων αντικειμένων και ασχολούνται με τα χωροχρονικά δεδομένα μόνο ως συλλογή γραμμικών τμημάτων στο χώρο των 2+1 διαστάσεων (όπως το SETI [CEP03] και το 3D R-tree [TVS96]), παραβλέπουν την ανάγκη για διαδικασίες διαγραφής· παρά το γεγονός ότι η διαγραφή ενός γραμμικού τμήματος από μία βάση δεδομένων τροχιών μπορεί να ακούγεται ανούσια, η διαγραφή μιας ολόκληρης τροχιάς είναι μία πολύ χρήσιμη διαδικασία την οποία πρέπει να υποστηρίζει οποιοδήποτε ευρετήριο τροχιάς στον πραγματικό κόσμο. Η ίδια ανάγκη για διατήρηση τροχιάς προκύπτει όταν ασχολούμαστε και με μηχανισμούς συμπίεσης, οι οποίοι όπως θα δούμε στο επόμενο κεφάλαιο, εξ' ορισμού προϋποθέτουν ότι αντιμετωπίζουμε την κάθε τροχιά ως ανεξάρτητο αντικείμενο.

Δύο δομές ευρετηρίου που παρουσιάζονται στη [PJT00], πιο συγκεκριμένα το Χωροχρονικό R-δέντρο (STR-tree) και το Trajectory Bundle δένδρο (TB-tree), προσπαθούν να καλύψουν αυτές τις ανάγκες και να υποστηρίξουν αποτελεσματικά διαδικασίες που είναι βασισμένες στην έννοια της τροχιά. Το αποτέλεσμα αυτής της εργασίας ήταν ότι το TB-δέντρο θα μπορούσε να υποστηρίξει μη παραδοσιακές επερωτήσεις πολύ πιο αποτελεσματικά από το 3D R-δέντρο και το STR-δέντρο. Δυστυχώς, παρά τα σαφή του πλεονεκτήματα στην επεξεργασία επερωτήσεων που είναι βασισμένες στη τροχιά, το TB-δέντρο έχει ένα πολύ σημαντικό μειονέκτημα λόγω της στρατηγικής που ακολουθεί κατή την εισαγωγή: τα νέα δεδομένα τροχιάς εισάγονται πάντα στη δεξιά «άκρη» του δέντρου, που σημαίνει ότι η απόδοσή του θα εξαρτάται σε μεγάλο βαθμό από την διάταξη εισαγωγής των δεδομένων. Ωστόσο, σε πραγματικές εφαρμογές, αυτή η υπόθεση δεν είναι κατ'ανάγκη πάντα αληθής. Για παράδειγμα, έστω μία εφαρμογή που πρέπει να υποστηρίξει εισαγωγές σε πραγματικό χρόνο και μία κατάσταση όπου το κινούμενο αντικείμενο εισέρχεται σε μία περιοχή όπου το σύστημα μετάδοσης θέσης δε λειτουργεί· τότε η τροχιά του θα μπορούσε να αποθηκευτεί τοπικά στο αντικείμενο και να

μεταδοθεί στον κεντρικό εξυπηρετητή – όπου λειτουργεί το ευρετήριο – σε κατοπινό χρόνο. Εν τω μεταξύ, άλλα κινούμενα αντικείμενα θα μπορούσαν να έχουν μεταδώσει τις θέσεις τους, παραβιάζοντας την ανωτέρω υπόθεση του TB-δέντρου. Επιπλέον, η δομή του TB-δέντρου δεν είναι κατάλληλη για να υποστηρίξει λειτουργίες διαγραφής και συμπίεσης· μία διαγραφή τροχιάς θα αφήσει «κενά» στους κόμβους και η συμπίεση τροχιών όπως θα συζητήσουμε στη συνέχεια, σημαίνει ότι το ευρετήριο θα πρέπει να αντιμετωπίζει δεδομένα τα οποία εισάγονται με μη χρονολογική σειρά.

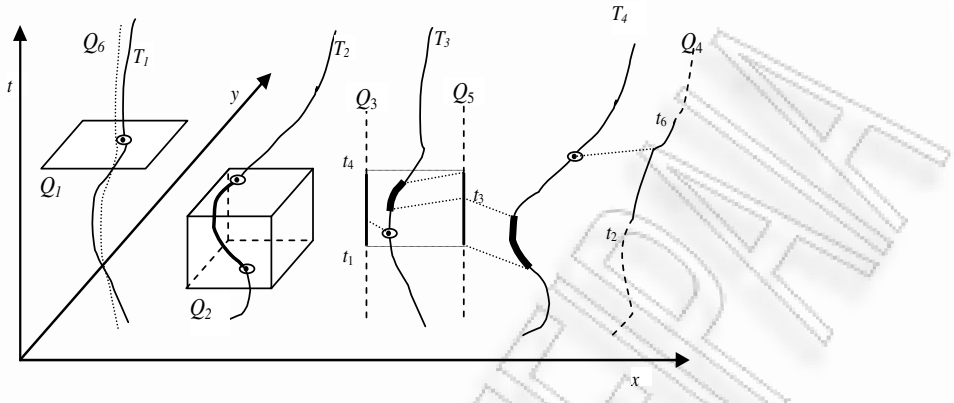
Μία ακόμη ενδιαφέρουσα προσέγγιση που αφορά τη δεικτοδότηση χωροχρονικών τροχιών, προκύπτει αναγνωρίζοντας ότι οι τροχιές κατά πάσα πιθανότητα περιορίζονται από ένα δίκτυο. Όπως επισημαίνεται στο [KGT99], η ύπαρξη περιορισμών στο χώρο στον οποίο τα αντικείμενα κινούνται είναι μία συνθήκη που μπορεί να χρησιμοποιηθεί για τη βελτίωση της απόδοσης των χωροχρονικών ευρετηρίων. Στην πράξη, αυτό συμβαίνει στις περισσότερες πραγματικές εφαρμογές: τα αεροπλάνα πετούν σε αεροδιαδρόμους, τα αυτοκίνητα και οι πεζοί κινούνται σε οδικά δίκτυα, ενώ τα τρένα έχουν σταθερές τροχιές σε σιδηροδρομικά δίκτυα. Αυτές οι ειδικές συνθήκες (περιορισμοί στην κίνηση) αποτελούν αντικείμενο ερευνητικού ενδιαφέροντος στις [KGT99], [PTKZ02].

Πιο συγκεκριμένα, βάσει της [KGT99], το πεδίο της τροχιάς ενός αντικειμένου που κινείται σε ένα δίκτυο δεν είναι ο χώρος των 2+1 διαστάσεων, αλλά μάλλον ένας χώρος με 1.5 διαστάσεις, καθώς τα γραμμικά τμήματα που περιλαμβάνουν το δίκτυο μπορούν να αποθηκευτούν σε ένα συμβατικό ευρετήριο χωρικών δεδομένων (όπως το R-δέντρο). Έτσι, η δεικτοδότηση των αντικειμένων που κινούνται σε ένα δίκτυο περιορίζεται σε πρόβλημα μονοδιάστατης δεικτοδότησης. Στην [KGT99], το πρόβλημα της δεικτοδότησης τροχιών που περιορίζονται από το δίκτυο μελετάται από πιο θεωρητικής σκοπιάς αντί να προτείνεται μία μέθοδος προσπέλασης που θα μπορούσε να χρησιμοποιηθεί σε πραγματικές εφαρμογές. Από την άλλη πλευρά, ακολουθώντας τις κατευθύνσεις που δίνονται στην [KGT99], στην παρούσα διατριβή, καταδεικνύουμε πως αυτές μπορούν να υλοποιηθούν στην πράξη μέσω ανάπτυξης νέων μεθόδων προσπέλασης για την δεικτοδότηση τροχιών υπό περιορισμούς δικτύου.

1.3.2. Προηγμένη Επεξεργασία Επερωτήσεων

Ο τομέας της προηγμένης επεξεργασίας επερωτήσεων έχει ως στόχο την ανάπτυξη εξειδικευμένων τεχνικών επεξεργασίας για την εκτέλεση προηγμένων επερωτήσεων, που μπορούν (ή και όχι) να εκμεταλλευτούν υπάρχουσες δομές ευρετηρίων που υποστηρίζουν πιο παραδοσιακές επερωτήσεις. Εδώ θα πρέπει να επισημάνουμε ότι, οι επερωτήσεις της μορφής «*βρες όλα τα αντικείμενα που εντοπίζονται σε ένα δεδομένο χώρο σε ένα συγκεκριμένο χρονικό διάστημα*», ήτοι επερωτήσεις *εύρους (range)* (Q_2 στο Σχήμα 1.3), θεωρούνται παραδοσιακές επερωτήσεις και εξ' ορισμού υποστηρίζονται από οποιοδήποτε ευρετήριο· στην ίδια κατηγορία εμπίπτουν επίσης και οι επερωτήσεις της μορφής «*βρες όλες τις θέσεις των αντικειμένων σε μία δεδομένη περιοχή σε μία συγκεκριμένη χρονική στιγμή*», που ονομάζονται επερωτήσεις χρονικής στιγμής (*timeslice*) και αποτελούν εξειδίκευση των απλών επερωτήσεων εύρους έχοντας μηδενική διάρκεια ζωής (Q_1 στο Σχήμα 1.3) . Η εκτέλεση επερωτήσεων εύρους είναι συνήθως μία απλή διαδικασία· για παράδειγμα, η εκτέλεση μιας επερώτησης εύρους σε δομές που ομοιάζουν στο R-δέντρο (όπως, το 3D R-δέντρο [TVS96], τα TB και τα STR-δέντρα [PJT00] και το TB*-δέντρο που προτείνεται σε αυτή τη διατριβή) και αποθηκεύουν δεδομένα

ιστορικών τροχιών είναι μία απλή επέκταση του αλγορίθμου `FindLeaf` που προτάθηκε αρχικά στην [Gut84] στον 3D χώρο που σχηματίζεται από τις δύο χωρικές και τη μία χρονική διάσταση.



Σχήμα 1.3: Επερωτήσεις σε βάσεεις δεδομένων τροχιών

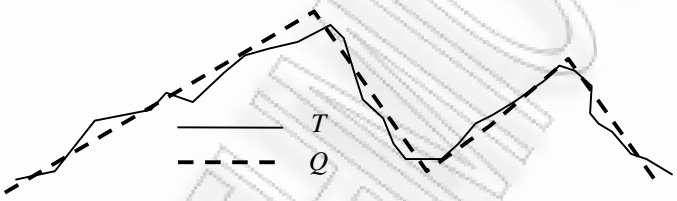
Αφ'ετέρου, υπάρχει μία σειρά από χωροχρονικούς τελεστές, που απαιτούν πιο εξελιγμένες τεχνικές επεξεργασίας επερωτήσεων· συχνά αυτοί οι τελεστές είναι οι προεκτάσεις των αντίστοιχων χωρικών. Μεταξύ αυτών έχει εισαχθεί και μία σημαντική κατηγορία επερωτήσεων στις MODs απευθείας από το πεδίο των χωρικών βάσεων δεδομένων, η γνωστή ως *αναζήτηση των k πλησιέστερων γειτόνων (k-nearest neighbor (k-NN) search)*, όπου μας ενδιαφέρει η εύρεση των *k* πλησιέστερων τροχιών σε ένα προκαθορισμένο αντικείμενο επερώτησης *Q*. Εξ' όσων γνωρίζουμε, η βιβλιογραφία βάσεων δεδομένων που αφορά σε τέτοιες επερωτήσεις ασχολείται κυρίως είτε με σταθερά ([RKV95], [CF98], [HS99]) είτε με συνεχώς κινούμενα σημεία επερωτήσεων ([SR01], [TPS02]) σε βάσεις δεδομένων ακίνητων χωρικών αντικειμένων ή επερωτήσεις επάνω στις μελλοντικές ή τρέχουσες θέσεις συνεχώς κινούμενων σημείων ([BJKS02], [TP02], [ISS03], [YPK05], [XMA05], [MHP05]). Προφανώς, αυτοί οι τύποι επερωτήσεων δεν καλύπτουν την αναζήτηση του πλησιέστερου γείτονα σε ιστορικές τροχιές. Έτσι, μία από τις προκλήσεις που αντιμετωπίζονται στο πεδίο των βάσεων δεδομένων τροχιών είναι η ανάπτυξη μηχανισμών για την εκτέλεση αναζητήσεων *k*-NN σε MODs εκμεταλλευόμενοι χωροχρονικά ευρετήρια που αποθηκεύουν ιστορικές πληροφορίες.

Επιπλέον, η πολυπλοκότητα των σχετικών δεδομένων μας οδηγεί στην ταξινόμηση των τελεστών του δυνητικού πλησιέστερου γείτονα στις MODs που αποθηκεύουν ιστορικά δεδομένα τροχιών ως εξής: (a) βάσει της φύσης του αντικειμένου επερώτησης, που μπορεί να είναι είτε *σταθερό* είτε *κινούμενο* σημείο (δηλαδή μία ακόμα τροχιά που δεν περιέχεται στην MOD) και, (b) σύμφωνα με το αποτέλεσμα που ζητά ο τελεστής, δηλαδή μεταξύ του πλησιέστερου αντικειμένου στην επερώτηση κατά τη διάρκεια ζωής της επερώτησης και του πλησιέστερου / πλησιέστερων σε οποιαδήποτε χρονική στιγμή της· αυτή η τελευταία είναι η κατηγορία των *ιστορικών επερωτήσεων συνεχούς πλησιέστερου γείτονα/ historical continuous nearest neighbor queries*.

Για να κατανοήσουμε καλύτερα την προηγούμενη ταξινόμηση, ας θυμηθούμε το Σχήμα 1.3 που αναπαριστά μία βάση δεδομένων με τέσσερις τροχιές $\{T_1, T_2, T_3, T_4\}$ και αρκετές επερωτήσεις που τίθενται προς τη βάση. Η επερώτηση Q_3 ζητά την πλησιέστερη τροχιά προς το αντικείμενο της επερώτησης (που είναι ένα σταθερό σημείο) κατά τη χρονική περίοδο $[t_1, t_4]$· αυτή είναι η απλή περίπτωση και η απάντηση στην επερώτηση είναι η τροχιά T_3. Ομοίως, το Q_4 είναι ισοδύναμο του Q_3, με τη μόνη διαφορά ότι το αντικείμενο της επερώτησης είναι μία άλλη τροχιά, που δεν περιέχεται στη

βάση δεδομένων· σε αυτή την περίπτωση, η απάντηση είναι η τροχιά T_4. Έστω τώρα η επερώτηση Q_5 που είναι ιστορική επερώτηση συνεχούς πλησιέστερου γείτονα· σε αυτή την περίπτωση το αποτέλεσμα της επερώτησης θα πρέπει να είναι μία λίστα από πλειάδες στοιχείων που περιέχουν τις πλησιέστερες τροχιές καθώς και την χρονική περίοδο κατά την οποία η συγκεκριμένη τροχιά ήταν η πλησιέστερη, δηλαδή$\{(T_4, [t_1,t_3)), [T_3, [t_3, t_4])\}$.

Ένας ακόμα ενδιαφέρων τύπος επερώτησης που είναι χρήσιμος στην αναζήτηση των MODs προκύπτει από το λεγόμενο πρόβλημα της *ομοιότητας τροχιών (trajectory similarity)* που έχει ως στόχο την εξεύρεση «παρόμοιων» τροχιών κινούμενων αντικειμένων. Για την αποτελεσματική αντιμετώπιση αυτών των επερωτήσεων, οι MODs θα πρέπει να περιλαμβάνουν μεθόδους για την αναζήτηση της *ομοιότερης τροχιάς (Most-Similar-Trajectory - MST*) η οποία συζητείται στην [The03]· παράδειγμα επερώτησης MST είναι το Q_6 στο Σχήμα 1.3, που ανακαλεί ως τροχιά με τη μεγαλύτερη ομοιότητα την T_1. Η αναζήτηση της ομοιότερης τροχιάς είναι ένα σχετικά νέο θέμα στη βιβλιογραφία· η πλειονότητα των μεθόδων που προτείνονται μέχρι τώρα βασίζονται είτε σε αντίστοιχες εργασίες του χώρου της ανάλυσης χρονοσειρών είτε στο μοντέλο της μεγαλύτερης κοινής υποακολουθίας (Longest Common Subsequence - LCSS) [VKG02] και της προσφάτως προταθείσης απόστασης επεξεργασίας σε πραγματικές ακολουθίες (Edit Distance on Real Sequence - EDR) [COO05].



Σχήμα 1.4. Τροχιές με διαφορετικό ρυθμό δειγματοληψίας

Ωστόσο, η πλειοψηφία των προτεινόμενων μεθόδων, είτε αγνοούν την χρονική διάσταση της κίνησης, υπολογίζοντας συνεπώς την χωρική (και όχι τη χωροχρονική) ομοιότητα μεταξύ των τροχιών, ή υποθέτουν ότι οι τροχιές έχουν τον ίδιο ρυθμό δειγματοληψίας. Για να δώσουμε ένα παράδειγμα για το πρόβλημα που προκύπτει όταν έχουμε διαφορετικούς ρυθμούς δειγματοληψίας, ας δούμε το Σχήμα 1.4 που αναπαριστά δύο τροχιές T και Q που η θέση τους καταγράφεται με διαφορετικούς ρυθμούς· παρά το γεγονός ότι προφανώς οι δύο τροχιές είναι παρόμοιες, οι μέθοδοι που βασίζονται στο μοντέλο LCSS ή το EDR δεν μπορούν να ανιχνεύσουν τέτοιας μορφής ομοιότητα διότι προσπαθούν να ταιριάξουν τις θέσεις των αντικειμένων μία προς μία, γεγονός το οποίο σαφώς και δεν συμβαίνει στο παραπάνω παράδειγμα του πραγματικού κόσμου. Επιπλέον, η πλειονότητα των προτεινόμενων προσεγγίσεων εκμεταλλεύεται εξειδικευμένες δομές ευρετηρίων για το κλάδεμα του χώρου αναζήτησης και την ανάκτηση της ομοιότερης τροχιάς σε μια τροχιά επερώτησης. Συνεπώς, μία από τις προκλήσεις που αντιμετωπίζονται στον τομέα των βάσεων δεδομένων τροχιάς είναι η ανάπτυξη μηχανισμών για την εκτέλεση αναζήτησης k-MST σε MODs εκμεταλλευόμενοι τα υπάρχοντα χωροχρονικά ευρετήρια που υποστηρίζουν και άλλους τύπους επερωτήσεων.

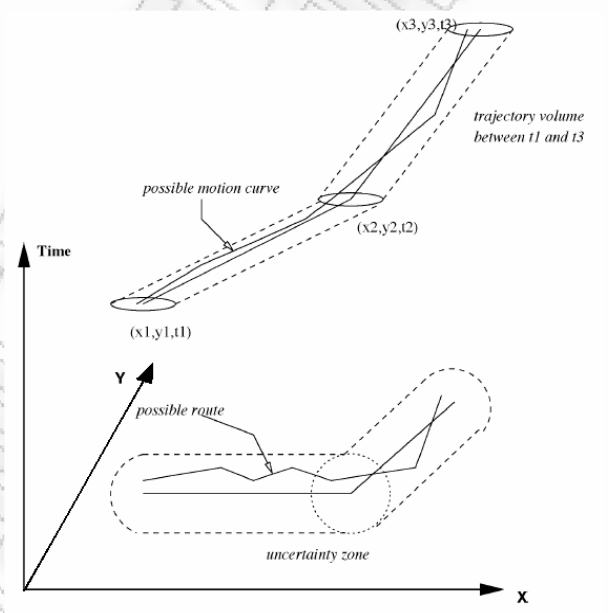
1.3.3. Υποστήριξη Αβεβαιότητας

Στη βιβλιογραφία, η αβεβαιότητα ορίζεται ως η μέτρηση της διαφοράς μεταξύ των πραγματικών περιεχομένων μιας βάσης δεδομένων και των περιεχομένων που θα είχε δημιουργήσει ο χρήστης ή η

εφαρμογή μέσω άμεσης και ακριβούς παρατήρησης της πραγματικότητας [ZG02]. Τα κάτωθι στοιχεία μπορεί ν' αποτελέσουν πηγή αβεβαιότητας:

- Ατελής παρατήρηση του πραγματικού κόσμου,
- Ελλιπής γλώσσα αναπαράστασης,
- Άγνοια, οκνηρία ή έλλειψη αποτελεσματικότητας.

Οι Pfoser και Jensen [PJ99] προτείνουν μία αναπαράσταση της αβεβαιότητας θέσης λόγω των σφαλμάτων μέτρησης και δειγματοληψίας, τα οποία εμπίπτουν στην πρώτη και τρίτη από τις παραπάνω πηγές σφαλμάτων, αντίστοιχα. Σύμφωνα με την [PJ99] η χωρική προβολή της τροχιάς ενός αντικειμένου μπορεί να αναπαρασταθεί ως μία 2D ελλειπτική επιφάνεια, που ορίζεται από δύο συνεχόμενες καταγεγραμμένες θέσεις. Από την άλλη πλευρά, στις [TWHC04], [TWZC02] περιγράφεται ένα μοντέλο που αντιλαμβάνεται ταυτόχρονα και τα δύο είδη αβεβαιότητας. Στο εν λόγω μοντέλο, εισάγεται ένα *κατώφλι αβεβαιότητας (uncertainty threshold)* που υποδηλώνει την μέγιστη απόσταση του αντικειμένου από την υποτιθέμενη θέση στην τροχιά. Πιο συγκεκριμένα, δεδομένων των καταγεγραμμένων σημείων, και μετά την εφαρμογή γραμμικής παρεμβολής μεταξύ των, το μοντέλο αυτό για κάθε σημείο της τροχιάς προσδιορίζει ένα δίσκο, παράλληλο στο επίπεδο *XY*, ακτίνας ίσης του κατωφλίου. Αν πάρουμε όλους τους δίσκους μαζί στον 3D χωροχρόνο, σχηματίζεται ένας «σωλήνας» γύρω από την πολυγραμμή (polyline) που συνδέει τα καταγεγραμμένα σημεία (Σχήμα 1.5). Το κατώφλι αυτό περιλαμβάνει και την αβεβαιότητα παρεμβολής και τα σφάλματα μέτρησης, ενώ δεν κάνει καμία διάκριση μεταξύ των σημείων δειγματοληψίας και των σημείων που προκύπτουν από την παρεμβολή.



Σχήμα 1.5. Μοντέλο αβεβαιότητας κινούμενου αντικειμένου [TWHC04]

Η βιβλιογραφία για τη διαχείριση της αβεβαιότητας θέσης των χωροχρονικών αντικειμένων μέχρι τούδε, πέρα από θέματα αναπαράστασης της αβεβαιότητας [Tra03], [TWHC04], [WSCY99], ασχολείται και με πιθανοτικούς αλγορίθμους [TWHC04], [TWZC02], [CKP04] που επεξεργάζονται επερωτήσεις παρουσία αβεβαιότητας, εκτιμώντας την πιθανότητα που έχει κάθε τροχιά να περιλαμβάνεται στο αποτέλεσμα της επερώτησης. Από την άλλη πλευρά, υπάρχουν περιπτώσεις όπου

ο χρήστης θα προτιμούσε να γνωρίζει την επίδραση του σφάλματος στα αποτελέσματα της επερώτησης, χωρίς να χρειαστεί να εκτελέσει την επερώτηση. Έστω για παράδειγμα η ακόλουθη πραγματική κατάσταση, εμπνευσμένη από το παράδειγμα των αναδυόμενων ανοικτών αγορών [Ioa07]: έστω ένας χρήστης που επιθυμεί να θέσει μία επερώτηση σε αρκετές κατανεμημένες πηγές δεδομένων που παρέχουν δεδομένα στους συνδρομητές τους και περιέχουν τα ίδια χωροχρονικά αντικείμενα (ήτοι τροχιές) που αναπαρίστανται σε διαφορετικά επίπεδα αβεβαιότητας λόγω διαφορετικών μεθόδων μέτρησης και κατά συνέπεια διαφορετικών κατωφλίων αβεβαιότητας αντίστοιχα· παρόλο που το κριτήριο που χρησιμοποιείται για να επιλέξουμε μεταξύ των διαφόρων πηγών δεδομένων είναι η βελτιστοποίηση δηλαδή η ελαχιστοποίηση, της αβεβαιότητας που εισάγεται στα τελικά αποτελέσματα της επερώτησης, οι πηγές δεδομένων κατά τη φάση της διαπραγμάτευσης [Ioa07] παρουσιάζουν στους δυνητικούς πελάτες / χρήστες μόνο συγκεντρωτικά δεδομένα. Συνεπώς ο μόνος τρόπος για να αποφασίσουμε ως προς την αβεβαιότητα των αποτελεσμάτων είναι η παρουσία ενός μοντέλου που εξυπηρετεί αυτό το σκοπό, βάσει των συγκεντρωτικών μόνον δεδομένων που δίνονται από τους παρόχους.

Ένα άλλο πρόβλημα που μας εξάπτει το ενδιαφέρον και σχετίζεται με αυτό που μόλις παρουσιάστηκε, είναι να καθορίσουμε τη *μέγιστη επιτρεπόμενη* (αν)ακρίβεια των δεδομένων τροχιάς που θα τροφοδοτήσουν μία MOD δεδομένης της απαιτούμενης ακρίβειας στα αποτελέσματα των επερωτήσεων χρονικής στιγμής. Κατόπιν, μπορούμε να καθοδηγήσουμε τους χρήστες μέσω του DBMS στη χρήση της κατάλληλης, περισσότερο / λιγότερο ακριβούς – που σημαίνει επίσης και περισσότερο ή λιγότερο δαπανηρής – μεθόδου εντοπισμού που θα χρησιμοποιηθεί για τα δεδομένα που θα τροφοδοτήσουν το σύστημα.

Αμφότερες οι παραπάνω απαιτήσεις θα μπορούσαν να ικανοποιηθούν μέσω ενός μοντέλου που προβλέπει το σφάλμα που εισάγεται στα αποτελέσματα των επερωτήσεων βάσει γνωστών ιδιοτήτων των επερωτήσεων και του συνόλου δεδομένων (όπως για παράδειγμα το πλήθος των δεδομένων), χωρίς να εκτελείται η επερώτηση· επιπλέον, ένα τέτοιο μοντέλο θα μπορούσε να χρησιμοποιηθεί σε έναν διαδραστικό δημιουργό / βελτιστοποιητή επερωτήσεων, που θα ενημερώνει το χρήστη για την επίδραση της αβεβαιότητας στα αποτελέσματα της επερώτησης, μαζί με άλλα στοιχεία, όπως η επιλεκτικότητα της επερώτησης, ο αναμενόμενος χρόνος εκτέλεσης κλπ. Εξ' όσων γνωρίζουμε, δεν υπάρχει θεωρητική μελέτη για τη μοντελοποίηση του σφάλματος που εισάγεται σε αποτελέσματα χωροχρονικών επερωτήσεων λόγω της αβεβαιότητας της τροχιάς· επομένως παραμένει ανοικτό ερευνητικό πρόβλημα στον τομέα των χωροχρονικών βάσεων δεδομένων.

1.3.4. Συμπίεση Τροχιών

Όπως αναφέρεται στην [MB04], αναμένεται ότι οι πανταχού παρούσες συσκευές εντοπισμού θέσης αργά ή γρήγορα θα αρχίσουν να παράγουν μια ροή δεδομένων θέσεων άνευ προηγουμένου. Ένας τόσο μεγάλος αριθμός δεδομένων θα οδηγήσει αναπόφευκτα σε προβλήματα αποθήκευσης, μετάδοσης, υπολογισμών και απεικόνισης. Εξ' ου και προκύπτει η ανάγκη τεχνικών συμπίεσης. Ωστόσο, οι μελέτες στο συγκεκριμένο τομέα είναι σχετικά περιορισμένες [CWT03], [MB04], [PPS06], [PPS06a], [PPS07], και κατευθύνονται κυρίως από αντίστοιχες τεχνικές που προτείνονται στον τομέα της απλούστευσης γραμμών, της χαρτογραφικής γενίκευσης και της συμπίεσης χρονοσειρών. Βάσει της [MB04] οι στόχοι για την συμπίεση των δεδομένων τροχιάς είναι να παράγει:

- μία μόνιμη μείωση στο μεγέθος των δεδομένων·
- μια σειρά δεδομένων που θα επιτρέπει ακόμα αρκετούς υπολογισμούς σε αποδεκτή (χαμηλή) πολυπλοκότητα·
- μια σειρά δεδομένων με γνωστά, μικρά περιθώρια σφάλματος, τα οποία κατά προτίμηση είναι παραμετρικά προσαρμόσιμα.

Κατά συνέπεια μας ενδιαφέρουν τεχνικές συμπίεσης με απώλειες (lossy) που αποκλείουν περιττές πληροφορίες υπό πολύ καλά καθορισμένα όρια σφάλματος.

Ειδικά όσον αφορά στο θέμα του σφάλματος που εισάγεται στα δεδομένα που προκύπτουν από τις εν λόγω τεχνικές συμπίεσης, η μόνη σχετική εργασία [MB04] μας δίνει ένα τύπο που εκτιμά το μέσο σφάλμα της προσεγγιστικής τροχιάς σε όρους της απόστασης από την αρχική ροή δεδομένων. Αφ'ετέρου, υπάρχουν κι άλλα είδη σφαλμάτων που θα μπορούσαν να βοηθήσουν το χρήστη μιας MOD να αποφασίσει για την ποιότητα των συμπιεσμένων δεδομένων. Για παράδειγμα, είναι πολύ πιο ουσιαστικό να δοθούν στο χρήστη πληροφορίες σχετικά με το μέσο σφάλμα που εισάγεται στα αποτελέσματα επερωτήσεων για τα συμπιεσμένα δεδομένα. Συνεπώς, προκύπτει η ανάγκη ενός αναλυτικού μοντέλου που εκτιμά το σφάλμα λόγω συμπίεσης στα αποτελέσματα χωροχρονικών επερωτήσεων.

Ένα τέτοιο μοντέλο θα μπορούσε να χρησιμοποιηθεί αμέσως μετά την συμπίεση ενός συνόλου δεδομένων τροχιών προκειμένου να δοθεί στο χρήστη το μέσο σφάλμα που εισάγεται στα αποτελέσματα χωροχρονικών επερωτήσεων διαφόρων μεγεθών· θα μπορούσαμε συνεπώς να το εκμεταλλευτούμε ως επιπλέον κριτήριο για το χρήστη προκειμένου ν' αποφασίσει αν τα συμπιεσμένα δεδομένα είναι κατάλληλα για τις ανάγκες του / της και πιθανόν ν' αποφασίσει για διάφορους βαθμούς συμπίεσης κ.ο.κ.. Επιπλέον, θα μπορούσε να χρησιμοποιηθεί ώστε να βελτιωθεί η αποτελεσματικότητα των προτεινόμενων λύσεων σχετικά με τη συμπίεση τροχιών· δεδομένου ότι ένα μοντέλο της μορφής αυτής θα μας δώσει τις πραγματικές παραμέτρους από τις οποίες εξαρτάται το σφάλμα, θα μπορούσαν στη συνέχεια να οδηγήσουν σε διαισθητικές κατευθύνσεις προς τη χρήση πιο εξελιγμένων / αποτελεσματικών λύσεων. Η πρόκληση λοιπόν που έχουμε ν' αντιμετωπίσουμε ως προς τη συμπίεση τροχιών είναι να βρεθεί ένα θεωρητικό μοντέλο που να εκτιμά το σφάλμα λόγω συμπίεσης στα αποτελέσματα των χωροχρονικών επερωτήσεων και επιπροσθέτως να το προσαρμόσουμε στα πλαίσια των MODs.

1.4. Συνεισφορά της Διατριβής

Η διατριβή αυτή παρουσιάζει αρκετές εργασίες που είναι απαραίτητες για την αποτελεσματική *Διαχείριση των Δεδομένων Τροχιών Κινούμενων Αντικειμένων.*

Ο απώτερος στόχος της διεξαγόμενης έρευνας είναι να μας δώσει αποτελεσματικούς μηχανισμούς που επιτρέπουν στις Βάσεις Δεδομένων Κινούμενων Αντικειμένων την αποτελεσματική αποθήκευση και εκτέλεση επερωτήσεων σε δεδομένα ιστορικών τροχιών· άρα η έρευνα ασχολείται με τη δεικτοδότηση, την προηγμένη επεξεργασία επερωτήσεων, την υποστήριξη της αβεβαιότητας και με θέματα συμπίεσης τροχιών.

Στη συνέχεια, θα αναφερθούμε στις συνεισφορές της διατριβής, ομαδοποιημένες ανά θέμα. Θα πρέπει εδώ να επισημάνουμε, ότι η καινοτομία στην προσέγγιση μας καθορίζεται σε κάθε κεφάλαιο,

μέσω της παρουσίασης των αντίστοιχων υπαρχόντων σχετικών εργασιών. Επιλέξαμε την προσέγγιση αυτή, αντί της συνολικής παρουσίασης των σχετικών εργασιών, λόγω της ποικιλομορφίας των θεμάτων σε μία απόπειρα να διευκολύνουμε την ανάγνωση του κειμένου.

Δεικτοδότηση. Για την αντιμετώπιση των απαιτήσεων δεικτοδότησης που παρουσιάστηκαν παραπάνω, στην παρούσα διατριβή εισάγονται δύο νέα ευρετήρια, πιο συγκεκριμένα το TB*-δέντρο και το FNR-δέντρο. Το TB*-δέντρο είναι μία επέκταση του TB-δέντρου που μας δίνει τη δυνατότητα να υποστηρίξουμε *μη χρονολογικές εισαγωγές*· είναι πιο συμπαγές, βελτιώνει την απόδοση σε θέματα χρόνου κατασκευής, ενώ υπερτερεί από άποψη ταχύτητας απόκρισης του προκάτοχού του στην πλειονότητα των επερωτήσεων που υποστηρίζει. Εκτός από τους αλγορίθμους κατασκευής και επεξεργασίας επερωτήσεων, το TB*-δέντρο υποστηρίζει διαγραφές τροχιάς, ενώ η δομή του επιτρέπει επιπλέον την υποστήριξη αλγορίθμων συμπίεσης τροχιάς, δύο ιδιότητες που δεν υποστηρίζονται από το αρχικό TB-δέντρο. Παρόλα αυτά είναι απαραίτητο να διευκρινίσουμε ότι το προτεινόμενο TB*-δέντρο, δεν εκμεταλλεύεται τις ειδικές συνθήκες που διέπουν τα αντικείμενα όταν κινούνται σε σταθερά δίκτυα, αλλά δεικτοδοτεί αντικείμενα που κινούνται ελεύθερα στο 2D χώρο.

Από την άλλη κάτω από το σενάριο κίνησης υπό περιορισμούς δικτύου η διατριβή αυτή δίνει ένα νέο ευρετήριο, που ονομάζεται R-δέντρο σταθερού δικτύου (Fixed Network R-tree - FNR-tree) και αποτελεί επέκταση του πολύ γνωστού R-δέντρου [Gut84]. Οι γενικές ιδέες στις οποίες βασίζεται το FNR-δένδρο παρουσιάζονται σε αδρές γραμμές στο [Fre02], παρόλα αυτά χωρίς οποιαδήποτε εφαρμογή ή πειραματική αξιολόγηση της προτεινόμενης μεθόδου. Το FNR-δέντρο μπορεί να αναπαρασταθεί εν συντομία ως ένα δάσος 1D R-δέντρων πάνω σε ένα 2D R-δέντρο. Το 2D R-δέντρο χρησιμοποιείται για να δεικτοδοτήσουμε τα χωρικά δεδομένα του διαγράμματος του δικτύου (ήτοι δρόμους που αποτελούνται από γραμμικά τμήματα), ενώ τα 1D R-δέντρα χρησιμοποιούνται για τη δεικτοδότηση του χρονικού διαστήματος της κίνησης κάθε αντικειμένου σε ένα δεδομένο τμήμα του δικτύου. Όπως θα αποδειχθεί πειραματικά, το προτεινόμενο FNR-δέντρο ξεπερνά σε απόδοση όλους τους ανταγωνιστές του σε γενικές επερωτήσεις εύρους, κάτι που αντισταθμίζει το κόστος της έλλειψης ενός μηχανισμού που διατηρεί τις τροχιές.

Τα αποτελέσματα των ανωτέρω θεμάτων παρουσιάζονται στο Κεφάλαιο 2. Προκαταρτικά αποτελέσματα έχουν ήδη δημοσιευθεί στις [Fre03], [FT06].

Προηγμένη Επεξεργασία Επερωτήσεων: Αναζήτηση Πλησιέστερου Γείτονα. Για να υποστηρίξουμε αποτελεσματικά την αναζήτηση πλησιέστερου γείτονα σε τροχιές κινούμενων αντικειμένων προτείνουμε κατ' αρχάς μία σειρά νέων μετρικών που είναι απαραίτητες για τον τρόπο με τον οποίο διατάσσονται και κλαδεύονται τα δεδομένων κατά την εκτέλεση των αντίστοιχων αλγορίθμων. Πιο συγκεκριμένα, ο ορισμός της μετρικής ελάχιστης απόστασης *MINDIST* μεταξύ σημείων και ορθογωνίων παραλληλογράμμων, που προτάθηκε αρχικώς στην [RKV95] και επεκτάθηκε στην [TPS02], επεκτείνεται προκειμένου οι αλγόριθμοι μας να υπολογίσουν την ελάχιστη απόσταση μεταξύ τροχιών και ορθογωνίων παραλληλογράμμων αποτελεσματικά. Κατόπιν προτείνουμε αλγορίθμους επεξεργασίας επερωτήσεων για την εκτέλεση αναζητήσεων NN στα χωροχρονικά ευρετήρια που αποθηκεύουν ιστορικές πληροφορίες κινούμενων αντικειμένων. Μεταξύ των υποψηφίων

χωροχρονικών ευρετηρίων, εκμεταλλευόμαστε τα πιο συνηθισμένα ευρετήρια που είναι και αυτά που υποστηρίζουν την απεριόριστη κίνηση δηλαδή, δομές που ομοιάζουν στα R-δέντρα όπως το 3D R-δέντρο [TVS96], το TB-δέντρο [PJT00] και το προτεινόμενο σε αυτή τη διατριβή TB*-δέντρο. Η περιγραφή των αλγορίθμων μας για τις διάφορες επερωτήσεις εξαρτάται από τον τύπο του αντικείμενου επερωτήσεως (σημείο ή τροχιά) καθώς και από το κατά πόσο η ίδια η επερώτηση είναι συνεχής ή όχι. Πιο συγκεκριμένα, παρουσιάζουμε αποτελεσματικούς αλγορίθμους «πρώτα στο βαθύτερο» (depth-first) και «πρώτα στο καλύτερο» (best-first) για ιστορικές επερωτήσεις NN καθώς και αλγορίθμους πρώτα στο βαθύτερο για τις συνεχείς αντίστοιχές τους. Όλοι οι προτεινόμενοι αλγόριθμοι γενικεύονται για να βρούμε τους k πλησιέστερους γείτονες. Τέλος, εκτελούμε μία πλήρη σειρά πειραμάτων σε μεγάλα συνθετικά και πραγματικά σύνολα δεδομένων που αποδεικνύουν ότι οι αλγόριθμοι έχουν υψηλή απόδοση και δυνατότητα κλιμάκωσης, μετρούμενης σε σχέση με τον αριθμό των κόμβων που προσπελαύνονται, με το χρόνο εκτέλεσης και με το κλαδεμένο χώρο.

Τα αποτελέσματά μας στα παραπάνω θέματα παρουσιάζονται στο Κεφάλαιο 3. Προκαταρτικά αποτελέσματα έχουν ήδη δημοσιευθεί στις [FGPT05], [FGPT07], [PFGT08].

Προηγμένη Επεξεργασία Επερωτήσεων: Αναζήτηση Ομοιότητας. Στη διατριβή αυτή αντιμετωπίζουμε τα όσα αναφέρονται στο θέμα της αναζήτησης ομοιότητας, μέσω αποτελεσματικής υποστήριξης της αναζήτησης k-MST σε MODs που αποθηκεύουν δεδομένα ιστορικών τροχιών και δεικτοδοτούνται από δομές παρόμοιες των R-δέντρων. Πιο συγκεκριμένα, υποστηρίζουμε την αναζήτηση k-MST καθορίζοντας μία *μετρική ανομοιότητας* που ονομάζεται *DISSIM* για τη μέτρηση της χωροχρονικής ανομοιότητας μεταξύ δύο τροχιών· η μετρική αυτή έχει επίσης προταθεί ανεξάρτητα στην [NP06] και μπορούμε να την θεωρήσουμε ως τη μέση απόσταση μεταξύ των δύο τροχιών στο πέρασμα του χρόνου. Ακολούθως προτείνουμε μία αποτελεσματική προσεγγιστική μέθοδο για να ξεπερασθεί το κόστος του υπολογισμού της *DISSIM*, ενώ, στη συνέχεια, αναπτύσσουμε μια σειρά νέων μετρικών καθώς και αρκετά σχετικά λήμματα, που χρησιμοποιούνται για σκοπούς διάταξης και κλαδέματος από τους προτεινόμενους αλγορίθμους αναζήτησης ομοιότερης τροχιάς. Πιο συγκεκριμένα χρησιμοποιώντας αυτές τις μετρικές, προτείνουμε ένα «πρώτα στο βαθύτερο» και ένα «πρώτα στο καλύτερο» αλγόριθμο για την εκτέλεση της αναζήτησης k-MST σε δομές παρόμοιες των R-δέντρων που αποθηκεύουν δεδομένα ιστορικών τροχιών. Για την ολοκλήρωση του συγκεκριμένου θέματος διεξάγουμε μια πλήρη σειρά πειραμάτων σε μεγάλα συνθετικά και πραγματικά σύνολα δεδομένων που αποδεικνύουν ότι οι αλγόριθμοι έχουν υψηλή δυνατότητα κλιμάκωσης και απόδοση, μετρούμενη σε σχέση με τον αριθμό των κόμβων που προσπελαύνονται και με το κλαδεμένο χώρο. Τέλος, δείχνουμε ότι η προτεινόμενη μετρική ομοιότητας ανακτά με αποτελεσματικό τρόπο χωροχρονικά παρόμοιες τροχιές σε περιπτώσεις όπου ανάλογες προσεγγίσεις αποτυγχάνουν.

Θα πρέπει εδώ να τονίσουμε ότι όλοι οι προτεινόμενοι αλγόριθμοι δεν απαιτούν κάποια συγκεκριμένη δομή ευρετηρίου και μπορούν να εφαρμοσθούν απευθείας σε οποιοδήποτε μέλος της οικογενείας των R-δέντρων που χρησιμοποιούνται για τη δεικτοδότηση τροχιών, όπως το 3D R-δέντρο [TVS96], το TB-δέντρο [PJT00] και το TB*-δέντρο που προτείνεται σε αυτή τη διατριβή. Εξ' όσων γνωρίζουμε, η πρόταση αυτής της διατριβής είναι η πρώτη που παρέχει τεχνικές για ένα χωροχρονικό

ευρετήριο για την υποστήριξη επερωτήσεων χωροχρονικού εύρους, τοπολογικών, πλησιέστερου γείτονα και βάσει ομοιότητας.

Τα αποτελέσματά μας στα παραπάνω θέματα παρουσιάζονται στο Κεφάλαιο 4. Προκαταρτικά αποτελέσματα έχουν ήδη δημοσιευθεί στην [FGT07].

Υποστήριξη Αβεβαιότητας: Τα προβλήματα σχετικά με τη διαχείριση της αβεβαιότητας που επισημάναμε στην προηγούμενη ενότητα καλύπτονται αρχικώς από δύο λήμματα, που εκτιμούν το μέσο αριθμό λανθασμένων θετικών και λανθασμένων αρνητικών κατά την εκτέλεση επερωτήσεων χρονικής στιγμής (timeslice) σε ομοιόμορφα κατανεμημένες αβέβαιες τροχιές η μοντελοποίηση των οποίων γίνεται μέσω της πρότασης του [TWHC04]· και τα δύο σφάλματα εξαρτώνται από την ακτίνα του κυλινδρικού όγκου (δηλαδή το κατώφλι αβεβαιότητας) και την περίμετρο του παραθύρου επερώτησης χρονικής στιγμής, αντί της επιφάνειας. Κατόπιν, για να απαλλαγούμε από την υπόθεση της ομοιόμορφης αβεβαιότητας θέσης (που προκύπτει απευθείας από το μοντέλο της [TWHC04]) και για να χρησιμοποιήσουμε την διμεταβλητή κανονική κατανομή που εμφανίζεται στον πραγματικό κόσμο [PTJ05], γίνεται μία αποτελεσματική προσέγγισή της με την κατανομή διαφοράς ομοιομορφίας (uniform difference distribution). Τα αποτελέσματα προσεγγίζουν σε ικανοποιητικό βαθμό την αρχική ανάλυση. Η επέκταση του μοντέλου ώστε να καλύπτει και τις αυθαίρετα κατανεμημένες τροχιές και τις διάφορες κατανομές ακτινών αβεβαιότητας γίνεται μέσω της χρήσης πρωτότυπων χωροχρονικών και άλλων επαυξημένων ιστογραμμάτων. Κατόπιν εκτελούμε μια πλήρη σειρά πειραμάτων που καταδεικνύουν την ορθότητα και την ακρίβεια και της ανάλυσης. Τέλος, αποδεικνύεται πως τα αποτελέσματα της ανάλυσης μπορούν να εφαρμοσθούν σε χωρικά σύνολα δεδομένων: οι λύσεις που προτείνονται εφαρμόζονται σε ένα εμπορικό Σύστημα Διαχείρισης Χωρικών Βάσεων Δεδομένων (Spatial Database Management System - SDBMS), πιο συγκεκριμένα, την PostgreSQL [Post08b] με τη χωρική επέκταση PostGIS [Post08a]. Εδώ, θα πρέπει να επισημάνουμε ότι τα πιο συνηθισμένα χωρικά ιστογράμματα, που χρησιμοποιούνται ήδη σε εμπορικά SDBMS για την εκτίμηση της επιλεκτικότητας των επερωτήσεων, υποστηρίζουν το προτεινόμενο μοντέλο χωρίς περαιτέρω προσθήκες.

Τα αποτελέσματά μας στα παραπάνω θέματα παρουσιάζονται στο Κεφάλαιο 5. Προκαταρτικά αποτελέσματα έχουν ήδη δημοσιευθεί στην [FGT08].

Συμπίεση Τροχιών: Προκειμένου να καλύψουμε τα θέματα που προέκυψαν από την προηγούμενη συζήτηση σχετικά με τη συμπίεση τροχιών, περιγράφουμε πρώτα δύο τύπους σφαλμάτων (πιο συγκεκριμένα τα λανθασμένα αρνητικά και λανθασμένα θετικά) κατά την εκτέλεση επερωτήσεων χρονικής στιγμής σε συμπιεσμένες τροχιές και αποδεικνύουμε ένα λήμμα που εκτιμά το μέσο αριθμό των παραπάνω τύπων σφαλμάτων. Αποδεικνύεται ότι ο μέσος αριθμός λανθασμένων αποτελεσμάτων και των δύο τύπων εξαρτάται από τη Σύγχρονη Ευκλείδεια Απόσταση (Synchronous Euclidean Distance - *SED*) [CWT03], [MB04], [PPS06], [PPS06a] κατά τους *x*- και *y*- άξονες μεταξύ της αρχικής και της συμπιεσμένης τροχιάς και την περίμετρο (αντί της επιφάνειας) του παραθύρου επερώτησης. Ακολούθως, αποδεικνύουμε πως το κόστος του υπολογισμού του τύπου που αναπτύξαμε μπορεί να μειωθεί σε μία μικρή επιπρόσθετη επιβάρυνση στο χρησιμοποιούμενο αλγόριθμο συμπίεσης, ενώ αναφέρουμε και πώς το αναλυτικό μοντέλο που αναπτύξαμε συνεισφέρει σε πιο αποτελεσματικούς

αλγόριθμους συμπίεσης. Τέλος, εκτελούμε μια πλήρη σειρά πειραμάτων σε συνθετικά και πραγματικά δεδομένα που καταδεικνύουν την εύκολη εφαρμογή, την ορθότητα και την ακρίβεια της ανάλυσης μας. Αξίζει τον κόπο να σημειώσουμε ότι η προεξέχουσα εφαρμογή του προτεινόμενου μοντέλου βασίζεται στα στοιχεία που μας παρέχει προς την ανάπτυξη πιο αποτελεσματικών αλγορίθμων συμπίεσης σε σχέση με αυτούς που ήδη γνωρίζουμε από τη βιβλιογραφία βάσεων δεδομένων.

Τα αποτελέσματά μας για τα παραπάνω θέματα παρουσιάζονται στο Κεφάλαιο 6. Προκαταρτικά αποτελέσματα έχουν ήδη δημοσιευθεί στο [FT07].

Συνοψίζοντας, οι κύριες συνεισφορές της έρευνάς μας είναι:

- Η ανάπτυξη δύο πρωτότυπων χωροχρονικών ευρετηρίων, που ονομάζονται TB*- και FNR-δέντρο, με το πρώτο να ενισχύει το γνωστό TB-δέντρο προς την υποστήριξη πιο ρεαλιστικών σεναρίων λειτουργίας, ενώ το δεύτερο εκμεταλλεύεται τους περιορισμούς δικτύου, ξεπερνώντας σε απόδοση όλα τα υπόλοιπα προς σύγκριση ευρετήρια.
- Η πρόταση αρκετών κλιμακούμενων και αποδοτικών αλγορίθμων για την αναζήτηση πλησιέστερου γείτονα σε δομές παρόμοιες του R-δέντρου που αποθηκεύουν πληροφορίες ιστορικών τροχιών.
- Η ανάπτυξη δύο αλγορίθμων για αναζήτηση της πιο όμοιας τροχιάς σε δομές που ομοιάζουν σε R-δέντρο και αποθηκεύουν δεδομένα ιστορικών τροχιών. Εδώ, αξίζει τον κόπο να αναφέρουμε ότι η χρήση των προτεινόμενων αλγορίθμων αναζήτησης NN και MST, δίνει τη δυνατότητα σε δομές που ομοιάζουν σε R-δέντρα να υποστηρίξουν ένα ευρύ φάσμα χωροχρονικών επερωτήσεων.
- Η πρόταση ενός αναλυτικού μοντέλου που εκτιμά το αποτέλεσμα της αβεβαιότητας σε επερωτήσεις χρονικής στιγμής για ομοιόμορφα δεδομένα τροχιών, καθώς και η επέκτασή του για να καλύπτει αυθαίρετα κατανεμημένες τροχιές με τη βοήθεια ιστογραμμάτων· το ίδιο μοντέλο επιδεικνύει εξαιρετικές εφαρμογές σε σταθερά χωρικά δεδομένα, ενώ μπορεί να χρησιμοποιηθεί απευθείας στα υφιστάμενα SDBMS.
- Η ανάπτυξη ενός αναλυτικού μοντέλου που εκτιμά τα αποτελέσματα της συμπίεσης τροχιάς σε χωροχρονικές επερωτήσεις.

1.5. Επισκόπηση των Χρησιμοποιούμενων Δεδομένων Τροχιών

Κατά την εκπόνηση αυτής της διατριβής πειραματιστήκαμε με μία ποικιλία πραγματικών και συνθετικών δεδομένων. Πιο συγκεκριμένα, χρησιμοποιήσαμε δύο σύνολα πραγματικών δεδομένων, και διάφορα συνθετικά δεδομένα που παρήχθησαν από την γεννήτρια χωροχρονικών δεδομένων GSTD [TSN99], από μία γεννήτρια δεδομένων υπό περιορισμούς δικτύου της [Bri02] καθώς και μία εξειδικευμένη γεννήτρια τροχιών που αναπτύχθηκε για τους σκοπούς του [FGT07]. Οι λεπτομέρειες των χρησιμοποιούμενων συνόλων δεδομένων δίνονται στον παρακάτω πίνακα (Πίνακας 1.1).

1.5.1. Πραγματικές Τροχιές

Ο προέλευση των δύο πραγματικών συνόλων δεδομένων ήταν από ένα στόλο φορτηγών(*Trucks*) και ένα στόλο σχολικών λεωφορείων (*Buses*), και παρουσιάζονται στο Σχήμα 1.6(a) and (b), αντίστοιχα. Τα δύο πραγματικά σύνολα δεδομένων αποτελούνται από 276 (112203) και 145 (66096) τροχιές

(ευθύγραμμα τμήματα), αντίστοιχα. Και τα δύο σύνολα δεδομένων είναι διαθέσιμα στη διαδικτυακή διεύθυνση http://www.rtreeportal.org.

Πίνακας 1.1: Συγκεντρωτικές πληροφορίες για τα πραγματικά και συνθετικά δεδομένα

Σύνολο Δεδομένων	*# τροχιές*	*# εγγραφές*
Real Data (Trucks)	276	112K
Real Data (Buses)	145	66K
GSTD 100	100	485K
GSTD 250	250	1213K
GSTD 500	500	2426K
GSTD 1000	1000	4850K
GSTD 2000	2000	9701K
NG 200	200	106K
NG 400	400	213K
NG 800	800	417K
NG 1200	1200	626K
NG 1600	1600	831K
NG 2000	2000	1043K

(a) ένας στόλος φορτηγών (b) ένας στόλος σχολικών λεωφορείων (c) συνθετικά δεδομένα από τη γεννήτρια GSTD

Σχήμα 1.6: Δείγματα από πραγματικά και συνθετικά χωροχρονικά δεδομένα

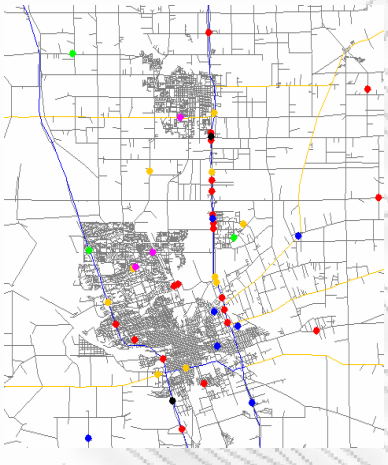
1.5.2. Συνθετικές Τροχιές που Προσομοιώνουν μη Περιορισμένη Κίνηση

Για το σκοπό της παραγωγής τροχιών που κινούνται στον μη περιορισμένο χώρο, χρησιμοποιήσαμε τη γεννήτρια τροχιών δεδομένων GSTD [TSN99]. Ένα δείγμα από τα δεδομένα του GSTD παρουσιάζεται στο Σχήμα 1.6(c). Οι συνθετικές τροχιές που παρήχθησαν με αυτόν τον τρόπο αντιστοιχούν σε 100, 250, 500, 1000 και 2000 κινούμενα αντικείμενα, που καταλήγουν σε σύνολα δεδομένων με 500K, 1250K, 2500K, 5000K, και 10000K εγγραφές από ευθύγραμμα τμήματα (η θέση κάθε αντικειμένου λήφθηκε περίπου 5000 φορές), χτίζοντας επομένως ευρετήρια μεγέθους μέχρι 500 Mbytes (η περίπτωση του 3D R-δένδρου για το GSTD 2000 σύνολο δεδομένων). Σε σχέση με τις υπόλοιπες παραμέτρους της γεννήτριας GSTD, η αρχική κατανομή των σημείων ήταν κανονική (Gaussian), ενώ η κίνησή τους καθοριζόταν από μία τυχαία κατανομή.

1.5.3. Συνθετικές Τροχιές που Προσομοιώνουν Κίνηση υπό Περιορισμούς Δικτύου

Σε σχέση με την περίπτωση όπου τα αντικείμενα είναι περιορισμένα να κινούνται επάνω σε ένα δίκτυο, τα πειράματά μας βασίζονται σε συνθετικά δεδομένα που παρήχθησαν από την γεννήτρια χωροχρονικών δεδομένων υπό περιορισμούς δικτύου που παρουσιάστηκε στο [Bri02], καθώς και το

πραγματικό οδικό δίκτυο του San Joaquin ([Bri02] Σχήμα 1.7). Παραγάγαμε τα σύνολα δεδομένων τροχιών *NG* που αποτελούνται από 200, 400, 800, 1200, 1600 και 2000 κινούμενα αντικείμενα, και όπου η θέση κάθε αντικειμένου λήφθηκε 400 φορές. Ενώ το εξαγόμενο της γεννήτριας ήταν στη μορφή (*id*, *t*, *x*, *y*), στα πειράματά μας θέλαμε να χρησιμοποιήσουμε τέτοια δεδομένα μόνο στην περίπτωση που (*x*, *y*) είναι οι συντεταγμένες ενός κόμβου του δικτύου. Γι αυτό το σκοπό, η γεννήτρια μετατράπηκε έτσι ώστε να παράγει εγγραφές της μορφής (*id*, *t*, *x*, *y*) κάθε φορά που ένα κινούμενο αντικείμενο περνούσε επάνω από έναν κόμβο του δικτύου. Ο μέγιστος όγκος δεδομένων, σε όρους ευθυγράμμων τμημάτων, που παράχθηκε από την εν λόγω γεννήτρια ήταν περίπου 1M εγγραφές και προέκυψε για τα 2000 κινούμενα αντικείμενα.



Σχήμα 1.7: Το πραγματικό οδικό δίκτυο του San Joaquin, μαζί με ένα δείγμα των παραγόμενων δεδομένων

1.6. Περίγραμμα της Διατριβής

Η διατριβή αναπτύσσεται ως εξής: Στο Κεφάλαιο 2 προτείνουμε και αξιολογούμε δύο πρωτότυπα ευρετήρια για χωροχρονικές τροχιές. Τα Κεφάλαια 3 και 4 παρουσιάζουν τους αλγορίθμους που χρησιμοποιούνται για την αναζήτηση πλησιέστερου γείτονα και της ομοιότερης τροχιάς αντίστοιχα, σε δεδομένα ιστορικών τροχιών. Τα Κεφάλαια 5 και 6 εισάγουν αναλυτικά μοντέλα, το πρώτο για την πρόβλεψη του αποτελέσματος της αβεβαιότητας σε χωροχρονικές επερωτήσεις, και το δεύτερο για την εκτίμηση του αποτελέσματος της συμπίεσης τροχιών σε χωροχρονικές επερωτήσεις. Τέλος, με το Κεφάλαιο 7 ολοκληρώνεται η διατριβή σε μία σύνοψη των αποτελεσμάτων και μία παρουσίαση των μελλοντικών ερευνητικών κατευθύνσεων.

2. Δεικτοδότηση Τροχιών

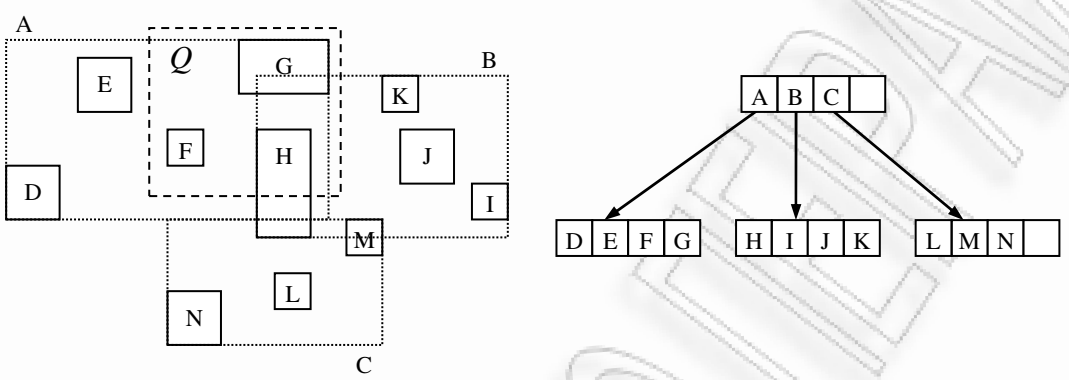
Στο κεφάλαιο αυτό εστιάζουμε στο πρόβλημα της δεικτοδότησης στις βάσεις δεδομένων τροχιών και παρουσιάζουμε τις δύο προτάσεις μας, το TB*-δέντρο και το FNR-δέντρο. Το κεφάλαιο διαμορφώνεται ως εξής: Η Ενότητα 2.1 εισάγει τα θέματα που σχετίζονται με τη δεικτοδότηση χωροχρονικών τροχιών ενώ, η Ενότητα 2.2 εξετάζει τις σχετικές εργασίες. Η Ενότητα 2.3 παρουσιάζει τη δομή και τους αλγορίθμους για τη διατήρηση και την αναζήτηση στο TB*-δέντρο και η Ενότητα 2.4 περιέχει τη δομή και τους αλγορίθμους του FNR-δέντρου. Οι Ενότητες 2.5 και 2.6 παρουσιάζουν την πειραματική μας μελέτη σε απεριόριστο και περιορισμένο από το δίκτυο χώρο, αντίστοιχα, και τέλος η Ενότητα 2.7 ολοκληρώνει το κεφάλαιο με την παρουσίαση των συμπερασμάτων.

2.1. Εισαγωγή

Όπως και στις παραδοσιακές βάσεις δεδομένων, οι επερωτήσεις σε MODs θα μπορούσαν να γίνουν ιδιαίτερα δαπανηρές λόγω της φύσης των δεδομένων και της πολυπλοκότητας των αλγορίθμων επεξεργασίας επερωτήσεων. Δεδομένου δε ότι οι αντιλαμβανόμενες τη θέση συσκευές είναι πλέον σχεδόν πανταχού παρούσες, οι βάσεις δεδομένων τροχιών, αργά ή γρήγορα, θα βρεθούν αντιμέτωπες με τεράστιο όγκο δεδομένων. Κατά συνέπεια προκύπτει ότι η απόδοση παρουσία τεράστιων μεγεθών δεδομένων, θα αποτελέσει ουσιαστικό πρόβλημα για τις βάσεις δεδομένων τροχιών. Δεδομένου ότι η διάταξη απέχει πολύ από τη φύση των γεωγραφικών (πολυδιάστατων) δεδομένων, τα παραδοσιακά ευρετήρια τύπου B-δέντρου δεν χρησιμεύουν στις χωρικές (και κατά συνέπεια στις χωροχρονικές) βάσεις δεδομένων. Στον τομέα των χωρικών βάσεων δεδομένων, το R-δέντρο που προτείνεται από τον Guttman [Gut84] είναι «πανταχού παρόν», με εφαρμογές που καλύπτουν τα Συστήματα Γεωγραφικών Πληροφοριών (Geographical Information Systems - GIS) και το Σχεδιασμό με τη βοήθεια Υπολογιστή (Computer Aided Design - CAD) μέχρι και τα Συστήματα Διαχείρισης Εικόνων και Πολυμέσων (Image and Multimedia Management Systems) [MNPT05].

Το R-δέντρο μπορεί να θεωρηθεί επέκταση του B-δέντρου σε *n*-διάστατους χώρους. Όπως και το B-δέντρο, το R-δέντρο είναι ένα ισοσταθμισμένο δέντρο με τις εγγραφές του ευρετηρίου στα φύλλα του (leaf nodes) που περιλαμβάνουν δείκτες (pointers) προς τα πραγματικά δεδομένα. Οι εγγραφές των φύλλων είναι της μορφής <*id*, *MBB*>, όπου *id* είναι ένας ταυτοποιητής (identifier) που δείχνει στο πραγματικό αντικείμενο και το *MBB* (Ελάχιστο Περιβάλλον Κουτί/Minimum Bounding Box) είναι ένα *n*-διάστατο διάστημα. Οι εγγραφές στους εσωτερικούς κόμβους είναι της μορφής <*ptr*, *MBB*>, όπου *ptr* είναι ένας δείκτης προς τον απόγονο κόμβο (child node) και *MBB* το περιβάλλον κουτί που καλύπτει όλους τους απογόνους κόμβους. Ένας κόμβος στο δέντρο αντιστοιχεί σε μία φυσική σελίδα

δίσκου (ή σε ένα μπλοκ δίσκου που είναι το θεμελιώδες στοιχείο στο οποίο οργανώνεται η δομή αποθήκευσης του δίσκου) και περιέχει μεταξύ *m* και *M* εγγραφές (*M* είναι η χωρητικότητα του κόμβου και *m* είναι μία παράμετρος συντονισμού – συνήθως το *m* ορίζεται στα *M*/2 γεγονός που διασφαλίζει ότι η χρήση του χώρου είναι τουλάχιστον στο 50%). Σε αντίθεση με το B-δέντρο, τα MBBs των κόμβων που ανήκουν στο ίδιο επίπεδο επιτρέπεται να αλληλεπικαλύπτονται. Στο Σχήμα 2.1 βλέπουμε ένα σύνολο χωρικών αντικειμένων και το αντίστοιχο R-δέντρο.



Σχήμα 2.1: Παράδειγμα χωρικών δεδομένων, τα Ελάχιστα Περιβάλλοντα Κουτιά (MBBs) τους, μία επερώτηση εύρους και το αντίστοιχο R-δέντρο [MNPT05].

Στο τομέα της *χωροχρονικής δεικτοδότησης*, οι παραλλαγές και οι επεκτάσεις του R-δέντρου περιλαμβάνουν, μεταξύ άλλων, 3D R-δέντρα [TVS96], TB-δέντρα και STR-δέντρα [PJT00], δέντρα Οκταγωνικού-Πρίσματος (Octagon-Prism (OP)) [ZSI02], PA-δέντρα [NR07], MON-δέντρα [AG05], ενώ το SETI [CEP03] είναι μία υβριδική τεχνική που βασίζεται σε R-δέντρα και τεχνικές διαμέρισης. Θα τα εξετάσουμε διεξοδικά στις επόμενες ενότητες. Επιπλέον, εφόσον το ενδιαφέρον μας στη διατριβή αυτή εστιάζεται σε ιστορικές MODs, περιορίζουμε τη συζήτησή μας σε τεχνικές δεικτοδότησης που καταγράφουν προηγούμενες θέσεις. Για όποιον ενδιαφέρεται για τη δεικτοδότηση τρέχουσας θέσης και ανυσμάτων κίνησης, μερικές αρκετά ενδιαφέρουσες εργασίες είναι οι [SJ02], [SJLL00], [TPS03], και [XP03].

Πίνακας 2.1: Ταξινόμηση χωροχρονικών επερωτήσεων [Pfo02]

Τύπος Επερώτησης		*Λειτουργία*
Επερωτήσεις βάσει συντεταγμένων		επυκαλύπτεται, βρίσκεται μέσα, πλησιέστερος γείτονας, κλπ.
Επερωτήσεις βάσει τροχιάς	Τοπολογικές επερωτήσεις	εισάγεται, αφήνει, διασχίζει, παρακάμπτει, κλπ.
	Επερωτήσεις πλοήγησης	Διανυόμενη απόσταση, καλυπτόμενη περιοχή, ταχύτητα, κατεύθυνση, σταθμευμένο, κλπ.

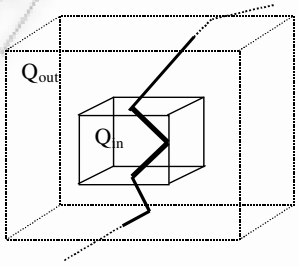
2.1.1. Προδιαγραφές Δεικτοδότησης Τροχιών

Όπως τονίστηκε στο [PJT00], η μεγάλη πλειοψηφία των προτεινόμενων χωροχρονικών ευρετηρίων παραβλέπουν τις προκλήσεις που τίθενται από τη φύση των δεδομένων τροχιάς και δεικτοδοτούν συλλογές γραμμικών τμημάτων στο χωροχρονικό διάστημα, μόνο για την επεξεργασία παραδοσιακών επερωτήσεων *βασισμένων στις συντεταγμένες* (όπως για παράδειγμα επερωτήσεις εύρους και χρονικής

στιγμής), αγνοώντας ταυτόχρονα άλλες χρήσιμες επερωτήσεις, όπως τοπολογικές επερωτήσεις και επερωτήσεις πλοήγησης που είναι *βασισμένες στη τροχιά*. Πιο συγκεκριμένα, επερωτήσεις της μορφής *«βρες όλα τα αντικείμενα που βρίσκονται σε μία συγκεκριμένη περιοχή σε ένα συγκεκριμένο χρονικό διάστημα»* γενικεύουν την επερώτηση χωρικού εύρους της μορφής *«βρες όλα τα αντικείμενα σε μία δεδομένη περιοχή»* και δεν λαμβάνουν υπ' όψιν την έννοια της τροχιάς· γι' αυτό και καλούνται «βάσει συντεταγμένων» (*coordinate-based)* [PJT00]. Οι επερωτήσεις της μορφής «*βρες τις θέσεις όλων των αντικειμένων σε μία δεδομένη περιοχή σε μία συγκεκριμένη χρονική στιγμή*», που ονομάζονται επερωτήσεις *χρονικής στιγμής (timeslice)*, αποτελούν ένα ειδικό τύπο επερωτήσεων εύρους όπου η χρονική έκταση θεωρείται μηδενική. Μία άλλη άμεση επέκταση απλών χωρικών επερωτήσεων στο τομέα των χωροχρονικών εφαρμογών περιλαμβάνει επερωτήσεις *πλησιέστερου γείτονα* της μορφής *«βρες το πλησιέστερο κινούμενο αντικείμενο σε ένα αντικείμενο επερώτησης στη διάρκεια συγκεκριμένου χρονικού διαστήματος»*. Επιπλέον, στην περίπτωση χωροχρονικών επερωτήσεων πλησιέστερου γείτονα, το αντικείμενο της επερώτησης θα μπορούσε να είναι ένα 2D σημείο ή μία ακόμα τροχιά κινούμενου αντικειμένου, ενώ η επερώτηση θα μας έδινε το πλησιέστερο αντικείμενο στην επερώτηση σε οποιαδήποτε χρονική στιγμή ενός χρονικού διαστήματος, ή, σε κάθε χρονική στιγμή του χρονικού διαστήματος της επερώτησης (ιστορικές συνεχείς επερωτήσεις).

Επιπλέον, η [PJT00] προτείνει να ονομάζονται *βάσει τροχιάς* (*trajectory-based)* οι επερωτήσεις που απαιτούν τη γνώση όλης – ή τουλάχιστον ενός τμήματος – της τροχιάς του αντικειμένου προκειμένου να τις επεξεργαστούμε. Τέτοιες επερωτήσεις είναι αυτές που ασχολούνται με τοπολογικές σχέσεις (είσοδος, αναχώρηση κλπ.) και αυτές που παρέχουν προκύπτουσες πληροφορίες για την πλοήγηση ενός αντικειμένου (μέση ταχύτητα, απόσταση που διανύθηκε κλπ.). Ο Πίνακας 2.1 συνοψίζει τους παραπάνω δύο τύπους επερωτήσεων.

Ο συνδυασμός επερωτήσεων εύρους και τοπολογικών επερωτήσεων μας δίνει έναν άλλο τύπο επερωτήσεων που καλούνται *συνδυαστικές επερωτήσεις* (*combined queries).* Ως παράδειγμα [PJT00], ας δούμε την εξής επερώτηση «*Ποιες ήταν οι τροχιές των αντικειμένων αφ' ης στιγμής έφυγαν από την οδό Tucson μεταξύ 7π.μ. και 8 π.μ. σήμερα, για την επόμενη ώρα*», που πρώτα εντοπίζει τις τροχιές που περιέχονται σε ένα εσωτερικό παράθυρο επερώτησης εύρους (*οδός Tucson*, *μεταξύ 7πμ και 8 πμ σήμερα*, Q_{in} στο Σχήμα 2.2) και στη συνέχεια ανακτά τα τμήματα της τροχιάς των αντικειμένων που περιέχονται σε ένα εξωτερικό παράθυρο επερώτησης (*στην επόμενη ώρα*, Q_{out} στο Σχήμα 2.2).



Σχήμα 2.2: Επερωτήσεις συνδυασμένης αναζήτησης

Στα πλαίσια ενός άλλου ερευνητικού πεδίου, η [MB04] ασχολήθηκε με την ανάγκη εξεύρεσης αποτελεσματικών μηχανισμών συμπίεσης τροχιών· βάσει της ίδιας εργασίας, αναμένεται ότι όλες οι αντιλαμβανόμενες τη θέση συσκευές θα αρχίσουν τελικά να παράγουν μία ροή δεδομένων από χωροχρονικές θέσεις άνευ προηγουμένου. Αργά ή γρήγορα ένας τέτοιος τεράστιος όγκος δεδομένων

θα δημιουργήσει προκλήσεις ως προς την αποθήκευση, τη διαβίβαση, τη διαχείριση και την απεικόνιση (display). Συνεπώς προκύπτει η ανάγκη τεχνικών συμπίεσης. Ωστόσο, τα υφιστάμενα χωροχρονικά ευρετήρια που δεν διατηρούν τις τροχιές των κινούμενων αντικειμένων και αντιμετωπίζουν τα χωροχρονικά δεδομένα ως συλλογές γραμμικών τμημάτων στο χώρο των 2+1 διαστάσεων (όπως το SETI [CEP03] και το 3D R-δέντρο [TVS96]), παραβλέπουν την ανάγκη συμπίεσης, κάτι το οποίο εξ' ορισμού σημαίνει ότι κάθε τροχιά θα θεωρείται ένα ολοκληρωμένο αντικείμενο. Η ίδια ανάγκη για διατήρηση της τροχιάς ανακύπτει όταν ασχολούμαστε με διαδικασίες διαγραφής· παρά το γεγονός ότι η διαγραφή ενός γραμμικού τμήματος από τη βάση δεδομένων τροχιών μπορεί να φαίνεται ανούσια, η διαγραφή μιας ολόκληρης τροχιάς είναι μία πολύ χρήσιμη διαδικασία η οποία πρέπει να υποστηρίζεται από οποιοδήποτε ευρετήριο τροχιών στον πραγματικό κόσμο.

Στην [PJT00] παρουσιάζονται δύο δομές ευρετηρίων, το Χωροχρονικό R-δέντρο (STR-δέντρο) και το Trajectory Bundle tree (TB-δέντρο), που προσπαθούν να καλύψουν αυτές τις ανάγκες και να υποστηρίξουν αποτελεσματικά διαδικασίες βασισμένες στην έννοια της τροχιάς όπως την επεξεργασία τοπολογικών επερωτήσεων. Το αποτέλεσμα αυτής της εργασίας είναι ότι το TB-δέντρο θα μπορούσε να υποστηρίξει και μη παραδοσιακές επερωτήσεις πολύ πιο αποτελεσματικά από το παραδοσιακό 3D R-δέντρο και το STR-δέντρο. Δυστυχώς, παρά τα σαφή του πλεονεκτήματα σε επεξεργασία επερωτήσεων που βασίζονται στην τροχιά, το TB-δέντρο έχει ένα πολύ ουσιαστικό μειονέκτημα: λόγω της στρατηγικής που υιοθετεί για την εισαγωγή νέων δεδομένων τροχιών, αυτά εισάγονται πάντα στη δεξιά «άκρη» του δέντρου, που σημαίνει ότι η απόδοσή του εξαρτάται σε πολύ μεγάλο βαθμό από την σειρά εισαγωγής των δεδομένων. Παρόλα αυτά, σε πραγματικές εφαρμογές, αυτή η υπόθεση δεν είναι πάντα απαραίτητα σωστή. Επί παραδείγματι, σε μία εφαρμογή όπου οι εισαγωγές γίνονται σε πραγματικό χρόνο, αν το κινούμενο αντικείμενο εισέλθει σε μία περιοχή όπου το σύστημα μετάδοσης θέσης δεν λειτουργεί, η τροχιά του θα μπορούσε να αποθηκευθεί τοπικά στο αντικείμενο και να διαβιβασθεί στον κεντρικό εξυπηρετητή – όπου και λειτουργεί το ευρετήριο - αργότερα· εντωμεταξύ, άλλα κινούμενα αντικείμενα θα μπορούσαν να έχουν διαβιβάσει τη θέση τους, παραβιάζοντας έτσι την παραπάνω υπόθεση του TB-δέντρου. Επιπλέον, η δομή του TB-δέντρου δεν είναι κατάλληλη για να υποστηρίξει λειτουργίες διαγραφής και συμπίεσης· τυχόν διαγραφή μιας τροχιάς θα άφηνε «κενά» στους κόμβους ενώ η συμπίεση τροχιών όπως θα συζητήσουμε στη συνέχεια, σημαίνει ότι το ευρετήριο θα πρέπει να χειρίζεται δεδομένα που εισάγονται με μη χρονολογική σειρά.

Μία άλλη προσέγγιση που είναι αρκετά ενδιαφέρουσα σχετικά με τη δεικτοδότηση χωροχρονικών τροχιών, προκύπτει αναγνωρίζοντας ότι οι τροχιές κατά πάσα πιθανότητα θα είναι περιορισμένες σε ένα δίκτυο. Όπως επισημαίνεται στην [KGT99], η ύπαρξη περιορισμών στο χώρο στον οποίο τα κινούμενα αντικείμενα πραγματοποιούν την κίνησή τους είναι μία προϋπόθεση που μπορεί να χρησιμοποιηθεί για τη βελτίωση της απόδοσης χωροχρονικών ευρετηρίων. Στην πράξη, αυτό συμβαίνει στις περισσότερες πραγματικές εφαρμογές: αεροπλάνα που κινούνται σε αεροδιάδρομους, αυτοκίνητα και πεζοί που κινούνται σε οδικά δίκτυα, ενώ τα τρένα έχουν σταθερές τροχιές σε σιδηροδρομικά δίκτυα. Αυτές οι ειδικές συνθήκες (περιορισμοί στην κίνηση) αποτέλεσαν αντικείμενο ερευνητικού ενδιαφέροντος στις [KGT99], [PTKZ02]. Πιο συγκεκριμένα, σύμφωνα με τους Kollios et al. [KGT99], το πεδίο αναφοράς της τροχιάς των κινούμενων αντικειμένων σε ένα

δίκτυο δεν είναι ο χώρος των 2+1 διαστάσεων, αλλά ένας χώρος με 1.5 διαστάσεις, καθώς τα γραμμικά τμήματα που περιλαμβάνουν το δίκτυο μπορούν να αποθηκευτούν σε ένα συμβατικό ευρετήριο χωρικών δεδομένων (όπως το R-δέντρο). Τότε, η δεικτοδότηση των αντικειμένων που κινούνται σε ένα τέτοιο δίκτυο δεν είναι παρά ένα 1D πρόβλημα. Στην [KGT99], το πρόβλημα της δεικτοδότησης τροχιών με περιορισμούς στο δίκτυο μελετάται υπό μία πιο θεωρητική σκοπιά χωρίς να προτείνεται μία μέθοδος προσπέλασης που θα μπορούσε να χρησιμοποιηθεί σε πραγματικές εφαρμογές. Από την άλλη πλευρά, ακολουθώντας τις οδηγίες που δίνονται στην [KGT99], στις επόμενες ενότητες, δείχνουμε πως η πρόταση αυτή μπορεί να υλοποιηθεί και να χρησιμοποιηθεί για την ανάπτυξη πρωτότυπων μεθόδων προσπέλασης που χρησιμεύουν για τη δεικτοδότηση δεδομένων τροχιών με περιορισμούς στο δίκτυο.

2.1.2. Τι προτείνεται

Για την αντιμετώπιση των παραπάνω απαιτήσεων, στην παρούσα διατριβή, προτείνονται ανεξάρτητα δύο πρωτότυπα ευρετήρια, ήτοι, το TB*-δέντρο και το FNR-δέντρο. Πιο συγκεκριμένα, το TB*-δέντρο είναι επέκταση του TB-δέντρου που ξεπερνά το βασιμό μειονέκτημα του προκατόχου του, δηλαδή, την υποστήριξη της ανάγκης για *μη χρονολογικές εισαγωγές*, διατηρώντας ταυτόχρονα όλες τις «επιθυμητές» του ιδιότητες. Επιπλέον, πέρα από τους αλγορίθμους που παρέχονται για τη δόμησή του και την επεξεργασία επερωτήσεων, το TB*-δέντρο υποστηρίζει *διαγραφές τροχιών*, ενώ η δομή του επιτρέπει και την υποστήριξη αλγορίθμων *συμπίεσης τροχιών*. Η δομή και οι αλγόριθμοι του TB*-δέντρου θα περιγραφούν στις επόμενες ενότητες και θα ακολουθήσει μία πειραματική μελέτη που αποκαλύπτει τις θετικές και αρνητικές πτυχές του προτεινόμενου ευρετηρίου. Εδώ είναι απαραίτητο να διευκρινίσουμε ότι το προτεινόμενο TB*-δέντρο, δεν εκμεταλλεύεται τις ειδικές συνθήκες τις οποίες έχουν τα αντικείμενα όταν κινούνται σε σταθερά δίκτυα· αντιθέτως, δεικτοδοτεί αντικείμενα που κινούνται ελεύθερα στο 2D χώρο.

Από την άλλη μεριά, κάτω από το σενάριο των περιορισμών δικτύου η παρούσα διατριβή προτείνει ένα καινοτόμο ευρετήριο, που ονομάζεται R-δέντρο Σταθερού Δικτύου (Fixed Network R-tree - FNR-tree) και είναι μία προέκταση του πολύ γνωστού R-δέντρου [Gut84]. Μπορούμε να περιγράψουμε το FNR-δέντρο σαν ένα δάσος από 1D R-δέντρα πάνω από ένα 2D R-δέντρο. Το 2D R-δέντρο χρησιμοποιείται για τη δεικτοδότηση των χωρικών δεδομένων του διαγράμματος του δικτύου (π.χ. δρόμων που αποτελούνται από γραμμικά τμήματα), ενώ τα 1D R-δέντρα χρησιμοποιούνται για τη δεικτοδότηση του χρονικού διαστήματος της κίνησης κάθε αντικειμένου σ' ένα δεδομένο τμήμα του δικτύου. Όπως θα δείξουμε πειραματικά στις επόμενες ενότητες, το προτεινόμενο FNR-δέντρο υπερτερεί τόσο του TB- όσο και του 3D R-δέντρου σε γενικές επερωτήσεις που βασίζονται στις συντεταγμένες· ωστόσο αυτή η αποτελεσματικότητα του FNR-δέντρου σε επερωτήσεις που βασίζονται στις συντεταγμένες επιβαρύνεται από το γεγονός ότι δεν υπάρχει ένας μηχανισμός διατήρησης της τροχιάς και συνεπώς δεν υπάρχει η δυνατότητα υποστήριξης επερωτήσεων που είναι βασισμένες στη τροχιά.

2.2. Σχετικές Εργασίες

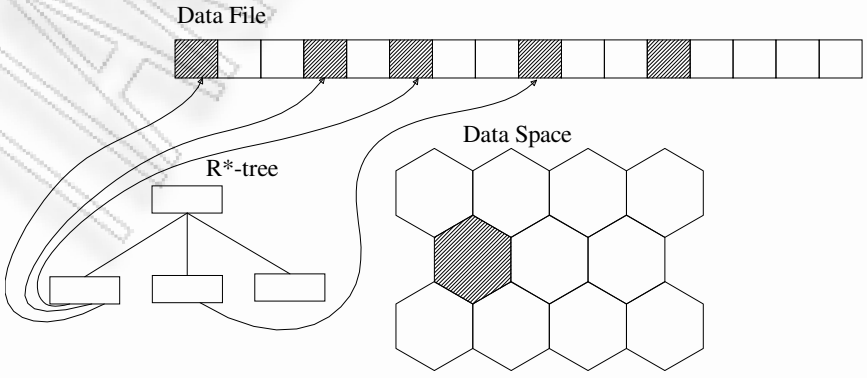
Στη συνέχεια, εξετάζουμε εν συντομία τις εργασίες που σχετίζονται με τον τομέα της δεικτοδότησης ιστορικών τροχιών κινούμενων αντικειμένων. Θα πρέπει εδώ να επισημάνουμε ότι δεν

περιλαμβάνουμε όλες αυτές τις δομές στην πειραματική μας μελέτη δεδομένου ότι στην πλειονότητά τους προτάθηκαν στη διάρκεια της εκπόνησης της παρούσας διατριβής· ωστόσο, κάποιες από τις εργασίες που εξετάζουμε παραθέτουν στοιχεία και συγκρίνονται με μία πρώιμη εκδοχή του FNR-δέντρου που παρουσιάζεται στην [Fre03], ενώ οι υπόλοιπες αξιολογούνται επίσης σε αντιπαράθεση με το αρχικό 3D R-δέντρο [TVS96] και το TB-δέντρο [PJT00]. Αναφερόμαστε πρώτα σε δομές που δεικτοδοτούν αντικείμενα που κινούνται σε απεριόριστο χώρο, ενώ κατόπιν, παρουσιάζουμε κάποιες προσεγγίσεις με περιορισμούς δικτύου.

2.2.1. Δεικτοδότηση Τροχιών Κινούμενων Αντικειμένων σε Απεριόριστο Χώρο

Οι Zhu et al. [ZSI02] πρότειναν μια πρώτη βελτίωση του TB-δέντρου προτείνοντας το *Δέντρο Οκταγωνικού Πρίσματος* (OP-δένδρο)· τα OP-δέντρα χρησιμοποιούν προσεγγίσεις οκταγώνου αντί για MBBs. Βάσει των πειραμάτων που διεξήχθησαν, αποδεικνύεται ότι τα OP-δέντρα υπερέχουν του αρχικού TB-δέντρου τόσο ως προς επερωτήσεις βασισμένες στις συντεταγμένες όσο και βασισμένες στη τροχιά. Θα πρέπει εδώ να σταθούμε στο γεγονός ότι οι τροποποιήσεις που πτροτείνουμε στις επόμενες παραγράφους στη δομή TB*-δέντρου σε σχέση με το αρχικό TB-δέντρο (δηλαδή αντικατάσταση 3D γραμμικών τμημάτων από 3D σημεία και η τροποποιημένη στρατηγική εισαγωγής) μπορούν να εφαρμόζονται απευθείας στα πλαίσια του OP-δέντρου αντικαθιστώντας τις προσεγγίσεις MBB με οκτάγωνα.

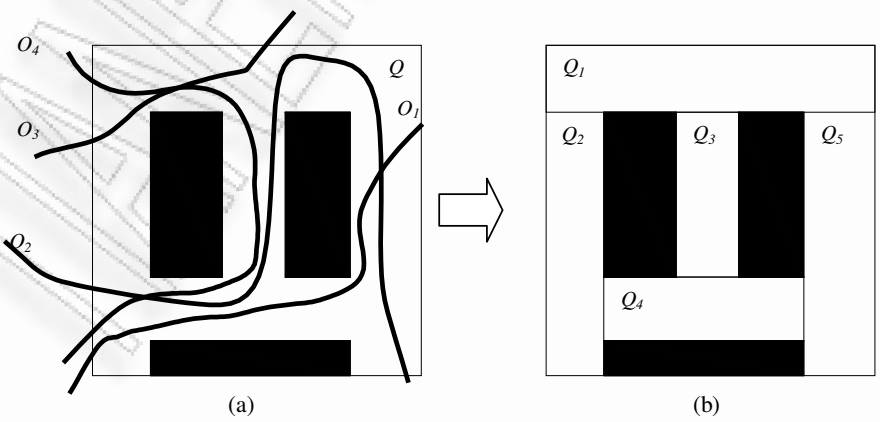
Το *Κλιμακούμενο και Αποδοτικό Ευρετήριο Τροχιάς (Scalable and Efficient Trajectory Index* - SETI) που παρουσιάζεται στο [CEP03] είναι μία υβριδική δομή που δεικτοδοτεί τροχιές σε δύο επίπεδα ξεχωρίζοντας τη χωρική από τη χρονική διάσταση. Το SETI χωρίζει το 2D χώρο σε ξένα μεταξύ τους εξαγωνικά κελιά που παραμένουν σταθερά στη διάρκεια ζωής της δομής (ενώ κι άλλες στρατηγικές χωρικής διαμέρισης θα μπορούσαν να χρησιμοποιηθούν) διότι αναγνωρίζει ότι οι τροχιές των κινούμενων αντικειμένων επεκτείνονται συνεχώς προς τη χρονική διάσταση ενώ τα χωρικά όρια παραμένουν σταθερά ή μεταβάλλονται σπανίως. Κάθε κελί σε λογικό επίπεδο περιλαμβάνει μόνο τα τμήματα της τροχιάς που είναι πλήρως εντός του κελιού, ενώ στην περίπτωση ενός τμήματος τροχιάς που διασχίζει τα όρια του κελιού, χωρίζεται και στη συνέχεια εισάγεται και στα δύο κελιά. Σε φυσικό επίπεδο, τα τμήματα της τροχιάς εισάγονται σε ένα αρχείο δεδομένων· κάθε σελίδα του αρχείου περιλαμβάνει τμήματα από μόνο ένα κελί. Κατόπιν, αντιστοιχίζεται το καθορισμένο κελί με ένα χρονικό ευρετήριο, (ήτοι, ένα 1D R-δέντρο) που δεικτοδοτεί τα χρονικά διαστήματα κάθε συγκεκριμένου κελιού στο αρχείο δεδομένων. Το Σχήμα 2.3 συνοψίζει τη δομή SETI.



Σχήμα 2.3: Η δομή SETI [CEP03]

Οι αλγόριθμοι εισαγωγής και αναζήτησης ακολουθούν μία προσέγγιση πολλαπλών βημάτων που αποτελείται από χωρική επιλογή, χρονική επιλογή και διαλογή. Πιο συγκεκριμένα, στη διάρκεια κάθε εισαγωγής, ο αλγόριθμος εντοπίζει το κελί στο οποίο πρέπει να εισαχθεί το τμήμα (λαμβάνοντας υπ' όψιν και πιθανούς διαχωρισμό του τμήματος μεταξύ κελιών) και κατόπιν το εισάγει στην αντίστοιχη σελίδα του αρχείου δεδομένων, ενημερώνοντας ταυτόχρονα την αντίστοιχη εγγραφή του 1D R-δέντρου (εάν απαιτείται). Παρά το γεγονός ότι όπως παρουσιάστηκε στην πειραματική μελέτη του [CEP03], το SETI σαφώς υπερέχει του 3D R-δέντρου και του TB-δέντρου σε επερωτήσεις χρονικού διαστήματος και χρονικής στιγμής, δεν μπορεί να χρησιμοποιηθεί για την επεξεργασία επερωτήσεων που είναι βασισμένες στη τροχιά. Αυτό οφείλεται στο γεγονός ότι τα γραμμικά τμήματα των τροχιών οργανώνονται εντός του ευρετηρίου μόνο βάσει των χωρικών και χρονικών τους σχέσεων· επομένως, τα διαδοχικά γραμμικά τμήματα της ίδιας τροχιάς μπορούν να τοποθετηθούν σε διαφορετικές σελίδες του δίσκου. Συνεπώς, στην χειρότερη περίπτωση για την ανάκτηση μιας τροχιάς θα πρέπει να διαβάσουμε μία σελίδα δίσκου για κάθε γραμμικό τμήμα τροχιάς. Επιπλέον, η εργασία του [CEP03] δεν δίνει κανέναν αλγόριθμο επεξεργασίας επερώτησης πλησιέστερου γείτονα, ενώ η ανάπτυξη ενός αποτελεσματικού αλγορίθμου για την υποστήριξη τέτοιων επερωτήσεων δεν είναι και τόσο απλή.

Οι Pfoser et al. [PJ01] χρησιμοποιούν τους περιορισμούς που τίθενται στην κίνηση των αντικειμένων από την υφιστάμενη υποδομή για τη βελτίωση της απόδοσης των χωροχρονικών επερωτήσεων δεδομένου ενός υφιστάμενου χωροχρονικού ευρετηρίου· η στρατηγική τους δεν επηρεάζει τη δομή του ίδιου του ευρετηρίου. Αντίθετα, η [PJ01] υιοθετεί ένα επιπλέον βήμα προεπεξεργασίας πριν την εκτέλεση κάθε επερώτησης. Εν προκειμένω, δεδομένου ότι η υποδομή (π.χ. κτήρια) σπανίως ενημερώνεται, μπορεί να δεικτοδοτηθεί μέσω ενός συμβατικού χωρικού ευρετηρίου όπως για παράδειγμα το R-δέντρο. Αφ' ετέρου, για τη δεικτοδότηση τροχιών κινούμενων αντικειμένων θα μπορούσε να χρησιμοποιηθεί ένα γενικό χωροχρονικό ευρετήριο, όπως το TB-δέντρο [PJT00] ή το 3D R-δέντρο [TVS96]. Έτσι, ένα βήμα προ επεξεργασίας της επερώτησης, διαιρεί το αρχικό παράθυρο σε αρκετά μικρότερα παράθυρα, από τα οποία έχουν αποκλεισθεί οι περιοχές που καλύπτονται από την υποδομή (Σχήμα 2.4). Κάθε μία από τις μικρότερες επερωτήσεις εκτελείται στο (γενικό χωροχρονικό) ευρετήριο και μας δίνει μία σειρά από υποψήφια αντικείμενα, των οποίων γίνεται διαλογή βάσει του αρχικού παραθύρου επερώτησης.



Σχήμα 2.4: Το αρχικό παράθυρο επερώτησης *Q* (a) αναλύεται σε μία σειρά από μικρότερα παράθυρα επερωτήσεων *Q1*, *Q2*,.. (b) σε σχέση με τα στοιχεία υποδομής (μαύρα πλαίσια) [PJ01].

Στην ίδια εργασία [PJ01], δίνεται ένας αλγόριθμος για την εφαρμογή του βήματος προ επεξεργασίας, βάσει των όσων παρουσιάστηκαν στην [KF93]. Σύμφωνα με την [KF93], ο αριθμός των προσπελάσεων κόμβων που απαιτούνται από ένα R-δέντρο για να απαντήσουμε μία επερώτηση παραθύρου, εξαρτάται από την επιφάνεια του παραθύρου και από το μήκος του ανά διάσταση. Συνεπώς, αυτό που μας απασχολεί δεν είναι μόνο η ελαχιστοποίηση της επιφάνειας του παραθύρου επερώτησης (που επιτυγχάνεται αφαιρώντας το τμήμα που περιέχει την υποδομή από το αρχικό παράθυρο) αλλά και η ελαχιστοποίηση της περιμέτρου του. Στην αντίστοιχη αξιολόγηση, έγινε σύγκριση της απόδοσης των δύο χωροχρονικών ευρετηρίων (TB- και 3D R-δέντρο) με χρήση του βήματος προ επεξεργασίας (δηλαδή διαιρώντας το αρχικό παράθυρο σε μικρότερα παράθυρα) και χωρίς, και αποδείχθηκε ότι η απόδοση της επερώτησης βελτιώθηκε και για τα δύο ευρετήρια με τη χρήση του βήματος αυτού.

Προσφάτως, μελετήθηκε πως μπορούμε να χωρίσουμε τροχιές με βέλτιστο τρόπο για τη βελτίωση της απόδοσης επερωτήσεων εύρους [HKTG02], [HKTG06]. Οι Hadjieleftheriou et al. [HKTG02] χρησιμοποιούν μία μερικώς σταθερή δομή, το PPR-δέντρο, που προσπαθεί να αντιμετωπίσει το πρόβλημα του νεκρού χώρου που γεννάται από τις προσεγγίσεις των τροχιών κινούμενων αντικειμένων από MBBs. Ο νεκρός χώρος ορίζεται ως ο χώρος σε μία προσέγγιση MBB που δεν καλύπτει στην πράξη κανένα αντικείμενο που περιέχεται σε αυτό. Οι [HKTG02] εισάγουν τεχνητές ενημερώσεις της θέσης των αντικειμένων που χωρίζουν τις τροχιές σε μικρότερα στοιχεία, μειώνοντας με αυτόν τον τρόπο το νεκρό χώρο· χρησιμοποιούν μη γραμμικές συναρτήσεις για την περιγραφή τροχιών κινούμενων αντικειμένων, που αρχικώς δεικτοδοτούνται από το PPR-δέντρο. Η μελέτη αυτή επεκτείνεται στην [HKTG06] όπου αντί του PPR-δέντρου χρησιμοποιείται ένα Multi-Version R-δέντρο, όπως αυτό που προτείνεται στην [TPS03], που οδηγεί σε ένα σχήμα δεικτοδότησης βελτιωμένης απόδοσης. Επίσης, οι προτεινόμενοι αλγόριθμοι για την αντιμετώπιση του προβλήματος του νεκρού χώρου που εισάγονται στα MBBs μπορούν να χρησιμοποιηθούν σε συνδυασμό με οποιοδήποτε χωροχρονικό αρχείο δεδομένων, όπως το R-δέντρο και τις παραλλαγές του.

Ωστόσο, η πιο πολλά υποσχόμενη προσέγγιση ως προς τη δεικτοδότηση τροχιών κινούμενων αντικειμένων σε απεριόριστο χώρο είναι αυτή που παρουσιάζεται στην [NR07]· σύμφωνα με την [NR07], επειδή τα MBBs δεν μπορούν να συλλάβουν την ομαλότητα των πραγματικών δεδομένων τροχιών, προτείνεται οι τροχιές να προσεγγίζονται μέσω μιας ακολουθίας συναρτήσεων κίνησης που περιγράφονται από ένα απλό συνεχές πολυώνυμο. Στη συνέχεια εισάγεται το PA-δέντρο, ένα παραμετρικό ευρετήριο που δεικτοδοτεί τα πολυώνυμα που προκύπτουν· τα PA-δέντρα ομοιάζουν με R-δέντρα, με τη κύρια διαφορά ότι οι εγγραφές αποτελούνται από πολυωνυμικούς συντελεστές, αντί των MBBs. Βάσει της πειραματικής μελέτης που παρουσιάστηκε, το PA-δέντρο υπερέχει τόσο του MVR-δέντρου [HKTG06] όσο και του SETI [CEP03] στην πλειονότητα των πειραματικών ρυθμίσεων.

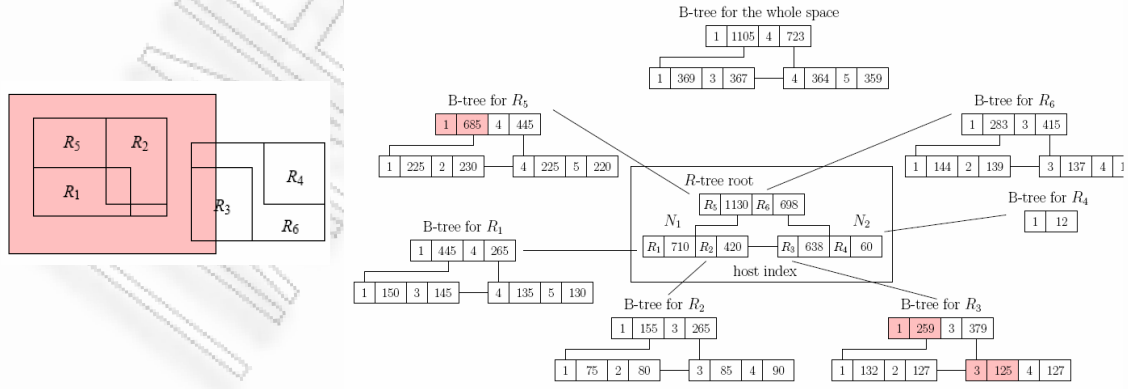
2.2.2. Δεικτοδότηση Τροχιών Κινούμενων Αντικειμένων σε Σταθερά Δίκτυα

Η πρώτη πρόταση που λαμβάνει υπόψη κινούμενα αντικείμενα υπό περιορισμούς δικτύου ήταν η εργασία των Papadias et al. σε [PTKZ02] που υιοθέτησε αυτήν την υπόθεση, προκειμένου να δημιουργήσει μια δομή που απαντά σε χωροχρονικές συγκεντρωτικές επερωτήσεις του τύπου «*βρες το συνολικό αριθμό αντικειμένων στις περιοχές που τέμνουν κάποια παράθυρα q_s στη διάρκεια ενός χρονικού διαστήματος q_t*». Όπως και στο FNR-δέντρο, το προτεινόμενο συγκεντρωτικό R-B-δέντρο

(aRB-δέντρο) ακολουθεί την πρόταση της [KGT99] και δίνει ένα συνδυασμό R- και B-δέντρων βασισμένα στην εξής ιδέα: οι γραμμές στο δίκτυο αποθηκεύονται μόνο μία φορά και δεικτοδοτούνται από ένα R-δέντρο. Κατόπιν, σε κάθε εσωτερικό κόμβο και κόμβο φύλλου του R-δέντρου, τοποθετείται ένας δείκτης, που αποθηκεύει ιστορικά συγκεντρωτικά δεδομένα για το συγκεκριμένο χωρικό αντικείμενο (π.χ. το MBB του κόμβου).

Πιο συγκεκριμένα, η προσέγγιση αυτή βασίζεται σε δύο τύπους ευρετηρίων: *ένα ευρετήριο υποδοχής* (*host index)*, που διαχειρίζεται τις χωρικές επιφάνειες της περιοχής και συσχετίζει μία συγκεντρωτική πληροφορία σε αυτές τις περιοχές με όλα τα χρονικά αποτυπώματα στην βασική σχέση και κάποια *ευρετήρια μέτρων (measure indexes)* (ένα για κάθε εγγραφή του ευρετηρίου υποδοχής), που είναι συγκεντρωτικές χρονικές δομές και αποθηκεύουν τις τιμές των μετρήσεων στο πέρασμα της χρόνου. Για μία σειρά σταθερών περιοχών, οι γράφοντες καθορίζουν το *συγκεντρωτικό R-B-δέντρο* (aRB-δέντρο), που υιοθετεί ένα R-δέντρο με σύνοψη πληροφοριών ως ευρετήριο υποδοχής και ένα B-δέντρο που περιλαμβάνει χρονικώς μεταβαλλόμενα συγκεντρωτικά δεδομένα, ως ευρετήριο μέτρων.

Όπως έχουμε ήδη αναφέρει, το aRB-δέντρο είναι κατάλληλο για την αποτελεσματική επεξεργασία *συγκεντρωτικών επερωτήσεων παραθύρου* δηλαδή, για τον υπολογισμό του συγκεντρωτικού μέτρου των περιοχών που τέμνουν ένα συγκεκριμένο παράθυρο. Πράγματι, για κόμβους που περικλείονται εντελώς στην επερώτηση παραθύρου, το συγκεντρωτικό μέτρο διατίθεται ήδη και αποφεύγεται έτσι η κάθοδος σε αυτούς τους κόμβους. Ως συνέπεια, η συγκεντρωτική επεξεργασία γίνεται ταχύτερα. Επί παραδείγματι, ας υπολογίσουμε τον αριθμό των τηλεφωνημάτων εντός της σκιασμένης περιοχής στο Σχήμα 2.5(a) σε ένα χρονικό διάστημα $[T_1,T_3]$ χρησιμοποιώντας το aRB-δέντρο που φαίνεται στο Σχήμα 2.5(b). Επειδή το R_5 περιλαμβάνεται πλήρως στην επερώτηση παραθύρου δεν υφίσταται απαίτηση ανάλυσης των R_1 και R_2 εξ' ου και προσπελαύνουμε το B-δέντρο για το R_5. Η πρώτη εγγραφή της ρίζας αυτού του B-δέντρου περιλαμβάνει το μέτρο για το διάστημα $[T_1, T_3]$ που είναι η τιμή που μας ενδιαφέρει. Αντ' αυτού, για να έχουμε το σύνολο των τηλεφωνικών κλήσεων στο διάστημα $[T_1, T_3]$ για το R_3 πρέπει να δούμε μια εγγραφή της ρίζας του B-δέντρου για το R_3 και ένα φύλλο (οι έγχρωμοι κόμβοι). Το Σχήμα 2.5 δίνει ένα παράδειγμα της δομής του aRB-δέντρου.



(a) (b)

Σχήμα 2.5: (a) Παράδειγμα δεδομένων και (b) το αντίστοιχο aRB-tree [PTKZ02]

Η [AG05] προτείνει μια παραλλαγή του FNR-δέντρου, που ονομάζεται *Δέντρο Κινούμενων Αντικειμένων σε Δίκτυα/Moving Objects in Networks tree* (MON-δέντρο) και εκμεταλλεύεται την ίδια

ιδιότητα του χωρικού δικτύου. Αντί να χρησιμοποιεί ένα 1D R-δέντρο για κάθε φύλλο του 2D R-δέντρου, το MON-δέντρο χρησιμοποιεί ένα 2D R-δέντρο για κάθε πολυγραμμή του χωρικού δικτύου. Το MON-δέντρο υπερτερεί σημαντικά του 3D R-δέντρου και του FNR-δέντρου, σε επερωτήσεις χρονικού διαστήματος και χρονικής στιγμής και προς το παρόν θεωρείται τεχνολογία αιχμής. Παρόλα αυτά, και εδώ διαπιστώνεται το μειονέκτημα της μη ύπαρξης δυνατότητας αποτελεσματικής επεξεργασίας επερωτήσεων που βασίζονται στη τροχιά.

Μία άλλη ενδιαφέρουσα μεθοδολογία για το ίδιο θέμα (δηλαδή, τη δεικτοδότηση κινούμενων αντικειμένων σε δίκτυα) παρουσιάζεται στην [PJ03]. Η εν λόγω προσέγγιση προτείνει το μετασχηματισμό του σχετικού δικτύου από δύο στη μία διάσταση μέσω ταξινόμησης των ακμών του δικτύου βάσει των τιμών Hilbert. Οι τιμές Hilbert είναι μία προσέγγιση για τη διάταξη του 2D χώρου· καθορίζονται εφαρμόζοντας μία καμπύλη Hilbert, που καλύπτει όλο το 2D χώρο, αντιστοιχώντας κάθε 2D σημείο σε ένα 1D σημείο [WD04]. Έτσι, το πρόβλημα της δεικτοδότησης τριών (δηλαδή 2 χωρικών και 1 χρονικής) διαστάσεων περιορίζεται στο πρόβλημα της δεικτοδότησης δύο (δηλαδή 1 χωρικής και 1 χρονικής) διαστάσεων, που μπορεί να αντιμετωπιστεί αποτελεσματικά χρησιμοποιώντας οποιοδήποτε από τα γνωστά απλά χωρικά ευρετήρια όπως το R-δέντρο που υποστηρίζεται από τα υπάρχοντα DBMS. Μετά από αυτό, κάθε επερώτηση εύρους πρέπει να μετασχηματιστεί στον 1D χώρο, δίνοντας μας έτσι κάποια 2D (χωρικά και χρονικά) ορθογώνια παραλληλόγραμμα, τα οποία στη συνέχεια αντιπαρατίθενται στο R-δέντρο. Η τεχνική χρησιμοποιεί επίσης ένα R-δέντρο για τη δεικτοδότηση του σχετικού δικτύου για να επιταχύνουμε τη διαδικασία μετασχηματισμού της επερώτησης. Η πειραματική μελέτη που παρουσιάζεται στην [PJ03] αποδεικνύει ότι η προτεινόμενη μέθοδος υπερέχει σαφώς της 3D προσέγγισης (π.χ. το 3D R-δέντρο, που αντιμετωπίζει το χρόνο ως μία επιπλέον χωρική διάσταση) καθώς αυξάνεται το μέγεθος της επερώτησης· η αντίστοιχη πειραματική μελέτη δεν περιλαμβάνει ούτε το FNR- ούτε το MON-δέντρο. Επιπλέον, δεν υπάρχει προφανής τρόπος ως προς το πώς η συγκεκριμένη προσέγγιση [PJ03] μπορεί να επεξεργαστεί επερωτήσεις βασισμένες στη τροχιά.

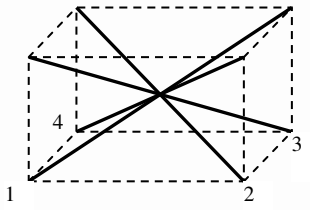
2.3. Δεικτοδότηση Τροχιών Κινούμενων Αντικειμένων σε Απεριόριστο Χώρο

Πριν περιγράψουμε διεξοδικά τη δομή και τους αλγορίθμους του TB*-δέντρου, είναι απαραίτητο να κάνουμε μία σύντομη εισαγωγή για το απλό TB-δέντρο στο οποίο βασίζεται.

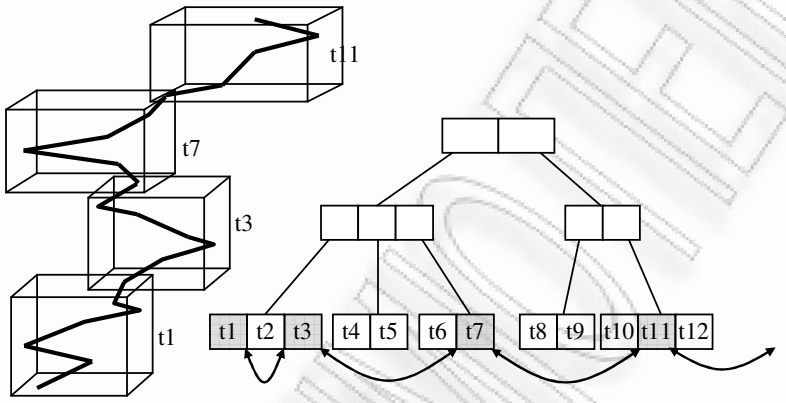
2.3.1. Το TB-δέντρο

Στην πράξη, το πρώτο ευρετήριο που προτάθηκε για την υποστήριξη επερωτήσεων βασισμένων στην τροχιά ήταν το Trajectory Bundle δέντρο (TB-δέντρο) [PJT00], που διαφέρει ουσιαστικά από τις υπόλοιπες μεθόδους χωροχρονικής δεικτοδότησης κυρίως λόγω της στρατηγικής εισαγωγής και διάσπασης. Όπως και το αρχικό R-δέντρο, το TB-δέντρο είναι ένα ισοσταθμισμένο δέντρο με τις εγγραφές στα φύλλα, με την ιδιαιτερότητα ότι περιέχουν εγγραφές από την ίδια τροχιά και είναι της μορφής <*MBB, Orientation*>, όπου *MBB* είναι το 3D περιβάλλον κουτί του 3D γραμμικού τμήματος που ανήκει στην τροχιά ενός αντικειμένου (και θεωρεί το χρόνο ως την τρίτη διάσταση) και *Orientation* είναι ένα σημάδι (*flag*) που χρησιμοποιείται για την ανακατασκευή του πραγματικού 3D γραμμικού τμήματος εντός του MBB μεταξύ των τεσσάρων δυνατών εναλλακτικών λύσεων (δες

Σχήμα 2.6). Επειδή κάθε φύλλο περιλαμβάνει εγγραφές της ίδιας τροχιάς, η *ταυτότητα (id)* του αντικειμένου μπορεί να αποθηκευτεί μία φορά στην κεφαλίδα του φύλλου.



Σχήμα 2.6: Εναλλακτικοί τρόποι με τους οποίους ένα 3D γραμμικό τμήμα μπορεί να περιέχεται σε ένα MBB



Σχήμα 2.7: Η δομή του TB-δέντρου

Αντίθετα με την πλειονότητα των παραλλαγών του R-δέντρου, ο αλγόριθμος εισαγωγής του TB-δέντρου δεν βασίζεται στις χωρικές και χρονικές σχέσεις των κινούμενων αντικειμένων αλλά εξαρτάται μόνο από την ταυτότητα (*id*) του κινούμενου αντικειμένου. Όταν εισάγονται νέα γραμμικά τμήματα, ο αλγόριθμος αναζητά το φύλλο που περιλαμβάνει την τελευταία εγγραφή της ίδιας τροχιάς και απλώς καταχωρεί σε αυτό τη νέα εγγραφή, σχηματίζοντας με αυτό τον τρόπο φύλλα που περιλαμβάνουν γραμμικά τμήματα από μία τροχιά. Εάν το φύλλο είναι πλήρες, τότε δημιουργείται ένα καινούργιο και εισάγεται στη δεξιά άκρη του δέντρου. Για κάθε τροχιά, μία διπλά συνδεδεμένη λίστα συνδέει τα φύλλα που περιλαμβάνουν τα τμήματα της μεταξύ τους (Σχήμα 2.7), οδηγώντας μας σε μία δομή που μπορεί να απαντήσει σε επερωτήσεις βασισμένες στη τροχιά αποτελεσματικά.

Από την άλλη πλευρά, το TB-δέντρο αποδίδει μέτρια σε επερωτήσεις εύρους όπως αποδεικνύεται στην [PJT00] γιατί ο τρόπος οργάνωση των δεδομένων δεν αποσκοπεί στο να κρατήσει μαζί εγγραφές που είναι κοντά στο 2D χώρο. Ένα δεύτερο, ίσως πιο σημαντικό μειονέκτημα είναι ότι ο αλγόριθμος κατασκευής θεωρεί ότι οι θέσεις των κινούμενων αντικειμένων κατά πάσα πιθανότητα εισάγονται με χρονολογική σειρά και έτσι δεν ευνοεί την εισαγωγή μιας θέσης σε χρόνο t_i όταν η τελευταία θέση ενός αντικειμένου που έχει ήδη εισαχθεί στο ευρετήριο, αντιστοιχεί στο χρονικό αποτύπωμα $t_j > t_i$. Ωστόσο, σε πραγματικές εφαρμογές, η υπόθεση αυτή δεν είναι απαραιτήτως σωστή. Όπως αναφέρθηκε ήδη, αν υποθέσουμε ότι ένα αντικείμενο εισέρχεται σε μία περιοχή όπου το σύστημα μετάδοσης θέσης δεν λειτουργεί, η τροχιά του θα μπορούσε να αποθηκευτεί τοπικά στο αντικείμενο και να διαβιβασθεί αργότερα · εντωμεταξύ άλλα κινούμενα αντικείμενα μπορεί να έχουν διαβιβάσει τη θέση τους, παραβιάζοντας συνεπώς την ανωτέρω υπόθεση του TB-δέντρου.

Στην επόμενη ενότητα, δεδομένου ότι αναγνωρίζουμε τα βασικά πλεονεκτήματα του TB-δέντρου για τη διατήρηση τροχιάς, αναπτύσσουμε ένα νέο ευρετήριο, που ονομάζεται TB*-δέντρο, που ξεπερνά τα μειονεκτήματα του προκατόχου ενώ διατηρεί ταυτόχρονα όλες τις «επιθυμητές» του ιδιότητες.

2.3.2. Το TB*-δέντρο

Η ανάγκη ενός ευρετηρίου που υποστηρίζει την εισαγωγή θέσεων αντικειμένων ανεξάρτητα, η απαίτηση για υποστήριξη διαγραφών, για διατήρηση τροχιάς και για αποτελεσματικότητα τόσο σε επερωτήσεις βασισμένες στις συντεταγμένες όσο και σε επερωτήσεις βασισμένες στη τροχιά είναι οι βασικές προϋποθέσεις που πρέπει να ικανοποιεί το νέο ευρετήριο. Στη συνέχεια, παρουσιάζουμε τη δομή του TB*-δέντρου καθώς και τους αλγορίθμους για την εισαγωγή, διαγραφή, συμπίεση και την εκτέλεση επερωτήσεων σε τροχιές κινούμενων αντικειμένων.

Θα πρέπει να επισημάνουμε ότι, αντίθετα από το αρχικό TB-δέντρο, το TB*-δέντρο δεν ενδιαφέρεται για το κατά πόσο οι εγγραφές εισάγονται με χρονολογική σειρά. Υπάρχει επίσης μια υπόθεση για την ίδια την τροχιά (που επίσης ισχύει και στο TB-δέντρο): οι εισαγωγές εγγραφών που ανήκουν στην ίδια τροχιά γίνονται με χρονολογική σειρά, δηλαδή το ευρετήριο δεν επιτρέπει την εισαγωγή μιας θέσης στο χρονικό σημείο t_i όταν η τελευταία θέση που έχει ήδη εισαχθεί στο ευρετήριο *για το ίδιο αντικείμενο*, ήταν στο $t_j > t_i$. Ακόμα κι αυτό μπορούμε να το αποδεσμεύσουμε από την αρχική μας υπόθεση, όπως θα δείξουμε στην ενότητα 2.3.2.2.1

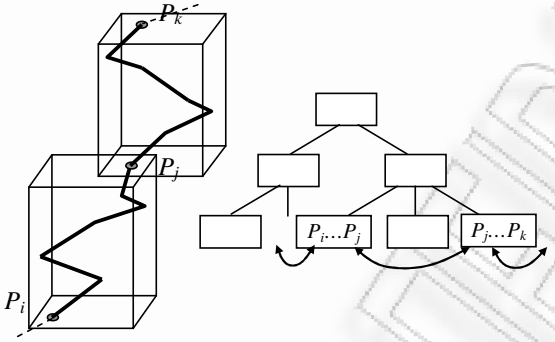
2.3.2.1. Η Δομή του TB-δέντρου*

Στο αρχικό TB-δέντρο, κάθε φορά που ένα κινούμενο αντικείμενο ενημερώνει τη θέση του, εισάγεται ένα νέο 3D γραμμικό τμήμα χρησιμοποιώντας τον αλγόριθμο εισαγωγής που περιγράφεται στο [PJT00]. Το γεγονός αυτό οδηγεί σε αποθήκευση κάθε 3D σημείου της τροχιάς του κινούμενου αντικειμένου δύο φορές: μία φορά ως τελικό σημείο και μία ως αρχικό σημείο. Ενώ κάτι τέτοιο θα ήταν απαραίτητο για μια δομή που αποθηκεύει εγγραφές από διαφορετικές τροχιές στα φύλλα του (π.χ. το 3D R-δέντρο [TVS96] και το STR-δέντρο [PJT00]), είναι σπατάλη χώρου στο TB-δέντρο: εξ' ορισμού, τα γραμμικά τμήματα που αποθηκεύονται στον ίδιο κόμβο φύλλου ανήκουν στην ίδια τροχιά.

Έτσι, αντί για 3D γραμμικά τμήματα, τα φύλλα του TB*-δέντρου αποθηκεύουν 3D σημεία που σχηματίζουν μία 3D πολυγραμμή και αντιπροσωπεύει ένα τμήμα της ακριβούς τροχιάς του αντικειμένου. Επιπλέον, επειδή η *ταυτότητα* του αντικειμένου αποθηκεύεται μία φορά στην κεφαλίδα του φύλλου, οι εγγραφές του φύλλων του TB*-δέντρου αποτελούνται μόνο από 3D σημεία (το σημάδι *Orientation* είναι περιττό). Τα μόνα 3D σημεία που εμφανίζονται δύο φορές στο δένδρο είναι τα σημεία στο τέλος ενός φύλλου και στην αρχή του επόμενου (Σχήμα 2.8). Ενώ αυτά συμβαίνουν στο επίπεδο των φύλλων, η δομή των εσωτερικών κόμβων παραμένει η ίδια με αυτή του αρχικού TB-δέντρου.

Επισήμως, οι κόμβοι φύλλου του TB*-δέντρου είναι της μορφής $\langle header, \{P_i\}\rangle$, όπου κάθε $P_i = \langle t_i, x_i, y_i\rangle$ και $header = \langle$*ταυτότητα*, *#εγγραφών*, *ptr*$\rangle$ (μ' άλλα λόγια, ο ταυτοποιητής του αντικειμένου, ο αριθμός των εγγραφών κόμβου και ένας δείκτης στον πρόγονο κόμβο). Από την άλλη μεριά, οι εσωτερικοί κόμβοι είναι της μορφής $\langle header, \{E_i\}\rangle$, όπου κάθε $E_i = \langle MBB_i, ptr_i\rangle$ με MBB_i το περιβάλλον 3D κουτί του απόγονου κόμβου με δείκτη ptr_i και η κεφαλίδα $header = \langle$*#εγγραφών*, *ptr*$\rangle$

απλώς αποθηκεύει τον αριθμό των εγγραφών του κόμβου και ένα δείκτη στον πρόγονο κόμβο. Επιπλέον, όπως και στο SETI [CEP03] και προκειμένου να υποστηρίξει υψηλούς ρυθμούς εισαγωγής, το TB*-δέντρο χρησιμοποιεί μία δομή κατακερματισμού κύριας μνήμης πρώτης γραμμής, που διατηρεί πλειάδες της μορφής $\langle id, P_{curr}, N_{curr} \rangle$ με τον ταυτοποιητή του αντικειμένου *id*, την πιο πρόσφατη θέση του $P_{curr} = \langle t_{curr}, x_{curr}, y_{curr} \rangle$ και ένα δείκτη N_{curr} στο φύλλο που περιλαμβάνει το P_{curr}.



Σχήμα 2.8: Τα μόνα σημεία που εμφανίζονται δύο φορές στο TB*-δέντρο είναι το αρχικό και το τελικό κάθε φύλλου.

2.3.2.2. Οι Αλγόριθμοι του TB-δέντρου*

Στη συνέχεια, δίνουμε τους αλγορίθμους για την διατήρηση του ευρετηρίου, δηλαδή τον αλγόριθμο εισαγωγής νέων θέσεων, τον αλγόριθμο διαγραφής τροχιών και τον αλγόριθμο συμπίεσης του ευρετηρίου. Όσον αφορά την επεξεργασία επερωτήσεων, οι αλγόριθμοι για την επεξεργασία επερωτήσεων εύρους, βασισμένων στη τροχιά και συνδυαστικών, είναι πανομοιότυποι με αυτούς που παρουσιάζονται στην [PJT00] για το αρχικό TB-δέντρο. Επιπλέον, οι αλγόριθμοι που χρησιμοποιούνται για την προηγμένη επεξεργασία επερωτήσεων, όπως επερωτήσεων πλησιέστερου γείτονα και της ομοιότερης τροχιάς, θα εξετασθούν στα επόμενα κεφάλαια. Μολαταύτα, για λόγους πληρότητας, συμπεριλαμβάνουμε τον αλγόριθμο αναζήτησης εύρους στη συζήτησή μας, που είναι ουσιαστικά ο αλγόριθμος `FindLeaf` που προτάθηκε αρχικώς στην [Gut84] για το απλό R-δέντρο.

2.3.2.2.1. Εισαγωγή νέων Τμημάτων Τροχιάς

Ο αλγόριθμος εισαγωγής του TB*-δέντρου εκτελείται κάθε φορά που ένα κινούμενο αντικείμενο με ταυτότητα *id* διαβιβάζει τη (νέα) του θέση P_{curr}, κάνοντας έτσι, με τη βοήθεια της δομής πρώτης γραμμής, μία νέα εγγραφή που θα εισαχθεί στο δένδρο με ρίζα *Root*. Ο αλγόριθμος *Insert* παρουσιάζεται σε ψευδοκώδικα στο Σχήμα 2.9. Ο παρών ψευδοκώδικας περιλαμβάνει σχόλια που εξηγούν κάθε βήμα του αλγορίθμου. Σημειώστε μόνο ότι πρόκειται για μία εγγραφή, την P_{curr}, που εισάγεται στο ευρετήριο, με εξαίρεση την περίπτωση πλήρους κόμβου όπου ο αλγόριθμος οδηγεί στη δημιουργία ενός νέου κόμβου με δύο εγγραφές, με την πλέον πρόσφατη ήδη δεικτοδοτημένη, P_{prev}, και τη νέα θέση, P_{curr}. Επίσης, προσθέτοντας τη δομή πρώτης γραμμής, η εύρεση του κατάλληλου φύλλου αποδεικνύεται απλή διαδικασία (σε αντίθεση με τον μάλλον ακριβό αλγόριθμο `FindNode` για το TB-δέντρο που περιγράφεται στην [PJT00]).

Όταν γεμίσει ένα φύλλο ο αλγόριθμος εισαγωγής υφίσταται μία μεγάλη τροποποίηση σε σύγκριση με αυτόν του αρχικού TB-δέντρου (Σχήμα 2.10)· ο αλγόριθμος εντοπίζει την προγονή εγγραφή του φύλλου και τη διαγράφει από το δέντρο χρησιμοποιώντας τον κλασσικό αλγόριθμο

Delete του Guttman για το R-δέντρο [Gut84]. Στη συνέχεια, η εγγραφή επανεισάγεται στο δέντρο, χρησιμοποιώντας τον αλγόριθμο Insert του Guttman, αλλά τοποθετείται σε υψηλότερη θέση στο δέντρο (στο επίπεδο πάνω από το επίπεδο του φύλλου), έτσι ώστε το φύλλο που φέρει μαζί της η εγγραφή να βρίσκεται στο ίδιο επίπεδο με τα υπόλοιπα φύλλα – μία τεχνική που χρησιμοποιείται επίσης στον αρχικό αλγόριθμο Delete του R-δέντρου. Με αυτή την τεχνική, όταν ένα φύλλο γεμίσει, πηγαίνει σε μία «καλύτερη» θέση, από άποψη χωρικής γειτονίας, δεδομένου ότι ο αλγόριθμος Insert του Guttman χρησιμοποιεί το κριτήριο της μικρότερης μεγέθυνσης (*least enlargement criterion*) προκειμένου να βρει τον κόμβο στον οποίο θα πραγματοποιήσει την εισαγωγή. Αυτή η τεχνική «διαγραφής και επανεισαγωγής», που χρησιμοποιείται αρχικά στο R*-δέντρο [BKSS90], είναι ο λόγος που ονομάζουμε αυτό το προτεινόμενο ευρετήριο, TB*-δέντρο.

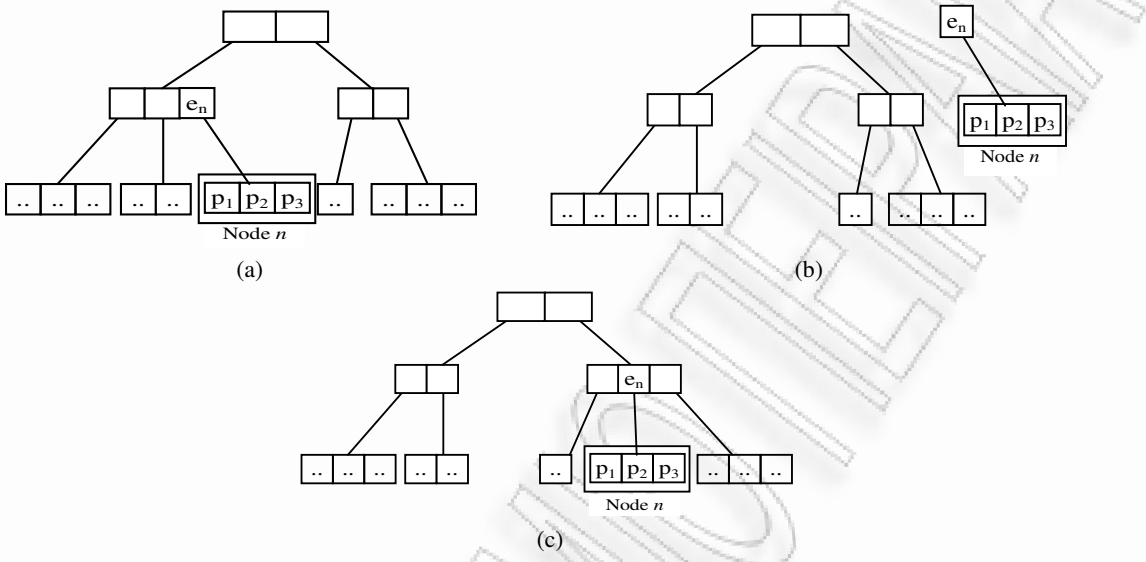
```
 1.  Algorithm Insert(node Root, int Id, 3D Point Pcurr)
 2.     // Algorithm TB*-tree Insert
 3.     // Find leaf node NN containing previous segment
 4.     NN = FrontLine(Id).LastNode
 5.     Pprev = FrontLine(Id).Pcurr
 6.     // If NN exists and has space, insert Pcurr in it and propagate
 7.     // changes upwards using Guttman's AdjustTree
 8.     IF NN exists
 9.        IF NN has space
10.           Insert Pcurr in node NN
11.           AdjustTree (NN)
12.           // If, after the insertion of Pcurr, node NN becomes full,
13.           // delete and reinsert its entry in parent node using
14.           // Guttman's delete and insert algorithms
15.           IF NN is full
16.              PN = NN.Parent
17.              PE = PN.Entry_pointing_to(NN)
18.              Delete (Root, PE)
19.              Insert (Root, PE)
20.           ENDIF
21.        ELSE
22.           // Otherwise, create a new node, insert Pprev and Pcurr in
23.           // the new node and update the front-line
24.           NNode=InsertInNewNode(Root,Pprev,Pcurr)
25.           FrontLine(Id).LastNode = NNode
26.        ENDIF
27.     ELSE
28.        NNode = InsertInNewNode(Root,Pprev,Pcurr)
29.        FrontLine(Id).LastNode = NNode
30.     ENDIF
31.     FrontLine(Id).Pcurr = Pcurr
```

Σχήμα 2.9: Ο αλγόριθμος Insert του TB*-δέντρου

Μια άλλη μεγάλη διαφορά σε σχέση με το αρχικό TB-δέντρο αφορά στη δημιουργία νέων φύλλων και την επιλογή της θέσης όπου θα τοποθετηθούν αυτά. Για το σκοπό αυτό, αναπτύχθηκε ένας νέος αλγόριθμος που καλείται InsertInNewNode (ψευδοκώδικας στο Σχήμα 2.11), που χρησιμοποιεί τους αλγορίθμους του Guttman ChooseLeaf και AdjustTree [Gut84]. Όπως αναφέρθηκε ήδη, ο αλγόριθμος αρχικά τοποθετεί δύο σημεία, P_{prev} και P_{curr}, στο νέο φύλλο (δες Σχήμα 2.8).

Αντίθετα από τη δομή του TB-δέντρου, ο αλγόριθμος InsertInNewNode του TB*-δέντρου βρίσκει το φύλλο δίπλα στο οποίο θα πρέπει να τοποθετηθεί το νέο φύλλο χρησιμοποιώντας το κριτήριο της μικρότερης διεύρυνσης (αλγόριθμος ChooseLeaf του Guttman). Κατόπιν, καλείται ο

αλγόριθμος AdjustTree με ορίσματα και τα δύο φύλλα – τόσο αυτό που επιστρέφεται από τον ChooseLeaf όσο και το καινούργιο – όπως θα συνέβαινε και αν ο κόμβος που επιστρεφόταν από το ChooseLeaf είχε προηγουμένως διασπαστεί. Τέλος, αν η διαδικασία προκαλέσει τη διάσπαση της ρίζας, το δέντρο μεγαλώνει δημιουργώντας μία νέα ρίζα της οποίας οι απόγονοι είναι οι δύο κόμβοι που προκύπτουν.



Σχήμα 2.10: Η στρατηγική που ακολουθείται όταν ένας φύλλο γεμίσει: (a) Το φύλλο *n* γεμίζει (b) Η εγγραφή e_n διαγράφεται από το δέντρο και (c) Η εγγραφή e_n επανεισάγεται στο δέντρο.

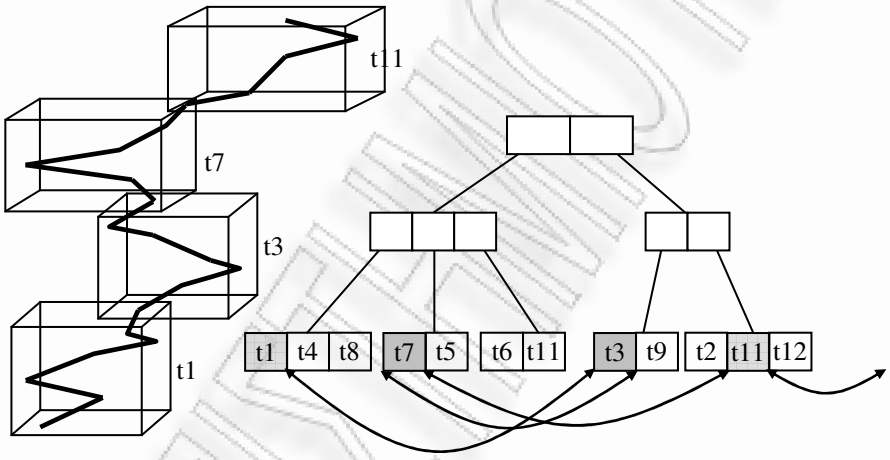
```
 1.  Algorithm InsertInNewNode(node Root, 3D Point P_prev, 3D Point P_curr)
 2.     // Algorithm TB*-tree InsertInNewNode
 3.     Create New Leaf Node NNode
 4.     Insert P_prev in node NNode
 5.     Insert P_curr in node NNode
 6.     // Find Position for the new Node using Guttman's ChooseLeaf
 7.     L = ChooseLeaf (Root, (P_prev,P_curr))
 8.     // Propagate changes upward
 9.     AdjustTree (L, NNode)
10.     // Grow tree taller
11.     IF AdjustTree caused the Root to split
12.        Create a new Root NRoot
13.        Insert first resulted node in NRoot
14.        Insert second resulted node in NRoot
15.     ENDIF
16.     // Return the new Node
17.     RETURN NNode
```

Σχήμα 2.11: Ο αλγόριθμος InsertInNewNode

Όσον αφορά την υπόθεση του TB*-δέντρου ότι οι εγγραφές που ανήκουν στην ίδια τροχιά εισάγονται με χρονολογική σειρά, αυτό συμβαίνει μόνο για να παραμείνει η διαδικασία εισαγωγής απλή (εισάγεται μια νέα θέση στον «τρέχοντα» κόμβο – όπως υποδεικνύεται από τη δομή πρώτης γραμμής– ή σε ένα νέο κόμβο, ενημερώνοντας ανάλογα την δομή κατακερματισμού). Σε περίπτωση που θέλουμε να αποδεσμευτούμε από την αρχική υπόθεση, απαιτείται μία αναζήτηση προς τα πίσω στην διπλά συνδεδεμένη λίστα κόμβων (που ξεκινά από τον «τρέχοντα» κόμβο) ώστε η νέα ('παρωχημένη') εγγραφή να εισαχθεί στον κατάλληλο κόμβο, και επειδή όλοι οι κόμβοι πριν από τον 'τρέχοντα' κόμβο στη λίστα είναι εξ' ορισμού πλήρεις, μία εγγραφή πρέπει να μετακινηθεί από κάθε

κόμβο στον επόμενο κόμβο στην αλυσίδα που ξεκινά από τον κόμβο στον οποίο εισήχθη η εγγραφή και τελειώνει στον 'τρέχοντα' κόμβο.

Τέλος, μία τεχνική προσωρινής μνήμης (buffering) χρησιμοποιείται για τη βελτιστοποίηση της διαδικασίας εισαγωγής. Πιο συγκεκριμένα, πέρα από τη χρήση ενός παραδοσιακού μηχανισμού προσωρινής μνήμης (όπως του LRU), η δομή του TB*-δέντρου μπορεί να χρησιμοποιήσει μία *επιπλέον προσωρινή μνήμη*, από τούδε και στο εξής καλείται *Προσωρινή Μνήμη Τελευταίας Σελίδας (Last Page (LP) Buffer)*, στην οποία αποθηκεύονται όλα τα φύλλα που δεν έχουν συμπληρωθεί μέχρι τώρα, και είναι ένα για κάθε ξεχωριστό κινούμενο αντικείμενο· κατόπιν όταν κάθε φύλλο συμπληρώνεται, αποθηκεύεται στο δίσκο (πραγματοποιώντας επομένως μόνο μία πρόσβαση), και το επόμενο (νέο) φύλλο της ίδιας τροχιάς παίρνει τη θέση του στην προσωρινή μνήμη LP. Συνεπώς, το μέγεθος της προσωρινής μνήμης LP θα είναι πάντα ίσο με τον αριθμό των αντικειμένων που δεικτοδοτούνται επί του παρόντος από το TB*-δέντρο. Όπως θα αποδειχθεί στα πειράματα, η προσωρινή μνήμη LP μειώνει δραματικά τον αριθμό των προσπελάσεων στο δίσκο που απαιτούνται για την εισαγωγή νέων δεδομένων στο ευρετήριο.



Σχήμα 2.12: Η δομή του TB*-δέντρου

Η γενική εικόνα του TB*-δέντρου απεικονίζεται στο Σχήμα 2.12. Σε σύγκριση με το TB-δέντρο (Σχήμα 2.7), είναι σαφές ότι οι κόμβοι φύλλου που ανήκουν στην ίδια τροχιά δεν τοποθετούνται πλέον με αύξουσα χρονική σειρά (π.χ. από αριστερά προς τα δεξιά), αλλά τοποθετούνται σε θέσεις που καθορίζονται από το κριτήριο μικρότερης διεύρυνσης.

2.3.2.2.2. Διαγραφή Τροχιών

Οι διαγραφές συχνά παραμελούνται όταν προτείνονται μέθοδοι δεικτοδότησης για τροχιές κινούμενων αντικειμένων, με κύριο επιχείρημα ότι η διαγραφή ενός 3D γραμμικού τμήματος δεν έχει νόημα. Παρόλο που μπορούμε να υποθέσουμε ότι εννοιολογικά κάτι τέτοιο είναι σωστό (οι θέσεις που μεταδίδονται καταγράφονται, συνεπώς υπάρχουν), η διαγραφή ολόκληρης της τροχιάς ενός αντικειμένου έχει νόημα (οι τροχιές αντικειμένων που δεν είναι πλέον χρήσιμες μπορούν να διαγραφούν από το ευρετήριο). Άρα πρέπει να δώσουμε ένα αποτελεσματικό αλγόριθμο που να υποστηρίζει τις διαγραφές τροχιών αντικειμένων.

Ο αλγόριθμος `DeleteTrajectory`, που φαίνεται στο Σχήμα 2.13, μπορεί να χρησιμοποιηθεί στο TB*-δέντρο προκειμένου να διαγράψει την τροχιά του κινούμενου αντικειμένου με ταυτότητα *Id*.

Ο αλγόριθμος αρχικά εντοπίζει το 'τρέχον' φύλλο *N* του αντικειμένου με ταυτότητα *id*. Κατόπιν, αφαιρεί την εγγραφή του προγόνου του *N* από τον πρόγονο κόμβο εκτελώντας τον αλγόριθμο `Delete` [Gut84] του R-tree και ακολουθεί την αλυσίδα προς τα πίσω σε κόμβους που περιλαμβάνουν τμήματα της ίδιας τροχιάς, διαγράφοντας το ένα μετά το άλλο. Αν παραστεί ανάγκη, βάσει του αλγορίθμου `Delete`, οι κόμβοι αναδιατάσσονται π.χ. αν ο αριθμός των εγγραφών πέφτει κάτω από το κατώφλι *m=M/2*, το δέντρο μπορεί ακόμα και να υποχρεωθεί να συμπτυχθεί.

```
 1.  Algorithm DeleteTrajectory(int Id)
 2.     // Algorithm TB*-tree DeleteTrajectory
 3.     // Find latest trajectory leaf node N
 4.     N = FrontLine(Id).LastNode
 5.     // Delete leaf node N's parent entry using Guttman's Delete
 6.     // Algorithm and follow the pointers to the trajectory's previous
 7.     // leaf nodes deleting also their parent entries
 8.     DO UNTIL N Is NULL
 9.        PN = N.Parent
10.        PE = PN.Entry_pointing_to(N)
11.        Delete (Root, PE)
12.        N = N.PreviousLeaf
13.     LOOP
```

Σχήμα 2.13: Ο Αλγόριθμος `DeleteTrajectory`

Η δομή του TB*-δέντρου φαίνεται ιδανική για την παροχή ενός τέτοιου αλγορίθμου: έχοντας εντοπίσει έστω και ένα γραμμικό τμήμα που ανήκει στην τροχιά ενός αντικειμένου, θα μπορούσε κανείς να ακολουθήσει τις διπλά συνδεδεμένες λίστες για να ανακτήσει όλη την τροχιά και να διαγράψει τα φύλλα που το αποτελούν. Από την άλλη πλευρά, το αρχικό TB-δέντρο δεν μπορεί να υποστηρίξει διαγραφές τροχιάς: οι διαγραφές τροχιάς έχουν ως αποτέλεσμα διαγραφές εγγραφών σε εσωτερικούς κόμβους του δένδρου που είτε απαιτούν τεχνικές σύμπτυξης (όπως τον αλγόριθμο `CondenseTree` [Gut84]) ή αφήνουν οπές στους κόμβους. Όπως και να' χει, οι 'επιθυμητές' ιδιότητες του TB-δέντρου χάνονται (όλοι οι κόμβοι φύλλου εκτός των 'τρεχουσών' είναι πλήρεις· υπάρχει μία χρονολογική σειρά κλπ.).

Όσο για τις άλλες δομές δεικτοδότησης (όπως το 3D R-δέντρο [TVS96], το STR-δέντρο [PJT00], το SETI [CEP03]), εξ' ορισμού δεν διαθέτουν κάποιο μηχανισμό για την αποτελεσματική ανάκτηση ολόκληρης της τροχιάς ενός αντικειμένου· έτσι, για να υποστηρίξουν διαγραφές τροχιάς πρέπει να απαντήσουν σε ακολουθίες επερωτήσεων εύρους όπως περιγράφεται στην [PJT00] για τη συνδυαστική αναζήτηση του 3D R-δέντρου και του STR-δέντρου – μία πολύ ακριβή προσέγγιση όπως δείχνεται στην [PJT00].

2.3.2.2.3. Συμπίεση του Ευρετηρίου

Ενώ το αρχικό TB-δέντρο ικανοποιεί την απαίτηση διατήρησης της τροχιάς έτσι ώστε να χρησιμοποιήσει τον αλγόριθμο συμπίεσης τροχιάς TD-TR που προτείνεται στην [MB04], ένας τέτοιος αλγόριθμος θα πρέπει να διαβάζει κάθε δεικτοδοτημένη τροχιά μία προς μία, να τη συμπιέζει και τέλος να τροφοδοτεί ένα νέο TB-δέντρο με τη συμπιεσμένη τροχιά. Ωστόσο, επειδή το TB-δέντρο τοποθετεί τις νέες εγγραφές πάντα στη δεξιά «άκρη» του δέντρου, μια τέτοια προσέγγιση θα τοποθετεί ολόκληρες τροχιές σε αυτή την πλευρά του δέντρου χωρίς να λαμβάνει υπ' όψιν την χρονική τους διάταξη και θα μας οδηγεί σε ένα δένδρο με κόμβους που έχουν μεγάλη χρονική αλληλοεπικάλυψη, και τελικά στη μείωση της απόδοσής του. Συνεπώς, προκειμένου να ξεπεράσουμε αυτό το

μειονέκτημα, θα πρέπει να χρησιμοποιήσουμε ενδιάμεσα βήματα ανακτώντας όλες τις τροχιές που δεικτοδοτούνται από το TB-δέντρο, δημιουργώντας όλες τις νέες συμπιεσμένες τροχιές, ταξινομώντας τες με χρονολογική σειρά και τέλος τροφοδοτώντας το νέο TB-δέντρο. Μολαταύτα, μια τέτοια τεχνική θα σήμαινε επεξεργασία όλου του ευρετηρίου στην κύρια μνήμη, ή ανάπτυξη εξειδικευμένων αλγορίθμων για την αποτελεσματική της αντιμετώπιση. Απο την άλλη, ο αλγόριθμος OW-TR που βασίζεται στο ανοιγόμενο παράθυρο και παρουσιάζεται στην [MB04] θα μπορούσε να είναι μια κάποια λύση· όμως, μια τέτοια προσέγγιση θα οδηγούσε στη χρήση ενός λιγότερο αποτελεσματικού αλγορίθμου συμπίεσης τόσο σε όρους ποιότητας όσο και συμπίεσης.

```
 1.  Algorithm CompressIndex(double Threshold, TB*-tree TB)
 2.     // Algorithm TB*-tree CompressIndex
 3.     // Create a new TB*-tree
 4.     NTB = New TB*-Tree
 5.     FOR EACH Id IN TB.Trajectories
 6.     // Find latest trajectory leaf node N
 7.        N = FrontLine(Id).LastNode
 8.        // Create a new Trajectory retrieve all of its entries
 9.        Traj = New Trajectory
10.        DO UNTIL N Is NULL
11.           Traj.Add N.Segments
12.           N = N.PreviousLeaf
13.        LOOP
14.        // Apply the top-down spatiotemporal compression algorithm
15.        // TD-TR in the Trajectory with the given threshold
16.        TD-TR (Traj, Threshold)
17.        // Insert in the new TB*-tree each point P of the compressed
18.        // trajectory
19.        FOR EACH P IN Traj
20.           Insert NTB.Root, Id, P
21.        NEXT
22.     NEXT
```

Σχήμα 2.14: Ο Αλγόριθμος `CompressIndex`

Αντίθετα, το προτεινόμενο TB*-δέντρο δεν παρουσιάζει κανένα από αυτά τα μειονεκτήματα. Ο αλγόριθμος εισαγωγής υποστηρίζει προσθήκες τροχιάς με μη χρονολογική σειρά. Γι' αυτό και στο Σχήμα 2.14 παρουσιάζουμε έναν απλό αλγόριθμο που συμπιέζει ένα TB*-δέντρο χρησιμοποιώντας τον αλγόριθμο TD-TR [MB04]. Ο αλγόριθμος ξεκινάει δημιουργώντας ένα νέο TB*-δέντρο, και στη συνέχεια χρησιμοποιώντας την κατακερματισμένη δομή, προσπελαύνει τον τελευταίο κόμβο κάθε τροχιάς. Κατόπιν, ακολουθώντας τους δείκτες των προηγούμενων φύλλων, ανακτά όλη την τροχιά στην οποία εφαρμόζεται ο αλγόριθμος [MB04] με το δεδομένο κατώφλι. Τέλος, ο αλγόριθμος τροφοδοτεί το νέο TB*-δέντρο με τη συμπιεσμένη τροχιά και επαναλαμβάνει την ίδια διαδικασία για τις υπόλοιπες τροχιές μέχρι να τις έχει επεξεργαστεί όλες.

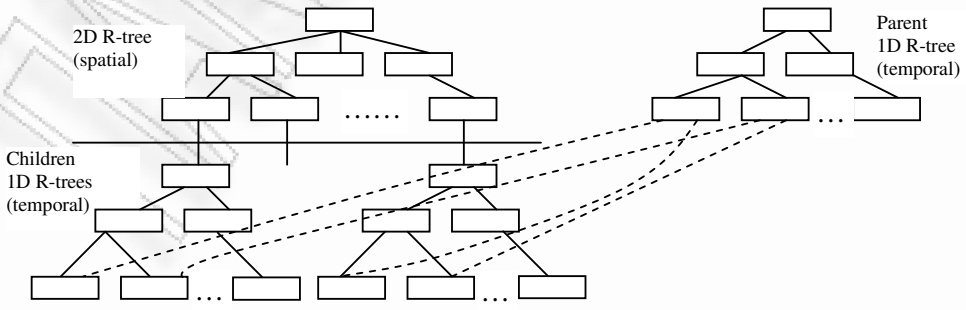
2.3.2.2.4. Εκτέλεση Επερωτήσεων στο TB-δέντρο*

Όπως αναφέρθηκε ήδη, επειδή και το TB και το TB*-δέντρο βασίζονται στο πασίγνωστο R-δέντρο, οι αντίστοιχοι αλγόριθμοι αναζήτησης εύρους ακολουθούν τον αλγόριθμο `FindLeaf` που παρουσιάστηκε αρχικά στην [Gut84]. Ο αλγόριθμος αυτός εξετάζει αναδρομικά τους κόμβους του δέντρου, απορρίπτοντας κόμβους με MBBs που δεν αλληλεπικαλύπτονται με το παράθυρο επερώτησης, και ταυτόχρονα ακολουθεί τους δείκτες από τις εγγραφές με MBBs που αλληλεπικαλύπτονται με το παράθυρο της ερώτησης προς τους αντίστοιχους απογόνους κόμβους μέχρι να βρεθούν όλα τα υποψήφια φύλλα που πιθανόν περιέχουν αντικείμενα που περιέχονται στην

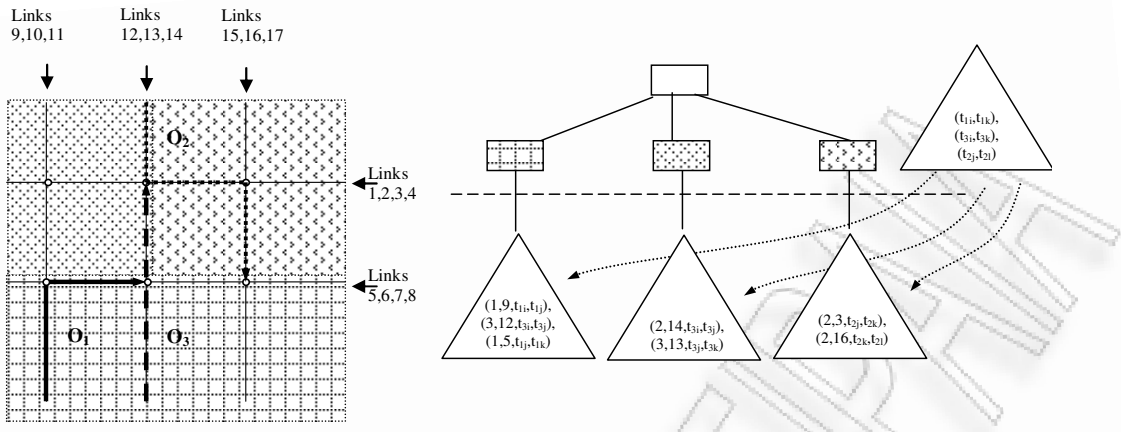
απάντηση της ερώτησης. Ακολουθώντας το παράδειγμα που εμφανίζεται στο Σχήμα 2.1 για τα χωρικά αντικείμενα, ας θεωρήσουμε μια επερώτηση εύρους *Q* που εκτελείται στο 2D R-δέντρο. Ο αλγόριθμος ξεκινά εξετάζοντας τη ρίζα του δέντρου, ελέγχοντας κατά πόσο τα MBB των εγγραφών της ρίζας αλληλεπικαλύπτονται με την *Q*. Αν το MBB μίας εγγραφής αλληλεπικαλύπτεται με την *Q*, ο αλγόριθμος ακολουθεί το δείκτη στον αντίστοιχο απόγονο κόμβο (εγγραφές *A* και *B* στο παράδειγμά μας), όπου επαναλαμβάνει την ίδια διαδικασία. Αν ο αλγόριθμος φτάσει σ΄ ένα φύλλο, οι εγγραφές του φύλλου ελέγχονται ως προς την *Q* και αν τα MBB τους αλληλεπικαλύπτονται, ο αλγόριθμος αναφέρει τις ταυτότητές τους (αντικείμενα *F* και *G* όταν ο αλγόριθμος εξετάζει τον κόμβο φύλλου *A* και αντικείμενο *H* όταν βρίσκεται στον κόμβο *B*). Η επέκταση του παραπάνω αλγορίθμου στο χωροχρονικό πλαίσιο είναι μία αρκετά ξεκάθαρη διαδικασία διότι κάθε 2D MBB απλώς αντικαθίσταται από τα αντίστοιχα 3D MBB των αντικειμένων, κόμβων ή επερωτήσεων.

2.4. Δεικτοδότηση Τροχιών Αντικειμένων Κινούμενων σε Σταθερά Δίκτυα

Όπως αναφέρθηκε ήδη, ακολουθώντας τις προτάσεις της [KGT99], στην παρούσα διατριβή προτείνουμε το FNR-δέντρο, μία προέκταση του πασίγνωστου R-δέντρου [Gut84], που σχεδιάστηκε για την δεικτοδότηση αντικειμένων που κινούνται σε σταθερά δίκτυα. Το FNR-δέντρο μπορεί να θεωρηθεί ως ένα δάσος αρκετών 1D R-δέντρων πάνω από ένα μόνο 2D R-δέντρο. Το 2D R-δέντρο χρησιμοποιείται για τη δεικτοδότηση των χωρικών δεδομένων του δικτύου (δηλαδή δρόμων που αποτελούνται από γραμμικά τμήματα), ενώ κάθε ένα από τα (χρονικά) 1D R-δέντρα (που από εδώ και πέρα ονομάζονται *«Απόγονα 1D R-δέντρα»*) αντιστοιχεί σε ένα φύλλο του 2D R-δέντρου και δεικτοδοτεί τα χρονικά διαστήματα στη διάρκεια των οποίων τα κινούμενα αντικείμενα κινήθηκαν στους συνδέσμους (ακμές) του δικτύου που βρίσκονται μέσα στο αντίστοιχο φύλλο του 2D R-δέντρου. Έτσι, το (χωρικό) 2D R-δέντρο παραμένει σταθερό στη διάρκεια της ζωής του FNR-δέντρου – φτάνει να μην υπάρχουν αλλαγές στο δίκτυο. Ένα επιπλέον (χρονικό) 1D R-δέντρο (που από εδώ και στο εξής θα ονομάζεται *«Πρόγονο 1D R-δέντρο»*) χρησιμοποιείται για να δεικτοδοτήσει τα φύλλα όλων των απογόνων 1D R-δέντρων σε σχέση με τη διάρκεια ζωής τους. Έτσι, το χρονικό διάστημα κάθε φύλλου των 1D R-δέντρων εισάγεται μαζί με ένα δείκτη στο φύλλο ως νέα εγγραφή στο πρόγονο 1D R-δέντρο. Η συνολική δομή του FNR-δέντρου παρουσιάζεται στο Σχήμα 2.15, ενώ το Σχήμα 2.16 (b) περιλαμβάνει ένα παράδειγμα βασισμένο στην διαμόρφωση των αντικειμένων παρουσιάζεται Σχήμα 2.16 (a).



Σχήμα 2.15: Η δομή του FNR-δέντρου

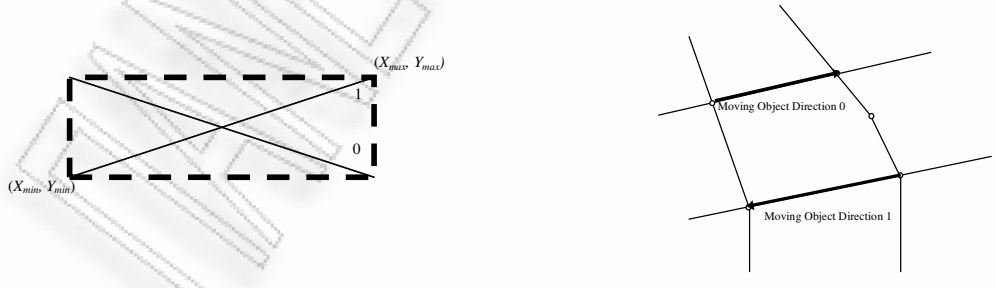


(a) (b)

Σχήμα 2.16: Ένα παράδειγμα FNR-δέντρου: (a) τροχιές τριών αντικειμένων σε οδικό δίκτυο και (b) το αντίστοιχο FNR-δέντρο

2.4.1. Η Δομή του FNR-δέντρου

Όπως αναφέρθηκε ήδη, το FNR-δέντρο μπορεί να θεωρηθεί ως ένα 2D R-δέντρο που δεικτοδοτεί τα γραμμικά τμήματα του δικτύου μαζί με ένα δάσος 1D R-δέντρων που δεικτοδοτούν χρονικά διαστήματα. Ακολουθώντας τη συνήθη δομή του R-δέντρου, οι εσωτερικοί κόμβοι του 2D R-δέντρου είναι της μορφής $\langle header, \{ptr_i, MBB_i\}\rangle$, όπου κάθε $MBB_i = \langle x_{min\text{-}i}, y_{min\text{-}i}, x_{max\text{-}i}, y_{max\text{-}i}\rangle$ και $header = \langle$*ταυτότητα*, *#εγγραφών*, *ptr*$\rangle$. Από την άλλη μεριά, η δομή των φύλλων του 2D R-δέντρου είναι λίγο τροποποιημένη σε σχέση με το συμβατικό R-δέντρο· επισήμως, τα φύλλα είναι της μορφής $\langle header, \{link_i, MBB_i, orientation\}\rangle$ και $header = \langle$*ταυτότητα*, *#εγγραφών*, *ptr*, $ptr_{child\text{-}R\text{-}tree}\rangle$. Βάσει αυτής της μορφής, ο δείκτης που συνήθως βρίσκεται εντός της κάθε εγγραφής του φύλλου έχει αντικατασταθεί από ένα σημάδι '*orientation*' (0/1) που περιγράφει την ακριβή γεωμετρία του γραμμικού τμήματος εντός του MBB (Σχήμα 2.17(a)). Μία παρόμοια προσέγγιση ακολουθείται στην [PJT00] για την αναπαράσταση γραμμικών τμημάτων τροχιών σε 3D R-δέντρα [TVS96]. Επιπλέον, κάθε φύλλο του 2D R-δέντρου περιλαμβάνει ένα δείκτη ($ptr_{child\text{-}R\text{-}tree}$) που δείχνει στη ρίζα του αντίστοιχου απόγονου 1D R-δέντρου.



(a) (b)

Σχήμα 2.17: (a) Το σημάδι 'orientation' στις εγγραφές του 2D R-δέντρου· (b) το σημάδι 'direction' στις εγγραφές των 1D R-δέντρων

Όσον αφορά τα 1D R-δέντρα, οι εσωτερικοί τους κόμβοι είναι της μορφής $\langle header, \{ptr_i, MBB_i\}\rangle$, ενώ τα φύλλα είναι ελαφρώς διαφορετικά: $\langle header, \{Object\text{-}id_i, Link\text{-}id_i, MBB_i, direction\}\rangle$,

και $MBB_i=\langle t_{in}, t_{out}\rangle$ είναι το χρονικό διάστημα εντός του οποίου το αντικείμενο με ταυτότητα *Object-id*$_i$ κινείται στο γραμμικό τμήμα με ταυτότητα *Link-id*$_i$, που περιλαμβάνεται στο αντίστοιχο φύλλο του 2D R-δέντρου. Το *Direction* είναι ένα ακόμα σημάδι (με τιμές 0/1) που περιγράφει την κατεύθυνση του κινούμενου αντικειμένου (Σχήμα 2.17 (b)). Πιο συγκεκριμένα, το *direction* καθορίζεται στο 0 (1) όταν το κινούμενο αντικείμενο εισέρχεται στο γραμμικό τμήμα από τον αριστερότερο - (δεξιότερο) κόμβο. Στην ειδική περίπτωση που το γραμμικό τμήμα είναι κάθετο, το *direction* καθορίζεται στο 0 (1) για αντικείμενα που εισέρχονται στο γραμμικό τμήμα από τον κατώτερο (ανώτερο) κόμβο. Τέλος, οι κεφαλίδες των φύλλων είναι της μορφής *header* = $\langle$*ταυτότητα*, *#εγγραφών*, *ptr*, $ptr_{parent\text{-}R\text{-}tree\text{-}node}\rangle$ όπου το $ptr_{parent\text{-}R\text{-}tree\text{-}node}$ σημαίνει ότι δείχνουμε απευθείας από κάθε 1D φύλλο R-δέντρου στο αντίστοιχο φύλλο του 2D R-δέντρου.

Η δομή του πρόγονου 1D R-δέντρου είναι παρόμοια με την παραπάνω δομή. Παρόλο που οι εσωτερικοί κόμβοι παραμένουν πανομοιότυποι με τους προηγούμενους, τα φύλλα του διαφέρουν σε κάποιο βαθμό: είναι της μορφής $\langle$*header*, $\{ptr_{child\text{-}R\text{-}tree\text{-}node}, MBB_i\}\rangle$ *με* $MBB_i=\langle t_{min}, t_{max}\rangle$, και το $ptr_{child\text{-}R\text{-}tree\text{-}node}$ να δείχνει στον αντίστοιχο φύλλο στο δάσος των απογόνων 1D δέντρων (Σχήμα 2.15).

2.4.2. Οι Αλγόριθμοι του FNR-δέντρου

Στην συνέχεια, δίνουμε αλγορίθμους για την εισαγωγή μιας νέας εγγραφής στο FNR-δέντρο (ενότητα 2.4.2.1) καθώς και την αναζήτηση στο FNR-δέντρο μ' ένα χωροχρονικό παράθυρο επερώτησης (ενότητα 2.4.2.2).

```
 1.  Algorithm FNR_tree_Insert(object_id, xstart, ystart, xend,
 2.  yend, tin, tout)
 3.     // Search the line segment with Link_id in the 2D R-tree
 4.     // that object_id leaves
 5.     Link_id = 2D_R_tree_search(xstart, ystart, xend, yend)
 6.     // follow the pointer from leaf node that contains link_id
 7.     // to the corresponding 1D R-tree, RT
 8.     RT=link_id.Child_R_tree
 9.     // Insert the time interval into the 1D R-tree
10.     // Let ND be the leaf node where the input was inserted
11.     RT.Insert_most_recent(tin, tout, object_id, link_id, ND)
12.     // If necessary, update the Parent 1D R-tree by inserting or
13.     // updating the MBB of node ND
14.     IF ND is a new node caused by the insertion
15.        Parent_1D_R_tree_insert(ND.MBB, ND.ptr)
16.     ELSEIF the ND.MBB was modified
17.        Parent_1D_R_tree_delete(ND.MBB, ND.ptr)
18.        Parent_1D_R_tree_insert(ND.MBB, ND.ptr)
19.     ENDIF
```

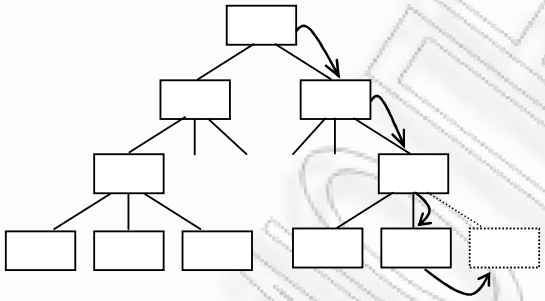
Σχήμα 2.18: Αλγόριθμος Εισαγωγής FNR-δέντρου

2.4.2.1. Εισαγωγή νέων Τμημάτων Τροχιάς

Ο αλγόριθμος εισαγωγής του FNR-δέντρου εκτελείται κάθε φορά που ένα κινούμενο αντικείμενο με *Object-id*$_i$ φεύγει από ένα γραμμικό τμήμα του δικτύου, που αναπαρίσταται από το αντίστοιχο MBB (x_{start}, y_{start}, x_{end}, y_{end}) και το σημάδι *direction*. Η λίστα των ορισμάτων περιλαμβάνει επίσης το χρονικό διάστημα (t_{in}, t_{out}) στη διάρκεια του οποίου το *Object-id*$_i$ μετακινείται στο γραμμικό τμήμα. Ο αλγόριθμος εισαγωγής φαίνεται στο Σχήμα 2.18.

Σε αυτό τον αλγόριθμο, `R_tree_insert`, `R_tree_delete` και `R_tree_search` είναι οι κλασσικοί αλγόριθμοι που περιγράφονται στην [Gut84] για τη διατήρηση και αναζήτηση σε ένα (1D ή 2D) R-δέντρο. Από την άλλη πλευρά, για την εισαγωγή που γίνεται στη γραμμή 11 (λαμβάνοντας υπ'

όψιν τα 1D απόγονα R-δέντρα του FNR-δέντρου), παρατηρούμε ότι τα 1D χρονικά διαστήματα εισάγονται στο δέντρο με αύξουσα σειρά γιατί ο χρόνος είναι μονοτονικός. Το γεγονός αυτό μας οδηγεί στην ακόλουθη τροποποίηση του αλγόριθμου εισαγωγής του R-δέντρου, που από τούδε και στο εξής ονομάζεται `Insert_most_recent` και περιγράφεται στο Σχήμα 2.19. Κάθε νέα εγγραφή απλά εισάγεται στο πιο πρόσφατο (δεξιότερο) φύλλο του 1D R-δέντρου. Σε περίπτωση που ένας κόμβος είναι πλήρης, δημιουργείται ένας νέος κόμβος και η εγγραφή εισάγεται σ' αυτόν. Το νέο φύλλο εισάγεται στη δομή ως συγγενής κόμβος του (προηγουμένως) πιο πρόσφατου φύλλου. Ως τέτοιο, θα μπορούσε να προκαλέσει διάδοση της αύξησης των κόμβων προς τα πάνω χρησιμοποιώντας τον αλγόριθμο `AdjustTree`, που περιγράφεται επίσης και στην [Gut84]. Το αποτέλεσμα αυτής της τεχνική εισαγωγής είναι 1D R-δέντρα με σχεδόν πλήρη φύλλα και πολύ μικρή αλληλοεπικάλυψη.

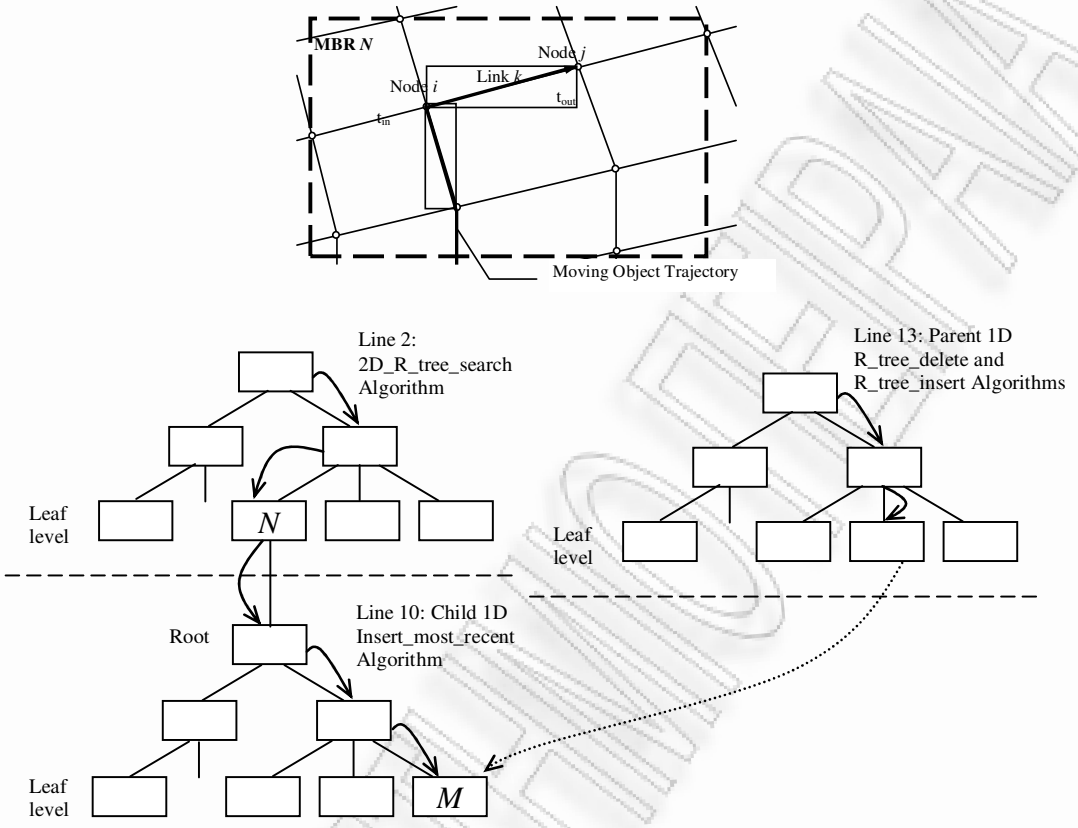


Σχήμα 2.19: Οι νέες εγγραφές εισάγονται πάντα στον δεξιότερο κόμβο κάθε 1D R-δέντρου όταν οι εισαγωγές γίνονται με χρονολογική σειρά

Ωστόσο, δεδομένης της συζήτησης στην ενότητα 2.3.2, η στρατηγική του `Insert_most_recent` μπορεί να θεωρηθεί μειονέκτημα σε αρκετές εφαρμογές όπου οι εισαγωγές νέων τμημάτων τροχιάς δεν ακολουθούν απαραίτητα τη μονοσήμαντη διάταξη του χρόνου. Για να αντιμετωπίσουμε αυτή την απαίτηση μη χρονολογικών εισαγωγών, το FNR-δέντρο μπορεί να υλοποιηθεί χρησιμοποιώντας τον απλό αλγόριθμο `R_tree_insert` [Gut84] στη γραμμή 11, μία προσέγγιση που δίνει τη δυνατότητα στο ευρετήριο να χειριστεί αποτελεσματικά τροχιές αντικειμένων που εισάγονται με αυθαίρετη χρονική σειρά. Ολοκληρώνοντας, ανάλογα με την εφαρμογή, το FNR-δέντρο μπορεί ή όχι να υποστηρίξει μη χρονολογικές εισαγωγές απλά τροποποιώντας τη γραμμή 10 του αλγορίθμου `FNR_tree_insert`. Τέλος, οι αλγόριθμοι εισαγωγής και διαγραφής που χρησιμοποιήθηκαν στις γραμμές 14-18 είναι οι συμβατικοί αλγόριθμοι του R-δέντρου [Gut84] κυρίως λόγω των ενημερώσεων που απαιτούνται (διαγραφές και επανεισαγωγές).

Όπως φαίνεται στο Σχήμα 2.20, ο αλγόριθμος εισαγωγής εκτελείται όταν το κινούμενο αντικείμενο φτάσει σ' ένα κόμβο (*Node j*) του δικτύου. Το πρώτο βήμα (γραμμή 4) απαιτεί μία χωρική αναζήτηση στο 2D R-δέντρο (με τις συντεταγμένες των κόμβων *i* και *j* ως ορίσματα) προκειμένου να βρούμε το σύνδεσμο *k*, που περικλείεται από το MBB του φύλλου *N* του 2D R-δέντρου. Στη συνέχεια, ακολουθούμε το δείκτη στο αντίστοιχο 1D R-δέντρο, στο οποίο εισάγουμε μια νέα εγγραφή (t_{in}, t_{out}, *object-id*, *link-id*). Ανάλογα με την πολιτική εισαγωγής, η νέα εισαγωγή τοποθετείται στον φύλλο *M* στο δεξιότερο άκρο του 1D R-δέντρου (ή στον κόμβο που προσδιορίζεται από τον αλγόριθμο `R_tree_insert`). Οι πιθανές τροποποιήσεις στη δομή λόγω αυτής της εισαγωγής (το MBB αυτού του φύλλου μπορεί να ενημερωθεί, μπορεί να δημιουργηθεί ένα νέο φύλλο), διαδίδονται προς τα πάνω. Μια τέτοια τροποποίηση προκαλεί ενημερώσεις και στο πρόγονο 1D R-δέντρο· η ενημέρωση του

MBB ενός φύλλου στο 1D απόγονο R-δέντρο προκαλεί διαγραφή και επανεισαγωγή της αντίστοιχης εγγραφής στον πρόγονο 1D R-δέντρο, ενώ η δημιουργία ενός νέου κόμβου στο απόγονο 1D R-δέντρο προκαλεί εισαγωγή μιας νέας εγγραφής στο πρόγονο R-δέντρο.



Σχήμα 2.20: Εισαγωγή μιας νέας εγγραφής στο FNR-δέντρο

```
1.   Algorithm FNR_tree_Search_from_2D(x_min, x_max, y_min, y_max, t_min, t_max)
 2.     // Search in the 2D R-tree with the 2D interval
 3.     // (x_min, x_max, y_min, y_max) retrieving the Links contained in it
 4.     Links = 2D_R_tree_search(x_min, x_max, y_min, y_max)
 5.     // follow the pointers from leaf nodes ND containing the Links
 6.     // to the corresponding 1D R-trees, RT
 7.     FOR EACH ND containing any of Links
 8.        RT=ND.Child_R_tree
 9.        // Search each one of the corresponding 1D R-trees
10.        Candidates=RT.R_tree_search(tmin, tmax)
11.        // Refinement
12.        // If ND2 is completely contained inside (x_min, y_min, x_max,
13.        // y_max) all entries of ND2 are also inside
14.        IF ND2.MBB is inside (xmin, ymin, xmax, ymax)
15.           RETURN all entries in Candidates
16.        ELSE // ND2 is partially inside (x_min, y_min, x_max, y_max)
17.           FOR EACH Entry IN Candidates
18.              IF Links(Entry.Link_id).MBB is inside (x_min, x_max, y_min, y_max)
19.                 RETURN the Entry
20.              ENDIF
21.           NEXT
22.        ENDIF
23.     NEXT
```

Σχήμα 2.21: Ο Αλγόριθμος του FNR-δέντρου `Search-from-2D-R-tree`

2.4.2.2. Επεξεργασία Επερωτήσεων στο FNR-δέντρο

Η δομή του FNR-δέντρου προσφέρει την ευελιξία να χρησιμοποιήσουμε δύο διαφορετικούς αλγόριθμους για τους διάφορους τύπους επερωτήσεων. Τα συγκριτικά πλεονεκτήματα θα παρουσιαστούν εν συνεχεία μέσω παραδειγμάτων στη μελέτη απόδοσης.

`Search-from-2D-R-tree`: Ο πρώτος αλγόριθμος, που φαίνεται στο Σχήμα 2.21, ξεκινά από τη ρίζα του 2D R-δέντρου, εντοπίζει τις εγγραφές του δένδρου που ικανοποιούν τους χωρικούς περιορισμούς της επερώτησης και στη συνέχεια ακολουθώντας τον / τους δείκτη(ες) στο / στα αντίστοιχο(α) 1D R-δέντρο(α), ελέγχει αν υπάρχουν εγγραφές που ικανοποιούν και το χρονικό περιορισμό της επερώτησης. Τέλος, ένα βήμα διαλογής διασφαλίζει ότι ο αλγόριθμος επιστρέφει μόνο τις εγγραφές που ικανοποιούν συγχρόνως τα χωρικά και χρονικά κριτήρια της ερώτησης.

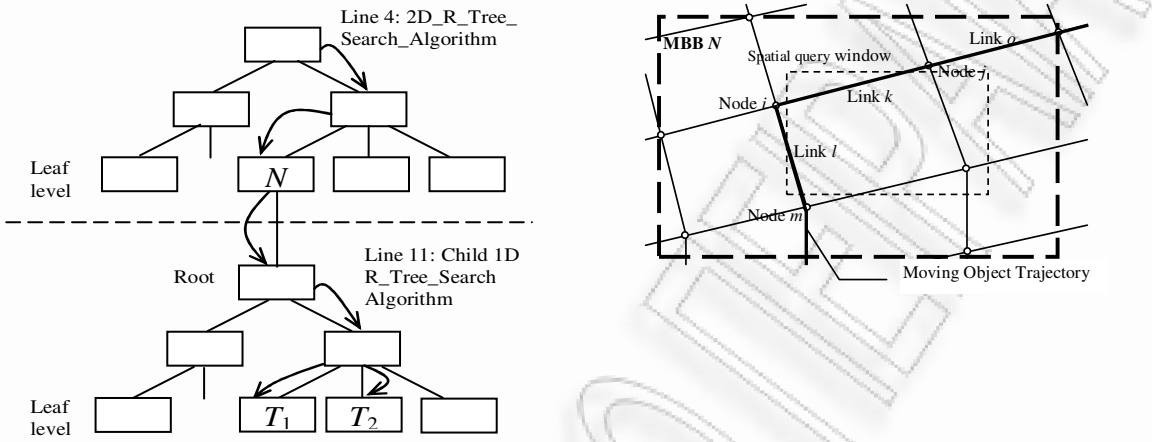
`Search-from-Parent-1D-R-tree`: Ο δεύτερος αλγόριθμος αναζήτησης του FNR-δέντρου χρησιμοποιεί το 1D Πρόγονο R-δέντρο και φαίνεται στο Σχήμα 2.22. Ξεκινά από τη ρίζα του Πρόγονου 1D R-δέντρου και εντοπίζει τις εγγραφές που ικανοποιούν τους χρονικούς περιορισμούς της επερώτησης. Κατόπιν, ακολουθώντας τους δείκτες, βρίσκει τα φύλλα των απογόνων 1D R-δέντρων που περιλαμβάνουν τις εγγραφές που ικανοποιούν τους χρονικούς περιορισμούς της επερώτησης και τα αντίστοιχα φύλλα του 2D R-δέντρου. Τέλος, ένα βήμα διαλογής εγγυάται ότι ο αλγόριθμος επιστρέφει μόνο τις εγγραφές που ικανοποιούν και τα χρονικά και τα χωρικά κριτήρια της ερώτησης.

```
1.  Algorithm FNR_tree_Search_from_Parent_1D(x_min, x_max, y_min, y_max, t_min, t_max)
2.     // Search in the Parent 1D R-tree with the 1D interval
3.     // (t_min, t_max) retrieving the entries overlapping it
4.     PEntries = Parent_1D_R_tree_search(x_min, x_max, y_min, y_max)
5.     // follow the pointer to the children 1D R-tree Leaf Nodes ND1
6.     FOR EACH PEntry IN PEntries
7.        ND1=Pentry.Child_1D_R_Tree_Leaf
8.        // follow the pointer to the parent 2D R-tree Leaf Node
9.        // ND2 to get spatial extent
10.       ND2=ND1.Parent_2D_R_Tree_Leaf
11.       // Refinement
12.       IF ND2.MBB is outside (x_min, x_max, y_min, y_max)
13.          Reject ND2
14.       ELSEIF ND2.MBB is inside (x_min, x_max, y_min, y_max)
15.          RETURN all entries of ND1
16.       ELSE // ND2 is partially inside (x_min, y_min, x_max, y_max)
17.          FOR EACH Entry IN ND1
18.           IF ND2.Links(Entry.Link_id).MBB is inside(x_min, x_max, y_min, y_max)
19.                RETURN the Entry
20.           ENDIF
21.          NEXT
22.       ENDIF
23.    NEXT
```

Σχήμα 2.22: Ο Αλγόριθμος του FNR-δέντρου `Search-from-Parent-1D-R-tree`

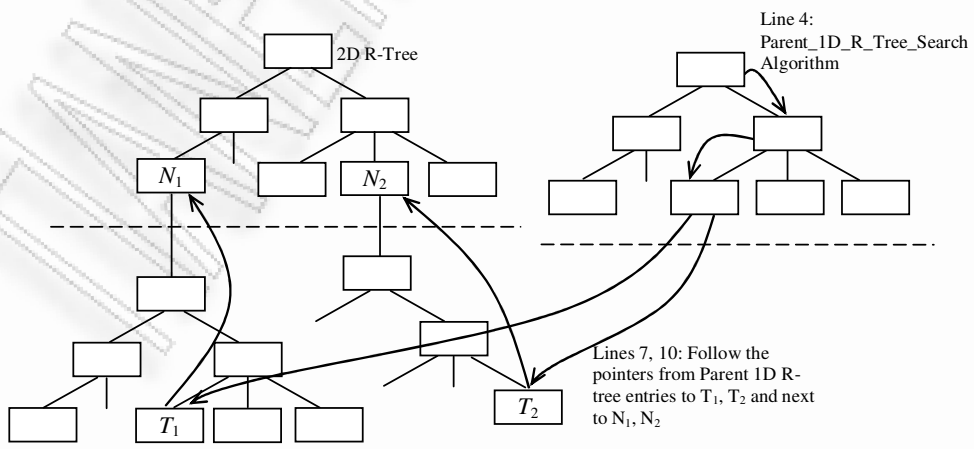
Έστω ότι κάνουμε μία αναζήτηση στο FNR-δέντρο με ένα χωροχρονικό παράθυρο επερώτησης (x_1, y_1, x_2, y_2, t_1, t_2) χρησιμοποιώντας τον πρώτο αλγόριθμο αναζήτησης, `Search-from-2D-R-tree` (Σχήμα 2.23). Το πρώτο βήμα απαιτεί μία χωρική αναζήτηση (για (x_1, y_1, x_2, y_2)) στο 2D R-δέντρο ώστε να εντοπίσουμε τα γραμμικά τμήματα και τα αντίστοιχα φύλλα του 2D R-δέντρου (στο παράδειγμά μας, φύλλο *N*) που αλληλεπικαλύπτονται με τη χωρικής συνιστώσα του παραθύρου επερώτησης. Στη συνέχεια, εκτελείται αναζήτηση του διαστήματος (t_1, t_2) σε όλα τα 1D R-δέντρα που αντιστοιχούν στα φύλλα του πρώτου βήματος. Στο παράδειγμά μας, η αναζήτηση μας οδηγεί στα

φύλλα T_1 και T_2 του 1D R-δέντρου που περιλαμβάνουν (μεταξύ άλλων) τους συνδέσμους *k*, *l* και *o*. Στο τελικό βήμα, ανακτούμε από την κύρια μνήμη τις συντεταγμένες κάθε συνδέσμου που επιλέξαμε στο δεύτερο βήμα και - επειδή το φύλλο *N* του 2D R-δέντρου αλληλεπικαλύπτεται μερικώς με το χωρικό παράθυρο της επερώτησης – ελέγχουμε και απορρίπτουμε αυτούς που βρίσκονται εκτός του χωρικού παραθύρου επερώτησης (στο παράδειγμά μας, σύνδεσμος *o*).



Σχήμα 2.23: Αναζήτηση του FNR-δέντρου με χρήση του Αλγορίθμου `Search-from-2D-R-tree`

Για να επιδείξουμε το δεύτερο αλγόριθμο αναζήτησης (`Search-from-Parent-1D-R-tree`), έστω και πάλι ένα χωροχρονικό παράθυρο επερώτησης (x_1, y_1, x_2, y_2, t_1, t_2) (Σχήμα 2.24). Το πρώτο βήμα του αλγορίθμου απαιτεί μία αναζήτηση στο Πρόγονο 1D R-δέντρο με τα (t_1, t_2) ως ορίσματα για να εντοπίσουμε τις εγγραφές στα φύλλα του που αλληλεπικαλύπτονται με αυτό το διάστημα. Κατόπιν, ακολουθώντας τους δείκτες στα φύλλα του απόγονου 1D R-δέντρου, εντοπίζουμε τους κόμβους που περιέχουν τις εγγραφές του 1D R-δέντρου που ικανοποιούν τους χρονικούς περιορισμούς της επερώτησης (φύλλα T_1, T_2). Αυτοί οι κόμβοι ανήκουν σε διαφορετικά απόγονα 1D R-δέντρα, που αντιστοιχούν σε διαφορετικά φύλλα του 2D R-δέντρου, οι οποίοι εντοπίζονται ακολουθώντας τους δείκτες προς αυτούς (κόμβοι φύλλου N_1, N_2). Τέλος, ελέγχουμε κατά πόσο οι εγγραφές που περιέχονται στους κόμβους N_1 και N_2 ικανοποιούν τον χωρικό περιορισμό της επερώτησης (x_1, y_1, x_2, y_2).



Σχήμα 2.24: Αναζήτηση στο FNR-δέντρο χρησιμοποιώντας τον Αλγόριθμο `Search-from-Parent-1D-R-tree`

Αν έχουμε κάτι να προσάψουμε στο δεύτερο αλγόριθμο αναζήτησης είναι το γεγονός ότι ενδιαφέρεται μόνο για τη χρονική θέση των δεδομένων και εφαρμόζει μία χωρική επιλογή στο τελευταίο βήμα μόνο. Από την άλλη πλευρά, θα περιμέναμε ο αλγόριθμος αυτός να είναι αποτελεσματικός σε περιπτώσεις όπου έχουν σημασία μόνο οι χρονικοί περιορισμοί της επερώτησης. Η συμπεριφορά αυτή παρουσιάζεται στην πειραματική ενότητα που ακολουθεί.

Αυτή τη στιγμή, αξίζει τον κόπο να επισημάνουμε ότι *το FNR-δέντρο θα είναι λειτουργικό και χωρίς την παρουσία του Πρόγονου 1D R-δέντρου*. Σε αυτή την περίπτωση, η μόνη τροποποίηση του αλγόριθμου εισαγωγής του FRN-δέντρου θα είναι η απουσία των γραμμών 11-18 (Σχήμα 2.18). Επιπλέον, ο πρώτος αλγόριθμος αναζήτησης (Σχήμα 2.21) θα εκτελεστεί «ως έχει». Γι' αυτό, η κατασκευή και η λειτουργία του FNR-δέντρου θα είναι δυνατή χωρίς την παρουσία του Πρόγονου 1D R-δέντρου.

Ωστόσο, η λειτουργία του δεύτερου αλγόριθμου αναζήτησης (Σχήμα 2.22) απαιτεί την ύπαρξη του Προγόνου 1D R-δέντρου. Αυτή η δομή μπορεί να κατασκευαστεί σε οποιαδήποτε χρονική στιγμή της λειτουργίας του FNR-δέντρου χρησιμοποιώντας τον αλγόριθμο κατασκευής που φαίνεται στο Σχήμα 2.25· αυτός ο αλγόριθμος μπορεί να προσπελάσει όλη τη δομή του FNR-δέντρου και απλά εισάγει τις χρονικές εκτάσεις όλων των φύλλων των απογόνων 1D R-δέντρων ως εγγραφές στο Πρόγονο 1D R-δέντρο.

```
 1.  Algorithm FNR_tree_Parent_1D_R_Tree_Construction
 2.     Create a new 1D R-tree
 3.     // Access the 2D R-tree. Use the 2D R-tree structure and locate
 4.     // every leaf node named ND2
 5.     FOR EACH Leaf Node ND2 IN 2D R-tree
 6.        // Follow the pointer to the child 1D R-tree RT
 7.        ND1=ND2.Child_1D_R_Tree_Leaf
 8.        // Access all the child 1D R-tree and insert the leaf
 9.        // nodes in the Parent 1D R-tree
10.        FOR EACH Leaf Node ND1 IN RT
11.           // Execute R-tree-insert algorithm in the Parent // 1D R-tree
12.           // and insert ND1 as a new entry
13.           Parent_1D_R_tree_insert(ND.MBB, ND.ptr)
14.        NEXT
15.     NEXT
```

Σχήμα 2.25: Ο Αλγόριθμος του FNR-δέντρου `Parent-1D-R-Tree-Construction`

2.5. Πειραματική Μελέτη: Απεριόριστη Κίνηση

Για να αξιολογήσουμε την απόδοση του TB*-δένδρου, υλοποιήσαμε τη δομή και τους αλγορίθμους που προτείνουμε, και το συγκρίναμε με το αρχικό TB-δέντρο [PJT00], καθώς επίσης και με το παραδοσιακό 3D R-δέντρο [TVS96].

2.5.1. Πειραματικό Πλαίσιο

Επιλέξαμε ως μέγεθος σελίδας και για τα τρία δέντρα τα 4 KB με αποτέλεσμα τη μέγιστη χωρητικότητα (*M*) για το TB*-δέντρο να είναι 338 και 145, για τα φύλλα και τους εσωτερικούς κόμβους, αντίστοιχα. Πέρα από την προσωρινή μνήμη LP που εισήχθη, χρησιμοποιήσαμε μία (μεταβλητού μεγέθους) προσωρινή μνήμη LRU χωρητικότητας 10% του μεγέθους ευρετηρίου, με μέγιστη χωρητικότητα 1000 σελίδες. Τα πειράματα εκτελέστηκαν σε PC με Microsoft Windows XP, επεξεργαστή AMD Athlon 64 3GHz, 1 GB RAM και σκληρό δίσκο αρκετών GB .

Δεδομένου ότι θέλαμε να μελετήσουμε τη συμπεριφορά των ευρετηρίων με κλιμακούμενο μέγεθος δεδομένων, χρησιμοποιήσαμε τα συνθετικά σύνολα δεδομένων GSTD που περιγράφηκαν στην παράγραφο 1.5.2 (τα σύνολα των πραγματικών δεδομένων της παραγράφου 1.5.1 που διαθέτουμε, έχουν σταθερό και μάλλον περιορισμένο μέγεθος για τους σκοπούς της μελέτης μας). Επιπλέον, χρησιμοποιήσαμε δύο διαφορετικές στρατηγικές για την εισαγωγή των συνόλων δεδομένων στις δύο δομές. Η πρώτη στρατηγική απαιτεί να διαταχθεί το σύνολο δεδομένων που θα εισαχθεί χρονικά. Αυτό συμβαίνει συνήθως σε online χωροχρονικές εφαρμογές, όπου, λόγω χρονικής μονοτονικότητας, περιμένουμε τα δεδομένα τροχιών να συλλέγονται και να εισάγονται στο ευρετήριο με αύξουσα χρονική σειρά (στο εξής, οργάνωση 'χρόνου' στα πειράματα που ακολουθούν). Η δεύτερη στρατηγική δεν κάνει αυτή την υπόθεση, με την προϋπόθεση ότι τα δεδομένα τροχιάς ενός κινούμενου αντικειμένου εισάγονται με χρονολογική σειρά. Αυτή είναι η περίπτωση όπου τα κινούμενα αντικείμενα καταγράφουν μεν τη θέση τους αλλά δεν διατηρούν online επικοινωνία με τον κεντρικό εξυπηρετητή που διατηρεί το ευρετήριο· το αντίθετο μάλιστα, τα αντικείμενα αποστέλλουν τις θέσεις τους μόλις υπάρχει η δυνατότητα (π.χ. όταν είναι στην ακτίνα της συσκευής που χρησιμοποιείται για τη μετάδοση), ή σε προγραμματισμένες χρονικές στιγμές. Το ίδιο ισχύει και όταν το ευρετήριο δημιουργείται μετά τη συμπίεση ενός άλλου ήδη υπάρχοντος ευρετηρίου ή οποιουδήποτε άλλου αρχείου που περιέχει πληροφορίες τροχιάς. Έτσι, για να προσομοιώσουμε και τις δύο προηγούμενες καταστάσεις, τα δεδομένα των τροχιών εισάγονται στο ευρετήριο σε αύξουσα σειρά της ταυτότητας / χρόνου των κινούμενων αντικειμένων (στο εξής οργάνωση 'ταυτότητα / χρόνος' στα πειράματα που ακολουθούν).

Πίνακας 2.2: Αποτελέσματα όσον αφορά το μέγεθος των δένδρων (συνθετικά δεδομένα GSTD)

Δεδομένα	*Μέγεθος ευρετηρίου σε σελίδες των 4 KB*		
	3D R-δένδρο	*TB-δένδρο*	*TB*-δένδρο*
GSTD 100	6253	3054	1522
GSTD 250	15471	7649	3808
GSTD 500	30937	15301	7597
GSTD 1000	61864	30588	15156
GSTD 2000	122703	61170	30557

2.5.2. Αποτελέσματα για το Μέγεθος του Δέντρου και το Κόστος Εισαγωγής

Τα μεγέθη των ευρετηρίων που δημιουργήθηκαν φαίνονται στους δύο πίνακες (Πίνακας 2.2 και Πίνακας 2.3). Είναι σαφές ότι το μέγεθος του TB*-δέντρου είναι σχεδόν το μισό του μεγέθους του TB-δέντρου και σχεδόν το 15% του μεγέθους του κλασσικού 3D R-δέντρου. Επίσης, η κατάληψη του παρεχόμενου χώρου (space utilization) τόσο για το TB όσο και για το TB* είναι υψηλή όπως αναμένεται: περίπου 99% και 96%, αντίστοιχα, ενώ η αντίστοιχη τιμή του 3D R-δέντρου είναι 56%, που είναι τυπική τιμή για τα R-δέντρα. Αποδεικνύεται συνεπώς ότι το TB*-δέντρο είναι μία πολύ συμπαγής δομή ευρετηρίου, που υπερτερεί και των δύο ανταγωνιστών του.

Ο Πίνακας 2.3 παρουσιάζει επίσης τα αποτελέσματα των προσπελάσεων κόμβων ανά εισαγωγή για όλα τα σύνολα δεδομένων. Κάθε εισαγωγή ενός νέο γραμμικού τμήματος τροχιάς στο TB*-δέντρο απαιτούνται κατά μέσο όρο 1.4 προσπελάσεις κόμβων. Ο λόγος γι' αυτή την πρώτης τάξης απόδοση

είναι η χρήση της δομής κατακερματισμού στη κύρια μνήμη, που «δείχνει» απευθείας τον κόμβο στον οποίο πρέπει να εισαχθεί η νέα εγγραφή, καθώς και η παρουσία της προσωρινής μνήμης LP. Αντιθέτως, το TB-δέντρο και το 3D R-δέντρο απαιτούν μεγαλύτερο αριθμό προσπελάσεων ανά εισαγωγή, 4.0 και 2.3, αντίστοιχα· αυτό οφείλεται στον αλγόριθμο `FindNode` του TB-δέντρου, που ακολουθεί μια διαδρομή πολλών κατευθύνσεων (και όχι μοναδικής κατεύθυνσης όπως κάνει ο αλγόριθμος `ChooseLeaf` του R-δέντρου) για να βρει τον κατάλληλο κόμβο για να τοποθετήσει τη νέα εγγραφή.

Πίνακας 2.3: Μέγεθος ευρετηρίου, χρησιμοποίηση χώρου και αριθμός προσπελάσεων κόμβων ανά εισαγωγή στο σύνολο δεδομένων *GSTD 2000*

	3D R-δένδρο	*TB-δένδρο*	*TB*-δένδρο*
Μέγεθος ευρετηρίου (KB ανά αντικείμενο)	99.6	30.6	15.2
Εκμετάλλευση χώρου	56%	99%	96%
Προσπελάσεις κόμβων ανά εισαγωγή (μέσο)	2.3	4.0 (1.2)	1.4

Εδώ, πρέπει να τονίσουμε, ότι ο αρχικός αλγόριθμος εισαγωγής του TB-δέντρου μπορεί να τροποποιηθεί και να χρησιμοποιήσει τη δομή πρώτης γραμμής που εισάγεται στο TB*-δέντρο και να αντικαταστήσουμε το βήμα που εκτελεί τον αλγόριθμο `FindNode`, απλά, ακολουθώντας τον δείκτη μέχρι τον 'τρέχοντα' κόμβο φύλλου του κινούμενου αντικειμένου. Το αρχικό TB-δέντρο μπορεί επίσης να χρησιμοποιήσει την προσωρινή μνήμη LP, η οποία αντίθετα δεν μπορεί να χρησιμοποιηθεί στην περίπτωση ευρετηρίων που δεν προσανατολίζονται σε τροχιές (όπως το 3D R-δέντρο), γιατί μια τέτοια στρατηγική θα σήμαινε ότι η προσωρινή μνήμη LP θα έπρεπε να κρατήσει όλα τα φύλλα του ευρετηρίου. Έτσι, για να επιδείξουμε την επίδραση αυτών των βελτιώσεων (πρώτης γραμμής και προσωρινής μνήμης LP) στη συμπεριφορά του απλού TB-δέντρου τα χρησιμοποιήσαμε στα πειράματά μας και λάβαμε έτσι τον δεύτερο αριθμό (1.2) που δείχνει ο Πίνακας 2.3. Όπως φαίνεται, οι τεχνικές αυτές βελτιώνουν δραστικά την απόδοση εισαγωγής, επιτρέποντας στο απλό TB-δέντρο να υποστηρίξει υψηλούς ρυθμού εισαγωγών.

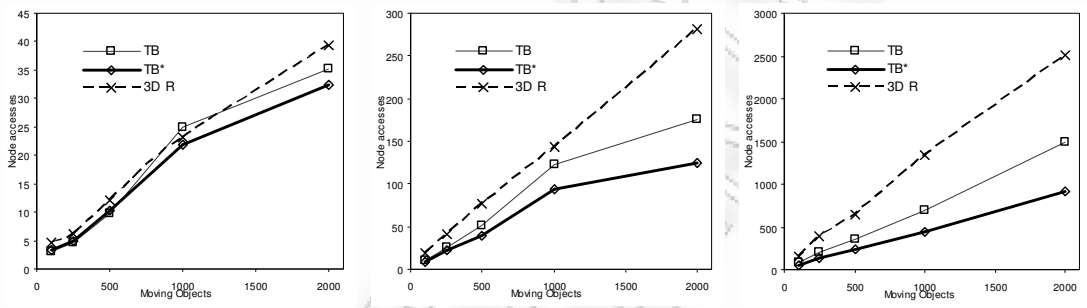
2.5.3. Αποτελέσματα ως προς το Κόστος Αναζήτησης

Χρησιμοποιήθηκαν επερωτήσεις εύρους, χρονικής στιγμής και συνδυαστικές για την αξιολόγηση της απόδοσης του TB*-δέντρου. Πιο συγκεκριμένα, χρησιμοποιήσαμε τις ακόλουθες ομάδες πέντε επερωτήσεων (Q_1 – Q_5):

- Q_1–Q_3: τρεις ομάδες 500 κυβικών παραθύρων επερώτησης με εύρος 0.01%, 0.1% και 1% του συνολικού χώρου, αντίστοιχα, αυξάνοντας τον αριθμό των κινούμενων αντικειμένων (σύνολα δεδομένων GSTD 100 – GSTD 2000).
- Q_4: μία ομάδα 500 παραθύρων επερωτήσεων χρονικής στιγμής με το 100% της έκτασης των χωρικών διαστάσεων και μηδενική χρονική έκταση, αυξάνοντας τον αριθμό των κινούμενων αντικειμένων (σύνολα δεδομένων GSTD 100 – GSTD 2000).
- Q_5: μία ομάδα 500 συνδυαστικών επερωτήσεων με εσωτερικό παράθυρο 0.01% και εξωτερικό 1% του συνολικού χώρου, αυξάνοντας τον αριθμό των κινούμενων αντικειμένων (σύνολα δεδομένων GSTD 100 – GSTD 2000)

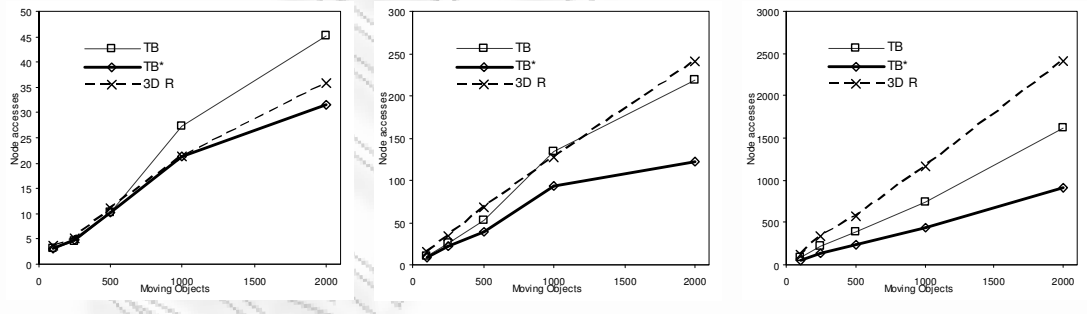
2.5.3.1. Αποτελέσματα σε Επερωτήσεις Εύρους

Το Σχήμα 2.26 παρουσιάζει το μέσο αριθμό προσπελάσεων κόμβων ανά επερώτηση για διάφορες επερωτήσεις εύρους και σύνολα δεδομένων. Πιο συγκεκριμένα, το Σχήμα 2.26 δείχνει το μέσο αριθμό προσπελάσεων κόμβων για επερωτήσεις εύρους μ' ένα κυβικό παράθυρο 0.01%, 0.1% και 1% του συνολικού χώρου στα συνθετικά δεδομένα που εισάγονται στις δομές οργανωμένα κατά χρονολογική σειρά ανεξάρτητα της ταυτότητας ('*χρόνος*'), ενώ το Σχήμα 2.27 δείχνει το μέσο αριθμό προσπελάσεων κόμβων για τις ίδιες επερωτήσεις εύρους στα ίδια δεδομένα όταν αυτά εισάγονται στις δομές οργανωμένα κατά την ταυτότητά τους και στη συνέχεια κατά τη χρονολογική σειρά ('*ταυτότητα / χρόνος*'). Είναι σαφές ότι το TB*-δέντρο έχει ανώτερη απόδοση σε επερωτήσεις εύρους και από τους δύο ανταγωνιστές του ως προς τις επερωτήσεις με μέγεθος 0.1% (Q_2) και 1% (Q_3) του συνολικού χώρου και για τις δύο διαφορετικές οργανώσεις του τρόπου εισαγωγής. Όσον αφορά τις επερωτήσεις με μικρότερο μέγεθος (0.01% του συνολικού χώρου, Q_1) και την οργάνωση των δεδομένων με μόνο χρονολογική σειρά, η απόδοση του TB*-δέντρου είναι μόνον οριακά καλύτερη από το αρχικό TB-δέντρο, μια διαφορά που γίνεται σαφέστερη όσο ο πληθυσμός του συνόλου δεδομένων μεγαλώνει.



(a) Q_1 (b) Q_2 (c) Q_3

Σχήμα 2.26: Επερωτήσεις Q_1 – Q_3 σε δεδομένα που έχουν εισαχθεί οργανωμένα κατά τον χρόνο



(a) Q_1 (b) Q_2 (c) Q_3

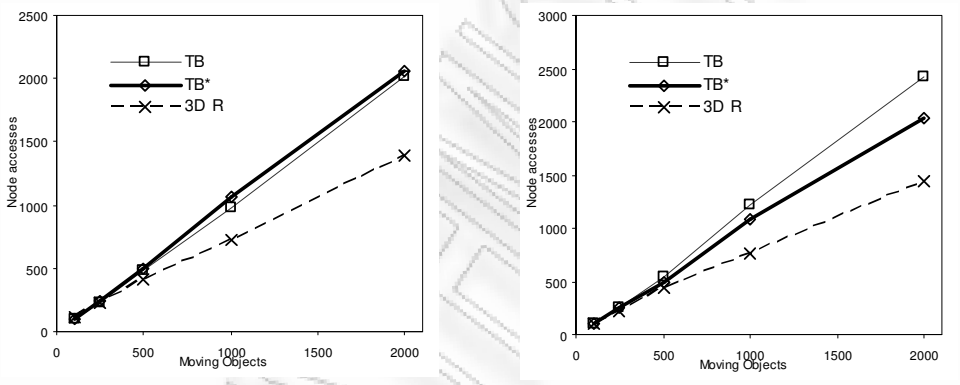
Σχήμα 2.27: Επερωτήσεις Q_1 – Q_3 σε δεδομένα που έχουν εισαχθεί οργανωμένα βάσει ταυτότητας / χρόνου

Από τη σύγκριση μεταξύ των παραπάνω σχημάτων (Σχήμα 2.26 και Σχήμα 2.27) προκύπτει επιπλέον το συμπέρασμα ότι ενώ το TB*-δέντρο και το 3D R-δέντρο δείχνουν περίπου την ίδια συμπεριφορά μεταξύ των δύο διαφορετικών στρατηγικών εισαγωγής (η κλίση των γραμμών του TB*-δέντρου και του 3D R-δέντρου είναι περίπου η ίδια στο Σχήμα 2.26 και στο Σχήμα 2.27), η συμπεριφορά του αρχικού TB-δέντρου επηρεάζεται σημαντικά, γεγονός που στη δεύτερη περίπτωση οδηγεί σε ένα δέντρο με δραστικά μειωμένη απόδοση. Η συμπεριφορά αυτή είναι αναμενόμενη για το

TB-δέντρο διότι η βασική υπόθεση στην οποία βασίζεται η αποτελεσματικότητα του δέντρου σε επερωτήσεις εύρους, δηλαδή η εισαγωγή νέων εγγραφών με χρονολογική σειρά, δεν ισχύει πλέον.

2.5.3.2. Αποτελέσματα σε Επερωτήσεις Χρονικής Στιγμής

Το Σχήμα 2.28(a) παρουσιάζει το μέσο αριθμό προσπελάσεων κόμβων για επερωτήσεις χρονικής στιγμής με 100% χωρικό εύρος σε κάθε χωρική διάσταση (π.χ. βρες όλα τα αντικείμενα σε ένα συγκεκριμένο χρονικό αποτύπωμα) όταν τα δεδομένα εισάγονται οργανωμένα με καθαρά χρονολογική σειρά. Όπως, αποδεικνύεται στην πρώτη περίπτωση, το αρχικό TB-δέντρο υπερτερεί μόνο οριακά του TB*-δέντρου, συμπεριφορά που είναι αναμενόμενη, δεδομένου ότι το αρχικό TB-δέντρο εκμεταλλεύεται πλήρως τη μονοτονικότητα του χρόνου και αποθηκεύει τροχιές αντικειμένων λαμβάνοντας υπ' όψιν μόνο τη σειρά με την οποία εισάγονται στο ευρετήριο. Ωστόσο, το γεγονός αυτό αποδεικνύεται ότι είναι μειονέκτημα όταν τα δεδομένα δεν εισάγονται με καθαρά χρονολογική σειρά (Σχήμα 2.28(b))· σε αυτή την περίπτωση, το TB*-δέντρο έχει καλύτερη απόδοση, συμπεριφορά που είναι ανάλογη με αυτή που επιδεικνύεται από αυτή τη δομή όταν τα δεδομένα εισάγονται με αυστηρά χρονολογική σειρά.



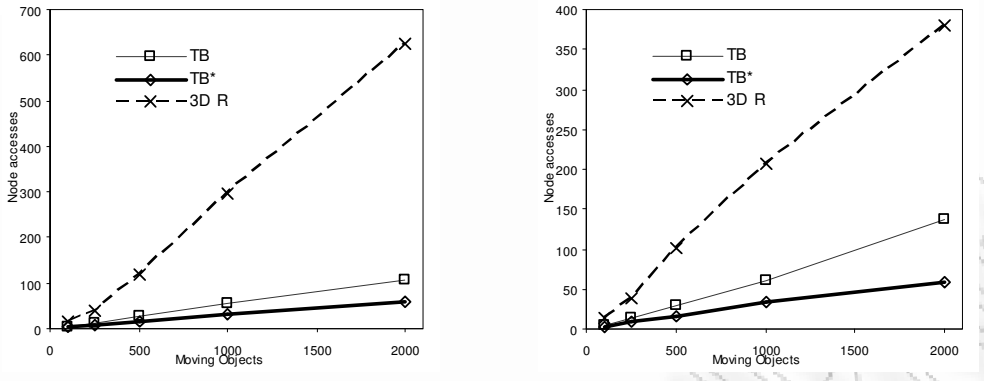
(a) (b)

Σχήμα 2.28: Επερωτήσεις Q_4 σε δεδομένα που έχουν εισαχθεί οργανωμένα βάσει (a) χρόνου, (b) ταυτότητας / χρόνου

Συμπερασματικά, η απόδοση του TB*-δέντρου σε επερωτήσεις χρονικής στιγμής μειώνεται σε σύγκριση με την απόδοσή του σε επερωτήσεις εύρους, παρά το γεγονός ότι και πάλι υπερτερεί του TB-δέντρου όταν τα δεδομένα εισάγονται σε δομές οργανωμένα βάσει 'ταυτότητας / χρόνου'.

2.5.3.3. Αποτελέσματα σε Συνδυαστικές Επερωτήσεις

Το Σχήμα 2.29 δείχνει το μέσο αριθμό προσπελάσεων κόμβων που απαιτούνται από το TB-, το TB*- και το 3D R-δέντρο για να απαντήσουμε σε συνδυαστικές επερωτήσεις. Και στα δύο σχήματα το TB*-δέντρο υπερτερεί του TB- και του 3D R-δέντρου ανεξάρτητα του πλήθους του συνόλου δεδομένων. Επιπλέον, βάσει όλων των προηγούμενων πειραμάτων, η διαφορά στην απόδοση μεταξύ των TB- και TB*-δένδρων αυξάνεται υπέρ του TB*-δέντρου με το πλήθος του συνόλου δεδομένων. Επιπλέον, στη δεύτερη περίπτωση (Σχήμα 2.29(b)) όπου τα δεδομένα εισάγονται στα δέντρα με οργάνωση 'ταυτότητας / χρόνου', το TB*-δέντρο παρουσιάζει ακόμα καλύτερη απόδοση σε σχέση με τον προκάτοχό του.



(a) (b)

Σχήμα 2.29: Οι συνδυαστικές επερωτήσεις, (Q_5) σε δεδομένα που έχουν εισαχθεί οργανωμένα βάσει (a) χρόνου, (b) ταυτότητας / χρόνου

2.5.4. Σύνοψη των Πειραμάτων

Τα πειράματα που διεξήχθησαν για την αξιολόγηση της απόδοσης του προτεινόμενου TB*-δέντρου σε σχέση με το αρχικό TB-δέντρο και το 3D R-δέντρο κατέδειξαν ότι το προτεινόμενο ευρετήριο υποστηρίζει αποτελεσματικά τις επερωτήσεις εύρους. Πιο συγκεκριμένα, όταν ασχολούμαστε με σχετικά μεγάλες εκτάσεις επερωτήσεων (0.1% και 1% του συνολικού χώρου) το TB*-δέντρο πάντοτε υπερέχει του TB-δέντρου και του 3D R-δέντρου, ενώ σε μικρότερες εκτάσεις επερωτήσεων (0.01% του συνολικού χώρου) είναι οριακά καλύτερο από τους ανταγωνιστές του, ένα πλεονέκτημα που γίνεται σαφέστερο όσο αυξάνεται το πλήθος του συνόλου δεδομένων. Αντιθέτως σε επερωτήσεις χρονικής στιγμής το TB*-δέντρο φαίνεται να έχει μειωμένη απόδοση και απαιτεί κάπως περισσότερες προσπελάσεις σελίδων για να επεξεργαστεί μια επερώτηση χρονικής στιγμής· σε αυτή την περίπτωση, ο 'κότινος ελιάς' απονέμεται στο 3D R-δέντρο. Τέλος, για τις συνδυαστικές επερωτήσεις το νέο ευρετήριο παρουσιάζει ανώτερη απόδοση από το αρχικό TB-δέντρο (αλλά και το 3D R-δέντρο) σε όλες τις περιπτώσεις. Επιπλέον, η υπεροχή του προτεινόμενου ευρετηρίου εδραιώνεται στην γενική περίπτωση που τα ευρετήρια δημιουργούνται εισάγοντας τα δεδομένα των τροχιών με μη χρονολογική σειρά (δηλαδή, με την οργάνωση 'ταυτότητας / χρόνου'), μία περίπτωση αναμενόμενη σε πραγματικές εφαρμογές και όταν το ευρετήριο δημιουργείται μετά από συμπίεση ενός συνόλου δεδομένων. Υπ' αυτές τις προϋποθέσεις, το TB*-δέντρο είναι πάντα πολύ πιο αποτελεσματικό από το TB-δέντρο. Ως προς το μέγεθος του TB*-δέντρου, η εκμετάλλευση χώρου, όπως και στο αρχικό TB-δέντρο, αγγίζει το 96% και το μέσο μέγεθος ανά κινούμενο αντικείμενο είναι το ήμισυ του TB-δέντρου. Συνολικά, το TB*-δέντρο υποστηρίζει υψηλούς ρυθμούς εισαγωγής επειδή ο αλγόριθμος εισαγωγής είναι πολύ γρήγορος, είναι πιο συμπαγές από τους ανταγωνιστές του, συμπεριφέρεται καλά σε μη χρονολογικές εισαγωγές τροχιάς που εμφανίζονται σε πραγματικά περιβάλλοντα και υποστηρίζει διαγραφές τροχιάς και συμπίεση τροχιάς αποτελεσματικά.

2.6. Πειραματική Μελέτη: Κίνηση με Περιορισμούς Δικτύου

Για να καθορίσουμε τις προϋποθέσεις βάσει των οποίων το FNR-δέντρο είναι αποτελεσματικό, το συγκρίναμε με άλλες μεθόδους χωροχρονικής ευρετηριοποίησης, πιο συγκεκριμένα το 3D R-, το TB- και το TB*-δέντρο.

2.6.1. Πειραματικό Πλαίσιο

Για να αξιολογήσουμε την απόδοση του FNR-δέντρου, υλοποιήσαμε τη δομή του στην κύρια μνήμη, προσομοιάζοντας τη συμπεριφορά του. Επιλέξαμε μέγεθος σελίδας για όλα τα δέντρα τα 4096 bytes, αποκτώντας τις εξής ρυθμίσεις χωρητικότητας για το FNR-δέντρο: (a) για το 2D R-δέντρο χωρητικότητες 193 και 202, για φύλλα και εσωτερικούς κόμβους, αντίστοιχα· (b) για τα 1D R-δέντρα 290 και 339, για φύλλα και εσωτερικούς κόμβους, αντίστοιχα· (c) για το Πρόγονο 1D R-δέντρο, χωρητικότητα 339 και για φύλλα και για εσωτερικούς κόμβους. Σε σύγκριση με τη χωρητικότητα των ανταγωνιστών του, το FNR-δέντρο είναι τόσο συμπαγές όσο και το TB*-δέντρο (υπερτερώντας των άλλων δύο ευρετηρίων). Χρησιμοποιήσαμε επίσης μία (μεταβλητού μεγέθους) προσωρινή μνήμη LRU χωρητικότητας 10% του μεγέθους ευρετηρίου, με μέγιστη χωρητικότητα 1000 σελίδες. Τα πειράματα εκτελέστηκαν σε PC με Microsoft Windows XP με επεξεργαστή AMD Athlon 64 3GHz, 1 GB RAM. Για να πειραματιστούμε με κλιμακούμενους όγκους δεδομένων και να μελετήσουμε τη συμπεριφορά του ευρετηρίου κάτω από διάφορες συνθήκες χρησιμοποιήσαμε τα συνθετικά σύνολα δεδομένων NG που εισήχθησαν στην ενότητας 1.5.3 (και πάλι, τα πραγματικά δεδομένα της παραγράφου 1.5.1 που διαθέτουμε δεν είναι κατάλληλα για τους σκοπούς της μελέτης μας επειδή έχουν σταθερό, μάλλον περιορισμένο μέγεθος).

Πίνακας 2.4: Αποτελέσματα για το μέγεθος των δένδρων (συνθετικά σύνολα δεδομένων NG με περιορισμό στο δίκτυο)

Δεδομένα	*Μέγεθος ευρετηρίου σε σελίδες των 4 KB*			
	FNR-δένδρο	*3D R-δένδρο*	*TB-δένδρο*	*TB*-δένδρο*
NG 200	769	1204	770	424
NG 400	1139	2397	1533	848
NG 800	1850	4603	3040	1669
NG 1200	2575	7001	4499	2495
NG 1600	3281	9234	5972	3310
NG 2000	4030	11636	7455	4158

2.6.2. Αποτελέσματα για το Μέγεθος του Δέντρου και το Κόστος Εισαγωγής

Το μέγεθος των ευρετηρίων παρουσιάζεται στον παραπάνω πίνακα (Πίνακας 2.4). Όπως αναφέρεται εκεί, το FNR-δέντρο είναι πολύ μικρότερο του 3D R-δέντρου και του TB-δέντρου. Η αναλογία μεταξύ του μεγέθους του ευρετηρίου του FNR και του 3D R-δέντρου ποικίλει μεταξύ του 0.30 και του 0.45 για μεγάλο αριθμό κινούμενων αντικειμένων. Για παράδειγμα, το μέγεθος του FNR-δέντρου για 2000 κινούμενα αντικείμενα είναι περίπου 16 MB, ενώ το μέγεθος του αντίστοιχου 3D R-δέντρου είναι 48 MB. Μόνο το TB*-δέντρο φαίνεται να έχει συγκρίσιμο μέγεθος με το FNR-δέντρο.

Πίνακας 2.5: Μέγεθος ευρετηρίου, εκμετάλλευση χώρου και προσπελάσεις κόμβων ανά εισαγωγή στο σύνολο δεδομένων *NG 2000*

	FNR-δένδρο	*3D R-δένδρο*	*TB-δένδρο*	*TB*-δένδρο*
Μέγεθος ευρετηρίου (KB ανά αντικείμενο)	8.1	24.0	14.9	8.3
Εκμετάλλευση χώρου	92%	64%	86%	75%
Προσπελάσεις κόμβων ανά εισαγωγή (μέσο)	2.0	2.1	4.0	1.04

Παρόμοια αποτελέσματα παρουσιάζονται σχετικά και με την εκμετάλλευση του διατιθέμενου χώρου στο ευρετήριο. Όπως φαίνεται στον παραπάνω πίνακα (Πίνακας 2.5), η χρησιμοποίηση του χώρου στο 3D R-δέντρο είναι περίπου η συνήθης της τάξης του 65%, η οποία παραμένει σταθερή ανεξάρτητα από τον αριθμό των κινούμενων αντικειμένων. Παρομοίως, η εκμετάλλευση του χώρου για το TB- και το TB*-δέντρο είναι περίπου 86% και 75%, αντίστοιχα, ποσοστά που παραμένουν ανεπηρέαστα από τον αριθμό των κινούμενων αντικειμένων. Παρόλα αυτά, θα πρέπει να επισημάνουμε ότι η εκμετάλλευση χώρου του TB- και του TB*-δέντρου επηρεάζεται κυρίως από τον αριθμό των ληφθέντων θέσεων για κάθε τροχιά, που είναι σχετικά χαμηλός, περίπου 500 κορυφές ανά τροχιά· λαμβάνοντας υπ' όψιν ότι κάθε φύλλο περιλαμβάνει περίπου 300 εγγραφές, διαφαίνεται σαφώς ότι κάθε τροχιά θα πρέπει να καταλαμβάνει 2 φύλλα, το πρώτο θα είναι πλήρες και το άλλο μισογεμάτο. Αντιθέτως, η χρησιμοποίηση του χώρου, του FNR-δέντρου αυξάνεται με το πλήθος του συνόλου δεδομένων. Έτσι, η χρησιμοποίηση του χώρου του FNR-δέντρου με 200 κινούμενα αντικείμενα είναι 65 %, ενώ, για 1200 και πάνω αγγίζει το 92%.

Όσον αφορά τα αποτελέσματα για τον αριθμό προσπελάσεων κόμβων ανά εισαγωγή, κάθε εισαγωγή ενός 3D γραμμικού τμήματος στο FNR-δέντρο απαιτεί ένα μέσο κόστος 2.0 προσπελάσεων, ενώ μία εισαγωγή στο 3D R-δέντρο απαιτεί κατά μέσο όρο 2.1. Είναι σαφές, ότι η προ-αναζήτηση στο 2D R-δέντρο για να βρούμε το 1D R-δέντρο του νεοεισαχθέντος γραμμικού τμήματος δεν προσθέτει σημαντικό επιπρόσθετο φόρτο, κυρίως λόγω της επίδρασης της προσωρινής μνήμης LRU. Όσον αφορά το TB και το TB*-δέντρο, εξακολουθούν να επιδεικνύουν την ίδια συμπεριφορά που παρατηρήθηκε στα προηγούμενα πειράματα σε απεριόριστο χώρο και είναι ικανά να υποστηρίξουν υψηλούς ρυθμούς εισαγωγών.

2.6.3. Αποτελέσματα ως προς το Κόστος Αναζήτησης

Διάφορες επερωτήσεις εύρους και χρονικής στιγμής χρησιμοποιήθηκαν για να αξιολογήσουν την απόδοση του FNR-δέντρου. Και οι δύο τύπου ερωτήσεων εκτελέσθηκαν έναντι του FNR-, του 3D R-, του TB- και του TB*-δέντρου δεικτοδοτώντας τα σύνολα δεδομένων NG. Πιο συγκεκριμένα χρησιμοποιήσαμε σύνολα των 500 επερωτήσεων με τα ακόλουθα παράθυρα επερωτήσεων:

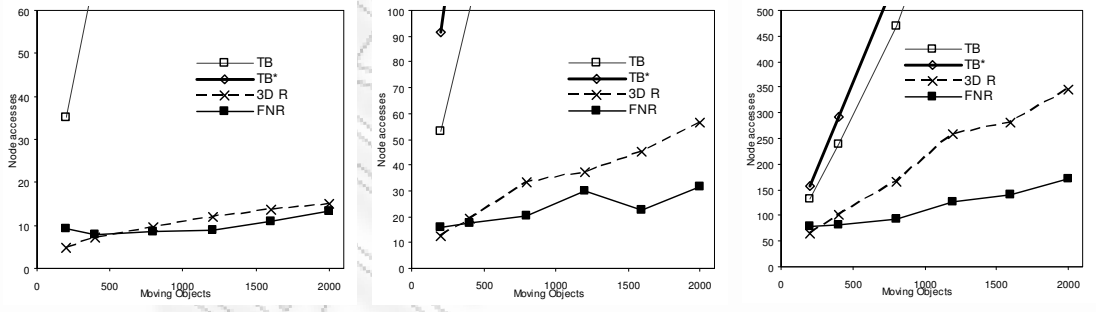
- Q_1–Q_3: τρεις ομάδες 500 κυβικών παραθύρων επερώτησης με εύρος 0.01%, 0.1% και 1% του συνολικού χώρου, αντίστοιχα, αυξάνοντας τον αριθμό των κινούμενων αντικειμένων.
- Q_4–Q_6: τρία παράθυρα επερωτήσεων χρονικής στιγμής με εύρος 1%, 10% και 100% που εκτείνονται σε κάθε χωρική διάσταση και μηδενική χρονική έκταση.

Χρησιμοποιήσαμε τον αλγόριθμο `Search-from-2D-R-tree` (Σχήμα 2.21) του FNR-δέντρου για την εκτέλεση όλων των παραπάνω επερωτήσεων και επιπλέον δοκιμάσαμε το `Search-from-Parent-1D-R-tree` (Σχήμα 2.22) στο Q_6 δηλαδή με 100% έκταση σε κάθε χωρική διάσταση.

2.6.3.1. Αποτελέσματα σε Επερωτήσεις Εύρους

Το Σχήμα 2.30 δείχνει το μέσο αριθμό προσπελάσεων κόμβων ανά επερώτηση για διάφορα μεγέθη επερωτήσεων και σύνολα δεδομένων. Πιο συγκεκριμένα, τα (a), (b) και (c), στο Σχήμα 2.30 δείχνουν το μέσο αριθμό προσπελάσεων κόμβων για επερωτήσεις εύρους με παράθυρο μεγέθους 0.01%, 0.1% και 1% του συνολικού χώρου. Όπως αποδεικνύεται σαφώς, το FNR-δέντρο έχει υψηλότερη απόδοση

σε επερωτήσεις εύρους σε σχέση με όλους τους ανταγωνιστές του για πλήθος δεδομένων πάνω από ένα κατώφλι, σε όλα τα μεγέθη επερωτήσεων. Το σημείο ισορροπίας μετά από το οποίο το FNR-δέντρο υπερτερεί των υπολοίπων εξαρτάται από το μέγεθος επερώτησης. Πιο συγκεκριμένα το σημείο ισορροπίας είναι περίπου στα 400 κινούμενα αντικείμενα για μικρά μεγέθη επερωτήσεων (0.01%), ενώ μεγαλύτερα μεγέθη επερωτήσεων έχουν ως αποτέλεσμα μικρότερη τιμή για το σημείο ισορροπίας, στα περίπου 200 κινούμενα αντικείμενα. Όσον αφορά τις υπόλοιπες δομές, το 3D R-δέντρο αποδίδει πάντα πολύ καλύτερα από το TB και το TB*-δέντρο όταν χρησιμοποιούμε τα δεδομένα που κινούνται στα δίκτυα· σημειώστε ότι τα διαγράμματα μπορεί να μην περιλαμβάνουν όλες τις καμπύλες και των τεσσάρων ευρετηρίων, λόγω του γεγονότος ότι δεν περιλαμβάνονται στην δεδομένη κλίμακα απεικόνισης (δηλαδή έχουν τιμές μεγαλύτερες από αυτές που απεικονίζονται στον *y*-άξονα του διαγράμματος). Πρέπει να επισημάνουμε παρόλα αυτά, ότι η παρατήρηση που αφορά την απόδοση του 3D R-, του TB και του TB*-δέντρου δεν μπορεί να γενικευθεί. Πιο συγκεκριμένα, η κακή απόδοση που επιδεικνύουν το TB και το TB*-δέντρο στα πειράματα, οφείλεται κυρίως στο μικρό αριθμό θέσεων σε κάθε τροχιά, που υποχρεώνει κάθε τροχιά να διαιρεθεί μόνο μεταξύ δύο κόμβων του δέντρου. Αναμενόμενο είναι (όπως έχουν δείξει και τα προηγούμενα πειράματα) ότι όσο μεγαλώνει η χρονική έκταση των τροχιών και προστίθενται κι άλλες χρονικά αποτυπωμένες θέσεις σε κάθε τροχιά, η απόδοση του TB και του TB*-δέντρου θα τείνουν σε πιο 'φυσιολογικές' τιμές όπως αυτές που παρουσιάστηκαν στα πειράματα στον απεριόριστο χώρο. Ωστόσο, το εργαλείο που παρέχεται από την [Bri02] και χρησιμοποιείται για την παραγωγή των συνθετικών τροχιών με περιορισμούς δικτύων, δεν μπορεί να μας δώσει τροχιές μεγαλύτερου μήκους· συνεπώς δεν μπορούμε να χρησιμοποιήσουμε μεγαλύτερα (μακρύτερα στη χρονική διάσταση) σύνολα δεδομένων.



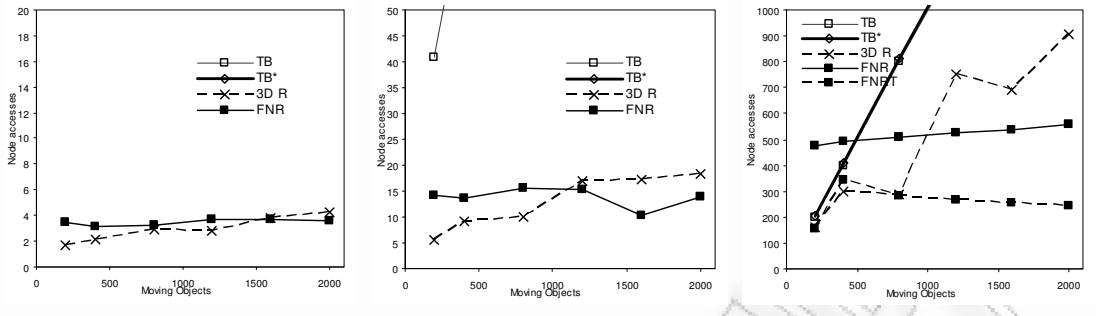
(a) Q_1 (b) Q_2 (c) Q_3

Σχήμα 2.30: Επερωτήσεις $Q_1 - Q_3$

2.6.3.2. Αποτελέσματα σε Επερωτήσεις Χρονικής Στιγμής

Η απόδοση του FNR-δέντρου σε επερωτήσεις χρονικής στιγμής μειώνεται σε σύγκριση με την απόδοση του σε επερωτήσεις εύρους, παρόλο που εξακολουθεί να υπερτερεί των ανταγωνιστών του στις περισσότερες περιπτώσεις. Το Σχήμα 2.31 δείχνει το μέσο αριθμό προσπελάσεων κόμβων για επερωτήσεις χρονικής στιγμής με διάφορα σύνολα δεδομένων και χωρικές εκτάσεις. Πιο συγκεκριμένα, το FNR-δέντρο παρουσιάζει καλύτερη απόδοση από τους ανταγωνιστές του από ένα συγκεκριμένο αριθμό κινούμενων αντικειμένων και πάνω· το σημείο ισορροπίας εξαρτάται από το μέγεθος της επερώτησης και είναι περίπου 1600 για μέγεθος επερώτησης 1%, 1200 για 10% και 1000 για 100% σε κάθε χωρική διάσταση. Οι άλλοι δύο ανταγωνιστές (TB και TB*-δέντρα) εξακολουθούν

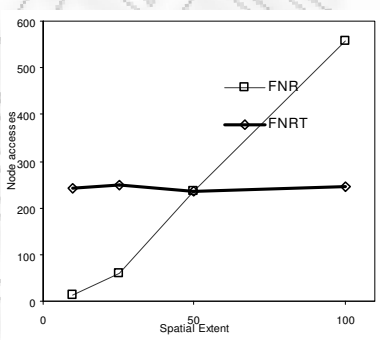
να παρουσιάζουν τα ίδια μειονεκτήματα με το προηγούμενο πείραμα που αφορά τις γενικές επερωτήσεις εύρους.



(a) (b) (c)

Σχήμα 2.31: Επερωτήσεις Q_4 – Q_6

Επιπλέον, η γραμμή που σημειώνεται με FNRT στο Σχήμα 2.31(c) παριστάνει την απόδοση του FNR-δέντρου χρησιμοποιώντας το δεύτερο αλγόριθμο αναζήτησης, που σε αυτό τον ειδικό τύπο επερωτήσεων υπερτερεί του πρώτου. Πιο συγκεκριμένα, η χρήση του δεύτερου αλγορίθμου μετατοπίζει το σημείο ισορροπίας από το οποίο το FNR-δέντρο είναι καλύτερο από το 3D R-δέντρο από τα 1000 στα 800 κινούμενα αντικείμενα. Για μια άμεση σύγκριση μεταξύ των δύο αλγορίθμων αναζήτησης του FNR-δέντρου, παρουσιάζουμε επίσης και το Σχήμα 2.32. Εκεί, φαίνεται σαφώς ότι ο μέσος όρος των προσβάσεων κόμβων του δεύτερου αλγορίθμου αναζήτησης – για 2000 αντικείμενα – παραμένει σταθερός ανεξάρτητα από την χωρική έκταση της επερώτησης, ενώ το κόστος του πρώτου αλγορίθμου αναζήτησης μεγαλώνει υπογραμμικά με τη χωρική έκταση.



Σχήμα 2.32: Επερωτήσεις χρονικής στιγμής με αύξουσα χωρική έκταση στο FNR-δέντρο με 2000 κινούμενα αντικείμενα

2.6.4. Σύνοψη των Πειραμάτων

Τα πειράματα που εκτελέσαμε για να αποτιμήσουμε την απόδοση του FNR-δένδρου έδειξαν ότι υποστηρίζει ερωτήσεις εύρους και χρονικής στιγμής πολύ πιο αποδοτικά από τους ανταγωνιστές τους. Ειδικά για τις επερωτήσεις χρονικής στιγμής, το FNR-δένδρο μόνο υπό συνθήκες υπερκαλύπτει την απόδοση του 3D R-δένδρου. Αυτό συμβαίνει όταν ο πληθυσμός των δεικτοδοτούμενων δεδομένων ξεπερνάει τις 1000 τροχιές. Επιπλέον, καθορίζουμε τις συνθήκες κάτω από τις οποίες ο δεύτερος αλγόριθμος αναζήτησης του FNR-δένδρου είναι πιο αποδοτικός από τον πρώτο. Αποδεικνύεται ότι οι ερωτήσεις χρονικής στιγμής με χωρικό εύρος μεγαλύτερο του 50% σε κάθε διάσταση υποστηρίζονται πιο αποδοτικά από τον αλγόριθμο `Search-from-Parent-1D-R-tree`. Όσον αφορά το μέγεθος του ευρετηρίου, η εκμετάλλευση του χώρου του FNR-δένδρου μπορεί να φτάσει το 92%. Το μέσο

μέγεθος ανά κινούμενο αντικείμενο είναι συγκρίσιμο μόνο με αυτό του TB*-δένδρου και μπορεί να γίνει τρεις φορές μικρότερο από το αντίστοιχο μέγεθος του 3D R-δένδρου. Τέλος, ο μέσος αριθμός προσβάσεων ανά εισαγωγή στο FNR-δένδρο είναι καλύτερος από αυτόν του TB-δένδρου και στην ίδια τάξη μεγέθους με αυτόν του απλού 3D R-δένδρου.

2.7. Συμπεράσματα

Ο τομέας της δεικτοδότησης χωροχρονικών δεδομένων υπήρξε ιδιαίτερα δραστήριος την τελευταία δεκαετία. Ενώ η μεγάλη πλειονότητα των χωροχρονικών εφαρμογών στον πραγματικό κόσμο αφορά σε αντικείμενα που παράγουν δεδομένα τροχιών, ένα μεγάλο τμήμα των αναπτυσσόμενων ευρετηρίων αγνοεί τις προκλήσεις που προκύπτουν από τη φύση αυτών των δεδομένων και απλά δεικτοδοτεί συλλογές γραμμικών τμημάτων στο χωροχρονικό χώρο, και διαχειρίζεται μόνο παραδοσιακές επερωτήσεις βασισμένες στις συντεταγμένες. Επιπλέον, επειδή πάμπολλες από αυτές τις εφαρμογές προϋποθέτουν ότι ο χώρος στον οποίο κινούνται τα αντικείμενα είναι περιορισμένος σε δίκτυα (συστήματα διαχείρισης στόλων κ.ο.κ.), τα χωροχρονικά ευρετήρια θα έπρεπε να εκμεταλλεύονται αυτή την ιδιότητα για να γίνουν πιο αποτελεσματικά όπως προτάθηκε στην [KGT99].

Το πρώτο ευρετήριο που προτάθηκε για την αποτελεσματική υποστήριξη επερωτήσεων βασισμένες στην τροχιά, το TB-δέντρο [PJT00], ήταν θεμελιωδώς διαφορετικό από άλλες μεθόδους χωροχρονικής προσπέλασης δεδομένων διότι πρότεινε την ομαδοποίηση των γραμμικών τμημάτων στα ίδια φύλλα, βάσει της τροχιάς στην οποίαν ανήκουν και ουχί της χωρικής ή χρονικής τους εγγύτητας. Όμως, το TB-δέντρο διαθέτει κάποια μειονεκτήματα το κυριότερο εκ των οποίων είναι ότι η απόδοσή του εξαρτάται από τη σειρά με την οποία εισάγονται σε αυτό τα δεδομένα. Πολλώ δε μάλλον, που ενώ οι περιορισμοί στην κίνηση έχουν αποτελέσει αντικείμενο έρευνας [KGT99], [PTKZ02], [Pfo02], [PJ01], μέχρι πρόσφατα δεν έχει υπάρξει πρόταση για χωροχρονική μέθοδο προσπέλασης κατάλληλη για αντικείμενα που κινούνται σε σταθερά δίκτυα.

Σε αυτό το κεφάλαιο, η τεχνολογία αιχμής προχωρά σε δύο ανεξάρτητες κατευθύνσεις:

- Στην πρώτη περίπτωση, για αντικείμενα που κινούνται ελεύθερα στο χώρο, αναγνωρίζοντας τα βασικά πλεονεκτήματα του TB-δέντρου, προχωράμε ένα βήμα περαιτέρω και προτείνουμε ένα πρωτότυπο ευρετήριο, που ονομάζεται TB*-δέντρο. Το προτεινόμενο ευρετήριο ξεπερνά τα κύρια μειονεκτήματα του προκατόχου του ενώ ταυτόχρονα διατηρεί όλες τις 'επιθυμητές' του ιδιότητες. Πιο συγκεκριμένα, υποστηρίζει εισαγωγές, διαγραφές και συμπίεση τροχιών ενώ η επεξεργασία των επερωτήσεων γίνεται με χρήση των αλγορίθμων που παρέχονται στην [PJT00].
- Στη δεύτερη περίπτωση, για αντικείμενα περιορισμένα να κινούνται σε δίκτυα, προτείνεται μια νέα τεχνική δεικτοδότησης, που ονομάζεται Fixed Network R-tree (FNR-δέντρο). Η γενική ιδέα που περιγράφει το FNR-δέντρο είναι ένα δάσος αρκετών 1D R-δέντρων [Gut84] πάνω από ένα 2D R-δέντρο. Το 2D R-δέντρο χρησιμοποιείται για τη δεικτοδότηση των χωρικών δεδομένων του δικτύου (δηλαδή δρόμους που αποτελούνται από γραμμικά τμήματα), ενώ τα 1D R-δέντρα χρησιμοποιούνται για τη δεικτοδότηση του χρονικού διαστήματος της κίνησης κάθε αντικειμένου σε ένα δεδομένο τμήμα του δικτύου. Επιπλέον,

τα φύλλα όλων των 1D δέντρων δεικτοδοτούνται από ένα άλλο 1D R-δέντρο που χρησιμοποιείται για να απαντήσει σε επερωτήσεις χωρίς χωρική έκταση.

Τα πειράματα που διεξήχθησαν για την αξιολόγηση της απόδοσης του προτεινόμενου TB*-δέντρου σε σχέση με το αρχικό TB-δέντρο και το 3D R-δέντρο κατέδειξαν ότι το προτεινόμενο ευρετήριο υποστηρίζει επερωτήσεις εύρους και συνδυαστικές αποτελεσματικά. Η υπεροχή του προτεινόμενου ευρετηρίου εδραιώνεται στην γενική περίπτωση που τα ευρετήρια δημιουργούνται εισάγοντας τα δεδομένα τροχιάς με μη χρονολογική σειρά (οργάνωση 'ταυτότητας / χρόνου'), μία περίπτωση αναμενόμενη σε πραγματικές εφαρμογές και όταν το ευρετήριο δημιουργείται μετά από συμπίεση ενός συνόλου δεδομένων. Το TB*-δέντρο είναι πιο συμπαγές από τον προκάτοχό του, υποστηρίζει υψηλούς ρυθμούς εισαγωγής, συμπεριφέρεται καλά σε μη χρονολογικές εισαγωγές δεδομένων τροχιών που εμφανίζονται σε πραγματικό περιβάλλον και υποστηρίζει διαγραφές και συμπίεση τροχιών αποτελεσματικά.

Επίσης συγκρίναμε πειραματικά το FNR-δέντρο με το TB*-δέντρο, το παραδοσιακό 3D R-δέντρο [TVS96] και το TB-δέντρο [PJT00]. Για διάφορα σύνολα δεδομένων και επερωτήσεις εύρους, το FNR-δέντρο αποδείχθηκε ότι υπερέχει όλων των ανταγωνιστών του στην αθρόα πλειονότητα των περιπτώσεων. Το FNR-δέντρο έχει υψηλή χρησιμοποίηση χώρου, μικρότερο μέγεθος ανά κινούμενο αντικείμενο και υποστηρίζει επερωτήσεις εύρους πολύ πιο αποτελεσματικά. Σε γενικές γραμμές, θεωρούμε ότι το FNR-δέντρο είναι ιδανική μέθοδος προσπέλασης για εφαρμογές διαχείρισης στόλου. Ωστόσο, το FNR-δέντρο μπορεί να χρησιμοποιηθεί μόνο κάτω από το σενάριο κίνησης περιορισμένης σε δίκτυο.

3. Προηγμένη Επεξεργασία Επερωτήσεων Τροχιών: Αναζήτηση Πλησιέστερου Γείτονα

Στο κεφάλαιο αυτό δίνουμε μία σειρά από αλγορίθμους για την πραγματοποίηση αναζητήσεων πλησιέστερου γείτονα σε τροχιές κινούμενων αντικειμένων, χρησιμοποιώντας δομές τύπου R-δέντρου που αποθηκεύουν δεδομένα ιστορικών τροχιών. Το κεφάλαιο οργανώνεται ως εξής. Η Ενότητα 3.1 είναι μια εισαγωγή στο βασικό θέμα αυτού του κεφαλαίου. Οι σχετικές εργασίες παρουσιάζονται στην Ενότητα 3.2, ενώ η Ενότητα 3.2 εισάγει, σε περιληπτικό επίπεδο, μία σειρά *k*-NN αλγορίθμων σε τροχιές κινούμενων αντικειμένων, όπως και τις μετρικές που υποστηρίζουν τις στρατηγικές διάταξης και κλαδέματος κατά την αναζήτηση στο δέντρο. Οι Ενότητες 3.4 και 3.5 αποτελούν τον πυρήνα του κεφαλαίου περιγράφοντας με λεπτομέρειες τους αλγορίθμους επεξεργασίας επερωτήσεων για την εκτέλεση αναζητήσεων πλησιέστερου γείτονα σε δεδομένα ιστορικών τροχιών μαζί με τα συνεχή αντίστοιχά τους· οι αλγόριθμοι που παρουσιάζονται βασίζονται στο παράδειγμα «πρώτα στο βαθύτερο» (depth-first) και «πρώτα στο καλύτερο» (best-first), χρησιμοποιώντας δομές που ομοιάζουν στο R-δέντρο. Η Ενότητα 3.6 παρουσιάζει τα αποτελέσματα της πειραματικής μας μελέτης και η Ενότητα 3.6.5 παρουσιάζει τα συμπεράσματά μας.

3.1. Εισαγωγή

Η έρευνα στον τομέα της προηγμένης επεξεργασίας επερωτήσεων σε βάσεις δεδομένων ιστορικών χωροχρονικών τροχιών κατευθύνεται από σχετικές εργασίες στο τομέα των (σταθερών) χωρικών βάσεων δεδομένων. Για παράδειγμα, επερωτήσεις της μορφής «*Βρες όλα τα αντικείμενα που βρίσκονται σε μία δεδομένη περιοχή στη διάρκεια ενός συγκεκριμένου χρονικού διαστήματος*» γενικεύουν την αντίστοιχη χωρική επερώτηση εύρους της μορφής «*Βρες όλα τα αντικείμενα σε μία δεδομένη περιοχή*». Αυτές οι επερωτήσεις θεωρούνται βασικές, δεδομένου ότι τα προτεινόμενα ευρετήρια θα πρέπει εξ' ορισμού να τις υποστηρίζουν. Από την άλλη πλευρά, άλλοι χωρικοί τελεστές, όπως ο τελεστής *πλησιέστερου γείτονα* [RKV95] *(nearest neighbor – NN)* και ο τελεστής *χωρικής σύνδεσης βασισμένης στην απόσταση (distance join)* [HS99], θεωρούνται προηγμένοι, διότι απαιτούν πιο εξελιγμένες τεχνικές επεξεργασίας για την αποτελεσματική τους εκτέλεση. Επιπλέον, οι προηγμένες τεχνικές που απαντούν επερωτήσεις αυτής της μορφής μπορεί να λάβουν υπ' όψιν τους την παρουσία ενός χωροχρονικού ευρετηρίου αλλά μπορεί και όχι. Απ' την άλλη πάλι, στις MODs συνήθως έχουμε ν' αντιμετωπίσουμε τεράστιο όγκο ιστορικών δεδομένων από μεγάλο αριθμό κινούμενων αντικειμένων. Δεδομένου, ότι μία επεκτάσιμη DBMS που υποστηρίζει κινούμενα

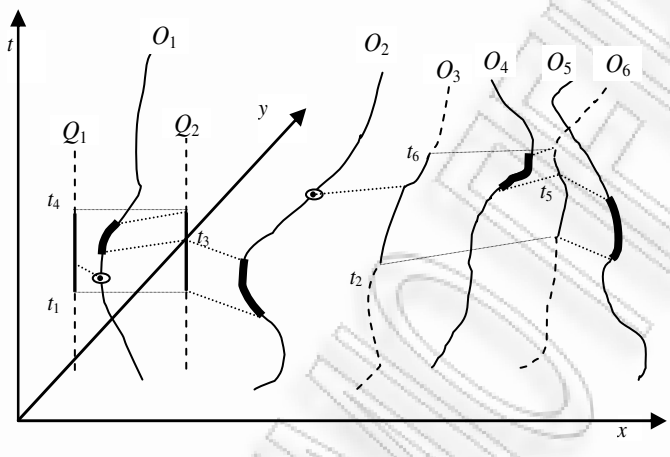
αντικείμενα, πρέπει να παρουσιάζει ικανοποιητική συμπεριφορά στην εκτέλεση χωροχρονικών επερωτήσεων, στη συνέχεια, περιορίζουμε τη συζήτησή μας σε τεχνικές προηγμένης επεξεργασίας επερωτήσεων υπό την προοπτική της πρώτης υπόθεσης, δηλαδή, θεωρώντας ότι λειτουργεί μιας κάποιας μορφής χωροχρονικό ευρετήριο.

Μια σημαντική κατηγορία επερωτήσεων που είναι σαφώς χρήσιμες στην επεξεργασία των MODs είναι οι λεγόμενες επερωτήσεις του *k*-οστού πλησιέστερου γείτονα (*k* nearest neighbor - *k*-NN), όπου μας ενδιαφέρει η εξεύρεση των *k* πλησιέστερων τροχιών σε ένα προκαθορισμένο αντικείμενο επερώτησης *Q*. Εξ' όσων γνωρίζουμε, στη βιβλιογραφία εξετάζονται επερωτήσεις που ασχολούνται κυρίως είτε με σταθερά ([RKV95], [CF98], [HS99]) είτε με συνεχώς κινούμενα σημεία επερώτησης ([SR01], [TPS02]) σε σταθερά σύνολα δεδομένων, ή επερωτήσεις σχετικά με τις μελλοντικές ή τρέχουσες θέσεις συνεχώς κινούμενων σημείων ([BJKS02], [TP02], [ISS03], [YPK05], [XMA05], [MHP05]). Προφανώς, οι επερωτήσεις αυτής της μορφής δεν καλύπτουν την αναζήτηση NN σε ιστορικές τροχιές.

Έτσι, μία από τις προκλήσεις που έχουμε δεχτεί σε αυτή τη διατριβή είναι η περιγραφή διαφόρων μηχανισμών για την εκτέλεση αναζήτησης *k*-NN σε MODs που τα ιστορικά δεδομένα είναι αποθηκευμένα σε κατάλληλα χωροχρονικά ευρετήρια. Για να κατανοήσουμε το πρόβλημα, έστω μια εφαρμογή που καταγράφει τη θέση σπανίων ειδών άγριων ζώων. Μια τέτοια εφαρμογή αποτελείται από μία MOD που αποθηκεύει τα δεδομένα που αφορούν στη θέση, καθώς κι ένα χωροχρονικό ευρετήριο για την αναζήτηση και την απάντηση επερωτήσεων *k*-NN με πιο αποτελεσματικό τρόπο. Οι εμπειρογνώμονες στον τομέα θα είχαν σίγουρα κάτι να αποκομίσουν αν μπορούσαν να θέσουν μια επερώτηση της μορφής «*Βρες τις πλησιέστερες τροχιές ενός ζώου σε κάποιο σταθερό σημείο (εργαστήριο, πηγή τροφής ή άλλα μη μεταναστευτικά είδη) από το οποίο πέρασε το εν λόγω είδος κατά τη διάρκεια του μηνός Μαρτίου*». Έστω τώρα ότι επιθυμία του εμπειρογνώμονα είναι να θέσει την ίδια επερώτηση με τη διαφορά ότι το αντικείμενο *Q* της επερώτησης δεν είναι ένα σταθερό σημείο αλλά ένα κινούμενο ζώο που κινείται από τη θέση P_1 στη P_2 στην διάρκεια ενός χρονικού διαστήματος. Η επερώτηση αυτή μας δίνει το έναυσμα να συνάγουμε μια πιο γενική επερώτηση όπου ο ειδικός μπορεί να θελήσει να θέσει μια άλλη τροχιά του ιδίου ή σχετικού είδους σαν αντικείμενο επερώτησης *Q*. Αυτονόητο είναι ότι μέσω αυτών των τύπων επερωτήσεων ο ειδικός μπορεί να συνάγει συνήθειες και πρότυπα κίνησης άγριων ζώων ή παρεκκλίσεις από τη φυσική μετανάστευση, που θα μπορούσαν να είναι αλληλένδετες με περιβαλλοντικές ή / και οικολογικές αλλαγές ή καταστροφές. Δεδομένου δε ότι οι χρήστες των MOD συνήθως ενδιαφέρονται για συνεχείς τύπους επερωτήσεων, οι δύο επερωτήσεις που συζητήσαμε προηγουμένως επεκτείνονται στις συνεχείς αντίστοιχές τους. Στη συνεχή τους παραλλαγή, κάθε επερώτηση μας δίνει μία χρονικά μεταβαλλόμενη τιμή (που υποδηλώνει την πλησιέστερη απόσταση, η οποία εξαρτάται από το χρόνο) καθώς και μία σειρά ταυτοτήτων τροχιών και τα κατάλληλα χρονικά διαστήματα για τα οποία ισχύει κάθε κινούμενο αντικείμενο $\{O_1[t_1, t_2), O_2[t_2, t_3), \ldots\}$.

Για να θέσουμε το πρόβλημα σε ένα πιο ανθρωποκεντρικό πλαίσιο, έστω μία εφαρμογή που αναλύει τις δυναμικές αστικών και περιφερειακών συστημάτων. Πρόθεσή μας εδώ είναι να συνδράμουμε στην ανάπτυξη χωροχρονικών συστημάτων υποστήριξης αποφάσεων (spatio-temporal decision support systems - STDSS) που χρησιμοποιούν επαγγελματίες στον τομέα του σχεδιασμού.

Μια τέτοια περίπτωση απαιτεί παρόμοιες μεθοδολογίες για την κατανόηση, στο χώρο και στο χρόνο, των αλληλεπιδράσεων της πορείας της ζωής των ατόμων. Η πορεία της ζωής των περισσοτέρων ανθρώπων δημιουργείται γύρω από δύο αλληλένδετα διαδοχικά γεγονότα: μία τροχιά διαμονής και μία επαγγελματική καριέρα. Αυτά τα πρότυπα γεγονότων έγιναν πιο περίπλοκα τις τελευταίες δεκαετίες, δημιουργώντας νέες προκλήσεις για όσους ασχολούνται με τον αστικό και περιφερειακό σχεδιασμό. Πιστεύουμε ότι ένας ειδικός μπορεί να εκμεταλλευτεί τα χαρακτηριστικά των αλγορίθμων επεξεργασίας επερωτήσεων πλησιέστερου γείτονα και να τα χρησιμοποιήσει για την ανάλυση της πορείας της ζωής των ανθρώπων.



Σχήμα 3.1: Επερωτήσεις NN σε τροχιές κινούμενων αντικειμένων

Για να γίνουν πιο κατανοητά τα προηγούμενα παραδείγματα, έστω το Σχήμα 3.1 που αναπαριστά τις τροχιές έξι κινούμενων ζώων των {O_1, O_2, O_3, O_4, O_5, O_6} καθώς και δύο στατικών σημείων (Q_1 και Q_2) που αντιπροσωπεύουν δύο πηγές τροφής. Τώρα, έστω οι ακόλουθες επερωτήσεις που παρουσιάζονται και στο Σχήμα 3.1 (Οι επερωτήσεις 2 και 4 είναι οι συνεχείς αντίστοιχες των Επερωτήσεων 1 και 3, αντίστοιχα):

- Επερώτηση 1. «*Βρες ποιο ζώο ήταν πλησιέστερο στην σταθερή πηγή τροφής Q_1 κατά το χρονικό διάστημα $[t_1,t_4]$*», με αποτέλεσμα το ζώο O_1.
- Επερώτηση 2. «*Βρες ποιο ζώο ήταν πλησιέστερο στην σταθερή πηγή τροφής Q_2 σε οποιαδήποτε χρονική στιγμή του χρονικού διαστήματος $[t_1,t_4]$*», με αποτέλεσμα μια λίστα αντικειμένων: το O_2 για το διάστημα $[t_1,t_3)$ και το O_1 για το διάστημα $[t_3,t_4]$.
- Επερώτηση 3. «*Βρες ποιο ζώο ήταν πλησιέστερο στο ζώο O_3 κατά το χρονικό διάστημα $[t_2,t_6]$*», με αποτέλεσμα το O_2.
- Επερώτηση 4. «*Βρες ποιο ζώο ήταν πλησιέστερο στο ζώο O_6 σε οποιαδήποτε χρονική στιγμή του χρονικού διαστήματος $[t_2,t_6]$*», με αποτέλεσμα μια λίστα αντικειμένων: το O_5 για το διάστημα $[t_2,t_5)$ και το O_4 για το διάστημα $[t_5,t_6]$.

Σε αντίθεση με τις παραδοσιακές βάσεις δεδομένων, οι MODs έχουν το εξής χαρακτηριστικό: αρκετές χωροχρονικές επερωτήσεις είναι εκ φύσεως συνεχείς. Σε αντίθεση με τις στιγμιαίες (snapshot) επερωτήσεις, τις οποίες καλούμε μόνο άπαξ, οι συνεχείς επερωτήσεις απαιτούν συνεχή αξιολόγηση καθώς το αποτέλεσμα της επερώτησης είναι άκυρο μετά από σύντομο χρονικό διάστημα. Αν θέσουμε την προηγούμενη συζήτηση υπό την προοπτική των ιστορικών τροχιών, παρόλο που οι επερωτήσεις 2 και 4 είναι συνεχείς εκ φύσεως (*σε οποιαδήποτε χρονική στιγμή*) δεν μπορούν να χαρακτηρισθούν

αμιγείς συνεχείς επερωτήσεις· σε σχέση με τη μηχανή της βάσης δεδομένων, μία συνεχής επερώτηση είναι αυτή που υποβάλλεται άπαξ στην βάση δεδομένων και παραμένει ενεργή, ενημερώνοντας συνεχώς το αποτέλεσμα της επερώτησης με την πάροδο του χρόνου, μέχρι να δηλωθεί ότι ολοκληρώθηκε είτε μέσω μηνύματος του χρήστη είτε μέσω της προκαθορισμένης διάρκειας ζωής της επερώτησης [BW01], [HXL05], [MXA04]. Υπ' αυτήν την έννοια, οι επερωτήσεις 2 και 4 είναι στιγμιαίες (snapshot) επερωτήσεις. Ωστόσο, για να τις διαφοροποιήσουμε από τις επερωτήσεις 1 και 3 καθώς επίσης και από τις αμιγείς συνεχείς επερωτήσεις, από εδώ και στο εξής θα ονομάζονται *Ιστορικές Επερωτήσεις Συνεχούς Πλησιέστερου Γείτονα* (*Historical Continuous NN queries - HCNN).*

Για να συνοψίσουμε την προηγούμενη συζήτηση, τα κύρια στοιχεία στον παρόν κεφάλαιο είναι τα εξής:

- Προτείνουμε νέες μετρικές για την υποστήριξη των στρατηγικών διάταξης και κλαδέματος της αναζήτησης. Πιο συγκεκριμένα, ο ορισμός της μετρικής ελάχιστης απόστασης *MINDIST* μεταξύ σημείων και ορθογωνίων παραλληλογράμμων, ο οποίος προτάθηκε αρχικώς στην [RKV95] και επεκτάθηκε στην [TPS02], επεκτείνεται περαιτέρω προκειμένου οι αλγόριθμοί μας να υπολογίζουν την ελάχιστη απόσταση μεταξύ τροχιών και ορθογωνίων παραλληλογράμμων αποτελεσματικά.
- Προτείνουμε αλγορίθμους επεξεργασίας επερωτήσεων NN σε χωροχρονικά ευρετήρια που αποθηκεύουν ιστορικά δεδομένα κινούμενων αντικειμένων. Εκμεταλλευόμαστε τα πιο συνηθισμένα χωροχρονικά ευρετήρια, που υποστηρίζουν απεριόριστη κίνηση, ήτοι δομές παρόμοιες του R-δέντρου όπως το 3D R-δέντρο [TVS96], το TB-δέντρο [PJT00] και το TB*-δέντρο που προτείνεται σε αυτή τη διατριβή. Η περιγραφή των αλγορίθμων μας εξαρτάται από τον τύπου του αντικειμένου της επερώτησης (σημείο ή τροχιά) καθώς επίσης και από το κατά πόσο η ίδια η επερώτηση είναι ιστορικά συνεχής ή όχι. Πιο συγκεκριμένα, παρουσιάζουμε αποτελεσματικούς αλγορίθμους «πρώτα στο βαθύτερο» και «πρώτα στο καλύτερο» (αυξητικούς) για ιστορικές επερωτήσεις NN καθώς και αλγορίθμους «πρώτα στο βαθύτερο» για τις συνεχείς αντίστοιχές τους. Όλοι οι προτεινόμενοι αλγόριθμοι γενικεύονται για την εύρεση των k πλησιέστερων γειτόνων.
- Διεξάγουμε μία πλήρη σειρά πειραμάτων σε μεγάλα συνθετικά και πραγματικά σύνολα δεδομένων αποδεικνύοντας ότι οι αλγόριθμοι έχουν υψηλή δυνατότητα κλιμάκωσης και αποτελεσματικότητα σε όρους κόμβων που προσπελαύνονται, χρόνου εκτέλεσης και χώρου που κλαδεύεται.

Πρέπει να επισημάνουμε εδώ ότι οι προτεινόμενοι αλγόριθμοι δεν απαιτούν μια συγκεκριμένη δομή ευρετηρίου και μπορούν να εφαρμοσθούν απευθείας σε οποιοδήποτε μέλος της οικογενείας των R-δέντρων που χρησιμοποιείται για τη δεικτοδότηση τροχιών σαν αυτές που παρουσιάζονται στο προηγούμενο κεφάλαιο.

3.2. Σχετικές Εργασίες

Κατά την τελευταία δεκαετία, οι επερωτήσεις NN έχουν προμηθεύσει την κοινότητα των χωροχρονικών βάσεων δεδομένων με μια σειρά από ενδιαφέροντα αξιοσημείωτα ερευνητικά ζητήματα. Υπάρχει ήδη πληθώρα μεθόδων για την αποτελεσματική επεξεργασία των επερωτήσεων

NN για σταθερά σημεία επερώτησης, με πιο σημαντικό κατά πάσα πιθανότητα τον αλγόριθμο *επίσκεψης κλαδιού και περιορισμού* (branch-and-bound) του R-δέντρου που πρότειναν οι Roussopoulos et al. [RKV95] για την εύρεση του πλησιέστερου γείτονα ενός σταθερού σημείου. Ο αλγόριθμος χρησιμοποιεί δύο μετρικές, την *MINDIST* και την *MINMAXDIST*, για να κλαδέψει και να διατάξει τους κόμβους του δένδρου. Πιο συγκεκριμένα, ξεκινώντας από τη ρίζα του δέντρου, ο αλγόριθμος εντοπίζει την εγγραφή με την ελάχιστη απόσταση από το σημείο της επερώτησης (με χρήση των παραπάνω μετρικών). Η επεξεργασία επαναλαμβάνεται συνεχώς μέχρις ότου να φτάσουμε στο επίπεδο των φύλλων, στο οποίο βρίσκεται ο πρώτος υποψήφιος πλησιέστερος γείτονας. Επιστρέφοντας από την αναδρομική αυτή διαδικασία, εξετάζονται μόνον οι εγγραφές με ελάχιστη απόσταση μικρότερη της απόστασης του πλησιέστερου γείτονα που έχει ήδη εντοπισθεί. Η παραπάνω διαδικασία γενικεύθηκε για να υποστηρίξει επερωτήσεις *k*-NN. Αργότερα, οι Cheung και Fu [CF98] απέδειξαν ότι, δεδομένης της διάταξης βάσει *MINDIST*, το κλάδεμα που λαμβάνεται από την [RKV95] μπορεί να διατηρηθεί χωρίς τη χρήση μετρικής *MINMAXDIST* (ο υπολογισμός της οποίας είναι δαπανηρός υπολογιστικά).

Οι Hjaltason και Samet [HS99] παρουσίασαν ένα γενικό αυξητικό αλγόριθμο NN, που χρησιμοποιεί έναν «πρώτα στον καλύτερο» αλγόριθμο διάσχισης της δομής του R-δέντρου. Κατά τη διαδικασία απόφασης ποιου κόμβου του δέντρου θα επισκεφτεί στη συνέχεια, ο προτεινόμενος αλγόριθμος επιλέγει τον κόμβο με την ελάχιστη απόσταση επί του συνόλου των κόμβων που πρέπει ακόμα να εξεταστούν. Για να το επιτύχουμε, ο αλγόριθμος χρησιμοποιεί μία ουρά (queue) προτεραιότητας όπου οι κόμβοι του δέντρου αποθηκεύονται με αύξουσα σειρά ανάλογα με την απόσταση από το αντικείμενο της επερώτησης. Ο αλγόριθμος «πρώτα στον καλύτερο» υπερτερεί του αλγορίθμου των Roussopoulos et al. σε όρους κλαδεμένου χώρου. Επιπλέον, αφ' ης στιγμής βρεθεί ο *k*-οστός NN, ο *k*+1-οστός NN μπορεί να ανακτηθεί χωρίς μεγάλο επιπλέον κόστος, δεδομένου ότι ο αλγόριθμος είναι αυξητικός. Το βασικό μειονέκτημα του αλγόριθμου «πρώτα στον καλύτερο» είναι ότι η απόδοσή του εξαρτάται από το μέγεθος της ουράς προτεραιότητας. Σε περίπτωση που η ουρά προτεραιότητας γίνει πολύ μεγάλη, ο χρόνος εκτέλεσης του αλγορίθμου αυξάνεται πολύ γρήγορα.

Ο πρώτος αλγόριθμος για την αναζήτηση του *k* NN σε ένα κινούμενο σημείο επερώτησης προτάθηκε στην [SR01]. Ο αλγόριθμος υποθέτει ότι τα σημεία ενδιαφέροντος είναι σταθερά και οι θέσεις τους (γνωστές εκ των προτέρων) αποθηκεύονται σε μία δομή τύπου R-δέντρου. Υποθέτουμε ότι υπάρχει μία διακριτή χρονική διάσταση, συνεπώς εφαρμόζεται μία τεχνική περιοδικής δειγματοληψίας στο ίχνος του κινούμενου σημείου επερώτησης. Η θέση του σημείου επερώτησης που βρίσκεται μεταξύ δύο διαδοχικών δειγματοληπτούμενων θέσεων εκτιμάται χρησιμοποιώντας γραμμικές ή πολυωνυμικές splines. Το να εκτελέσουμε μια επερώτηση NN για κάθε δειγματοληπτούμενο σημείο του ίχνους της επερώτησης δεν είναι καθόλου αποτελεσματικό, έτσι ο προτεινόμενος αλγόριθμος υιοθετεί μία προοδευτική προσέγγιση, που βασίζεται στη παρατήρηση ότι όταν δύο σημεία επερώτησης είναι κοντά, τα αποτελέσματα της αναζήτησης *k*-NN σε αυτές τις θέσεις θα πρέπει να είναι ανάλογα. Συνεπώς, όταν υπολογίζουμε τα αποτελέσματα για μία δειγματοληπτούμενη θέση, ο αλγόριθμος προσπαθεί να εκμεταλλευτεί τις πληροφορίες που παρέχονται από τα αποτελέσματα των προηγούμενων δειγμάτων. Το βασικό μειονέκτημα αυτής της προσέγγισης είναι ότι η ακρίβεια των αποτελεσμάτων εξαρτάται από το ρυθμό δειγματοληψίας. Επιπλέον, υπάρχει σημαντική υπολογιστική επιβάρυνση.

Μία τεχνική που αποφεύγει τα μειονεκτήματα της δειγματοληψίας βασίζεται σε χρονικά παραμετροποιημένες (TP) επερωτήσεις [TP02]. Οι TP επερωτήσεις ανακαλούν το τρέχον αποτέλεσμα όταν δίνεται η επερώτηση, την περιόδου για την οποία ισχύει το αποτέλεσμα και το σύνολο των αντικειμένων που προκαλούν την εκπνοή του αποτελέσματος. Δεδομένου του τρέχοντος αποτελέσματος και του συνόλου των αντικειμένων που επηρεάζουν την ισχύ του, το επόμενο αποτέλεσμα μπορεί να υπολογιστεί αυξητικά. Η σημασία των TP επερωτήσεων είναι διττή: i) ως ανεξάρτητες μέθοδοι, είναι κατάλληλες για εφαρμογές που περιλαμβάνουν δυναμικά περιβάλλοντα, όπου κάθε αποτέλεσμα είναι έγκυρο για μία συγκεκριμένη χρονική περίοδο και ii) βρίσκονται στον πυρήνα πολύ πιο πολύπλοκων μηχανισμών, όπως οι Συνεχείς επερωτήσεις NN (CNN). Το κύριο μειονέκτημα της χρήσης επερωτήσεων TP για την επεξεργασία μιας επερώτησης CNN είναι ότι πρέπει να εκτελεστούν αρκετές επερωτήσεις NN. Έτσι, το κόστος της μεθόδου είναι απαγορευτικό για μεγάλα σύνολα δεδομένων.

Χρησιμοποιώντας τη δομή του TPR-δέντρου (Χρονικά Παραμετροποιημένο/Time Parameterized R-Δέντρο) [SJLL00], οι Benetis et al. [BJKS02] παρουσίασαν αποτελεσματικές λύσεις για επερωτήσεις NN και RNN (Αντίστροφου Πλησιέστερου Γείτονα/Reverse Nearest Neighbor). (Μια επερώτηση RNN μας δίνει όλα τα αντικείμενα των οποίων το αντικείμενο της επερώτησης είναι ο πλησιέστερος γείτονας). Ο προτεινόμενος αλγόριθμος αρχικά δημιουργήθηκε για να απαντήσει σε συνεχείς επερωτήσεις RNN, διότι οι παλιοί αλγόριθμοι RNN δημιουργήθηκαν με την υπόθεση ότι το σημείο επερώτησης είναι σταθερό. Οι αλγόριθμοι τόσο για τις επερωτήσεις NN όσο και για τις RNN στην [BJKS02] αναφέρονται σε μελλοντικές (εκτιμώμενες) θέσεις της επερώτησης και σε σημεία δεδομένων, που υποθέτουμε ότι κινούνται συνεχώς στο επίπεδο. Στην ίδια εργασία, προτείνεται κι ένας αλγόριθμος που απαντά σε επερωτήσεις CNN.

Οι Tao et al. [TPS02] μελέτησαν επίσης τις επερωτήσεις CNN και πρότειναν έναν αλγόριθμο βασισμένου στο R-δέντρο (για κινούμενα σημεία επερώτησης και σταθερά σημεία δεδομένων) που αποφεύγει τις κακοτοπιές των προηγούμενων (αποτυχία ανάκτησης όλων των απαντήσεων και υψηλό κόστος επεξεργασίας). Η προτεινόμενη ευριστική μέθοδος κλαδέματος του δέντρου εκμεταλλεύεται τη μετρική *MINDIST* που παρουσιάζεται στην [RKV95]. Σε κάθε εγγραφή των φύλλων, ο αλγόριθμος εστιάζει στον ακριβή υπολογισμό των σημείων διαίρεσης, δηλαδή, των σημείων του τμήματος της επερώτησης που υποδεικνύουν ότι υπάρχει αλλαγή γείτονα. Παρουσιάστηκε μία θεωρητική ανάλυση της βέλτιστης απόδοσης για αλγορίθμους CNN και προτάθηκαν μοντέλα κόστους για την εκτίμηση του αριθμού των κόμβων που προσπελαύνονται. Επιπλέον, ο αλγόριθμος CNN επεκτάθηκε για την περίπτωση των *k* γειτόνων.

Βάσει των επερωτήσεων TP που προτάθηκαν στην [TP02], οι Iwerks et al. [ISS03] περιέγραψαν μια τεχνική που εστιάζει στη διατήρηση των επερωτήσεων CNN (για μελλοντικές προβλεπόμενες θέσεις) παρουσία ενημερώσεων για τα κινούμενα τμήματα, όπου η κίνηση των σημείων αναπαρίσταται ως συνάρτηση του χρόνου. Παρουσιάστηκε επίσης μια νέα προσέγγιση που φιλτράρει τον αριθμό των αντικειμένων που πρέπει να ληφθούν υπ' όψιν όταν διατηρείται μία μελλοντική επερώτηση CNN.

Προσφάτως, στον ίδιο τομέα, οι Xiong et al. [XMA05], πρότειναν μία μέθοδο για κλιμακούμενη επεξεργασία επερωτήσεων CNN σε χωροχρονικές βάσεις δεδομένων. Προτείνουν ένα γενικό πλαίσιο για την επεξεργασία μεγάλων αριθμών ταυτόχρονων επερωτήσεων *k*-CNN με σταθερές ή κινούμενες

επερωτήσεις σε σταθερά ή (επί του παρόντος) κινούμενα σύνολα δεδομένων χωρίς να γίνεται καμία υπόθεση ως προς τις τροχιές του αντικειμένου. Αντίθετα με άλλες προτάσεις, η λύση τους για να υποστηρίξει υψηλούς ρυθμούς ενημέρωσης δεν βασίζεται στο R-δέντρο αλλά σε μία απλή δομή πλέγματος (grid) που διατηρείται στο δίσκο. Μία παρόμοια μέθοδος προτάθηκε επίσης από τους Yu et al. [YPK05] για την παρακολούθηση των επερωτήσεων k-CNN σε (επί του παρόντος) κινούμενα αντικείμενα χωρίς να γίνεται καμία υπόθεση για τις τροχιές των αντικειμένων. Η μέθοδος χρησιμοποιεί επίσης ευρετήρια πλέγματος (κύριας μνήμης) που δεικτοδοτούν κινούμενα αντικείμενα και επερωτήσεις, και αποδεικνύεται ότι υπερέχει των λύσεων που βασίζονται σε R-δέντρα. Οι Mouratidis et al. [MHP05] επίσης αποδεσμευόμενοι από την υπόθεση ότι οι τροχιές κινούμενων αντικειμένων είναι πλήρως προβλέψιμες από τις παραμέτρους κίνησής τους, προτείνουν μία συνολική τεχνική για την αποτελεσματική παρακολούθηση συνεχών επερωτήσεων NN. Η προτεινόμενη μέθοδος, που καλείται *μέθοδος εννοιολογική κατάτμησης και παρακολούθησης* (conceptual partitioning monitoring method - CPM), χρησιμοποιεί επίσης μια δομή πλέγματος και επιτυγχάνει χαμηλό χρόνο εκτέλεσης διότι χειρίζεται τις ενημερώσεις των θέσεων μόνο των κινουμένων αντικειμένων που πέφτουν κοντά σε κάποια επερώτηση. Τα πειραματικά αποτελέσματα που παρουσιάζονται στην [MHP05] δείχνουν ότι η μέθοδος CPM υπερτερεί των τεχνικών που παρουσιάζονται στις [XMA05] και [YPK05].

Οι Shahabi et al. [SKS03] παρουσίασαν τον πρώτο αλγόριθμο για την επεξεργασία των επερωτήσεων *k*-NN για κινούμενα αντικείμενα σε οδικά δίκτυα. Ο προτεινόμενος αλγόριθμος, που χρησιμοποιεί την απόσταση του δικτύου μεταξύ δύο θέσεων αντί της Ευκλείδειας, βασίζεται στην μετατροπή του δικτύου σε χώρο υψηλότερης διάστασης, στον οποίο μπορούν να εφαρμοσθούν απλούστερες συναρτήσεις απόστασης. Χρησιμοποιώντας αυτό το χώρο, προτείνονται αποτελεσματικές τεχνικές για την εύρεση της συντομότερης διαδρομής μεταξύ δύο σημείων στο οδικό δίκτυο. Η ανωτέρω διαδικασία, που χρησιμοποιείται για την περίπτωση σταθερών σημείων, τροποποιείται ελαφρώς προκειμένου να υποστηρίξει την περίπτωση κινούμενων σημείων επερωτήσεων.

Αναγνωρίζοντας τα πλεονεκτήματα των παραπάνω θεμελιωδών τεχνικών στην παρούσα διατριβή παρουσιάζεται η πρώτη πλήρης αντιμετώπιση ιστορικών επερωτήσεων NN σε τροχιές κινούμενων αντικειμένων που δεικτοδοτούνται από χωροχρονικά ευρετήρια.

Πίνακας 3.1: Πίνακας συμβόλων

Σύμβολο	*Περιγραφή*
D	βάση δεδομένων τροχιών
O_i	η ταυτότητα ενός κινούμενου αντικειμένου
T	μία δεικτοδοτούμενη τροχιά
$T_{i,k}$	το k-οστό ευθύγραμμο τμήμα του T_i
$x_{i,k}$, $y_{i,k}$, $t_{i,k}$	οι συντεταγμένες της τροχιάς T_i κατά το χρονικό αποτύπωμα t_k
Q_p, Q_T, Q_{per}	ένα σημείο επερώτησης, μία τροχιά επερώτησης και μία περίοδος επερώτησης$[t_{start}, t_{end}]$

3.3. Διατύπωση Προβλήματος και Μετρικές για Αναζήτηση Πλησιέστερου Γείτονα

Καθορίζουμε πρώτα τα ερωτήματα NN που θα εξετάσουμε σε αυτή τη διατριβή. Κατόπιν, παρουσιάζονται οι μετρικές που απαιτούνται για τη διατύπωση της στρατηγικής διάταξης και

κλαδέματος της αναζήτησης. Τέλος, παρατίθεται μία αναλυτική μέθοδος για τη διατύπωση της συνάρτησης της απόστασης ως προς το χρόνο μεταξύ δύο αντικειμένων που κινούνται ταυτόχρονα με σταθερή ταχύτητα και κατεύθυνση (ήτοι μεταξύ δύο διαδοχικών τυχαίων σημείων), καθώς και η ελάχιστη τιμή της· τα αποτελέσματα και των δύο αναλύσεων μας είναι απαραίτητα για τους αλγορίθμους που παρουσιάζονται στην επόμενη ενότητα. Ο Πίνακας 3.1 παρουσιάζει τα σύμβολα που χρησιμοποιούνται στο υπόλοιπο του κεφαλαίου.

3.3.1. Διατύπωση Προβλήματος

Έστω D μία βάση δεδομένων N κινούμενων αντικειμένων με ταυτότητες (ids) αντικειμένων $\{O_1, O_2, \ldots, O_N\}$ και των αντίστοιχων τροχιών $\{T_1, T_2, \ldots, T_N\}$. Έχουμε ήδη δηλώσει ότι οι επερωτήσεις NN αναζητούν την πλησιέστερη τροχιά προς ένα αντικείμενο επερώτησης Q. Στην περίπτωση μας, διακρίνουμε δύο τύπους αντικειμένων επερώτησης: το Q_p, ένα σημείο (x, y) που παραμένει σταθερό κατά τη χρονική περίοδο της επερώτησης $Q_{per}[t_{start}, t_{end}]$, και το Q_T, ένα κινούμενο αντικείμενο με τροχιά T. Λαμβάνοντας υπόψη και τα παραπάνω, καθορίζουμε τους ακόλουθους δύο τύπους επερωτήσεων NN:

- *Η επερώτηση NN_Q_p* (D, Q_p, Q_{per}) αναζητά στη βάση δεδομένων D τον NN ενός σημείου Q_p που παραμένει σταθερό κατά τη χρονική περίοδο Q_{per}, και μας δίνει το σημείο p_c, το πλησιέστερο σημείο στο Q_p από το οποίο πέρασε ένα κινούμενο αντικείμενο O_i κατά τη χρονική περίοδο Q_{per}, καθώς και την αντίστοιχη ελάχιστη απόσταση.
- *Η επερώτηση NN_Q_T* (D, Q_T, Q_{per}) είναι παρόμοια με την προηγούμενη με τη μόνη διαφορά ότι το αντικείμενο της επερώτησης στη δεδομένη περίπτωση είναι ένα κινούμενο αντικείμενο με τροχιά Q_T.

Η επέκταση των παραπάνω επερωτήσεων στις ιστορικές συνεχείς αντίστοιχές τους ποικίλει ως προς τα αποτελέσματα των αλγορίθμων. Στη συνεχή περίπτωση, κάθε επερώτηση μας δίνει έναν χρονικά μεταβαλλόμενο πραγματικό αριθμό, καθώς η πλησιέστερη απόσταση εξαρτάται από το χρόνο. Εισάγουμε τους ακόλουθους τύπους ιστορικών CNN επερωτήσεων:

- Η επερώτηση *HCNN_Q_p* (D, Q_p, Q_{per}) σε ένα σημείο Q_p που παραμένει σταθερό σε μία χρονική περίοδο Q_{per} μας δίνει μία σειρά από τριάδες που αποτελούνται από έναν χρονικά μεταβαλλόμενο πραγματικό αριθμό R_i (χρονικά μεταβαλόμενη απόσταση), ένα κινούμενο αντικείμενο O_i (που ανήκει στη βάση δεδομένων D) και την αντίστοιχη χρονική περίοδο $[t_{i\text{-}start}, t_{i\text{-}end})$ για την οποία ισχύει η πλησιέστερη απόσταση μεταξύ Q_p και O_i. Αυτές οι χρονικά μεταβαλλόμενες πραγματικές τιμές R_i είναι, σε οποιαδήποτε χρονική στιγμή της διάρκειας ζωής τους, μικρότερες ή ίσες προς την απόσταση μεταξύ οποιουδήποτε κινούμενου αντικειμένου O_j στη D και του σημείου επερώτησης Q_p. Οι χρονικές περίοδοι $[t_{i\text{-}start}, t_{i\text{-}end})$ είναι ξένες μεταξύ τους και η ένωσή τους σχηματίζει το Q_{per}.
- Ομοίως, η *HCNN_Q_T* (D, Q_T, Q_{per}) διαφέρει, σε σχέση με την προηγούμενη, στο σημείο επερώτησης που σε αυτή την περίπτωση είναι ένα κινούμενο αντικείμενο με τροχιά Q_T. Οι αντίστοιχες χρονικά μεταβαλλόμενες πραγματικές τιμές R_i είναι, σε κάθε χρονική στιγμή της διάρκειας ζωής τους, μικρότερες ή ίσες προς την απόσταση μεταξύ οποιουδήποτε κινούμενου αντικειμένου O_j και της τροχιάς επερώτησης Q_T. Οι αντίστοιχες χρονικές περίοδοι $[t_{i\text{-}start}, t_{i\text{-}end})$ είναι ξένες μεταξύ τους και η ένωσή τους μας δίνει το Q_{per}.

Οι ανωτέρω τέσσερις επερωτήσεις γενικεύονται για να μας δώσουν τις αντίστοιχες επερωτήσεις *k*-οστού NN. Η γενίκευση των πρώτων δύο επερωτήσεων προκύπτει απλά ζητώντας το 1^o , 2^o , ….*k*-οστό πλησιέστερο σημείο – σε σχέση με ένα σημείο επερώτησης ή μία τροχιά επερώτησης– από το οποίο πέρασε ένα κινούμενο αντικείμενο O_i κατά τη χρονική περίοδο $Q_{per,}$ εξαιρώντας ταυτόχρονα τα σημεία που ανήκουν σε ένα κινούμενο αντικείμενο που έχει ήδη εντοπισθεί ως το *j* –οστό πλησιέστερο ($1 \le j < k$). Οι ιστορικές συνεχείς επερωτήσεις γενικεύονται για να μας δώσουν τους *k*-HCNN ζητώντας τις *k* λίστες τριάδων $\{R_i, [t_{i\text{-}start}, t_{i\text{-}end}), O_i\}$. Κατόπιν, για οποιαδήποτε χρονική στιγμή στη διάρκεια της χρονικής περιόδου Q_{per}, η *i*-στη σειρά ($1 \le i \le k$) θα περιλαμβάνει το *i*-στο στη σειρά κινούμενο αντικείμενο NN (σε σχέση με το σημείο επερώτησης ή την τροχιά της επερώτησης) σε αυτή τη χρονική στιγμή.

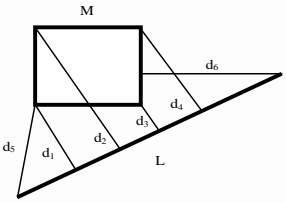
Για να δώσουμε ένα παράδειγμα των προτεινόμενων επεκτάσεων *k*-οστού NN, ας ξαναδούμε το Σχήμα 3.1. Αναζητώντας τις μορφές 2-NN των τεσσάρων επερωτήσεων (Επερώτηση 1, 2, 3 και 4) που παρουσιάζονται στην Ενότητα 1, θα έχουμε τα εξής αποτελέσματα:

- Επερώτηση 1 (ιστορική μη συνεχής): O_1 ($1^{ος}$ NN) και O_2 ($2^{ος}$ NN)
- Επερώτηση 2 (ιστορική συνεχής): η σειρά 1-NN περιλαμβάνει τον O_2 για το διάστημα $[t_1,t_3)$ και τον O_1 για το διάστημα $[t_3,t_4]$· η σειρά 2-NN περιλαμβάνει τον O_1 για το διάστημα $[t_1,t_3)$ και τον O_2 για το διάστημα $[t_3,t_4]$
- Επερώτηση 3 (ιστορική μη συνεχής): O_2 ($1^{ος}$ NN) and O_4 ($2^{ος}$ NN)
- Επερώτηση 4 (ιστορική συνεχής): η σειρά 1-NN περιλαμβάνει τον O_5 για το διάστημα $[t_2,t_5)$ και τον O_4 για το διάστημα $[t_5,t_6]$· η σειρά 2-NN περιλαμβάνει τον O_4 για το διάστημα $[t_2,t_5)$ και τον O_5 για το διάστημα $[t_5,t_6]$.

3.3.2. Μετρικές

Βασιζόμαστε στον ορισμό της μετρικής ελάχιστης απόστασης (*MINDIST*) που παρουσιάζεται στο [RKV95] μεταξύ σημείων και ορθογωνίων παραλληλογράμμων, προκειμένου να υπολογίσουμε, την ελάχιστη απόσταση μεταξύ γραμμικών τμημάτων και ορθογωνίων παραλληλογράμμων και την ελάχιστη απόσταση μεταξύ τροχιών και ορθογωνίων που απαιτούνται για την εφαρμογή των αλγορίθμων που συζητήσαμε στα παραπάνω.

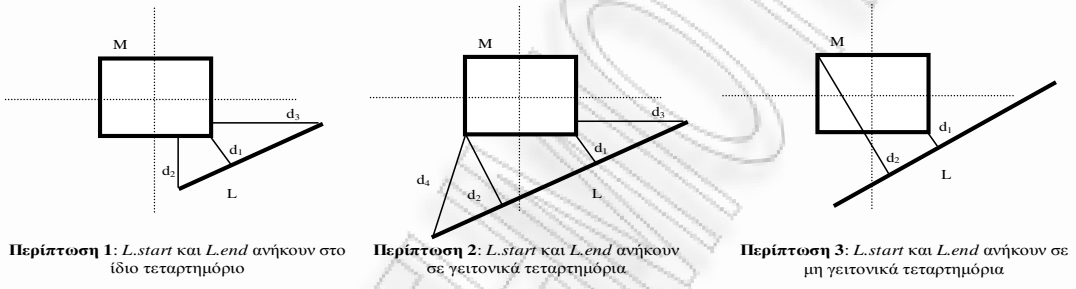
Αρχικώς, στο [RKV95], οι Roussopoulos et al. όρισαν την Ελάχιστη Απόσταση (*MINDIST*) μεταξύ ενός σημείου *P* κι ενός ορθογωνίου παραλληλογράμμου *R* στο n-διάστατο χώρο ως το τετράγωνο της Ευκλείδειας απόστασης μεταξύ του *P* και της πλησιέστερης κορυφής του *R*, αν το *P* βρίσκεται εκτός του *R* (ή μηδέν, αν το *P* βρίσκεται εντός του *R*). Στη συνέχεια οι Tao et al. [TPS02] πρότειναν μία μέθοδο για τον υπολογισμό της *MINDIST* μεταξύ ενός 2D γραμμικού τμήματος *L* κι ενός ορθογωνίου παραλληλογράμμου M (Σχήμα 3.2).



Σχήμα 3.2: Υπολογισμός της *MINDIST* μεταξύ ενός γραμμικού τμήματος και ενός ορθογωνίου παραλληλογράμμου [TPS02]

Η μέθοδος υπολογισμού της *MINDIST* στο [TPS02] αρχικά καθορίζει κατά πόσο το *L* τέμνει το *M*· αν το τέμνει, η *MINDIST* είναι μηδέν. Ειδάλλως, επιλέγεται η μικρότερη μεταξύ έξι αποστάσεων, πιο συγκεκριμένα των τεσσάρων αποστάσεων από την κάθε γωνία του *M* και του *L* (*d*1, *d*2, *d*3, *d*4) και των δύο ελάχιστων αποστάσεων από την αρχή και το τέλος του *L* στο *M* (*d*5, *d*6). Συνεπώς, ο υπολογισμός της *MINDIST* μεταξύ ενός γραμμικού τμήματος και ενός ορθογωνίου παραλληλογράμμου περιλαμβάνει έναν έλεγχο για το κατά πόσο τέμνονται, τέσσερις υπολογισμούς της απόστασης από ευθύγραμμο τμήμα προς σημείο και δύο υπολογισμούς *MINDIST* από σημείο προς παραλληλόγραμμο.

Στην παρούσα διατριβή προτείνεται μία πιο αποτελεσματική μέθοδος υπολογισμού της *MINDIST* μεταξύ ενός γραμμικού τμήματος *L* και ενός ορθογωνίου παραλληλογράμμου *M* (Σχήμα 3.3). Όπως και προηγουμένως, αν το *L* τέμνει το *M*, τότε η *MINDIST* είναι προφανώς μηδέν. Ειδάλλως, χωρίζουμε το χώρο σε τέσσερα τεταρτημόρια χρησιμοποιώντας τους δύο άξονες που τέμνονται στο κέντρο του *M* και καθορίζουμε τα τεταρτημόρια *Qs* και *Qe* στα οποία βρίσκονται το αρχικό σημείο (*L.start*) και το τελικό σημείο (*L.end*) του *L*, αντίστοιχα.



Περίπτωση 1: *L.start* και *L.end* ανήκουν στο ίδιο τεταρτημόριο

Περίπτωση 2: *L.start* και *L.end* ανήκουν σε γειτονικά τεταρτημόρια

Περίπτωση 3: *L.start* και *L.end* ανήκουν σε μη γειτονικά τεταρτημόρια

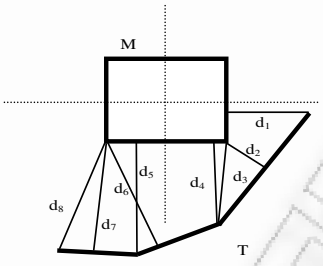
Σχήμα 3.3: Η προτεινόμενη μέθοδος υπολογισμού της *MINDIST* μεταξύ ενός γραμμικού τμήματος και ενός ορθογωνίου παραλληλογράμμου

Κατόπιν, η *MINDIST* είναι η ελάχιστη μεταξύ των:

- **Περίπτωση 1** (τα δύο ακραία σημεία του γραμμικού τμήματος ανήκουν στο ίδιο τεταρτημόριο (*Qs*)): (i) η *MINDIST* μεταξύ της γωνίας του *M* στο *Qs* και του *L*, (ii) η *MINDIST* μεταξύ του *L.start* και του *M*, και (iii) η *MINDIST* μεταξύ του *L.end* και του *M*.
- **Περίπτωση 2** (το *L.start* και το *L.end* ανήκουν σε παρακείμενα τεταρτημόρια το *Qs* και το *Qe*, αντίστοιχα): (i) η *MINDIST* μεταξύ της γωνίας του *M* στο *Qs* και του *L*, (ii) η *MINDIST* μεταξύ της γωνίας του *M* στο *Qe* και του *L*, και (iii) η *MINDIST* μεταξύ του *L.start* και του *M* ή (iv) η *MINDIST* μεταξύ του *L.end* και του *M*.
- **Περίπτωση 3** (Τα *L.start* και *L.end* ανήκουν σε μη παρακείμενα τεταρτημόρια τα *Qs* και *Qe*, αντίστοιχα):δύο *MINDIST* μεταξύ των δύο γωνιών του *M*, που δεν ανήκουν ούτε στο *Qs* ούτε στο *Qe*, και το *L*.

Η μέθοδος αυτή χρησιμοποιεί έναν μικρότερο αριθμό υπολογισμών αποστάσεων (σημείο προς γραμμικό τμήμα και σημείο προς ορθογώνιο παραλληλόγραμμο) σε σύγκριση με τον αντίστοιχο αλγόριθμο στο [TPS02]. Η χειρότερη περίπτωση του προτεινόμενου υπολογισμού της *MINDIST* περιλαμβάνει τον καθορισμό του τεταρτημορίου στο οποίο ανήκουν το αρχικό και το τελικό σημείο του γραμμικού τμήματος, δύο υπολογισμούς απόστασης σημείο προς γραμμικό τμήμα και δύο υπολογισμούς σημείο προς ορθογώνιο παραλληλόγραμμο, ενώ ο αντίστοιχος αλγόριθμος του [TPS02]

χρησιμοποιεί τέσσερις υπολογισμούς σημείο προς γραμμικό τμήμα και δύο σημείο προς ορθογώνιο παραλληλόγραμμο. Συνεπώς, ο προτεινόμενος υπολογισμός της *MINDIST*, στη χειρότερη περίπτωση, καθορίζει το τεταρτημόριο του αρχικού και του τελικού σημείου αντί να εκτελεί 2 επιπλέον υπολογισμούς σημείο προς γραμμικό τμήμα. Η αποδοτικότητα της προτεινόμενης βελτίωσης στον υπολογισμό της *MINDIST* για γραμμικά τμήματα και τροχιές θα αποδειχθεί και στη πειραματική μας μελέτη.

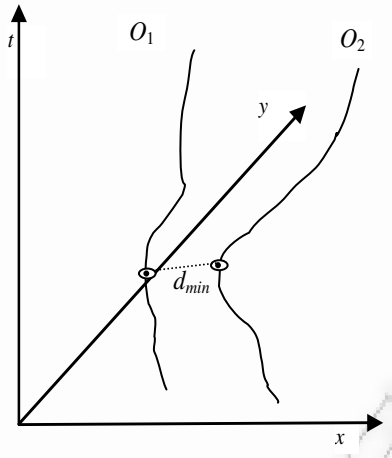


Σχήμα 3.4: Η προτεινόμενη μέθοδος υπολογισμού της *MINDIST* μεταξύ της προβολή μιας τροχιάς στο επίπεδο και ενός ορθογωνίου παραλληλογράμμου

Τέλος, προεκτείνουμε τον ανωτέρω τρόπο υπολογισμού για να υπολογίσουμε τη μετρική *MINDIST* μεταξύ της προβολής μιας τροχιάς *T* στο επίπεδο και ενός ορθογωνίου παραλληλογράμμου *M* (Σχήμα 3.4) και παρέχουμε τη μετρική *MINDIST_Trajectory_Rectangle*. Δεδομένου ότι μία τέτοια προβολή μπορεί να θεωρηθεί σαν μία συλλογή 2D γραμμικών τμημάτων, η *MINDIST_Trajectory_Rectangle* μεταξύ της προβολής μιας τροχιάς και ενός παραλληλογράμμου μπορεί να υπολογιστεί ως το ελάχιστό όλων των *MINDIST* μεταξύ του ορθογωνίου παραλληλογράμμου και κάθε γραμμικού τμήματος της διαδρομής. Η αποτελεσματικότητα του υπολογισμού ενισχύεται απλώς και μόνο αποφεύγοντας τους διπλούς υπολογισμούς για το τεταρτημόριο και την *MINDIST* της αρχής και του τέλους των γειτονικών γραμμικών τμημάτων σε σχέση με το παράθυρο της ερώτησης.

3.3.3. Καθορισμός της Συνάρτησης Απόστασης Μεταξύ Ταυτόχρονα Κινουμένων Τροχιών

Πριν περάσουμε στον πυρήνα του παρόντος κεφαλαίου που περιγράφει τους αντίστοιχους αλγορίθμους, θα πρέπει να επισημάνουμε ότι οποιοσδήποτε αλγόριθμος ανακτά από μια βάση δεδομένων την πλησιέστερη τροχιά σε ένα αντικείμενο επερώτησης (δηλαδή, την περίπτωση $NN\_Q_T(D, Q_T, Q_{per})$), θα πρέπει να υπολογίσει την ελάχιστη απόσταση μεταξύ δύο ταυτόχρονα κινούμενων τροχιών· επιπλέον, δεδομένου ότι η αναζήτηση του Ιστορικά Συνεχούς Πλησιέστερου Γείτονα $HCNN\_Q_T(D, Q_T, Q_{per})$ ανακτά χρονικά μεταβαλλόμενες πραγματικές τιμές R_i που περιγράφουν την απόσταση μεταξύ του αντικειμένου της επερώτησης και των πλησιέστερων τροχιών στη βάση δεδομένων σε οποιαδήποτε χρονική στιγμή της περιόδου επερώτησης Q_{per}, προκύπτει ότι αυτές οι χρονικά μεταβαλλόμενες πραγματικές τιμές θα πρέπει να είναι συναρτήσεις της απόστασης ως προς το χρόνο μεταξύ των αντίστοιχων τροχιών (ή μεταξύ των τροχιών και του σημείου). Από την άλλη πάλι, δεδομένου ότι οι τροχιές κινούμενων αντικειμένων μπορούν να μοντελοποιηθούν ως σειρές συνεχόμενων 3D γραμμικών τμημάτων (δηλαδή 3D πολυγραμμών), αυτή η ελάχιστη απόσταση μπορεί να ερμηνευθεί ως η ελάχιστη απόσταση μεταξύ δύο 3D γραμμικών τμημάτων.



Σχήμα 3.5: Ελάχιστη σύγχρονη ευκλείδεια («οριζόντια») απόσταση μεταξύ δύο τροχιών

Η συνάρτηση της Σύγχρονης Ευκλείδειας Απόστασης (ήτοι, «οριζόντια» στο Σχήμα 3.5) μεταξύ δύο 3D γραμμικών τμημάτων, $P\left(\left(P_{1x},P_{1y},t_1\right),\left(P_{2x},P_{2y},t_2\right)\right)$ και $Q\left(\left(Q_{1x},Q_{1y},t_1\right),\left(Q_{2x},Q_{2y},t_2\right)\right)$ είναι:

$$Dist(t)=\sqrt{\left(Q_x(t)-P_x(t)\right)^2+\left(Q_y(t)-P_y(t)\right)^2} \tag{3.1}$$

Αν αντικαταστήσουμε τα $Q_x(t)=Q_{1x}+\left(Q_{2x}-Q_{1x}\right)\cdot\Delta t$, $Q_y(t)=Q_{1y}+\left(Q_{2y}-Q_{1y}\right)\cdot\Delta t$, $P_x(t)=P_{1x}+\left(P_{2x}-P_{1x}\right)\cdot\Delta t$, $P_y(t)=P_{1y}+\left(P_{2y}-P_{1y}\right)\cdot\Delta t$, όπου $\Delta t=\frac{t-t_1}{t_2-t_1}$, στην (3.1), έχουμε

$$Dist(t)=\sqrt{\left(Q_{1x}+\left(Q_{2x}-Q_{1x}\right)\cdot\Delta t-P_{1x}-\left(P_{2x}-P_{1x}\right)\cdot\Delta t\right)^2+\left(Q_{1y}+\left(Q_{2y}-Q_{1y}\right)\cdot\Delta t-P_{1y}-\left(P_{2y}-P_{1y}\right)\cdot\Delta t\right)^2}$$

Στην συνέχεια, χρησιμοποιούμε το τετράγωνο της Ευκλείδειας απόσταση για ευκολία.

$$Dist^2(t)=\left(Q_{1x}+\left(Q_{2x}-Q_{1x}\right)\cdot\Delta t-P_{1x}-\left(P_{2x}-P_{1x}\right)\cdot\Delta t\right)^2+\left(Q_{1y}+\left(Q_{2y}-Q_{1y}\right)\cdot\Delta t-P_{1y}-\left(P_{2y}-P_{1y}\right)\cdot\Delta t\right)^2=$$

$$=\left(\left(Q_{2x}-Q_{1x}-P_{2x}+P_{1x}\right)^2+\left(Q_{2y}-Q_{1y}-P_{2y}+P_{1y}\right)^2\right)\cdot\Delta t^2+$$

$$+2\left(\left(Q_{2x}-Q_{1x}-P_{2x}+P_{1x}\right)\left(Q_{1x}-P_{1x}\right)+\left(Q_{2y}-Q_{1y}-P_{2y}+P_{1y}\right)\left(Q_{1y}-P_{1y}\right)\right)\cdot\Delta t+\left(Q_{1x}-P_{1x}\right)^2+\left(Q_{1y}-P_{1y}\right)^2$$

Θέτοντας

$$A=\left(Q_{2x}-Q_{1x}-P_{2x}+P_{1x}\right)^2+\left(Q_{2y}-Q_{1y}-P_{2y}+P_{1y}\right)^2, \tag{3.2}$$

$$B=2\left(\left(Q_{2x}-Q_{1x}-P_{2x}+P_{1x}\right)\left(Q_{1x}-P_{1x}\right)+\left(Q_{2y}-Q_{1y}-P_{2y}+P_{1y}\right)\left(Q_{1y}-P_{1y}\right)\right), \tag{3.3}$$

$$C=\left(Q_{1x}-P_{1x}\right)^2+\left(Q_{1y}-P_{1y}\right)^2, \tag{3.4}$$

η Σύγχρονη Ευκλείδεια «οριζόντια» συνάρτηση της απόστασης δύο 3D γραμμικών τμημάτων υπολογίζεται ως εξής:

$$Dist^2(t)=\frac{A}{(t_2-t_1)^2}t^2+\left(\frac{B}{t_2-t_1}-\frac{2A\cdot t_1}{(t_2-t_1)^2}\right)t+\frac{A\cdot t_1^2}{(t_2-t_1)^2}-\frac{B\cdot t_1}{t_2-t_1}+C, \tag{3.5}$$

όπου τα A, B, C προκύπτουν από τις (3.2), (3.3), και (3.4), αντίστοιχα.

Σύμφωνα με την (3.5), το τετράγωνο της Σύγχρονης Ευκλείδειας «οριζόντιας» συνάρτησης της απόστασης δύο 3D γραμμικών τμημάτων έχει την τετραγωνική μορφή $Dist(t)=a\cdot t^2+b\cdot t+c$, που ελαχιστοποιείται στη τιμή $Dist_{\min}=c-\frac{b^2}{4a}$ με $t=-\frac{b}{2a}$. Έτσι, στην περίπτωσή μας

$$Dist^2{}_{\min} = \frac{A \cdot t_1^2}{(t_2 - t_1)^2} - \frac{B \cdot t_1}{t_2 - t_1} + C - \frac{\left(\frac{B}{t_2 - t_1} - \frac{2A \cdot t_1}{(t_2 - t_1)^2} \right)^2}{\frac{4A}{(t_2 - t_1)^2}} \quad (3.6)$$

με

$$t = \frac{\frac{2A \cdot t_1}{(t_2 - t_1)^2} - \frac{B}{t_2 - t_1}}{\frac{2A}{(t_2 - t_1)^2}} \quad (3.7)$$

όπου τα *A*, *B*, *C* προκύπτουν από τις (3.2), (3.3), και (3.4), αντιστοίχως.

Θα πρέπει εδώ να τονίσουμε ότι ο τύπος (3.6) μπορεί να χρησιμοποιηθεί στην περίπτωση που το *t* που υπολογίζεται από τον (3.7) βρίσκεται εντός της χρονικής περιόδου επερώτησης $Q_{per}[t_{start}, t_{end}]$. Ειδάλλως, διακρίνουμε τις ακόλουθες δύο περιπτώσεις:

- αν $t \leq t_{start}$, τότε η ελάχιστη Σύγχρονη Ευκλείδεια «οριζόντια» απόσταση προκύπτει από τον τύπο (3.5) με $t = t_{start}$
- αν $t \geq t_{end}$, τότε η ελάχιστη Σύγχρονη Ευκλείδεια «οριζόντια» απόσταση προκύπτει από τον τύπο (3.5) με $t = t_{end}$.

3.4. Αλγόριθμοι για Επερωτήσεις Πλησιέστερου Γείτονα σε Τροχιές

Σε αυτήν την ενότητα εισάγονται διεξοδικότερα αρκετοί αλγόριθμοι, που απαντούν στους πρώτους δύο (ιστορικούς μη συνεχείς) τύπους επερωτήσεων NN που παρουσιάζονται στην Ενότητα 3.3.1, ενώ στη συνέχεια γενικεύονται για να υποστηρίξουν τις αντίστοιχες επερωτήσεις *k*-NN. Ακολουθούνται και οι δύο προσεγγίσεις που χρησιμοποιούνται παραδοσιακά για την επεξεργασία επερωτήσεων πλησιέστερου γείτονα σε χωρικά δεδομένα, δηλαδή οι προσεγγίσεις «Πρώτα στο Βαθύτερο» (Depth-First) [RKV95] και «Πρώτα στο Καλύτερο» (Best-First) [HS99]. Παρουσιάζουμε λοιπόν, πρώτα του αλγορίθμους «πρώτα στο βαθύτερο» και στην συνέχεια τους «πρώτα στο καλύτερο».

3.4.1. Μη αυξητικοί («Πρώτα στο Βαθύτερο») Αλγόριθμοι NN σε Τροχιές

Παρακάτω παρουσιάζονται μη αυξητικοί αλγόριθμοι που απαντούν στους δύο πρώτους (ιστορικούς μη συνεχείς) τύπους επερωτήσεων NN που παρουσιάζονται στην Ενότητα 3.3.1 και γενικεύονται για να υποστηρίξουν τις αντίστοιχες επερωτήσεις *k*-NN.

3.4.1.1. Μη Αυξητικός Αλγόριθμος NN για Σταθερά Αντικείμενα Επερώτησης

Ο μη αυξητικός αλγόριθμος NN για σταθερά αντικείμενα επερώτησης (αλγόριθμος `PointNNSearch` στο Σχήμα 3.6), μας δίνει τη δυνατότητα να απαντήσουμε σε επερωτήσεις NN για σταθερό αντικείμενο επερώτησης Q_p, στη διάρκεια ενός συγκεκριμένου χρονικού διαστήματος επερώτησης $Q_{per}[t_{start}, t_{end}]$. Ο αλγόριθμος χρησιμοποιεί την ίδια ευριστική με τις [RKV95] και [CF98], κλαδεύοντας συγχρόνως το χώρο αναζήτησης βάσει του Q_{per}.

Ο αλγόριθμος προσπελαύνει το δέντρο (που δεικτοδοτεί τις τροχιές των κινούμενων αντικειμένων) με τακτική «πρώτα στο βαθύτερο» κλαδεύοντας τους κόμβους του δέντρου βάσει του Q_{per} απορρίπτοντας αυτούς που είναι πλήρως εκτός του διαστήματος. Στο επίπεδο των φύλλων, ο αλγόριθμος εξετάζει τις εγγραφές του φύλλου ελέγχοντας κατά πόσον η χρονική διάρκεια μιας εγγραφής αλληλεπικαλύπτεται με την Q_{per} (Γραμμή 7)· αν η χρονική συνιστώσα της εγγραφής είναι

πλήρως εντός της Q_{per}, ο αλγόριθμος υπολογίζει την πραγματική Ευκλείδεια απόσταση μεταξύ του Q και της (χωρικής συνιστώσας της) εγγραφής· ειδάλλως, αν η χρονική συνιστώσα της εγγραφής είναι μόνο μερικώς εντός της Q_{per}, εφαρμόζεται γραμμική παρεμβολή για να υπολογίσουμε το τμήμα της εγγραφής που βρίσκεται εντός της Q_{per} (Γραμμή 9) και να υπολογίσουμε την Ευκλείδεια απόσταση μεταξύ του Q και αυτού του τμήματος της εγγραφής. Όταν επιλεγεί ένας υποψήφιος *πλησιέστερος γείτονας* (*nearest*), ο αλγόριθμος, οπισθοδρομώντας στο ανώτερο επίπεδο κλαδεύει τους κόμβους στην ενεργή λίστα του κλαδιού (Γραμμή 27) εφαρμόζοντας την ευριστική *MINDIST* [RKV95] [CF98].

```
 1. Algorithm PointNNSearch(node N, point Q, period Q_per, struct Nearest)
 2.    IF N Is Leaf
 3.    // Iterate through leaf entries computing Euclidean
 4.    //distance from point Q
 5.       FOR EACH Entry E IN N
 6.          // If entry is (fully or partially) inside the period
 7.          IF Q_per Overlaps (E.T_S, E.T_E)
 8.             // Compute entry's spatial extent inside the period
 9.             nE = Interpolate(E, Max(Q_per.T_S, E.T_S), Min(Q_per.T_E, E.T_E))
10.             // Compute Entry's actual distance from Q.
11.             // Update Nearest if necessary
12.             Dist = Euclidean_Dist_2D(Q, nE)
13.             IF Dist < Nearest.Dist Update Nearest with nE, Dist
14.          ENDIF
15.       NEXT
16.    ELSE
17.       // Generate Node's branch list with entries overlapping
18.       // the query period
19.       BranchList = GenBranchList(Q, N, Q_per)
20.       // Sort active branch List by MinDist
21.       SortBranchList(BranchList)
22.       // Iterate through active branch List
23.       FOR EACH Entry E IN BranchList
24.          // Visit Child Nodes
25.          PointNNSearch(E.ChildNode, Q, Q_per, Nearest)
26.          // Apply MinDist heuristic to do pruning
27.          PruneBranchList(BranchList)
28.       NEXT
29.    ENDIF
```

Σχήμα 3.6: Αλγόριθμος αναζήτησης ιστορικού NN για σταθερά σημεία επερώτησης

3.4.1.2. Μη Αυξητικός Αλγόριθμος NN για Κινούμενα Αντικείμενα Επερώτησης

Ο αλγόριθμος `PointNNSearch` μπορεί να τροποποιηθεί για να υποστηρίξει το δεύτερο τύπο επερώτησης NN όπου το αντικείμενο της επερώτησης είναι μία τροχιά ενός κινούμενου σημείου (αλγόριθμος `TrajectoryNNSearch`, Σχήμα 3.7). Στο επίπεδο των φύλλων, ο αλγόριθμος υπολογίζει την ελάχιστη Ευκλείδεια απόσταση μεταξύ των εγγραφών του φύλλου και κάθε τμήματος της τροχιάς επερώτησης χρησιμοποιώντας τη συνάρτηση `Min_Horizontal_Dist` (Γραμμή 10), που υπολογίζει την ελάχιστη Σύγχρονη Ευκλείδεια απόσταση μεταξύ δύο 3D γραμμικών τμημάτων, εφαρμόζοντας την εξίσωση (3.6) ή (3.5), σύμφωνα με τη σχετική συζήτηση που προηγήθηκε. Επιπλέον, για κάθε τμήμα της τροχιάς επερώτησης *QE* και πριν τον υπολογισμό της απόστασης από την τρέχουσα εγγραφή φύλλου κάνουμε πρώτα παρεμβολή προκειμένου να πάρουμε μια πλειάδα τμήματος εγγραφής – τμήματος επερώτησης με πανομοιότυπη χρονική έκταση (Γραμμές 8, 9). Για να μειώσει τον αριθμό αξιολογήσεων των χρονικών αλληλοεπικαλύψεων μεταξύ των εγγραφών του φύλλου και των τμημάτων της τροχιάς, ο αλγόριθμός μας χρησιμοποιεί μία μέθοδο σάρωσης επιπέδου (plane sweep method), που εξετάζει τις εγγραφές του φύλλου και τα τμήματα της τροχιάς επερώτησης

σε ένα πέρασμα κατά τη χρονική τους διάσταση (Γραμμές 5, 6, 7). Αυτό σημαίνει ότι οι εγγραφές του φύλλου θα πρέπει προηγουμένως να έχουν ταξινομηθεί βάσει της χρονικής τους έκτασης (Γραμμή 4), εκτός κι αν η δομή που χρησιμοποιείται τα αποθηκεύει με χρονική σειρά ούτως ή άλλως (όπως π.χ. το TB-δέντρο).

```
 1.  Algorithm TrajectoryNNSearch(node N, trajectory Q, period Q_per,
                                  struct Nearest)
 2.     Q = Interpolate(Q, Max(Q.T_S, Q_per.T_S), Min(Q.T_E, Q_per.T_E))
 3.     IF N Is Leaf
 4.        Sort(N, T_S) // Sort A-Z Entries in Node N by their Tstart
 5.        FOR EACH Entry E IN N
 6.           Find next query trajectory entry QS with QS.T_e<N.T_S; QE=QS
 7.           DO UNTIL QE.T_s > E.T_e
 8.              nE = Interpolate(E, Max(QE.T_S, E.T_S), Min(QE.T_E, E.T_E))
 9.              nQE = Interpolate(QE, Max(QE.T_S, E.T_S), Min(QE.T_E, E.T_E))
10.              Dist = Min_Horizontal_Dist(nQE, nE)
11.              IF Dist < Nearest.Dist Update Nearest with nE, Dist
12.           NEXT query entry QE
13.           Return QE in the query entry QS
14.        NEXT
15.     ELSE
16.        BranchList = GenTrajectoryBranchList(Q, N)
17.        SortBranchList(BranchList)
18.        FOR EACH Entry E IN BranchList
19.           TrajectoryNNSearch(E.ChildNode, E.Trajectory, Nearest)
20.           PruneBranchList(BranchList)
21.        NEXT
22.     ENDIF
```

Σχήμα 3.7: Αλγόριθμος αναζήτησης ιστορικού NN για κινούμενα σημεία επερώτησης

```
 1.   Algorithm genTrajectoryBranchList(node N, trajectory Q)
 2.      FOR EACH Entry E IN N
 3.         // If entry is partially inside the trajectory lifetime
 4.         IF (Q.T_S, Q.T_E) Overlaps (E.T_S, E.T_E)
 5.            // Compute trajectory's spatial extent inside E's lifetime
 6.            nQ = Interpolate(Q, Max(Q.T_S, E.T_S), Min(Q.T_E, E.T_E))
 7.            // Compute MinDist between the trajectory and the rectangle
 8.            Dist=MinDist_Trajectory_Rectangle(nQ, E)
 9.            // Add the rectangle along with its calculated distance and
10.            // the interpolated trajectory in the list
11.            List.Add(E, Dist, nQ)
12.         ENDIF
13.      NEXT
14.      RETURN List
```

Σχήμα 3.8: Αλγόριθμος δημιουργίας της λίστας του κλαδιού του κόμβου *N* βάσει της τροχιάς *Q*

Στα εσωτερικά επίπεδα του δένδρου, ο αλγόριθμος χρησιμοποιεί τη συνάρτηση `GenTrajectoryBranchList` (ψευδοκώδικας στο Σχήμα 3.8) αντί του `GenBranchList`. Η συνάρτηση `GenTrajectoryBranchList` χρησιμοποιεί τη μετρική `MinDist_Trajectory_Rectangle` που εισήχθηκε στην Ενότητα 3.3.2 για να υπολογίσουμε την *MINDIST* μεταξύ της τροχιάς της επερώτησης και του ορθογωνίου παραλληλεπιπέδου κάθε εγγραφής του κόμβου *N*. Εδώ, πρέπει να τονίσουμε ότι δεν χρειάζεται να υπολογίσουμε την `MinDist_Trajectory_Rectangle` σε σχέση με την πραγματική τροχιάς *Q* της επερώτησης, αλλά σε σχέση με το τμήμα του *Q* που βρίσκεται εντός της χρονικής έκτασης του περιβάλλοντος ορθογωνίου του *N* και για να γίνει αυτό (εάν απαιτείται) κάνουμε παρεμβολή για να πάρουμε τη νέα τροχιά επερώτησης *nQ* (Γραμμή 6). Η τροχιά *nQ* αποθηκεύεται και στην *Branchlist* μαζί με την

αντίστοιχη εγγραφή του κόμβου και την υπολογιζόμενη απόσταση (Γραμμή 8). Επειδή όλοι οι κόμβοι στο υποδέντρο του *N* χωρικά και χρονικά περιέχονται στον *N*, η τροχιά *nQ* μπορεί να χρησιμοποιηθεί σαν τροχιά επερώτησης για τους κόμβους στο επόμενο επίπεδο εντός του υποδέντρου, γεγονός που μας επιτρέπει ν' αποφύγουμε περιττούς υπολογισμούς.

3.4.1.3. Επέκταση σε μη Αυξητικούς Αλγορίθμους k-NN

Όπως και στην [RKV95], γενικεύουμε τους παραπάνω δύο αλγορίθμους για την αναζήτηση των *k*-NN λαμβάνοντας υπ' όψιν τα εξής:

- Χρήση μια προσωρινής μνήμης (buffer) το πολύ *k* (τρεχουσών) πλησιέστερων αντικειμένων ταξινομημένων βάσει της πραγματικής τους απόστασης από το αντικείμενο επερώτησης (σημείο ή τροχιά).
- Κλάδεμα βάσει της απόστασης του (τρέχοντος) πιο απομακρυσμένου πλησιέστερου αντικειμένου στην προσωρινή μνήμη.
- Ενημέρωση της απόστασης κάθε κινούμενου αντικειμένου εντός της προσωρινής μνήμης όταν γίνεται εξέταση ενός κόμβου που περιλαμβάνει μια εγγραφή του ιδίου αντικειμένου εγγύτερη στο αντικείμενο της επερώτησης.

3.4.2. Αυξητικοί («Πρώτα στον Καλύτερο») Αλγόριθμοι NN σε Τροχιές

Σε συνέχεια της προηγούμενης ενότητας, παρουσιάζουμε τώρα τους ανάλογους «πρώτα στον καλύτερο» αλγορίθμους και κατόπιν τους γενικεύουμε για υποστήριξη των αντίστοιχων επερωτήσεων *k*-NN.

```
 1.  Algorithm IncPointNNSearch(R-tree R, 2D point Q, time period Q_per)
 2.     EnQueue Queue, R.RootNode, 0
 3.     DO WHILE Queue.Count > 0
 4.        Element = DeQueue(Queue)
 5.        IF Element Is MovingObjectEntry
 6.           RETURN Element as the next nearest object
 7.        ELSEIF Element Is Leaf
 8.           // Iterate through leaf entries computing Euclidean
 9.           // distance from Q
10.           FOR EACH Entry E IN leaf node Element
11.              // If entry is (fully or partially) inside the period
12.              IF Q_per Overlaps (E.T_S, E.T_E)
13.                 // Compute entry's spatial extent inside the period
14.                 nE = Interpolate(E,Max(Q_per.T_S E.T_S),Min(Q_per.T_E,E.T_E))
15.                 // Compute Entry's actual distance from Q.
16.                 Dist = Euclidean_Dist_2D(Q, nE)
17.                 EnQueue Queue, nE, Dist
18.              ENDIF
19.           NEXT
20.        ELSE // Element is a non leaf node
21.           // Iterate through node entries computing their
22.           // minimum distance from Q
23.           FOR EACH Entry E IN node Element
24.              // If entry is (fully or partially) inside the period
25.              IF Q_per Overlaps (E.T_S, E.T_E)
26.                 // Compute Entry's MinDist from Q.
27.                 Dist = MinDist(Q, E)
28.                 EnQueue Queue, E, Dist
29.              ENDIF
30.           NEXT
31.        ENDIF
32.     LOOP
```

Σχήμα 3.9: Αυξητικός αλγόριθμος ιστορικής Αναζήτησης NN για σταθερά σημεία επερώτησης

3.4.2.1. Αυξητικός Αλγόριθμος NN για Σταθερά Αντικείμενα Επερώτησης

Ο προτεινόμενος αλγόριθμος βασίζεται στον αλγόριθμο NN για σταθερά αντικείμενα που παρουσιάζεται στην [HS99] και διασχίζει το δέντρο με τρόπο «πρώτα στον καλύτερο». Ο αλγόριθμος χρησιμοποιεί μια ουρά προτεραιότητας (queue), στην οποία οι εγγραφές των φύλλων αποθηκεύονται με αύξουσα σειρά βάσει της απόστασής τους από το αντικείμενο της επερώτησης.

Το Σχήμα 3.9 παρουσιάζει αναλυτικά τον αλγόριθμο `IncPointNNSearch`. Στην Γραμμή 1, αρχικοποιείται η ουρά προτεραιότητας. Στη Γραμμή 6, αναφέρεται από τον αλγόριθμο το επόμενο πλησιέστερο αντικείμενο. Όπως και στον αντίστοιχο αλγόριθμο «πρώτα στο βαθύτερο» που περιγράφεται στην Ενότητα 3.4.1.1, στο επίπεδο των φύλλων ο αλγόριθμος διατρέχει τις εγγραφές του φύλλου ελέγχοντας κατά πόσο η διάρκεια ζωής μιας εγγραφής αλληλεπικαλύπτεται με τη χρονική περίοδο της επερώτησης Q_{per} (Γραμμή 10)· αν η χρονική συνιστώσα της εγγραφής βρίσκεται πλήρως εντός της Q_{per}, ο αλγόριθμος υπολογίζει την πραγματική Ευκλείδεια απόσταση μεταξύ του Q και της (χωρικής συνιστώσας της) εγγραφής· ειδάλλως, αν η χρονική συνιστώσα της εγγραφής βρίσκεται μόνο μερικώς εντός της Q_{per}, εφαρμόζεται γραμμική παρεμβολή ώστε να υπολογίσουμε το τμήμα της εγγραφής που είναι εντός της Q_{per} (Γραμμή 14) και υπολογίζει την Ευκλείδεια απόσταση μεταξύ του Q και του τμήματος αυτής της εγγραφής (Γραμμή 16). Στη Γραμμή 17, η εγγραφή του φύλλου εισάγεται στην ουρά μαζί με την πραγματική της απόσταση από το αντικείμενο της τροχιάς. Στα εσωτερικά επίπεδα του δένδρου (Γραμμές 23-30), ο αλγόριθμος απλώς υπολογίζει τη *MINDIST* μεταξύ του αντικειμένου της επερώτησης και της εγγραφή του κάθε κόμβου που αλληλεπικαλύπτεται με την περίοδο της επερώτησης Q_{per} και στη συνέχεια εισάγει αυτή την εγγραφή μαζί με την τιμή *MINDIST* στην ουρά.

3.4.2.2. Αυξητικός Αλγόριθμος NN για Κινούμενα Αντικείμενα Επερώτησης

Ο αλγόριθμος `IncPointNNSearch` που προτείνεται παραπάνω μπορεί να τροποποιηθεί ελαφρώς προκειμένου να υποστηρίξει το δεύτερο τύπο επερώτησης NN όπου το αντικείμενο της επερώτησης είναι μία τροχιά κινούμενου σημείου οδηγόντας μας στον αλγόριθμο `IncTrajectoryNNSearch` που εμφανίζεται στο Σχήμα 3.10. Για αυτό το σκοπό γίνονται οι εξής αλλαγές: πρώτον, όπως και στον αντίστοιχο αλγόριθμο «πρώτα στο βαθύτερο» (Ενότητα 3.4.1.1), στο επίπεδο των φύλλων, ο αλγόριθμος υπολογίζει την ελάχιστη «οριζόντια» Ευκλείδεια απόσταση μεταξύ κάθε εγγραφής του φύλλου και του κάθε τμήματος της τροχιάς επερώτησης Q, χρησιμοποιώντας τη συνάρτηση `Min_Horizontal_Dist` (Γραμμή 15) και την εξίσωση (3.6). Χρησιμοποιούμε επίσης τον ίδιο αλγόριθμο σάρωσης επιπέδου, προκειμένου να προσδιορίσουμε ποιες εγγραφές του φύλλου και τμήματα της Q αλληλεπικαλύπτονται στη χρονική τους διάσταση και κατόπιν υπολογίζουμε την απόσταση μεταξύ αυτών που όντως αλληλεπικαλύπτονται (Γραμμές 10-12).

Στα εσωτερικά επίπεδα του δένδρου, ο αλγόριθμος χρησιμοποιεί τη μετρική *MinDist_Trajectory_Rectangle* που εισήχθη στην ενότητα 3.3.2 για τον υπολογισμό της *MINDIST* μεταξύ της τροχιάς επερώτησης και του ορθογωνίου παραλληλεπιπέδου κάθε εγγραφής του κόμβου (Γραμμή 24). Όπως και στον αλγόριθμο `TrajectoryNNSearch`, εάν απαιτείται, κάνουμε παρεμβολή για να πάρουμε το *nQ,* που είναι το τμήμα της Q που βρίσκεται εντός της χρονικής έκτασης του περιβάλλοντος ορθογωνίου παραλληλογράμμου κάθε εγγραφής του κόμβου (Γραμμή 23) και κατόπιν την αποθηκεύουμε στην ουρά μαζί με την αντίστοιχη εγγραφή του κόμβου και την

υπολογιζόμενη απόσταση (Γραμμή 25). Επειδή όλοι οι κόμβοι στο υποδέντρο του *N* χωρικά και χρονικά περιέχονται στο *N*, η τροχιά *nQ* μπορεί να χρησιμοποιηθεί περαιτέρω ως η τροχιά της επερώτησης για τους κόμβους του επόμενου επιπέδου εντός του υποδέντρου, επιτρέποντάς μας έτσι να αποφύγουμε περιττούς υπολογισμούς.

```
 1.  Algorithm IncTrajectoryNNSearch(R-tree R,trajectory Q, period Q_per)
 2.     Q = Interpolate(Q, Max(Q.T_S, Q_per.T_S), Min(Q.T_E, Q_per.T_E))
 3.     EnQueue Queue, R.RootNode, Q, 0
 4.     DO WHILE Queue.Count > 0
 5.        DeQueue(Queue, Element, Q)
 6.        IF Element Is MovingObjectEntry
 7.           RETURN Element as the next nearest object
 8.        ELSEIF Element Is Leaf
 9.           Sort(Element, T_S)// Sort A-Z Entries in Node by their T_start
10.           FOR EACH Entry E IN leaf node Element
11.              Find next query trajectory entry QS with QS.T_e<N.T_S; QE=QS
12.              DO UNTIL QE.T_s > E.T_e
13.                 nE = Interpolate(E,Max(QE.T_S, E.T_S),Min(QE.T_E, E.T_E))
14.                 nQE = Interpolate(QE,Max(QE.T_S, E.T_S),Min(QE.T_E, E.T_E))
15.                 Dist = Min_Horizontal_Dist(nQE, nE)
16.                 EnQueue Queue, nE, Dist
17.              NEXT query entry QE
18.              Return QE in the query entry QS
19.           NEXT
20.        ELSE
21.           FOR EACH Entry E IN node Element
22.              IF (Q.T_S, Q.T_E) Overlaps (E.T_S, E.T_E)
23.                 nQ = Interpolate(Q, Max(Q.T_S, E.T_S), Min(Q.T_E, E.T_E))
24.                 Dist = MinDist_Trajectory_Rectangle(nQ, E)
25.                 EnQueue Queue, E, Dist, nQ
26.              ENDIF
27.           NEXT
28.        ENDIF
29.     LOOP
```

Σχήμα 3.10: Αυξητικός αλγόριθμος ιστορικής αναζήτησης ΝΝ για κινούμενα σημεία επερώτησης

3.4.2.3. Επέκταση σε Αυξητικούς k-NN αλγορίθμους

Οι αλγόριθμοι που περιγράφονται στις Ενότητες 3.4.2.1 και 3.4.2.2 είναι αυξητικοί υπό την έννοια ο *k*-στος ΝΝ μπορεί να προκύψει με ελάχιστη επιπρόσθετη επεξεργασία αφ' ης στιγμής έχει βρεθεί ο (*k*-1)-στος ΝΝ. Θυμηθείτε π.χ. τον `IncTrajectoryNNSearch` στο Σχήμα 3.10. Αφού βρεθεί ο $1^{ος}$ ΝΝ, την επόμενη φορά που η υπόθεση της Γραμμής 4 είναι αληθής, θα έχουμε βρει το $2^{ο}$ ΝΝ.

Εδώ, θα πρέπει να επισημάνουμε ότι οι δύο διαφορετικές στρατηγικές που χρησιμοποιούνται για τους ιστορικούς μη συνεχείς αλγορίθμους ΝΝ φαίνεται να έχουν πλεονεκτήματα αλλά και μειονεκτήματα. Όπως αναφέρθηκε ήδη, ενώ η προσέγγιση «πρώτα στον καλύτερο» έχει πάντα ως αποτέλεσμα την εξέταση λιγότερων κόμβων και λιγότερους υπολογισμούς απόστασης, η απόδοσή του εξαρτάται σε μεγάλο βαθμό από το μέγεθος της ουράς προτεραιότητας· όπως θα δείξουμε σαφώς στα πειράματα, αυτό το μειονέκτημα μπορεί να οδηγήσει τους αυξητικούς αλγορίθμους σε χειρότερη απόδοση από τους αλγορίθμους «πρώτα στο βαθύτερο» από πλευράς χρόνου εκτέλεσης, παρά το γεγονός ότι απαιτείται να εξεταστούν λιγότεροι κόμβοι του δένδρου και να αξιολογηθούν λιγότερες αποστάσεις. Από την άλλη πλευρά, οι αυξητικοί αλγόριθμοι έχουν ένα εξαιρετικό πλεονέκτημα σε σχέση με τους «πρώτα στο βαθύτερο» και αυτό είναι η δυνατότητα ανάκτησης καθενός από τους *k* πλησιέστερους γείτονες με αύξουσα σειρά, ενώ η προσέγγιση «πρώτα στο βαθύτερο» προϋποθέτει πρότερη γνώση της παραμέτρου *k*.

3.5. Αλγόριθμοι για Ιστορικές Επερωτήσεις Συνεχούς Πλησιέστερου Γείτονα σε τροχιές

Στην ενότητα αυτή περιγράφουμε τους ιστορικούς συνεχείς ανάλογους αλγορίθμους με αυτούς της Ενότητας 3.4. Πιο συγκεκριμένα, θα ασχοληθούμε με τον τρίτο τύπο επερώτησης NN (την αναζήτηση του NN σε σχέση με ένα σταθερό σημείο επερώτησης σε οποιαδήποτε στιγμή σε δεδομένη χρονική περίοδο) και με τον τέταρτο τύπο επερώτησης NN (όπου το αντικείμενο της επερώτησης είναι η τροχιά ενός κινούμενου σημείου) και στη συνέχεια θα τους επεκτείνουμε προς αναζήτηση του *k*-NN.

```
 1.  Algorithm HContPointNNSearch(node N, 2D point Q, Period Q_per, List
                                  Nearests, Roof)
 2.     IF N Is Leaf
 3.        FOR EACH Entry E IN N
 4.           IF Q_per Overlaps (E.T_S, E.T_E)
 5.              nE = Interpolate(E, Max(Q_per.T_S, E.T_S), Min(Q_per.T_E,E.T_E))
 6.              MovingDist = ConstructMovingDistance(nE, Q)
 7.              IF MovingDist.D_min < Roof
 8.                 UpdateNearests(Nearests,MovingDist,Roof)
 9.              ENDIF
10.           ENDIF
11.        NEXT
12.     ELSE
13.        BranchList = GenBranchList(Q, N, Q_per)
14.        SortBranchList(BranchList)
15.        PruneHContBranchList(BranchList, Nearests, Roof)
16.        FOR EACH Entry E IN BranchList
17.           HContPointNNSearch(E.ChildNode, Q, Q_per, Nearests, Roof)
18.           PruneHContBranchList(BranchList, Nearests, Roof)
19.        NEXT
20.     ENDIF
```

Σχήμα 3.11: Αλγόριθμος αναζήτησης ιστορικού CNN για σταθερά σημεία επερώτησης

3.5.1. Αλγόριθμος HCNN για Σταθερά Αντικείμενα Επερώτησης

Αρχίζουμε την περιγραφή των αλγορίθμων με το τρίτο τύπο NN επερώτησης που αναζητεί τη βάση δεδομένων για το πλησιέστερη τροχιά σε ένα σταθερό σημείο επερώτησης σε κάθε χρονική στιγμή της διάρκειας μίας δοσμένης περιόδου επερώτησης. Ο αλγόριθμος `HContPointNNSearch` που προτείνεται για αυτό τον τύπο επερώτησης παρουσιάζεται στο Σχήμα 3.11.

Όλοι οι ιστορικοί συνεχείς αλγόριθμοι χρησιμοποιούν μία δομή *κινούμενης απόστασης* (*MovingDist* στο Σχήμα 3.11, Γραμμή 6), που αποθηκεύει τις παραμέτρους της συνάρτησης απόστασης (που υπολογίζεται με χρήση των συντελεστών της Εξ.(3.5)), καθώς και την χρονική έκταση της εγγραφής και τα σχετικά ελάχιστα και μέγιστα της συνάρτησης στη διάρκεια ζωής της. Επίσης αποθηκεύουμε την πραγματική εγγραφή εντός της δομής προκειμένου να μπορέσουμε να τη την επιστρέψουμε ως αποτέλεσμα της επερώτησης. Η συνάρτηση `ConstructMovingDistance` απλά υπολογίζει τη δομή αυτή (δηλαδή τις παραμέτρους της συνάρτησης απόστασης a, b, c, και την ελάχιστη D_{min} και μέγιστη D_{max} της συνάρτησης εντός της διάρκειας ζωής της εγγραφής, εφαρμόζοντας επίσης και τα όσα είπαμε στην Ενότητα 3.3.3).

Στη Γραμμή 6 εμφανίζεται ένα ενδιαφέρον σημείο του αλγορίθμου, εκεί όπου εισάγεται η δομή *Nearests*. Η *Nearests* είναι μία λίστα παρακείμενων κινούμενων αποστάσεων που καλύπτουν χρονικά την περίοδο Q_{per}. Η *Roof* είναι το μέγιστο των κινούμενων αποστάσεων που είναι αποθηκευμένες στη λίστα *Nearests* και χρησιμοποιείται ως κατώφλι για να απορρίψουμε γρήγορα τις εγγραφές (και να

κλαδέψουμε τα κλαδιά στο εσωτερικό επίπεδο του δένδρου) που έχουν ελάχιστη απόσταση μεγαλύτερη της *Roof* (κατά συνέπεια μεγαλύτερη όλων των κινούμενων αποστάσεων που είναι αποθηκευμένες στη λίστα *Nearests*). Η Ενότητα 3.5.3, κάνει μια πλήρη παρουσίαση του τρόπου με τον οποίο διατηρείται η λίστα *Nearests*.

Όταν βρισκόμαστε σε εσωτερικά επίπεδα, ο αλγόριθμος `HContPointNNSearch` κατά την οπισθοδρόμηση εφαρμόζει τον αλγόριθμο κλαδέματος `PruneHContBranchList` (Γραμμή 18), που κλαδεύει την ενεργή λίστα του κάθε κλαδιού χρησιμοποιώντας την ευριστική *MINDIST*: Πρώτα, συγκρίνει την *MINDIST* κάθε εγγραφής με τη *Roof* και στη συνέχεια υπολογίζει τη μέγιστη απόσταση εντός της λίστας *Nearests* στη διάρκεια ζωής της εγγραφής. Κατόπιν, κλαδεύει όλες τις εγγραφές που έχουν *MINDIST* μεγαλύτερη από αυτή που υπολογίζεται.

3.5.2. Αλγόριθμος HCNN για Κινούμενα Αντικείμενα Επερώτησης

Ο τέταρτος τύπος επερώτησης NN είναι η ιστορική συνεχής εκδοχή της επερώτησης NN όπου το αντικείμενο της επερώτησης είναι η τροχιά ενός κινούμενου σημείου. Ο αλγόριθμος `HContTrajNNSearch`, που χρησιμοποιείται για την επεξεργασία αυτού του τύπου επερώτησης φαίνεται στο Σχήμα 3.12.

```
 1.  Algorithm HContTrajNNSearch (node N, Trajectory Q, period Q_per, List
                                  Nearests, Roof)
 2.      Q = Interpolate(Q, Max(Q.T_S, Q_per.T_S), Min(Q.T_E, Q_per.T_E))
 3.      IF N Is Leaf
 4.         Sort(N, T_S)
 5.         FOR EACH Entry E IN N
 6.            FIND next query trajectory entry QS with QS.T_e<N.T_S; QE=QS
 7.            DO UNTIL QE.T_s > E.T_e
 8.               nE = Interpolate(E, Max(QE.T_S, E.T_S), Min(QE.T_E, E.T_E))
 9.               nQE = Interpolate(QE, Max(QE.T_S, E.T_S), Min(QE.T_E, E.T_E))
10.               MovingDist = ConstructMovingDistance(nE, nQE)
11.               IF MovingDist.D_min<Roof
12.                  UpdateNearests(Nearests,MovingDist,Roof)
13.               ENDIF
14.            NEXT query entry QE
15.            Return QE in the query entry QS
16.         NEXT
17.      ELSE
18.         BranchList = GenTrajectoryBranchList(Q, N)
19.         SortBranchList(BranchList)
20.         PruneHContBranchList(BranchList, Nearests, Roof)
21.         FOR EACH Entry E IN BranchList
22.            HContTrajNNSearch(E.ChildNode,E.Trajectory,Nearests, Roof)
23.            PruneHContBranchList(BranchList, Nearests, Roof)
24.         NEXT
25.      ENDIF
```

Σχήμα 3.12: Αλγόριθμος αναζήτησης ιστορικού CNN για κινούμενα σημεία επερώτησης

Ο `HContTrajNNSearch` διαφέρει από τον `HContPointNNSearch` μόνο σε δύο σημεία: Το πρώτο είναι, ότι στο επίπεδο των φύλλων, η συνάρτηση `ConstructMovingDistance` υπολογίζει την κινούμενη απόσταση μεταξύ δύο κινούμενων σημείων, αντί ενός κινούμενου κι ενός σταθερού σημείου (Γραμμή 10). Δεύτερον, στα εσωτερικά επίπεδα, η `GenBranchList` αντικαθίσταται από τη συνάρτηση `GenTrajectoryBranchList` που εισάγεται στην περιγραφή του αλγορίθμου `TrajectoryNNSearch` (Γραμμή 18). Επιπλέον, όπως και στην `TrajectoryNNSearch`, για κάθε συνδιασμό τμήματος της τροχιάς επερώτησης *QE και* εγγραφής

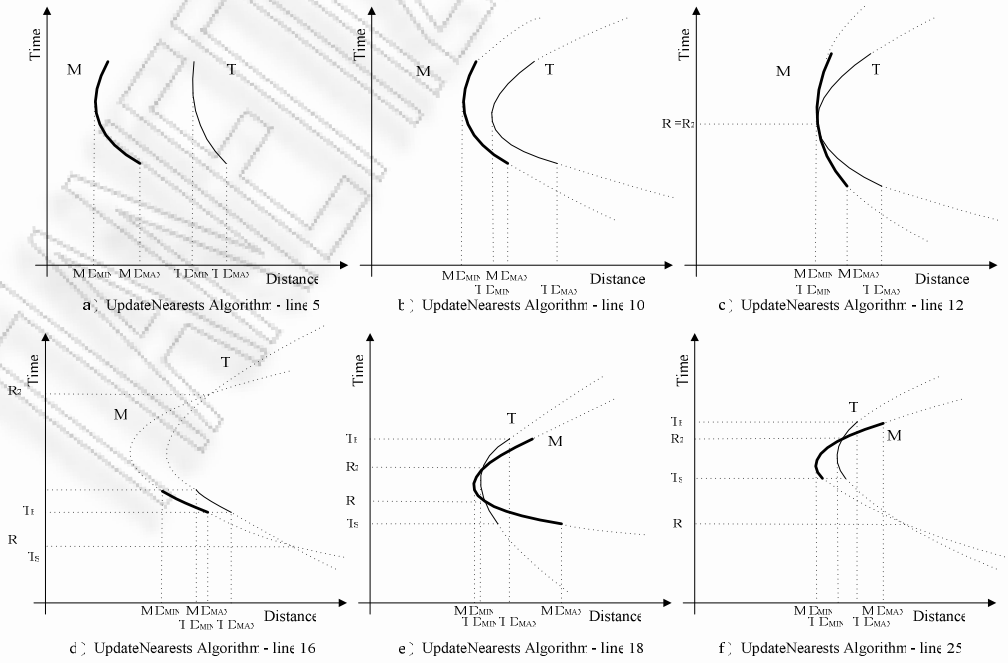
φύλλου, κάνουμε πρώτα παρεμβολή για να πάρουμε μια πλειάδα τμήματος εγγραφής – τμήματος επερώτησης με όμοια χρονική έκταση (γραμμές 8, 9). Επίσης χρησιμοποιούμε την ίδια μέθοδο σάρωσης επιπέδου, για να μειώσουμε το πλήθος των υπολογισμών απόστασης μεταξύ των ευθύγραμμων τμημάτων του *Q* και των εγγραφών του κάθε φύλλου (Γραμμές 5-7).

3.5.3. Διατήρηση της Λίστας *Nearests*

Ο ψευδοκώδικας της συνάρτησης `UpdateNearests`, που είναι υπεύθυνος για τη διατήρηση της λίστας *Nearests*, παρουσιάζεται στο Σχήμα 3.14. Πιο συγκεκριμένα, ο αλγόριθμος διατρέχει τη λίστα *Nearests* αναζητώντας τα στοιχεία της που αλληλεπικαλύπτονται χρονικά με την ελεγχόμενη εγγραφή (*CM*). Όταν βρει ένα τέτοιο στοιχείο, ο αλγόριθμος εφαρμόζει γραμμική παρεμβολή και στις δύο εγγραφές (αυτή που ελέγχεται και αυτή που βρίσκεται ήδη στη λίστα) δίνοντας δύο νέες εγγραφές που έχουν την ίδια χρονική έκταση (*M* και *T*). Κατόπιν, συγκρίνει τις δύο συναρτήσεις απόστασης για να προσδιορίσει αν η εγγραφή στον κατάλογο θα αντικατασταθεί ή όχι.

Το Σχήμα 3.13 εξηγεί γραφικά όλες τις πιθανές συγκρίσεις μεταξύ των παραβολών δύο συναρτήσεων κινούμενης απόστασης. Πιο συγκεκριμένα, το Σχήμα 3.13 (a) αντιστοιχεί στη γραμμή 6 του αλγορίθμου που παρουσιάζεται στο Σχήμα 3.14, όπου η μέγιστη απόσταση του *M* είναι μικρότερη από το ελάχιστο του *T*, οδηγώντας στην αντικατάσταση του *T* με το *M*. Ειδάλλως, μετά τον υπολογισμό της διακρίνουσας μεταξύ των συναρτήσεων απόστασης του *M* και του *T*, πρέπει να κάνουμε διάκριση μεταξύ των τριών παρακάτω περιπτώσεων:

- Διακρίνουσα μικρότερη του μηδενός, που σημαίνει ότι οι δύο συναρτήσεις *M* και *T* είναι ασύμπτωτες και δεν τέμνονται (Γραμμή 10)· πρέπει ελέγξουμε απλώς το ελάχιστο τους για να προσδιορίσουμε το συνολικό ελάχιστο (δες Σχήμα 3.13(b))
- Η διακρίνουσα είναι μηδέν, που σημαίνει ότι οι δύο συναρτήσεις εφάπτονται στο κοινό ελάχιστό τους (Γραμμή 12)· πρέπει να ελέγξουμε το μέγιστο τους για να καθορίσουμε το συνολικό ελάχιστο (δες Σχήμα 3.13(c))



Σχήμα 3.13: Γραφική αναπαράσταση των συγκρίσεων του αλγορίθμου `UpdateNearests`

```
 1.  Algorithm UpdateNearests (List Nearests, struct CM, Roof)
 2.     FOR EACH T IN Nearests
 3.        IF (T.TS, T.TE)Overlaps(CM.TS, CM.TE)
 4.           M=Interpolate(CM, Max(CM.TS, T.TS), Min(CM.TE, T.TE))
 5.           T=Interpolate(T, Max(CM.TS, T.TS), Min(CM.TE, T.TE))
 6.           IF M.DMax < T.DMin
 7.              Nearests.Replace T with M
 8.           ELSEIF M.DMax < T.DMax
 9.              D = Discriminant(M-T)
10.              IF D < 0
11.                 IF T.DMin > M.DMin THEN Nearests.Replace T with M
12.              ELSEIF D=0
13.                 IF T.DMax > M.DMax THEN Nearests.Replace T with M
14.              ELSE
15.                 RR1=Solution1(T - M); RR2=Solution2(T - M)
16.                 R1=Min(RR1,RR2); R2=Max(RR1,RR2)
17.                 IF R2<T.TS OR R1>T.TE
18.                    IF T.DMax > M.DMax THEN Nearests.Replace T with M
19.                 ELSEIF R2<T.TE AND R1>T.TS
20.                    IF M.Dmin < T.Dmin
21.                       M1=Part(M,,R1); M2=Part(M,R2); T1=Part(T,R1,R2)
22.                       Nearests.Replace T with (M1,T1,M2)
23.                    ELSE
24.                       T1=Part(T,,R1); T2=Part(T,R2); M1=Part(M,R1,R2)
25.                       Nearests.Replace T with (T1,T2,M1)
26.                    ENDIF
27.                 ELSE
28.                    IF M(R1 - 1)<T(R1 - 1)
29.                       M1=Part(M,,R1); T1=Part(T,R1)
30.                       Nearests.Replace T with (M1,T1)
31.                    ELSE
32.                       T1=Part(T,,R1); M1=Part(M,R1)
33.                       Nearests.Replace T with (T1,M1)
34.                    ENDIF
35.                 ENDIF
36.              ENDIF
37.           ENDIF
38.        ENDIF
39.        Roof=max(Roof,T.Dmax)
40.     NEXT
```

Σχήμα 3.14: Αλγόριθμος UpdateNearests

- Η διακρίνουσα είναι μεγαλύτερη του μηδενός, που σημαίνει ότι οι δύο συναρτήσεις τέμνονται σε δύο σημεία (Γραμμή 14). Σε αυτή την περίπτωση, πρέπει να προσδιορίσουμε αν αυτές οι χρονικές στιγμές βρίσκονται εντός της διάρκειας ζωής της εγγραφής. Εξ' ου και έχουμε τρεις επιπλέον υποκατηγορίες:
    - Και οι δύο λύσεις είναι εκτός της χρονικής έκτασης του *M* (και του *T*) (Γραμμή 17). Πρέπει απλώς να ελέγξουμε τα μέγιστά τους για να καθορίσουμε ποιο είναι το συνολικό ελάχιστο μέσα στο παρόν χρονικό διάστημα (δες Σχήμα 3.13(d))
    - Και οι δύο λύσεις είναι εντός της χρονικής έκτασης του *M* (και του *T*) (Γραμμή 19). Πρέπει να χωρίσουμε την εγγραφή σε 3 διαφορετικές εγγραφές (δες Σχήμα 3.13(e)) και καθορίζουμε το τμήμα του *T* που θα αντικατασταθεί από το *M*.

- Μόνο μία λύση είναι εντός της χρονικής έκτασης του *M* (Γραμμή 27). Πρέπει να χωρίσουμε την εγγραφή σε 2 διαφορετικές εγγραφές (δες Σχήμα 3.13(f)) και καθορίζουν το τμήμα του *T* που πρέπει να αντικατασταθεί από το *M*.

3.5.4. Επέκταση σε Αλγορίθμους *k*-HCNN

Οι δύο ιστορικοί συνεχείς αλγόριθμοι που προτείνονται παραπάνω μπορούν να γενικευθούν για την αναζήτηση των *k*- πλησιέστερων γειτόνων λαμβάνοντας υπ' όψιν τα εξής:

- Χρήση μιας προσωρινής μνήμης με *k* πλησιέστερες τρέχουσες λίστες *Nearests*
- Κλάδεμα βάσει της απόστασης από τις μακρύτερες σε απόσταση λίστες *Nearests* στην προσωρινή μνήμη – συνεπώς η *Roof* υπολογίζεται ως η μέγιστη απόσταση από την πιο απομακρυσμένη λίστα *Nearests*
- Επεξεργασία κάθε εγγραφής του δέντρου σε σχέση με την *i*-στη λίστα (με το *i* να αυξάνεται, από το 1 στο *k*) ελέγχοντας κατά πόσο θα πρέπει να βρίσκεται σε λίστα
- Όταν μία κινούμενη απόσταση αντικαθίσταται από μία νέα εγγραφή στην *i*-στη λίστα, αυτή πρέπει να ελεγχθεί ξανά έναντι της (*i*+1)-ης λίστας για να βρούμε κατά πόσο θα πρέπει να βρίσκεται σε αυτήν.

3.6. Πειραματική Μελέτη

Οι αλγόριθμοι που περιγράψαμε παραπάνω μπορούν να εφαρμοσθούν σε οποιαδήποτε δομή τύπου R-δέντρου που αποθηκεύει ιστορικά δεδομένα κινούμενων αντικειμένων όπως στο 3D R-δέντρο, στο STR-δέντρο [PJT00] στο TB-δέντρο [PJT00] και το TB*-δέντρο. Μεταξύ αυτών, εφαρμόζουμε τους προτεινόμενους αλγορίθμους στο 3D R-, το TB- και το TB*-δέντρο.

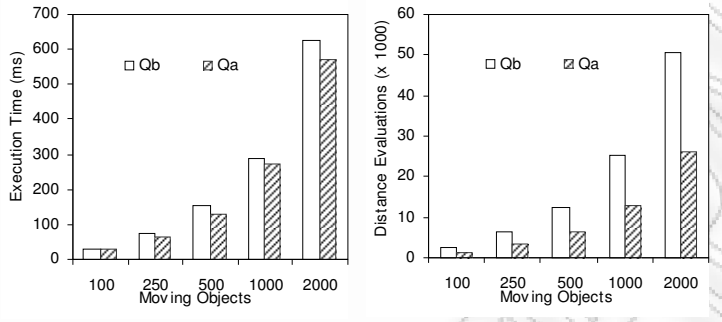
3.6.1. Πειραματικό Πλαίσιο

Όλοι οι αλγόριθμοι εφαρμόσθηκαν πάνω στην υλοποίηση των δομών τύπου R-δέντρου που παρουσιάστηκε στο προηγούμενο κεφάλαιο, χρησιμοποιώντας το περιβάλλον ανάπτυξης Microsoft Visual Basic. Τα πειράματα εκτελέστηκαν σε PC με Microsoft Windows XP, επεξεργαστή AMD Athlon 64 3GHz, 1 GB RAM, αρκετά GB χώρο στο δίσκο, σελίδα μεγέθους 4 KB και μία (μεταβλητού μεγέθους) προσωρινή μνήμη LRU χωρητικότητας 10% του μεγέθους ευρετηρίου, με μέγιστη χωρητικότητα 1000 σελίδες. Τέλος στα πειράματά μας χρησιμοποιήσαμε τα συνθετικά και πραγματικά σύνολα δεδομένων τροχιών που παρουσιάζονται στις παραγράφους 1.5.2 και 1.5.1 αντίστοιχα.

3.6.2. Αποτελέσματα στον Υπολογισμό της *MINDIST*

Για να αποδείξουμε την αποτελεσματικότητα του προτεινόμενου υπολογισμού *MINDIST* σε σχέση με αυτόν που παρουσιάζεται στην [TPS02], κάναμε μία σειρά πειραμάτων εκτελόντας 500 επερωτήσεις στα σύνολα δεδομένων GSTD που δεικτοδοτούνται από το TB-δέντρο χρησιμοποιώντας τον αλγόριθμο `TrajectoryNNSearch`· ωστόσο, παρόμοια αποτελέσματα συλλέγονται και όταν εφαρμόζουμε τον προτεινόμενο τρόπο υπολογισμού στις άλλες δύο εναλλακτικές, δηλαδή, το 3D R- και το TB*-δέντρο. Οι επερωτήσεις εκτελέστηκαν αρχικά με τον προτεινόμενο υπολογισμό *MINDIST*, σχηματίζοντας το σύνολο επερωτήσεων Q_a και κατόπιν με τον υπολογισμό της *MINDIST* που προτείνεται στην [TPS02], σχηματίζοντας το σύνολο επερωτήσεων Q_b. Το σύνολο των 500 αντικειμένων επερώτησης (τροχιών) παρήχθη χρησιμοποιώντας τη γεννήτρια GSTD με αρχική

κατανομή Gauss και μία κατανομή τυχαίας κίνησης. Τέλος, ένα τυχαίο τμήμα κάθε τροχιάς μήκους 1% επί του συνολικού της μήκους χρησιμοποιήθηκε ως τροχιά επερώτησης. Η απόδοση της κάθε επερώτησης μετρήθηκε σε όρους χρόνου εκτέλεσης και αριθμού αξιολογήσεων αποστάσεων μεταξύ σημείου και σημείου, σημείου και γραμμής, και σημείου και MBB.



(a) (b)

Σχήμα 3.15: (a) Χρόνος εκτέλεσης και (b) αριθμός αξιολογήσεων αποστάσεων για σύνολα επερωτήσεων Q_a και Q_b αυξάνοντας τον αριθμό των κινούμενων αντικειμένων

Το Σχήμα 3.15(a) παρουσιάζει το μέσο χρόνο εκτέλεσης για τα σύνολα επερωτήσεων Q_a και Q_b. Σαφώς, ο αλγόριθμος `TrajectoryNNSearch` με την προτεινόμενη βελτίωση στον υπολογισμό *MINDIST* υπερτερεί πάντα του αντίστοιχου υπολογισμού που προτείνεται στην [TPS02], σε όλα τα σύνολα δεδομένων. Η βελτίωση του χρόνου υπολογισμού κυμαίνεται μεταξύ 8% (στο σύνολο δεδομένων GSTD 100) και 17% (στο σύνολο δεδομένων GSTD 250). Η αποδοτικότητα της προτεινόμενης βελτίωσης στον υπολογισμό της *MINDIST* μπορεί να αποδειχθεί περαιτέρω στο Σχήμα 3.15(b), που παρουσιάζει τοπ πλήθος των αξιολογήσεων απόστασης για κάθε εναλλακτικό τρόπο υπολογισμού· το Σχήμα 3.15(b) δείχνει ότι ο προτεινόμενος υπολογισμός της *MINDIST* σε όλες τις περιπτώσεις απαιτεί σχεδόν τους μισούς από τους αντίστοιχους υπολογισμούς που γίνονται όταν εφαρμόζεται η λύση που προτείνεται στην [TPS02].

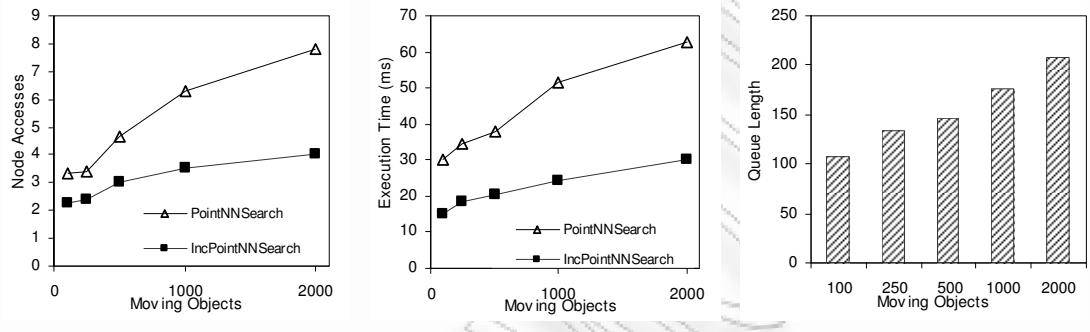
3.6.3. Αποτελέσματα για το Κόστος Αναζήτησης των Ιστορικών μη Συνεχών Αλγορίθμων

Η απόδοση των προτεινόμενων αλγορίθμων μετρήθηκε σε σχέση με τον αριθμό των προσπελάσεων κόμβων και του χρόνου εκτέλεσής τους. Χρησιμοποιήθηκαν αρκετές επερωτήσεις για να αξιολογήσουμε την απόδοση των προτεινόμενων αλγορίθμων στα συνθετικά και πραγματικά δεδομένα. Πιο συγκεκριμένα, χρησιμοποιήσαμε τα ακόλουθα σύνολα επερωτήσεων:

- Q_1: οι αλγόριθμοι `PointNNSearch` και `IncPointNNSearch` αξιολογήθηκαν με ένα σύνολο 500 επερωτήσεων NN αυξάνοντας των αριθμό κινούμενων αντικειμένων στα σύνολα δεδομένων GSTD που δεικτοδοτούνται από TB-, TB*- και 3D R-δέντρο. Οι επερωτήσεις χρησιμοποίησαν ένα τυχαίο σημείο στο 2D χώρο και μία χρονική περίοδο με έκταση το 1% της χρονικής διάστασης του Q_1.
- Q_2: οι αλγόριθμοι `TrajectoryNNSearch` και `IncPointNNSearch` αξιολογήθηκαν με ένα σύνολο 500 επερωτήσεων NN αυξάνοντας τον αριθμό των κινούμενων αντικειμένων στα σύνολα δεδομένων GSTD που δεικτοδοτούνται από TB-, TB*- και 3D R-δέντρο. Το σύνολο των 500 αντικειμένων (τροχιών) επερώτησης παρήχθη χρησιμοποιώντας το GSTD, μια

κατανομή Gauss για την αρχική θέση των σημείων και μία τυχαία κατανομή για την κίνησή τους. Επίσης, στο Q_2 χρησιμοποιήσαμε ένα τυχαίο τμήμα 1% της κάθε τροχιάς ως τροχιά επερώτησης.

- Q_3, Q_4: δύο σύνολα 500 k-NN επερωτήσεων στο σύνολο πραγματικών δεδομένων των Trucks αυξάνοντας τον αριθμό των k με σταθερό χρονικό διάστημα, και αυξάνοντας το μέγεθος του χρονικού διαστήματος (με σταθερό k=1), αντίστοιχα. Για τον αλγόριθμο `PointNNSearch` χρησιμοποιούμε ένα τυχαίο σημείο στο 2D χώρο με το 1% του χρόνου ως περίοδο επερώτησης, ενώ για τον αλγόριθμο `TrajectoryNNSearch` χρησιμοποιήσαμε ένα τυχαίο τμήμα μιας τυχαίας τροχιάς που ανήκει στο σύνολο δεδομένων Buses και καλύπτει χρονικά το 1% του χρόνου.

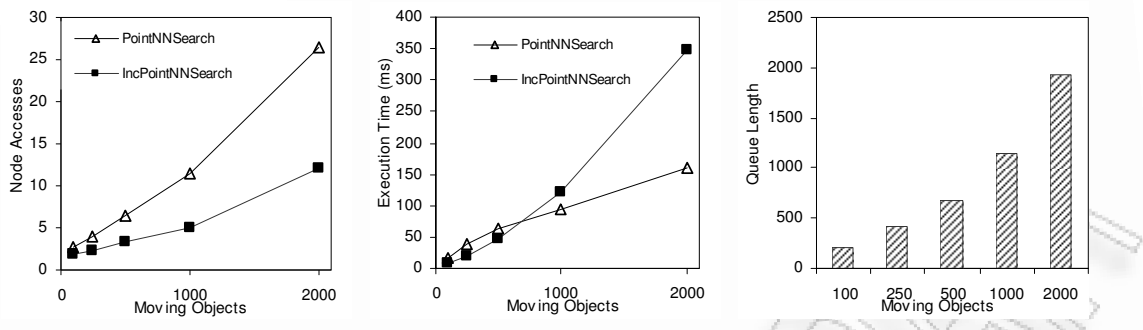


(a) (b) (c)

Σχήμα 3.16: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_1 που εκτελούν αναζήτηση NN σημείου σε 3D R-δέντρα με τα σύνολα δεδομένων GSTD

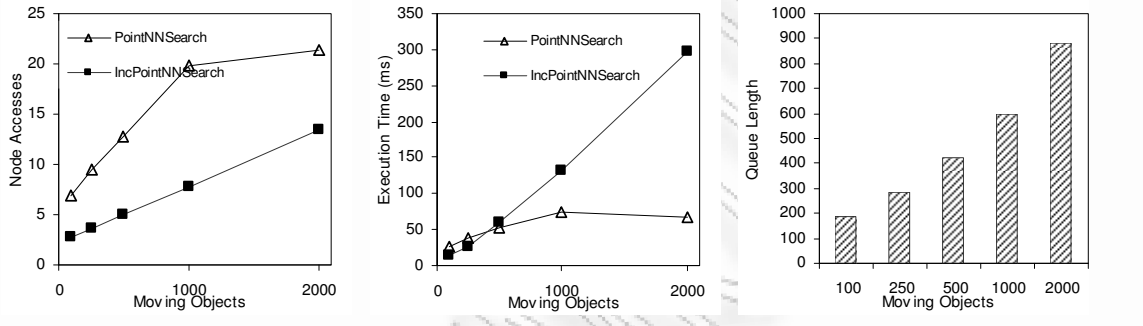
Το Σχήμα 3.16 απεικονίζει τα αποτελέσματα για το σύνολο επερώτησης Q_1 που αξιολογεί τους αλγορίθμους `PointNNSearch` και `IncPointNNSearch` στο 3D R-δέντρο, σε όρους (a) μέσου αριθμού προσπελάσεων κόμβου και (b) μέσου χρόνου εκτέλεσης ανά επερώτηση. Όπως φαίνεται σαφώς, η απόδοση και των δύο αλγορίθμων εξαρτάται υπογραμμικά από το πλήθος του συνόλου δεδομένων, υποβαθμίζεται δε (προσπελαύνει περισσότερες σελίδες) όσο αυξάνεται το πλήθος. Ένα άλλο συμπέρασμα το οποίο προκύπτει από τα ίδια διαγράμματα είναι ότι ο αλγόριθμος `IncPointNNSearch` υπερτερεί του αλγορίθμου `PointNNSearch` σε όλα τα σύνολα δεδομένων, τόσο ως προς τον αριθμό των κόμβων που προσπελαύνονται όσο και ως προς το χρόνο εκτέλεσης. Το Σχήμα 3.16(c) παριστάνει το μέσο μήκος (σε κόμβους) της ουράς που χρησιμοποιείται από τον `IncPointNNSearch` και αυξάνεται γραμμικά με το πλήθος του συνόλου δεδομένων.

Το σύνολο των επερωτήσεων Q_1 που εκτιμά την απόδοση των `PointNNSearch` και `IncPointNNSearch` εκτελέστηκε και στο TB-δέντρο και στο TB*-δέντρο και μας έδωσε τα αποτελέσματα που παρουσιάζονται στο Σχήμα 3.17 και το Σχήμα 3.18, αντίστοιχα. Παρόλο, που ακριβώς όπως αναφέραμε για το 3D R-δέντρο, ο `IncPointNNSearch` υπερτερεί του `PointNNSearch` όσον αφορά το μέσο αριθμό προσπελάσεων κόμβων ανά επερώτηση σε όλα τα σύνολα δεδομένων (Σχήμα 3.17(a) και Σχήμα 3.18(a)), ο πραγματικός χρόνος που απαιτείται για την εκτέλεση της κάθε επερώτησης (Σχήμα 3.17(b) και Σχήμα 3.18(b)) από τον `IncPointNNSearch`, αυξάνεται γρηγορότερα από τον αντίστοιχο χρόνο εκτέλεσης του `PointNNSearch`, που οδηγεί σε υπεροχή του μη αυξητικού αλγόριθμου καθώς αυξάνεται το πλήθος του συνόλου δεδομένων.



(a) (b) (c)

Σχήμα 3.17: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_1 που εκτελούν αναζήτηση NN σημείου σε TB-δέντρα με τα σύνολα δεδομένων GSTD



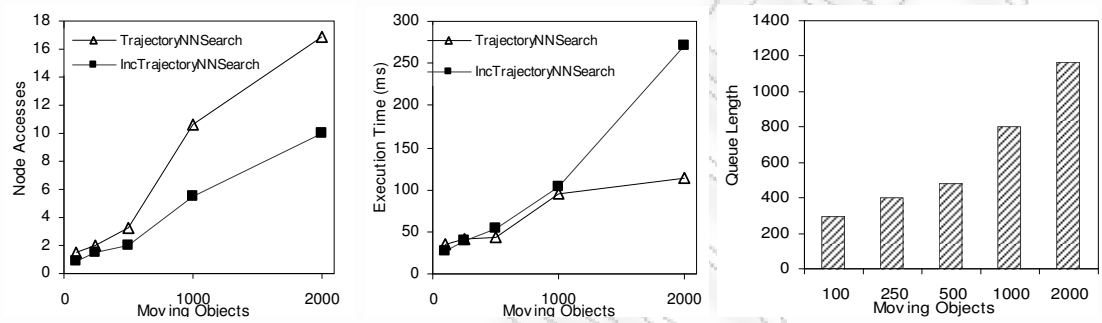
(a) (b) (c)

Σχήμα 3.18: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_1 που εκτελούν αναζήτηση NN σημείου σε TB*-δέντρα με τα σύνολα δεδομένων GSTD

Η ίδια ακριβώς τάση με αυτήν που παρουσιάστηκε για το χρόνο εκτέλεσης του `IncPointNNSearch` παρουσιάζεται στο Σχήμα 3.17(c) και το Σχήμα 3.18(c) που παριστάνουν το μήκος της ουράς που χρησιμοποιήθηκε από τον αντίστοιχο αλγόριθμο. Πιο συγκεκριμένα, ο `PointNNSearch` υπερτερεί του αυξητικού αντίστοιχού του όταν το μέσο μήκος της ουράς υπερβαίνει ένα συγκεκριμένο αριθμό κόμβων (περίπου 400 κόμβους στο σύνολο δεδομένων GSTD 500). Το παραπάνω συμπέρασμα μπορεί να επαληθευτεί επίσης από τα αποτελέσματα του 3D R-δέντρου, όπου το μήκος της ουράς είναι πάντα λιγότερο του 400, και μας οδηγεί σε υπεροχή του αυξητικού αλγορίθμου. Όσον αφορά τη σύγκριση μεταξύ της απόδοσης του TB, του TB* και του 3D R-δέντρου, το τελευταίο υπερβαίνει σε απόδοση των άλλων δύο καθώς αυξάνεται το πλήθος του συνόλου δεδομένων, κάτι αντίστοιχο με αυτό που αναφέρθηκε στην [PJT00] για τις απλές επερωτήσεις εύρους μικρής έκτασης· από την άλλη πάλι, το αρχικό TB-δέντρο φαίνεται να υπερέχει οριακά του TB*-δέντρου που αναπτύξαμε στην παρούσα διατριβή.

Το Σχήμα 3.19 παρουσιάζει τα αποτελέσματα για το σύνολο επερώτησης Q_2 που αξιολογεί την απόδοση των αλγορίθμων `TrajectoryNNSearch` και `IncTrajectoryNNSearch` στο 3D R-δέντρο, σε όρους (a) μέσου αριθμού προσπελάσεων κόμβων, και (b) μέσου χρόνου εκτέλεσης ανά επερώτηση. Η απόδοση και των δύο αλγορίθμων έχει γραμμική εξάρτηση με το πλήθος του συνόλου δεδομένων και υποβαθμίζεται όσο αυτό αυξάνεται. Παρόλο που ο `IncTrajectoryNNSearch` υπερτερεί του `TrajectoryNNSearch` σε όλα τα σύνολα δεδομένων όσον αφορά τον αριθμό των προσπελάσεων κόμβων, ο μέσος χρόνος εκτέλεσης του αυξητικού αλγορίθμου γίνεται μεγαλύτερος

από τον αντίστοιχο χρόνο του μη αυξητικού, όσο αυξάνεται το πλήθος του συνόλου δεδομένων. Το μέσο μήκος ουράς που χρησιμοποιείται από τον `IncTrajectoryNNSearch`, φαίνεται επίσης στο Σχήμα 3.19(c)· βάσει των αποτελεσμάτων για το χρόνο εκτέλεσης του αυξητικού αλγορίθμου, το μήκος της ουράς προτεραιότητας αυξάνεται γραμμικά με το πλήθος του συνόλου δεδομένων. Η διεύρυνση του μήκους της ουράς είναι επίσης υπεύθυνη για την συμπεριφορά που επιδεικνύεται σχετικά με τη σύγκριση του χρόνου εκτέλεσης μεταξύ του `TrajectoryNNSearch` και του αλγορίθμου `IncTrajectoryNNSearch`· καθώς αυξάνεται το μήκος της ουράς, κάθε ενημέρωση γίνεται πιο ακριβή διαδικασία γεγονός που οδηγεί σε υποβάθμιση της απόδοσης του αντίστοιχου αλγορίθμου.



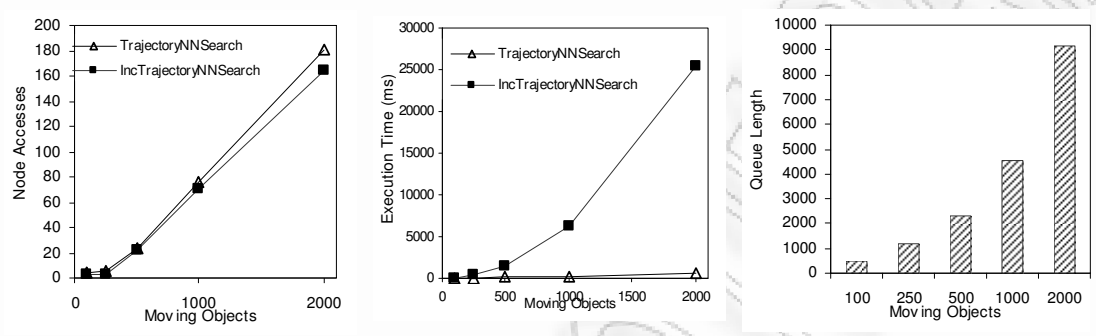
(a) (b) (c)

Σχήμα 3.19: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_2 που εκτελούν αναζήτηση NN τροχιάς σε 3D R -δέντρα με τα σύνολα δεδομένων GSTD

Όσον αφορά τη σύγκριση της απόδοσης των αυξητικών αλγορίθμων που παρουσιάζονται στο Σχήμα 3.16 και στο Σχήμα 3.19 μας οδηγεί στην παρατήρηση ότι ενώ στην πρώτη περίπτωση, λιγότερες προσπελάσεις κόμβων οδηγούν σε μικρότερο χρόνο εκτέλεσης (σε σχέση με τον αυξητικό αλγόριθμο), στη δεύτερη περίπτωση ο χρόνος εκτέλεσης του αυξητικού αλγορίθμου γίνεται μεγαλύτερος από τον αντίστοιχο του μη αυξητικού. Το γεγονός αυτό εξηγείται παρατηρώντας τα αντίστοιχα μήκη της ουράς: στην πρώτη περίπτωση το μήκος ουράς δεν υπερβαίνει τα 200 αντικείμενα (δηλαδή, λιγότερο από μία τυπική *BranchList*), ενώ στη δεύτερη περίπτωση, η ουρά προτεραιότητας περιλαμβάνει χιλιάδες αντικείμενα γεγονός που οδηγεί σε μείωση της απόδοσης του αλγορίθμου.

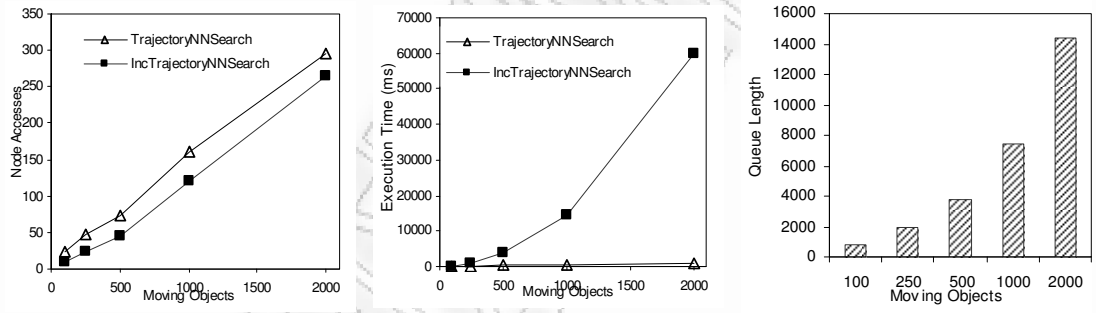
Τα αποτελέσματα του συνόλου επερωτήσεων Q_2 στο TB-δέντρο και το TB*-δέντρο παρουσιάζονται στο Σχήμα 3.20 και Σχήμα 3.21, αντίστοιχα. Ενώ ο `IncTrajectoryNNSearch` υπερτερεί πάντα του `TrajectoryNNSearch` όσον αφορά τον μέσο αριθμό προσπελάσεων κόμβων (Σχήμα 3.20(a) και Σχήμα 3.21(a)), η ανισότητά τους δεν είναι τόσο ουσιαστική όσο αυτή που αναφέρθηκε για το 3D R-δέντρο. Επιπλέον, ο πραγματικός χρόνος εκτέλεσης του αυξητικού αλγορίθμου (Σχήμα 3.20(b) και Σχήμα 3.21(b)) είναι πάντα πολύ μεγαλύτερος του αντίστοιχου χρόνου εκτέλεσης του μη αυξητικού. Τα αποτελέσματα αυτά εξηγούνται από τα εξής δύο γεγονότα. Το πρώτο είναι ότι ο πραγματικός χρόνος εκτέλεσης του αυξητικού αλγορίθμου εξαρτάται σε μεγάλο βαθμό από το αντίστοιχο μήκος της ουράς προτεραιότητας που, όπως φαίνεται στο Σχήμα 3.20(c) και Σχήμα 3.21(c), υπερβαίνει τους 1000 κόμβους για το σύνολο δεδομένων GSTD 250 φθάνοντας τους 9000 κόμβους για το σύνολο δεδομένων GSTD 2000 που δεικτοδοτείται από το TB-δέντρο, ενώ στην περίπτωση του TB*-δέντρου το πλήθος της ουράς αγγίζει πολύ μεγαλύτερες τιμές (14000). Το δεύτερο είναι ότι το TB- και το TB*-δέντρο ομαδοποιούν στο ίδιο φύλλο εγγραφές που ανήκουν στην ίδια

τροχιά, εκμεταλλευόμενα μόνο τη χρονική σειρά με την οποία γίνεται η εισαγωγή των εγγραφών, αγνοώντας ταυτόχρονα οποιαδήποτε σχέση χωρικής εγγύτητας μεταξύ τους. Αυτή η στρατηγική εισαγωγής οδηγεί σε κόμβους με υψηλή χωρική (και χαμηλή χρονική) αλληλοεπικάλυψη, που σημαίνει ότι οι εσωτερικοί κόμβοι συχνά θα διασχίζουν την τροχιά της επερώτησης και η αντίστοιχη *MINDIST* θα είναι μηδέν. Έτσι, θα πρέπει να εξετάσουμε αυτούς τους εσωτερικούς κόμβους διότι η *MINDIST* τους ισούται με μηδέν και οδηγούν εντός της ουράς, με αποτέλεσμα να χάνεται το πλεονέκτημα του αυξητικού αλγορίθμου.



(a) (b) (c)

Σχήμα 3.20: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_2 που εκτελούν αναζήτηση NN τροχιάς σε TB-δέντρα με τα σύνολα δεδομένων GSTD



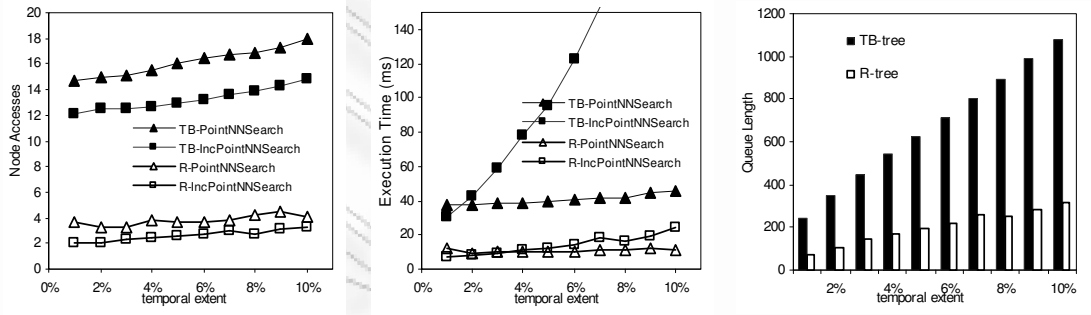
(a) (b) (c)

Σχήμα 3.21: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_2 που εκτελούν αναζήτηση NN τροχιάς σε TB*-δέντρα με τα σύνολα δεδομένων GSTD

Για τους ίδιους λόγους επηρεάζεται και η σύγκριση της απόδοσης μεταξύ του TB-, του TB*- και του 3D R-δέντρου, όπου το τελευταίο υπερτερεί των άλλων δύο καθώς αυξάνεται το πλήθος του συνόλου δεδομένων. Επιπλέον, το πλεονέκτημα του αρχικού TB-δέντρου σε σχέση με το TB*-δέντρο που αποκαλύφθηκε στις σημειακές επερωτήσεις NN, γίνεται σαφέστερο εδώ, όπου το τελευταίο πάντα αποδίδει χειρότερα από το πρώτο. Διαφαίνεται συνεπώς σαφώς ότι η δομή του TB*-δέντρου δεν είναι κατάλληλη για επερωτήσεις NN. Αυτό οφείλεται κυρίως στο γεγονός ότι το TB*-δέντρο περιέχει ευρύτερα MBBs (δεδομένου ότι η χωρητικότητα φύλλου είναι σχεδόν διπλάσια του αρχικού TB-δέντρου), που μας οδηγεί σε υψηλότερη αλληλοεπικάλυψη κόμβων και χαμηλότερη χωρική διακριτοποίηση· η ίδια τάση διαπιστώθηκε και στην αρχική εργασία του [PJT00] για το TB-δέντρο, στην περίπτωση επερωτήσεων εύρους μικρής έκτασης (1% σε κάθε διάσταση ήτοι 0.0001% του συνολικού χώρου), όπου η υψηλή χρησιμοποίηση χώρου του TB-δέντρου γίνεται μειονέκτημα που επηρεάζει την απόδοσή του. Αυτή η ομοιότητα μεταξύ επερωτήσεων μικρού εύρους και πλησιέστερου

γείτονα δικαιολογείται από τα αποτελέσματα της [TZPM04], όπου το κόστος εκτέλεσης επερωτήσεων NN εκτιμάται προσεγγίζοντας τον κύκλο εγγύτητας $C(q, R)$, δηλαδή τον κύκλο εντός του οποίου πραγματοποιείται η αναζήτηση, με κέντρο το σημείο επερώτησης q και ακτίνα R την απόστασή του q από τον k-στο πλησιέστερο γείτονα, με ένα ορθογώνιο εγγύτητας ίσης επιφάνειας. Έτσι, όσο περισσότερα αντικείμενα βρίσκονται στο ευρετήριο, τόσο μικρότερη είναι η ακτίνα του k-στου NN και τόσο μικρότερο ο αντίστοιχος κύκλος εγγύτητας· τέλος το αντίστοιχο της επερώτησης NN αποδεικνύεται [TZPM04] ότι προσεγγίζεται επαρκώς από μία επερώτηση εύρους με μικρή έκταση (και συνολική επιφάνεια ίση με την επιφάνεια του $C(q, R)$). Λόγω των παραπάνω, καθώς επίσης και για λόγους σαφήνειας της παρουσίασης το TB*-δέντρο δεν θα περιλαμβάνεται από εδώ και πέρα στην πειραματική μελέτη για τις ιστορικές επερωτήσεις μη συνεχών NN. Ωστόσο, τα υπόλοιπα πειράματα που διεξήχθησαν επιβεβαιώνουν την τάση που παρατηρήσαμε μέχρι τούδε και καταδεικνύουν ότι το TB*-δέντρο έχει πάντα χειρότερη απόδοση από τους άλλους δύο ανταγωνιστές του.

Η απόδοση των αλγορίθμων για ιστορικές μη συνεχείς επερωτήσεις NN σημείου αυξάνοντας τη χρονική έκταση της επερώτησης, σε όρους μέσου αριθμού προσπελάσεων κόμβων και μέσου χρόνου εκτέλεσης ανά επερώτηση, παρουσιάζεται στο Σχήμα 3.22 για το 3D R-δέντρο και το TB-δέντρο, όπου και τα δύο δεικτοδοτούν το σύνολο δεδομένων Trucks. Σαφώς, και στα δύο ευρετήρια, ο αριθμός των προσπελάσεων κόμβων που απαιτούνται για την επεξεργασία μιας επερώτησης NN, αυξάνεται γραμμικά με την χρονική έκταση της επερώτησης και ο `IncPointNNSearch` είναι πάντοτε υποδεέστερος του `PointNNSearch`. Όσον αφορά στο χρόνο εκτέλεσης και τα δύο ευρετήρια παρουσιάζουν την ίδια συμπεριφορά ως ένα σημείο ισορροπίας όπου ο υπεργραμμικά αυξανόμενος χρόνος εκτέλεσης του `IncPointNNSearch` (συνέπεια του αυξανόμενου μήκους της ουράς που φαίνεται στο Σχήμα 3.22 (c)) ισούται με το γραμμικά αυξανόμενο χρόνο εκτέλεσης του αλγορίθμου `PointNNSearch`. Για το TB-δέντρο, το σημείο ισορροπίας είναι γύρω στο 1.5% της χρονικής έκτασης ενώ στο 3D R-δέντρο αυξάνεται γύρω στο 3.5%.

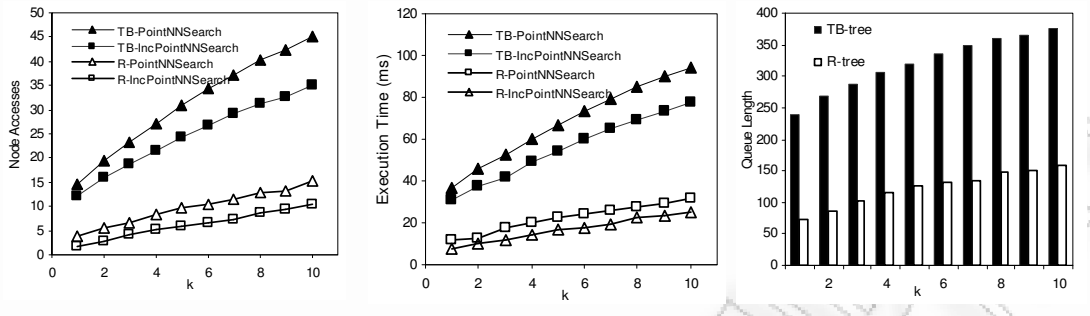


(a) (b) (c)

Σχήμα 3.22: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_3 που εκτελούν αναζήτηση NN σημείου σε 3D R- και TB-δέντρα με τα σύνολα δεδομένων Trucks

Το Σχήμα 3.23 παρουσιάζει το μέσο αριθμό προσπελάσεων κόμβων και του χρόνου εκτέλεσης ανά σημείο επερώτησης, αυξάνοντας τον αριθμό των k, στο σύνολο δεδομένων Trucks που δεικτοδοτείται από το 3D R-δέντρο και το TB-δέντρο. Και στα δύο ευρετήρια είναι σαφές ότι ο αυξητικός αλγόριθμος υπερτερεί του `PointNNSearch` και ως προς το μέσο αριθμό προσπελάσεων και ως προς το χρόνο εκτέλεσης. Χρησιμοποιώντας το 3D R-δέντρο, η απόδοση και των δύο

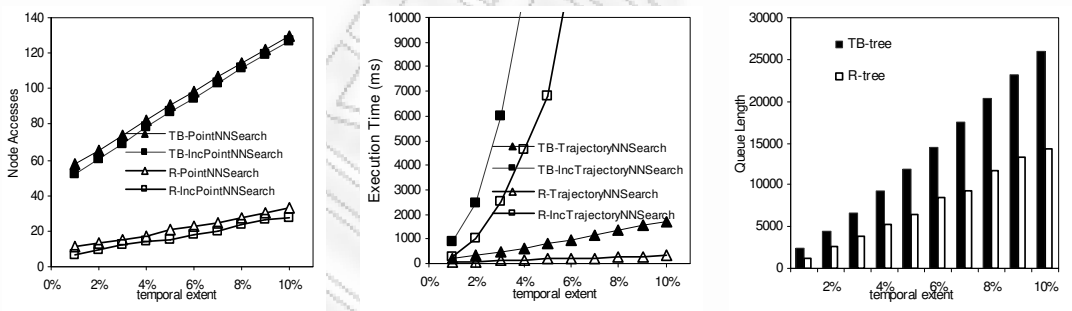
αλγορίθμων επιδεινώνεται γραμμικά από τον αριθμό των *k*, ενώ όταν χρησιμοποιείται το TB-δέντρο η μείωση είναι υπογραμμική.



(a) (b) (c)

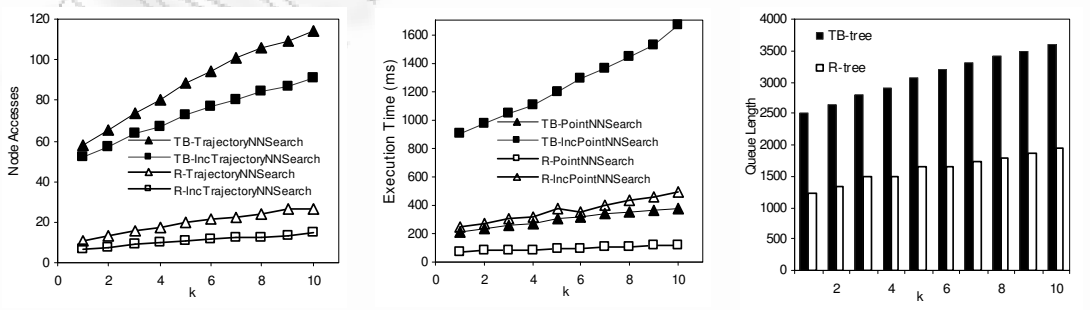
Σχήμα 3.23: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_3 που εκτελούν αναζήτηση σημείου NN σε 3D R- και TB-δέντρα με τα σύνολα δεδομένων Trucks

Τα αποτελέσματα των αλγορίθμων για ιστορικές μη συνεχείς επερωτήσεις NN τροχιάς αυξάνοντας την χρονική έκταση επερώτησης στο 3D R-δέντρο και TB-δέντρο που δεικτοδοτούν το σύνολο δεδομένων Trucks φαίνονται στο Σχήμα 3.24. Για άλλη μια φορά, ο αριθμός των προσπελάσεων κόμβων που απαιτείται για την επεξεργασία μιας επερώτησης NN και με τους δύο αλγορίθμους και στα δύο ευρετήρια, αυξάνεται γραμμικά με την χρονική έκταση της επερώτησης. Ωστόσο, όσον αφορά το χρόνο εκτέλεσης, η απόδοση του αυξητικού αλγορίθμου αυξάνεται υπεργραμμικά με τη χρονική έκταση ως συνέπεια του υπερβολικά μεγάλου μήκους της ουράς προτεραιότητας (Σχήμα 3.24(c)).



(a) (b) (c)

Σχήμα 3.24: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_4 που εκτελούν αναζήτηση NN τροχιάς σε 3D R- και TB-δέντρα με τα σύνολα δεδομένων Trucks



(a) (b) (c)

Σχήμα 3.25: (a) Προσπελάσεις κόμβων, (b) χρόνος εκτέλεσης και (c) μήκος ουράς σε επερωτήσεις Q_4 που εκτελούν αναζήτηση NN τροχιάς σε 3D R-δέντρα με τα σύνολα δεδομένων Trucks

Η απόδοση των αλγορίθμων για ιστορικές μη συνεχείς επερωτήσεις NN τροχιάς αυξάνοντας τον αριθμό των k στο σύνολο δεδομένων Trucks φαίνεται στο Σχήμα 3.25 όπου ο αλγόριθμος `TrajectoryNNSearch` υπερτερεί του αυξητικού αντίστοιχού του σε όρους χρόνου εκτέλεσης, με την αντίστοιχη ουρά να περιλαμβάνει άνω των 1000 κόμβων.

3.6.4. Αποτελέσματα για το Κόστος Αναζήτησης των Ιστορικών Συνεχών Αλγορίθμων

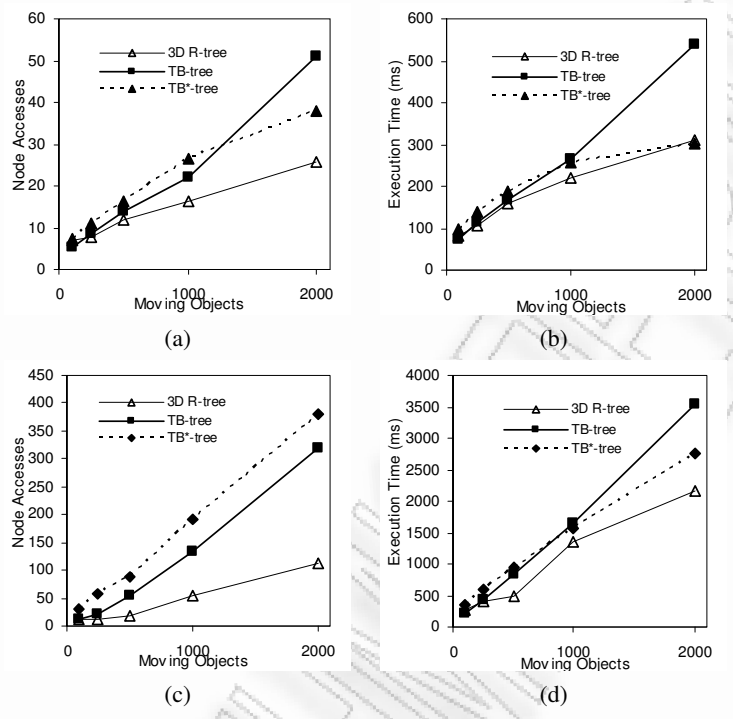
Οι αλγόριθμοι αναζήτησης ιστορικά συνεχούς NN αξιολογήθηκαν ως προς τον αριθμό των προσπελάσεων σε κόμβους και το χρόνο εκτέλεσής τους, με τα ακόλουθα σύνολα επερώτησης:

- Q_5: ο αλγόριθμος `HContPointNNSearch` αξιολογήθηκε με μία σειρά 500 επερωτήσεων NN αυξάνοντας τον αριθμό των κινούμενων αντικειμένων στα σύνολα δεδομένων GSTD που δεικτοδοτούνται από τα TB-, TB* και 3D R-δέντρο όπως αυτό που έγινε για το σύνολο επερώτησης Q_1.
- Q_6: ο αλγόριθμος `HContTrajectoryNNSearch` αξιολογήθηκε με ένα σύνολο 500 επερωτήσεων NN αυξάνοντας τον αριθμό κινούμενων αντικειμένων στα σύνολα δεδομένων GSTD που δεικτοδοτούνται από το TB-, TB* και το 3D R-δέντρο όπως αυτό που έγινε για το σύνολο επερώτησης Q_2.
- Q_7, Q_8: δύο σύνολα 500 k-NN επερωτήσεων στο σύνολο δεδομένων Buses αυξάνοντας τον αριθμό του k με σταθερό χρονικό διάστημα, και αυξάνοντας το μέγεθος του χρονικού διαστήματος (με σταθερό k = 1), αντίστοιχα. Για τον αλγόριθμο `HContPointNNSearch` χρησιμοποιήσαμε ένα τυχαίο σημείο στο 2D χώρο με 1% του συνολικού χρόνου ως περίοδο επερώτησης, ενώ για τον αλγόριθμο `HContTrajectoryNNSearch` χρησιμοποιήσαμε ένα τυχαίο τμήμα μιας τυχαίας τροχιάς που ανήκει στο σύνολο δεδομένων Trucks, που καλύπτει χρονικά το 1% του χρόνου.

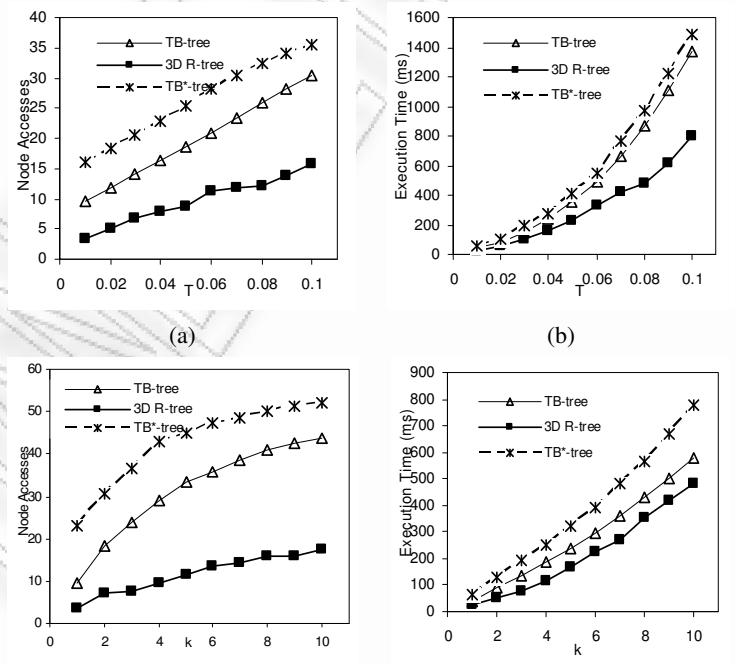
Το Σχήμα 3.26(a) και (b) παρουσιάζει τα αποτελέσματα του αλγορίθμου `HContPointNNSearch` στα σύνολα δεδομένων GSTD αυξάνοντας τον αριθμό των κινούμενων αντικειμένων όσον αφορά (a) το μέσο αριθμό προσπελάσεων κόμβων και (b) το μέσο χρόνο εκτέλεσης ανά επερώτηση. Ακριβώς όπως και στον μη συνεχή αντίστοιχό του, η απόδοση του αλγορίθμου εξαρτάται γραμμικά από το πλήθος του συνόλου δεδομένων και επιδεινώνεται καθώς το πλήθος αυξάνεται, ενώ ο μέσος χρόνος εκτέλεσης και για τα δύο ευρετήρια ακολουθεί την ίδια τάση με τον μέσο αριθμό των κόμβων που εξετάστηκαν. Ένα άλλο αποτέλεσμα που προκύπτει είναι ότι όσο το πλήθος μεγαλώνει, το 3D R-δέντρο υπερτερεί του TB-δέντρου και του TB*-δέντρου, ακολουθώντας την ίδια τάση που παρουσιάζεται στην [PJT00] για απλές επερωτήσεις εύρους μικρής έκτασης. Παρόμοια αποτελέσματα παρουσιάζονται στο Σχήμα 3.26 (c) και (d) όπου ο αλγόριθμος `HContTrajectoryNNSearch` εκτελείται στα σύνολα δεδομένων GSTD.

Η σύγκριση μεταξύ των αλγορίθμων για ιστορική μη συνεχή αναζήτηση του NN με τους συνεχείς ομολόγους τους (δηλαδή Σχήμα 3.16 και Σχήμα 3.17 σε σύγκριση με το Σχήμα 3.26 (a) και (b), και το Σχήμα 3.19 και Σχήμα 3.20 σε σύγκριση με το Σχήμα 3.26(c) και (d)), δείχνει ότι οι συνεχείς αλγόριθμοι είναι πολύ πιο ακριβοί από τους μη συνεχείς. Το συμπέρασμα αυτό είναι αναμενόμενο αφού οι αλγόριθμοι για ιστορική αναζήτηση του συνεχούς NN δεν χρησιμοποιούν μία μόνο απόσταση για το κλάδεμα του χώρου αναζήτησης· αντ' αυτού χρησιμοποιούν μία λίστα

κινούμενων αποστάσεων, που σε γενικές γραμμές αποθηκεύει μεγαλύτερες αποστάσεις από την ελάχιστη απόσταση. Στην πράξη, οι μη συνεχείς αλγόριθμοι κλαδεύουν το χώρο αναζήτησης με την ελάχιστη απόσταση που αποθηκεύεται στη λίστα *Nearests*, εκτελώντας συνεπώς κλάδεμα πιο αποτελεσματικά από το συνεχή αντίστοιχό τους

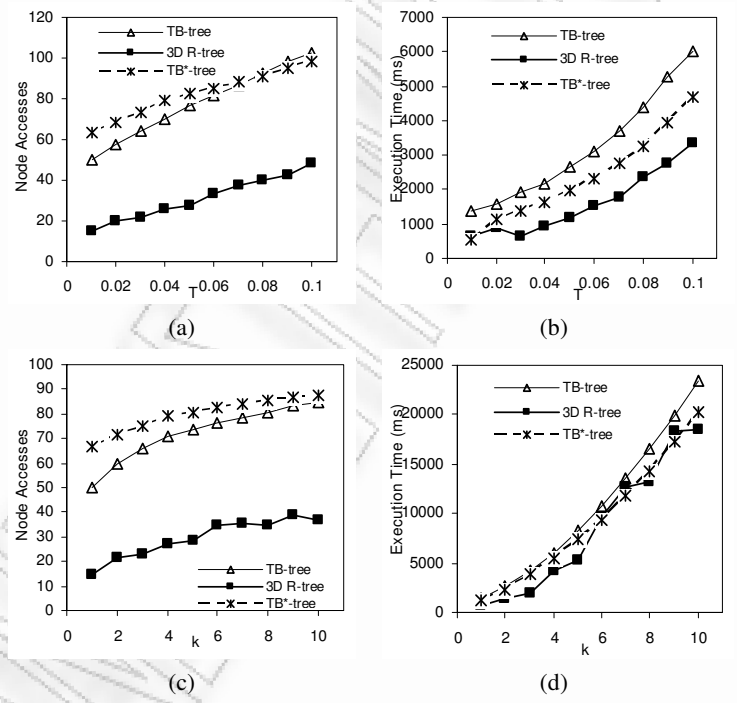


Σχήμα 3.26: Προσπελάσεις κόμβων και χρόνος εκτέλεσης στις επερωτήσεις Q_5 (a, b) και Q_6 (c, d) στο 3D R-, στο TB- και στο TB*-δέντρο αυξάνοντας τον αριθμό των κινούμενων αντικειμένων



Σχήμα 3.27: Προσπελάσεις κόμβων και χρόνος εκτέλεσης στις επερωτήσεις Q_7 (a, b) και Q_8 (c, d) στο ευρετήριο του 3D R-, του TB- και του TB*-δέντρου αυξάνοντας τη χρονική έκταση επερώτησης

Η κλιμάκωση των ιστορικά συνεχών αλγορίθμων με τη χρονική έκταση της επερώτησης παρουσιάζεται στο Σχήμα 3.27. Και οι δύο αλγόριθμοι (`HContPointNNSearch` και `HContTrajectoryNNSearch`) εκτελέστηκαν στο πραγματικό σύνολο δεδομένων Buses που δεικτοδοτείται από το TB-, το TB*- και το 3D R-δέντρο. Από το Σχήμα 3.27(a) και το (c) είναι σαφές ότι η απόδοση και των δύο αλγορίθμων σε όρους προσπελάσεων κόμβων είναι υπογραμμική σε σχέση με τη χρονική έκταση της επερώτησης. Ωστόσο, ο πραγματικός χρόνος εκτέλεσης που απαιτείται από κάθε επερώτηση αυξάνεται υπεργραμμικά με την χρονική έκταση της επερώτησης σα συνέπεια του αυξανόμενου μήκους του αποτελέσματος της επερώτησης (τη λίστα *Nearests*). Η απόδοση των ιστορικά συνεχών αλγορίθμων NN αυξάνοντας τον αριθμό των *k* σε σχέση με το σύνολο δεδομένων Buses που δεικτοδοτούνται από το TB, το TB*- και το 3D R-δέντρο φαίνεται στο Σχήμα 3.28. Όπως προκύπτει από το Σχήμα 3.28 (a) και (c), ο μέσος αριθμός προσπελάσεων κόμβων που απαιτείται για την επεξεργασία μίας *k*-HCNN επερώτησης σημείου ή τροχιάς αυξάνεται υπογραμμικά με το *k*. Όμως, ο πραγματικός χρόνος εκτέλεσης που παρουσιάζεται στο Σχήμα 3.28 (b) και (d) αυξάνεται υπεργραμμικά με το *k*, όπως και με τη χρονική έκταση, σαν συνέπεια του αυξανόμενου μεγέθους του αποτελέσματος της επερώτησης (τις *k* λίστες *Nearests*).



Σχήμα 3.28: Προσπελάσεις κόμβων και χρόνος εκτέλεσης στις επερωτήσεις Q_7 (a, b) και Q_8 (c, d) στο ευρετήριο του 3D R-, του TB- και του TB*-δέντρου αυξάνοντας των αριθμό των *k*

3.6.5. Σύνοψη των Πειραμάτων

Οι περισσότεροι από τους αλγορίθμους που παρουσιάζουμε, σε όρους προσπελάσεων κόμβων, είναι γραμμικοί ή υπογραμμικοί με τις κύριες παραμέτρους της πειραματικής μας μελέτης: το πλήθος του συνόλου δεδομένων, τη χρονική έκταση της επερώτησης και τον αριθμό *k* των ζητούμενων NN. Ωστόσο, ο χρόνος εκτέλεσης των αλγορίθμων `IncPointNNSearch` και `IncTrajectoryNNSearch` φαίνεται να αυξάνεται υπεργραμμικά με τη χρονική έκταση της επερώτησης ως αποτέλεσμα του αυξανόμενου μήκους της ουράς που χρησιμοποιείται, όπως και ο

χρόνος εκτέλεσης των `HContPointNNSearch` και `HContTrajectoryNNSearch`, που έχουν την ίδια τάση σε σχέση με την χρονική έκταση και τον αριθμό του *k*, σαν συνέπεια του αυξανόμενου μήκους της λίστας *Nearests*.

Πίνακας 3.2: Πραγματικός δεικτοδοτημένος χώρος που προσπελαύνεται από κάθε αλγόριθμο NN για το σύνολο δεδομένων GSTD 2000

Αλγόριθμος	*3D R-δέντρο*	*TB-δέντρο*	*TB*-δέντρο*
PointNNSearch	0.006%	0.022%	0.070%
IncPointNNSearch	0.003%	0.010%	0.044%
TrajectoryNNSearch	0.014%	0.148%	0.963%
IncTrajectoryNNSearch	0.008%	0.134%	0.868%
HContPointNNSearch	0.016%	0.042%	0.124%
HContTrajectoryNNSearch	0.053%	0.259%	1.248%

Ο Πίνακας 3.2 συνοψίζει την ισχύ κλαδέματος του αλγόριθμού μας παρουσιάζοντας σε ποσοστιαίες μονάδες το δεικτοδοτούμενο χώρο που προσπελαύνεται προκειμένου να εκτελέσουμε όλους τους προτεινόμενους αλγόριθμους με *k*=1 και χρονική έκταση του 1% του δεικτοδοτούμενου χρόνου. Όπως μπορούμε να συνάγουμε οι αλγόριθμοί μας παρουσιάζουν υψηλή ικανότητα κλαδέματος, περιβάλλοντας καλά το χώρο που θα αναζητηθεί προκειμένου να απαντήσουμε σε επερωτήσεις NN και HCNN, εκτός από την περίπτωση του TB*-δέντρου που, συνολικά, φαίνεται να μην είναι καλή επιλογή για τις επερωτήσεις NN.

3.7. Συμπεράσματα

Οι επερωτήσεις NN είναι στο επίκεντρο της χωρικής και χωροχρονικής αναζήτησης βάσεων δεδομένων στη διάρκεια της τελευταίας δεκαετίας. Η πλειονότητα των αλγορίθμων που επεξεργάζονται τέτοιες επερωτήσεις μέχρι τούδε ασχολούνται είτε με σταθερά είτε με κινούμενα σημεία επερώτησης σε σταθερές ή μελλοντικές (προβλεπόμενες) θέσεις ενός συνόλου συνεχώς κινούμενων σημείων. Στην παρούσα εργασία αναγνωρίζοντας τη συνεισφορά των υφιστάμενων εργασιών, παρουσιάσαμε την πρώτη πλήρη επεξεργασία ιστορικών επερωτήσεων NN σε τροχιές κινούμενων αντικειμένων που αποθηκεύονται σε δομές τύπου R-δέντρου.

Βασιζόμενοι στις πρωτότυπες μετρικές μας που υποστηρίζουν τη στρατηγική διάταξης και κλαδέματος που ακολουθείται κατά την αναζήτηση, παρουσιάσαμε αλγορίθμους που απαντούν στις επερωτήσεις NN και HCNN για σταθερά σημεία επερώτησης ή τροχιές και τα γενικεύσαμε για την αναζήτηση των *k* πλησιέστερων γειτόνων. Οι αλγόριθμοι μπορούν να εφαρμοσθούν στις παραλλαγές του R-δέντρου που δεικτοδοτούν δεδομένα τροχιών, μεταξύ των οποίων χρησιμοποιήσαμε, για τη μελέτη απόδοσης, το 3D R-δέντρο, το TB-δέντρο και το TB*-δέντρο. Πέρα από την υλοποίηση των προτεινόμενων αλγορίθμων σε δομές που ομοιάζουν στο R-δένδρο που χρησιμοποιήθηκαν για τους σκοπούς της πειραματικής μας μελέτης, οι αλγόριθμοι `IncPointNNSearch` and `IncTrajectoryNNSearch` έχουν ήδη υλοποιηθεί στο Αντικειμενο-Σχεσιακό Σύστημα Διαχείρισης Βάσης Δεδομένων (Object-Relational DBMS - ORDBMS) και ενσωματωθεί στη μηχανή HERMES [PFGT08], η οποία έχει επίσης επεκταθεί ώστε να περιλαμβάνει το TB-tree [PJT00].

Για να μετρήσουμε την απόδοση των αλγορίθμων μας κάναμε μία εκτεταμένη πειραματική μελέτη σε συνθετικά και πραγματικά σύνολα δεδομένων. Καταρχήν, αποδείξαμε ότι η βελτίωσή μας στον υπολογισμό της *MINDIST* μπορεί να αυξήσει επαρκώς την απόδοση των προτεινόμενων αλγορίθμων. Για τους ιστορικούς μη συνεχείς αλγορίθμους, αποδείχθηκε ότι ενώ η αυξητική (πρώτα στον καλύτερο) προσέγγιση είναι πάντα λιγότερο δαπανηρή από τη μη αυξητική (πρώτα στο βαθύτερο) αντίστοιχη σε όρους προσπελάσεων κόμβων, ο πραγματικός χρόνος εκτέλεσης εξαρτάται από το μήκος της χρησιμοποιούμενης ουράς προτεραιότητας. Σε γενικές γραμμές, η προσέγγιση «πρώτα στον καλύτερο» υπερτερεί του ανταγωνιστή της μόνο για σημειακές επερωτήσεις NN σε μικρή χρονική έκταση (μικρότερη του 2-4% ανάλογα με το ευρετήριο που χρησιμοποιείται και για οποιοδήποτε k), ενώ σε όλες τις υπόλοιπες περιπτώσεις η προσέγγιση «πρώτα στο βαθύτερο» απαιτεί λιγότερο χρόνο εκτέλεσης. Αυτό το μειονέκτημα του αυξητικού αλγορίθμου οφείλεται κυρίως στο μήκος της ουράς προτεραιότητας που μπορεί να γίνει τεράστιο, ιδιαίτερα στην περίπτωση του TB-δέντρου και του TB*-δέντρου. Όσον αφορά τη σύγκριση μεταξύ των χρησιμοποιούμενων ευρετηρίων, το 3D R-δέντρο υπερτερεί του TB-δέντρου σε όρους προσπελάσεων κόμβων και χρόνου εκτέλεσης, ενώ διαπιστώνουμε ότι το προτεινόμενο TB*-δέντρο της παρούσας διατριβής δεν είναι η πλέον κατάλληλη επιλογή όταν ασχολούμαστε με επερωτήσεις NN.

4. Προηγμένη Επεξεργασία Επερωτήσεων Τροχιών: Αναζήτηση Ομοιότητας

Σκοπός του κεφαλαίου αυτού είναι να παρουσιάσουμε τους αλγορίθμους για αναζήτηση ομοιότητας σε δομές τύπου R-δέντρου που αποθηκεύουν ιστορικές τροχιές κινούμενων αντικειμένων. Η δομή του κεφαλαίου έχει ως εξής: η Ενότητα 4.1 αιτιολογεί το κεφάλαιο και παρουσιάζει τις αρχικές ιδέες. Οι υπάρχουσες σχετικές εργασίες παρουσιάζονται στην Ενότητα 4.2. Η Ενότητα 4.3 εισάγει επισήμως πλέον τον κύριο στόχο του κεφαλαίου αυτού και εξετάζει ενδελεχώς τις μετρικές που χρησιμοποιούνται για αναζήτηση της ομοιότερης τροχιάς (Most Similar Trajectory - MST), καθώς και αυτές που χρησιμοποιούνται για την να υποστηρίξουμε τις στρατηγικές αναζήτησης κλαδέματος και διάταξης. Οι Ενότητα 4.4 αποτελεί τον πυρήνα του κεφαλαίου περιγράφοντας λεπτομερώς τους αλγορίθμους για την εκτέλεση της αναζήτησης MST σε ιστορικά δεδομένα τροχιάς· οι αλγόριθμοι που παρουσιάζονται βασίζονται στο παράδειγμα «πρώτα στο βαθύτερο» και «πρώτα στον καλύτερο», χρησιμοποιώντας δομές τύπου R-δέντρου. Η Ενότητα 4.5 παρουσιάζει τα αποτελέσματα της πειραματικής μελέτης ενώ τα συμπεράσματά μας παρουσιάζονται στην Ενότητα 4.6.

4.1. Εισαγωγή

Ένας επιπλέον ενδιαφέρων τύπος επερώτησης που είναι χρήσιμος στην αναζήτηση σε MODs προκύπτει από το λεγόμενο πρόβλημα *ομοιότητας τροχιάς*, που στόχο έχει την ανεύρεση «παρόμοιων» τροχιών κινούμενων αντικειμένων. Για να κατανοήσουμε το πρόβλημα, ας δούμε το εξής παράδειγμα. Έστω ότι το δίκτυο του μετρό μιας πόλης πρόσφατα επεκτάθηκε και έθεσε σε λειτουργία μια νέα γραμμή, για να εξυπηρετήσει ένα μεγάλο τμήμα των προαστίων. Η επέκταση του δικτύου του μετρό προϋποθέτει τον επανασχεδιασμό του υφιστάμενου δικτύου μεταφοράς (λεωφορεία, τραμ, τρόλεϊ κλπ.). Θα μπορούσαμε συνεπώς να συνδράμουμε τους εμπειρογνώμονες στον τομέα αν τους δίναμε τη δυνατότητα να θέσουν επερωτήσεις ως προς την ομοιότητα μεταξύ των τροχιών των υπαρχόντων μέσων μεταφοράς και της νέας γραμμής μετρό. Έτσι, θα είναι σε θέση, για παράδειγμα, να αλλάξουν το ωράριο μιας γραμμής λεωφορείου αν κάποια μέρα συμπίπτει με το ωράριο της νέας γραμμής μετρό, ή ακόμα και να την καταργήσουν. Για να αντιμετωπίσουν τέτοιες επερωτήσεις αποτελεσματικά, τα συστήματα MOD θα πρέπει να περιλαμβάνουν μεθόδους για να απαντούν στη λεγόμενη αναζήτηση *Ομοιότερης Τροχιάς* (MST) που αναφέρεται και στην [The03].

Η αναζήτηση ομοιότητας τροχιάς είναι ένα σχετικά νέο θέμα στη βιβλιογραφία· η πλειονότητα των μεθόδων που προτείνονται μέχρι τώρα βασίζονται είτε στο πλαίσιο της ανάλυσης χρονοσειρών και

το μοντέλο της Μεγαλύτερης Κοινής Υποακολουθίας (Longest Common Subsequence - LCSS) [VKG02], ή στην πρόσφατα προταθείσα Απόσταση Επεξεργασίας σε Πραγματικές Ακολουθίες (Edit Distance on Real Sequence - EDR) [COO05]. Ωστόσο, όλες αυτές οι μέθοδοι έχουν το βασικό μειονέκτημα ότι είτε αγνοούν τη χρονική διάσταση της κίνησης, υπολογίζοντας συνεπώς τη χωρική (και όχι τη χωροχρονική) ομοιότητα μεταξύ τροχιών, ή υποθέτουν ότι οι τροχιές είναι του ιδίου μήκους και έχουν τον ίδιο ρυθμό δειγματοληψίας. Για να δώσουμε ένα παράδειγμα του προβλήματος που προκύπτει όταν έχουμε διαφορετικούς ρυθμούς δειγματοληψίας, ας θυμηθούμε το Σχήμα 1.4 που παρουσιάζει δύο τροχιές T και Q με τη θέση τους να δειγματοληπτείται με διαφορετικούς ρυθμούς. Ενώ η Q και η T δειγματοληπτούνται 4 και 32 φορές αντίστοιχα, έχουν περίπου το ίδιο μήκος διασχίζοντας την ίδια περιοχή. Παρόλο που οι δύο τροχιές είναι προφανώς παρόμοιες, οι μέθοδοι που βασίζονται στο μοντέλο LCSS ή το EDR δεν μπορούν να ανιχνεύσουν αυτής της μορφής την ομοιότητα καθώς προσπαθούν να συνταιριάξουν τις θέσεις τροχιάς μία προς μία, γεγονός το οποίο σαφώς δεν συμβαίνει στο παραπάνω παράδειγμα (του πραγματικού κόσμου). Επιπλέον, η πλειονότητα των προτεινόμενων προσεγγίσεων εκμεταλλεύονται εξειδικευμένες δομές ευρετηρίου για να κλαδέψουν το χώρο αναζήτησης και να ανακτήσουν την ομοιότερη τροχιά σε μία τροχιά επερώτησης.

Η πρόκληση λοιπόν που έχουμε δεχτεί στην παρούσα διατριβή, είναι να υποστηρίξουμε αποτελεσματικά την αναζήτηση της k-οστής ομοιότερης τροχιάς (k-MST) σε MODs που αποθηκεύουν ιστορικές πληροφορίες τροχιών, και δεικτοδοτούνται από δομές τύπου R-δέντρου. Οι κύριες συνεισφορές του κεφαλαίου είναι οι εξής:

- Ορίζεται μία μετρική ανομοιότητας (*DISSIM*) για τη μέτρηση της χωροχρονικής ανομοιότητας μεταξύ δύο τροχιών· η μετρική αυτή μπορεί να θεωρηθεί ως η μέση απόσταση μεταξύ των δύο τροχιών συν τω χρόνω, και έχει επίσης ανεξάρτητα προταθεί από την [NP06]. Κατόπιν προτείνουμε μία αποτελεσματική μέθοδο προσέγγισης για να υπερβούμε το δαπανηρό υπολογισμό της.
- Προτείνεται ένα σύνολο πρωτότυπων μετρικών (*MINDISSIM, PESDISSIM,* OPTDISSIM) καθώς και αρκετά σχετικά λήμματα που στη συνέχεια χρησιμοποιούνται για κλάδεμα από τους δύο αλγορίθμους αναζήτησης ομοιότερης τροχιάς. Πιο συγκεκριμένα, χρησιμοποιώντας αυτές τις μετρικές, προτείνουμε ένα αλγόριθμο επεξεργασίας επερώτησης «πρώτα στο βαθύτερο» και έναν «πρώτα στον καλύτερο» για να εκτελέσουμε αναζήτηση k-MST σε δομές τύπου R-δέντρου που αποθηκεύουν ιστορικές πληροφορίες τροχιών.
- Διεξάγουμε μία πλήρη σειρά πειραμάτων σε μεγάλα συνθετικά και πραγματικά σύνολα δεδομένων αποδεικνύοντας ότι οι αλγόριθμοι είναι υψηλής κλιμάκωσης και αποτελεσματικοί σε όρους προσπελάσεων κόμβων, χρόνου εκτέλεσης και κλαδεμένου χώρου. Αποδεικνύουμε δε ότι η προτεινόμενη μετρική ομοιότητας (DISSIM) ανακτά χωροχρονικά παρόμοιες τροχιές με αποτελεσματικό τρόπο σε περιπτώσεις που ανάλογες εργασίες αποτυγχάνουν.

Θα πρέπει, δε και πάλι, να τονίσουμε ότι όλοι οι προτεινόμενοι αλγόριθμοι δεν απαιτούν μια συγκεκριμένη δομή ευρετηρίου και μπορούν να εφαρμοσθούν απευθείας σε οποιοδήποτε μέλος της οικογένειας του R-δέντρου χρησιμοποιείται για τη δεικτοδότηση τροχιών, όπως το 3D R-δέντρο, το TB-δέντρο [PJT00] και το TB*-δέντρο που προτείνεται σε αυτή τη διατριβή. Εξ' όσων γνωρίζουμε, η πρόταση αυτής της διατριβής είναι η πρώτη που παρέχει τεχνικές προκειμένου ένα χωροχρονικό

ευρετήριο να υποστηρίζει τόσο κλασσικές επερωτήσεις εύρους, τοπολογικές όσο και βάσει ομοιότητας.

4.2. Σχετικές Εργασίες

Η αναζήτηση ομοιότητας έχει μελετηθεί αρκετά στον τομέα της ανάλυσης χρονοσειρών. Ως μέτρο προσεγγιστικού ταιριάσματος, οι Agrawal et al. [AFS93] πρότειναν τη χρήση του Διακριτού Μετασχηματισμού Fourier (Discrete Fourier Transformation) (DFT). Μία εναλλακτική μέθοδος ταιριάσματος χρονοσειρών μέσω μείωσης διαστάσεων προτάθηκε από τους Chan και Fu [CF99], χρησιμοποιώντας το Διακριτό Μετασχηματισμό Κυματιδίων (Discrete Wavelet Transformation) (DWT). Για να συγκρίνουν ακολουθίες με διαφορετικά μήκη, οι Berndt και Clifford [BC96] χρησιμοποίησαν την τεχνική Δυναμικής Παραμόρφωσης Χρόνου (Dynamic Time Warping - DTW) που επέτρεψε στις ακολουθίες να τεντωθούν κατά μήκος του χρονικού άξονα για την ελαχιστοποίηση της απόστασης μεταξύ των ακολουθιών. Παρά το γεγονός ότι η DTW προϋποθέτει πολύ υψηλό κόστος υπολογισμού, είναι πολύ πιο εύρωστη ως προς το θόρυβο.

Στην [YAS03] παρουσιάστηκε μία μέθοδος δεικτοδότησης για την επεξεργασία επερωτήσεων ομοιότητας βάσει σχήματος (shape-based) σε βάσεις δεδομένων τροχιών. Η προτεινόμενη μέθοδος βασίστηκε στην Ευκλείδεια Απόσταση. Ωστόσο μπορεί να εφαρμοσθεί μόνο σε τροχιές ίδιου μήκους που ισχύουν για το ίδιο χρονικό διάστημα. Οι Cai και Ng [CN04] πρότειναν τη χρήση των πολυωνύμων Chebyshev για την προσέγγιση και τη δεικτοδότηση τροχιών για ταίριασμα ομοιότητας. Και πάλι όμως, η μέθοδος αυτή έπασχε από την απαίτηση ότι οι τροχιές πρέπει να είναι του ιδίου μήκους (σε όρους αριθμού χωροχρονικών σημείων από τα οποία αποτελούνται).

Οι Vlachos et al. [VGD04] παρουσίασαν ένα μέτρο της απόστασης που μας επιτρέπει να βρούμε παρόμοιες τροχιές με μετασχηματισμούς μετάφρασης, κλιμάκωσης και περιστροφής. Το πρώτο βήμα της μεθόδου τους ήταν ο μετασχηματισμός κάθε τροχιάς σε ένα περιστροφικά αμετάβλητο χώρο. Για τον υπολογισμό της απόστασης μεταξύ δύο τροχιών στο νέο χώρο αμετάβλητης περιστροφής, χρησιμοποιήθηκε η τεχνική DTW.

Οι Sakurai et al. [SYF05] πρότειναν μία βελτιωμένη εκδοχή της DTW, τη μέθοδο Ταχείας αναζήτησης για Δυναμική Παραμόρφωση Χρόνου (Dynamic Time Warping) (FTW), βάσει μιας νέας μικρότερης περιοριστικής μετρικής για την προσέγγιση της απόστασης παραμόρφωσης χρόνου. Απέδειξαν ότι η FTW μπορεί να κλαδέψει ένα σημαντικό τμήμα του χώρου αναζήτησης, με αποτέλεσμα σημαντική μείωση στο κόστος αναζήτησης.

Προσφάτως, οι Lin και Su [LS05] μελέτησαν το πρόβλημα της αναζήτησης ομοιότητας ανεξάρτητα του χρόνου στις τροχιές κινούμενων αντικειμένων. Η μονόδρομη συνάρτηση απόστασης (One Way Distance - OWD) εισάγεται για τη σύγκριση των χωρικών σχημάτων των τροχιών μαζί με τους ανάλογους αλγόριθμους για τον υπολογισμό της OWD. Η πειραματική μελέτη τους καταδεικνύει ότι η χρήση της συνάρτησης OWD υπερτερεί του αλγορίθμου DTW ως προς την ακρίβεια και την απόδοση.

Αρκετές προσεγγίσεις βασίζονται στο μέτρο ομοιότητας της Μεγαλύτερης Κοινής Υποακολουθίας (Longest Common Sub Sequence - LCSS). Η μετρική LCSS ταιριάζει δύο ακολουθίες επιτρέποντας τους να τεντωθούν, χωρίς αναδιάταξη της ακολουθίας των στοιχείων, αλλά επιτρέποντας

να μην ταιριαστούν κάποια στοιχεία (που είναι το βασικό πλεονέκτημά της σε σύγκριση με την Ευκλείδεια απόσταση και την DTW). Συνεπώς, η LCSS μπορεί να αντιμετωπίσει αποτελεσματικά σκοπέλους, θόρυβο, και διάφορους συντελεστές κλιμάκωσης. Οι Vlachos et al. [VKG02] υιοθέτησαν τη χρήση της μεθόδου LCSS. Μέσω εισαγωγής δύο μέτρων ομοιότητας που επιτρέπουν τέντωμα στο χρόνο και μεταφράσεις αντίστοιχα, οι γράφοντες πρότειναν μη μετρικές συναρτήσεις ομοιότητας, που ήταν ιδιαίτερα ανθεκτικές παρουσία θορύβου και μας έδωσαν μια ευφυή έννοια της ομοιότητας μεταξύ τροχιών προσδίδοντας περισσότερο βάρος στα παρόμοια τμήματα των τροχιών. Επιπλέον, παρουσιάστηκε μία αποτελεσματική δομή ευρετηρίου (που βασίζεται στην ιεραρχική συσταδοποίηση) για επερωτήσεις ομοιότητας. Ωστόσο, όπως θα αποδείξουμε στην πειραματική μας μελέτη, η προτεινόμενη μέθοδος πάσχει όταν οι τροχιές έχουν διαφορετικούς ρυθμούς δειγματοληψίας.

Στην [COO05] παρουσιάστηκε ένα μέτρο απόστασης, που καλείται Απόσταση Επεξεργασίας σε Πραγματικές Ακολουθίες (Edit Distance on Real Sequences - EDR). Αυτό το μέτρο απόστασης, που βασίζεται στη απλή απόσταση επεξεργασίας (edit distance), αποδείχθηκε ότι είναι πιο σθεναρό από τις DTW και LCSS σε τροχιές με θόρυβο. Η αποτελεσματικότητα αυτού του μέτρου απόστασης βελτιώθηκε από την εφαρμογή τριών στρατηγικών κλαδέματος, που μείωσε το αντίστοιχο υπολογιστικό κόστος για υπολογισμούς μεταξύ της επερώτησης και των τροχιών δεδομένων χωρίς την εισαγωγή ψευδών απορρίψεων. Από την άλλη πλευρά, όπως και στην LCSS, η EDR καθορίζει χωρική ομοιότητα μόνο, αγνοώντας το χρόνο, ενώ οι τροχιές με διαφορετικούς ρυθμούς δειγματοληψίας δεν μπορούν να αντιμετωπισθούν αποτελεσματικά, όπως θα αποδειχθεί στην πειραματική μελέτη. Επιπλέον, τόσο οι [VKG02] όσο και [COO05] προτείνουν τη χρήση ευρετηρίων ειδικά για το κλάδεμα του χώρου αναζήτησης για να υποστηρίξουν αποτελεσματικά την αναζήτηση *k*-MST.

Οι Keogh et al. [KWX+06] παρουσίασαν έναν αλγόριθμο (βάσει της συνάρτησης *LB_Keogh* που εισήχθη στην [Keo02]), που μείωσε δραματικά τη χρονική πολυπλοκότητα του υπολογισμού της μέτρησης Ευκλείδειας απόστασης. Η επιτάχυνση αυτή έγινε μεγαλύτερη όταν επετράπη η δεικτοδότηση. Ωστόσο, ο ανωτέρω αλγόριθμος, που γενικεύθηκε σε άλλα μέτρα απόστασης, όπως το DTW και το LCSS, μπορούσε να εφαρμοσθεί μόνο σε 2D σχήματα.

Προσφάτως, οι Pelekis et al. [PKM+07] ασχολήθηκαν με το πρόβλημα της ομοιότητας τροχιών κάτω από μία διαφορετική οπτική. Αντίθετα με άλλες εργασίες που χρησιμοποιούν μία γενικευμένη μετρική ομοιότητας που αγνοεί την χρονική διάσταση, εισάγουν ένα πλαίσιο αναφοράς αποτελούμενο από ένα σύνολο τελεστών απόστασης που βασίζονται σε πρωτογενείς (χώρος και χρόνος), όσο και σε δευτερογενείς παραμέτρους των τροχιών (όπως η ταχύτητα και η απόσταση). Ως αποτέλεσμα, οι [PKM+07] καθορίζουν διαφορετικές μετρικές ομοιότητας για κάθε είδος ομοιότητα μεταξύ τροχιών: χωρική, χρονική, χωροχρονική, βασισμένη στην ταχύτητα και ομοιότητα κατεύθυνσης. Η καινοτομία της προσέγγισης βασίζεται όχι μόνο στην παροχή των ποιοτικά διαφορετικών μέσων αποτίμησης της ομοιότητας μεταξύ τροχιών, αλλά επιπλέον, στην υποστήριξη της συσταδοποίησης τροχιών και εργασιών εξόρυξης και κατηγοριοποίησης, οι οποίες σαφώς απαιτούν έναν τρόπο ποσοτικοποίησης της απόσταση μεταξύ των τροχιών. Για κάθε μία από τις προτεινόμενες αποστάσεις παρέχονται παραμετρικοί αλγόριθμοι, των οποίων η απόδοση εκτιμάται από μία εκτενή πειραματική μελέτη.

Αναγνωρίζοντας τη συνεισφορά των παραπάνω προτάσεων, στη συνέχεια προτείνουμε πρωτότυπες μετρικές και αλγορίθμους για την αναζήτηση *k*-MST σε δομές τύπου R-δέντρου.

Πίνακας 4.1: Πίνακας συμβόλων

Σύμβολο	*Περιγραφή*
D	βάση δεδομένων τροχιών
O_i	ταυτότητα κινούμενου αντικειμένου
T, Q	δεικτοδοτημένη τροχιά και τροχιά επερώτησης
T_k, Q_k	το k-οστό γραμμικό τμήμα της τροχιάς T ή Q
x_k, y_k, t_k	οι συντεταγμένες της τροχιάς T σε ένα χρονικό αποτύπωμα t_k
$Dist_{Q,T}(t)$	χρονικές συναρτήσεις της Σύγχρονης Ευκλείδειας Απόστασης μεταξύ τροχιών Q και T
a, b, c	συντελεστές του τριωνύμου $Dist_{Q,T}(t)$
$E_{Q,T}$	Σφάλμα υπολογισμού της ανομοιότητας μεταξύ τροχιών
$Dist$	Ευκλείδια απόσταση μεταξύ τροχιών
V	σχετική ταχύτητα μεταξύ κινούμενων αντικειμένων
N	κόμβος R-δέντρου
$MINDIST(Q,N)$	ελάχιστη απόσταση μεταξύ Q και N
V_{max}	το άθροισμα της μέγιστης ταχύτητας των δεικτοδοτημένων τροχιών συν τη μέγιστη ταχύτητα της τροχιάς επερώτησης
S_R	το σύνολο των γραμμικών τμημάτων που έχουν ήδη ανακτηθεί από το ευρετήριο
S_C	το σύνολο των τροχιών με γραμμικά τμήματα που έχουν ήδη ανακτηθεί από το ευρετήριο αλλά δεν έχουν ακόμα ολοκληρωθεί στη δεδομένη χρονική περίοδο.

4.3. Διατύπωση Προβλήματος και Μετρικές για Αναζήτηση Ομοιότερης Τροχιάς

Σ' αυτή την ενότητα ορίζεται η έννοια των επερωτήσεων Ομοιότερης Τροχιάς σε σχέση με ένα μέτρο ανομοιότητας και κατόπιν εισάγεται και τυπικά η έννοια της χωροχρονικής ανομοιότητας που χρησιμοποιείται στην προσέγγιση της παρούσας διατριβής. Τέλος καθορίζεται μια σειρά μετρικών και ευριστικών για αναζήτηση MST που χρησιμοποιούνται στους αλγόριθμους που παρουσιάζονται στην διατριβή. Ο Πίνακας 4.1 παρουσιάζει τα σύμβολα που χρησιμοποιούνται στο υπόλοιπο του κεφαλαίου.

4.3.1. Διατύπωση Προβλήματος

Έστω D μια βάση δεδομένων N κινουμένων αντικειμένων με ταυτότητες αντικειμένων $\{O_1, O_2, \ldots, O_N\}$ υποθέτοντας γραμμική παρεμβολή μεταξύ των δειγματοληπτούμενων σημείων τους. Η τροχιά T ενός κινούμενου αντικειμένου O_i αποτελείται από n-1 3D γραμμικά τμήματα $\{T_1, T_2, .. T_{n-1}\}$. Κάθε 3D γραμμικό τμήμα T_k είναι της μορφής $((x_k, y_k, t_k), (x_{k+1}, y_{k+1}, t_{k+1}))$, όπου $t_0 \le t_k < t_{k+1} \le now$. Λαμβάνοντας υπ' όψιν ότι έχουν προταθεί πόλλα μέτρα ομοιότητας στη βιβλιογραφία όπως συζητήθηκε στην προηγούμενη ενότητα, ο ορισμός μιας επερώτησης MST θα πρέπει να είναι όσο το δυνατόν πιο γενικός. Συνεπώς και τύποις ορίζουμε την αναζήτηση MST ως ανεξάρτητη του υποκείμενου μέτρου ομοιότητας:

Ορισμός 4.1: *Αν έχουμε μία τροχιά επερώτησης Q, μία βάσης δεδομένων τροχιών D και ένα μέτρο DSIM που μετρά την ανομοιότητα μεταξύ δύο τροχιών, η επερώτηση ομοιότερης τροχιάς είναι η επερώτηση*

$$MST(D,Q) = \left(T, DSIM(Q,T)\right) : DSIM(Q,T) \le DSIM(Q,T') \forall T' \in D \quad (4.1)$$

που αναζητά στη βάση δεδομένων D την τροχιά T που έχει την ελάχιστη ανομοιότητα με την τροχιά επερώτησης Q μεταξύ όλων των τροχιών στη D, καθώς και την ανάλογη τιμή ανομοιότητας.

Ωστόσο, όσον αφορά στο μέτρο ομοιότητας, η πλειονότητα των σχετικών εργασιών στον τομέα της αναζήτησης ομοιότητας τροχιάς, είτε αγνοεί τη χρονική διάσταση της κίνησης, υπολογίζοντας τη χωρική ομοιότητα μεταξύ τροχιών είτε υποθέτει ότι οι τροχιές έχουν τα ίδια μήκη (σε όρους του αριθμού των χωροχρονικών σημείων από τα οποία αποτελούνται) και τον ίδιο ρυθμό δειγματοληψίας. Για να υπερβούμε αυτά τα προσκόμματα, μπορούμε να γενικεύσουμε τη πασίγνωστη μετρική Ευκλείδειας Απόστασης και να δώσουμε την έννοια της χωροχρονικής *ανομοιότητας* μεταξύ δύο τροχιών *T* και *Q* που ισχύουν και οι δύο στη διάρκεια ενός συγκεκριμένου χρονικού διαστήματος [t_1, t_n], ολοκληρώνοντας την Ευκλείδεια απόσταση τους στο χρόνο.

Ορισμός 4.2*: Η ανομοιότητα DISSIM(Q, T) μεταξύ των τροχιών Q και T που ισχύουν κατά τη χρονική περίοδο [t_1, t_n] ορίζεται ως το ορισμένο ολοκλήρωμα της χρονικής συνάρτησης της Ευκλείδειας απόστασης μεταξύ των δύο τροχιών κατά την ίδια χρονική περίοδο:*

$$DISSIM\left(Q,T\right)=\int_{t_1}^{t_n} Dist_{Q,T}\left(t\right)dt\,, \tag{4.2}$$

όπου DistQ,T(t) είναι η συνάρτηση της Ευκλείδειας απόστασης μεταξύ των τροχιών Q και T στο χρόνο.

Ωστόσο, επειδή κάθε τροχιά παριστάνεται από ένα σύνολο διακριτών σημείων όπου εφαρμόζεται γραμμική παρεμβολή ενδιαμέσως, ο ορισμός της ανομοιότητας μετατρέπεται ως εξής:

$$DISSIM\left(Q,T\right)=\sum_{k=1}^{n-1}\int_{t_k}^{t_{k+1}} Dist_{Q,T}\left(t\right)dt\,, \tag{4.3}$$

όπου t_k είναι τα χρονικά αποτυπώματα στα οποία τα αντικείμενα *T* και *Q* κατέγραψαν τη θέσης τους. Προφανώς, σε εφαρμογές στον πραγματικό κόσμο, οι ρυθμοί δειγματοληψίας τροχιών μπορεί να ποικίλουν, με αποτέλεσμα τροχιές με θέσεις που δειγματοληπτούνται σε διαφορετικά χρονικά αποτυπώματα· παρόλα αυτά, αν θεωρήσουμε δύο τροχιές με αυτό το χαρακτηριστικό, η θέση του πρώτου αντικειμένου στη χρονική στιγμή κατά την οποία το δεύτερο κατέγραψε τη θέση του μπορεί να βρεθεί προσεγγιστικά εφαρμόζοντας γραμμική παρεμβολή.

Η Ευκλείδεια απόσταση μεταξύ δύο σημείων που κινούνται σε γραμμικές χρονικές συναρτήσεις μεταξύ διαδοχικών χρονικών αποτυπωμάτων, ορίστηκε στην Εξ.(3.5) και είναι :

$$Dist_{Q,T}\left(t\right)=\sqrt{at^2+bt+c}\,, \tag{4.4}$$

όπου *a*, *b*, *c* είναι οι συντελεστές του τριωνύμου (πραγματικοί αριθμοί, $a\geq 0$).

Για να υπολογίσουμε το ολοκλήρωμα της $Dist_{Q,T}(t)$, διακρίνουμε τις εξής δύο περιπτώσεις για την τιμή του μη αρνητικού συντελεστή a :

- $a=0$. Όπως αποδεικνύεται στην [MB04], αυτό σημαίνει ότι $b=0$. Άρα,

$$\int_{t_k}^{t_{k+1}} Dist_{Q,T}\left(t\right)dt=\frac{\sqrt{c}}{t_{k+1}-t_k} \tag{4.5}$$

- $a>0$. Βάσει του [MB04]:

$$\int_{t_k}^{t_{k+1}} Dist_{Q,T}\left(t\right)dt=\left|\frac{2at+b}{4a}\sqrt{at^2+bt+c}-\frac{b^2-4ac}{8a\sqrt{a}}\operatorname{arcsinh}\left(\frac{2at+b}{\sqrt{4ac-b^2}}\right)\right|_{t_i}^{t_{i+1}} \tag{4.6}$$

Για να αποφύγουμε αυτό τον τόσο πολύπλοκο υπολογισμό, χρησιμοποιούμε τον Κανόνα του Τραπεζίου (Trapezoid Rule) για τον υπολογισμό του ολοκληρώματος, που μας δίνει το ακόλουθο λήμμα.

Λήμμα 4.1*: Η τιμή της ανομοιότητα μεταξύ δύο σημείων που κινούνται γραμμικά με το χρόνο μπορεί να δοθεί προσεγγιστικά από τον ακόλουθο τύπο:*

$$DISSIM\left(Q,T\right)\approx DISSIM_{approx}\left(Q,T\right)=\frac{1}{2}\sum_{k=1}^{n-1}\left(\left(Dist_{Q,T}\left(t_k\right)+Dist_{Q,T}\left(t_{k+1}\right)\right)\cdot\left(t_{k+1}-t_k\right)\right) \tag{4.7}$$

με το σφάλμα της προσέγγισης, που εξαρτάται από τις τιμές t_k, t_{k+1}, να έχει όριο:

$$E_{Q,T}\le\sum_{k=1}^{n-1}\begin{cases}\frac{\left(t_{k+1}-t_k\right)^3}{12}\left|Dist_{Q,T}^{(2)}\left(-b/2a\right)\right| & ,\alpha\nu\ t_k\le -b/2a\le t_{k+1}\\ \frac{\left(t_{k+1}-t_k\right)^3}{12}\left|Dist_{Q,T}^{(2)}\left(t_{k+1}\right)\right| & ,\alpha\nu\ t_k<t_{k+1}<-b/2a\\ \frac{\left(t_{k+1}-t_k\right)^3}{12}\left|Dist_{Q,T}^{(2)}\left(t_k\right)\right| & ,\alpha\nu\ -b/2a<t_k<t_{k+1}\end{cases} \tag{4.8}$$

Απόδειξη: Σύμφωνα με τη μέθοδο του τραπεζίου, η τραπεζοειδής προσέγγιση $T_n\left(f\right)$ της $\int_{x_0}^{x_n}f\left(x\right)dx$ που σχετίζεται με τη διαμέριση $x_0<x_1<...<x_n$ δίνεται από τον:

$$T_n\left(f\right)=\frac{1}{2}\left(x_n-x_0\right)\cdot\left[f\left(x_0\right)+2f\left(x_1\right)+...+2f\left(x_{n-1}\right)+f\left(x_n\right)\right] \tag{4.9}$$

Αν η $f^{(2)}\left(x\right)$ είναι συνεχής στο $[x_0,x_n]$, τότε το όριο σφάλματος $E_n\left(f\right)$ στον τραπεζοειδή κανόνα δίνεται ως εξής:

$$E_n\left(f\right)\le\frac{\left(x_n-x_0\right)^3}{12n^2}\left|f^{(2)}\left(M\right)\right|, \tag{4.10}$$

όπου $f^{(2)}\left(M\right)$ είναι η μέγιστη τιμή της $f^{(2)}\left(x\right)$ στο $[x_0,x_n]$, δηλαδή,

$$\left|f^{(2)}\left(M\right)\right|\ge\left|f^{(2)}\left(x\right)\right|\forall\,x\in[x_0,x_n] \tag{4.11}$$

Στην περίπτωσή μας, αν θέσουμε $n=1$, τελικά υπολογίζουμε:

$$\int_{t_k}^{t_{k+1}}Dist_{Q,T}\left(t\right)dt\approx\frac{1}{2}\left(Dist_{Q,T}\left(t_k\right)+Dist_{Q,T}\left(t_{k+1}\right)\right)\cdot\left(t_{k+1}-t_k\right) \tag{4.12}$$

με όριο σφάλματος της προσέγγισής μας:

$$E_{Q_k,T_k}\le\frac{\left(t_{k+1}-t_k\right)^3}{12}\left|Dist_{Q,T}^{(2)}\left(M\right)\right|, \tag{4.13}$$

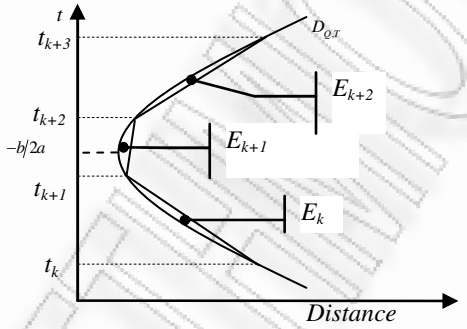
όπου $Dist_{Q,T}^{(2)}\left(M\right)$ είναι η μέγιστη τιμή της $Dist_{Q,T}^{(2)}\left(t\right)$ in $[t_k,t_{k+1}]$. Έτσι, προσδιορίζουμε τη μέγιστη τιμή της $Dist_{Q,T}^{(2)}\left(t\right)=\frac{4ac-b^2}{4\left(at^2+bt+c\right)^{3/2}}$ στο $[t_k,t_{k+1}]$. Επειδή η πρώτη παράγωγος της $Dist_{Q,T}^{(2)}\left(t\right)$, $Dist_{Q,T}^{(3)}\left(t\right)$ μηδενίζεται στο $t=-b/2a$ και $D_{Q,T}^{(4)}\left(-b/2a\right)=\frac{-3a\left(4a\right)^{5/2}}{4\left(4ac-b^2\right)^{3/2}}\le 0$ (επειδή $a\ge 0$), η μεγαλύτερη τιμή της $Dist_{Q,T}^{(2)}(t)$ στον $\mathbb{R}$ είναι $Dist_{Q,T}^{(2)}\left(-b/2a\right)$. Τέλος, διακρίνουμε τις εξής τρεις περιπτώσεις:

- $t_k\le -b/2a\le t_{k+1}$. Σ' αυτή την περίπτωση, $Dist_{Q,T}^{(2)}\left(M\right)=Dist_{Q,T}^{(2)}\left(-b/2a\right)$ και το σφάλμα είναι $E_{Q_k,T_k}\le\frac{\left(t_{k+1}-t_k\right)^3}{12}\left|Dist_{Q,T}^{(2)}\left(-b/2a\right)\right|$.

- $t_k < t_{k+1} < b/2a$. Σ' αυτή την περίπτωση, $Dist^{(2)}_{Q,T}(M) = Dist^{(2)}_{Q,T}(t_{k+1})$ και το σφάλμα είναι $E_{Q_k,T_k} \le \frac{(t_{k+1}-t_k)^3}{12}\left|Dist^{(2)}_{Q,T}(t_{k+1})\right|$.
- $-b/2a < t_k < t_{k+1}$. Σ' αυτή την περίπτωση, $Dist^{(2)}_{Q,T}(M) = Dist^{(2)}_{Q,T}(t_k)$ και το σφάλμα είναι $E_{Q_k,T_k} \le \frac{(t_{k+1}-t_k)^3}{12}\left|Dist^{(2)}_{Q,T}(t)\right|$.

Τέλος, προσθέτοντας τις *n*-1 εξισώσεις του υπολογισμού σφάλματος ανομοιότητας κατά μέρη, υποδηλώνεται ότι το προσεγγιστικό σφάλμα $E_{Q,T}$ υπολογίζεται όπως παρουσιάστηκε στο Λήμμα 4.1■

Το Σχήμα 4.1 παρουσιάζει την τραπεζοειδή προσέγγιση δείχνοντας το προσεγγιστικό σφάλμα *E* στις τρεις παραπάνω περιπτώσεις: η τιμή του $-b/2a$ είναι το σημείο καμπής της $Dist^{(2)}_{Q,T}$· το E_k υπολογίζεται βάση της τιμής της $Dist^{(2)}_{Q,T}(t_{k+1})$ (περίπτωση b), το E_{k+1} υπολογίζεται βάση της τιμής της $Dist^{(2)}_{Q,T}(-b/2a)$ (περίπτωση a) και το E_{k+2} υπολογίζεται βάσει της $Dist^{(2)}_{Q,T}(t_{k+2})$ (περίπτωση c).



Σχήμα 4.1: Μέθοδος του τραπεζίου

Μέχρι τούδε έχουμε ορίσει την ανομοιότητα μεταξύ δύο τροχιών (Ορισμός 4.2) και έχουμε προσεγγίσει τη μέτρησή της με ένα λιγότερο δαπανηρό υπολογισμό και ένα όριο σφάλματος. Όπως αναφέρθηκε ήδη, η θέση των χρονικών αποτυπωμάτων που δεν καταγράφονται προσεγγίζεται από τη γραμμική παρεμβολή μεταξύ διαδοχικά καταγεγραμμένων σημείων (η υποστήριξη μη γραμμικής π.χ. τοξωτής, κίνησης έχει αφεθεί ως ανοικτό θέμα). Στη συνέχεια, θα δώσουμε μια σειρά από μετρικές που θα χρησιμοποιηθούν στους αλγορίθμους αναζήτησης MST.

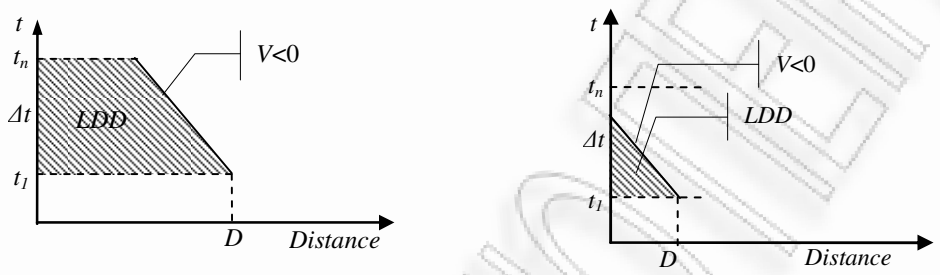
4.3.2. Μετρικές Εξαρτώμενες από την Ταχύτητα

Στην ενότητα αυτή καθορίζουμε δύο μετρικές, πιο συγκεκριμένα τις *OPTDISSIM* και *PESDISSIM* και δίνουμε αρκετά λήμματα που θα χρησιμοποιηθούν για σκοπούς κλαδέματος κατά την Αναζήτηση MST. Πριν προχωρήσουμε στον πυρήνα της ενότητας, ορίζουμε την *Γραμμικά Εξαρτώμενη Ανομοιότητα* (*Linearly Depended Dissimilarity - LDD*) που χρησιμοποιείται για τον ορισμό των μετρικών μας:

Ορισμός 4.3: *Η Γραμμικά Εξαρτώμενη Ανομοιότητα (LDD) μεταξύ δύο κινούμενων αντικειμένων με αρχική απόσταση D που κινούνται συγγραμμικά με σχετική ταχύτητα V κατά την περίοδο* $\Delta t = [t_1, t_n]$, *δίνεται από την ακόλουθη έκφραση*

$$LDD(D,V,\Delta t)=\begin{cases}\Delta t\cdot\left(D+\dfrac{V\cdot\Delta t}{2}\right) & ,\alpha\nu\ D+V\cdot\Delta t\geq 0\\ \dfrac{D^2}{2|V|} & ,\alpha\lambda\lambda\iota\acute{\omega}\varsigma\end{cases} \tag{4.14}$$

Η σχετική ταχύτητα *V* είναι ένα αρνητικός (θετικός) αριθμός όταν η απόσταση μεταξύ των δύο αντικειμένων μειώνεται (αυξάνεται, αντίστοιχα). Για να δείξουμε αυτό τον ορισμό, έστω το Σχήμα 4.2 όπου η *LDD* περιγράφεται ως η σκιασμένη επιφάνεια που περικλείεται από την κεκλιμένη γραμμή που παριστάνει τη συνάρτηση της απόστασης μεταξύ δύο αντικειμένων που κινούνται το ένα προς το άλλο με σχετική ταχύτητα *V*, με τις οριζόντιες γραμμές t_1 και t_n να ορίζουν το Δt. Οι δύο περιπτώσεις του ορισμού *LDD* παρουσιάζονται στο Σχήμα 4.2(a) και στο Σχήμα 4.2(b), αντίστοιχα.



(a) (b)

Σχήμα 4.2: Ορισμός της *LDD*

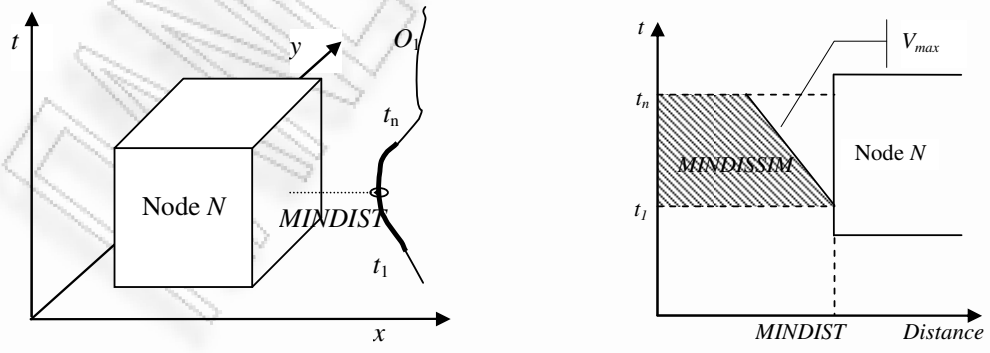
Έχοντας ορίσει την *LDD*, μπορούμε να συνεχίσουμε με τον ορισμό της πρώτης μετρικής που χρησιμοποιείται στις στρατηγικές διάταξης και κλαδέματος κατά τη διαδικασία της αναζήτησης:

Ορισμός 4.4: *Η ελάχιστη DISSIM (MINDISSIM) στη διάρκεια της περιόδου* $\Delta t=[t_1,t_n]$ *μεταξύ μιας τροχιάς που δεικτοδοτείται από μία δομή τύπου R-δέντρου με ένα γραμμικό τμήμα που βρίσκεται εντός ενός κόμβου ευρετηρίου N και μίας τροχιάς επερώτησης Q, ορίζεται ως:*

$$MINDISSIM(Q,N,\Delta t)=LDD\left(MINDIST(Q,N),V_{\max},\Delta t\right) \tag{4.15}$$

όπου V_{max} *είναι το άθροισμα της (a) μέγιστης ταχύτητας των δεικτοδοτημένων τροχιών και (b) της μέγιστης ταχύτητας της τροχιάς επερώτησης.*

Η μετρική μπορεί να χρησιμοποιηθεί για σκοπούς διάταξης και κλαδέματος λόγω του εξής λήμματος.



(a) (b)

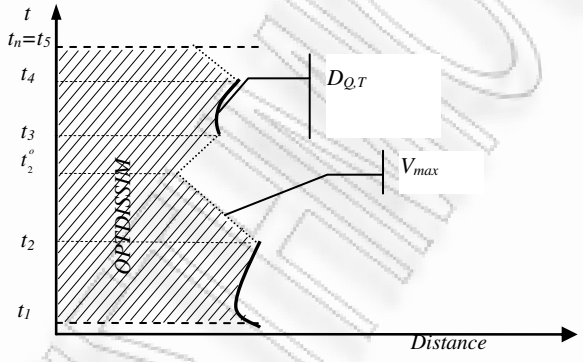
Σχήμα 4.3: Ορισμός της *MINDISSIM*

Λήμμα 4.2: *Η DISSIM μεταξύ μιας τροχιάς που δεικτοδοτείται από μία δομή τύπου R-δέντρου που περιέχεται μερικώς εντός ενός κόμβου του ευρετηρίου N και μιας τροχιάς επερώτησης Q στη διάρκεια*

μιας περιόδου $[t_1,t_n]$ *δεν μπορεί να έχει DISSIM μικρότερη από την αντίστοιχη MINDISSIM του κόμβου.*

Απόδειξη: Βάσει του παραπάνω ορισμού, η *MINDISSIM* αντιστοιχεί στην *DISSIM* ενός κινούμενου αντικειμένου που βρίσκεται εντός του *N* για μία *μόνο* χρονική στιγμή και κατόπιν μετακινείται προς την τροχιά επερώτησης με τη μέγιστη δυνατή ταχύτητα (η σκιασμένη επιφάνεια στην περιοχή στο Σχήμα 4.3(b)). Προφανώς, οποιοδήποτε άλλο αντικείμενο με τουλάχιστον ένα γραμμικό τμήμα εντός της *N* θα πλησιάζει την τροχιά επερώτησης με ταχύτητα μικρότερη ή ίση της V_{max}, αυξάνοντας συνεπώς τη σκιασμένη τραπεζοειδή επιφάνεια στο Σχήμα 4.3 (ήτοι η κλίση της κεκλιμένης γραμμής στο Σχήμα 4.3(b) θα είναι μεγαλύτερη). Επιπλέον, αν το αντικείμενο παραμείνει εντός του *N* για πάνω από μία χρονική στιγμή, η κεκλιμένη γραμμή θα καλύπτει ένα τμήμα του διαστήματος επερώτησης (και όχι το σύνολο), γεγονός που θα οδηγεί σε μία περιοχή με μεγαλύτερη επιφάνεια. ■

Ανάλογα με την παρουσία ή απουσία ευρετηρίου, κάθε αλγόριθμος που θα χρησιμοποιείται για αναζήτηση MST θα πρέπει να υπολογίζει την ανομοιότητα μιας τροχιάς επερώτησης και αρκετών (δεικτοδοτημένων ή μη) τροχιών· προφανώς, σε οποιαδήποτε χρονική στιγμή της εκτέλεσής του ένας τέτοιος αλγόριθμος θα έχει ανακτήσει αρκετά τμήματα υποψήφιων MSTs.



Σχήμα 4.4: Ορισμός της *OPTDISSIM*

Παρόλο που δεν μπορούμε να υπολογίσουμε την ακριβή *DISSIM* αυτών των τροχιών που έχουν ανακτηθεί μερικώς από την τροχιά επερώτησης, μπορούμε με ασφάλεια να εκτιμήσουμε ένα κάτω όριο γι' αυτήν, το οποίο καλείται *OPTDISSIM*. Έστω, για παράδειγμα, το Σχήμα 4.4 που δείχνει την *OPTDISSIM* μιας μερικώς ανακτηθείσας υποψήφιας τροχιάς *T* από την τροχιά επερώτησης *Q*. Η *OPTDISSIM* αποτελείται μερικώς από την ανομοιότητα των εγγραφών που έχουν ήδη ανακτηθεί από το ευρετήριο (η σκιασμένη περιοχή κατά τα χρονικά διαστήματα $[t_1,t_2]$ και $[t_3,t_4]$). Όσον αφορά την περίοδο $[t_4,t_5]$, η μικρότερη δυνατή ανομοιότητα δίνεται αν υποθέσουμε ότι το κινούμενο αντικείμενο ξεκίνησε από τη θέση του στο t_4 προσεγγίζοντας το αντικείμενο επερώτησης με τη μέγιστη δυνατή ταχύτητα (η κεκλιμένη γραμμή μεταξύ t_4 και t_5). Τέλος, όταν ασχολούμαστε με τα ενδιάμεσα χρονικά διαστήματα όπως το $[t_2,t_3]$, πρέπει να υπολογίσουμε τη χρονική στιγμή t_2^o στην οποία το αντικείμενο σταμάτησε την κίνησή του προς την τροχιά επερώτησης (η κεκλιμένη γραμμή μεταξύ t_2 και t_2^o) και κατόπιν επέστρεψε στη γνωστή του θέση κατά τη χρονική στιγμή t_3 (η κεκλιμένη γραμμή μεταξύ t_2^o και t_3). Τώρα μπορούμε και επισήμως πλέον να δώσουμε τον ορισμό της *OPTDISSIM*:

Ορισμός 4.5: *Η πιο αισιόδοξη DISSIM (OPTDISSIM) μεταξύ μιας τροχιάς επερώτησης Q και μιας δεικτοδοτημένης τροχιάς T με γραμμικά τμήματα που έχουν ανακτηθεί μερικώς από το ευρετήριο, κατά την περίοδο* $[t_1, t_n]$*, ορίζεται ως:*

$$OPTDISSIM(Q,T,t_1,t_n) = \sum_{k=1}^{n-1} \begin{cases} DISSIM(Q_k, T_k) & ,\alpha\nu\ T_k \in S_R; \\ LDD\left(Dist_{Q,T}(t_{k+1}), -V_{\max}, (t_{k+1} - t_k)\right) & ,\alpha\nu\ T_k \notin S_R, k = 1; \\ LDD\left(Dist_{Q,T}(t_k), -V_{\max}, (t_{k+1} - t_k)\right) & ,\alpha\nu\ T_k \notin S_R, k = n-1; \\ LDD\left(Dist_{Q,T}(t_k), -V_{\max}, (t_k^o - t_k)\right) + & \\ LDD\left(Dist_{Q,T}(t_{k+1}), V_{\max}, (t_{k+1} - t_k^o)\right) & ,\alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \quad (4.16)$$

όπου $Dist_{Q,T}$ *είναι η χρονική συνάρτηση της απόστασης μεταξύ των τροχιών Q και T,* S_R *είναι το σύνολο των γραμμικών τμημάτων που έχουν ήδη ανακτηθεί από το ευρετήριο,* V_{max} *είναι το άθροισμα της μέγιστης ταχύτητας των δεικτοδοτημένων τροχιών συν τη μέγιστη ταχύτητα της τροχιάς επερώτησης, και το* t_k^o *εκφράζεται ως:*

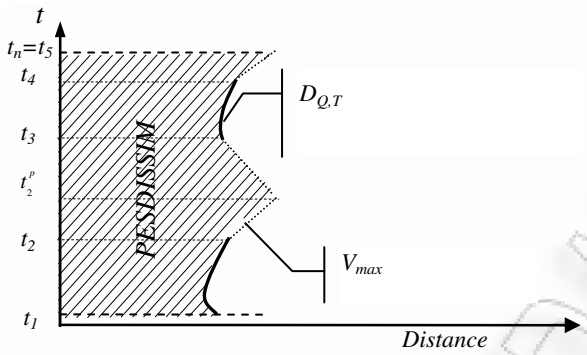
$$t_k^o = \frac{1}{2}\left(t_k + t_{k+1} + \frac{\left(D_{Q,T}(t_{k+1}) - D_{Q,T}(t_k)\right)}{V_{\max}} \right) \quad (4.17)$$

Αν θυμηθούμε το Σχήμα 4.4, η τιμή του t_k^o χρησιμοποιεί απευθείας το γεγονός ότι η κλίση των δύο κεκλιμένων γραμμών μεταξύ $[t_2, t_2^o]$ και $[t_2^o, t_3]$ είναι η ίδια και ισούται με V_{max}. Αφού ορίσαμε την *OPTDISSIM*, μπορούμε να δώσουμε το ακόλουθο λήμμα, το οποίο θα αποβεί χρήσιμο και για σκοπούς κλαδέματος:

Λήμμα 4.3: *Μια τροχιά που δεικτοδοτείται από μια δομή τύπου R-δέντρου με γραμμικά τμήματα μερικώς ανακτηθέντα από το ευρετήριο δεν μπορεί να έχει μικρότερη DISSIM από μια τροχιά επερώτησης Q κατά τη διάρκεια μιας περιόδου* $[t_1, t_n]$ *από την αντίστοιχη OPTDISSIM.*

Απόδειξη: Βάσει του παραπάνω ορισμού, η *OPTDISSIM* είναι το άθροισμα της *DISSIM* των εγγραφών τροχιάς που έχουν ήδη ανακτηθεί από το ευρετήριο (που ανήκουν στο σύνολο S_R), μία τιμή που είναι σταθερή, συν την *DISSIM* ενός αντικειμένου που προσέγγισε την τροχιά επερώτησης με τη μέγιστη δυνατή ταχύτητα (V_{max}) κατά τα χρονικά διαστήματα που δεν έχουν ακόμα ανακτηθεί από το ευρετήριο, με τον περιορισμό ότι το αντικείμενο πρέπει να βρεθεί σε δεδομένες θέσεις στην αρχή και / ή το τέλος του διαστήματος. Συνεπώς, μιας και τα δύο αντικείμενα προσεγγίζουν το ένα το άλλο με τη μέγιστη δυνατή ταχύτητα κατά τη διάρκεια αυτών των περιόδων, η απόσταση μεταξύ τους ελαχιστοποιείται· ελαχιστοποιείται έτσι και το αντίστοιχο ολοκλήρωμα και κατά συνέπεια και η ανομοιότητα τους. ■

Αντιστοίχως, υιοθετώντας το ίδιο σενάριο όπου ένας αλγόριθμος MST έχει ανακτήσει μερικώς μόνο κάποιες τροχιές, μπορούμε να εκτιμήσουμε ένα ανώτατο όριο, για τη *DISSIM* ανάμεσα στην επερώτηση και μία μερικώς ανακτηθείσα τροχιά, που καλείται *PESDISSIM*. Όπως φαίνεται στο Σχήμα 4.5, η *PESDISSIM* λειτουργεί με ανάλογο τρόπο με την *OPTDISSIM* με τη διαφορά ότι κατά τα χρονικά διαστήματα όπου η κίνηση του αντικειμένου δεν είναι γνωστή, υποθέτουμε ότι το αντικείμενο απομακρίνεται (και δεν προσεγγίζει) από την τροχιά επερώτησης με τη μέγιστη δυνατή ταχύτητα V_{max}. Κατά τον ίδιο τρόπο, ορίζουμε και τυπικά την *PESDISSIM*:



Σχήμα 4.5: Ορισμός της *PESDISSIM*

Ορισμός 4.6: *Η πιο απαισιόδοξη DISSIM (PESDISSIM) μεταξύ μιας τροχιάς επερώτησης Q και μιας δεικτοδοτημένης τροχιάς T με γραμμικά τμήματα μερικώς ανακτηθέντα από το ευρετήριο, κατά την περίοδο* $[t_1,t_n]$*,ορίζεται ως:*

$$PESDISSIM(Q,T,t_1,t_n)=\sum_{k=1}^{n-1}\begin{cases} DISSIM(Q_k,T_k) & ,\alpha\nu\ T_k\in S_R; \\ LDD\left(D_{Q,T}(t_{k+1}),V_{\max},(t_{k+1}-t_k)\right) & ,\alpha\nu\ T_k\notin S_R,k=1; \\ LDD\left(D_{Q,T}(t_k),V_{\max},(t_{k+1}-t_k)\right) & ,\alpha\nu\ T_k\notin S_R,k=n-1; \\ LDD\left(D_{Q,T}(t_k),V_{\max},(t_k^p-t_k)\right)+ & \\ LDD\left(D_{Q,T}(t_{k+1}),-V_{\max},(t_{k+1}-t_k^p)\right) & ,\alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \quad (4.18)$$

όπου $D_{Q,T}$*,* S_R *και* V_{max} *είναι όπως ορίζονται στους προηγούμενους ορισμούς, και το* t_k^p *εκφράζεται ως εξής:*

$$t_k^p=\frac{1}{2}\left(t_k+t_{k+1}+\frac{\left(D_{Q,T}(t_k)-D_{Q,T}(t_{k+1})\right)}{V_{\max}}\right) \quad (4.19)$$

Το ακόλουθο λήμμα προκύπτει απευθείας από τον ορισμό της PESDISSIM.

Λήμμα 4.4: *Μια τροχιά που δεικτοδοτείται από μία δομή τύπου R-δέντρου με γραμμικά τμήματα μερικώς ανακτηθέντα από το ευρετήριο δεν μπορεί να έχει μεγαλύτερη DISSIM από μία τροχιά επερώτησης Q κατά την περίοδο* $[t_1,t_n]$ *από την αντίστοιχη PESDISSIM.*

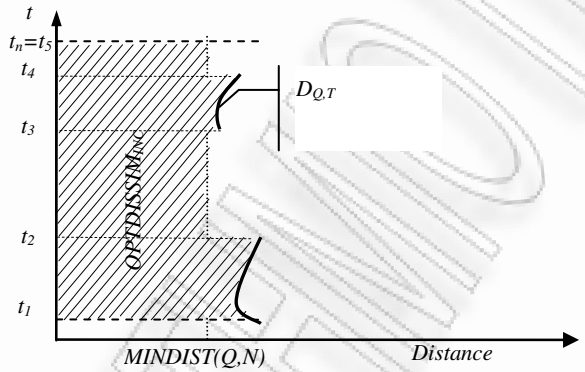
Απόδειξη: Βάσει του παραπάνω ορισμού, η *PESDISSIM* είναι το άθροισμα της *DISSIM* των εγγραφών τροχιάς που έχουν ήδη ανακτηθεί από το ευρετήριο (που ανήκουν στο σύνολο S_R), τιμή που είναι σταθερή, συν την *DISSIM* ενός αντικειμένου που αποκλίνει την τροχιά επερώτησης με τη μέγιστη δυνατή ταχύτητα (V_{max}) κατά τα χρονικά διαστήματα που δεν έχουν ακόμα ανακτηθεί από το ευρετήριο, με τον περιορισμό ότι το αντικείμενο πρέπει να βρεθεί σε δεδομένες θέσεις στην αρχή και / ή το τέλος του χρονικού διαστήματος. Συνεπώς, η απόσταση μεταξύ των δύο τροχιών κατά τις περιόδους αυτές μεγιστοποιείται, μεγιστοποιώντας έτσι και την ανομοιότητα τους. ■

4.3.3. Μετρικές Ανεξάρτητες της Ταχύτητας

Η χρησιμοποίηση των μετρικών που ορίσαμε προηγουμένων σε έναν αλγόριθμο αναζήτησης MST μπορεί να ενισχύσει σημαντικά την απόδοσή του κλαδεύοντας αρκετές υποψήφιες τροχιές. Ωστόσο, αυτές οι μετρικές είναι σχετικά χαλαρές, διότι βασίζονται στη μέγιστη ταχύτητα V_{max} που από θεωρητικής σκοπιάς, θα μπορούσε να είναι τάξεις μεγέθους υψηλότερη από τη μέση ταχύτητα των αντικειμένων. Συνεπώς, πρέπει να καθορίσουμε άλλες μετρικές που δεν επηρεάζονται από την V_{max},

που να υποστηρίζουν αλγόριθμους αναζήτησης MST ανεξάρτητους της ταχύτητας. Αυτές οι μετρικές μπορούν να αναπτυχθούν όταν ένας αλγόριθμος MST αναφέρει του κόμβους του ευρετηρίου με αύξουσα σειρά της *MINDIST* τους από την τροχιά επερώτησης. Προφανώς, η υπόθεση αυτή είναι λογική αν λάβουμε υπ' όψιν δομές τύπου R-δέντρου όπου μπορεί να χρησιμοποιηθεί μια στρατηγική «πρώτα στον καλύτερο» όπως αυτή που προτείνεται στην [HS99].

Έστω για παράδειγμα, το Σχήμα 4.6 που περιγράφει την *DISSIM* μιας μερικώς ανακτηθείσας υποψήφιας τροχιάς *T* από την τροχιά επερώτησης *Q*. Σύμφωνα με την προηγούμενη συζήτησή μας, η *DISSIM* μεταξύ $[t_1,t_2]$ και $[t_3,t_4]$ ορίζεται με ακρίβεια. Σε αυτή την περίπτωση ωστόσο, μπορούμε να χρησιμοποιήσουμε το γεγονός ότι οι κόμβοι ευρετηρίου προσπελαύνονται με αύξουσα σειρά της *MINDIST* τους από την τροχιά επερώτησης. Κατά συνέπεια, οποιοδήποτε γραμμικό τμήμα δεν έχει ανακτηθεί ακόμα από το ευρετήριο, δεν μπορεί να είναι εγγύτερο στην *Q* από την *MINDIST*(*Q*,*N*) όπου *N* είναι ο επόμενος κόμβος του ευρετηρίου στην ουρά και το κατώτερο όριο της *DISSIM* γίνεται η σκιασμένη περιοχή στο Σχήμα 4.6.



Σχήμα 4.6: Ορισμός της $OPTDISSIM_{INC}$

Πιο τυπικά, ορίζουμε την $OPTDISSIM_{INC}$ ως εξής:

Ορισμός 4.7: *Αν υποθέσουμε ότι οι κόμβοι του ευρετηρίου αναφέρονται με αύξουσα σειρά της MINDIST τους από την τροχιά επερώτησης, η πιο αισιόδοξη DISSIM μεταξύ μιας τροχιάς επερώτησης Q και μιας δεικτοδοτημένης τροχιάς T στη διάρκεια μιας περιόδου $[t_1,t_n]$ που έχει ένα γραμμικό τμήμα εντός ενός κόμβου δέντρου N, εκφράζεται ως εξής:*

$$OPTDISSIM_{INC}(Q,T,N,t_1,t_n)=\sum_{k=1}^{n-1}\begin{cases} DISSIM(Q_k,T_k) & , \ \alpha\nu\ T_k\in S_R; \\ MINDIST(N,T)\cdot(t_{k+1}-t_k) & , \ \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \quad (4.20)$$

όπου S_R είναι το σύνολο των ήδη ανακτηθέντων από το ευρετήριο γραμμικών τμημάτων.

Χρησιμοποιώντας τον παραπάνω ορισμό της $OPTDISSIM_{INC}$, μπορούμε επίσης να ορίσουμε την ελάχιστη *DISSIM* ενός κόμβου ευρετηρίου *N*:

Ορισμός 4.8: *Αν υποθέσουμε ότι οι κόμβοι ευρετηρίου αναφέρονται με αύξουσα σειρά της MINDIST τους από την τροχιά επερώτησης, η ελάχιστη DISSIM μεταξύ μιας τροχιάς T, που δεικτοδοτείται από μία δομή τύπου R-δέντρου που έχει ένα γραμμικό τμήμα σε ένα κόμβο N, και μια τροχιά επερώτησης Q κατά την περίοδο $[t_1,t_n]$, ορίζεται ως:*

$$MINDISSIM_{INC}(Q,N,t_1,t_n) = \min\begin{cases} MINDIST(Q,N)\cdot(t_n - t_1) \\ OPTDISSIM_{INC}(Q,T,N,t_1,t_n), \quad \forall T \in S_C \end{cases} \tag{4.21}$$

όπου S_C, είναι το σύνολο των τροχιών με γραμμικά τμήματα που έχουν ήδη ανακτηθεί από το ευρετήριο αλλά δεν έχουν ακόμα ολοκληρωθεί εντός της περιόδου $[t_1,t_n]$.

Λήμμα 4.5: *Αν υποθέσουμε ότι οι κόμβοι ευρετηρίου αναφέρονται με αύξουσα σειρά της MINDIST τους από μια τροχιά επερώτησης Q, μια τροχιά που είναι μερικώς αποθηκευμένη σε ένα κόμβο N δεν μπορεί να έχει μικρότερη DISSIM από την Q κατά τη χρονική περίοδο $[t_1,t_n]$ από την αντίστοιχη $MINDISSIM_{INC}$ του κόμβου.*

Απόδειξη: Οποιοδήποτε γραμμικό τμήμα εντός του *N* βρίσκεται σε μία τροχιά που είτε ανήκει στο S_c είτε όχι. Στην πρώτη περίπτωση, θεωρώντας ότι οι κόμβοι αναφέρονται με αύξουσα σειρά, οι εγγραφές τροχιάς που δεν έχουν ακόμα ανακτηθεί δεν μπορούν να βρίσκονται εγγύτερα στο αντικείμενο της επερώτησης από τη *MINDIST* του κόμβου στον οποίο ανήκουν. Έτσι, η ελάχιστη ανομοιότητα ενός αντικειμένου του S_c είναι το άθροισμα της ανομοιότητας των εγγραφών που έχουν ήδη ανακτηθεί από το ευρετήριο, συν την ανομοιότητα ενός αντικειμένου που είναι όσο κοντά όσο η *MINDIST* στην τροχιά επερώτησης κατά το υπόλοιπο χρονικό διάστημα της επερώτησης – ένα άθροισμα που αντιστοιχεί στον ορισμό του $OPTDISSIM_{INC}$. Στη δεύτερη περίπτωση, όπου η τροχιά δεν ανήκει στο S_c, το γραμμικό τμήμα δεν μπορεί ν' ανήκει σε ένα αντικείμενο που έχει ανακτηθεί πλήρως από το ευρετήριο γιατί αυτό θα οδηγούσε σε επανάληψη των γραμμικών τμημάτων στο ευρετήριο. Έτσι το γραμμικό τμήμα ανήκει σε ένα κινούμενο αντικείμενο χωρίς ανακτηθέντα τμήματα από τους κόμβους που προσπελάστηκαν προηγουμένως και δεν μπορεί να είναι εγγύτερα στην τροχιά επερώτησης από την *MINDIST*. Έτσι στην καλύτερη περίπτωση, η απόστασή του από το αντικείμενο της επερώτησης κατά την περίοδο της επερώτησης είναι ίσο με τη *MINDIST* και η *DISSIM* του είναι ίση με την $MINDIST(Q,N)\cdot\Delta t$. ■

4.3.4. Ευριστικές

Τα λήμματα που παρουσιάστηκαν στις προηγούμενες ενότητες υποστηρίζουν στις ακόλουθες ευριστικές που χρησιμοποιούνται απευθείας στους αλγόριθμους Αναζήτησης MST που θα παρουσιαστούν στις επόμενες Ενότητες.

- **Ευριστική 1**: Κάθε γραμμικό τμήμα τροχιάς που περιέχεται σε ένα κόμβο ευρετηρίου τύπου R-δέντρου με *MINDISSIM* μεγαλύτερη της τρέχουσας ομοιότερης (ήτοι, αυτής με τη μικρότερη υπολογισμένη *DISSIM* - ή *PESDISSIM* αν δεν υπάρχει μία πλήρως υπολογισμένη *DISSIM*) ανήκει σε ένα κινούμενο αντικείμενο που δεν μπορεί να είναι ομοιότερο της τροχιάς επερώτησης από το τρέχον ομοιότερο· επομένως ο κόμβος μπορεί να κλαδευτεί από τη λίστα των υποψηφίων.
- **Ευριστική 2**: Κάθε τροχιά με *OPTDISSIM* μεγαλύτερη της τρέχουσας ομοιότερης δεν μπορεί να είναι ομοιότερη της τροχιάς επερώτησης από την τρέχουσα ομοιότερη· άρα μπορεί να κλαδευτεί από τη λίστα των υποψηφίων.
- **Ευριστική 3**: Όταν αναφέρονται κόμβοι φύλλου και εσωτερικοί κόμβοι με αύξουσα σειρά της *MINDIST* τους από την τροχιά επερώτησης, κάθε γραμμικό τμήμα της τροχιάς που περιέχεται σ΄ ένα κόμβο με $MINDISSIM_{INC}$ μεγαλύτερη της τρέχουσας ομοιότερης ανήκει σε ένα κινούμενο αντικείμενο που δεν μπορεί να είναι ομοιότερο στην τροχιά επερώτησης, εξ'

ου και ο κόμβος μπορεί να κλαδευτεί από τη λίστα των υποψηφίων. Επιπλέον, επειδή κάθε κόμβος που αναφέρεται μετά από αυτόν που επεξεργαστήκαμε θα έχει *MINDIST* μεγαλύτερη ή ίση της *MINDIST* του παρόντος κόμβου (Ορισμός 4.8), το ίδιο θα ισχύει και για τις αντίστοιχες τιμές της $MINDISSIM_{INC}$. Ως αποτέλεσμα, όλοι αυτοί οι κόμβοι θα έχουν $MINDISSIM_{INC}$ μεγαλύτερη από την τρέχουσα ομοιότερη και ο αλγόριθμος μπορεί να τερματιστεί γιατί όλοι οι υπόλοιποι κόμβοι μπορούν να κλαδευτούν.

4.4. Αλγόριθμοι για Αναζήτηση της *k*-Ομοιότερης Τροχιάς

Στην ενότητα αυτή περιγράφουμε ενδελεχώς τους αλγόριθμους που απαντούν σε επερωτήσεις MST χρησιμοποιώντας τις τρεις ευριστικές που περιγράφηκαν στην προηγούμενη ενότητα και κατόπιν τους γενικεύουμε ώστε να υποστηρίξουμε επερωτήσεις *k*-MST. Παρέχουμε δύο εναλλακτικές: μία «πρώτα στο βαθύτερο» και μία «πρώτα στον καλύτερο» η δεύτερη εκ των οποίων υποθέτει ότι οι κόμβοι ευρετηρίου αναφέρονται με αύξουσα σειρά της *MINDIST* τους.

4.4.1. Αλγόριθμος Αναζήτησης MST «πρώτα στο βαθύτερο»

Ο πρώτος αλγόριθμος (`DFMSTSearch` που φαίνεται στο Σχήμα 4.7) προσπελαύνει τη δενδρική δομή με τρόπο «πρώτα στο βαθύτερο», κλαδεύοντας τους κόμβους του δέντρου που δεν πληρούν το χρονικό περιορισμό της τροχιάς επερώτησης, όπως συμβαίνει και στην ενότητα 3.4.1.1 σχετικά με τις επερωτήσεις πλησιέστερης γείτονα. Ο αλγόριθμος ξεκινά κάνοντας παρεμβολή για να μας δώσει το τμήμα της τροχιάς επερώτησης που βρίσκεται πλήρως εντός της χρονικής έκτασης της επερώτησης. Σε γενικές γραμμές, ο αλγόριθμος χρησιμοποιεί την *DISSIM* μεταξύ τροχιών ως μετρική απόστασης και όχι την Ευκλείδεια Απόσταση· στα εσωτερικά επίπεδα του ευρετηρίου η μετρική *MINDISSIM* χρησιμοποιείται για να ταξινομήσει τη λίστα των κλαδιών και να την κλαδέψει χρησιμοποιώντας την ευριστική 1 κατά την οπισθοδρόμηση του αλγορίθμου (γραμμές 31-36). Στο επίπεδο των φύλλων, ο αλγόριθμος χρησιμοποιεί τρεις δομές κατακερματισμού στη κύρια μνήμη: Μία με τις πλήρεις τροχιές (*Completed*), μία με τις μερικώς ανακτηθείσες τροχιές (*Valid*) και μία με τις μερικώς ανακτηθείσες που παρόλα αυτά έχουν ήδη απορριφθεί (*Rejected*) τροχιές. Οι *Completed* και *Valid* δομές κύριας μνήμης αποθηκεύουν *λίστες* που η καθεμία περιέχει τα χρονικά διαστήματα του κινούμενου αντικειμένου που έχουν ανακτηθεί, και τις αποστάσεις στην αρχή και το τέλους του κάθε ανακτηθέντος τμήματος, τη (μερική) *DISSIM* του, τον αντίστοιχο υπολογισμό του σφάλματος και τις τιμές των *OPTDISSIM* και *PESDISSIM*. Η δομή της *Rejected* περιέχει μόνο *ταυτότητες* τροχιάς.

Όταν γίνεται επεξεργασίας μιας εγγραφής φύλλου, ο αλγόριθμος ελέγχει αν η εγγραφή ανήκει σε ένα *Rejected* κινούμενο αντικείμενο (απλά χρησιμοποιώντας την ταυτότητά του) και το απορρίπτει αν ανήκει (γραμμή 6). Εν συνεχεία ελέγχει κατά πόσο η εγγραφή ανήκει σε ένα *Valid* κινούμενο αντικείμενο και αν ανήκει ανακτά τη λίστα του *L*· αλλιώς δημιουργεί μία νέα λίστα και την προσθέτει στην *Valid* (γραμμές 7-11). Ο αλγόριθμος χρησιμοποιεί μία μέθοδο σάρωσης επιπέδου που εξετάζει τις εγγραφές των φύλλων και τα τμήματα της τροχιάς στη χρονική τους διάσταση σε ένα μόνο πέρασμα. Κάτι τέτοιο απαιτεί οι εγγραφές του φύλλου να έχουν προηγουμένως ταξινομηθεί βάσει της χρονικής τους σειράς (γραμμή 4), εκτός και αν η αντίστοιχη δομή δέντρου (όπως το TB-δέντρο και το TB^*-δέντρο) τις αποθηκεύει με χρονική σειρά ούτως ή άλλως.

```
 1.  Algorithm DFMSTSearch(node N, trajectory Q, time period Q_per, struct
                           MSim, Hash Valid, Hash Completed, Hash Rejected)
 2.      Q = Interpolate(Q, Max(Q.T_S, Q_per.T_S), Min(Q.T_E, Q_per.T_E))
 3.      IF N is leaf
 4.         Sort(N,TS)
 5.         FOR EACH leaf entry E IN leaf node N
 6.            IF Rejected not contains E.Id
 7.               IF Valid contains E.Id
 8.                  retrieve list L
 9.               ELSE
10.                  create list L; Add L in Valid
11.               ENDIF
12.               Find next query entry QS with QS.T_e<N.T_S; QE=QS
13.               DO UNTIL QE.T_s > E.T_e
14.                  Interpolate to produce nE, nQE in period (T1,T2)
15.                  Add (T1,T2) in L
16.                  Calc(DISSIM,PESDISSIM,OPTDISSIM,ERR)
17.                  IF L is completed
18.                     Move L from Valid to Completed
19.                     IF DISSIM<MSim.DISSIM Update MSim with nE.Id,DISSIM
20.                  ELSE
21.                     IF PESDISSIM<MSim.DISSIM
22.                        Update MSim with nE.Id,PESDISSIM
23.                     ENDIF
24.                     IF OPTDISSIM>MSim.DISSIM
25.                        Move L from Valid to Rejected
26.                     ENDIF
27.                  ENDIF
28.            NEXT query entry QE
29.            Return QE in the query entry QS
30.         NEXT
31.      ELSE
32.         BranchList=GenTrajectoryBranchList(Q,N)
33.         SortBranchList(BranchList)
34.         FOR EACH entry E IN BranchList
35.            DFMSTSearch E.ChNode, E.Trajectory, MSim
36.            PruneBranchList(BranchList)
37.         NEXT
38.      ENDIF
```

Σχήμα 4.7: Αλγόριθμος αναζήτησης ομοιότερης τροχιάς «πρώτα στο βαθύτερο» (αλγόριθμος DFMSTSearch)

Όταν μία εγγραφή φύλλου και αλληλεπικαλύπτεται με τη τροχιά επερώτησης στη χρονική τους διάσταση ο αλγόριθμος προσθέτει την αντίστοιχη περίοδο στη λίστα *L* (γραμμές 14-15), και υπολογίζει τις *DISSIM*, *OPTDISSIM* και *PESDISSIM,* μαζί με το αντίστοιχο σφάλμα υπολογισμού (γραμμή 16). Αν η λίστα *L* συμπληρωθεί, αφαιρείται από τη *Valid* και προστίθεται στην *Completed*, η *DISSIM* της ελέγχεται βάσει της τρέχουσας ομοιότερης και, αν είναι μικρότερη, παίρνει τη θέση της στη μεταβλητή που αποθηκεύει τη τρέχουσα τιμή της πιο όμοιας τροχιάς *MSim* (γραμμές 18-19). Στην περίπτωση που η *L* δεν έχει ακόμα συμπληρωθεί, η *PESDISSIM* της ελέγχεται σε σχέση με την τρέχουσα ομοιότερη και εάν είναι μικρότερη, παίρνει τη θέση της στην *MSim* (γραμμές 21-24)· η *OPTDISSIM* της συγκρίνεται επίσης με την τρέχουσα ομοιότερη και εάν είναι μεγαλύτερη, η λίστα μετακινείται από τη *Valid* στη *Rejected* με εφαρμογή της ευριστικής 2 (γραμμές 24-26).

```
 1.  Algorithm BFMSTSearch (R-tree R, trajectory Q, period Q_per)
 2.     Q = Interpolate(Q, Max(Q.T_S, Q_per.T_S), Min(Q.T_E, Q_per.T_E))
 3.     EnQueue Queue, R.RootNode, 0, Q
 4.     DO WHILE Queue.Count > 0
 5.        Element = DeQueue(Queue); N=Element.Node; Q=Element.QueryTrajectory
 6.        IF Completed.Count>0
 7.           IF MINDISSIM_INC(Q,N)>MSim.DISSIM RETURN MSim
 8.        ELSE
 9.           IF N is leaf node
10.              Sort(N, TS)
11.              FOR EACH leaf entry E IN leaf node N
12.                 IF Rejected not contains E.Id
13.                    IF Valid contains E.Id
14.                       retrieve list L
15.                    ELSE
16.                       create list L; Add L in Valid
17.                    ENDIF
18.                    Find next query entry QS in Q with QS.T_e<N.T_S; QE=QS
19.                    DO UNTIL QE.T_S > E.T_e
20.                       Interpolate to produce nE, nQE period (T1,T2)
21.                       Add (T1,T2) in L
22.                       Calc DISSIM, PESDISSIM, OPTDISSIM, ERR
23.                       IF L is completed
24.                          Move L from Valid to Completed
25.                       ENDIF
26.                       IF DISSIM<MSim.DISSIM
27.                          Update Msim with nE, DISSIM
28.                       ELSE
29.                          IF PESDISSIM<MSim.DISSIM
30.                             Update MSim with nE, PESDISSIM
31.                          ENDIF
32.                          IF OPTDISSIM>MSim.DISSIM
33.                             Move L from Valid to Rejected
34.                          ENDIF
35.                       ENDIF
36.                    NEXT query entry QE
37.                 ENDIF
38.                 Return QE in the query entry QS
39.              NEXT
40.           ELSE
41.              FOR EACH entry E IN the node Element
42.                 IF (Q.T_S,Q.T_E) Overlaps (E.T_S,E.T_E)
43.                    Interpolate to produce nQE in period (T1,T2)
44.                    Dist = MinDist_Trajectory_Rectangle(nQ, E)
45.                    EnQueue Queue, E, Dist, nQ
46.                 ENDIF
47.              NEXT
48.           ENDIF
49.        ENDIF
50.     LOOP
```

Σχήμα 4.8. Αλγόριθμος αναζήτηση ομοιότερης τροχιάς «πρώτα στον καλύτερο» (αλγόριθμος BFMSTSearch)

4.4.2. Αλγόριθμος Αναζήτησης MST «πρώτα στον καλύτερο»

Ο δεύτερος αλγόριθμος (BFMSTSearch, φαίνεται στο Σχήμα 4.8) προσπελαύνει τη δενδρική δομή με τρόπο «πρώτα στον καλύτερο», υπολογίζοντας τις κατάλληλες *MINDIST*s μεταξύ της τροχιάς επερώτησης και των κόμβων του δέντρου, αναφέροντας έτσι φύλλα και εσωτερικούς κόμβους σε αύξουσα σειρά της *MINDIST* τους από την τροχιά επερώτησης.

Και πάλι, ο αλγόριθμος ξεκινά εφαρμόζοντας παρεμβολή για την παραγωγή του τμήματος της τροχιάς επερώτησης που βρίσκεται πλήρως εντός της χρονικής έκτασης της επερώτησης. Εν συνεχεία, κατά την επεξεργασία ενός εσωτερικού κόμβου (γραμμές 35-39), ο αλγόριθμος υπολογίζει τη

MINDIST μεταξύ του κόμβου και του τμήματος της τροχιάς επερώτησης *Q* που βρίσκεται εντός της χρονικής έκτασης του κόμβου χρησιμοποιώντας τη μετρική *MinDist_Trajectory_Rectangle* (που χρησιμοποιείται επίσης στους αλγορίθμους εύρεσης του πλησιέστερου γείτονα τροχιάς) και κατόπιν μπαίνει στην ουρά. Κατά την επεξεργασία των φύλλων (γραμμές 9-30), ο αλγόριθμος επεξεργάζεται εγγραφές με τον ίδιο ακριβώς τρόπο με τον αλγόριθμο `DFMSTSearch`. Και στις δύο περιπτώσεις επεξεργασίας ενός φύλλου ή εσωτερικού κόμβου, ο αλγόριθμος ελέγχει πρώτα αν η $MINDISSIM_{INC}$ είναι μεγαλύτερη από την τρέχουσα ομοιότερη και αν είναι, ο αλγόριθμος τερματίζεται εφαρμόζοντας την ευριστική 3 και επιστρέφει την τρέχουσα ομοιότερη ως απάντηση της επερώτησης (γραμμές 5-7). Σημειώστε ότι προκειμένου να αποφύγουμε τον υπολογισμό όλων των $OPTDISSIM_{INC}$ τιμών που περιλαμβάνονται στον ορισμό της $MINDISSIM_{INC}$ (δηλαδή όλων των $T \in S_C$ στον ορισμό 6), ελέγχουμε πρώτα κατά πόσον η τιμή $MINDIST(Q,N)\cdot(t_n - t_1)$ του κόμβου είναι μικρότερη της τρέχουσας ομοιότερης. Σ' αυτή την περίπτωση, ο υπολογισμός των τιμών $OPTDISSIM_{INC}$ παραλείπεται, αφού η τιμή της $MINDISSIM_{INC}$ θα είναι μικρότερη της τρέχουσας ομοιότερης ανεξάρτητα των τιμών της $OPTDISSIM_{INC}$.

4.4.3. Επέκταση σε *k*-MST αλγόριθμους

Όπως στην [RKV95] και στην εργασία μας για επερωτήσεις πλησιέστερου γείτονα που παρουσιάστηκε στο προηγούμενο κεφάλαιο, γενικεύουμε τους δύο ανωτέρω αλγόριθμους για να υποστηρίζουν την αναζήτηση της *k*-ομοιότερης τροχιάς λαμβάνοντας υπ' όψιν τα εξής:

- χρήση μιας προσωρινής μνήμης από το πολύ *k* (τρεχουσών) ομοιότερων τροχιών που ταξινομούνται βάσει της πραγματικής τους ανομοιότητας με την τροχιά επερώτησης, και
- είτε κλάδεμα βάσει της ανομοιότητας του πιο ανόμοιου αντικειμένου στην προσωρινή μνήμη, όταν επεκτείνεται ο αλγόριθμος `DFMSTSearch`, είτε
- ολοκλήρωση της εκτέλεσης του αλγόριθμου όταν γίνει επεξεργασία ενός κόμβου με $MINDISSIM_{INC}$ μεγαλύτερη της ανομοιότητας του πιο ανόμοιου αντικειμένου στην προσωρινή μνήμη, όταν επεκτείνεται ο αλγόριθμος `BFMSTSearch`.

4.4.4. Διαχείριση Σφάλματος

Και οι δύο παραπάνω αλγόριθμοι υπολογίζουν τις ανομοιότητες μεταξύ τροχιάς επερώτησης και δεικτοδοτημένων τροχιών χρησιμοποιώντας την προσέγγιση που παρουσιάστηκε στο Λήμμα 4.1 υπολογίζοντας ταυτόχρονα το αντίστοιχο όριο του σφάλματος (σημειώνεται ως *ERR* και στο Σχήμα 4.7 και στο Σχήμα 4.8). Ωστόσο, περά από τον υπολογισμό του, η χρήση του ορίου του σφάλματος είναι θεμελιώδης προκειμένου να υπολογίσουμε ακριβή και ορθά αποτελέσματα, κάτι το οποίο δεν περιγράφεται ρητά στους δύο αλγόριθμους για λόγους σαφήνειας. Στην πράξη, πρέπει να εισαγάγουμε τρεις τροποποιήσεις και στους δύο αλγόριθμους για να συμπεριλάβουμε το ρόλο του προσεγγιστικού σφάλματος:

- Μία υποψήφια ομοιότερη τροχιά, που δεν έχει ακόμα ολοκληρωθεί, συγκρίνεται με την τρέχουσα *k*-στη ομοιότερη χρησιμοποιώντας την τιμή του *PESDISSIM-ERR*.
- Μια συμπληρωμένη υποψήφια ομοιότερη τροχιά συγκρίνεται με την τρέχουσα *k*-στη ομοιότερη χρησιμοποιώντας την τιμή *DISSIM-ERR*.

- Αντί της χρήσης μιας *k*-στης ομοιότερης, απαιτείται η χρήση μιας προσωρινής μνήμης των υποψήφιων *k*-στων ομοιότερων τροχιών. Αυτές θα είναι όλες οι τροχιές με *DISSIM* μεγαλύτερη από την *k*-στη ομοιότερη και *DISSIM-ERR* μικρότερο.

Σύμφωνα με τα προηγούμενα και οι δύο αλγόριθμοι μπορεί να τερματιστούν με *m*>*k* υποψήφιες ομοιότερες τροχιές. Σε τέτοιες περιπτώσεις, απαιτείται ένα βήμα επεξεργασία μετά την εκτέλεση και των δύο αλγορίθμων προκειμένου να προσδιορίσουμε τις οριστικές *k* ομοιότερες τροχιές υπολογίζοντας την πραγματική ανομοιότητα κάθε υποψήφιας τροχιάς σε σχέση με την τροχιά επερώτησης. Παρόλο, που υπολογιστικά η διαδικασία είναι επίπονη, γίνεται μόνον όταν η προσωρινή μνήμη που αποθηκεύει τις υποψήφιες τροχιές περιέχει παραπάνω από μία τροχιά, ή όταν η σειρά με την οποία αναφέρονται οι τροχιές από την *k*-προσωρινή μνήμη μπορεί να επηρεαστεί από το σφάλμα υπολογισμού της ομοιότητας κάθε τροχιάς.

4.5. Πειραματική Μελέτη

Οι δύο αλγόριθμοι `DFMSTSearch` και `BFMSTSearch` που παρουσιάστηκαν μπορούν να εφαρμοσθούν σε οποιαδήποτε δομή τύπου R-δέντρου που αποθηκεύει ιστορικές πληροφορίες κινούμενων αντικειμένων όπως στο 3D R-δέντρο, στο STR-δέντρο [PJT00], στο TB-δέντρο [PJT00] και στο TB*-δέντρο. Μεταξύ αυτών, εφαρμόσαμε τους αλγόριθμους χρησιμοποιώντας το 3D R-, το TB- και το TB*-δέντρο.

4.5.1. Πειραματικό Πλαίσιο

Κατά τη διάρκεια των πειραμάτων, χρησιμοποιήσαμε σελίδα μεγέθους 4 KB και μία (μεταβλητού μεγέθους) προσωρινή μνήμη χωρητικότητας 10% του μεγέθους ευρετηρίου, με μέγιστη χωρητικότητα 1000 σελίδες. Τα πειράματα εκτελέστηκαν σε PC με Microsoft Windows XP με επεξεργαστή AMD Athlon 64 3GHz, 512 MB RAM και χώρο δίσκο αρκετά GB.

Όσον αφορά τα σύνολα δεδομένων που χρησιμοποιήθηκαν για το σκοπό της μελέτης, τα πραγματικά σύνολα δεδομένων που χρησιμοποιούνται από ανάλογες εργασίες ομοιότητας τροχιάς ([COO05], [VKG02]), δεν είναι κατάλληλα για τους στόχους μας λόγω του γεγονότος ότι αποτελούνται από 2D προβολές τροχιών χωρίς καμία πληροφορία για τα δειγματοληπτούμενα χρονικά αποτυπώματα· λογικό βέβαια, αν αναλογιστεί κανείς ότι η ομοιότητα που μετράται σε εκείνες τις εργασίες εξαρτάται μόνο από τη χωρική και όχι από τη χωροχρονική ομοιότητα τροχιάς. Για το λόγο αυτό χρησιμοποιήσαμε το σύνολο δεδομένων *Trucks* (ενότητα 1.5.1) ώστε να αξιολογήσουμε την ποιότητα της προτεινόμενης μέτρησης ομοιότητας (ενότητα 4.5.2). Ωστόσο, επειδή αυτό το σύνολο δεδομένων είναι σχετικά μικρό (273 τροχιές και 112203 γραμμικά τμήματα), δεν μπορούσε να μας δώσει την πραγματική απόδοση των αλγόριθμων· συνεπώς, η μελέτη απόδοσης (ενότητα 4.5.3) έγινε χρησιμοποιώντας συνθετικά σύνολα δεδομένων που δημιουργήθηκαν από μία γεννήτρια που αναπτύχθηκε ειδικά για τους σκοπούς της μελέτης και βασίστηκε στη γεννήτρια δεδομένων GSTD [TSN99]. Ο κύριος σκοπός για τον οποίο χρησιμοποιήθηκε αυτή η ειδικής κατασκευής γεννήτρια δεδομένων και όχι η ευρύτατα χρησιμοποιούμενη GSTD, είναι ότι μία θεμελιώδης παράμετρος που επηρεάζει την απόδοση των προτεινόμενων αλγόριθμων είναι η σχέση μεταξύ της μέσης και της μέγιστης ταχύτητας των κινούμενων αντικειμένων που δεικτοδοτούνται από το δέντρο, κάτι που δεν μπορεί να ελεγχθεί από τη GSTD.

Για να επιτύχουμε κλιμάκωση στον όγκο των δεδομένων, δημιουργήσαμε συνθετικές τροχιές των 100, 250, 500 και 1000 κινούμενων αντικειμένων με αποτέλεσμα σύνολα δεδομένων των 200K, 500K, 1000K, και 2000K εγγραφών, αντίστοιχα (η θέση κάθε αντικειμένου αποτέλεσε αντικείμενο δειγματοληψίας περίπου 2000 φορές), δημιουργώντας έτσι ευρετήρια μεγέθους μέχρι και 100 MB. Ο λόγος μέγιστη / μέση ταχύτητα (που στο υπόλοιπο του κεφαλαίου συμβολίζεται με *MMS*) των συνόλων δεδομένων τίθεται εξ' ορισμού στο 10, που είναι λογική τιμή αν λάβουμε υπ' όψιν εφαρμογές στον πραγματικό κόσμο όπου οι τροχιές αντιπροσωπεύουν πεζούς ή κινούμενα οχήματα. Παρόλα αυτά, για να μελετήσουμε την ευαισθησία των αλγόριθμων ως προς αυτή την παράμετρο, παρήχθησαν 3 επιπλέον σύνολα 500 κινούμενων αντικειμένων με τον *MMS* στο 2, 5 και 20, αντίστοιχα. Για τις υπόλοιπες παραμέτρους της γεννήτριας, η αρχική κατανομή και η κατεύθυνση των αντικειμένων σε όλες τις περιπτώσεις ήταν τυχαία, ενώ η ταχύτητα τους δίνεται από μία κανονική ή λογαριθμική κανονική κατανομή ανάλογα με τον επιθυμητό *MMS*. Ο Πίνακας 4.2 παρουσιάζει συνοπτικές πληροφορίες για τα πραγματικά και τα παραγόμενα σύνολα δεδομένων και τα αντίστοιχα ευρετήρια. Σημειώστε ότι κάθε συνθετικό σύνολο δεδομένων συμβολίζεται με βάση το πλήθος του και το *MMS* του (π.χ. το $S_{0100,10}$ αποτελείται από 100 τροχιές με το *MMS* να ισούται με 10).

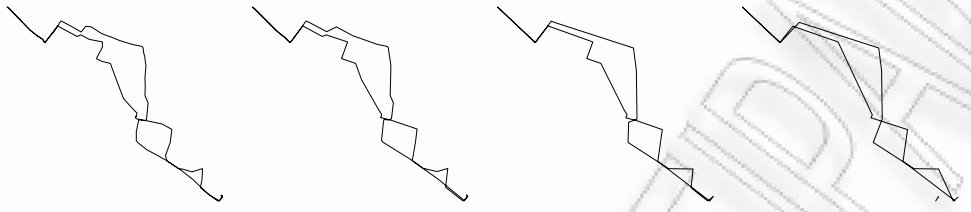
Πίνακας 4.2: Συνοπτικές πληροφορίες συνόλων δεδομένων

Σύνολο δεδομένων	*# Αντικειμένων*	*MMS*	*# Εγγραφών (x1K)*	*Κατανομή Ταχύτητας*			*Μέγεθος Ευρετηρίου (MB)*		
				Τύπος	*μ*	*Σ*	*3D R- δέντρο*	*TB-δέντρο*	*TB*-δέντρο*
Trucks	*276*	*16*	*112*	*Real data*			*3.2*	*1.8*	*1.0*
$S_{0100,10}$	*100*	*10*	*200*	*Lognormal*	*1*	*0.6*	*10.7*	*5.2*	*2.4*
$S_{0250,10}$	*250*	*10*	*500*	*Lognormal*	*1*	*0.6*	*25.8*	*13.1*	*6.1*
$S_{0500,10}$	*500*	*10*	*1000*	*Lognormal*	*1*	*0.6*	*51.0*	*26.2*	*12.2*
$S_{1000,10}$	*1000*	*10*	*2000*	*Lognormal*	*1*	*0.6*	*99.1*	*52.4*	*24.5*
$S_{0500,2}$	*500*	*2*	*1000*	*Normal*	*3*	*1.0*	*51.4*	*26.2*	*12.2*
$S_{0500,5}$	*500*	*5*	*1000*	*Lognormal*	*1*	*0.4*	*51.1*	*26.2*	*12.2*
$S_{0500,20}$	*500*	*20*	*1000*	*Lognormal*	*1*	*0.8*	*50.3*	*26.2*	*12.2*

4.5.2. Πειράματα ως προς την Ποιότητα

Για να αξιολογήσουμε την ποιότητα της προτεινόμενης μετρικής ομοιότητας κάναμε μία εκτεταμένη σειρά πειραμάτων χρησιμοποιώντας το πραγματικό σύνολο δεδομένων *Trucks*. Όλες οι τροχιές του συνόλου δεδομένων συμπιέστηκαν χρησιμοποιώντας τον αλγόριθμο TD-TR της [MB04] παράγοντας συνεπώς τεχνητές τροχιές, που ήταν παρόμοιες (αλλά όχι πανομοιότυπες) με αυτές του αρχικού συνόλου δεδομένων. Μετά, χρησιμοποιήσαμε κάθε συμπιεσμένη τροχιά για να κάνουμε επερωτήσεις στο αρχικό σύνολο δεδομένων, με προοπτική ο αλγόριθμος να μας δώσει την αντίστοιχη αρχική τροχιά ως ομοιότερη. Τρέξαμε ένα σύνολο επερωτήσεων θέτοντας k=1 και μετρήσαμε πόσες φορές η επερώτηση δεν μας έδωσε την αρχική τροχιά ως ομοιότερη. Κλιμακώσαμε επίσης την τιμής της παραμέτρου (κατωφλίου p) του TD-TR από 0.1% έως 10% του μήκους κάθε τροχιάς, για να επιτύχουμε διαφορετικές τιμές ομοιότητας δεδομένου ότι μία αυξανόμενη παράμετρος TD-TR παράγει μία συμπιεσμένη τροχιά με λιγότερα δειγματοληπτούμενα σημεία και μεγαλύτερη ανομοιότητα ως

προς την αρχική τροχιά. Επί παραδείγματι, το Σχήμα 4.9 δείχνει (a) μία αρχική τροχιά και τις τροχιές που παράγονται χρησιμοποιώντας τον αλγόριθμοTD-TR με (b, c, d) διάφορες τιμές του *p*. Μια βασική παρατήρηση που προκύπτει από το Σχήμα 4.9 είναι ότι ενώ το γενικό σχήμα της τροχιάς παραμένει ανεπηρέαστο με την αύξηση της τιμής του *p*, ο αριθμός των κορυφών που καθορίζουν την τροχιά μειώνεται και οι κατά τόπους λεπτομέρειες εξαφανίζονται.



p=0 *(168 κορυφές)* (a)

p=0.1 % *(65 κορυφές)* (b)

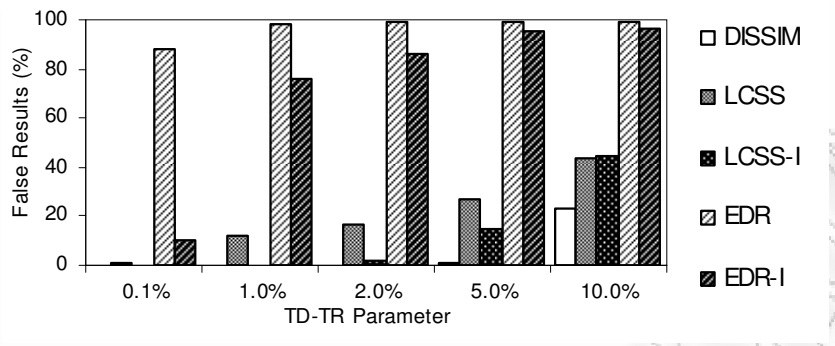
p=1 % *(29 κορυφές)* (c)

p=2 % *(22 κορυφές)* (d)

Σχήμα 4.9: Διαφορετικός βαθμός συμπίεσης σε μία τροχιά

Από τις σχετικές εργασίες επιλέξαμε να κάνουμε τα ίδια πειράματα χρησιμοποιώντας τις LCSS [VKG02] και EDR [COO05] ως μέτρα ομοιότητας. Δεν έχουμε συμπεριλάβει την DTW [BC96] στην πειραματική μας μελέτη, καθώς τόσο η LCSS όσο και η EDR αποδείχθηκε ότι υπερέχουν [VKG02], [COO05]. Καθορίζουμε την τιμή της παραμέτρου ε γι' αυτές τις δύο μετρήσεις στο ένα τέταρτο της μέγιστης τυπικής απόκλισης των τροχιών, που οδηγεί στα βέλτιστα αποτελέσματα συσταδοποίησης, όπως αναφέρεται στην [COO05]. Επίσης κανονικοποιήσαμε το σύνολο δεδομένων όπως προτείνεται στην ίδια εργασία. Επιπλέον, για να είναι δίκαιη η σύγκριση, κάναμε μία προφανή βελτίωση στις LCSS και EDR, προσθέτοντας χειροκίνητα σημεία στην συμπιεσμένη τροχιά (την τροχιά επερώτησης) με γραμμική παρεμβολή στα χρονικά αποτυπώματα στα οποία η ελεγχόμενη τροχιά του συνόλου δεδομένων έχει σημεία δειγματοληψίας. Ονομάσαμε αυτές τις βελτιωμένες εκδοχές LCSS-I και EDR-I αντίστοιχα.

Τα αποτελέσματα των πειραμάτων που αξιολογούν την ποιότητα της προτεινόμενης μετρικής ομοιότητας φαίνονται στο Σχήμα 4.10. Προφανώς, η προτεινόμενη μέτρηση ανομοιότητας (*DISSIM*) υπερέχει και των δύο ανταγωνιστών της σε όλες τις περιπτώσεις και ως προς τις βελτιωμένες τους εκδοχές. Στην πράξη, στο μεγαλύτερο τμήμα των πειραμάτων, η *DISSIM* ορθώς εντοπίζει την αρχική τροχιά από την οποία έχει παραχθεί η επερώτηση. Αφ' ετέρου, παράγει λανθασμένες απαντήσεις μόνο όταν η τιμή του *p* υπερβαίνει το 5%, επαληθεύοντας ότι είναι μία πολύ εύρωστη μετρική ομοιότητας. Η LCSS (και LCSS-I) επιτυγχάνουν επίσης καλή ποιότητα ταξινομώντας ορθά την τροχιά επερώτησης στην πλειονότητα των πειραματικών ρυθμίσεων · παρόλα αυτά, είναι πάντα λιγότερο ακριβής από την *DISSIM*. Για τις EDR και EDR-I, αποδεικνύεται ότι για τιμές *p* μεγαλύτερες του 1% αποτυγχάνουν εντελώς να περιγράψουν την ομοιότητα μεταξύ τροχιών, δεδομένου ότι οι λανθασμένες απαντήσεις υπερβαίνουν το 60%.



Σχήμα 4.10: Λανθασμένα αποτελέσματα αυξάνοντας την τιμή της παραμέτρου TD-TR

Ο λόγος για την κακή απόδοση που εμφανίζει η μέτρηση ομοιότητας EDR σε αυτά τα πειράματα εξηγείται λαμβάνοντας υπ' όψιν τον ορισμό της: η EDR είναι ο αριθμός των διαδικασιών εισαγωγής, διαγραφής ή αντικατάστασης (insert, delete, or replace) που απαιτούνται για τη μετατροπή της τροχιάς A σε B [COO05]. Έτσι, αν υποθέσουμε ότι n είναι ο αριθμός των κορυφών στην A και m είναι ο αριθμός των κορυφών στην (συμπιεσμένη) A_c, η $EDR(A, A_c) \geq n - m$ αφού απαιτείται να προστεθούν τουλάχιστον n-m κορυφές στην A_c ώστε να μετατραπεί στην A. Για ένα αυθαίρετο σύνολο δεδομένων τροχιάς T με k κορυφές οι οποίες χωρικά απέχουν της A, αποδεικνύεται εύκολα ότι η EDR μεταξύ των T και A_c είναι το πολύ $\max(m, k)$. Συνεπώς, αν ένα σύνολο δεδομένων περιέχει μια τροχιά T με k κορυφές και $\max(m,k) \leq n - m$, π.χ. μια τροχιά που αποτελείται από ένα μικρότερο αριθμό κορυφών, τότε ισχύει επίσης ότι $EDR(T, A_c) \leq EDR(A, A_c)$.

4.5.3. Πειράματα ως προς την Απόδοση

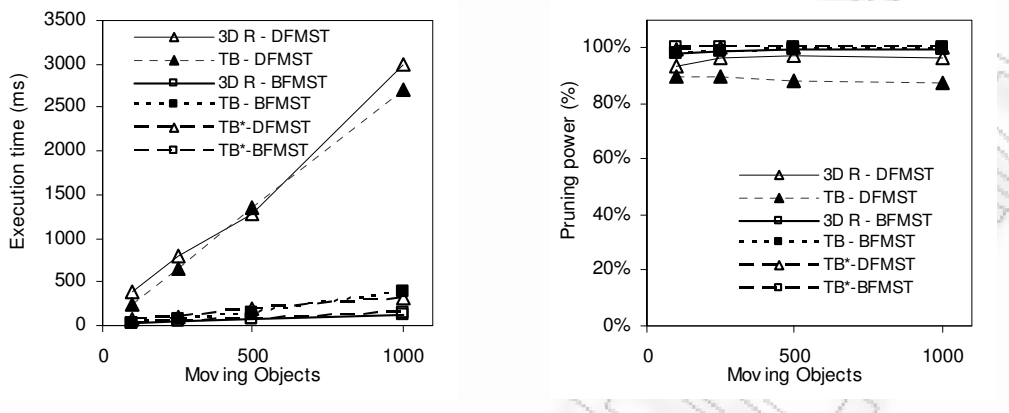
Και οι δύο αλγόριθμοι αξιολογήθηκαν με τέσσερα σύνολα των 500 επερωτήσεων βάσει των ρυθμίσεων που παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 4.3). Τα αποτελέσματα των αλγορίθμων σε σχέση με το πλήθος (Q_1), το *MMS* (Q_2), το μήκος επερώτησης (Q_3) και το k (Q_4) αξιολογήθηκαν χρησιμοποιώντας το 3D R-, το TB- και το TB*-δέντρο.

Πίνακας 4.3: Ρυθμίσεις Επερωτήσεων

Σύνολο Επερωτήσεων	*Σύνολα δεδομένων*	*Τροχιά επερώτησης (ως τμήμα μιας τυχαίας τροχιάς δεδομένων)*	*k*
Q_1	$S_{0100,10} \ldots S_{1000,10}$	*5%*	*1*
Q_2	$S_{0500,02} \ldots S_{0500,20}$	*5%*	*1*
Q_3	$S_{0500,10}$	*1% … 100%*	*1*
Q_4	$S_{0500,10}$	*5%*	*1..10*

Το Σχήμα 4.11 δείχνει το χρόνο εκτέλεσης και τον επιτευχθέντα κλαδεμένο χώρο για το σύνολο επερώτησης Q_1 (κλιμακούμενο με το πλήθος του συνόλου δεδομένων) αξιολογώντας τους αλγορίθμους `DFMSTSearh` and `BFMSTSearch`. Είναι σαφές ότι, ο αλγόριθμος `BFMSTSearch` υπερέχει του `DFMSTSearch` όσον αφορά το χρόνο εκτέλεσης, ενώ και οι δύο επιδεικνύουν πολύ καλή ισχύ κλαδέματος (άνω του 80% σε όλα τα πειράματα). Ο λόγος για την ελαφρά χειρότερη ισχύ κλαδέματος του αλγορίθμου `DFMSTSearch` είναι η επίδραση της παραμέτρου V_{max} στον ορισμό των μετρικών *MINDISSIM*, *OPTDISSIM* και *PESDISSIM*, γεγονός που οδηγεί σε σχετικά «χαλαρές» ευριστικές.

Καθώς αυξάνεται ο αριθμός των κινούμενων αντικειμένων, η V_{max} γίνεται αρκετές φορές μεγαλύτερη από την ταχύτητα της πλειονότητας των κινούμενων αντικειμένων, μειώνοντας την αποτελεσματικότητα των παραπάνω μετρικών και κατά συνέπεια της απόδοσης του αλγορίθμου MST .

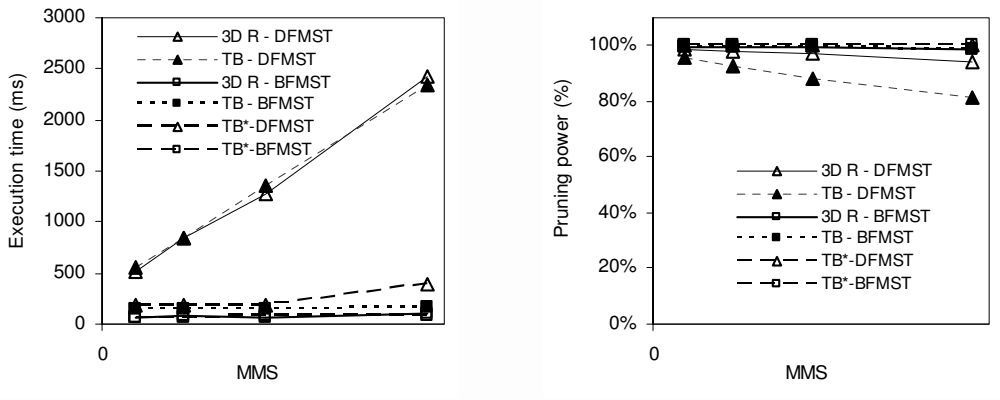


(a) (b)

Σχήμα 4.11: Κλιμάκωση με το πλήθος του συνόλου δεδομένων (Q1)

Μία άλλη παρατήρηση που προκύπτει από το Σχήμα 4.11 είναι ότι ενώ ο αλγόριθμος `DFMSTSearch` παρουσιάζει καλή ισχύ κλαδέματος, ο χρόνος εκτέλεσής του δεν ακολουθεί ανάλογη συμπεριφορά. Αυτό εξηγείται από το γεγονός ότι το κλάδεμα στον αλγόριθμο MST οφείλεται κυρίως στην ευριστική *OPTDISSIM*, που απαιτεί ο αλγόριθμος να διαβάζει τις εγγραφές των φύλλων και να τις απορρίπτει χωρίς να τις επεξεργάζεται (αν οι *ταυτότητές τους* ανήκουν σε ένα κινούμενο αντικείμενο *Rejected*). Αφ' ετέρου, ο αλγόριθμος `BFMSTSearch` κλαδεύει κυρίως μέσω της ευριστικής $MINDISSIM_{INC}$ η οποία απορρίπτει απευθείας όλους τους κόμβους του δέντρου που δεν έχουν ακόμα υποστεί επεξεργασία.

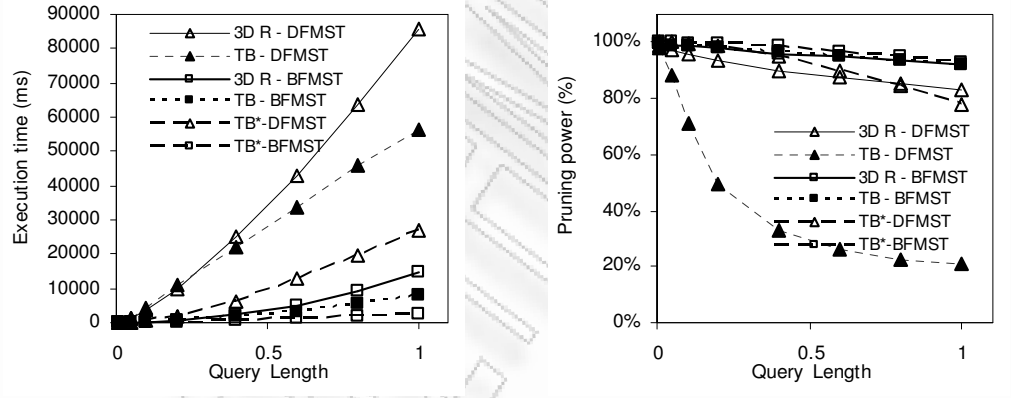
Η επιρροή της παραμέτρου V_{max} γίνεται εμφανής στο δεύτερο σύνολο πειραμάτων με το σύνολο επερώτησης Q_2, κλιμακούμενο με την τιμή του *MMS* (Σχήμα 4.12). Όπως διαπιστώνουμε, ο χρόνος εκτέλεσης του αλγορίθμου `DFMSTSearch` αυξάνεται γραμμικά με την τιμή του *MMS*. Ο χρόνος εκτέλεσης του αλγορίθμου `BFMSTSearch` παραμένει σταθερός, αφού δεν χρησιμοποιεί την ταχύτητα των αντικειμένων. Και πάλι, και οι δύο αλγόριθμοι επιτυγχάνουν πολύ καλό κλάδεμα του χώρου αναζήτησης.



(a) (b)

Σχήμα 4.12: Κλιμάκωση με το *MMS* (Q2)

Ανάλογα συμπεράσματα με τα παραπάνω συνάγονται και από το σύνολο επερωτήσεων Q_3, που δείχνει την κλιμάκωση των αλγορίθμων με το μήκος της επερώτησης (Σχήμα 4.13). Σε αυτή την περίπτωση θα πρέπει να τονίσουμε την «κακή» συμπεριφορά του TB-δέντρου ως προς την ισχύ κλαδέματός του, καθώς αυξάνεται με τη διάρκεια της περιόδου επερώτησης. Η παρατήρηση αυτή εξηγείται αν λάβουμε υπ' όψιν τον αλγόριθμο εισαγωγής του TB-δέντρου που αποθηκεύει σε κάθε κόμβο φύλλου τμήματα που ανήκουν στην ίδια τροχιά. Το κύριο μειονέκτημα αυτού είναι ότι τμήματα με χωρική εγγύτητα από διαφορετικές τροχιές αποθηκεύονται σε διαφορετικούς κόμβους. Άρα, το TB-δέντρο διατηρεί τη χρονική διάταξη των θέσεων των κινούμενων αντικειμένων, ενώ αγνοεί το χωρικό τους καταμερισμό. Ωστόσο όσο η χρονική έκταση της επερώτησης αυξάνεται, η ισχύς κλαδέματος του TB-δέντρου επιδεινώνεται, διότι το μειονέκτημα του ακατάλληλου χωρικού καταμερισμού των θέσεων των κινούμενων αντικειμένων υπερβαίνει το όφελος που αποκομίζεται από την αποθήκευση τους με χρονική σειρά. Εδώ, θα πρέπει επίσης να επισημάνουμε το πλεονέκτημα του TB*-δέντρου που φαίνεται να είναι ο αδιαμφισβήτητος νικητής σε όλες τις πειραματικές ρυθμίσεις. Προκύπτει ότι η στρατηγική «διαγραφή και επανεισαγωγή» ("delete και re-insert") που υιοθετείται από το TB*-δέντρο σε συνδυασμό με την αυξημένη χωρητικότητα των κόμβων του είναι επαρκώς αποτελεσματική στην αναζήτηση ομοιότητας.



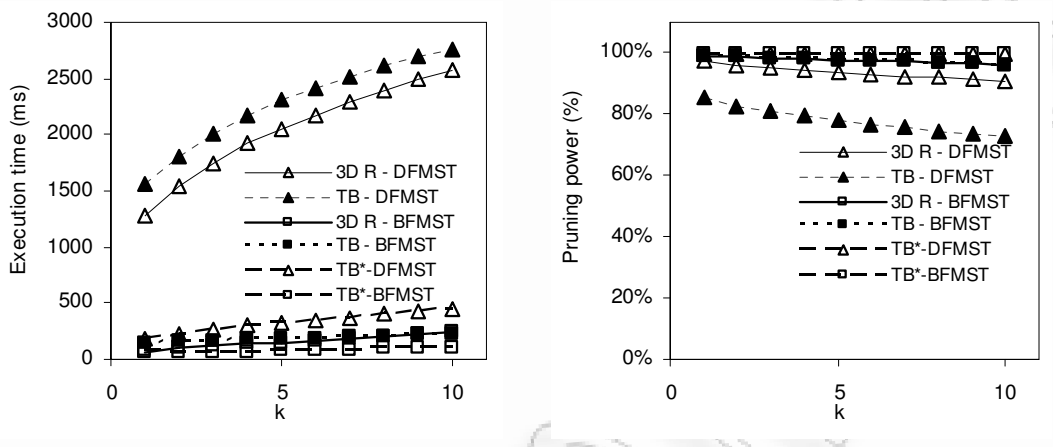
(a) (b)

Σχήμα 4.13: Κλιμάκωση με το μήκος επερώτησης (Q3)

Τέλος, το Σχήμα 4.14 παρουσιάζει τη συμπεριφορά των προτεινόμενων αλγορίθμων ως προς τον αριθμό των ομοιότερων τροχιών που ζητούνται. Και πάλι ο `BFMSTSearch` επιτυγχάνει υψηλή ισχύ κλαδέματος και μικρό χρόνο εκτέλεσης, ενώ η απόδοσή του μειώνεται με το k σε μικρή σχετικά αναλογία. Από την άλλη πλευρά, παρόλο που αλγόριθμος `DFMSTSearch` επιδεικνύει καλή ισχύ κλαδέματος – αλλά πάντοτε χειρότερη από αυτή του `BFMSTSearch` – ο χρόνος εκτέλεσης είναι αρκετές φορές υψηλότερος του αντίστοιχου χρόνου εκτέλεσης του ανταγωνιστή του, για τον ίδιο λόγο με προηγουμένως.

Συνοψίζοντας τα αποτελέσματα της πειραματικής μας μελέτης, και οι δύο αλγόριθμοι επιδεικνύουν υψηλή ισχύ κλαδέματος ενώ ο `BFMSTSearch` πάντα υπερέχει του `DFMSTSearch` κατά αρκετές τάξεις μεγέθους. Η ισχύς κλαδέματος του `DFMSTSearch` εξαρτάται από το μέγεθος του συνόλου δεδομένων, το *MMS*, το μήκος της επερώτησης και τον αριθμό του k, ενώ αυτή του `BFMSTSearch` εξαρτάται μόνον από το χρονικό μήκος (διάρκεια) της επερώτησης και (σε μικρότερο

βαθμό) από των αριθμό του *k*. Επιπλέον, ο `BFMSTSearch` πάντα επιτυγχάνει ισχύ κλαδέματος άνω του 90%. Όσον αφορά στο χρόνο εκτέλεσης, ο `BFMSTSearch` πάντοτε υπερέχει του `DFMSTSearch` λόγω της χρήσης της ευριστικής $MINDISSIM_{INC}$ που απορρίπτει απευθείας όλους τους κόμβους δέντρου που δεν έχουν υποστεί επεξεργασία μέχρι να ολοκληρωθεί.



(a) (b)

Σχήμα 4.14: Κλιμάκωση με τον αριθμό των *k* ομοιότερων τροχιών (Q4)

4.6. Συμπεράσματα

Οι υπάρχουσες σχετικές εργασίες επάνω σε επερωτήσεις ομοιότητας τροχιών βασίζονται στο πλαίσιο της ανάλυσης χρονοσειρών ή του μοντέλου LCSS [VKG02] και του προσφάτως προταθέντος EDR [COO05]. Ωστόσο, το βασικό μειονέκτημα όλων αυτών των μεθόδων είναι ότι είτε αγνοούν τη χρονική διάσταση της κίνησης, υπολογίζοντας συνεπώς τη χωρική (και όχι τη χωροχρονική) ομοιότητα μεταξύ των τροχιών, ή υποθέτουν ότι οι τροχιές είναι του ίδιου μήκους και έχουν τον ίδιο ρυθμό δειγματοληψίας. Επιπλέον, η πλειονότητα των προτεινόμενων προσεγγίσεων εκμεταλλεύονται εξειδικευμένες δομές ευρετηρίου για να κλαδέψουν το χώρο αναζήτησης και να ανακτήσουν την ομοιότερη σε μία τροχιά επερώτησης.

Στη διατριβή αυτή αποδεσμευτήκαμε από αυτές τις υποθέσεις ορίζοντας μια νέα μετρική, που ονομάζεται *DISSIM* και κατόπιν παρουσιάσαμε μία πλήρη επεξεργασία των ιστορικών επερωτήσεων MST σε τροχιές κινούμενων αντικειμένων που αποθηκεύονται σε δομές παρόμοιες με το R-δέντρο αποφεύγοντας τα μειονεκτήματα των υπαρχόντων μεθόδων. Προτείναμε μία σειρά από μετρικές, βάσει απλών εννοιών της τροχιάς, όπως της μέγιστης ταχύτητας του συνόλου δεδομένων, και για την καθεμία από αυτές ακολούθησε ένα λήμμα προς υποστήριξη των στρατηγικών διάταξης και κλαδέματος των αλγορίθμων μας. Κατόπιν παρουσιάσαμε δύο αλγορίθμους MST σε τροχιές που δεικτοδοτούνται από δομές τύπου R-δέντρου ακολουθώντας τα παραδείγματα του «πρώτα στο βαθύτερο» [RKV95] και «πρώτα στον καλύτερο» [HS99].

Για αρκετά συνθετικά και πραγματικά σύνολα δεδομένων τροχιάς, αποδείξαμε την ανωτερότητα της προτεινόμενης μετρικής *DISSIM* σε αντιπαράθεση με άλλους ανταγωνιστές [VKG02], [COO05], από άποψη ποιότητας, ενώ οι αλγόριθμοι μας παρουσιάζουν υψηλή ισχύ κλαδέματος κατά την επεξεργασία επερωτήσεων MST, γεγονός που επαληθεύεται επίσης και για την περίπτωση των *k*-MST. Μεταξύ των προτεινόμενων αλγόριθμων, ο `BFMSTSearch` που ακολουθεί το παράδειγμα του «πρώτα

στον καλύτερο» [HS99] φαίνεται πιο πολλά υποσχόμενος δεδομένου ότι παρουσιάζει καλύτερη απόδοση σε σχέση με τον ανταγωνιστή του `DFMSTSearch`· πιο συγκεκριμένα, επιδεικνύει γραμμική συμπεριφορά από πλευράς χρόνου εκτέλεσης και προσπελάσεων κόμβων, ενώ η ισχύς κλαδέματος που παρέχει είναι άνω του 90% σε όλες τις περιπτώσεις που εξετάσαμε κατά την πειραματική μελέτη (ενώ η ισχύς κλαδέματος του `DFMSTSearch` υποβαθμίζεται σε πολύ χαμηλές τιμές καθώς αυξάνεται το χρονικό μήκος της επερώτησης).

Θα πρέπει εδώ να επισημάνουμε ότι οι προτεινόμενοι αλγόριθμοι δεν απαιτούν κάποια συγκεκριμένη δομή ευρετηρίου και μπορούν να εφαρμοσθούν απευθείας σε οποιοδήποτε μέλος της οικογένειας των R-δέντρων που χρησιμοποιείται για τη δεικτοδότηση τροχιών, όπως για παράδειγμα το 3D R-δέντρο, το TB-δέντρο και το TB*-δέντρο που χρησιμοποιούνται στην εφαρμογή μας.

5. Διαχείριση της Επίδρασης της Αβεβαιότητας Θέσης σε Βάσεις Δεδομένων Τροχιών

Στο κεφάλαιο αυτό παρέχουμε το θεωρητικό μας μοντέλο για την εκτίμηση της επίδρασης της αβεβαιότητας σε χωροχρονικές επερωτήσεις. Το κεφάλαιο διαρθρώνεται ως εξής. Η ενότητα 5.1 δίνει το σκοπό του κεφαλαίου. Η Ενότητα 5.2 παρουσιάζει τις σχετικές εργασίες, ενώ η Ενότητα 5.3 περιγράφει τη θεωρητική ανάλυση της επίδρασης της αβεβαιότητας βάσει διαφόρων υποθέσεων ομοιομορφίας. Στην Ενότητα 5.4 επεκτείνεται το προτεινόμενο μοντέλο για να υποστηρίξει μη ομοιόμορφες κατανομές στις παραμέτρους του προβλήματος. Η Ενότητα 5.5 αξιολογεί την ακρίβεια του μοντέλου μέσω μιας εκτεταμένης πειραματικής μελέτης σε συνθετικά και πραγματικά σύνολα δεδομένων, ενώ η Ενότητα 5.6 συζητά τη χρήση του προτεινόμενου μοντέλου στο πλαίσιο των χωρικών βάσεων δεδομένων. Τέλος, η Ενότητα 5.7 μας δίνει τα συμπεράσματα του κεφαλαίου.

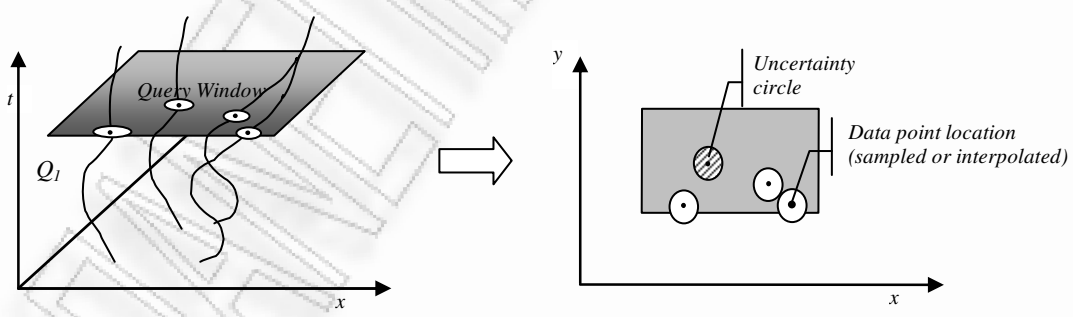
5.1. Εισαγωγή

Μία κοινή υπόθεση που υιοθετείται σε χωρικές και χωροχρονικές βάσεις δεδομένων είναι ότι η θέση των αντικειμένων είναι γνωστή επακριβώς. Ωστόσο, διάφοροι λόγοι, όπως επί παραδείγματι σφάλματα GPS και δειγματοληψίας, μπορούν να επηρεάσουν την ακρίβεια των θέσεων των τροχιών που καταγράφονται, δεδομένου ότι τα δεδομένα θέσεων που λαμβάνονται από συσκευές μέτρησης είναι εγγενώς ανακριβή. Επιπλέον, αρκετές πρόσφατες εργασίες [BS03], [CZBP06], [GL05] προτείνουν ότι θα πρέπει να προστατεύεται το «απόρρητο» της θέσης των κινητών χρηστών προσθέτοντας έναν ελεγχόμενο βαθμό θορύβου στη μετρούμενη θέση κάθε αντικειμένου. Κατά συνέπεια, όλα αυτά τα σφάλματα εισαγάγουν ένα παράγοντα αβεβαιότητας στις απαντήσεις των παραδοσιακών επερωτήσεων επάνω σε τέτοιου είδους δεδομένα.

Η βιβλιογραφία για διαχείριση της αβεβαιότητας θέσης χωροχρονικών αντικειμένων μέχρι τούδε, ασχολείται είτε με θέματα αναπαράστασης της αβεβαιότητας [Tra03], [TWHC04], [WSCY99] ή με πιθανοτικούς αλγορίθμους [CKP04] που επεξεργάζονται επερωτήσεις παρουσία αβεβαιότητας, εκτιμώντας την πιθανότητα κάθε τροχιά να συμπεριλαμβάνεται στα αποτελέσματα της επερώτησης. Από την άλλη πλευρά, σε αυτή τη διατριβή θεωρούμε ότι υπάρχουν περιπτώσεις όπου ο χρήστης θα προτιμούσε να γνωρίζει την επίδραση του σφάλματος μέτρησης στα αποτελέσματα της επερώτησης, χωρίς να εκτελείται στην πράξη η επερώτηση. Η πρόκληση συνεπώς στην οποία καλούμαστε ν' ανταποκριθούμε σε αυτό το κεφάλαιο, είναι *να δώσουμε ένα θεωρητικό πλαίσιο που να εκτιμά το σφάλμα που εισάγεται λόγω της αβεβαιότητας των θέσεων των κινούμενων αντικειμένων στα*

αποτελέσματα των χωροχρονικών επερωτήσεων. Μεταξύ των χωροχρονικών τύπων επερωτήσεων, το ενδιαφέρον μας εστιάζεται στις *επερωτήσεις χρονικής στιγμής*, που μπορούν να χρησιμοποιηθούν για την ανάκτηση των κινούμενων αντικειμένων σε ένα δεδομένο χρονικό σημείο στο παρελθόν και μπορεί να θεωρηθούν ως ειδική περίπτωση των χωροχρονικών επερωτήσεων εύρους, με τη χρονική τους έκταση ίση με μηδέν [PJT00]. Αυτός ο τύπος επερώτησης μπορεί να θεωρηθεί επίσης ως συνδυασμός χωρικής (ήτοι παραθύρου επερώτησης *W*) και χρονικής (ήτοι χρονικού αποτυπώματος *t*) συνιστώσας. Όπως θα συζητήσουμε στο Κεφάλαιο 7, η επέκταση του μοντέλου που παρέχεται από αυτή την εργασία για να υποστηρίξει τις επερωτήσεις εύρους με μη μηδενική χρονική έκταση σε καμία περίπτωση δεν μπορεί να θεωρηθεί απλοϊκός στόχος και το αφήνουμε ως μελλοντική εργασία. Εξ' όσων γνωρίζουμε, η διατριβή μας είναι η πρώτη που ασχολείται με το εν λόγω πρόβλημα.

Για το σκοπό αυτό, αρχικώς υιοθετούμε το μοντέλο που προτείνεται στις [TWHC04], [TWZC02] σχετικά με την αβεβαιότητα των δεδομένων τροχιάς. Πιο συγκεκριμένα η [TWHC04] προτείνει ότι οι τροχιές των κινούμενων αντικειμένων θα πρέπει να μοντελοποιούνται ως 3D κυλινδρικοί όγκοι γύρω από την εντοπισμένη τροχιά (θυμηθείτε το Σχήμα 1.5)· έτσι, όταν εκτελείται μια επερώτηση χρονικής στιγμής στη βάση δεδομένων τροχιών, η αρχική τροχιά μετατρέπεται σε ένα σημείο και ο κυλινδρικός όγκος σε δίσκο (Σχήμα 5.1) που καλείται *δίσκος αβεβαιότητας* και η ακριβής θέση του κινούμενου αντικειμένου σε αυτό το συγκεκριμένο χρονικό αποτύπωμα θεωρείται ότι κατανέμεται ομοιόμορφα εντός του δίσκου. Παρόλο που το μοντέλο που προτείνεται στην [TWHC04] (και κατά συνέπεια, την ομοιόμορφη κατανομή) μπορούμε να το (την) υποθέσουμε όταν εισάγουμε με τεχνητό τρόπο αβεβαιότητα στη θέση των κινούμενων αντικειμένων όπως προτείνεται από την [BS03], [CZBP06], [GL05], είναι μάλλον ουτοπικό να χρησιμοποιηθεί για να περιγράψει το πραγματικό σφάλμα μέτρησης και δειγματοληψίας που εισάγεται από διάφορες συσκευές και μεθόδους παρεμβολής που χρησιμοποιούνται για τον υπολογισμό της θέσης του κινούμενου αντικειμένου μεταξύ διαδοχικών θέσεων δειγματοληψίας. Συνεπώς, στη συνέχεια χρησιμοποιούμε άλλες στατιστικές κατανομές [Lei95], [PTJ05] και επαυξημένα ιστογράμματα προκειμένου να υποστηρίξουμε πιο ρεαλιστικά σενάρια κατανομής αβεβαιότητας.



Σχήμα 5.1: Διατύπωση Προβλήματος

Το μοντέλο που περιγράφεται σε αυτή τη διατριβή μπορεί να χρησιμοποιηθεί σε MODs για να εκτιμήσουμε το μέσο αριθμό λανθασμένων αποτελεσμάτων σε αποτελέσματα επερωτήσεων λόγω της αβεβαιότητας θέσης που υπάρχει στα χωροχρονικά δεδομένα · έτσι, θα μπορούσε να χρησιμοποιηθεί σε ένα διαδραστικό γραφικό δημιουργό / αναλυτή επερωτήσεων, παρέχοντας μία προσέγγιση του ποσοστού των λανθασμένων αποτελεσμάτων λόγω της αβεβαιότητας θέσης καθώς και άλλων εκτιμήσεων, όπως η επιλεκτικότητα, ο χρόνος εκτέλεσης κλπ. Επιπλέον, η προτεινόμενη μεθοδολογία

μπορεί να χρησιμοποιηθεί απευθείας στα υπάρχοντα Συστήματα Διαχείρισης Χωρικών Βάσεων Δεδομένων (Spatial Database Management Systems - SDBMS) προκειμένου να καλύψουν τις ίδιες ανάγκες · στην πράξη, η πλειονότητα των τεχνικών που αναπτύσσονται σε αυτό το κεφάλαιο μπορεί να χρησιμοποιηθεί απευθείας στο πλαίσιο των παραδοσιακών Χωρικών Βάσεων Δεδομένων, δεδομένου ότι η φέτα (slice) μιας χωροχρονικής βάσης δεδομένων στην πραγματικότητα μας δίνει ένα στιγμιότυπο (snapshot) ενός συνόλου σταθερών χωρικών αντικειμένων (Σχήμα 5.1).

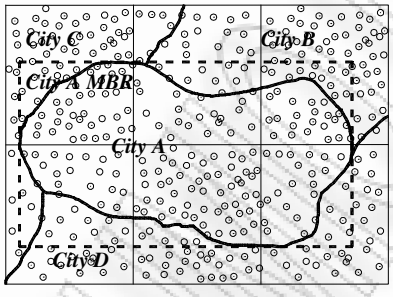
Ένα πιο ζωντανό παράδειγμα που αποδεικνύει πόσες εφαρμογές έχει το προτεινόμενο μοντέλο λαμβάνεται από την εξής κατάσταση, που έχουμε εμπνευστεί από το παράδειγμα των αναδυόμενων ανοικτών αγορών [Ioa07]: έστω ένας χρήστης που θέλει να θέσει μια επερώτηση χρονικής στιγμής σε αρκετές κατανεμημένες πηγές δεδομένων που παρέχουν δεδομένα στους συνδρομητές τους και περιλαμβάνουν τις ίδιες τροχιές, αλλά σε διάφορα επίπεδα αβεβαιότητας, λόγω των διαφορετικών μεθόδων μέτρησης και κατά συνέπεια, διαφορετικών σφαλμάτων· παρά το γεγονός ότι το κριτήριο που χρησιμοποιείται για να επιλέξουμε μεταξύ των είναι η βελτιστοποίηση, ήτοι η ελαχιστοποίηση της αβεβαιότητας που εισάγεται στα τελικά αποτελέσματα της επερώτησης, οι πηγές δεδομένων στη διάρκεια του διαπραγματευτικού βήματος [Ioa07] παρέχουν στους εν δυνάμει πελάτες-χρήστες μόνο συγκεντρωτικά δεδομένα. Υπ' αυτές τις συνθήκες, μόνο το μοντέλο που προτείνεται σε αυτή τη διατριβή μπορεί να προσδώσει στο βελτιστοποιητή της επερώτησης, ο οποίος βρίσκεται στην πλευρά του χρήστη, το σφάλμα που εισάγεται στα αποτελέσματα της επερώτησης για κάθε διαφορετική πηγή δεδομένων.

Επιπλέον, το προτεινόμενο μοντέλο μπορεί να χρησιμοποιηθεί προκειμένου να καθορίσουμε τη *μέγιστη επιτρεπόμενη* (αν)ακρίβεια των τροχιών που θα τροφοδοτήσουν μια MOD (και κατά συνέπεια, ένα SDBMS) δεδομένης της απαιτούμενης ακρίβειας στα αποτελέσματα των επερωτήσεων χρονικής στιγμής (αντίστοιχα, εύρους). Έτσι, το DBMS μπορεί να κατευθύνει τους χρήστες ως προς την κατάλληλη, λιγότερο ή περισσότερο ακριβή – γεγονός που σημαίνει και λιγότερο / περισσότερο δαπανηρή – μέθοδο εντοπισμού θέσης που θα χρησιμοποιηθεί για να συλλεχθούν τα δεδομένα που θα τροφοδοτήσουν το σύστημα.

Πέραν αυτών, η πλέον εξέχουσα εφαρμογή του αναπτυχθέντος μοντέλου είναι επάνω σε συνόψεις, που περιλαμβάνουν μόνο συγκεντρωτικές πληροφορίες αντί για τα πραγματικά αντικείμενα π.χ. τον αριθμό των διακριτών τροχιών εντός μιας χωρικής περιοχής σε ένα χρονικό αποτύπωμα (ή τον αριθμό των χωρικών αντικειμένων εντός μιας δεδομένης χωρικής περιοχής, στην περίπτωση των απλών χωρικών δεδομένων). Έστω, για παράδειγμα, η περίπτωση των Αποθηκών Δεδομένων Τροχιών (Trajectory Data Warehouse - TDW) [MFN+08], όπου η συνάθροιση μπορεί να παρουσιάζει σχέσεις μερικής περιεκτικότητας αντί για τις σχέσεις ολικής περιεκτικότητας που έχουμε συνήθως στις συμβατικές αποθήκες δεδομένων · δηλαδή, ένα χωρικό κελί μπορεί να περιέχεται σε μία πόλη Α κατά 30% και μία πόλη Β κατά 70%. Δεδομένου ότι οι πληροφορίες προ της συνάθροισης αποθηκεύονται μόνο στο κατώτερο επίπεδο της ιεραρχίας της διάστασης «θέση» της αποθήκης δεδομένων ήτοι στα κελιά ή βασικά κυβοειδή, μία διαδικασία γυρίσματος (roll-up) σε επίπεδο *πόλης* σε ένα συγκεκριμένο χρονικό αποτύπωμα, θα συνάθροιζε, μεταξύ άλλων, και έναν αριθμό κελιών που περιέχονται μερικώς στην κάθε πόλη. Η παραπάνω κατάσταση φαίνεται στο Σχήμα 5.2, που παρουσιάζονται τα όρια μεταξύ τεσσάρων πόλεων, A, B, C και D, καθώς και ένα στιγμιότυπο (snapshot) ενός συνόλου αβέβαιων

τροχιών (που έχουν μετατραπεί σε σημεία δεδομένων με τους δίσκους αβεβαιότητάς τους), και ένα κανονικό πλέγμα που αναπαριστά τα κελιά που περιέχουν τις πληροφορίες προ της συγκέντρωσης.

Σε αυτό το πλαίσιο, δεν μπορούμε να εφαρμόσουμε την επιλογή των πιθανοτικών επερωτήσεων, διότι απαιτείται η παρουσία των πραγματικών δεδομένων με την κατανομή της αβεβαιότητάς τους. Από την άλλη πλευρά, το μοντέλο που αναπτύχθηκε σε αυτή τη διατριβή μπορεί και πάλι να εφαρμοσθεί απευθείας χρησιμοποιώντας συγκεντρωτικές πληροφορίες, δηλαδή τον αριθμό των αντικειμένων και των ακτινών του δίσκου αβεβαιότητας ή της τυπικής απόκλισης της κανονικής κατανομής, δίνοντας τελικά μια προσέγγιση του σφάλματος που εισάγεται στα συναθροισμένα αποτελέσματα. Πιο συγκεκριμένα, δεδομένου ότι το μοντέλο μας μπορεί να προσδιορίσει το αποτέλεσμα της αβεβαιότητας θέσης στο *Ελάχιστο Περιβάλλον Παραλληλόγραμμο (Minimum Bounding Rectangle* - MBB) μιας πόλης *A* θεωρώντας το ως επερώτηση εύρους, τότε μπορεί να προσεγγίσει την επίδραση στο πραγματικό χωρικό αντικείμενο *A*, εμπλέκοντας μόνο το πλήθος κάθε κελιού, το MBB και την ακτίνα αβεβαιότητας.



Σχήμα 5.2: Μερική περιεκτικότητα σε αποθήκες δεδομένων τροχιών

Εξ' όσων γνωρίζουμε, δεν υπάρχει θεωρητική μελέτη για τη μοντελοποίηση του σφάλματος που εισάγεται σε χωροχρονικά (ή χωρικά) αποτελέσματα επερώτησης σε όρους λανθασμένων αποτελεσμάτων λόγω της αβεβαιότητας των τροχιών (ή των χωρικών αντικειμένων). Για να σκιαγραφήσουμε τα βασικά θέματα που θα αντιμετωπίσουμε, οι κύριες συνεισφορές του κεφαλαίου είναι οι εξής:

- Αποδεικνύονται δύο λήμματα, που εκτιμούν τον αριθμό των λανθασμένων θετικών και λανθασμένων αρνητικών που προκύπτουν από επερωτήσεις χρονικής στιγμής σε ομοιόμορφα κατανεμημένες αβέβαιες τροχιές που μοντελοποιούνται μέσω της πρότασης [TWHC04] · και τα δύο σφάλματα εξαρτώνται από την ακτίνα του κυλινδρικού όγκου και της περιμέτρου του παραθύρου επερώτησης χρονικής στιγμής και όχι από την επιφάνεια του.
- Για να αποδεσμευτούμε από την υπόθεση της ομοιομορφίας αβεβαιότητας θέσης (η οποία προκύπτει απευθείας από το μοντέλο της [TWHC04]) και να χρησιμοποιήσουμε τη διμεταβλητή κανονική κατανομή που εμφανίζεται στον πραγματικό κόσμο [Lei95], [PTJ05], την προσεγγίζουμε αποτελεσματικά με την κατανομή διαφοράς ομοιομορφίας. Τα αποτελέσματα είναι αρκετά κοντά στην αρχική ανάλυση.
- Χρησιμοποιούμε καινοτόμα χωροχρονικά και άλλα επαυξημένα ιστογράμματα για να εκτιμήσουμε το μέσο αριθμό λανθασμένων αποτελεσμάτων όταν αποδεσμευόμαστε από την υπόθεση ομοιομορφίας της κατανομής των αντικειμένων στο χώρο, καθώς και για να υποστηρίξουμε διάφορες κατανομές της ακτίνας αβεβαιότητας. Η ίδια μεθοδολογία

χρησιμοποιείται επίσης και σε άλλες μορφές συνοπτικών δεδομένων, π.χ. αποθήκες δεδομένων, για να περιγράψουμε το αποτέλεσμα της αβεβαιότητας.

- Εκτελείται μία πλήρης σειρά πειραμάτων που αποδεικνύουν την ορθότητα και την ακρίβεια της ανάλυσης.
- Τέλος, δείχνεται πως μπορούν να εφαρμοσθούν τα αποτελέσματα της ανάλυσης σε χωρικά σύνολα δεδομένων: οι λύσεις που προτείνονται σε αυτό το κεφάλαιο εφαρμόζονται σε ένα εμπορικό SDBMS, εν προκειμένω, το PostgreSQL [Post08a] με χωρική επέκταση PostGIS [Post08b]. Αξίζει εδώ να τονίσουμε ότι τα πιο συνηθισμένα χωρικά ιστογράμματα, τα οποία χρησιμοποιούνται ήδη σε SDBMS για την εκτίμηση της επιλεκτικότητας των επερωτήσεων, υποστηρίζουν το προτεινόμενο μοντέλο χωρίς επιπλέον απαιτήσεις.

5.2. Σχετικές Εργασίες

Οι Wolfson et al. [WSCY99] αντιμετωπίζουν το πρόβλημα της ανακρίβειας στη θέση των κινούμενων αντικειμένων προτείνοντας μια σειρά πολιτικών ενημέρωσης της βάσης δεδομένων που αποθηκεύει τις θέσεις των αντικειμένων. Η βασική ιδέα είναι ότι η βάση δεδομένων ενημερώνεται όποτε η απόσταση μεταξύ της πραγματικής θέσης ενός αντικειμένου και της τιμής που είναι αποθηκευμένη στη βάση δεδομένων υπερβαίνει ένα κατώφλι. Κατ' αυτόν τον τρόπο, εισάγεται ένας παράγοντας αβεβαιότητας στη θέση του κάθε αντικειμένου, καθόσον τα αντικείμενα βρίσκονται σε απόσταση 1Km από τις τελευταίες καταγεγραμμένες θέσεις. Υιοθετώντας τη χρήση των συναρτήσεων πυκνότητας πιθανότητας (probability density function - *pdf*), περιγράφουν έναν αλγόριθμο που επεξεργάζεται μία πιθανοτική επερώτηση χωρικού εύρους που εφαρμόζεται στην παραπάνω βάση δεδομένων. Το αποτέλεσμα αυτής της μορφής επερώτησης, που μας δίνει το σύνολο των αντικειμένων που είναι εντός μιας περιοχής R, αποτελείται από ζεύγη της μορφής (O_i, P_i), όπου P_i είναι η πιθανότητα το αντικείμενο O_i να τέμνει την περιοχή επερώτησης R. Οι Cheng et al. [CKP04] υιοθετούν τον ορισμό της πιθανοτικής επερώτησης που εισάγεται στην [WSCY99] και τον προεκτείνουν στην περίπτωση των επερωτήσεων πλησιέστερου γείτονα (NN).

Οι Pfoser και Jensen [PJ99] προτείνουν μία αναπαράσταση της αβεβαιότητας θέσης λόγω σφαλμάτων μέτρησης και δειγματοληψίας. Εκεί, η χωρική προβολή της τροχιάς ενός αντικειμένου μοντελοποιείται ως μία 2D ελλειπτική επιφάνεια, που ορίζεται από δύο διαδοχικές θέσεις του αντικειμένου. Παρουσιάζουν επίσης την επίδραση της αβεβαιότητας θέσης στην επεξεργασία των πιθανοτικών επερωτήσεων εύρους και προτείνουν μία μέθοδο προεπιλογής και διαλογής για την απάντησή τους.

Η αβεβαιότητα θέσης των κινούμενων αντικειμένων αναφέρεται επίσης από τους Trajcevski et al. [Tra03], [TWHC04], όπου μια τροχιά ενός αντικειμένου μοντελοποιείται ως ένας 3D κυλινδρικός όγκος γύρω από την τροχιά. Στην ίδια εργασία, εισάγονται δύο κατηγορίες τελεστών για τροχιές επερωτήσεων με αβεβαιότητα, σε επερωτήσεις χωροχρονικού σημείου και εύρους, ενώ επιπλέον παρουσιάζονται και αποτελεσματικοί αλγόριθμοι για την εφαρμογή των προτεινόμενων τελεστών.

Οι Ni et al. [NRB03] προτείνουν ένα πιθανοτικό μοντέλο χωρικών δεδομένων για την ακρίβεια της θέσης πολυγωνικών δεδομένων. Βάσει αυτού του μοντέλου, κάθε πολύγωνο χωρίζεται σε ξένα μεταξύ τους κομμάτια. Κάθε κομμάτι είναι μία σειρά κορυφών με πλήρως συσχετισμένες

αβεβαιότητες θέσης. Βάσει του παραπάνω μοντέλου, περιγράφεται ένας πιθανοτικός αλγόριθμος χωρικής ένωσης, στον οποίο τα ζεύγη του αντικειμένου του αποτελέσματος συσχετίζονται με την πιθανότητα τομής μεταξύ κάθε ζεύγους. Εισάγεται μία παραλλαγή του R-δέντρου, που ονομάζεται πιθανοτικό R-δέντρο, προς υποστήριξη της πιθανοτικής προεπιλογής (filtering) του αλγορίθμου ένωσης, στον οποίο κάθε προσέγγιση MBB του δένδρου επαυξάνεται με την κατανομή πιθανότητας του ορίου του MBB.

Οι Cheng et al. [CXP+04] ερεύνησαν το πρόβλημα της δεικτοδότησης αβέβαιων δεδομένων για να απαντήσουν αποτελεσματικά σε κατωφλιωμένες πιθανοτικές επερωτήσεις, στις οποίες το αποτέλεσμα της επερώτησης περιλαμβάνει μόνο σημεία με πιθανότητα που υπερβαίνει ένα δεδομένο κατώφλι. Προτείνονται δύο δομές δεικτοδότησης. Η ισχύς κλαδέματος του πρώτου ευρετηρίου βασίζεται στη χρήση των αβέβαιων πληροφοριών που επαυξάνουν στους εσωτερικούς κόμβους του ευρετηρίου, ενώ στο δεύτερο ευρετήριο τα σημεία δεδομένων με παρόμοιους βαθμούς αβεβαιότητας συσταδοποιούνται μαζί. Προσφάτως οι Tao et al. [TCX+05] μελέτησαν έναν παρόμοιο τύπο επερώτησης, την πιθανοτική επερώτηση εύρους, που ανακτά τα αντικείμενα που εμφανίζονται σε μία ορθογώνια επιφάνεια με πιθανότητα τουλάχιστον μιας προκαθορισμένης τιμής. Εισήγαγαν μία ολότελα δυναμική δομή ευρετηρίου για αβέβαια δεδομένα. Η δομή αυτή, που ονομάζεται U-δέντρο, διατηρεί «βοηθητικές πληροφορίες» σε όλα του τα επίπεδα για τα αντίστοιχα δεικτοδοτημένα αντικείμενα που μπορούν να χρησιμοποιηθούν για την επαλήθευση της παρουσίας ενός αντικειμένου στα αποτελέσματα μιας πιθανοτικής επερώτησης εύρους, χωρίς τον υπολογισμό της δαπανηρής από πλευράς υπολογισμών πιθανότητας εμφάνισης.

Οι Dai et al. [DYM+05] μελέτησαν το πρόβλημα της αξιολόγησης χωρικών επερωτήσεων για υπαρξιακά αβέβαια δεδομένα· σ' αυτή την περίπτωση, η αβεβαιότητα δεν αναφέρεται σε θέσεις των αντικειμένων αλλά στην ύπαρξή τους. Οι γράφοντες καθορίζουν δύο πιθανοτικούς τύπους επερωτήσεων: τις λεγόμενες επερωτήσεις κατωφλίου και ταξινόμησης (thresholding and ranking) στην οποία το αποτέλεσμα ελέγχεται είτε θέτοντας ένα κατώφλι στα αποτελέσματα χαμηλής πιθανότητας ή ταξινομώντας τα και επιλέγοντας αυτά με τη υψηλότερη πιθανότητα αντίστοιχα. Εν συνεχεία, παρουσιάζονται πιθανοτικές παραλλαγές των επερωτήσεων χωρικού εύρους και NN για αντικείμενα που δεικτοδοτούνται από ένα 2D ευρετήριο, όπως το R-δέντρο. Τέλος, για να βελτιώσουν την αποτελεσματικότητα των προτεινόμενων αλγορίθμων τους, προτείνουν μια προέκταση του R-δέντρου, στην οποία οι εγγραφές των εσωτερικών κόμβων επαυξάνονται με την μέγιστη υπαρξιακή πιθανότητα των αντικειμένων που δεικτοδοτούνται από αυτές.

Η μελέτη που ίσως σχετίζεται περισσότερο με τη δουλειά μας είναι αυτή των Yu και Mehrotra [YM03], όπου συζητείται το αποτέλεσμα της αβεβαιότητας σε πιθανοτικές χωρικές επερωτήσεις, ομοίως προς τα όσα παρουσιάζονται στην [NRB03]. Μέσω μιας θεωρητικής ανάλυσης, παρέχεται μία νέα τεχνική που μπορεί να χρησιμοποιηθεί ώστε να καθορίσει την απαιτούμενη ακρίβεια των δεδομένων που θα συλλεχθούν, προκειμένου να δοθεί μία πιθανοτική εγγύηση για την αβεβαιότητα στις απαντήσεις χωρικών επερωτήσεων. Το πρώτο αποτέλεσμα της ανάλυσης είναι το πλήθος των τριών υποσυνόλων του αποτελέσματος μίας επερώτησης εύρους, πιο συγκεκριμένα των συνόλων *MUST*, *MAY* και *ANS*: το *MUST* είναι το σύνολο των αντικειμένων που σε κάθε περίπτωση «πρέπει» να εντοπισθούν εντός του παραθύρου της επερώτησης, το *MAY* είναι το σύνολο των αντικειμένων που

«ίσως» εντοπισθούν εντός του παραθύρου επερώτησης και το *ANS* είναι η προσεγγιστική απάντηση του συνόλου των αντικειμένων των οποίων οι καταγεγραμμένες θέσεις είναι στην περιοχή της επερώτησης. Το δεύτερο αποτέλεσμα είναι μία μέθοδος για να καθορίσουμε τη μεγαλύτερη δυνατή ανακρίβεια δηλαδή την *ακτίνα αβεβαιότητας* της ανάλυσής μας, δεδομένου ότι η απάντηση σε μία τυχαία επερώτηση *COUNT* θα πρέπει να περιλαμβάνει μία αβεβαιότητα $\delta \leq \delta_0$, ήτοι το πλήθος του συνόλου *MAY* να είναι λιγότερο από μία τιμή, με πιθανότητα $P \geq P_0$.

Συγκρίνοντας το προτεινόμενο σε αυτή τη διατριβή μοντέλο με την [YM03], η πρώτη παρατήρηση είναι ότι οι αριθμοί E_N και E_P των λανθασμένων αποτελεσμάτων είναι στην πράξη μία *διαλογή (refinement),* ήτοι ένα υποσύνολο του συνόλου *MAY* που εκτιμάται από την [YM03] και δεν γίνεται να απομακρύνουμε απευθείας την υπερεκτίμηση που δίνεται από την [YM03] εκτός εάν χρησιμοποιείται το μοντέλο μας· αυτή η υπερεκτίμηση παρουσιάζεται σαφώς στα πειραματικά αποτελέσματα που παρουσιάζονται στην Ενότητα 5.6.2.1. Μία δεύτερη παρατήρηση είναι ότι το μοντέλο που παρουσιάστηκε στην [YM03] βασίζεται στην υπόθεση της ομοιομορφίας, ενώ η μελέτη μας ασχολείται με πιο ρεαλιστικές απαιτήσεις.

5.3. Μοντελοποίηση του Σφάλματος λόγω Αβεβαιότητας Θέσης

Έστω ένα σύνολο δεδομένων P που αποτελείται από N τροχιές T_i, $i = 1, \ldots, N$, που είναι ομοιόμορφα κατανεμημένες στο μοναδιαίο χωροχρονικό διάστημα $S = [0,1] \times [0,1] \times [0,1]$, δηλαδή, όλες οι διαστάσεις κανονικοποιούνται στο διάστημα μεταξύ του 0 και του 1. Αρχικά δίνουμε την έννοια των ομοιόμορφα κατανεμημένων τροχιών: *ένα σύνολο τροχιών P είναι ομοιόμορφα κατανεμημένο όταν και μόνο όταν οι θέσεις των κινούμενων αντικειμένων που λαμβάνονται από ένα στιγμιότυπο του P σε ένα τυχαίο χρονικό αποτύπωμα παράγει ένα σύνολο σημείων $T_{i,k}$, $i = 1, \ldots, N$, και k=1…now το οποίο είναι ομοιόμορφα κατανεμημένο*. Επιπλέον, το προϊόν της ίδιας στιγμιαίας αποτύπωσης του S είναι ο χώρος $S_k = [0,1] \times [0,1]$.

Βάσει της [TWHC04], οι τροχιές των κινούμενων αντικείμενων θα πρέπει να μοντελοποιούνται ως κυλινδρικοί όγκοι σταθερής ακτίνας d γύρω από τις πραγματικές δειγματοληπτούμενες θέσεις των κινούμενων αντικειμένων και την αντίστοιχη παρεμβαλλόμενη τροχιά. Έτσι, μια στιγμιαία εικόνα της τροχιάς T_i στο χρονικό αποτύπωμα t_k παράγει έναν *δίσκο αβεβαιότητας* με κέντρο $T_{i,k}$ και ακτίνα d, εντός της οποίας η *πραγματική* θέση $T_{i,k}^{\dagger}$ της τροχιάς T_i στο χρονικό αποτύπωμα t_k, είναι ομοιόμορφα κατανεμημένη. Έστω επίσης R το σύνολο όλων των επερωτήσεων χρονικής στιγμής που τέθηκαν στο σύνολο δεδομένων P, R_k το υποσύνολο του R στο χρονικό αποτύπωμα t_k, και $R_{k,a\times b}$ το υποσύνολο του R_k που περιέχει όλες τις επερωτήσεις χρονικής στιγμής που έχουν πλευρές μήκους $2a$ και $2b$ κατά τον άξονα x- και y-, αντίστοιχα.

Δύο τύποι σφαλμάτων εισάγονται όταν εκτελείται μία επερώτηση χρονικής στιγμής $W_k \in R_{k,a\times b}$ στο σύνολο δεδομένων P:

- Το E_N είναι το σύνολο των *λανθασμένων αρνητικών (false negatives)*, ήτοι, τροχιών που περνούν από το παράθυρο επερώτησης αλλά δεν ανακτώνται· επισήμως, $E_N = \left\{ T_i \in P : T_{i,k} \notin W_j \mid T_{i,k}^{\dagger} \in W_j \right\}$, και

- E_P είναι το σύνολο των *λανθασμένων θετικών (false positives)*, ήτοι, τροχιές που ανακτώνται παρόλο που δεν περνούν από το παράθυρο επερώτησης· επισήμως, $E_P = \left\{ T_i \in P : T_{i,k} \in W_j \mid T_{i,k}^{\dagger} \notin W_j \right\}$.

Πίνακας 5.1: Πίνακας συμβόλων

Σύμβολο	*Περιγραφή*
S, P, N	το μοναδιαίο χωροχρονικό διάστημα δεδομένων $[0,1]\times[0,1]\times[0,1]$ το σύνολο των τροχιών και το πλήθος του (επίσης, πυκνότητα)
t_k, S_k	ένα χρονικό αποτύπωμα και η στιγμιαία εικόνα του S στο χρονικό αποτύπωμα t_k
T_i, $T_{i,k}$, $T_{i,k}^{\dagger}$, d	Μια τροχιά, η (καταγεγραμμένη ή παρεμβαλλόμενη) θέση του T_i στο χρονικό αποτύπωμα t_k, η πραγματική θέση του και η ακτίνα του δίσκου αβεβαιότητας
W_j, $W_{j,c1}$ - $W_{j,c4}$	το παράθυρο μιας επερώτησης χρονικής στιγμής και οι τέσσερις γωνίες του (με τη φορά των δεικτών του ρολογιού, με πρώτη την κάτω αριστερά)
$W_j^{x,L}, W_j^{x,U}, W_j^{y,L}, W_j^{y,U}$	οι μέγιστες και ελάχιστες τιμές του παραθύρου επερώτησης χρονικής στιγμής W_j κατά μήκος των αξόνων x- και y-.
R, R_k, $R_{k,a\times b}$	το σύνολο όλων των επερωτήσεων χρονικής στιγμής στο P, το υποσύνολο του που καλείται στο χρονικό αποτύπωμα t_k, και το υποσύνολό του με μήκος μισής πλευράς a και b κατά μήκος των αξόνων x- και y-, αντιστοίχως
$C(T_{i,k},d)$, $A_{i,j}$	ο δίσκος αβεβαιότητας της (καταγεγραμμένης ή παρεμβαλλόμενη) θέσης του T_i στο χρονικό αποτύπωμα t_k με ακτίνα d και το τμήμα της επιφάνειας του που βρίσκεται εντός (για την περίπτωση των λανθασμένων αρνητικών) ή εκτός (στην περίπτωση των λανθασμένων θετικών) W_j
$Dist\left(T_{i,k}, W_j\right)$	η ελάχιστη Ευκλείδεια απόσταση μεταξύ της (καταγεγραμμένης ή παρεμβαλλόμενης) θέσης του T_i στο χρονικό αποτύπωμα t_k και το όριο του W_j
r_x, r_y	η απόσταση προς το πλησιέστερο σημείο $T_{i,k}$ του ορίου του W_j κατά μήκος των αξόνων x- και y-, αντίστοιχα.
$A_{1x}(r_x, r_y)$ $A_{1y}(r_x, r_y)$	η επιφάνεια που περικλείεται από μία χορδή κατακόρυφη στον x- (ή y-) άξονα με r_x (ή r_y) απόσταση αντίστοιχα από το $T_{i,k}$ και το αντίστοιχο τόξο του δίσκου αβεβαιότητας
$A_2(r_x, r_y)$	η αλληλεπικαλυπτόμενη περιοχή μεταξύ του δίσκου αβεβαιότητας της $T_{i,k}$ και μίας γωνίας του παραθύρου που βρίσκεται εντός του δίσκου με συντεταγμένες r_x και r_y σε σχέση προς την $T_{i,k}$.
$V_{i,j}$, $V_{1x}(r_x, r_y)$, $V_{1y}(r_x, r_y)$, $V_2(r_x, r_y)$	οι όγκοι των κωνικών τμημάτων, που ισοδυναμούν με τις επιφάνειες $A_{i,j}$, $A_{1x}(r_x, r_y)$, $A_{1y}(r_x, r_y)$, $A_2(r_x, r_y)$ όταν ακολουθείται η υπόθεση διαφοράς ομοιομορφίας της αβεβαιότητας.
$AvgP_{i,P}(R_{k,a\times b})$, $AvgP_{i,N}(R_{k,a\times b})$	η μέση πιθανότητα μια τροχιά T_i να είναι λανθασμένη θετική (ή λανθασμένη αρνητική) σε σχέση προς όλα τα παράθυρα επερώτησης $W_i \in R_{k,a\times b}$
$E_P\left(R_{k,a\times b}\right)$, $E_N\left(R_{k,a\times b}\right)$	ο μέσος αριθμός λανθασμένων θετικών (ή λανθασμένων αρνητικών) στα αποτελέσματα μιας επερώτησης χρονικής στιγμής $W_j \in R_{k,a\times b}$

Το πρόβλημα είναι να γίνει μία όσο το δυνατόν πιο ακριβής εκτίμηση των λανθασμένων αρνητικών και λανθασμένων θετικών για ένα τυχαίο W_j στο χρονικό αποτύπωμα t_k, βασισμένο μόνο σε γνωστά παραμέτρους του συνόλου δεδομένων και της επερώτησης. Από τον παραπάνω ορισμό του προβλήματος, είναι σαφές ότι αρχικώς κάνουμε τέσσερις κύριες υποθέσεις:

- A_I – *υπόθεση ομοιομορφίας αβεβαιότητας θέσης (uncertainty uniformity assumption)*: η πραγματική θέση $T_{i,k}^{\dagger}$ της τροχιάς T_i στο χρονικό αποτύπωμα t_k κατανέμεται ομοιόμορφα εντός του δίσκου αβεβαιότητας $C(T_{i,k},d)$,
- A_{II} – *υπόθεση ομοιομορφίας δεδομένων (data uniformity assumption)*: οι τροχιές T_i (και κατά συνέπεια, τα σημεία $T_{i,k}$ στο χρονικό αποτύπωμα t_k), κατανέμονται ομοιόμορφα στο χώρο δεδομένων,
- A_{III} – *υπόθεση σταθερής ακτίνας αβεβαιότητας (constant uncertainty radius assumption)*: η ακτίνα d του κυλινδρικού όγκου (και κατά συνέπεια, του δίσκου αβεβαιότητας) είναι σταθερή, και, χωρίς να προκύπτει απευθείας από τον ορισμό του προβλήματος,
- A_{IV} – *υπόθεση μεγέθους αβεβαιότητας (uncertainty size assumption)*: η ακτίνα d είναι πάντα λιγότερη από το ήμισυ του μήκους της μικρότερης πλευράς του παραθύρου επερώτησης W_j.

Όσον αφορά τις πρώτες τρεις υποθέσεις (A_I - A_{III}), θα αποδεσμευτούν στην επέκταση του μοντέλου που θα παρουσιαστεί στην Ενότητα 5.4. Για την υπόθεση A_{IV}, θεωρούμε ότι αυτή είναι μια λογική ιδιότητα των εμπλεκομένων χωρικών αντικειμένων, διότι τα τυπικά μεγέθη του παραθύρου επερώτησης W_j είναι συνήθως τάξεις μεγέθους μεγαλύτερα από την d· για παράδειγμα, οι τροχιές που δειγματοληπτούνται με συσκευές GPS εισάγουν ένα σφάλμα μερικών μέτρων (συνήθως λιγότερο των 10m), ενώ τα παράθυρα επερώτησης σε πραγματικές εφαρμογές αναμένεται να είναι τουλάχιστον εκατοντάδες τετραγωνικά μέτρα.

Μετά την περιγραφή του πλαισίου εργασίας μας, στην επόμενη ενότητα αποδεικνύουμε δύο λήμματα που είναι θεμελιώδη για το μοντέλο μας. Ο Πίνακας 5.1 συνοψίζει τα σύμβολα που χρησιμοποιούνται στη συνέχεια του κεφαλαίου.

5.3.1. Εκτίμηση του Αριθμού των Λανθασμένων Αρνητικών

Στην ενότητα αυτή αποδεικνύουμε ένα λήμμα βάσει του οποίου γίνεται ο υπολογισμός του μέσου αριθμού λανθασμένων αρνητικών.

Λήμμα 5.1*: Ο μέσος αριθμός $E_N(R_{k,a\times b})$ των λανθασμένων αρνητικών στα αποτελέσματα μιας επερώτησης χρονικής στιγμής $W_j \in R_{k,a\times b}$ με πλευρές μήκους $2a$ και $2b$ στο χρονικό αποτύπωμα t_k σε ένα σύνολο δεδομένων τροχιών που ακολουθεί τις υποθέσεις ομοιομορφίας δεδομένων και ομοιομορφίας αβεβαιότητας δίνεται από τον τύπο:*

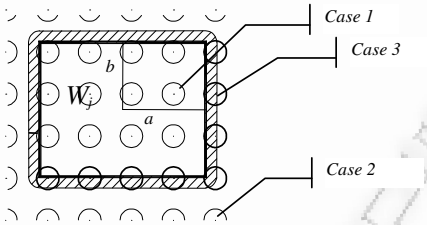
$$E_N\left(R_{k,a\times b}\right) = N \cdot \left(\frac{8d}{3\pi}(a+b) - \frac{d^2}{2\pi}\right) \qquad (5.1)$$

όπου d είναι η ακτίνα του δίσκου αβεβαιότητας.

Απόδειξη: Ο μέσος αριθμός $E_N(R_{k,a\times b})$ των τροχιών που είναι λανθασμένες αρνητικές στα αποτελέσματα της επερώτησης χρονικής στιγμής $W_j \in R_{k,a\times b}$, ήτοι, $T_{i,k} \notin W_j \mid T_{i,k}^{\dagger} \in W_j$, μπορεί να ληφθεί από τη μέση πιθανότητα $AvgP_{i,N}(R_{k,a\times b})$ μια αυθαίρετη τροχιά T_i να είναι λανθασμένη αρνητική σε σχέση με ένα αυθαίρετο παράθυρο επερώτησης $W_j \in R_{k,a\times b}$, πολλαπλασιασμένο με το συνολικό αριθμό N των τροχιών:

$$E_N\left(R_{k,a\times b}\right) = N \cdot AvgP_{i,N}\left(R_{k,a\times b}\right) \qquad (5.2)$$

Προφανώς, ο στόχος μας είναι να προσδιορίσουμε την $AvgP_{i,N}(R_{k,a\times b})$. Για να επιτύχουμε το στόχο μας, διατυπώνουμε την πιθανότητα να συμβαίνει το $T_{i,k} \notin W_j \mid T_{i,k}^{\dagger} \in W_j$. Αυτή η πιθανότητα δίνεται από το λόγο της επιφάνειας $A_{i,j}$ του τμήματος του δίσκου αβεβαιότητας $C(T_{i,k},d)$ που περιλαμβάνεται εντός του παραθύρου επερώτησης, προς στην συνολική επιφάνεια του $C(T_{i,k},d)$. Ωστόσο, το $A_{i,j}$ είναι μηδέν σε περιπτώσεις όπου το $C(T_{i,k},d)$ δεν αλληλεπικαλύπτεται με το όριο της επερώτησης.



Σχήμα 5.3: Στιγμιαία αποτύπωση των τροχιών που συνεισφέρουν στον αριθμό των λανθασμένων αρνητικών

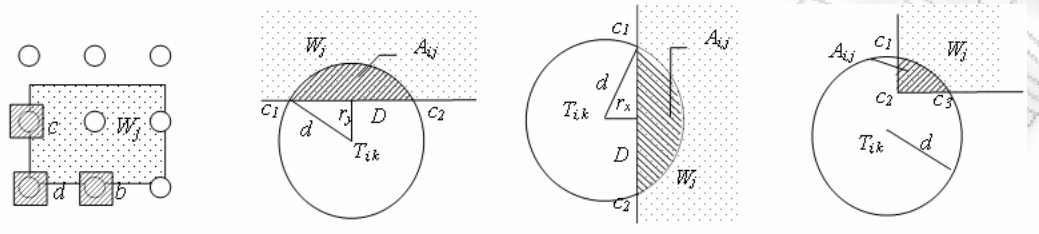
Το Σχήμα 5.3 απεικονίζει ένα παράθυρο επερώτησης χρονικής στιγμής W_j που επεκτείνεται από μία ζώνη επιρροής (buffer) πλάτους d, σε ένα υποσύνολο ομοιόμορφα κατανεμημένων σημείων, που αντιστοιχούν σε ένα στιγμιότυπο του P κοντά στο W_j στο χρονικό αποτύπωμα t_k: οι τροχιές που συμβολίζονται ως σημεία με δίσκους αβεβαιότητας και βρίσκονται εντός του παραθύρου επερώτησης, ήτοι αυτές που συμβολίζονται ως "case 1", δεν μπορούν να είναι λανθασμένες αρνητικές γιατί θα ανακτηθούν τελικά από την επερώτηση. Το ίδιο ισχύει και για τα σημεία με δίσκους αβεβαιότητας που βρίσκονται εκτός της περιοχής της ζώνης επιρροής και συμβολίζονται ως "case 2" στο Σχήμα 5.3. Η μόνη περίπτωση όπου η $T_{i,k}$ δεν ανακτάται από την επερώτηση ενώ η $T_{i,k}^{\dagger}$ μπορεί να βρεθεί εντός του W_j είναι όταν η $T_{i,k}$ βρίσκεται εντός της περιοχής της ζώνης επιρροής που περιβάλλει το W_j, που απεικονίζεται ως "case 3" στο Σχήμα 5.3.

Η παραπάνω συζήτηση εκφράζει το γεγονός ότι η τροχιά T_i είναι υποψήφια να είναι λανθασμένη αρνητική αν και μόνον αν η $T_{i,k}$ βρίσκεται εκτός του παραθύρου επερώτησης, ενώ ο αντίστοιχος δίσκος αβεβαιότητας $C(T_{i,k},d)$ τέμνει το όριο της επερώτησης. Εναλλακτικά, η $T_{i,k}$ θα πρέπει να βρίσκεται εντός της *περιοχής Minkowski (Minkowski region)* του W_j με ακτίνα d προκειμένου να είναι υποψήφια για να είναι λανθασμένη αρνητική· η περιοχή αυτή μπορεί να προσδιορισθεί επεκτείνοντας το W_j με απόσταση d προς όλες τις κατευθύνσεις [TZPM04]. Οι περιοχές Minkowski προκύπτουν απευθείας από την έννοια του *αθροίσματος Minkowski (Minkowski sum)* [AFH02] ανάμεσα στο παράθυρο επερώτησης W_j και ένα δίσκο ακτίνας d, ο οποίος, στην περίπτωσή μας, αποτελείται από ένα σύνολο γραμμικών τμημάτων και κυκλικών τόξων, που απεικονίζεται ως το όριο που είναι στο εξωτερικό του W_j στο Σχήμα 5.3. Τώρα, η πιθανότητα της τροχιάς T_i να είναι λανθασμένη αρνητική, σε σχέση μ' ένα παράθυρο επερώτησης W_j, είναι:

$$P\left(T_{i,k} \notin W_j \mid T_{i,k}^{\dagger} \in W_j\right) = \begin{cases} \dfrac{A_{i,j}}{\pi d^2}, & \alpha\nu\, T_{i,k} \notin W_j \;\kappa\alpha\iota\; Dist\left(T_{i,k}, W_j\right) \le d \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{5.3}$$

Η επιφάνεια $A_{i,j}$, που απεικονίζεται στο Σχήμα 5.4, καθορίζεται, λαμβάνοντας υπ' όψιν την υπόθεση μεγέθους αβεβαιότητας, διακρίνοντας μεταξύ τριών περιπτώσεων που απεικονίζονται στο Σχήμα 5.4(b) – (d): Στις πρώτες δύο περιπτώσεις, όπου η απόσταση μεταξύ $T_{i,k}$ και καθεμίας από τις δύο

γωνίες του W_j είναι μεγαλύτερη από το d, $A_{i,j}$ είναι το τμήμα του δίσκου αβεβαιότητας που περικλείεται από (a) τη χορδή c_1c_2 που σχηματίζεται από την πλευρά της επερώτησης και (b) το αντίστοιχο τόξο $\overset{\frown}{c_1c_2}$. Έτσι, μπορεί να υπολογιστεί ως το ολοκλήρωμα της συνάρτησης του μήκους χορδής D, που δίνεται ως έκφραση της απόστασής του, r_y ή r_x (ανάλογα με το ποια πλευρά της επερώτησης εξετάζεται) από το κέντρο του δίσκου.



(a) (b) (c) (d)

Σχήμα 5.4: Ο μοναδιαίος χώρος (a) και τρεις λεπτομέρειες του (b, c, d)

Έστω ότι η χορδή c_1c_2 είναι παράλληλη στον άξονα x (Σχήμα 5.4(b)), ισχύει ότι $D(r_x, r_y) = 2\sqrt{d^2 - r_y^2}$ και $A_{i,j} = A_{1y}(r_x, r_y) = \int_{r_y}^{d} D(r_x, r_y)dr_y = 2\int_{r_y}^{d} \sqrt{d^2 - r_y^2}\, dr_y$, που έχει ως αποτέλεσμα[1]:

$$A_{i,j} = A_{1y}(r_x, r_y) = d^2 \arctan\left[\sqrt{\left(\frac{d}{r_y}\right)^2 - 1}\right] - r_y\sqrt{d^2 - r_y^2} \tag{5.4}$$

Αντιστοίχως, έστω ότι η χορδή c_1c_2 είναι παράλληλη στον άξονα y (Σχήμα 5.4(c)), η επιφάνεια $A_{i,j} = A_{1x}(r_x, r_y)$ υπολογίζεται αντικαθιστώντας το r_y με το r_x στην Εξ.(5.4). Στην τρίτη περίπτωση, όπου η απόσταση μεταξύ του $T_{i,k}$ και μίας από τις τέσσερις γωνίες του W_j είναι λιγότερο του d (Σχήμα 5.4(d)), η $A_{i,j}$ μπορεί να προσδιοριστεί με παρόμοιο τρόπο δίνοντας τελικά:

$$A_{i,j} = A_2(r_x, r_y) =$$
$$\frac{1}{2}\left(d^2 \operatorname{arccot}\left(\sqrt{\frac{r_y}{R^2 - r_y^2}}\right) - d^2 \arctan\left(\sqrt{\frac{r_x}{R^2 - r_x^2}}\right) - r_y\sqrt{d^2 - r_y^2} - r_x\sqrt{d^2 - r_x^2} + 2r_xr_y\right) \tag{5.5}$$

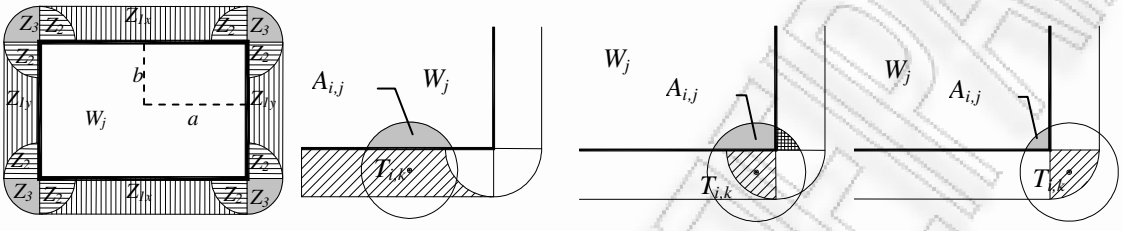
Η μέση, σε σχέση με οποιοδήποτε παράθυρο επερώτησης στην $R_{k,axb}$, πιθανότητα μιας τροχιάς T_i να είναι λανθασμένη αρνητική υπολογίζεται ολοκληρώνοντας την Εξ.(5.3) σε όλα τα πιθανά παράθυρα επερώτησης:

$$AvgP_{i,N}\left(R_{k,a\times b}\right) = \int_{W_j \in R_{k,a\times b}} P\left(T_{i,k} \notin W_j \mid T_{i,k}^{\dagger} \in W_j\right) dW = \iint_{S_k} P\left(T_{i,k} \notin W_j \mid T_{i,k}^{\dagger} \in W_j\right) dxdy \tag{5.6}$$

Για να υπολογίσουμε το παραπάνω ολοκλήρωμα, είναι απαραίτητο, να καθορίσουμε τις κύριες περιοχές εντός των οποίων η επιφάνεια $A_{i,j}$ μπορεί να εκφραστεί ως απλή συνάρτηση. Για να διευκολύνουμε τη συζήτηση, το Σχήμα 5.5(a) δείχνει το γεγονός ότι η επιφάνεια που καθορίζεται από την $Dist(T_{i,k},W_j) \le d$ μπορεί να χωριστεί σε τρία σύνολα ζωνών εντός των οποίων μπορεί να βρεθεί το σημείο $T_{i,k}$ σχετικά με τη θέση του παραθύρου επερώτησης: το πρώτο με τις κάθετες γραμμές, το δεύτερο με τις οριζόντιες και το σκιασμένο, που καλούνται Z_1, Z_2 και Z_3, αντίστοιχα. Οι περιοχές Z_1 περιέχουν τέτοια σημεία δεδομένων που η επιφάνεια που προκύπτει από την τομή της επιφάνειας αβεβαιότητάς τους με το W_j σχηματίζει ένα πλήρες κυκλικό τμήμα · εναλλακτικά, οι περιοχές Z_1 είναι

[1] Όλοι οι προηγμένοι υπολογισμοί στο κεφάλαιο αυτό εκτελέστηκαν χρησιμοποιώντας λογισμικό *Mathematica* [28].

ο γεωμετρικός τόπος των σημείων του χώρου που είναι εκτός του W_j, η απόσταση τους από το W_j είναι μικρότερη της d και η απόστασή τους από τις τέσσερις γωνίες του W_j είναι μεγαλύτερη της d. Οι περιοχές Z_2 είναι ο γεωμετρικός τόπος των σημείων του χώρου όπου τα σημεία είναι εκτός του W_j, η απόστασή τους από το W_j είναι μικρότερη της d και οι συντεταγμένες τους x και y είναι εντός της προβολής του W_j κατά μήκος του άξονα x- ή y-, αντιστοίχως· ομοίως, οι περιοχές Z_3 διαφέρουν μόνο ως προς το ότι οι συντεταγμένες x και y των σημείων τους είναι εκτός της προβολής του W_j κατά μήκος του x- ή y- άξονα.



Σχήμα 5.5: Περιοχές όπου η επιφάνεια $A_{i,j}$ που συνεισφέρει σε λανθασμένες αρνητικές απαντήσεις εκφράζεται ως απλή συνάρτηση

Οι περιοχές $Z_{1,j}$, $Z_{2,j}$ και $Z_{3,j}$ που σχετίζονται με το παράθυρο επερώτησης W_j ορίζονται και επισήμως από τους ακόλουθους τύπους:

$$Z_{1,j} = \left\{ T_i \in P : T_{i,k} \notin W_j \wedge Dist\left(T_{i,k}, W_j\right) \le d \wedge Dist\left(T_{i,k}, W_{j,c_i}\right) \ge d, i = 1..4 \right\} \quad (5.7)$$

$$Z_{2,j} = \left\{ \begin{array}{c} T_i \in P : T_{i,k} \notin W_j \wedge Dist\left(T_{i,k}, W_j\right) \le d \wedge Dist\left(T_{i,k}, W_{j,c_i}\right) \le d, i = 1..4 \wedge \\ \left(T_{i,k}^x \in \left[W_j^{x,L}, W_j^{x,U}\right] \vee T_{i,k}^y \in \left[W_j^{y,L}, W_j^{y,U}\right]\right) \end{array} \right\} \quad (5.8)$$

$$Z_{3,j} = \left\{ \begin{array}{c} T_i \in P : T_{i,k} \notin W_j \wedge Dist\left(T_{i,k}, W_j\right) \le d \wedge Dist\left(T_{i,k}, W_{j,c_i}\right) \le d, i = 1..4 \wedge \\ \left(T_{i,k}^x \notin \left[W_j^{x,L}, W_j^{x,U}\right] \wedge T_{i,k}^y \notin \left[W_j^{y,L}, W_j^{y,U}\right]\right) \end{array} \right\} \quad (5.9)$$

Όσον αφορά τις περιοχές τύπου Z_1, δηλαδή αυτές που συμβολίζονται Z_{1x} και Z_{1y} στο Σχήμα 5.5(b), η επιφάνεια $A_{i,j}$ μπορεί να υπολογιστεί χρησιμοποιώντας την Εξ.(5.4). Όταν η σχετική θέση των $T_{i,k}$ και W_j σχηματίζει μια περιοχή τύπου Z_2, η $A_{i,j}$ μπορεί να υπολογιστεί αφαιρώντας την μικρή περιοχή στην άνω δεξιά γωνία του δίσκου αβεβαιότητας (Σχήμα 5.5(c)), που δίνεται από την Εξ.(5.5), από το συνολικό δίσκο αβεβαιότητας που βρίσκεται πάνω από τη χαμηλότερη πλευρά επερώτησης (Εξ.(5.4)). Τέλος, για τα σημεία που αντιπροσωπεύουν τροχιές εντός περιοχών τύπου Z_3, όπως φαίνεται στο Σχήμα 5.5(d), η $A_{i,j}$ μπορεί να υπολογιστεί χρησιμοποιώντας την Εξ.(5.5). Συνοψίζοντας, το $T_{i,k}$ μπορεί να βρεθεί μέσα σε:

- μία από δύο ζώνες Z_{1x} (πάνω και κάτω στο Σχήμα 5.5(a)), και δύο ζώνες Z_{1y} (αριστερά και δεξιά στο Σχήμα 5.5(a))· σε αυτές τις περιπτώσεις το $A_{i,j}$ υπολογίζεται από τις A_{1x} και A_{1y}, αντιστοίχως,
- μία από τέσσερις ζώνες Z_3, μία για κάθε γωνία του παραθύρου της ερώτησης· σε αυτές τις περιπτώσεις, $A_{i,j}=A_2$,
- μία από τέσσερις ζώνες Z_2, για κάθε γωνία του παραθύρου της ερώτησης κατά μήκος του x- και άλλες τέσσερις κατά μήκος του y- άξονα· σε αυτές τις περιπτώσεις, $A_{i,j} = \left(A_{1x} - A_2\right)$ και $A_{i,j} = \left(A_{1y} - A_2\right)$, αντίστοιχα,

- σε οποιαδήποτε άλλη θέση· σε αυτή τη περίπτωση το $A_{i,j}$ μηδενίζεται.

Φέρνοντας στο μυαλό ότι (α) η Εξ.(5.6) ολοκληρώνει την τιμή $P\left(T_{i,k} \notin W_j \mid T_{i,k}^{\dagger} \in W_j\right) = A_{i,j} / \pi d^2$ σε όλο το χώρο S_k, και (β) η τιμή του $A_{i,j}$ είναι μηδέν σε οποιοδήποτε άλλο σημείο εκτός των ζωνών Z_1, Z_2, Z_3 όπου το $A_{i,j}$ παρέχεται σε όρους της σχετικής θέσης μεταξύ του $T_{i,k}$ και του W_j, δηλαδή των r_x και r_y, η Εξ.(5.6) μπορεί να ξαναγραφτεί όπως ακολουθεί:

$$AvgP_{i,N}\left(R_{k,a\times b}\right) = \frac{1}{\pi d^2}\left(\begin{array}{l} 2\iint_{Z_{1x}} A_{1x}(r_x,r_y)dr_y dr_x + 2\iint_{Z_{1y}} A_{1y}(r_x,r_y)dr_y dr_x + 4\iint_{Z_3} A_2(r_x,r_y)dr_y dr_x \\ 4\iint_{Z_2}\left(A_{1x}(r_x,r_y) - A_2(r_x,r_y)\right)dr_y dr_x + 4\iint_{Z_2}\left(A_{1y}(r_x,r_y) - A_2(r_x,r_y)\right)dr_y dr_x \end{array}\right) \Rightarrow$$

$$AvgP_{i,N}\left(R_{k,a\times b}\right) =$$

$$\frac{1}{\pi d^2}\left(2\iint_{Z_{1x}+2Z_2} A_{1x}(r_x,r_y)dr_y dr_x + 2\iint_{Z_{1y}+2Z_2} A_{1y}(r_x,r_y)dr_y dr_x - 4\iint_{Z_3} A_2(r_x,r_y)dr_y dr_x\right) \quad (5.10)$$

Οι δύο $Z_{1x} + 2Z_2$ περιοχές που εμπλέκονται στα παραπάνω ολοκληρώματα μπορούν να ειδωθούν ως το επάνω και κάτω παραλληλόγραμμο στο Σχήμα 5.5(a) που σχηματίζονται από το Z_{1x} και τις δύο Z_2 περιοχές που το περικλείουν, και το μέγεθός τους κατά μήκος των *x*- και *y*- αξόνων είναι 2*a* και *d*, αντίστοιχα. Το ίδιο επίσης ισχύει σε σχέση με τις δύο $Z_{1y} + 2Z_2$ περιοχές, με έκταση *d* και 2*b* κατά μήκος του *x*- και *y*- άξονα, αντίστοιχα. Σύμφωνα με αυτή τη θεώρηση, η παραπάνω σχέση μπορεί να ξαναγραφτεί ως ακολούθως:

$$AvgP_{i,N}\left(R_{k,a\times b}\right) =$$

$$\frac{1}{\pi d^2}\cdot\left(2\int_0^d\int_0^{2a} A_{1x}\left(r_x,r_y\right)dr_x dr_y + 2\int_0^{2b}\int_0^d A_{1y}\left(r_x,r_y\right)dr_x dr_y - 4\int_0^d\int_0^{\sqrt{d^2-x^2}} A_2\left(r_x,r_y\right)dr_x dr_y\right) \quad (5.11)$$

Αντικαθιστώντας $\int_0^d A_{1y}\left(r_x,r_y\right)dr_y = \int_0^d A_{1x}\left(r_x,r_y\right)dr_x$ με $\frac{2}{3}d^3$, και $\int_0^d\int_0^{\sqrt{d^2-x^2}} A_2\left(r_x,r_y\right)dr_y dr$ με $\frac{1}{8}d^4$ στην παραπάνω εξίσωση, παίρνουμε τον απλό τύπο:

$$AvgP_{i,N}\left(R_{k,a\times b}\right) = \frac{8d}{3\pi}(a+b) - \frac{d^2}{2\pi} \quad (5.12)$$

Αντικαθιστώντας την Εξ.(5.12) στην Εξ.(5.2) αποδεικνύεται το Λήμμα 5.1. ■

5.3.2. Εκτίμηση του Αριθμού των Λανθασμένων Θετικών

Στη συνέχεια αποδεικνύουμε ένα παρόμοιο λήμμα σχετικά με το μέσο αριθμό των λανθασμένων θετικών:

Λήμμα 5.2*: Ο μέσος αριθμός $E_P(R_{k,a\times b})$ των λανθασμένων θετικών στα αποτελέσματα μιας επερώτησης χρονικής στιγμής $W_j \in R_{k,a\times b}$ με πλευρές μήκους 2a και 2b στο χρονικό αποτύπωμα t_k σε ένα σύνολο δεδομένων τροχιάς που ακολουθεί τις υποθέσεις ομοιομορφίας δεδομένων και ομοιομορφίας αβεβαιότητας δίνεται από τον τύπο:*

$$E_P\left(R_{k,a\times b}\right) = N\cdot\left(\frac{8d}{3\pi}(a+b) - \frac{d^2}{2\pi}\right) \quad (5.13)$$

όπου d είναι η ακτίνα του δίσκου αβεβαιότητας.

Απόδειξη: Ο μέσος αριθμός $E_P(R_{k,a\times b})$ των τροχιών που είναι λανθασμένες θετικές στα αποτελέσματα μιας επερώτησης χρονικής στιγμής $W_j \in R_{k,a\times b}$, δηλαδή, $T_{i,k} \in W_j \mid T_{i,k}^{\dagger} \notin W_j$, μπορεί να ληφθεί από τη

μέση πιθανότητα $AvgP_{i,P}(R_{k,a\times b})$ μια αυθαίρετη τροχιά T_i να είναι λανθασμένη θετική σε σχέση με ένα αυθαίρετο παράθυρο επερώτησης $W_j \in R_{k,a\times b}$, πολλαπλασιάζοντάς το με το συνολικό αριθμό N τροχιών στο χώρο δεδομένων:

$$E_P\left(R_{k,a\times b}\right) = N \cdot AvgP_{i,P}\left(R_{k,a\times b}\right) \tag{5.14}$$

Κατόπιν, ακολουθώντας μεθοδολογία αντίστοιχη αυτής που ακολουθήσαμε στην απόδειξη του λήμματος 5.1, αποδεικνύεται ότι η πιθανότητα $T_{i,k} \in W_j \mid T_{i,k}^{\dagger} \notin W_j$ είναι:

$$P\left(T_{i,k} \in W_j \mid T_{i,k}^{\dagger} \notin W_j\right) = \begin{cases} \dfrac{A_{i,j}}{\pi d^2}, \; if\; T_{i,k} \in W_j \; and \; Dist\left(T_{i,k}, W_j\right) \le d \\ 0, \qquad otherwise \end{cases} \tag{5.15}$$

και η μέση, σε σχέση με οποιοδήποτε παράθυρο επερώτησης $R_{k,a\times b}$, πιθανότητα μια τροχιά T_i να είναι λανθασμένη θετική:

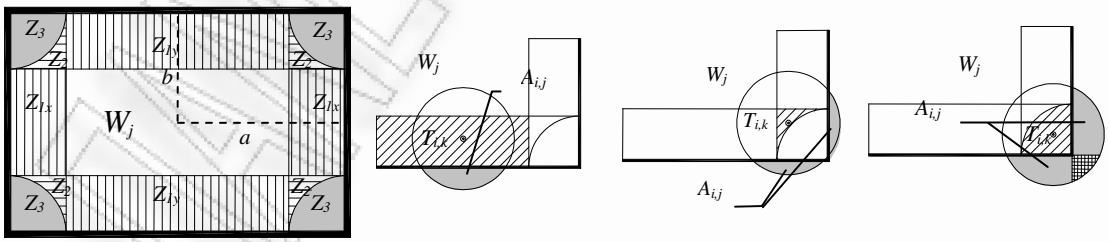
$$AvgP_{i,P}\left(R_{a\times b}\right) = \int\limits_{W_j \in R_{k,a\times b}} P\left(T_{i,k} \in W_j \mid T_{i,k}^{\dagger} \notin W_j\right) dW = \iint\limits_{S_k} P\left(T_{i,k} \in W_j \mid T_{i,k}^{\dagger} \notin W_j\right) dxdy \tag{5.16}$$

Το παραπάνω ολοκλήρωμα υπολογίζεται εκ νέου προσδιορίζοντας τις περιοχές εντός των οποίων η $A_{i,j}$ εκφράζεται ως απλή συνάρτηση. Αυτές οι ζώνες βρίσκονται εντός της περιοχής που σχηματίζεται από το αρχικό παράθυρο της ερώτησης και τη *διαφορά Minkowski* (*Minkowski difference*) του W_j με ένα δίσκο ακτίνας d [TWHC04]. Η διαφορά Minkowski, είναι συμπληρωματική του αθροίσματος Minkowski [TWHC04] που έχει εξεταστεί ενδελεχώς στον χώρο των γραφικών υπολογιστών, ενώ ο υπολογισμός της για κυρτά πολύγωνα είναι μία ξεκάθαρη διαδικασία που απαιτεί γραμμικό χρόνο εκτέλεσης [TWHC04]. Το Σχήμα 5.6(a) παρουσιάζει τα τρία σύνολα αυτών των ζωνών, πιο συγκεκριμένα των Z_1, Z_2 και Z_3, που μπορεί να οριστούν με τρόπο ανάλογο αυτού που χρησιμοποιήθηκε για τον υπολογισμό των λανθασμένων αρνητικών. Επισήμως:

$$Z_{1,j} = \left\{ \begin{array}{c} T_i \in P : T_{i,k} \in W_j \wedge Dist\left(T_{i,k}, W_j\right) \le d \wedge \\ \left(T_{i,k}^{x} \in \left[W_j^{x,L} + d, W_j^{x,U} - d\right] \vee T_{i,k}^{y} \in \left[W_j^{y,L} + d, W_j^{y,U} - d\right]\right) \end{array} \right\} \tag{5.17}$$

$$Z_{2,j} = \left\{ T_i \in P : T_{i,k} \in W_j \wedge T_{i,k} \notin Z_{1,j} \wedge Dist\left(T_{i,k}, W_j\right) \le d \wedge Dist\left(T_{i,k}, W_{j,c_i}\right) \ge d, i = 1..4 \right\} \tag{5.18}$$

$$Z_{3,j} = \left\{ T_i \in P : T_{i,k} \in W_j \wedge Dist\left(T_{i,k}, W_{j,c_i}\right) \le d, i = 1..4 \right\} \tag{5.19}$$



Σχήμα 5.6: Οι περιοχές όπου η επιφάνεια $A_{i,j}$ συνεισφέρει σε λανθασμένες θετικές απαντήσεις εκφράζεται ως απλή συνάρτηση

Όσον αφορά τις περιοχές Z_{1x} και Z_{1y}, η επιφάνεια $A_{i,j}$ υπολογίζεται χρησιμοποιώντας την Εξ.(5.4) (Σχήμα 5.6(b)). Όταν ασχολούμαστε με την περιοχή Z_2, η $A_{i,j}$ καθορίζεται προσθέτοντας την Εξ.(5.4) κατά μήκος των αξόνων x και y (Σχήμα 5.6(c)). Τέλος, τα σημεία που αντιπροσωπεύουν τις τροχιές

εντός του Z_3 υπολογίζονται επίσης προσθέτοντας την Εξ.(5.4) κατά μήκος των αξόνων *x* και *y* και αφαιρώντας την μικρή περιοχή στην κάτω αριστερά γωνία του δίσκου αβεβαιότητας (Σχήμα 5.6(d)), που δίνεται από την Εξ.(5.5). Συνοψίζοντας, το $T_{i,j}$ μπορεί να βρεθεί μέσα:

- σε μία από δύο ζώνες Z_{1x} (πάνω και κάτω στο Σχήμα 5.6(a))· σε αυτές τις περιπτώσεις, η $A_{i,j}$ υπολογίζεται από την A_{1x},
- μία από δύο ζώνες Z_{1y} (αριστερά και δεξιά στο Σχήμα 5.6(a))· σε αυτές τις περιπτώσεις, η $A_{i,j}$ υπολογίζεται από την A_{1y},
- μία από τέσσερις ζώνες Z_3, μία για κάθε γωνία του παραθύρου της ερώτησης· σε αυτές τις περιπτώσεις, $A_{i,j}=A_{1x}+A_{1y}$,
- μία από τέσσερις ζώνες Z_2, μία για κάθε γωνία του παραθύρου της ερώτησης; σε αυτές τις περιπτώσεις , $A_{i,j}=A_{1x}+A_{1y}-A_2$,

και η Εξ.(5.16) αναδιατυπώνεται ως εξής:

$$AvgP_{i,P}\left(R_{k,a\times b}\right)=$$
$$\frac{1}{\pi d^2}\cdot\begin{pmatrix}2\iint_{Z_{1y}}A_{1y}(r_x,r_y)dr_ydr_x+2\iint_{Z_{1x}}A_{1x}(r_x,r_y)dr_ydr_x+4\iint_{Z_2}\left(A_{1x}\left(r_x,r_y\right)+A_{1y}\left(r_x,r_y\right)\right)dr_ydr_x\\ 4\iint_{Z_3}\left(A_{1x}\left(r_x,r_y\right)+A_{1y}\left(r_x,r_y\right)-A_2\left(r_x,r_y\right)\right)dr_ydr_x\end{pmatrix} \quad (5.20)$$

που, μετά τους απαραίτητους υπολογισμούς, μας δίνει:

$$AvgP_{i,P}\left(R_{k,a\times b}\right)=\frac{8d}{3\pi}(a+b)-\frac{d^2}{2\pi} \quad (5.21)$$

Αντικαθιστώντας την Εξ.(5.21) στην Εξ.(5.14) έχουμε αποδείξει το Λήμμα 5.2.■

5.3.3. Συζήτηση

Συνοψίζοντας, το αναλυτικό μοντέλο για την πρόβλεψη των λανθασμένων θετικών και λανθασμένων αρνητικών όταν εκτελείται μία επερώτηση χρονικής στιγμής σε ομοιόμορφα κατανεμημένα δεδομένα τροχιών, αποτελείται από το Λήμμα 5.1 και το Λήμμα 5.2 που αποδείχθηκαν στις προηγούμενες ενότητες. Προκύπτει ότι ο μέσος αριθμός λανθασμένων αρνητικών και λανθασμένων θετικών μιας αυθαίρετης επερώτησης χρονικής στιγμής στο χρονικό αποτύπωμα t_k με γνωστό μέγεθος 2*a* και 2*b* κατά μήκος των αξόνων *x*- και *y*- αντίστοιχα, είναι μία συνάρτηση των *a*, *b*, της ακτίνα αβεβαιότητας *d* και του πλήθους *N* του συνόλου δεδομένων. Ένα αποτέλεσμα που προκύπτει ως επακόλουθο είναι ότι, θεωρητικά, ο μέσος αριθμός των λανθασμένων αρνητικών είναι ίσος με το μέσο αριθμό των λανθασμένων θετικών:

$$E_N\left(R_{k,a\times b}\right)=E_P\left(R_{k,a\times b}\right) \quad (5.22)$$

Ενώ εκ πρώτης όψεως ένα τέτοιο αποτέλεσμα ακούγεται περίεργο, αποδεικνύεται λογικό όταν λάβουμε υπ' όψιν ότι, αφ' ενός, ο αριθμός των τροχιών που συνεισφέρουν στον αριθμό των λανθασμένων αρνητικών, που απεικονίζεται ως η σκιασμένη περιοχή στο Σχήμα 5.5(a), είναι μεγαλύτερος από τον αντίστοιχο των λανθασμένων θετικών (Σχήμα 5.6(a)) και, αφ' ετέρου, η επιφάνεια $A_{i,j}$ του δίσκου αβεβαιότητας κάθε τροχιάς που συνεισφέρει στον αριθμό των λανθασμένων αρνητικών (Σχήμα 5.5(d)) είναι μικρότερη από την αντίστοιχη για τις λανθασμένες θετικές (Σχήμα 5.6(d)). Ο αναλυτικός μας υπολογισμός των of $E_N(R_{a\times b})$ και $E_P(R_{a\times b})$ αποδεικνύει ότι αυτοί οι δύο συμπληρωματικοί όροι τελικά καταλήγουν σε δύο ίσες τιμές για τον αριθμό των λανθασμένων θετικών και αρνητικών, καταλήγοντας στην Εξ.(5.22).

Επιπλέον, προκύπτει από τις Εξ.(5.1) και Εξ.(5.13) ότι ο μέσος αριθμός των λανθασμένων αρνητικών και λανθασμένων θετικών μιας επερώτησης χρονικής στιγμής εξαρτάται από την περίμετρο της επερώτησης ($a+b$) και όχι από την επιφάνειας επερώτησης ($a \cdot b$). Μία τελευταία παρατήρηση είναι ότι όταν το μοντέλο μας χρησιμοποιείται για να καθορίσει τη μέγιστη επιτρεπόμενη (αν)ακρίβεια των δεδομένων που θα τροφοδοτήσουν μία MOD, οι Εξ.(5.1) και Εξ.(5.13) μπορούν να λυθούν για την τιμή της ακτίνας αβεβαιότητας d, αν δοθούν οι τιμές της απαιτούμενης ακρίβειας σε όρους λανθασμένων αποτελεσμάτων και τυπικής έκτασης της επερώτησης.

Διαισθητικά, τα δύο τμήματα του πολλαπλασιαστή του N στην Εξ.(5.1) και την Εξ.(5.13), δηλαδή το $\frac{8d}{3\pi}(a+b)$ και το $\frac{d^2}{2\pi}$, αντιπροσωπεύουν τη συνεισφορά στο συνολικό αριθμό λανθασμένων αποτελεσμάτων, του μήκους της περιμέτρου επερώτησης και των τεσσάρων γωνιών του παραθύρου επερώτησης, αντίστοιχα. Η λεπτομέρεια αυτή θα αποβεί ιδιαίτερα χρήσιμη στην επόμενη ενότητα όταν θα χαλαρώσουμε την υπόθεση της ομοιομορφίας δεδομένων με τη βοήθεια των ιστογραμμάτων.

Τέλος, θα πρέπει εκ νέου να επισημάνουμε, ότι οι παραπάνω τύποι, καθώς και η πλειονότητα αυτών που παρουσιάζονται από τούδε και στο εξής, μπορεί να εφαρμοσθεί απευθείας στα απλά χωρικά δεδομένα χωρίς την ανάγκη οιασδήποτε τροποποίησης · αυτό λόγω του γεγονότος ότι μία επερώτηση χρονικής στιγμής σε ένα σύνολο τροχιών μπορεί να θεωρηθεί ως επερώτηση παραθύρου εύρους σε μία στιγμιαία αποτύπωση των τροχιών στο χρονικό αποτύπωμα που προσδιορίζεται από την επερώτηση χρονικής στιγμής. Έτσι, ο μέσος αριθμός $E_N(R_{a\times b})$ και $E_P(R_{a\times b})$ των λανθασμένων αρνητικών και λανθασμένων θετικών στα αποτελέσματα των επερωτήσεων παραθύρου εύρους με μήκος πλευρών $2a$ και $2b$ σε απλά χωρικά δεδομένα, με βάση τις τέσσερις υποθέσεις που αναφέρθηκαν στην αρχή της ενότητας είναι:

$$E_N\left(R_{a\times b}\right) = E_P\left(R_{a\times b}\right) = N \cdot \left(\frac{8d}{3\pi}(a+b) - \frac{d^2}{2\pi}\right) \tag{5.23}$$

Η ίδια αιτιολόγηση (και το ίδιο αποτέλεσμα) ισχύει και για όλους τους τύπους που αναπτύσσονται στις επόμενες ενότητες, όπου και χαλαρώνουμε τις υποθέσεις ομοιομορφίας. Θα αποδειχθεί, περαιτέρω, κατά την πειραματική μελέτη ότι το αναπτυχθέν μοντέλο έχει πολλές εφαρμογές σε εμπορικά SDBMS.

5.4. Χαλάρωση των Υποθέσεων Ομοιομορφίας

Στην ενότητα αυτή αποδεσμευόμαστε από τις τρεις υποθέσεις A_I – A_{III}, που κάναμε κατά την διατύπωση του προβλήματος στην Ενότητα 5.3. Αυτό θα γίνει σταδιακά με αύξουσα σειρά. Πρώτον θα δείξουμε πως υποστηρίζουμε μη ομοιόμορφες κατανομές αβεβαιότητας του πραγματικού κόσμου χαλαρώνοντας συνεπώς την A_I (Ενότητα 5.4.1), κατόπιν χρησιμοποιούμε χωροχρονικά ιστογράμματα για να χαλαρώσουμε την A_{II} (Ενότητα 5.4.2) και τέλος, δείχνουμε πως τα ιστογράμματα μπορούν να επαυξηθούν για να χαλαρώσουμε την A_{III} (Ενότητα 5.4.3).

5.4.1. Χαλάρωση της Υπόθεσης Ομοιμορφίας Αβεβαιότητας Θέσης

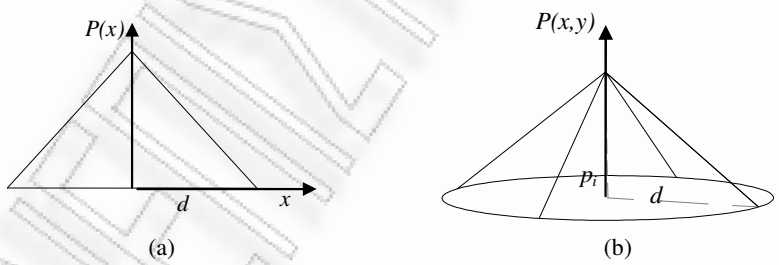
Η ανάλυση που έγινε στην Ενότητα 5.3 βασίστηκε στην υπόθεση ομοιόμορφης αβεβαιότητας θέσης, που σημαίνει ότι η ακριβής θέση κάθε σημείου τροχιάς σε ένα δεδομένο χρονικό αποτύπωμα είναι ομοιόμορφα κατανεμημένη εντός του δίσκου αβεβαιότητας με το σημείο που αντιπροσωπεύει την

τροχιά στο κέντρο και γνωστή ακτίνα. Μολαταύτα, σε αυτή την ενότητα επεκτείνουμε το προτεινόμενο μοντέλο προς μη ομοιόμορφες κατανομές της αβεβαιότητας θέσης. Η λογική που διέπει αυτή την επέκταση είναι ότι αν το ακριβές σημείο $T_{i,k}^{\dagger}$ βρίσκεται εντός μίας κυκλικής γειτονιάς του $T_{i,k}$, είναι πολύ πιο πιθανό η πιθανότητα μιας θέσης να είναι η ακριβής θέση του $T_{i,k}^{\dagger}$ να μειώνεται όσο αυξάνεται η απόστασή του από το $T_{i,k}$. Πολλώ δε μάλλον, που είναι γνωστό ότι το σφάλμα που σχετίζεται με θέσεις εντοπισμένες μέσω GPS κατανέμεται κανονικά, δηλαδή ακολουθώντας τη *διμεταβλητή κανονική κατανομή* με ασυσχέτιστες μεταβλητές *x* και *y*, που είναι η επέκταση της κανονικής κατανομής στο 2D χώρο [Lei95]· δεδομένου δε ότι η χρήση του GPS επιτρέπει υψηλούς ρυθμούς δειγματοληψίας το συνολικό σφάλμα σε αυτές τις περιπτώσεις κυριαρχείται από το σφάλμα που εισάγεται από τη συσκευή εντοπισμού. Όντως λοιπόν, το επιχείρημα ότι η αβεβαιότητα σε πραγματικά χωροχρονικά (και σταθερά χωρικά) δεδομένα τείνει να είναι ομοιόμορφα κατανεμημένη είναι εύλογο[CC07], [NRB03], [PTJ05].

Βάσει της προηγούμενης συζήτησης, ο στόχος της ενότητας αυτής είναι να χαλαρώσουμε την υπόθεση ομοιομορφίας στην αβεβαιότητα θέσης των κινούμενων αντικειμένων και να κάνουμε το προτεινόμενο μοντέλο να υποστηρίζει την διμεταβλητή κανονική κατανομή. Η αντίστοιχη συνάρτηση πυκνότητας πιθανότητας (probability density function - *pdf*), όταν οι μεταβλητές *x*, *y* δεν είναι συσχετισμένες, δίνεται από τον ακόλουθο τύπο [PTJ05]:

$$P_{BN}(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5.24}$$

όπου σ^2 είναι η διασπορά κατά μήκος των αξόνων *x*- και *y*- · άρα σ είναι η αντίστοιχη τυπική απόκλιση. Ωστόσο, ο υπολογισμός των αντίστοιχων τύπων όπως γίνεται στην Ενότητα 5.3 είναι μια επώδυνη διαδικασία που προϋποθέτει την ολοκλήρωση αρκετών εκθετικών συναρτήσεων.



Σχήμα 5.7: Κατανομή διαφοράς ομοιομορφίας σε (a) 1D και (b) 2D χώρο

Από την άλλη πλευρά, η συνάρτηση πυκνότητας της διμεταβλητής κανονικής κατανομής προσεγγίζεται αποτελεσματικά από τη *δυσδιάστατη κατανομή διαφοράς ομοιομορφίας (two-dimensional uniform difference distribution* - 2*d-UDD*) που είναι η επέκταση της *κατανομής διαφοράς ομοιομορφίας (uniform difference distribution)* στο δυσδιάστατο χώρο, ήτοι, η κατανομή της διαφοράς μεταξύ δύο ομοιόμορφα κατανεμημένων μεταβλητών στην [0, *d*]. Η *pdf* της 2d-UDD είναι:

$$P_{2d-UDD}(x,y) = \frac{3}{\pi d^2} \cdot \begin{cases} 1 - \frac{1}{d}\sqrt{x^2+y^2} & if \quad \sqrt{x^2+y^2} \le d \\ 0 & otherwise \end{cases} \tag{5.25}$$

που είναι η επέκταση στο δυσδιάστατο χώρο της κατανομής διαφοράς ομοιομορφίας με την εξής *pdf* :

$$P_{UDD}(x) = \frac{1}{d} \cdot \begin{cases} 1 - \frac{|x|}{d} & if \quad |x| \le d \\ 0 & otherwise \end{cases} \tag{5.26}$$

Και οι δύο κατανομές απεικονίζονται στο Σχήμα 5.7 που παρουσιάζει το περίγραμμα των *pdf*s τους· στην πράξη, η $P_{2d\text{-}UDD}$ σχηματίζει μία κωνική επιφάνεια με ακτίνα βάσης d και μοναδιαίο όγκο όπως φαίνεται στο Σχήμα 5.7(b).

Για την αναδιατύπωση του προτεινόμενου μοντέλου, η υπόθεση ομοιομορφίας αβεβαιότητας θέσης πρέπει να αντικατασταθεί από την ακόλουθη *υπόθεση διαφοράς ομοιμορφίας αβεβαιότητας θέσης* (*uncertainty uniformity difference assumption)*: η πραγματική θέση $T_{i,k}$ κάθε τροχιάς T_i στο χρονικό αποτύπωμα t_k δίνεται από την κατανομή $P_{2d\text{-}UDD}$ που περιγράφηκε παραπάνω. Βάσει αυτής της υπόθεσης, διατυπώνεται το ακόλουθο λήμμα:

Λήμμα 5.3*: Οι μέσοι αριθμοί $E_N(R_{k,a\times b})$ και $E_P(R_{k,a\times b})$ των λανθασμένων αρνητικών και λανθασμένων θετικών, αντίστοιχα, στα αποτελέσματα μίας επερώτησης χρονικής στιγμής $W_j \in R_{k,a\times b}$ με πλευρές μήκους 2a και 2b στο χρονικό αποτύπωμα t_k σε ένα σύνολο δεδομένων τροχιάς που ακολουθεί τις υποθέσεις ομοιομορφίας δεδομένων και διαφοράς ομοιομορφίας αβεβαιότητας θέσης δίνονται από τον τύπο:*

$$E_N\left(R_{k,a\times b}\right) = E_P\left(R_{k,a\times b}\right) = N \cdot \left(\frac{2d}{\pi}(a+b) - \frac{3d^2}{10\pi}\right) \tag{5.27}$$

όπου d είναι η ακτίνα του δίσκου αβεβαιότητας.

Απόδειξη: Τα $E_N(R_{k,a\times b})$ και $E_P(R_{k,a\times b})$ λαμβάνονται από τις μέσες πιθανότητες $AvgP_{i,N}(R_{k,a\times b})$ και $AvgP_{i,P}(R_{k,a\times b})$, αντίστοιχα, πολλαπλασιασμένες επί το συνολικό αριθμό N των αντικειμένων στο χώρο δεδομένων. Η πιθανότητα μία τροχιά T_i να είναι λανθασμένη αρνητική ή λανθασμένη θετική, σε σχέση με ένα παράθυρο επερώτησης W_j, στο χρονικό αποτύπωμα t_k είναι:

$$P\left(T_{i,k} \notin W_j \mid T_{i,k}^{\dagger} \in W_j\right) = \begin{cases} V_{i,j}, & if\ T_{i,k} \notin W_j\ and\ Dist\left(T_{i,k}, W_j\right) \le d \\ 0, & otherwise \end{cases} \tag{5.28}$$

αντίστοιχα

$$P\left(T_{i,k} \in W_j \mid T_{i,k}^{\dagger} \notin W_j\right) = \begin{cases} V_{i,j}, & if\ T_{i,k} \in W_j\ and\ Dist\left(T_{i,k}, W_j\right) \le d \\ 0, & otherwise \end{cases} \tag{5.29}$$

όπου $V_{i,j}$ είναι ο όγκος της *2d-UDD pdf* $P_{2d\text{-}UDD}$, που βρίσκεται πλήρως εντός ή εκτός του W_j, αντίστοιχα.

Ο όγκος $V_{i,j}$, της $P_{2d\text{-}UDD}$ που βρίσκεται εντός (εκτός, αντίστοιχα) του παραθύρου επερώτησης προσδιορίζεται ακολουθώντας την ίδια μεθοδολογία όπως στην απόδειξη του Λήμματος 5.1 (Λήμμα 5.2, αντίστοιχα), λαμβάνοντας επιπλέον υπ' όψιν την υπόθεση αβεβαιότητας μεγέθους. Πιο συγκεκριμένα, αν λάβουμε υπ' όψιν ότι τα (b) – (d) στο Σχήμα 5.4 παρουσιάζουν και την προβολή της $P_{2d\text{-}UDD}$ στο επίπεδο *x*-*y*, μπορούμε να τα χρησιμοποιήσουμε στη συζήτησή μας: στις πρώτες δύο περιπτώσεις (Σχήμα 5.4(b) και 4(c)) όπου η απόσταση μεταξύ του $T_{i,k}$ και καθεμιάς από τις τέσσερις γωνίες του W_j είναι μεγαλύτερη της d, το $V_{i,j}$ ισούται με το $V_{1x}(r_x,r_y)$ (ή $V_{1y}(r_x,r_y)$) που είναι το τμήμα της $P_{2d\text{-}UDD}$ που βρίσκεται πάνω (ή δεξιά, αντίστοιχα) του κάθετου επιπέδου που περνά από τη c_1c_2. Στην τρίτη περίπτωση, όπου η απόσταση μεταξύ της $T_{i,k}$ και μιας από τις τέσσερις γωνίες του W_j είναι μικρότερη του d (Σχήμα 5.4(d)), η $V_{i,j}$ ισούται με $V_2(r_x,r_y)$, που είναι το τμήμα της $P_{2d\text{-}UDD}$ που

βρίσκεται δεξιά του κάθετου επιπέδου που περνά από τη c_1c_2 και πάνω από αυτό που περνά από τη c_2c_3.

Η μέση, σε σχέση με οποιοδήποτε παράθυρο επερώτησης στο $R_{k,axb}$, πιθανότητα μια τροχιά T_i να είναι λανθασμένη αρνητική (λανθασμένη θετική, αντίστοιχα) στο χρονικό αποτύπωμα t_k υπολογίζεται ολοκληρώνοντας την Εξ.(5.28) (Εξ.(5.29), αντίστοιχα) σε όλες τις θέσεις επερώτησης όπως και στην Εξ.(5.6) (Εξ.(5.16), αντίστοιχα). Το αντίστοιχο ολοκλήρωμα υπολογίζεται με τον ίδιο τρόπο που ακολουθήσαμε και για την απόδειξη του Λήμματος 5.1 (Λήμμα 5.2, αντίστοιχα) αντικαθιστώντας τις τιμές των $A_{1x}(r_x,r_y)$, $A_{1y}(r_x,r_y)$ και $A_2(r_x,r_y)$ με $V_{1x}(r_x,r_y)$, $V_{1y}(r_x,r_y)$ και $V_2(r_x,r_y)$ στην Εξ.(5.11) (Εξ.(5.20), αντίστοιχα). Έτσι, αντικαθιστώντας $\int_0^d V_{1y}\left(r_x,r_y\right)dr_y = \int_0^d V_{1x}\left(r_x,r_y\right)dr_x = \frac{d}{2\pi}$, και $\int_0^d \int_0^{\sqrt{d^2-x^2}} V_2\left(r_x,r_y\right)dr_y dr_x = \frac{3d^2}{40\pi}$ στους αντίστοιχους τύπους και εκτελώντας τους απαραίτητους υπολογισμούς έχουμε:

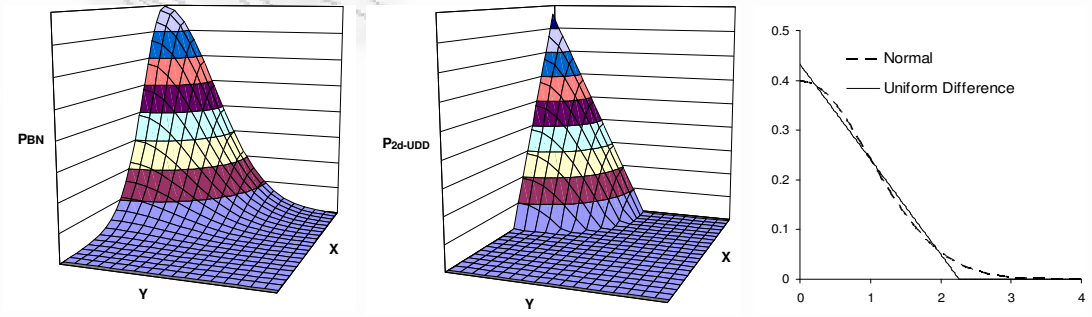
$$AvgP_{i,N}\left(R_{k,a\times b}\right) = \frac{2d}{\pi}\left(a+b\right) - \frac{3d^2}{10\pi} \tag{5.30}$$

και

$$AvgP_{i,N}\left(R_{k,a\times b}\right) = \frac{2d}{\pi}\left(a+b\right) - \frac{3d^2}{10\pi} \tag{5.31}$$

Πολλαπλασιάζοντας τους παραπάνω τύπους επί N αποδεικνύεται το Λήμμα 5.3.■

Ως εδώ, δεδομένου ότι η θέση του πραγματικού σημείου ακολουθεί την υπόθεση διαφοράς ομοιομορφίας αβεβαιότητας, το μοντέλο μας αποτελείται από την Εξ.(5.27), που μοιάζει πολύ με αυτές της Ενότητας 5.3 υπό την υπόθεση ομοιομορφίας αβεβαιότητας. Πιο συγκεκριμένα, αν η Εξ.(5.27) συγκριθεί με τις Εξ.(5.1) και Εξ.(5.13), οι τύποι διαφέρουν μόνο στα βάρη των μεταβλητών της συνάρτησης $d(a+b)$ και d^2. Παρόλο που το μοντέλο που περιγράφηκε παραπάνω δεν λαμβάνει άμεσα υπ' όψιν την διμεταβλητή κανονική κατανομή, μπορεί να χρησιμοποιηθεί για την αποτελεσματική προσέγγισή της. Έστω, για παράδειγμα, το Σχήμα 5.8 που παρουσιάζει την συνάρτηση πιθανότητας της διμεταβλητής κανονικής κατανομής με μη συσχετισμένες μεταβλητές (Σχήμα 5.8(a)), τη συνάρτηση πιθανότητας της 2d-UDD (Σχήμα 5.8(b)), και το περίγραμμα των δύο κατανομών στο 1D χώρο (Σχήμα 5.8(c))· οι δύο συναρτήσεις πιθανότητας αποδεικνύονται πολύ κοντά η μία με την άλλη. Άρα, μπορούμε να χρησιμοποιήσουμε ελάχιστα τετράγωνα και να εκτιμήσουμε την ακτίνα του κώνου που ταιριάζει καλύτερα στην «καμπάνα» του Gauss της διμεταβλητής κανονικής κατανομής.



(a) (b) (c)

Σχήμα 5.8: (a) Διμεταβλητή κανονική κατανομή (b) Δυσδιάστατη UDD, και, (c) καλύτερο ταίριασμα μεταξύ των δύο κατανομών σε μία διάσταση (c)

Επισήμως, διατυπώνεται το ακόλουθο λήμμα:

Λήμμα 5.4: *Η δυσδιάστατη κατανομή διαφοράς ομοιομορφίας που προσεγγίζει με το βέλτιστο τρόπο τη διμεταβλήτη κανονική κατανομή με μη συσχετισμένες μεταβλητές, λαμβάνεται όταν η ακτίνα d του δίσκου αβεβαιότητας είναι:*

$$d \approx 2.36533 \times \sigma \tag{5.32}$$

όπου σ είναι η τυπική απόκλιση της διμεταβλητής κανονικής κατανομής κατά μήκος των αξόνων x- και y.

Απόδειξη: Βάσει της Θεωρίας των Ελαχίστων Τετραγώνων, η βέλτιστη προσέγγιση μιας συνάρτησης f από μία άλλη συνάρτηση g στο ίδιο πεδίο D δίνεται ελαχιστοποιώντας το ολοκλήρωμα $\iint_D \left(f(x)-g(x)\right)^2 dx$ του τετραγώνου της διαφοράς τους κατά μήκος του D. Κατά συνέπεια, για να αποδείξουμε το λήμμα πρέπει να προσδιορίσουμε την τιμή του d που ελαχιστοποιεί το $\iint_{\mathbb{R}^2} \left(P_{2d-UDD}(x,y)-P_{BN}(x,y)\right)^2 dxdy$. Προς τούτο, ισχύει ότι:

$$\begin{gathered}\iint_{\mathbb{R}^2} \left(P_{2d-UDD}(x,y)-P_{BN}(x,y)\right)^2 dxdy = \\ \iint_{C(0,d)} \left(P_{2d-UDD}(x,y)-P_{BN}(x,y)\right)^2 dxdy + \iint_{\mathbb{R}^2-C(0,d)} \left(P_{2d-UDD}(x,y)-P_{BN}(x,y)\right)^2 dxdy\end{gathered} \tag{5.33}$$

όπου $C(0,d)$ είναι ο δίσκος με κέντρο (0,0) και ακτίνα d. Εφαρμόζοντας τις Εξ.(5.25) και Εξ.(5.24) στην Εξ.(5.33), έχουμε:

$$\begin{gathered}\iint_{\mathbb{R}^2} (P_{2d-UDD}(x,y)-P_{BN}(x,y))^2 dxdy = \\ \iint_{C(0,d)} \left(\frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} - \frac{3}{\pi d^2}\left(1-\frac{\sqrt{x^2+y^2}}{d}\right)\right)^2 dxdy + \iint_{\mathbb{R}^2-C(0,d)} \left(\frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} - 0\right)^2 dxdy\end{gathered} \tag{5.34}$$

Σ' αυτό το σημείο χρησιμοποιούμε το μετασχηματισμό από καρτεσιανές σε πολικές συντεταγμένες, που μετατρέπει τα (x, y) σε (ρ, θ) σύμφωνα με τον ακόλουθο τύπο:

$$\iint f(x,y)dxdy = \iint f(\rho\cos\theta, \rho\sin\theta)\rho d\rho d\theta \tag{5.35}$$

Αν εφαρμόσουμε τον παραπάνω μετασχηματισμό στην Εξ.(5.34), έχουμε:

$$\begin{gathered}\iint_{\mathbb{R}^2} \left(P_{2d-UDD}(x,y)-P_{BN}(x,y)\right)^2 dxdy = \\ \int_0^{2\pi}\int_0^{d} \left(\frac{1}{2\pi\sigma^2} e^{-\frac{\rho^2}{2\sigma^2}} - \frac{3}{\pi d^2}\left(1-\frac{\rho}{d}\right)\right)^2 \rho d\rho d\theta + \int_0^{2\pi}\int_d^{\infty} \left(\frac{\rho}{2\pi\sigma^2} e^{-\frac{\rho^2}{2\sigma^2}}\right)^2 \rho d\rho d\theta\end{gathered}$$

Άρα

$$\iint_{\mathbb{R}^2} \left(P_{2d-UDD}(x,y)-P_{BN}(x,y)\right)^2 dxdy = \frac{d^3 - 18d\sigma^2 + 12\sqrt{2\pi}\sigma^3 \operatorname{Erf}\left[\frac{d}{\sqrt{2}\sigma}\right]}{4d^3\pi\sigma^2} \tag{5.36}$$

όπου Erf[x] είναι η συνάρτηση του σφάλματος που εντοπίζεται όταν ολοκληρώνουμε την κανονική κατανομή. Κατόπιν, υπολογίζουμε την πρώτη παράγωγο της Εξ.(5.36) σε σχέση με τη d:

$$\frac{\partial \iint_{\mathbb{R}^2} \left(P_{2d-UDD}(x,y)-P_{BN}(x,y)\right)^2 dxdy}{\partial d} = -\frac{9d + 6de^{-\frac{d^2}{2\sigma^2}} - 9\sqrt{2\pi}\operatorname{Erf}\left[\frac{d}{\sqrt{2}\sigma}\right]}{d^4\pi} \tag{5.37}$$

και αντικαθιστώντας το d/σ με μία μεταβλητή a ($a \neq 0$), έχουμε ότι:

$$\frac{\partial \iint_{\mathbb{R}^2} \left(P_{2d-UDD}(x,y)-P_{BN}(x,y)\right)^2 dxdy}{\partial d} = \frac{9a + 6ae^{-\frac{a^2}{2}} - 9\sqrt{2\pi}\operatorname{Erf}\left[\frac{a}{\sqrt{2}}\right]}{ad^3\pi} \tag{5.38}$$

που μηδενίζεται όταν ο αριθμητής γίνει μηδέν. Άρα, η πρώτη παράγωγος της Εξ.(5.36) μηδενίζεται όταν

$$9a + 6ae^{-\frac{a^2}{2}} - 9\sqrt{2\pi}\,\mathrm{Erf}\left[\frac{a}{\sqrt{2}}\right] = 0 \tag{5.39}$$

Μετά από επίλυση με αριθμητικές μεθόδους στην Εξ.(5.39) προκύπτει ότι

$$a \approx 2.36533 \tag{5.40}$$

Δεδομένου ότι $a = d/\sigma$ αποδεικνύεται το Λήμμα 5.4. ■

Εν κατακλείδι, το προταθέν μοντέλο για κανονικά κατανεμημένη αβεβαιότητα αποτελείται από τις Εξ.(5.27) και Εξ.(5.32)· η τιμή της d που δίνεται από την Εξ.(5.32) μπορεί να χρησιμοποιηθεί απευθείας στην Εξ.(5.27) για να προσεγγίσουμε την κανονική κατανομή αρκετά αποτελεσματικά όπως θα αποδειχθεί αργότερα στην πειραματική μελέτη.

5.4.2. Χαλάρωση της Υπόθεσης Ομοιμορφίας Κατανομής Δεδομένων

Οι Ενότητες 5.3 και 5.4.1 υποθέτουν ότι οι τροχιές και κατά συνέπεια τα σημεία που λαμβάνονται από ένα στιγμιότυπο της βάσης δεδομένων είναι ομοιόμορφα κατανεμημένα στο χώρο. Στην ενότητα αυτή, χαλαρώνουμε την υπόθεση ομοιόμορφης κατανομής δεδομένων και εφαρμόζουμε την προτεινόμενη προσέγγιση σε αυθαίρετα κατανεμημένα δεδομένα με τη χρήση *ιστογραμμάτων* [Ioa93], [IP95]. Τα ιστογράμματα χρησιμοποιούνται ευρέως σε θέματα βελτιστοποίησης επερωτήσεων, όπως χωρική και χωροχρονική εκτίμηση επιλεκτικότητας [APR99], [TS96], [CC02], [HKT03], [TSP03], για να υπερκεράσουμε παρόμοιες υποθέσεις που γίνονται όταν εκτιμάται ο αριθμός των προσπελάσεων σελίδων δίσκου που απαιτούνται για την απάντηση μιας επερώτησης. Η υφέρπουσα ιδέα είναι ότι όταν τα δεδομένα περιλαμβάνονται σε ένα μικρό χώρο, μπορούν να θεωρηθούν ως ομοιόμορφα παρόλο που η κατανομή όλου του συνόλου δεδομένων μπορεί να διαφέρει αρκετά. Ο στόχος συνεπώς όταν χρησιμοποιούμε ιστογράμματα, είναι ο κατακερματισμός του χώρου σε μικρές περιοχές, που ονομάζονται *κάδοι (buckets),* που μπορεί να θεωρηθεί ότι περιέχουν ομοιόμορφα δεδομένα. Μεταξύ των όσων προτείνονται, υιοθετούμε την πρόταση του [APR99], διότι μπορεί να τροποποιηθεί απλά για να την εφαρμόσουμε στις απαιτήσεις μας.

Πιο συγκεκριμένα, οι Acharya et al. [APR99] παρουσιάζουν αρκετές τεχνικές διαμέρισης χώρου για την κατασκευή χωρικών ιστογραμμάτων που χρησιμοποιούνται για την εκτίμηση της επιλεκτικότητας των επερωτήσεων εύρους. Μεταξύ αυτών, έχει αποδειχθεί ότι η τεχνική *MinSkew* παρέχει τις πιο ακριβείς εκτιμήσεις επιλεκτικότητας για χωρικές επερωτήσεις. Η *MinSkew* είναι μία δυαδική τεχνική διαμέρισης χώρου (binary space partitioning -BSP) που χρησιμοποιεί τον ορισμό της χωρικής ασυμμετρίας (*spatial skew)*, που δίνεται στην [APR99]. Πιο συγκεκριμένα, *η χωρική ασυμμετρία ενός κάδου είναι η στατιστική διακύμανση των χωρικών πυκνοτήτων όλων των σημείων που ομαδοποιούνται εντός του κάδου και η χωρική ασυμμετρία όλου του συνόλου είναι το σταθμισμένο άθροισμα όλων των χωρικών ασυμμετριών όλων των κάδων*. Η προτεινόμενη τεχνική χρησιμοποιεί αρχικά ένα ομοιόμορφο πλέγμα περιοχών και τις χωρικές τους πυκνότητες ως είσοδο· άρα παράγει μία συμπαγή προσέγγιση των δεδομένων εισόδου στη θέση των αρχικών για να δημιουργήσει το ιστόγραμμα στη κύρια μνήμη. Κατόπιν, ο αλγόριθμος κατασκευής διαμερίζει κατ' επανάληψη το δεδομένο σύνολο περιοχών ώστε η χωρική ασυμμετρία να ελαχιστοποιείται σε κάθε βήμα μέχρι να μην είναι διαθέσιμοι άλλοι κάδοι για το ιστόγραμμα. Επειδή διαμερίζει μία υπάρχουσα περιοχή σε δύο, το

αποτέλεσμα είναι μία διαμέριση BSP. Ως αποτέλεσμα, το κατασκευασμένο ιστόγραμμα H είναι το σύνολο των n κάδων $H = \{B_i : \cup(B_i) = S \wedge \cap(B_i) = \varnothing\}$ και $B_i = \{[x_{i,L}, x_{i,U}], [y_{i,L}, y_{i,U}]\}$. Το κύριο πλεονέκτημα της προτεινόμενης τεχνικής είναι ότι οι αρχικές κυψέλες που ομαδοποιούνται στον ίδιο κάδο έχουν μικρή χωρική ασυμμετρία, δηλαδή διακύμανση. Αναμένεται επομένως ότι τα κελιά που περιέχονται μέσα στον ίδιο κάδο θα περιέχουν περίπου το ίδιο πλήθος δεδομένων· ως αποτέλεσμα, συνήθως γίνεται δεκτό ότι η κατανομή των δεδομένων μέσα στον κάθε κάδο B_i είναι ομοιόμορφη. Στην πραγματικότητα, αυτό η υπόθεση, όπως δείχνεται τόσο στο [APR99] όσο και στα δικά μας πειράματα, είναι μάλλον λογική, ακόμα και κάτω από την παρουσία ολικά ασύμμετρων δεδομένων, όπως το σύνολο δεδομένων *Charminar* [APR99].

Η κύρια χρήση των τοπικών ιστογραμμάτων είναι να παρέχουν εκτιμήσεις για την *τοπική πυκνότητα (local density)* του συνόλου δεδομένων, δεδομένης μιας χωρικής περιοχής. Προς τούτο, αρχικά καθορίζονται οι κάδοι που αλληλεπικαλύπτονται με τη χωρική επερώτηση και κατόπιν, υπολογίζεται η τοπική πυκνότητα μέσω του σταθμισμένης μέσης τιμής των πυκνοτήτων των κάδων που αλληλεπικαλύπτονται N_i. Αυτό γίνεται σταθμίζοντας την πυκνότητα N_i του κάθε κάδου B_i με την αντίστοιχη περιοχή A_i που καλύπτει μερικώς τη δεδομένη περιοχή, κανονικοποιημένη από τη συνολική επιφάνεια:

$$N' = \frac{1}{4ab} \sum_{i=1,..n} (N_i \cdot A_i) \tag{5.41}$$

Παρακάτω, δείχνεται πώς μπορεί να τροποποιηθεί καταλλήλως η δομή του ιστογράμματος MinSkew ώστε να υποστηρίζει χωροχρονικές επερωτήσεις χρονικής στιγμής.

5.4.2.1. Χωροχρονικά Ιστογράμματα για Επερωτήσεις Χρονικής Στιγμής

Το πρώτο βήμα για την κατασκευή ενός χωροχρονικού ιστογράμματος που υποστηρίζει την εκτίμηση της επιλεκτικότητας των επερωτήσεων χρονικής στιγμής, είναι να επαυξηθεί ο χώρος (το πεδίο αναφοράς των δεδομένων) που χρησιμοποιείται αρχικώς από την [APR99] ώστε να περιλαμβάνει και τη χρονική διάσταση. Άρα, το προτεινόμενο ιστόγραμμα είναι $H = \{B_i : \cup(B_i) = S \wedge \cap(B_i) = \varnothing\}$ και $B_i = \{[x_{i,L}, x_{i,U}], [y_{i,L}, y_{i,U}], [t_{i,L}, t_{i,U}]\}$. Η βασική ιδέα που μας επιτρέπει να χρησιμοποιήσουμε τη διαμέριση MinSkew για τους σκοπούς μας συνοψίζεται στα εξής: εφαρμογή ενός ομοιόμορφου πλέγματος n διαστημάτων σε κάθε χωρική διάσταση που σχηματίζει n^2 χωρικές περιοχές G_i,, επανάληψη του πλέγματος σε αρκετά ομοιόμορφα κατανεμημένα χρονικά αποτυπώματα t_k (k = 1...*now* – 1), και υπολογισμός του αριθμού των τροχιών $m_{i,k}$ που βρίσκονται εντός κάθε G_i σε κάθε t_k. Αν χρησιμοποιηθεί ένας αρκετά μεγάλος αριθμός t_k, τότε, ο αριθμός των τροχιών που βρίσκονται εντός της περιοχής G_i σε κάθε χρονικό αποτύπωμα κατά την περίοδο [t_k, t_{k+1}], μπορεί να θεωρηθεί ότι είναι ίσος με $m_{i,k}$· μ' άλλα λόγια, οι τροχιές μπορούν να θεωρηθούν σταθερές μεταξύ t_k και t_{k+1} αν το μήκος της περιόδου [t_k, t_{k+1}] είναι αρκετά μικρό. Κατόπιν μπορούμε να χρησιμοποιήσουμε απευθείας τον αλγόριθμο κατασκευής της [APR99] εφαρμόζοντας το ίδιο σύνολο ευριστικών ώστε να ελαχιστοποιήσουμε τη χωρική ασυμμετρία των κατασκευασμένων κάδων: στόχος και πάλι είναι να ομαδοποιήσουμε αρκετές περιοχές πλέγματος G_i σε αρκετά χρονικά αποτυπώματα t_k που έχουν παρόμοιες τιμές $m_{i,k}$, οδηγώντας έτσι σε μία ομαδοποίηση με όσο το δυνατόν μικρότερη χωρική ασυμμετρία.

Τέλος, η τοπική πυκνότητα *N*' της επερώτησης χρονικής στιγμής που εκτελείται σε ένα χρονικό αποτύπωμα t_k υπολογίζεται από το σταθμισμένο μέσο όρο των πυκνοτήτων των κάδων που αλληλεπικαλύπτονται N_i που είναι ταυτόχρονα σε ισχύ σε αυτό το συγκεκριμένο χρονικό αποτύπωμα:

$$N' = \frac{1}{4ab} \sum_{i:t_{i,L} \le t_k < t_{i,U}} (N_i \cdot A_i) \tag{5.42}$$

Η Εξ.(5.42) εκφράζει το γεγονός ότι η τοπική πυκνότητα *N*' υπολογίζεται από τη σταθμισμένη πυκνότητα N_i κάθε κάδου B_i με την αντίστοιχη περιοχή A_i που καλύπτει μερικώς τη δεδομένη περιοχή, κανονικοποιημένη από τη συνολική επιφάνεια.

5.4.2.2. Εκτίμηση της Επίδρασης της Αβεβαιότητας Χρησιμοποιώντας Χωροχρονικά Ιστογράμματα

Για να περάσουμε τώρα στο βασικό μας πρόβλημα, τα χωροχρονικά ιστογράμματα *MinSkew* μπορούν να χρησιμοποιηθούν για να εφαρμόσουμε την ανάλυσή μας σε μη ομοιόμορφα δεδομένα και να εκτιμήσουμε το σφάλμα που εισάγεται στα αποτελέσματα επερώτησης χωρίς να εκτελείται στην πράξη η επερώτηση. Πιο συγκεκριμένα, προτείνονται δύο εναλλακτικές προσεγγίσεις για την εκτίμηση των $E_P(R_{k,axb})$ και $E_N(R_{k,axb})$. Η πρώτη είναι να χρησιμοποιήσουμε απλά την εκτίμηση της τοπικής πυκνότητας που παράγεται από το χωροχρονικό ιστόγραμμα που έχουμε για την πυκνότητα που χρησιμοποιείται στο προτεινόμενο μοντέλο· αυτό μπορεί να επιτευχθεί αξιολογώντας τις Εξ.(5.1), Εξ.(5.13) ή Εξ.(5.27) χρησιμοποιώντας την τοπική πυκνότητα N', που προκύπτει από Εξ.(5.42), αντί της συνολικής πυκνότητας χώρου *N*.

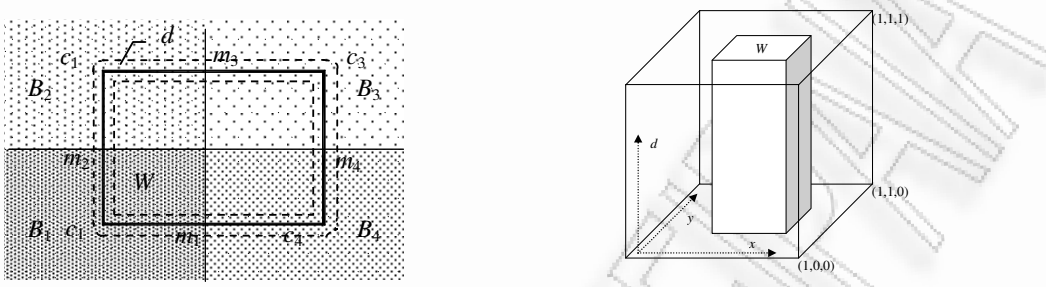
Ως εναλλακτική προσέγγιση, αντί του υπολογισμού της συνολικής τοπικής πυκνότητας N' για το συνολικό παράθυρο επερώτησης χρονικής στιγμής, μπορούμε να θεωρήσουμε τις διάφορες συνεισφορές των πλευρών του παραθύρου επερώτησης και των γωνιών του παραθύρου επερώτησης στο συνολικό αριθμό των λανθασμένων αποτελεσμάτων, όπως αναφέρθηκε στην Ενότητα 5.3.3. Συνεπώς, δεδομένου ενός χωροχρονικού ιστογράμματος που περιέχει *n* ξένους μεταξύ τους χωροχρονικούς κάδους B_i, η εκτίμηση του αριθμού των λανθασμένων θετικών και λανθασμένων αρνητικών στα αποτελέσματα μιας επερώτησης χρονικής στιγμής που καλείται στο χρονικό αποτύπωμα t_k βάσει της υπόθεσης ομοιομορφίας αβεβαιότητας, μπορεί να προσδιοριστεί χρησιμοποιώντας τον τύπο:

$$E_P\left(R_{k,a\times b}\right) = E_N\left(R_{k,a\times b}\right) = \sum_{i:t_{i,L} \le t_k < t_{i,U}} \left(N_i \cdot \left(\frac{2d}{3\pi} L_i - \frac{d^2}{8\pi} \cdot s_i \right) \right), \tag{5.43}$$

όπου L_i είναι το μήκος του τμήματος της περιμέτρου επερώτησης που αλληλεπικαλύπτεται χωρικά με το B_i και s_i είναι ο αριθμός των γωνιών του παραθύρου επερώτησης χρονικής στιγμής που βρίσκονται εντός B_i.

Η Εξ. (5.43) διατυπώνει το γεγονός ότι ο συνολικός αριθμός των λανθασμένων αρνητικών ή θετικών είναι το άθροισμα των συνεισφορών των διαφόρων συστατικών όπως συζητήθηκε στην Ενότητα 2.3. Πιο συγκεκριμένα, το τμήμα $\frac{2d}{3\pi} L_i$ της Εξ.(5.43) προκύπτει από το $\frac{8d}{3\pi}(a+b)$ της Εξ.(5.1) και Εξ.(5.13), επί το μήκος της περιμέτρου επερώτησης L_i που αλληλεπικαλύπτεται με τον κάδο B_i και διαιρείται με το συνολικό μήκος επερώτησης 4(*a*+*b*)· κατά τον ίδιο τρόπο, το τμήμα $\frac{d^2}{8\pi} s_i$

Εξ.(5.43) είναι ο μετασχηματισμός του τμήματος $\frac{d^2}{2\pi}$ της Εξ.(5.1) και Εξ.(5.13), επί του πραγματικού αριθμού των γωνιών του παραθύρου επερώτησης s_i που βρίσκονται χωρικά εντός του κάδου B_i, δια του συνολικού τους αριθμού, ήτοι, 4.



(a) (b)

Σχήμα 5.9: (a) Ένα παράθυρο επερώτησης χρονικής στιγμής σε μία στιγμιαία εικόνα του χωροχρονικού ιστογράμματος (b) ένα παράθυρο επερώτησης χρονικής στιγμής σε μία στιγμιαία εικόνα του επαυξημένου 4D χώρου.

Έστω, για παράδειγμα, το Σχήμα 5.9(a) που απεικονίζει την στιγμιαία εικόνα του παραθύρου επερώτησης χρονικής στιγμής W στο χρονικό αποτύπωμα t_k, που αλληλεπικαλύπτεται σε αυτό το συγκεκριμένο χρονικό αποτύπωμα με τέσσερεις κάδους του ιστογράμματος (B_1 … B_4). Επειδή τα λανθασμένα αποτελέσματα μπορούν να βρεθούν μόνο κοντά στο όριο του W, ο αριθμός λανθασμένων θετικών ή λανθασμένων αρνητικών στον κάδο B_1 εξαρτάται από το μήκος της περιμέτρου επερώτησης με την οποία αλληλεπικαλύπτεται, δηλαδή από το μήκος των γραμμών $|m_1c_1|+|c_1m_2|$ και τον αριθμό των γωνιών s_1=1. Αξίζει επίσης να επισημάνουμε ότι χρησιμοποιώντας την παραπάνω διαδικασία, το παράθυρο επερώτησης δεν κατακερματίζεται κατά μήκος των ορίων των κάδων ιστογραμμάτων, διότι μία τέτοια προσέγγιση θα αύξανε τη συνολική περίμετρο και κατά συνέπεια θα μείωνε την ακρίβεια του μοντέλου. Επιπλέον, στην περίπτωση της κατανομής αβεβαιότητας *2d-UDD*, ο τύπος για την εκτίμηση των λανθασμένων θετικών και λανθασμένων αρνητικών είναι:

$$E_P\left(R_{k,a\times b}\right)=E_N\left(R_{k,a\times b}\right)=\sum_{i:t_{i,L}\le t_k<t_{i,U}}\left(N_i\cdot\left(\frac{d}{2\pi}L_i-\frac{3d^2}{40\pi}\cdot s_i\right)\right), \tag{5.44}$$

Ο παραπάνω τύπος προκύπτει υπολογίζοντας τις συνεισφορές των πλευρών και γωνιών της επερώτησης στην Εξ.(5.27) κατά τρόπο ανάλογο με τα παραπάνω. Πιο συγκεκριμένα, το τμήμα $\frac{d}{2\pi}L_i$ της Εξ.(5.44) υπολογίζεται πολλαπλασιάζοντας το $\frac{2d}{\pi}(a+b)$ της Εξ.(5.27) επί του τμήματος της περιμέτρου της επερώτησης L_i που αλληλεπικαλύπτεται χωρικά με τον κάδο B_i, διαιρώντας το με το συνολικό μήκος επερώτησης 4(a+b), ενώ το τμήμα $\frac{3d^2}{40\pi}s_i$ της Εξ.(5.44) προκύπτει πολλαπλασιάζοντας το τμήμα $\frac{3d^2}{10\pi}$ της Εξ. (5.27) με τον πραγματικό αριθμό των γωνιών του παραθύρου επερώτησης s_i που βρίσκονται χωρικά εντός του κάδου B_i, δια του συνολικού τους αριθμού, ήτοι, 4.

Η ίδια μεθοδολογία μπορεί να εφαρμοσθεί σε οποιοδήποτε σχήμα αποθήκευσης δεδομένων που περιλαμβάνει συνοπτικές πληροφορίες, όπως σε κύβους δεδομένων σε αποθήκες δεδομένων τροχιάς (TDW). Επειδή ένας κύβος δεδομένων αποτελείται από ξένους μεταξύ του χωροχρονικούς κάδους, ήτοι, τα βασικά κυβοειδή, μαζί με συγκεντρωτικές πληροφορίες, οι Εξ.(5.43) και Εξ.(5.44), ανάλογα με τον τύπο της κατανομής αβεβαιότητας, μπορούν να εφαρμοσθούν σε λειτουργίες OLAP και να μας δώσουν μια εκτίμηση του συνολικού αριθμού λανθασμένων θετικών ή λανθασμένων αρνητικών. Για παράδειγμα, όταν συναθροίζουμε στοιχεία, από το επίπεδο του *κελιού* στο επίπεδο της *πόλης* όπως αναφέρθηκε στην εισαγωγή, ήτοι, εκτελώντας δηλαδή μία λειτουργία roll-up, το MBB μιας πόλης μπορεί να θεωρηθεί ως παράθυρο επερώτησης και να χρησιμοποιηθεί για την εκτίμηση των λανθασμένων αποτελεσμάτων που εισήχθησαν στη συνάθροιση στοιχείων. Δεδομένου, ωστόσο, ότι η πυκνότητα μεταξύ του ορίου της πραγματικής πόλης και του MBB της μπορεί να είναι πολύ διαφορετική, το N_i που περιλαμβάνεται στην Εξ.(5.43) ή Εξ.(5.44) θα πρέπει να προσδιοριστεί χρησιμοποιώντας την πραγματική περίμετρο του πολυγώνου της πόλης αντί του MBB της και τα μήκη L_i θα πρέπει να σταθμιστούν αναλόγως χρησιμοποιώντας το MBB και την περίμετρο του πολυγώνου. Η προσέγγιση αυτή θα αποτελέσει αντικείμενο δοκιμής στις επόμενες ενότητες για τα απλά χωρικά δεδομένα και όπως θα αποδειχθεί μας δίνει πολύ καλές εκτιμήσεις.

5.4.3. Χαλάρωση της Υπόθεσης Σταθερής Ακτίνας Αβεβαιότητας

Η τρίτη επέκταση του μοντέλου που παρουσιάζεται σε αυτή τη διατριβή για να υποστηρίξουμε σενάρια εφαρμογής στον πραγματικό κόσμο, είναι να ασχοληθούμε με σύνολα δεδομένων τροχιών που έχουν διαφορετικές τιμές ακτίνας αβεβαιότητας ή σταθερή απόκλιση για καθένα από αυτά. Έστω για παράδειγμα, m σύνολα P_j που περιλαμβάνουν N_j τροχιές το καθένα, που έχουν ληφθεί χρησιμοποιώντας διάφορες τεχνολογίες εντοπισμού θέσης, όπως GPS, εντοπισμό μέσω Wi-Fi κλπ. Τότε, η ένωση όλων των συνόλων $P = \bigcup_{i=1...m} \{P_j\}$ περιλαμβάνει τροχιές που έχουν αρκετές ακτίνες αβεβαιότητας ανάλογα με την αρχική πηγή δεδομένων της κάθε τροχιάς. Μία απλή προσέγγιση για να προσδιορίσουμε το σφάλμα E_P ή E_N που εισάγεται στα αποτελέσματα μιας επερώτησης χρονικής στιγμής στο P, είναι να υπολογίσουμε τα συγκεκριμένα σφάλματα $E_{P,j}$ ή $E_{N,j}$ για το κάθε P_j ξεχωριστά και μετά να αθροίσουμε τα σφάλματα που προκύπτουν. Πιο συγκεκριμένα,

$$E_P\left(R_{a\times b}\right) = \sum_{j=1..m} E_{P,j}\left(R_{a\times b}\right) \text{ και } E_N\left(R_{a\times b}\right) = \sum_{j=1..m} E_{N,j}\left(R_{a\times b}\right) \quad (5.45)$$

Μια τέτοια προσέγγιση λογικά θα έχει επιτυχή αποτελέσματα όταν έχουμε ομοιόμορφα κατανεμημένα δεδομένα. Ωστόσο, όταν έχουμε ασύμμετρα δεδομένα από τον πραγματικό κόσμο θα πρέπει να εφαρμοσθεί η μεθοδολογία που προτείνεται στην προηγούμενη ενότητα, που σημαίνει ότι θα πρέπει να διατηρήσουμε m διαφορετικά ιστογράμματα, ένα για κάθε διαφορετική τιμή της ακτίνας αβεβαιότητας. Μολαταύτα, στην παρούσα διατριβή παρέχεται μια πιο εξελιγμένη λύση ως απάντηση στην ανωτέρω πρόκληση. Εν προκειμένω, θα μπορούσαμε ν' επαυξήσουμε περαιτέρω το χωροχρονικό ιστόγραμμα που προτείνεται στην ενότητα 5.4.2.1, θεωρώντας την ακτίνα αβεβαιότητας ως την τέταρτη διάσταση. Μ' άλλα λόγια, προτείνουμε να χρησιμοποιηθεί το ιστόγραμμα MinSkew στον κανονικοποιημένο χώρο των 4 διαστάσεων δηλαδή τις δύο χωρικές διαστάσεις, τη χρονική και το μήκος της ακτίνας αβεβαιότητας d.

Πιο τυπικά, το προτεινόμενο ιστόγραμμα είναι $H = \left\{ B_i : \cup\left(B_i\right) = S \times [0,1] \wedge \cap\left(B_i\right) = \varnothing \right\}$ και $B_i = \left\{ \left[x_{i,L}, x_{i,U} \right], \left[y_{i,L}, y_{i,U} \right], \left[t_{i,L}, t_{i,U} \right], \left[d_{i,L}, d_{i,U} \right] \right\}$. Το ιστόγραμμα χτίζεται εφαρμόζοντας έναν ομοιόμορφο κάναβο στο $S \times [0,1]$ και μετρώντας τον αριθμών των σημείων που βρίσκονται μέσα σε κάθε κελί του 4D χώρου, και στη συνέχεια, διαμερίζοντας αναδρομικά το χώρο ,ελαχιστοποιώντας σε κάθε βήμα τη συνολική τιμή της χωρικής ασυμμετρίας. Ακολουθώντας την αντίστοιχη συζήτηση της προηγούμενης ενότητας, υποθέτεται ότι η κατανομή των δεδομένων μέσα σε κάθε 4D κάδο είναι ομοιόμορφη. Τότε, η εκτίμηση του αριθμού των λανθασμένων απαντήσεων στην περίπτωση της υπόθεσης ομοιόμορφης κατανομής θέσης είναι:

$$E_P\left(R_{a\times b}\right) = E_N\left(R_{a\times b}\right) = \sum_{i:t_{i,L} \le t_k < t_{i,U}} \left[\frac{N_i}{d_{i,U} - d_{i,L}} \cdot \int_{d_{i,L}}^{d_{i,U}} \left[\left(\frac{2d}{3\pi} L_i - \frac{d^2}{8\pi} s_i \right) dd \right] \right], \qquad (5.46)$$

όπου L_i είναι το μήκος της περιμέτρου της επερώτησης που αλληλεπικαλύπτεται με τον κάδο B_i στις δύο χωρικές διαστάσεις, s_i είναι ο αριθμός των γωνιών του παραθύρου που βρίσκεται μέσα στον κάδο B_i, και $d_{i,L}$, $d_{i,U}$ είναι οι ελάχιστη και μέγιστη τιμή της διάστασης του d στο B_i, αντίστοιχα. Η Εξ.(5.46) παράγεται απευθείας από όταν ολοκληρώνουμε την Εξ.(5.43) επάνω σε όλες τις πιθανές τιμές του d στον χώρο δεδομένων, λαμβάνοντας επίσης υπόψη ότι ο πραγματικός αριθμός των αντικειμένων που βρίσκονται σε κάθε φέτα της διάστασης του d είναι $N_i / \left(d_{i,U} - d_{i,L}\right)$ και $\left(d_{i,U} - d_{i,L}\right)$ είναι η έκταση του κάδου κατά μήκος της συγκεκριμένης διάστασης. Διαισθητικά, η παραπάνω εξίσωση εκφράζει το γεγονός ότι το συνολικό σφάλμα είναι το άθροισμα των σφαλμάτων που πραγματοποιούνται σε κάθε κάδο του ιστογράμματος που το παράθυρο της επερώτηση επικαλύπτει· επιπρόσθετα, σε αυτή τη περίπτωση, η χωρική συνιστώσα της ερώτησης W επαυξάνεται επίσης στην διάσταση του d, σχηματίζοντας ένα κουτί που καλύπτει εξολοκλήρου την διάσταση του d, όπως φαίνεται στο Σχήμα 5.9(b). Τελικά, η Εξ.(5.46), μετά τους αναγκαίους υπολογισμούς γίνεται:

$$E_P\left(R_{k,a\times b}\right) = E_N\left(R_{k,a\times b}\right) = \sum_{i:t_{i,L} \le t_k < t_{i,U}} \left[N_i \cdot \left(\frac{d_{i,U} + d_{i,L}}{3\pi} L_i - \frac{d_{i,U}^2 + d_{i,L}^2 + d_{i,L} d_{i,U}}{24\pi} s_i \right) \right], \qquad (5.47)$$

Ακολουθώντας μία παρόμοια προσέγγιση, η εκτίμηση του αριθμού των λανθασμένων αποτελεσμάτων στην περίπτωση της υπόθεσης διαφοράς ομοιομορφίας υπολογίζεται ως:

$$E_P\left(R_{k,a\times b}\right) = E_N\left(R_{k,a\times b}\right) = \sum_{i:t_{i,L} \le t_k < t_{i,U}} \left[N_i \cdot \left(\frac{d_{i,U} + d_{i,L}}{4\pi} L_i - \frac{d_{i,U}^2 + d_{i,L}^2 + d_{i,L} d_{i,U}}{40\pi} s_i \right) \right]. \qquad (5.48)$$

Η προτεινόμενη προσέγγιση έχει δύο βασικά πλεονεκτήματα σε σχέση με την εναλλακτική της διατήρησης m διαφορετικών ιστογραμμάτων ένα για κάθε σύνολο τροχιών· το πρώτο είναι ότι οι απαιτήσεις χώρου μειώνονται επαρκώς, κυρίως στην περίπτωση όπου ο αριθμός των διαφορετικών ακτινών αβεβαιότητας αυξάνεται σημαντικά. Ωστόσο, το πλέον σημαντικό πλεονέκτημα της πρότασης αυτής αποκαλύπτεται αν λάβουμε υπ' όψιν ότι τα δεδομένα που ανήκουν στην ίδια κατηγορία μπορεί να έχουν διαφορετική ακρίβεια· για παράδειγμα η αβεβαιότητα λόγω του GPS εξαρτάται από μεγάλο αριθμό παραμέτρων, όπως τον αριθμό των ορατών δορυφόρων, την παρεμβολή συχνότητας και τον αντικατοπτρισμό του δορυφορικού σήματος σε μεγάλες γυάλινες επιφάνειες εντός των αστικών περιοχών, που σημαίνει διαφορετική ακτίνα αβεβαιότητας για κάθε δειγματοληπτούμενο σημείο κάθε

τροχιάς· η απλοϊκή προσέγγιση δεν θα μπορούσε να εκπληρώσει τέτοιες απαιτήσεις διότι θα πρέπει να διατηρήσουμε ένα ξεχωριστό ιστόγραμμα για κάθε πιθανή τιμή της ακτίνας αβεβαιότητας. Από την άλλη πλευρά, η πρόταση μας μπορεί να απορροφήσει αυτές τις ανάγκες και να αντιμετωπίσει έναν απεριόριστο αριθμό διαφορετικών ακτινών χωρίς να αυξάνεται η απαίτηση χώρου μνήμης του κατασκευασμένου ιστογράμματος, δίνοντας ταυτόχρονα μία πολύ καλή εκτίμηση.

5.5. Πειραματική Μελέτη: Χωροχρονικά Δεδομένα

Στην ενότητα αυτή παρουσιάζονται αρκετά πειράματα προς απόδειξη της ορθότητας και ακρίβειας της προηγούμενης ανάλυσης χρησιμοποιώντας συνθετικά σύνολα δεδομένων τροχιών. Στην πειραματική μελέτη που ακολουθεί αποδεικνύεται η ορθότητα του αναλυτικού μοντέλου σε ομοιόμορφη κατανομή της αβεβαιότητας με τη βοήθεια των χωροχρονικών ιστογραμμάτων (Εξ.(5.43)), καθώς και της ευαισθησίας του ως προς τις σχετικές παραμέτρους, ήτοι, την ακτίνα αβεβαιότητας και το μήκος της περιμέτρου της επερώτησης.

Θα πρέπει ίσως εδώ να επισημάνουμε ότι για τυπικά μεγέθη επερώτησης και αβεβαιότητας (π.χ. επερωτήσεις 0.05×0.05 έως 0.30×0.30 στο μοναδιαίο χώρο και ακτίνα αβεβαιότητας στο 0.01), οι τύποι του προτεινόμενου μοντέλου μας δίνουν τιμές λανθασμένων αρνητικών / θετικών μεταξύ $0.0004\times N$ και $0.0025\times N$, που σημαίνει ότι για 2000 τροχιές αναμένουμε μεταξύ 0.8 και 5.0 τροχιές ως λανθασμένες θετικές / αρνητικές ανά επερώτηση. Άρα, είναι σαφές ότι για τυπικά μεγέθη επερώτησης και ακτίνες αβεβαιότητας, ο πληθυσμός του συνόλου δεδομένων θα πρέπει να είναι αρκετά μεγάλος ώστε να μας δώσει ένα σημαντικό αριθμό λανθασμένων αποτελεσμάτων, που θα μετρήσουμε και θα συγκρίνουμε με τα αποτελέσματα του προτεινόμενου μοντέλου. Ωστόσο, επειδή το πλήθος των τροχιών σε ένα σύνολο δεδομένων είναι συνήθως μικρό (από την άλλη πλευρά, το πραγματικό τους μέγεθος μπορεί να γίνει τεράστιο συν το χρόνω), στην επόμενη ενότητα θα εξετάσουμε τις λεπτομέρειες του αναπτυχθέντος μοντέλου, χρησιμοποιώντας συνθετικά και πραγματικά σύνολα χωρικών δεδομένων.

5.5.1. Πειραματικό Πλαίσιο

Η πειραματική μελέτη για χωροχρονικά δεδομένα βασίζεται στο συνθετικό σύνολο δεδομένων NG2000 (ενότητα 1.5.3). Κάθε τροχιά μοντελοποιήθηκε ως κυλινδρικός όγκος [TWHC04], ακολουθώντας την υπόθεση αβεβαιότητας της ομοιόμορφης κατανομής. Κατά τη διάρκεια κάθε πειράματος το σύνολο των δεδομένων επερωτήθηκε με 1000 τυχαία κατανεμημένες τετράγωνες, ήτοι, με $a=b$, επερωτήσεις χρονικής στιγμής. Κάθε επερώτηση αρχικώς ανέκτησε την παρεμβαλλόμενη θέση κάθε τροχιάς στο χρονικό αποτύπωμα που καθορίζεται από αυτή, και κατόπιν χρησιμοποιήθηκε η υπόθεση της [TWHC04] για να αποκαλυφθεί η πραγματική θέση του κάθε κινούμενου αντικειμένου σε αυτό το συγκεκριμένο χρονικό αποτύπωμα. Άρα, προέκυψε ένας αριθμός λανθασμένων αρνητικών και λανθασμένων θετικών, αφού τα αποτελέσματα της επερώτησης που συλλέχθηκαν από το πρώτο βήμα ήταν διαφορετικά από αυτά που προσδιορίστηκαν μετά το δεύτερο βήμα . Χρησιμοποιήσαμε για την εκτίμηση του ίδιου αριθμού του αναλυτικού μας μοντέλου που εκφράζεται από την Εξ.(5.43), δηλαδή, με τη βοήθεια των χωροχρονικών ιστογραμμάτων, όπως παρουσιάστηκαν σε προηγούμενη ενότητα (χαλαρώνοντας έτσι την υπόθεση A_{II})· η διαμέριση *MinSkew* του συνόλου δεδομένων που εξετάζουμε δημιουργήθηκε χρησιμοποιώντας ένα ομοιόμορφο πλέγμα με αρχικό μέγεθος $0.005\times0.005\times0.005$,

όπως συζητήθηκε στην [APR99]. Η ακτίνα του κυλινδρικού όγκου (ακτίνα αβεβαιότητας) κλιμακώθηκε μεταξύ 0.0005 και 0.02, ενώ το μήκος της πλευράς κάθε τετράγωνης επερώτησης κλιμακώθηκε μεταξύ 0.06 και 0.36· επιμήκη παράθυρα επερώτησης ανέφεραν παρόμοια συμπεριφορά. Τα πειράματά μας διεξήχθησαν σε ένα σταθμό εργασίας Windows XP με AMD Athlon 64 3GHz επεξεργαστή CPU, 1 GB κύριας μνήμης και αρκετά GB χώρου δίσκου.

5.5.2. Πειραματικά Αποτελέσματα

Χρησιμοποιήθηκαν δύο στατιστικές μετρήσεις για να αποδείξουμε τη συμπεριφορά του μοντέλου μας. Ο *μέσος αριθμός λανθασμένων αρνητικών και λανθασμένων θετικών*, $\overline{E_N}$ και $\overline{E_P}$, αντίστοιχα και το *μέσο απόλυτο σφάλμα στην εκτίμηση των λανθασμένων αρνητικών και λανθασμένων θετικών σε κάθε ανεξάρτητη επερώτηση*, $\overline{ES_N}$ και $\overline{ES_P}$, αντίστοιχα. Επισήμως, αυτές οι μετρήσεις ορίζονται ως:

$$\overline{E_N} = \frac{1}{n}\sum_{i=1..n} E_{N,i}\ ,\ \overline{E_P} = \frac{1}{n}\sum_{i=1..n} E_{P,i} \tag{5.49}$$

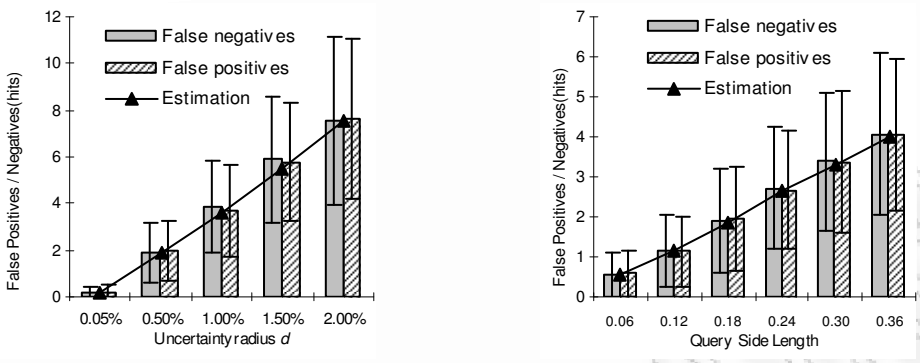
και,

$$\overline{ES_N} = \frac{1}{n}\sum_{i=1..n} \left|E_{N,i} - E_N\left(R_{k,a\times b}\right)\right|,\ \overline{ES_P} = \frac{1}{n}\sum_{i=1..n} \left|E_{P,i} - E_P\left(R_{k,a\times b}\right)\right| \tag{5.50}$$

όπου n είναι ο αριθμός των επερωτήσεων που εκτελέστηκαν και $E_{P,i}$ ($E_{N,i}$) ο πραγματικός αριθμός λανθασμένων θετικών (λανθασμένων αρνητικών, αντίστοιχα) στην i-στη επερώτηση. Κάνουμε διάκριση μεταξύ π.χ. $\overline{E_P}$ και $\overline{ES_P}$, ώστε να αποκαλύψουμε τις λεπτομέρειες της συμπεριφοράς του μοντέλου μας, όπως θα αποδειχθεί στα ακόλουθα πειράματα.

Στην πρώτη σειρά πειραμάτων τα συνθετικά σύνολα δεδομένων χρησιμοποιούνται προκειμένου να αποδειχθεί η ακρίβεια και η συμπεριφορά του παρουσιαζόμενου αναλυτικού μοντέλου κλιμακώνοντας τους δύο παράγοντες επιρροής: την ακτίνα d του δίσκου αβεβαιότητας και το μέγεθος (a, b) του παραθύρου επερώτησης. Σημειώστε ότι σε όλα τα σχήματα το μέγεθος επερώτησης εκτίθεται σε όρους του μήκους της πλευράς $2a = 2b$, π.χ. για πλευρικό μήκος επερώτησης 0.30, το μέγεθος του παραθύρου επερώτησης είναι ίσο με $0.30 \times 0.30 = 0.09$ του μοναδιαίου χώρου.

Πιο συγκεκριμένα, στο πρώτο πείραμα η τιμή του d κλιμακώνεται μεταξύ του 0.05% και 2% της έκτασης του χώρου κατά μήκος των αξόνων x- και y-, επερωτώντας το συνθετικό σύνολο δεδομένων, με σταθερό πλευρικό μήκος 0.18 (ήτοι, $a = b = 0.09$ που έχει ως αποτέλεσμα ένα παράθυρο επερώτησης μεγέθους 3.24% του χώρου δεδομένων). Τα αποτελέσματα αυτού του πειράματος φαίνονται στο Σχήμα 5.10(a)· ένα πρώτο αποτέλεσμα είναι ότι ο αριθμός των λανθασμένων θετικών και λανθασμένων αρνητικών αποδεικνύεται σχεδόν ίσος, επαληθεύοντας την ορθότητα του πορίσματος της Εξ.(5.22). Επιπλέον, οι εκτιμήσεις $E_P(R_{k,a\times b})$ και $E_N(R_{k,\times b})$ είναι πολύ ακριβείς σε σχέση με τα $\overline{E_P}$ και $\overline{E_N}$, με το σφάλμα να είναι πάντα κάτω του 6%, ενώ οι γραμμές σφάλματος σε κάθε στήλη του διαγράμματος, που αναπαριστούν τα $\overline{ES_P}$ και $\overline{ES_N}$, παρουσιάζουν χαμηλές έως μέτριες τιμές. Πιο συγκεκριμένα, το μέσο σφάλμα στην κάθε επερώτηση είναι περίπου 40% στην συντριπτική πλειονότητα των πειραματικών ρυθμίσεων ενώ αυξάνεται σημαντικά μόνο στην ακραία περίπτωση όπου η ακτίνα αβεβαιότητας d τίθεται στο ελάχιστο ($d = 0.05\%$).



(a) (b)

Σχήμα 5.10: Μέσες λανθασμένες αρνητικές / θετικές και οι εκτιμήσεις τους κλιμακούμενες με (a) *d* και (b) το μέγεθος επερώτησης (συνθετικά δεδομένα – ομοιόμορφη κατανομή αβεβαιότητας)

Παρόμοια αποτελέσματα διαπιστώνουμε στο δεύτερο πείραμα, που φαίνεται στο Σχήμα 5.10(b), και στο οποίο κλιμακώνεται το μέγεθος επερώτησης. Εν προκειμένω, η ακτίνα αβεβαιότητας τίθεται 0.5%, και το μήκος της πλευράς επερώτησης κλιμακώνεται μεταξύ 0.06 και 0.36, με αποτέλεσμα μεγέθη επερώτησης που καλύπτουν μεταξύ του 0.36% και 12.96% του χώρου δεδομένων. Όταν συγκρίνουμε την εκτίμηση του αριθμού των λανθασμένων αρνητικών και λανθασμένων θετικών και τις αντίστοιχες μέσες τιμές $\overline{E_P}$ και $\overline{E_N}$, το σφάλμα εκτίμησης που λαμβάνουμε είναι και πάλι κάτω του 6%, ανεξαρτήτως του μεγέθους επερώτησης, ενώ οι γραμμές σφάλματος σε κάθε στήλη του διαγράμματος (ήτοι, $\overline{ES_P}$ και $\overline{ES_N}$) παρουσιάζουν την ίδια τάση με προηγουμένως και κυμαίνονται γύρω στο 40%· και πάλι η μόνη περίπτωση στην οποία αγγίζουν υψηλές τιμές, είναι όταν και το σ και το μέγεθος επερώτησης τίθενται στις ελάχιστες τιμές τους.

Ενώ εκ πρώτης όψεως οι τιμές αυτές των $\overline{ES_P}$ και $\overline{ES_N}$ μπορεί να θεωρηθούν υψηλές, πρέπει να επισημάνουμε ότι το σφάλμα της εκτίμησης μειώνεται σημαντικά όσο αυξάνεται το πλήθος του συνόλου δεδομένων, γεγονός το οποίο θα αποδειχθεί για τα χωρικά δεδομένα στην επόμενη ενότητα. Τέλος, για να δικαιολογήσουμε την ακρίβεια των εκτιμήσεων, πρέπει να πούμε ότι οι τιμές των $\overline{ES_P}$ και $\overline{ES_N}$ ποτέ δεν υπερβαίνουν τα 2 λανθασμένα αποτελέσματα σε απόλυτες τιμές (π.χ. πραγματικές προς εκτιμώμενες λανθασμένες αρνητικές : 6 προς 8).

5.6. Πειραματική Μελέτη: Χωρικά Δεδομένα

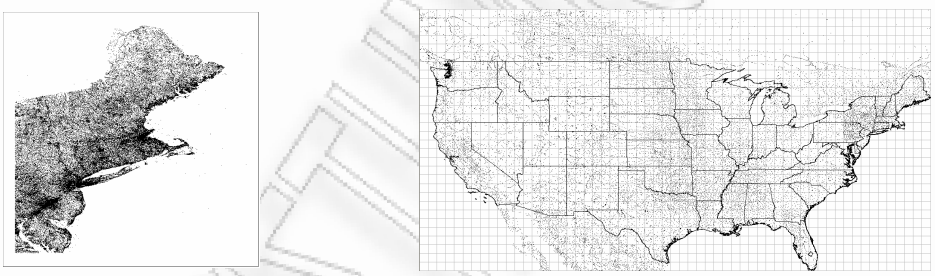
Στην ενότητα αυτή, ακολουθώντας τα προηγούμενα πειράματα για χωροχρονικά δεδομένα τροχιών, παρουσιάζουμε μια σειρά πειραμάτων χρησιμοποιώντας συνθετικά και πραγματικά (σταθερά) χωρικά δεδομένα ώστε να αποκαλυφθούν όλες οι λεπτομέρειες του προτεινόμενου μοντέλου, σε σύνολα δεδομένων με μεσαίο πλήθος (ωστόσο, αρκετά μεγαλύτερο από αυτό που είχαμε για την πειραματική μελέτη των χωροχρονικών δεδομένων), καθώς και την αποτελεσματικότητα των προτεινόμενων λύσεων. Επακριβώς, οι στόχοι της κάτωθι πειραματικής μελέτης είναι να:

- αποδείξουμε την ακρίβεια του απλού αναλυτικού μοντέλου (Εξ.(5.1) και Εξ.(5.13)), καθώς και την ευαισθησία του σε σχέση με τις ανάλογες παραμέτρους, ήτοι, την ακτίνα αβεβαιότητας, ή την τυπική απόκλιση και το μήκος της περιμέτρου επερώτησης.

- δείξουμε την ποιότητα της προσέγγισης της κανονικά κατανεμημένης αβεβαιότητας θέσης με την *2d-UDD* χρησιμοποιώντας το μοντέλο των Εξ.(5.27) και Εξ.(5.32)
- παρουσιάσουμε την ακρίβεια της εκτίμησης που παρέχεται από τα αναλυτικά μοντέλα– Εξ.(5.43), Εξ.(5.44), Εξ.(5.47), και Εξ.(5.48) – σε πραγματικά χωρικά δεδομένα χρησιμοποιώντας ιστογράμματα και επίσης να αποδείξουμε ότι πλεονεκτούν σε σχέση με την εναλλακτική του να χρησιμοποιείται το ιστόγραμμα ως εκτιμητής της πυκνότητας Εξ.(5.41),
- δείξουμε πως η πρόταση αυτού του κεφαλαίου μπορεί να χρησιμοποιηθεί στα πλαίσια των αποθηκών χωρικών δεδομένων, και,
- αποκαλύψουμε την αποτελεσματικότητα των λύσεων που προτείνουμε εφαρμόζοντάς τες σε ένα εμπορικό SDBMS.

5.6.1. Πειραματικό Πλαίσιο

Η πειραματική μελέτη αυτής της ενότητας για τα χωρικά δεδομένα βασίζεται σε συνθετικά και πραγματικά σύνολα σημειακών δεδομένων. Πιο συγκεκριμένα, τα χρησιμοποιούμενα σύνολα δεδομένων είναι τα εξής: ένα συνθετικό σύνολο δεδομένων (Rnd_0) 100K 2D σημείων τυχαία κατανεμημένων στον μοναδιαίο χώρο καθώς και δύο σύνολα δεδομένων, πιο συγκεκριμένα, τα σύνολα δεδομένων North East (*NE*) και Digital Chart of the World (*DCW*), που φαίνονται στο Σχήμα 5.11(a) και (b), αντίστοιχα.



(a) (b)

Σχήμα 5.11: Πραγματικά σύνολα δεδομένων: (a) North East και (b) Digital Chart of the World

Κατόπιν, όπως προτείνεται από τις [BS03], [CZBP06], [GL05], σε κάθε σύνολο δεδομένων προστίθεται ελεγχόμενος θόρυβος. Πιο συγκεκριμένα, η θέση κάθε σημείου και στα τρία σύνολα δεδομένων τροποποιείται προσθέτοντας θόρυβο, είτε ομοιόμορφα κατανεμημένο εντός ενός δίσκου αβεβαιότητας ακτίνας d , που μας δίνει το αντίστοιχο σύνολο δεδομένων *U-d*, ή ακολουθώντας μία διμεταβλήτη κανονική κατανομή με τυπική απόκλιση σ, που μας δίνει το αντίστοιχο σύνολο δεδομένων *N-σ* · για κάθε σύνολο δεδομένων *U-d* και *N-σ*, δημιουργήθηκαν πέντε διαφορετικά σύνολα δεδομένων τα $Rnd_{U\text{-}d\text{-}1}$, με $Rnd_{U\text{-}d\text{-}5}$, $NE_{U\text{-}d\text{-}1}$ με $NE_{U\text{-}d\text{-}5}$, και $DCW_{U\text{-}d\text{-}1}$ με $DCW_{U\text{-}d\text{-}5}$, και επίσης τα ίδια πέντε σύνολα δεδομένων για κάθε μία από τις περιπτώσεις $Rnd_{N\text{-}\sigma}$, $NE_{N\text{-}\sigma}$, $DCW_{N\text{-}\sigma}$. Επιπλέον, για να δοκιμάσουμε την ακρίβεια των εκτιμήσεων μας για τις ρυθμίσεις της Ενότητας 3.3, είχαμε το σύνολο δεδομένων $NE_{N\text{-}v0.02}$ στο οποίο προσθέσαμε θόρυβο ακολουθώντας τη διμεταβλήτη κανονική κατανομή με τη σ μεταξύ 0 και 0.02. Εκτός κι αν αναφέρεται διαφορετικά, όλα τα πειράματα που περιλαμβάνουν χωρικές επερωτήσεις έγιναν τρέχοντας 1000 τυχαία κατανεμημένα τετράγωνα, ήτοι, με επερωτήσεις *a=b*, και στα πέντε σύνολα δεδομένων της αντίστοιχης περίπτωσης· τα επιμήκη παράθυρα επερώτησης επέδειξαν ανάλογη συμπεριφορά. Όλα τα πειράματά μας διεξήχθησαν σε ένα σταθμό εργασίας Windows XP με AMD Athlon 64 3GHz επεξεργαστή CPU, 1 GB κύριας μνήμης και αρκετά GB

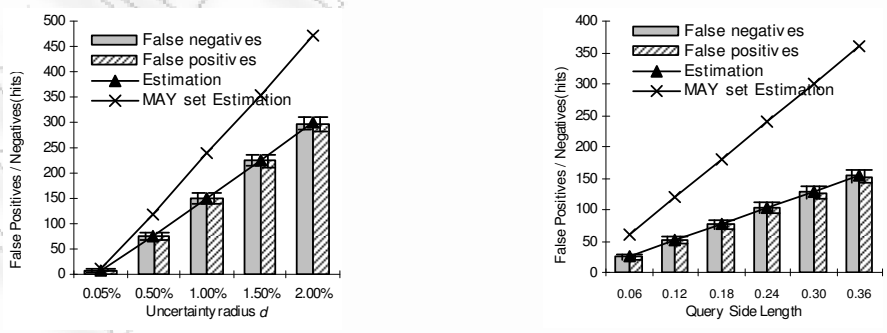
χώρου δίσκου · όλες οι μέθοδοι που αξιολογήθηκαν υλοποιήθηκαν σε VB.NET καθώς και στο περιβάλλον της PostgreSQL 8.2 [Post08a] με την επέκταση PostGIS 1.2.1 [Post08b] χρησιμοποιώντας τη PL/PgSQL.

5.6.2. Πειράματα ως προς την Ποιότητα

Όπως και στην πειραματική μελέτη για τα χωροχρονικά δεδομένα, σε αυτή την ενότητα χρησιμοποιούμε επίσης το μέσο αριθμό λανθασμένων αρνητικών και λανθασμένων θετικών, $\overline{E_N}$ και $\overline{E_P}$ (Εξ.(5.49)), καθώς και το μέσο απόλυτο σφάλμα στην εκτίμηση των λανθασμένων αρνητικών και λανθασμένων θετικών σε κάθε επερώτηση, $\overline{ES_N}$ και $\overline{ES_P}$ (Εξ.(5.50)).

5.6.2.1. Πειράματα σε Συνθετικά Δεδομένα Ακολουθώντας και τις τρεις Αρχικές Υποθέσεις A_I, A_{II}, A_{III}

Στην πρώτη σειρά πειραμάτων τα συνθετικά σύνολα δεδομένων χρησιμοποιούνται για να αποδείξουμε την ακρίβεια και τη συμπεριφορά του αναλυτικού μοντέλου κλιμακώνοντας τους δύο παράγοντες επιρροής όπως κάναμε και στην προηγούμενη ενότητα για τα χωροχρονικά δεδομένα. Στο πρώτο πείραμα η τιμή της d κλιμακώθηκε μεταξύ 0.05% και 2% της χωρικής έκτασης κατά μήκος των αξόνων x- και y-, επερωτώντας και το σύνολο δεδομένων Rnd_0 και το αντίστοιχο $Rnd_{u\text{-}d}$, με σταθερό μήκος πλευράς 0.18. Τα αποτελέσματα αυτού του πειράματος φαίνονται στο Σχήμα 5.12(a)· ένα πρώτο αποτέλεσμα είναι ότι οι εκτιμήσεις $E_P(R_{a\times b})$ και $E_N(R_{a\times b})$ είναι εξαιρετικά ακριβείς σε σχέση με τα $\overline{E_P}$ και $\overline{E_N}$, με σφάλμα πάντα μικρότερο του 3%, ενώ οι γραμμές σφάλματος σε κάθε στήλη του διαγράμματος που παριστάνουν τα $\overline{ES_P}$ και $\overline{ES_N}$, είναι σχετικά χαμηλές. Πιο συγκεκριμένα το μέσο σφάλμα στην κάθε επερώτηση είναι κάτω του 10% στην συντριπτική πλειοψηφία των πειραματικών ρυθμίσεων και αγγίζει το 29% σε μία μεμονωμένη ακραία περίπτωση όπου η ακτίνα αβεβαιότητας d τίθεται στο ελάχιστο (d = 0.05%). Επιβεβαιώνεται συνεπώς η αρχική πρόθεση της πειραματικής μελέτης για χωρικά σύνολα δεδομένων, δηλαδή, να αποδείξουμε ότι οι εκτιμήσεις που προκύπτουν από το προτεινόμενο αναλυτικό μοντέλο σε χωρικά σύνολα δεδομένων μεσαίου πλήθους είναι πολύ καλύτερες από αυτές που προκύπτουν σε σύνολα δεδομένων μικρού πλήθους (όπως αυτά που χρησιμοποιήθηκαν στην προηγούμενη ενότητα).



(a) (b)

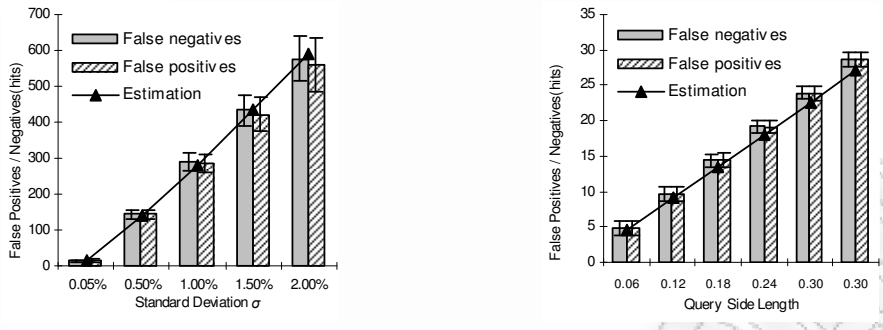
Σχήμα 5.12: Μέσες λανθασμένες αρνητικές / θετικές και οι εκτιμήσεις τους κλιμακούμενες με (a) d και (b) το μέγεθος επερώτησης (συνθετικά δεδομένα – ομοιόμορφη κατανομή αβεβαιότητας)

Στα ίδια πειράματα συμπεριλήφθηκε και η μεθοδολογία που παρέχεται από την [YM03], που εκτιμά το πλήθος του συνόλου *MAY*. Όπως έχουμε ήδη πει, το σύνολο *MAY* είναι στην πράξη ένα υπερσύνολο που περιλαμβάνει, μεταξύ άλλων, τα λανθασμένα αποτελέσματα που υπολογίζονται από την ανάλυσή μας · για να ξεπεράσουμε αυτό το πρόβλημα, εισάγουμε την υπόθεση ότι το 50% του συνόλου *MAY* είναι λανθασμένα αποτελέσματα, δηλαδή ένα αντικείμενο του συνόλου *MAY* έχει την ίδια πιθανότητα να είναι σωστό ή λανθασμένο αποτέλεσμα. Ωστόσο, όπως φαίνεται στο Σχήμα 5.12 από την καμπύλη *MAY set estimation*, η παραπάνω υπόθεση δεν οδηγεί σε σωστές εκτιμήσεις. Θα πρέπει να επισημάνουμε, ωστόσο, ότι ο στόχος της ανάλυσης που παρουσιάστηκε στην [YM03] δεν είναι να δώσει τον αριθμό των λανθασμένων αποτελεσμάτων όπως η ανάλυσή μας. Η υπόθεσή μας σχετικά με το τμήμα του συνόλου *MAY* που συνιστά λανθασμένα αποτελέσματα, ήτοι, το 50%, χρησιμοποιείται ελλείψει άλλων προτάσεων για το θέμα που να περιλαμβάνει η [YM03]. Επιπλέον, το Σχήμα 5.12(a) θα μπορούσε ίσως να οδηγήσει στην προϋπόθεση ένας απλός πολλαπλασιαστής στην εκτίμηση του συνόλου *MAY*, ήτοι, τη μείωση της αντίστοιχης καμπύλης στο Σχήμα 5.12, θα μπορούσε να την υποχρεώσει να δώσει καλύτερα αποτελέσματα. Και πάλι, βέβαια, για να προσδιορίσουμε τον πολλαπλασιαστή αυτό, θα πρέπει να ακολουθήσουμε τη μεθοδολογία που δίνεται στην ανάλυση που παρουσιάσαμε.

Παρόμοια αποτελέσματα προκύπτουν και στο δεύτερο πείραμα, που φαίνεται στο Σχήμα 5.12(b), όπου κλιμακώνεται το μέγεθος επερώτησης. Πιο συγκεκριμένα, η ακτίνα αβεβαιότητας τίθεται στο 0.5%, και το μήκος της πλευράς επερώτησης κλιμακώνεται μεταξύ 0.06 και 0.36, δίνοντας μας μεγέθη επερώτησης που καλύπτουν μεταξύ 0.36% και 12.96% του χώρου δεδομένων. Όταν συγκρίνουμε την εκτίμηση του αριθμού των λανθασμένων αρνητικών και λανθασμένων θετικών και τις αντίστοιχες μέσες τιμές $\overline{E_P}$ και $\overline{E_N}$, το αναφερθέν σφάλμα εκτίμησης είναι κάτω του 1%, ανεξαρτήτως του μεγέθους επερώτησης. Επιπλέον, η εκτίμηση βάσει του πλήθους του συνόλου *MAY*, για άλλη μια φορά δεν μπόρεσε να δώσει συγκρίσιμα αποτελέσματα · άρα, βάση της παρατήρησης ότι αυτή η εκτίμηση συστηματικά υπερεκτιμά τις $\overline{E_P}$ και $\overline{E_N}$, εξαιρείται από το υπόλοιπο της πειραματικής μελέτης. Όσον αφορά τις γραμμές σφάλματος σε κάθε στήλη του διαγράμματος, φαίνονται στις αντίστοιχες $\overline{ES_P}$ και $\overline{ES_N}$, είναι σχετικά μικρές καθώς στην πλειονότητα των πειραμάτων είναι κάτω του 16%· η μόνη περίπτωση στην οποία αγγίζουν υψηλότερες τιμές, ήτοι, 35%, συνέβη όταν και η *σ* και το μέγεθος της επερώτησης τέθηκαν στις ελάχιστες τιμές τους.

5.6.2.2. Πειράματα σε Συνθετικά Δεδομένα Χαλαρώνοντας την Υπόθεση A_I

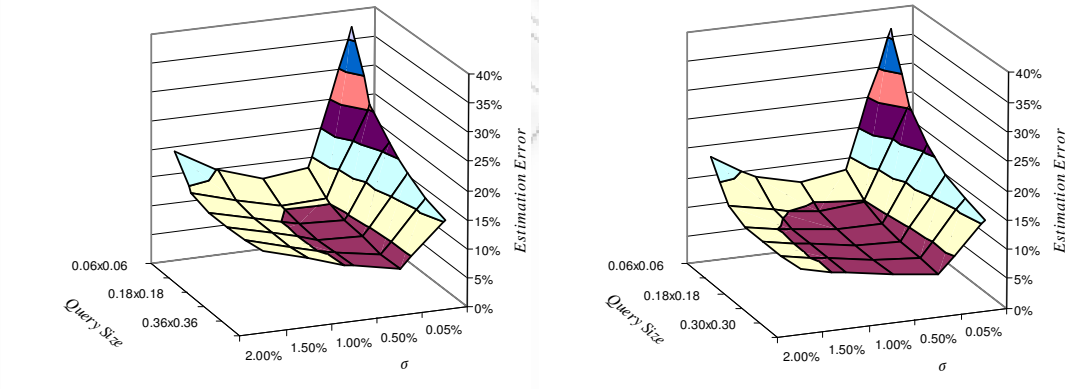
Για να αξιολογηθεί η ακρίβεια της εκτίμησης του αριθμού των λανθασμένων θετικών και λανθασμένων αρνητικών που υπολογίζονται από τις Εξ.(5.27), και Εξ.(5.32), εκτελέστηκε ένα πείραμα ανάλογο του συνόλου δεδομένων $Rnd_{N\text{-}\sigma}$ όπου κλιμακώθηκαν η *σ* και το μέγεθος επερώτησης. Τα αποτελέσματα αυτών των πειραμάτων φαίνονται στο Σχήμα 5.13 και είναι σαφές ότι το σφάλμα εκτίμησης για τις $\overline{E_P}$ και $\overline{E_N}$ είναι πάντα κάτω του 5%. Επιπλέον, ενώ οι γραμμές σφάλματος, που παριστάνουν τα $\overline{ES_P}$ και $\overline{ES_N}$, είναι σχετικά μικρές, συνήθως κάτω του 12 %, ενώ αγγίζουν το 36% μόνο στην περίπτωση όπου και η *d* και το μήκος της πλευράς επερώτησης τίθενται στις ελάχιστες τιμές τους.



(a) (b)

Σχήμα 5.13: Μέσες λανθασμένες αρνητικές / θετικές και οι εκτιμήσεις τους κλιμακούμενες με (a) σ και (b) το μέγεθος επερώτησης (συνθετικά δεδομένα – ομοιόμορφη κατανομή αβεβαιότητας)

Μία πιο εκτενής παρουσίαση του μέσου σφάλματος εκτίμησης σε κάθε ανεξάρτητη επερώτηση $\overline{ES_P}$ και $\overline{ES_N}$ γίνεται στο Σχήμα 5.14(a) και (b), ως ποσοστό του αριθμού των λανθασμένων θετικών και λανθασμένων αρνητικών, αντίστοιχα. Και τα δύο σχήματα δείχνουν ότι οι $\overline{ES_P}$ και $\overline{ES_N}$ ποικίλουν από χαμηλές τιμές, ήτοι, μικρότερες του 10% για υψηλές τιμές της σ, μέχρι υψηλότερες για πολύ χαμηλές τιμές του σ. Εξαρτώνται επίσης από το μέγεθος της επερώτησης, αυξανόμενες με τη μείωση του μεγέθους. Σε γενικές γραμμές, φαίνεται ότι οι $\overline{ES_P}$ και $\overline{ES_N}$ εξαρτώνται κατά βάση από την τυπική απόκλιση σ και, σε μικρότερο βαθμό από το μέγεθος επερώτησης. Επιπλέον, για μικρές τιμές της σ και μικρά μεγέθη επερώτησης, ενώ η εκτίμηση εξακολουθεί να παραμένει ακριβής σχετικά με τις $\overline{E_P}$ και $\overline{E_R}$ (Σχήμα 5.13(a) και (b), αντίστοιχα), οι $\overline{ES_P}$ και $\overline{ES_N}$ αυξάνονται σημαντικά αγγίζοντας το 40%.



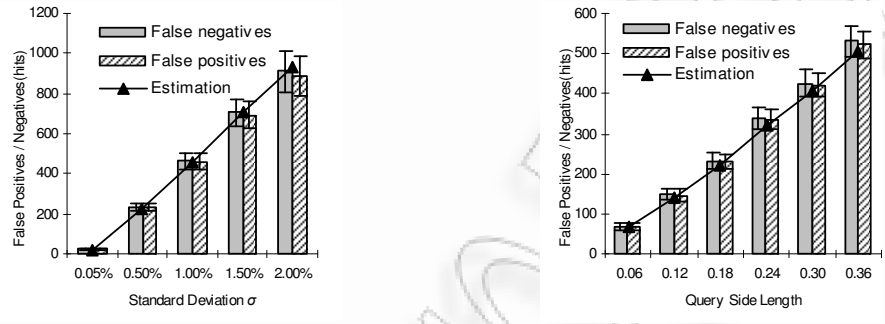
(a) (b)

Σχήμα 5.14: Μέσο σφάλμα εκτίμησης των (a) λανθασμένων θετικών $\overline{ES_P}$ και (b) λανθασμένων αρνητικών $\overline{ES_N}$, σε κάθε επερώτηση, κλιμακούμενες με το *d* και το μέγεθος επερώτησης (συνθετικά δεδομένα – ομοιόμορφη κατανομή αβεβαιότητας)

5.6.2.3. Πειράματα σε Πραγματικά Δεδομένα Χαλαρώνοντας την Υπόθεση A_{II}

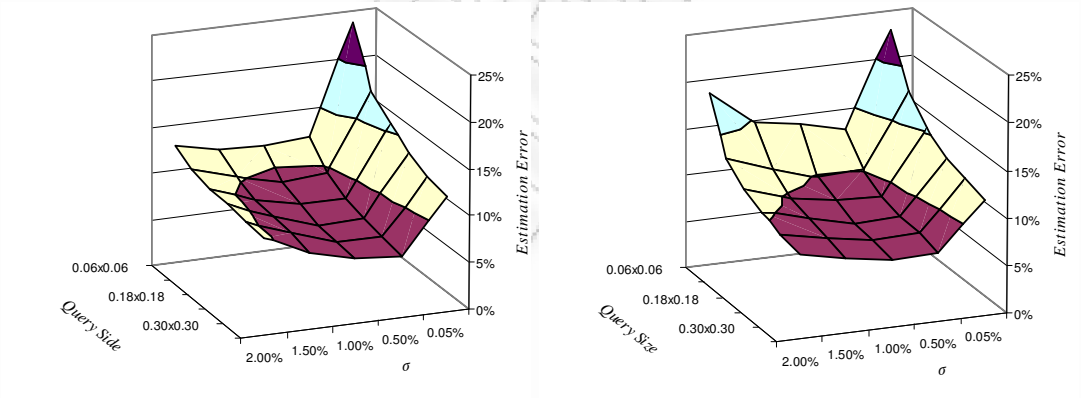
Για να υποστηρίξουμε πραγματικά, αυθαίρετα κατανεμημένα χωρικά δεδομένα μέσω χρήσης ιστογραμμάτων, χρησιμοποιήθηκε το σύνολο δεδομένων *NE* καθώς και τα αντίστοιχα σύνολα $NE_{N\text{-}\sigma}$. Στη συνέχεια, δημιουργήθηκε η διαμέριση *MinSkew* κάθε τροποποιημένου συνόλου δεδομένων

χρησιμοποιώντας ένα ομοιόμορφο πλέγμα αρχικού μεγέθους πλέγματος στα 0.001×0.001, όπως αναφέρθηκε στην [APR99]. Τα πειράματα στα σύνολα δεδομένων $NE_{U\text{-}d}$, ήτοι, με ομοιόμορφη κατανομή αβεβαιότητας, επέδειξαν ανάλογη συμπεριφορά και γι' αυτό παραλείπονται. Πιο συγκεκριμένα για να αξιολογήσουμε την ακρίβεια της ανάλυσης της ενότητας 5.4.2, ήτοι, την εκτίμηση των λανθασμένων αρνητικών και λανθασμένων θετικών χρησιμοποιώντας την Εξ.(5.44), χρησιμοποιήθηκαν τα σύνολα δεδομένων NE και $NE_{N\text{-}\sigma}$ για τα πειράματα, κλιμακώνοντας πρώτα τη σ με σταθερό μέγεθος επερώτησης 0.18×0.18, και κατόπιν κλιμακώνοντας το μέγεθος επερώτησης με σταθερό σ 0.5%.



(a) (b)

Σχήμα 5.15: Μέσες λανθασμένες αρνητικές / θετικές και οι εκτιμήσεις τους κλιμακούμενες με (a) σ και (b) το μέγεθος επερώτησης (πραγματικά δεδομένα – διμεταβλητή κανονική κατανομή αβεβαιότητας).



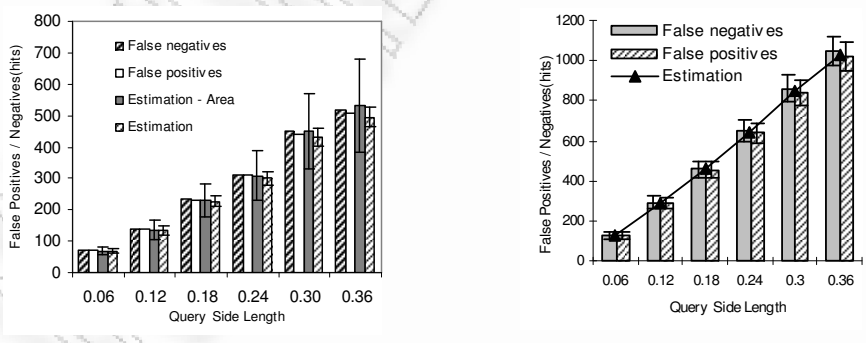
(a) (b)

Σχήμα 5.16: Μέσο σφάλμα εκτίμησης των (a) λανθασμένων θετικών $\overline{ES_P}$ και (b) λανθασμένων αρνητικών $\overline{ES_N}$, σε κάθε επερώτηση, κλιμακούμενες με το σ και το μέγεθος επερώτησης (πραγματικά δεδομένα– διμεταβλήτη κανονική κατανομή αβεβαιότητας).

Το Σχήμα 5.15 παριστάνει τις πραγματικές τιμές και τις τιμές εκτίμησης των λανθασμένων θετικών και λανθασμένων αρνητικών χρησιμοποιώντας τις παραπάνω πειραματικές ρυθμίσεις. Σαφώς, οι εκτιμήσεις είναι ακριβείς με το αναφερθέν σφάλμα πάντα κάτω του 6%. Επιπλέον, το μέσο απόλυτο σφάλμα της εκτίμησης κάθε επερώτησης $\overline{ES_P}$ και $\overline{ES_N}$, που φαίνεται στις γραμμές σφάλματος στο Σχήμα 5.15 και, με μεγαλύτερη λεπτομέρεια, στο Σχήμα 5.16(a) και (b), αντίστοιχα, είναι αρκετά μικρό καθώς είναι κάτω του 12% στην πλειονότητα των πειραματικών ρυθμίσεων. Είναι επίσης σαφές

ότι καθώς αυξάνεται το μέγεθος επερώτησης, οι $\overline{ES_P}$ και $\overline{ES_N}$ μειώνονται σε τιμές κάτω του 11%. Από την άλλη πλευρά, τα μικρά μεγέθη επερώτησης οδηγούν σε αύξηση των τιμών των $\overline{ES_P}$ και $\overline{ES_N}$, μεταξύ του 12% και 24% για μεγέθη επερώτησης 0.06×0.06, και πάλι όμως με μικρότερη κορυφή σφάλματος από αυτές που αναφέρονται για τυχαία δεδομένα χωρίς τη χρήση ιστογραμμάτων, π.χ. το 36% στο Σχήμα 5.14 προς το 24% στο Σχήμα 5.16. Η παραπάνω παρατήρηση εξηγείται από το γεγονός ότι τα ιστογράμματα δίνουν τοπικά πιο ακριβή τιμή του εκτιμώμενου σφάλματος, από το συνολικό τύπο για τα συνθετικά δεδομένα, διότι βοηθούν το μοντέλο να απορροφήσει τις τοπικές αλλαγές πυκνότητας των πραγματικών, αυθαίρετα κατανεμημένων, χωρικών δεδομένων. Εδώ θα πρέπει να επισημάνουμε ότι οι ρυθμίσεις αυτού του συγκεκριμένου πειράματος μας επιτρέπουν την άμεση σύγκριση με αυτό της Ενότητας 5.5, ήτοι και τα δύο χρησιμοποιούν ιστογράμματα για να επιτύχουν καλύτερες εκτιμήσεις· γίνεται συνεπώς σαφέστερο ότι η λειτουργικότητα του προταθέντος αναλυτικού μοντέλου σε σύνολα δεδομένων (τουλάχιστον) μεσαίου πλήθους είναι πολύ καλύτερη από αυτές που λαμβάνονται για σύνολα δεδομένων μικρού πλήθους.

Η επίδραση της ανάλυσης σε πραγματικά σύνολα δεδομένων με τη βοήθεια των ιστογραμμάτων αποδεικνύεται εκτελώντας μία σειρά πειραμάτων στα σύνολα δεδομένων *NE* και $NE_{N\text{-}\sigma}$, υπολογίζοντας το μοντέλο μας με δύο διαφορετικές προσεγγίσεις: (a) παράγοντας την τοπική πυκνότητα μέσω της Εξ.(5.42) και κατόπιν χρησιμοποιώντας την στην Εξ.(5.27), και, (b) χρησιμοποιώντας απευθείας την Εξ.(5.44). Στο πείραμα αυτό τίθεται σ = 0.5% και η πλευρά του παραθύρου επερώτησης κλιμακώνεται από 0.06 έως 0.36. Τα αντίστοιχα αποτελέσματα φαίνονται στο Σχήμα 5.17(a), που αποδεικνύει ότι παρόλο που η προσέγγιση (a), που σημειώνεται ως *Estimation – Area* στο Σχήμα 5.17(a), μας δίνει μια ακριβή μέση εκτίμηση, οι τιμές που λαμβάνονται για τις $\overline{ES_P}$ και $\overline{ES_N}$ είναι υψηλότερες αυτών που παράγονται από την προσέγγιση (b), που σημειώνεται ως *Estimation* στο Σχήμα 5.17(a). Αυτό επιβεβαιώνει ότι η κατάλληλη χρήση των ιστογραμμάτων στο μοντέλο μας είναι σύμφωνα με την ανάλυση στην Ενότητα 5.4.2 με απευθείας χρήση της Εξ.(5.44).



(a) (b)

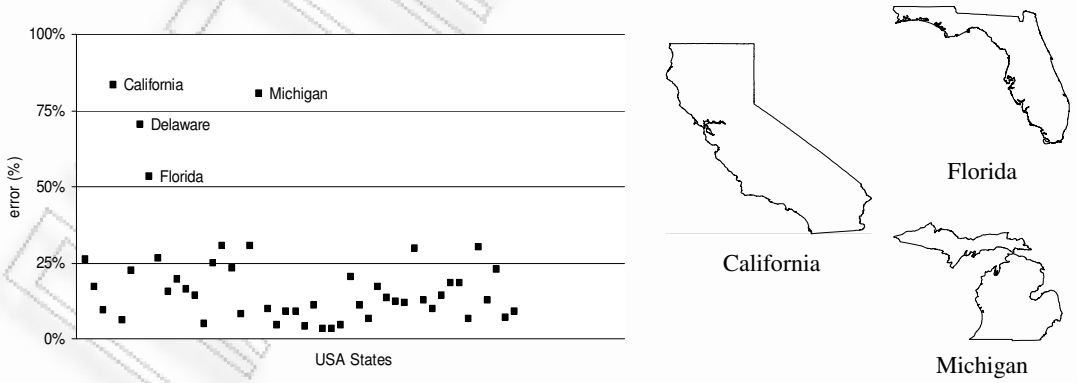
Σχήμα 5.17: (a) Μέσες λανθασμένες αρνητικές / θετικές και οι εκτιμήσεις τους σε κάθε επερώτηση χρησιμοποιώντας διαφορετικές ανεξάρτητες προσεγγίσεις (πραγματικά δεδομένα – διμεταβλητή κανονική κατανομή αβεβαιότητας). (b) Μέσες λανθασμένες αρνητικές / θετικές και οι εκτιμήσεις τους κλιμακούμενες με το μέγεθος επερώτησης (πραγματικά δεδομένα – διμεταβλητή κανονική κατανομή αβεβαιότητας).

5.6.2.4. Πειράματα σε Πραγματικά Δεδομένα Χαλαρώνοντας την Υπόθεση A_{III}

Για να αποδείξουμε τις υψηλής ποιότητας εκτιμήσεις που λαμβάνονται όταν χρησιμοποιείται η μεθοδολογία χωρικών αναλόγων των επαυξημένων ιστογραμμάτων της Ενότητας 5.4.3, εκτελέσθηκε ένα πείραμα χρησιμοποιώντας τα σύνολα δεδομένων *NE* και $NE_{N\text{-}\nu 0.02}$ · όπως αναφέρθηκε ήδη, το $NE_{N\text{-}\nu 0.02}$ περιλαμβάνει δεδομένα με *μεταβλητό* γνωστό μέγεθος της τυπικής απόκλισης *σ* μεταξύ 0 και 0.02. Κατόπιν κλιμακώσαμε την πλευρά του παραθύρου επερώτησης από 0.06 έως 0.36. Τα αντίστοιχα αποτελέσματα, που φαίνονται στο Σχήμα 5.17(b), αποδεικνύουν ότι δεν υπάρχει ουσιαστική διαφορά μεταξύ αυτής της περίπτωσης και αυτής στην οποία η *σ* έχει σταθερή τιμή (Σχήμα 5.15(b)) και οι εκτιμήσεις των $\overline{E_P}$, $\overline{E_N}$ είναι και πάλι πολύ ακριβείς. Επιπλέον, οι τιμές που λαμβάνονται για τις $\overline{ES_P}$ και $\overline{ES_N}$, ήτοι, οι γραμμές σφάλματος, κυμαίνονται μεταξύ 7% και 14%, ενώ στο Σχήμα 5.15(b) το αντίστοιχο σφάλμα κυμαινόταν μεταξύ 6% και 13%. Είναι συνεπώς σαφές ότι η ανάλυση της Ενότητας 5.4.3 για τις ακτίνες μεταβλητής αβεβαιότητας επαληθεύεται ότι είναι τουλάχιστον τόσο ακριβής όσο η αντίστοιχη ανάλυση της Ενότητας 5.4.2, που υποθέτει σταθερή ακτίνα αβεβαιότητας.

5.6.2.5. Πειράματα σε Αποθήκες Πραγματικών Δεδομένων

Για να δείξουμε την εφαρμογή του προτεινόμενου μοντέλου σε μία λειτουργία κύβου δεδομένων, χρησιμοποιήθηκαν τα σύνολα δεδομένων *DCW* και $DCW_{N\text{-}0.5}$ · ο προστιθέμενος θόρυβος Gauss στη θέση του κάθε σημείου έχει *σ* ίση με 0.5% της χωρικής έκτασης κατά μήκος του άξονα *x*-, δεδομένου ότι το μέγεθος του χώρου διαφέρει κατά μήκος των αξόνων *x*- και *y*-. Κατόπιν, εφαρμόζεται ένα ομοιόμορφο πλέγμα 60×30 κατά μήκος των αξόνων *x*- και *y*-, όπως φαίνεται στο Σχήμα 5.11(b), σχηματίζοντας 1800 κάδους που επικαλύπτουν το χάρτη των ΗΠΑ και μετράμε τον αριθμό των αντικειμένων που περιέχονται σε κάθε κελί. Στη συνέχεια εκτελείται μία διαδικασία *roll-up* σε επίπεδο *πολιτείας*, όπως αναφέρθηκε στην Ενότητα 5.4.2. Πιο συγκεκριμένα, η εκτίμηση των λανθασμένων θετικών και λανθασμένων αρνητικών υπολογίστηκε από τα MBBs των ΗΠΑ ως επερωτήσεις εύρους όπως αναφέρθηκε στην Ενότητα 5.4.2. Τέλος, τα αρχικά σύνολα δεδομένων χρησιμοποιήθηκαν για να προσδιορίσουμε τον πραγματικό αριθμό των λανθασμένων θετικών και λανθασμένων αρνητικών.



Σχήμα 5.18: (a) Σφάλμα μεταξύ του πραγματικού αριθμού λανθασμένων αποτελεσμάτων και οι εκτιμήσεις του στην διαδικασία roll-up από το επίπεδο των καλιών στο επίπεδο της πολιτείας στο χάρτη των ΗΠΑ, (b) μία κακή προσέγγιση του πολυγώνου μιας πολιτείας από το MBB της

Το σφάλμα μεταξύ του εκτιμώμενου και του πραγματικού αριθμού λανθασμένων αποτελεσμάτων που λαμβάνεται ως το άθροισμα των λανθασμένων θετικών και λανθασμένων αρνητικών φαίνεται στο Σχήμα 5.18(a). Σαφώς, το σφάλμα στην πλειονότητα των πολιτειών των ΗΠΑ, είναι κάτω του 25% ενώ ο πραγματικός σταθμισμένος μέσος όρος είναι 16%. Για τις τέσσερεις πολιτείες που σημειώνονται στο Σχήμα 5.18(a), το υψηλό σφάλμα που παρουσιάζεται οφείλεται είτε στο μικροσκοπικό μέγεθος του παραθύρου επερώτησης, ήτοι, την περίπτωση του Delaware, επαληθεύοντας το αποτέλεσμα ενός προηγούμενο πειράματος όπου το σφάλμα αυξάνεται όσο μειώνεται το μέγεθος της επερώτησης , ή το ακανόνιστο σχήμα του πραγματικού πολυγώνου επερώτησης που δεν προσεγγίζεται καλά από το MBB του, ήτοι, οι περιπτώσεις της California, της Florida και του Michigan. Τα σχήματά τους φαίνονται στο Σχήμα 5.18(b).

5.6.3. Πειράματα ως προς την Απόδοση

Το τελευταίο πείραμα που εκτελέστηκε για το θέμα περιλαμβάνει την απόδοση των προτεινόμενων λύσεων χρησιμοποιώντας μια εφαρμογή του προτεινόμενου μοντέλου στο PostgreSQL [Post08a] DBMS καθώς και στη χωρική επέκταση PostGIS [Post08b]. Επειδή το επιλεχθέν DBMS δεν υποστηρίζει εκ φύσεως τα χωρικά ιστογράμματα MinSkew [APR99], το επεκτείναμε προς αυτή την κατεύθυνση · επιπλέον, έχουμε συμπεριλάβει στην εφαρμογή μας το χωρικό ανάλογο του επαυξημένου ιστογράμματος που προτείνεται στην Ενότητα 5.4.3. Όλες οι μέθοδοι εφαρμόστηκαν ως συναρτήσεις του χωρικού DBMS σε γλώσσα PL/pgSQL· το αναπτυχθέν λογισμικό φέρεται σε μία βάση δεδομένων πρότυπο.

Πίνακας 5.2: Στατιστικά στοιχεία ιστογράμματος

	Σύνολο δεδομένων	*# Αντικειμένων*	*Μέγεθος πλέγματος*	*# κελιών πλέγματος*	*# κάδων*	*Χρόνος εκτέλεσης κατασκευής (sec)*
Ιστόγραμμα	$NE_{N\text{-}0.01\text{-}1}$	123K	0.001×0.001	920K	1K	21
Επαυξημένο Ιστόγραμμα	$NE_{N\text{-}v0.02}$	123K	0.005×0.005 ×0.0001	7078K	1K	29

Στο πρώτο πείραμα χρησιμοποιήθηκαν τα σύνολα δεδομένων $NE_{N\text{-}0.01\text{-}1}$ και $NE_{N\text{-}v0.02}$ και μετρήθηκε ο χρόνος που απαιτείται για την κατασκευή των απλών ιστογραμμάτων MinSkew και επαυξημένου MinSkew, αντίστοιχα· τα αποτελέσματα φαίνονται στον παραπάνω πίνακα (Πίνακας 5.2). Σαφώς, ο χρόνος επεξεργασίας είναι λογικός δεδομένου του γεγονότος ότι πρόκειται για διαδικασία εκτός σύνδεσης (off-line), που εκτελείται μόνο μία φορά· κατόπιν οι κάδοι ιστογραμμάτων που κατασκευάζονται αποθηκεύονται μονίμως σε ένα σχεσιακό πίνακα. Εδώ θα πρέπει να επισημάνουμε ότι ο αλγόριθμος κατασκευής MinSkew αρχικά τοποθετεί ένα κανονικό πλέγμα πάνω από το σύνολο δεδομένων, που στη συνέχεια χρησιμοποιείται αντί του αρχικού συνόλου δεδομένων, ο χρόνος που απαιτείται για την κατασκευή ενός ιστογράμματος MinSkew δεν εξαρτάται από το μέγεθος του συνόλου δεδομένων· αυτό επιβεβαιώνεται και στην αντίστοιχη πειραματική μελέτη της [APR99]. Συνεπώς, οι χρόνοι εκτέλεσης που φαίνονται στον παραπάνω πίνακα (Πίνακας 5.2) μπορούν να θεωρηθούν αντιπροσωπευτικοί, δεδομένων επίσης και των άλλων παραμέτρων των ιστογραμμάτων, ήτοι, τον αριθμό των κάδων και τον αριθμό των κελιών του πλέγματος που επιτίθεται επί του αρχικού του συνόλου δεδομένων.

Στο δεύτερο πείραμα χρησιμοποιήθηκαν το σύνολο δεδομένων $NE_{N\text{-}0.01\text{-}1}$ και 1000 τυχαία κατανεμημένες ορθογώνιες επερωτήσεις για την αξιολόγηση του μέσου χρόνου εκτέλεσης της συνάρτησης που εφαρμόζει το προτεινόμενο μοντέλο· το μέγεθος επερώτησης κλιμακώθηκε επίσης με τρόπο ανάλογο αυτού της Ενότητας 5.6.2 από 0.06×0.06 έως 0.36×0.36. Τα αντίστοιχα αποτελέσματα απέδειξαν ότι ανεξάρτητα από το μέγεθος επερώτησης, ο χρόνος εκτέλεσης που απαιτείται από το DBMS για την εκτίμηση των λανθασμένων αποτελεσμάτων που εισάγονται σε μία επερώτηση ήταν περίπου 16 ms, ενώ ο χρόνος που απαιτείται για την επεξεργασία της πραγματικής επερώτησης ήταν 120 ms. Σαφώς, η πρόταση της διατριβής μας μπορεί να χρησιμοποιηθεί ως εκτιμητής, διότι ο χρόνος εκτέλεσης περιορίζεται σε μερικά milliseconds, δεδομένου δε ότι η εκτέλεση της πραγματικής επερώτησης συνήθως απαιτεί χρόνο μία τάξη μεγέθους μεγαλύτερο. Επιπλέον, προκύπτει και το αναμενόμενο ότι δηλαδή η επιβάρυνση που εισάγεται από τον εκτιμητή είναι ανεξάρτητη του μεγέθους επερώτησης.

5.7. Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάσαμε ένα θεωρητικό μοντέλο που εκτιμά το σφάλμα που εισάγεται από την αβεβαιότητα θέσης κάθε αντικειμένου στα αποτελέσματα των χωροχρονικών επερωτήσεων χρονικής στιγμής, καθώς επίσης και σε απλές επερωτήσεις εύρους σε σταθερά χωρικά δεδομένα. Δώσαμε ένα κλειστό τύπο για το μέσο αριθμό των λανθασμένων αποτελεσμάτων, που ταξινομούνται ως λανθασμένα θετικά και λανθασμένα αρνητικά, βάσει τριών υποθέσεων: ομοιόμορφη αβεβαιότητα θέσης (ακολουθώντας το μοντέλο που προτάθηκε από την [TWHC04] για να περιγράψουμε την αβέβαιη θέση των τροχιών), ομοιόμορφα κατανεμημένα δεδομένα και σταθερή ακτίνα του δίσκου αβεβαιότητας. Κατόπιν, χαλαρώσαμε αυτές τις υποθέσεις προς πιο ρεαλιστικές ρυθμίσεις, χρησιμοποιώντας τη διμεταβλήτη κανονική κατανομή για την αβεβαιότητα θέσης και τα ιστογράμματα *MinSkew* για κατανομές δεδομένων και ακτίνας.

Η ακρίβεια του προτεινόμενου μοντέλου σε χωροχρονικά δεδομένα τροχιάς, καθώς επίσης και σε σταθερά χωρικά δεδομένα, αξιολογήθηκε μέσω εκτεταμένων πειραμάτων χρησιμοποιώντας διάφορα συνθετικά και πραγματικά χωροχρονικά και χωρικά σύνολα δεδομένων. Το μοντέλο μας παρουσιάζει υψηλή ακρίβεια με ένα μέσο σφάλμα στις $\overline{E_P}$ και $\overline{E_N}$ που δεν υπερβαίνει ποτέ το 6% τόσο για τυχαία συνθετικά ή πραγματικά χωροχρονικά και σταθερά χωρικά δεδομένα · όσον αφορά τα δεδομένα τροχιάς, το μοντέλο παρουσίασε μέτριες τιμές των $\overline{ES_P}$ και $\overline{ES_N}$ της τάξης του 40%. Ωστόσο, θα πρέπει και πάλι να επισημάνουμε ότι για τυπικά μεγέθη επερωτήσεων και αβεβαιότητας οι τύποι του προτεινόμενου μοντέλου παράγουν τιμές λανθασμένων αρνητικών / θετικών μεταξύ $0.0004\times N$ και $0.0025\times N$. Είναι συνεπώς σαφές ότι το πλήθος του συνόλου δεδομένων θα πρέπει να είναι αρκετά μεγάλο για να μας δώσει ένα σημαντικό αριθμό λανθασμένων αποτελεσμάτων και αυτό να επηρεάζει σημαντικά την ποιότητα των αποτελεσμάτων του μοντέλου.

Από την άλλη πάλι, σε περιπτώσεις όπου το πλήθος φτάνει σε κατάλληλες, δηλαδή υψηλές τιμές και την περίπτωση του τυχαίου συνόλου δεδομένων, η εκτίμηση του αριθμού των λανθασμένων αποτελεσμάτων είναι ακριβής ανεξάρτητα από την τιμή του μεγέθους επερώτησης και της ακτίνας d του δίσκου αβεβαιότητας ή της σ στην περίπτωση δεδομένων της κανονικά κατανεμημένης

αβεβαιότητας. Επιπλέον, αποδείχθηκε ότι απλές τροποποιήσεις στη μόνη μελέτη που ομοιάζει στα όσα παρουσιάστηκαν σε αυτή τη διατριβή [YM03], δεν μπόρεσαν να οδηγήσουν σε ακριβή εκτίμηση του μέσου αριθμού λανθασμένων αποτελεσμάτων. Τα πειράματα σε πραγματικά χωρικά δεδομένα καταδεικνύουν ακρίβεια ακόμα υψηλότερη από αυτήν που διαπιστώθηκε για τα συνθετικά δεδομένα, με πολύ χαμηλά σφάλματα $\overline{ES_P}$ και $\overline{ES_N}$, που αποδεικνύουν το πλεονέκτημα που εισάγεται από τη χρήση των ιστογραμμάτων, ακόμα και στην περίπτωση της μεταβλητής σ. Επιπλέον, επαληθεύεται ότι παρουσία των ιστογραμμάτων είναι σαφώς καταλληλότερη η χρήση του μοντέλου που εκφράζεται στις Εξ.(5.43) και Εξ.(5.44) από τη χρήση της τοπικής πυκνότητας που λαμβάνεται ως εκτίμηση από το ιστόγραμμα μέσω παραδοσιακών τρόπων, ήτοι, μέσω της Εξ.(5.41). Τα αποτελέσματα της εφαρμογής του προτεινόμενου μοντέλου σε χωρικούς κύβους δεδομένων και χωρικές λειτουργίες OLAP είναι επίσης πολλά υποσχόμενα. Τέλος, η εφαρμογή των προτεινόμενων λύσεων σε περιβάλλον του πραγματικού κόσμου απέδειξε την αποτελεσματικότητα της πρότασης αυτής όταν χρησιμοποιείται ως εκτιμητής, δεδομένου ότι ο χρόνος εκτέλεσής της είναι τυπικά μόνο ελάχιστα milliseconds.

Η εφαρμογή της πρότασής μας περιλαμβάνει τη βελτιστοποίηση επερωτήσεων στο σενάριο των ανοικτών αγορών [Ioa07], διαδραστικές επερωτήσεις βάσεων δεδομένων, ρυθμίσεις ανακρίβειας και λειτουργίες αποθηκών δεδομένων, όπως συζητήθηκε εκτενώς. Το προταθέν μοντέλο μπορεί να χρησιμοποιηθεί απευθείας σε συστήματα χωρικών βάσεων δεδομένων προκειμένου να δοθεί στους χρήστες η ακρίβεια των αποτελεσμάτων χωρικών επερωτήσεων βάσει μόνο γνωστών στοιχείων των συνόλων δεδομένων και χαρακτηριστικών της επερώτησης, ενώ τα συνηθισμένα ιστογράμματα που χρησιμοποιούνται ήδη σε χωρικές βάσεις δεδομένων για σκοπούς βελτιστοποίησης επερωτήσεων, μπορούν να εξυπηρετήσουν το μοντέλο μας χωρίς να υπάρχει ανάγκη περαιτέρω προσαρμογών.

6. Διαχείριση της Επίδρασης της Συμπίεσης Τροχιών στις Χωροχρονικές Επερωτήσεις

Ο σκοπός αυτού του κεφαλαίου είναι να δώσει μια ανάλυση της επίδρασης της συμπίεσης τροχιών στις χωροχρονικές επερωτήσεις. Το κεφάλαιο διαρθρώνεται ως εξής. Η Ενότητα 6.1 εισάγει βασικές έννοιες συμπίεσης τροχιών. Οι σχετικές εργασίες αναφέρονται στην Ενότητα 6.2. Η Ενότητα 6.3 αποτελεί τον πυρήνα του κεφαλαίου παρουσιάζοντας τη θεωρητική μας ανάλυση. Η Ενότητα 6.4 παρουσιάζει την πειραματική μας μελέτη, ενώ η Ενότητα 6.5 παρουσιάζει τα συμπεράσματα του κεφαλαίου.

6.1. Εισαγωγή

Οι υπάρχουσες εργασίες στο χώρο των MOD, επανειλημμένως επισημαίνουν ότι η πανταχού παρούσες γεωαναφερόμενες συσκευές (location-aware devices) θα αρχίσουν να παράγουν μία χωρίς προηγούμενο ροή δεδομένων θέσεων κινούμενων αντικειμένων συνοδευόμενες από το χρονικό τους αποτύπωμα. Κατά την τελευταία δεκαετία η κοινότητα των βάσεων δεδομένων συνεισφέρει συνεχώς στην ανάπτυξη πρωτότυπων σχημάτων δεικτοδότησης [AG05], [PJT00], [TP01] και προηγμένων τεχνικών επεξεργασίας επερωτήσεων, προκειμένου να αντιμετωπισθεί αυτή η υπερβολική ροή δεδομένων που παράγεται από τέτοιες συσκευές. Ωστόσο, αργά ή γρήγορα, τέτοιοι τεράστιοι όγκοι δεδομένων θα οδηγήσουν σε προκλήσεις αποθήκευσης και υπολογισμών. Εξ' ου και προκύπτει η ανάγκη για τεχνικές συμπίεσης τροχιών.

Οι στόχοι που πρέπει να επιτυγχάνουν οι τεχνικές συμπίεσης τροχιών σύμφωνα με την [MB04] είναι: να υπάρξει μόνιμη μείωση στο μέγεθος των δεδομένων, να υπάρχει μία σειρά δεδομένων που να μας επιτρέπει ακόμα αρκετούς υπολογισμούς σε αποδεκτή (χαμηλή) πολυπλοκότητα και τέλος, να ληφθεί μια σειρά δεδομένων με γνωστά, μικρά περιθώρια σφάλματος, που είναι κατά προτίμηση παραμετρικά προσαρμόσιμα. Κατά συνέπεια, μας ενδιαφέρουν τεχνικές συμπίεσης με ελεγχόμενες απώλειες που εξαφανίζουν κάποιες περιττές ή μη απαραίτητες πληροφορίες υπό καλά καθορισμένα όρια σφάλματος. Ωστόσο, οι υπάρχουσες μελέτες στον τομέα είναι σχετικά περιορισμένες [CWT03], [MB04], [PPS06], [PPS06a], [PPS07] και κατά κύριο λόγο κατευθύνονται από τις εξελίξεις στον τομέα της απλούστευσης γραμμών, της χαρτογραφικής γενίκευσης και της συμπίεσης χρονοσειρών.

Ειδικά στον τομέα των σφαλμάτων που εισάγονται στα παραγόμενα δεδομένα από τέτοιες τεχνικές συμπίεσης, η μόνη σχετική εργασία [MB04] μας δίνει ένα τύπο που εκτιμά το μέσο σφάλμα της τροχιάς που προσεγγίζουμε σε όρους της μέσης απόστασης από την αρχική ροή δεδομένων. Από

την άλλη πλευρά, σε αυτή τη διατριβή, θεωρούμε ότι αντί να παρέχουμε στο χρήστη μιας MOD με το μέσο σφάλμα στη θέση του κάθε (συμπιεσμένου) αντικειμένου σε κάθε χρονικό αποτύπωμα (που μπορεί επίσης να θεωρηθεί ως (αν)ακρίβεια δεδομένων), ο / η χρήστης θα προτιμούσε να ενημερώνεται για το μέσο σφάλμα που εισάγεται στα αποτελέσματα επερωτήσεων που εκτελούνται σε συμπιεσμένα δεδομένα. Συνεπώς, η πρόκληση που αποδεχόμαστε είναι να παρουσιάσουμε ένα θεωρητικό μοντέλο που εκτιμά το σφάλμα λόγω συμπίεσης στα αποτελέσματα των χωροχρονικών επερωτήσεων. Εξ' όσων γνωρίζουμε, αυτό είναι το πρώτο αναλυτικό μοντέλο της επίδρασης της συμπίεσης σε αποτελέσματα επερωτήσεων σε βάσεις δεδομένων τροχιών.

Παρακάτω συνοψίζονται τα βασικά στοιχεία των κύριων θεμάτων που παρουσιάζονται στο παρόν κεφάλαιο:

- Περιγράφουμε δύο τύπους σφαλμάτων (πιο συγκεκριμένα, λανθασμένα αρνητικά και λανθασμένα θετικά) όταν εκτελούμε επερωτήσεις χρονικής στιγμής σε συμπιεσμένες τροχιές και αποδεικνύουμε ένα λήμμα που εκτιμά το μέσο αριθμό των ανωτέρω τύπων σφαλμάτων. Αποδεικνύεται ότι ο μέσος αριθμός των λανθασμένων αποτελεσμάτων και των δύο τύπων σφαλμάτων εξαρτάται από τη *Σύγχρονη Ευκλείδεια Απόσταση* [CWT03], [MB04], [PPS06] κατά μήκος των αξόνων x- και y- μεταξύ της αρχικής και της συμπιεσμένης τροχιάς και της περιμέτρου (αντί της επιφάνειας) του παραθύρου επερώτησης.
- Αποδεικνύουμε πως το κόστος αξιολόγησης του τύπου που αναπτύσσουμε μπορεί να μειωθεί σε μία ελάχιστη επιπλέον επιβάρυνση επί του αλγορίθμου συμπίεσης που χρησιμοποιείται, ενώ συζητάμε πως το αναπτυχθέν αναλυτικό μοντέλο μας βοηθά να δώσουμε αποτελεσματικούς αλγορίθμους συμπίεσης.
- Τέλος, διεξάγουμε ένα σύνολο πειραμάτων σε συνθετικά και πραγματικά σύνολα δεδομένων τροχιάς που αποδεικνύουν ότι η ανάλυση εφαρμόζεται εύκολα, είναι ορθή και ακριβής.

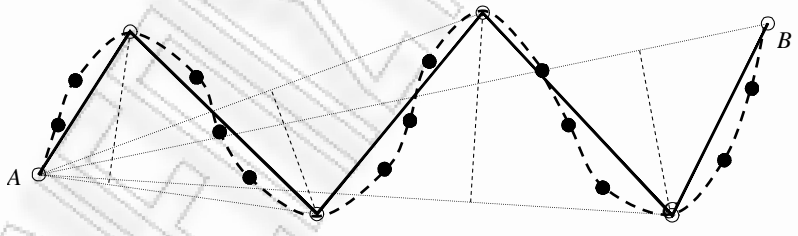
Το μοντέλο που περιγράφεται σε αυτό το κεφάλαιο μπορεί να χρησιμοποιηθεί σε MODs για να εκτιμήσουμε το μέσο αριθμό λανθασμένων απαντήσεων σε αποτελέσματα επερωτήσεων όταν συμπιέζονται τα δεδομένα των τροχιών. Επί παραδείγματι, θα μπορούσε να χρησιμοποιηθεί αμέσως μετά τη συμπίεση ενός συνόλου δεδομένων τροχιών για να δώσει στο χρήστη το μέσο σφάλμα που εισάγεται στα αποτελέσματα των χωροχρονικών επερωτήσεων διαφόρων μεγεθών· συνεπώς θα μπορούσαμε να το εκμεταλλευτούμε ως επιπλέον κριτήριο για το χρήστη προκειμένου να αποφασίσει κατά πόσο τα συμπιεσμένα δεδομένα είναι κατάλληλα για τις ανάγκες του / της και ει δυνατόν ν' αποφασίσει και για διαφορετικούς βαθμούς συμπίεσης κ.ο.κ. Επιπλέον, θα μπορούσε να χρησιμοποιηθεί για τη βελτίωση της αποτελεσματικότητας των προτεινόμενων αλγορίθμων συμπίεσης τροχιάς· δεδομένου ότι ένα μοντέλο αυτής της μορφής θα μας αποκάλυπτε τα πραγματικά μεγέθη από τα οποία εξαρτάται το σφάλμα, θα μπορούσε επίσης να παράσχει την διαίσθηση προς τη υλοποίηση πιο προηγμένων /αποτελεσματικών λύσεων.

6.2. Σχετικές Εργασίες

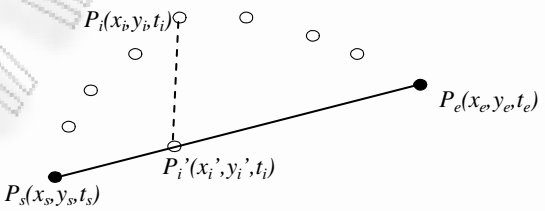
Στην ενότητα αυτή ασχολούμαστε πρωτίστως με τις τεχνικές που εισάγονται για τη συμπίεση τροχιών κατά τα τελευταία χρόνια, ενώ, στη συνέχεια εξετάζουμε τις σχετικές εργασίες στον τομέα της εκτίμησης και αντιμετώπισης του σφάλματος που εισάγεται από τις εν λόγω τεχνικές συμπίεσης.

6.2.1. Συμπίεση Τροχιών

Όπως αναφέρθηκε ήδη, οι υπάρχουσες μελέτες στο χώρο της συμπίεσης τροχιών κατευθύνονται κυρίως από τις σχετικές εργασίες στον τομέα της απλούστευσης γραμμών και τη συμπίεση χρονοσειρών. Οι Meratnia και By [MB04] εκμεταλλεύονται τους υπάρχοντες αλγορίθμους που χρησιμοποιούνται στον τομέα της γενίκευσης γραμμών, παρουσιάζοντας έναν αλγόριθμο από πάνω προς τα κάτω (top-down) και έναν αλγόριθμο ανοιγόμενου παραθύρου (opening window), που μπορούν να εφαρμοσθούν απευθείας σε χωροχρονικές τροχιές. Ο αλγόριθμος *top-down*, καλείται TD-TR, βασίζεται στον πολύ γνωστό αλγόριθμο των Douglas-Peucker [DP73] (Σχήμα 6.1) που εισήχθη από γεωγράφους στο χώρο της χαρτογραφίας. Ο αλγόριθμος αυτός υπολογίζει την κατακόρυφη απόσταση κάθε εσωτερικού σημείου από τη γραμμή που συνδέει το πρώτο και το τελευταίο σημείο της πολυγραμμής (γραμμή *AB* στο Σχήμα 6.1) και εντοπίζει το σημείο με τη μεγαλύτερη κατακόρυφη απόσταση (σημείο *C*). Κατόπιν, δημιουργεί τις γραμμές *AC* και *CB* και, αναδρομικά, ελέγχει τις νέες αυτές γραμμές σε σχέση με τα υπόλοιπα σημεία με την ίδια μέθοδο κ.ο.κ. Όταν η απόσταση όλων των υπόλοιπων σημείων από την γραμμή που μόλις εξετάσαμε είναι μικρότερη από ένα δεδομένο κατώφλι (π.χ. όλα τα σημεία που ακολουθούν τη *C* στη γραμμή *BC* στο Σχήμα 6.1) ο αλγόριθμος σταματά και μας επιστρέφει αυτό το γραμμικό τμήμα ως τμήμα της νέας - συμπιεσμένης - πολυγραμμής. Αναγνωρίζοντας το γεγονός ότι οι τροχιές είναι πολυγραμμές που εξελίσσονται χρονικά, ο αλγόριθμος που παρουσιάστηκε στην [MB04] αντικαθιστά την κάθετη απόσταση που χρησιμοποιείται στον αλγόριθμο DP με την επονομαζόμενη *Σύγχρονη Ευκλείδεια Απόσταση* (*Synchronous Euclidean Distance - SED*), που αναφέρεται επίσης στις [CWT03], [PPS06] και είναι η απόσταση μεταξύ του σημείου που εξετάζεται επί του παρόντος (P_i στο Σχήμα 6.2) και του σημείο της γραμμής (P_s, P_e) όπου θα βρισκόταν το κινούμενο αντικείμενο, αν υποθέσουμε ότι θα κινείται σε αυτή τη γραμμή, στη χρονική στιγμή t_i που καθορίζεται από το υπό εξέταση σημείο (P_i' στο Σχήμα 6.2).



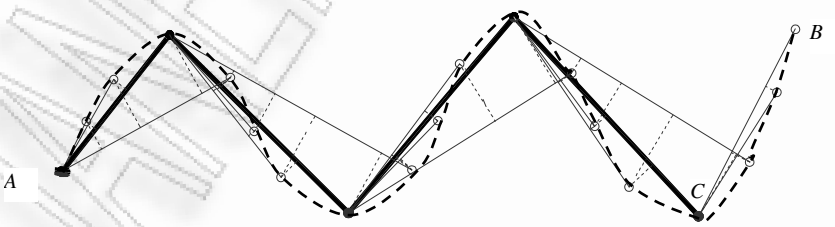
Σχήμα 6.1: Αλγόριθμος Top-down Douglas-Peucker που χρησιμοποιείται για τη Συμπίεση Τροχιών. Τα αρχικά σημεία δεδομένων παριστάνονται με κλειστούς κύκλους [MB04]



Σχήμα 6.2: Η σύγχρονη Ευκλείδεια Απόσταση (SED): Η απόσταση υπολογίζεται μεταξύ του υπό εξέταση σημείου (P_i) και του σημείου P_i' που καθορίζεται ως το σημείο στη γραμμή (P_s, P_e) τη χρονική στιγμή t_i [MB04]

Η χρονική πολυπλοκότητα του αρχικού αλγορίθμου Douglas-Peucker (στον οποίο βασίζεται ο αλγόριθμος TD-TR) είναι $O(N^2)$, με N τον αριθμό των αρχικών σημείων της τροχιάς, ενώ μπορεί να μειωθεί και σε $O(N\log N)$ εφαρμόζοντας την πρόταση που παρουσιάστηκε στην [HS92]. Παρόλο που η πειραματική μελέτη που παρουσιάστηκε στην [MB04] αποδεικνύει ότι ο αλγόριθμος TD-TR είναι σημαντικά καλύτερος από αυτόν του ανοιγόμενου παραθύρου (που παρουσιάζεται παρακάτω στην παρούσα ενότητα) από άποψη ποιότητας και συμπίεσης (διότι βελτιστοποιεί τη διαδικασία συμπίεσης συνολικά), ο αλγόριθμος TD-TR μειονεκτεί διότι δεν είναι αλγόριθμος πραγματικού χρόνου (on-line) και συνεπώς, δεν εφαρμόζεται σε νεοφερμένα τμήματα τροχιών τη στιγμή που τροφοδοτούν μία MOD. Αντιθέτως, απαιτεί εκ των προτέρων γνώση όλης της τροχιάς κινούμενου αντικειμένου.

Από την άλλη πλευρά, βάσει των συνθηκών που περιγράψαμε προηγουμένως για την λειτουργία πραγματικού χρόνου (on-line), η κατηγορία των αλγορίθμων *ανοιγόμενου παραθύρου* (*opening window* - OW) μπορεί να εφαρμοσθεί εύκολα. Οι αλγόριθμοι αυτοί ξεκινούν αγκυρώνοντας το πρώτο σημείο της τροχιάς και προσπαθούν να προσεγγίσουν τα ακόλουθα σημεία της ροής δεδομένων με ένα σταδιακά μεγαλύτερο τμήμα (Σχήμα 6.3). Αν όλες οι αποστάσεις των ακόλουθων σημείων της τροχίας από το τμήμα είναι κάτω του κατωφλίου που τίθεται, γίνεται μία απόπειρα να μετακινηθεί το τελικό σημείο του τμήματος μία θέση προς τα πάνω στη σειρά δεδομένων. Όταν υπάρχει τάση υπέρβασης του κατωφλίου, μπορούμε να εφαρμόσουμε δύο στρατηγικές: γίνεται τελικό σημείο, και ταυτόχρονα η άγκυρα του επόμενου τμήματος, είτε το σημείο που προκαλεί την παραβίαση (*Normal Opening Window*, *NOPW*) είτε το σημείο πριν από αυτή (*Before Opening Window*, *BOPW*). Αν δεν υπερβούμε το κατώφλι, μετακινούμαστε κατά μία θέση προς τα πάνω στη σειρά δεδομένων (ήτοι το παράθυρο ανοίγει κι άλλο) και ο αλγόριθμος συνεχίζει μέχρι να βρει και το τελευταίο σημείο της τροχιάς· τότε όλη η τροχιά μετατρέπεται σε μία γραμμική προσέγγιση. Ενώ στην αρχική κατηγορία αλγορίθμων OW κάθε απόσταση υπολογίζεται από το σημείο κάθετα προς το υπό εξέταση τμήμα, στον αλγόριθμο OPW-TR που παρουσιάζεται στην [MB04] αξιολογείται η *SED*. Παρόλο που οι αλγόριθμοι OW είναι ακριβοί υπολογιστικά – δεδομένου ότι η χρονική τους πολυπλοκότητα είναι $O(N^2)$ – είναι πολύ δημοφιλείς. Κι αυτό γιατί είναι αλγόριθμοι πραγματικού χρόνου και μπορούν να λειτουργήσουν σχετικά καλά παρουσία θορύβου (παρόλο που λειτουργούν μόνο με σχετικά σύντομες σειρές δεδομένων).



Σχήμα 6.3: Αλγόριθμος Ανοιγόμενου Παραθύρου που χρησιμοποιείται για Συμπίεση τροχιών. Τα αρχικά σημεία δεδομένων παριστάνονται με κλειστούς κύκλους [MB04]

Προσφάτως, οι Potamias et al. [PPS06] πρότειναν αρκετές τεχνικές που βασίζονται σε ομοιόμορφη και χωροχρονική δειγματοληψία για να συμπιέσουν ροές δεδομένων τροχιών, υπό διαφορετικές ρυθμίσεις διαθεσιμότητας μνήμης: σταθερή μνήμη, λογαριθμικά ή γραμμικά αυξανόμενη μνήμη, ή μνήμη που δεν είναι γνωστή εκ των προτέρων. Οι κύριες συνεισφορές είναι δύο αλγόριθμοι

συμπίεσης, πιο συγκεκριμένα, οι *STTrace* και *Thresholds*. Ο αλγόριθμος *STTrace*, χρησιμοποιεί ένα σταθερό ποσοστό μνήμης *M* για κάθε τροχιά. Ξεκινά εισάγοντας στην ανατεθειμένη μνήμη τις πρώτες *M* καταγεγραμμένες θέσεις, μαζί με τη *SED* κάθε θέσης σε σχέση με τον προκάτοχο και το διάδοχο στο δείγμα. Μόλις εξαντληθεί η ανατεθειμένη μνήμη και εξετάζεται ένα νέο σημείο για πιθανή εισαγωγή, γίνεται αναζήτηση στο δείγμα για το αντικείμενο με τη χαμηλότερη *SED*, που αντιπροσωπεύει τη λιγότερη δυνατή απώλεια πληροφοριών σε περίπτωση απόρριψης. Στη συνέχεια, ο αλγόριθμος ελέγχει κατά πόσο το εισαγόμενο σημείο έχει *SED* μεγαλύτερη από την ελάχιστη που έχει ήδη βρεθεί στο δείγμα και αν όντως ισχύει αυτό, το σημείο που επεξεργαζόμαστε επί του παρόντος εισάγεται στο δείγμα εις βάρος του σημείου με τη χαμηλότερη *SED*. Τέλος, οι *SED* των γειτονικών σημείων αυτού που αφαιρέθηκε υπολογίζονται εκ νέου, ενώ ξεκινά μία αναζήτηση στο δείγμα για τη νέα ελάχιστη *SED*. Ο προτεινόμενος αλγόριθμος μπορεί να εφαρμοσθεί εύκολα στην περίπτωση πολλαπλών τροχιών, απλώς υπολογίζοντας μία συνολική ελάχιστη *SED* όλων των τροχιών εντός της διατιθέμενης μνήμης.

Από την προηγούμενη συζήτηση προκύπτει ότι η πλειονότητα των προτεινόμενων αλγορίθμων συμπίεσης τροχιών βασίζουν την απόφασή τους στο κατά πόσο θα πρέπει να κρατήσουν ή να απορρίψουν ένα σημείο της αρχικής τροχιάς στην τιμή της *SED* μεταξύ της αρχικής και της συμπιεσμένης τροχιάς σε αυτό το συγκεκριμένο χρονικό αποτύπωμα. Κατά συνέπεια, μία μέθοδος για τον υπολογισμό του αποτελέσματος της συμπίεσης σε χωροχρονικές επερωτήσεις βάσει της τιμής της *SED* κατά μήκος των σημείων της αρχικής τροχιάς, δεν θα αποτελεί υπερβολική επιβάρυνση στον αλγόριθμο συμπίεσης, διότι θα απαιτεί μόνο την εκτέλεση μερικών επιπλέον πράξεων εντός του ίδιου αλγορίθμου.

6.2.2. Εκτίμηση Σφάλματος

Εξ' όσων γνωρίζουμε δεν υπάρχει θεωρητική μελέτη για τη μοντελοποίηση του σφάλματος που εισάγεται στα αποτελέσματα χωροχρονικών επερωτήσεων λόγω της συμπίεσης των τροχιών· η εργασία μας είναι η πρώτη πάνω σε αυτό το θέμα που καλύπτει την περίπτωση των χωροχρονικών επερωτήσεων χρονικής στιγμής. Ωστόσο, υπάρχουν δύο σχετικά θέματα: Το πρώτο είναι ο καθορισμός του σφάλματος που εισάγεται απευθείας σε κάθε τροχιά από τη συμπίεση [MB04], που είναι η μέση τιμή της *SED* μεταξύ μιας τροχιάς *p* και της προσέγγισής της *q* (χρησιμοποιείται και ο όρος σύγχρονο σφάλμα (synchronous error $E(q, p)$)). Η [MB04] παρέχει μία μέθοδο για τον υπολογισμό της μέσης τιμής ως συνάρτηση της απόστασης μεταξύ του *p* και του *q*. Το αποτέλεσμα αυτής της ανάλυσης μας οδηγεί σε έναν δαπανηρό τύπο, που παρέχει το μέσο σφάλμα (δηλαδή τη μέση απόσταση μεταξύ του *p* και του *q* κατά τη διάρκεια ζωής τους)· ωστόσο, δεν υπάρχει προφανής τρόπος για το πώς θα το χρησιμοποιήσουμε ώστε να καθορίσουμε το σφάλμα που εισάγεται στα αποτελέσματα της επερώτησης.

Το δεύτερο σχετικό ζήτημα είναι τα όσα γίνονται στα πλαίσια της διαχείρισης αβεβαιότητας τροχιάς, όπως στο [CKP04], [PJ99], [Tra03], [TWHC04]. Αυτό οφείλεται στο γεγονός ότι το σφάλμα που εισάγεται από τη συμπίεση μπορεί να θεωρηθεί επίσης αβεβαιότητα και έτσι οι σχετικές τεχνικές μπορούν να εφαρμοσθούν στο σύνολο δεδομένων που προκύπτει· έτσι τα όσα παρουσιάζονται στο προηγούμενο κεφάλαιο θα μπορούσαν να χρησιμοποιηθούν για το στόχο μας. Ωστόσο, αυτή η μεθοδολογία δεν μπορεί να χρησιμοποιηθεί άμεσα παρουσία των συμπιεσμένων δεδομένων τροχιών,

διότι ο καθορισμός της στατιστικής κατανομής της θέσης της συμπιεσμένης τροχιάς χρησιμοποιώντας πληροφορίες από την αρχική, είναι από μόνος του πολύ περίπλοκος.

Από την άλλη πλευρά, η προσέγγισή μας βασίζεται στο γεγονός ότι ο αλγόριθμος συμπίεσης εκμεταλλεύεται τη *SED* σε κάθε αρχικό σημείο της τροχιάς και έτσι εισάγει μία πολύ μικρή επιβάρυνση στον αλγόριθμο συμπίεσης.

Πίνακας 6.1: Πίνακας συμβόλων

Σύμβολο	*Περιγραφή*
$S, T^{\dagger}, T$	Ο μοναδιαίος χώρος, ένα σύνολο δεδομένων τροχιάς και το συμπιεσμένο αντίστοιχο του.
$T_i^{\dagger}, T_i$	μια αρχική τροχιά και η συμπιεσμένη αντίστοιχή της.
$R, R_{a\times b}, W_j$	το σύνολο όλων των επερωτήσεων χρονικής στιγμής στον S, το υποσύνολό του με πλευρές μήκους a και b κατά μήκος των αξόνων x- και y- και ένα παράθυρο επερώτησης χρονικής στιγμής.
n, m_i	το πλήθος του συνόλου δεδομένων T και ο αριθμός των δειγματοληπτούμενων σημείων εντός της τροχιάς $T_i^{\dagger}$.
$SED_i(t)$, $\delta x_i(t), \delta y_i(t)$	η συνάρτηση της Σύγχρονης Ευκλείδειας Απόστασης (SED) μεταξύ της τροχιάς $T_i^{\dagger}$ και της συμπιεσμένης αντίστοιχής της T_i και η προβολή της κατά μήκος των αξόνων x- και y-.
$t_{i,k}$, $SED_{i,k}$, $\delta x_{i,k}$, $\delta y_{i,k}$	το k-οστο χρονικό αποτύπωμα στο οποίο η τροχιά $T_i^{\dagger}$ έλαβε δείγμα της θέσης της, η Σύγχρονη Ευκλείδεια Απόσταση της από τη συμπιεσμένη της αντίστοιχη T_i στο ίδιο χρονικό αποτύπωμα και η προβολή της κατά μήκος των αξόνων x- και y-
$A_{i,j}$	η επιφάνεια στην οποία πρέπει να βρεθεί η κάτω αριστερά γωνία του W_j στο χρονικό αποτύπωμα t_j προκειμένου να ανακτήσει την τροχιά T_i ως λανθασμένη αρνητική (ή ως λανθασμένη θετική).
$AvgP_{i,N}(R_{a\times b})$, $AvgP_{i,P}(R_{a\times b})$	η μέση πιθανότητα όλες οι επερωτήσεις χρονικής στιγμής $W_j \in R_{a\times b}$, να ανακτήσουν την T_i ως λανθασμένη αρνητική (ή ως λανθασμένη θετική).
$E_N(R_{a\times b}), E_P(R_{a\times b})$	ο μέσος αριθμός λανθασμένων αρνητικών (ή λανθασμένων θετικών) στα αποτελέσματα μιας επερώτησης $W_i \in R_{a\times b}$.

6.3. Ανάλυση

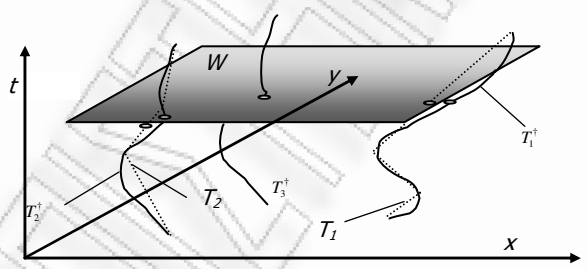
Ο πυρήνας της ανάλυσής μας είναι ένα λήμμα που παρέχει τον τύπο που χρησιμοποιείται για την εκτίμηση του μέσου όρου των λανθασμένων αποτελεσμάτων ανά επερώτηση που εκτελείται σε ένα σύνολο δεδομένων συμπιεσμένης τροχιάς. Όπως και στο προηγούμενο κεφάλαιο, και εδώ εστιάζουμε σε επερωτήσεις χρονικής στιγμής, που μπορούν να χρησιμοποιηθούν για την ανάκτηση των θέσεων κινούμενων αντικειμένων σε ένα δεδομένο χρονικό σημείο στο παρελθόν και μπορούν να θεωρηθούν ως ειδική περίπτωση των χωροχρονικών επερωτήσεων εύρους, με χρονική έκταση ίση με το μηδέν. Αυτός ο τύπος επερώτησης μπορεί να θεωρηθεί ως ο συνδυασμός μίας χωρικής (ήτοι παραθύρου επερώτησης W) και μίας χρονικής (ήτοι χρονικό αποτύπωμα t) συνιστώσας. Όπως θα συζητηθεί στο Κεφάλαιο 7, η επέκταση του μοντέλου μας για να υποστηρίξει τις επερωτήσεις εύρους με μη μηδενική χρονική έκταση δεν είναι σε καμία περίπτωση απλή, γι' αυτό και αφήνεται ως ανοιχτό θέμα.

Είναι σημαντικό να αναφέρουμε ότι το μοντέλο μας υποστηρίζει αυθαίρετα κατανεμημένα δεδομένα τροχιών χωρίς να ασχολείται με τα χαρακτηριστικά τους (π.χ. ρυθμός δειγματοληψίας, ταχύτητα, κατεύθυνση, ευελιξία). Συνεπώς, μπορεί να χρησιμοποιηθεί απευθείας στις MODs χωρίς

περαιτέρω τροποποιήσεις. Η μόνη υπόθεση που κάνουμε είναι ότι τα παράθυρα επερώτησης χρονικής στιγμής είναι ομοιόμορφα κατανεμημένα εντός του χώρου δεδομένων. Για να αποδεσμευτούμε από αυτή την υπόθεση, θα έπρεπε να μοντελοποιήσουμε μαθηματικώς την κατανομή των επερωτήσεω χρησιμοποιώντας και να τροποποιήσουμε την ακόλουθη ανάλυση, αναλόγως. Ο Πίνακας 6.1 συνοψίζει τα σύμβολα που χρησιμοποιούνται στο υπόλοιπο του κεφαλαίου.

Έστω ο μοναδιαίος 3D (ήτοι 2D χωρικός και 1D χρονικός) χώρος S που περιλαμβάνει ένα σύνολο $T^{\dagger}$, n τροχιών $T_i^{\dagger}$ και ένα σύνολο T με τις συμπιεσμένες αντίστοιχές τους T_i. Έστω επίσης R το σύνολο όλων των επερωτήσεων χρονικής στιγμής που τίθεται στα σύνολα δεδομένων $T^{\dagger}$ και T, και $R_{a\times b}$ το υποσύνολο του R που περιέχει όλες τις επερωτήσεις χρονικής στιγμής που έχουν πλευρές μήκους a και b κατά μήκος των αξόνων x- και y- αντίστοιχα. Εισάγονται δύο τύποι σφαλμάτων όταν εκτελείται μία επερώτηση χρονικής στιγμής $W_j \in R$ σε ένα σύνολο δεδομένων με τις ρυθμίσεις που περιγράψαμε προηγουμένως:

- *λανθασμένες αρνητικές* είναι οι τροχιές που αρχικώς ήταν επιλέξιμες για την επερώτηση αλλά οι συμπιεσμένες αντίστοιχές τους δεν ανακτήθηκαν· επισήμως το σύνολο των λανθασμένων αρνητικών $T_N \subseteq T$ ορίζεται ως $T_N = \left\{T_i \in T : T_i \notin W_j \mid T_i^{\dagger} \in W_j\right\}$·
- *λανθασμένες θετικές* είναι οι συμπιεσμένες τροχιές που ανακτώνται από την επερώτηση ενώ οι αρχικές αντίστοιχές τους δεν είναι επιλέξιμες· επισήμως, το σύνολο των λανθασμένων θετικών $T_P \subseteq T$ ορίζεται ως $T_P = \left\{T_i \in T : T_i \in W_j \mid T_i^{\dagger} \notin W_j\right\}$.



Σχήμα 6.4: Καθορισμός προβλήματος

Έστω για παράδειγμα το Σχήμα 6.4 που απεικονίζει ένα σύνολο n μη συμπιεσμένων τροχιών $T_i^{\dagger}$, καθώς και τις συμπιεσμένες αντίστοιχές τους T_i. Κάθε μη συμπιεσμένη τροχιά $T_i^{\dagger}$ αποτελείται από ένα σύνολο m_i σημείων που συνοδεύονται από το χρονικό τους αποτύπωμα, μεταξύ των οποίων εφαρμόζουμε γραμμική παρεμβολή. Το Σχήμα 6.4 απεικονίζει επίσης μία επερώτηση χρονικής στιγμής W· παρόλο που η W ανακτά τη συμπιεσμένη τροχιά T_1, η αρχική της αντίστοιχη $T_1^{\dagger}$ δεν τέμνει το παράθυρο επερώτησης, αποτελώντας μια λανθασμένη θετική. Αντιστρόφως, παρόλο που η αρχική τροχιά $T_2^{\dagger}$ τέμνει το W, η συμπιεσμένη αντίστοιχή της T_2 δεν είναι παρούσα στα αποτελέσματα της επερώτησης, αποτελώντας μια λανθασμένη αρνητική. Έχοντας περιγράψει το πλαίσιο της εργασίας μας, διατυπώνουμε το εξής λήμμα:

Λήμμα 6.1*: Ο μέσος αριθμός λανθασμένων αρνητικών $E_N(R_{a\times b})$ και λανθασμένων θετικών $E_P(R_{a\times b})$ στα αποτελέσματα μιας επερώτησης χρονικής στιγμής $W_j \in R_{a\times b}$ με πλευρές μήκους a και b κατά μήκος των*

αξόνων x-και y-, αντίστοιχα, σε ένα σύνολο δεδομένων συμπιεσμένων τροχιών δίνεται από τον ακόλουθο τύπο:

$$E_N\left(R_{a\times b}\right) = E_P\left(R_{a\times b}\right) = \sum_{i=1}^{n}\sum_{k=1}^{m_i-1}\frac{\left(t_{i,k+1}-t_{i,k}\right)}{(1+a)\cdot(1+b)}\cdot\left(\frac{b\left(\left|\delta x_{i,k}\right|+\left|\delta x_{i,k+1}\right|\right)}{2}+\frac{a\left(\left|\delta y_{i,k}\right|+\left|\delta y_{i,k+1}\right|\right)}{2}-\frac{e}{6}\right) \quad (6.1)$$

όπου $e = 2\left|\delta x_{i,k}\delta y_{i,k}\right| + 2\left|\delta x_{i,k+1}\delta y_{i,k+1}\right| + \left|\delta x_{i,k}\delta y_{i,k+1}\right| + \left|\delta x_{i,k+1}\delta y_{i,k}\right|$.

Η Εξ.(6.1) διατυπώνει το γεγονός ότι το μέσο σφάλμα στα αποτελέσματα μίας επερώτησης χρονικής στιγμής σε συμπιεσμένα δεδομένα τροχιών σχετίζεται άμεσα με τη σταθμισμένη μέση *SED* κατά μήκος των αξόνων *x*- και *y*- (δηλαδή, $\left(t_{i,k+1}-t_{i,k}\right)$ επί $\left|\delta x_{i,k}\right|+\left|\delta x_{i,k+1}\right|$ ή $\left|\delta y_{i,k}\right|+\left|\delta y_{i,k+1}\right|$) επί την αντίστοιχη αντίθετη διάσταση επερώτησης (δηλαδή, $b\left(\left|\delta x_{i,k}\right|+\left|\delta x_{i,k+1}\right|\right)$ και $a\left(\left|\delta y_{i,k}\right|+\left|\delta y_{i,k+1}\right|\right)$), ενώ το *e* είναι ένα άθροισμα ήσσονος σημασίας, δεδομένου ότι είναι το άθροισμα των προϊόντων μεταξύ $\left|\delta x_{i,k}\right|,\left|\delta x_{i,k+1}\right|,\left|\delta y_{i,k}\right|,\left|\delta y_{i,k+1}\right|$.

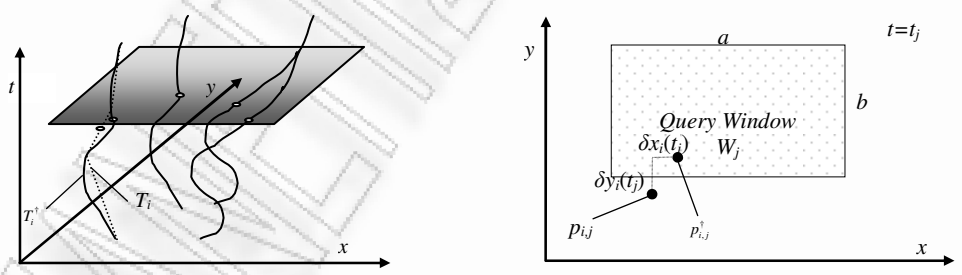
6.3.1. Απόδειξη του Λήμματος 6.1

Ο μέσος αριθμός $E_N(R_{a\times b})$ των τροχιών που είναι λανθασμένες αρνητικές στα αποτελέσματα μιας επερώτησης χρονικής στιγμής $W_j \in R_{a\times b}$, μπορεί να ληφθεί αθροίζοντας τις πιθανότητες $P\left(T_i \notin W_j \mid T_i^{\dagger} \in W_j\right)$ να είναι λανθασμένες αρνητικές όλες οι τροχιές του συνόλου δεδομένων T_i (i=1,…,n) σε σχέση με ένα αυθαίρετο παράθυρο επερώτησης χρονικής στιγμής $W_j \in R_{a\times b}$:

$$E_N\left(R_{a\times b}\right) = \sum_{i=1}^{n} AvgP_{i,N}\left(R_{a\times b}\right) \quad (6.2)$$

Ομοίως, ο μέσος αριθμός $E_P(R_{a\times b})$ *των* τροχιών που είναι λανθασμένες θετικές μπορεί να υπολογιστεί από τον ακόλουθο τύπο:

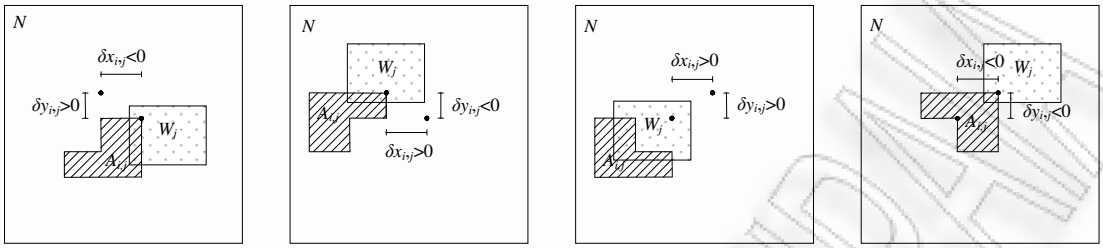
$$E_P\left(R_{a\times b}\right) = \sum_{i=1}^{n} AvgP_{i,P}\left(R_{a\times b}\right) \quad (6.3)$$



Σχήμα 6.5: Η τομή μιας τροχιάς $T_i^{\dagger}$ και της συμπιεσμένης της τροχιάς T_i, με το επίπεδο της επερώτησης χρονικής στιγμής στο χρονικό αποτύπωμα t_j.

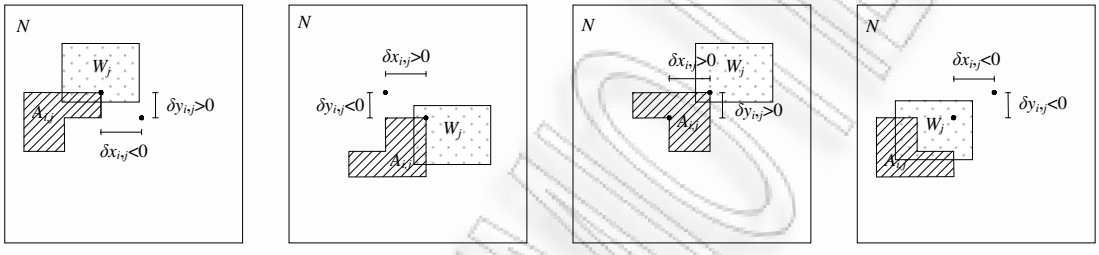
Εξ' ου και ο στόχος μας είναι να καθορίσουμε τη $AvgP_{i,N}(R_{a\times b})$ και τη $AvgP_{i,P}(R_{a\times b})$. Για το σκοπό αυτό, διατυπώνουμε την πιθανότητα μία τυχαία τροχιά να είναι λανθασμένη αρνητική (ή λανθασμένη θετική), σε σχέση με ένα αυθαίρετο παράθυρο επερώτησης χρονικής στιγμής $W_j \in R_{a\times b}$ που εκτελείται στο χρονικό αποτύπωμα t_j (ήτοι, $T_i \notin W_j \mid T_i^{\dagger} \in W_j$ και $T_i \in W_j \mid T_i^{\dagger} \notin W_j$, αντίστοιχα). Όπως παρουσιάστηκε επίσης και στο Σχήμα 6.5(b), η τομή των τροχιών T_i, $T_i^{\dagger}$ με το επίπεδο που καθορίζεται

από τη χρονική συνιστώσα του W_j (ήτοι το χρονικό αποτύπωμα t_j) θα σημειώνεται με δύο σημεία (τα σημεία $p_{i,j}$ και $p_{i,j}^{\dagger}$, αντίστοιχα στο Σχήμα 6.5(b)) που έχουν μεταξύ τους, αποστάσεις $\delta x_{i,j}$ και $\delta y_{i,j}$ κατά μήκος των αξόνων *x*- και *y*-, αντίστοιχα.



(a) $\delta x_{i,j} < 0 \wedge \delta y_{i,j} > 0$ (b) $\delta x_{i,j} > 0 \wedge \delta y_{i,j} < 0$ (c) $\delta x_{i,j} > 0 \wedge \delta y_{i,j} > 0$ (d) $\delta x_{i,j} < 0 \wedge \delta y_{i,j} < 0$

Σχήμα 6.6: Περιοχές εντός των οποίων πρέπει να βρεθεί η κάτω αριστερά γωνία του παραθύρου επερώτησης για να ανακτήσουμε την τροχιά T_i ως λανθασμένη αρνητική



(a) $\delta x_{i,j} < 0 \wedge \delta y_{i,j} > 0$ (b) $\delta x_{i,j} > 0 \wedge \delta y_{i,j} < 0$ (c) $\delta x_{i,j} > 0 \wedge \delta y_{i,j} > 0$ (d) $\delta x_{i,j} < 0 \wedge \delta y_{i,j} < 0$

Σχήμα 6.7: Περιοχές εντός των οποίων πρέπει να βρεθεί η κάτω αριστερά γωνία του παραθύρου επερώτησης για να ανακτήσουμε την τροχιά T_i ως λανθασμένη θετική

Για να υπολογίσουμε το ποσοστό των παραθύρων επερώτησης χρονικής στιγμής που θα ανακτήσουν την τροχιά T_i ως λανθασμένη αρνητική (λανθασμένη θετική) στο χρονικό αποτύπωμα t_j, πρέπει να διακρίνουμε μεταξύ τεσσάρων περιπτώσεων που αφορούν τα πρόσημα του $\delta x_{i,j}$ και του $\delta y_{i,j}$ όπως φαίνεται στο Σχήμα 6.6 (Σχήμα 6.7, αντίστοιχα). Η σκιασμένη (με πλάγιες γραμμές) περιοχή και στις τέσσερις περιπτώσεις παρουσιάζει την περιοχή εντός της οποίας πρέπει να βρεθεί η κάτω αριστερά γωνία του παραθύρου επερώτησης προκειμένου να ανακτηθεί η τροχιά T_i ως λανθασμένη αρνητική (ή λανθασμένη θετική, αντίστοιχα). Ωστόσο, όπως μπορεί να προκύψει εύκολα από τα σχήματα αυτά, η σκιασμένη περιοχή και στις τέσσερις περιπτώσεις, είναι ίση και για τις λανθασμένες αρνητικές και τις λανθασμένες θετικές και μπορεί να υπολογιστεί από την ακόλουθη εξίσωση:

$$A_{i,j} = a \cdot b - \left(a - \left|\delta x_{i,j}\right|\right) \cdot \left(b - \left|\delta y_{i,j}\right|\right) \tag{6.4}$$

Δεδομένου ότι το W_j ισχύει όταν βρεθεί (είτε μερικώς είτε ολικώς) εντός του μοναδιαίου χώρου, η κάτω αριστερή γωνία του παραθύρου επερώτησης πρέπει να βρεθεί εντός μιας περιοχής του χώρου επιφάνειας ίσης με $(1+a)\cdot(1+b)$. Έτσι, η πιθανότητα η τροχιά T_i να ανακτηθεί ως λανθασμένη αρνητική ή λανθασμένη θετική στο χρονικό αποτύπωμα t_j είναι:

$$P\left(T_i \notin W_j \mid T_i^{\dagger} \in W_j\right) = P\left(T_i \in W_j \mid T_i^{\dagger} \notin W_j\right) = \frac{A_{i,j}}{(1+a)\cdot(1+b)} = \frac{a \cdot b - \left(a - \left|\delta x_{i,j}\right|\right) \cdot \left(b - \left|\delta y_{i,j}\right|\right)}{(1+a)\cdot(1+b)} \tag{6.5}$$

Δεδομένης δε της υπόθεσής μας σχετικά με την κατανομή των παραθύρων επερώτησης, η μέση πιθανότητα μία τροχιά T_i να είναι λανθασμένη αρνητική σε σχέση με ένα αυθαίρετο παράθυρο

επερώτησης $W_j \in R_{a\times b}$ σε οποιοδήποτε χρονικό αποτύπωμα μπορεί να αποκτηθεί ολοκληρώνοντας την Εξ.(6.5) σε όλα τα χρονικά αποτυπώματα εντός του μοναδιαίου χώρου. Επειδή $P\left(T_i \notin W_j \mid T_i^{\dagger} \in W_j\right) = P\left(T_i \in W_j \mid T_i^{\dagger} \notin W_j\right)$, προκύπτει ότι:

$$AvgP_{i,N}\left(R_{a\times b}\right) = AvgP_{i,P}\left(R_{a\times b}\right) = \int_0^1 P\left(T_i \notin W_j \mid T_i^{\dagger} \in W_j\right) dt = \int_0^1 P\left(T_i \in W_j \mid T_i^{\dagger} \notin W_j\right) dt \qquad (6.6)$$

Ωστόσο, δεδομένου ότι κάθε αρχική τροχιά T_i είναι ένα σύνολο m_i δειγματοληπτούμενων σημείων που μεταξύ τους εφαρμόζεται γραμμική παρεμβολή, η Εξ.(6.6) μετατρέπεται ως εξής:

$$AvgP_{i,N}\left(R_{a\times b}\right) = AvgP_{i,P}\left(R_{a\times b}\right) =$$
$$\sum_{k=1}^{m_i-1} \frac{1}{t_{i,k+1}-t_{i,k}} \int_{t_k}^{t_{k+1}} P\left(T_i \notin W_j \mid T_i^{\dagger} \in W_j\right) dt = \sum_{k=1}^{m_i-1} \frac{1}{t_{i,k+1}-t_{i,k}} \int_{t_k}^{t_{k+1}} P\left(T_i \in W_j \mid T_i^{\dagger} \notin W_j\right) dt \qquad (6.7)$$

και η $\delta x_{i,j}$ και $\delta y_{i,j}$ μπορούν εύκολα να διατυπωθούν ως απλές συναρτήσεις του t όταν $t_{i,k} \le t \le t_{i,k+1}$, μεταξύ των δειγματοληπτούμενων σημείων:

$$\delta x_i(t) = \delta x_{i,k} + \left(t - t_{i,k}\right) \cdot \frac{\delta x_{i,k+1} - \delta x_{i,k}}{t_{i,k+1} - t_{i,k}}, \qquad \text{και} \qquad (6.8)$$

$$\delta y_i(t) = \delta y_{i,k} + \left(t - t_{i,k}\right) \cdot \frac{\delta y_{i,k+1} - \delta y_{i,k}}{t_{i,k+1} - t_{i,k}} \qquad (6.9)$$

Αν αντικαταστήσουμε την Εξ.(6.8), Εξ.(6.9) και Εξ.(6.5) στην Εξ.(6.7) και εκτελέσουμε τους απαραίτητους υπολογισμούς λαμβάνουμε τον ακόλουθο τύπο:

$$AvgP_{i,N}\left(R_{a\times b}\right) = AvgP_{i,P}\left(R_{a\times b}\right) =$$
$$\sum_{k=1}^{m_i-1} \frac{\left(t_{i,k+1}-t_{i,k}\right)}{(1+a)\cdot(1+b)} \cdot \left(\frac{b\left(\left|\delta x_{i,k}\right| + \left|\delta x_{i,k+1}\right|\right)}{2} + \frac{a\left(\left|\delta y_{i,k}\right| + \left|\delta y_{i,k+1}\right|\right)}{2} - \frac{2\left|\delta x_{i,k}\delta y_{i,k}\right| + 2\left|\delta x_{i,k+1}\delta y_{i,k+1}\right| + \left|\delta x_{i,k}\delta y_{i,k+1}\right| + \left|\delta x_{i,k+1}\delta y_{i,k}\right|}{6} \right) \qquad (6.10)$$

Τέλος, αντικαθιστώντας την Εξ.(6.10) στην Εξ.(6.2) και θέτοντας

$$e = 2\left|\delta x_{i,k}\delta y_{i,k}\right| + 2\left|\delta x_{i,k+1}\delta y_{i,k+1}\right| + \left|\delta x_{i,k}\delta y_{i,k+1}\right| + \left|\delta x_{i,k+1}\delta y_{i,k}\right| \qquad (6.11)$$

αποδεικνύεται το Λήμμα 6.1 ■

6.3.2. Συζήτηση για το Λήμμα 6.1

Η Εξ.(6.1), το κύριο αποτέλεσμα του Λήμματος 6.1, μπορεί να χρησιμοποιηθεί απευθείας για να εκτιμήσουμε το μέσο αριθμό λανθασμένων αρνητικών και λανθασμένων θετικών για παράθυρα επερώτησης χρονικής στιγμής με γνωστό μέγεθος κατά μήκος των αξόνων *x*- και *y*- (*a* και *b*, αντίστοιχα). Από τον τύπο εύκολα προκύπτει ότι ο μέσος αριθμός λανθασμένων αρνητικών στα αποτελέσματα μιας επερώτησης χρονικής στιγμής είναι ίσος με τον αντίστοιχο μέσο αριθμό λανθασμένων θετικών, ενώ οι τιμές τους εξαρτώνται κυρίως από την περίμετρο του παραθύρου επερώτησης (*a*+*b*) και όχι από την επιφάνεια του ($a \cdot b$). Ωστόσο, θα πρέπει να αναφερθεί ρητά ότι το Λήμμα 6.1 ισχύει μόνο για την περίπτωση των ομοιόμορφα κατανεμημένων παραθύρων επερωτήσεων· ο εκτιμώμενος μέσος αριθμός λανθασμένων αρνητικών και λανθασμένων θετικών χρησιμεύει ως μετρική που εκτιμά τις απώλειες δεδομένων λόγω συμπίεσης, αντί να παρέχει ακριβές αποτέλεσμα για τις ανεξάρτητες επερωτήσεις.

Ένα επιπλέον ενδιαφέρον αποτέλεσμα είναι ότι το σφάλμα που εισάγεται στα αποτελέσματα επερωτήσεων λόγω της συμπίεσης τροχιάς εξαρτάται από τις απόλυτες τιμές των $|\delta x_{i,k}|$ και $|\delta y_{i,k}|$ και όχι από τα τετράγωνά τους δηλαδή τα $\delta x_{i,k}^2$ και $\delta y_{i,k}^2$. Αυτό δεν είναι αναμενόμενο αποτέλεσμα και μας οδηγεί στην ακόλουθη συζήτηση. Πιο συγκεκριμένα, η Εξ.(6.1) δηλώνει ότι η ελαχιστοποίηση του σφάλματος που εισάγεται σε αποτελέσματα επερωτήσεων χρονικής στιγμής σε συμπιεσμένες τροχιές συμβαίνει όταν ελαχιστοποιείται η $|\delta x_{i,k}|+|\delta y_{i,k}|$, αντί της Σύγχρονης Ευκλείδειας Απόστασης (SED), που θεωρείται κριτήριο βελτιστοποίησης στην πλειονότητα των υφιστάμενων αλγορίθμων συμπίεσης τροχιάς. Αναμένεται συνεπώς ότι η χρήση της $|\delta x_{i,k}|+|\delta y_{i,k}|$ αντί της *SED* ως κριτήριο ελαχιστοποίησης στους αλγορίθμους συμπίεσης τροχιάς, θα μας οδηγήσει σε απλουστευμένες τροχιές που παράγουν μικρότερες τιμές σφάλματος στα αποτελέσματα των επερωτήσεων χρονικής στιγμής.

Προφανώς, η αξιολόγηση της Εξ.(6.1) είναι μία δαπανηρή διαδικασία· δεδομένου ότι προϋποθέτει ένα διπλό άθροισμα, η χρονική της πολυπλοκότητα είναι $O(n \cdot m)$ όπου n είναι ο αριθμός των τροχιών και m είναι ο (μέσος) αριθμός δειγματοληπτούμενων σημείων ανά τροχιά. Μ' άλλα λόγια, επειδή η Εξ.(6.1) περιλαμβάνει τον υπολογισμό των $\delta x_{i,k}$, $\delta y_{i,k}$, μεταξύ της κάθε πλειάδας των αρχικών και συμπιεσμένων τροχιών σε κάθε χρονικό αποτύπωμα για το οποίο λάβαμε αρχικώς δείγμα της τροχιάς, απαιτεί επεξεργασία ολόκληρου του αρχικού συνόλου δεδομένων καθώς και του συμπιεσμένου του αντίστοιχου. Από την άλλη πλευρά, όπως δηλώθηκε ήδη στην Ενότητα 6.2, η συντριπτική πλειοψηφία των προτεινόμενων αλγορίθμων συμπίεσης τροχιάς, βασίζουν την απόφασή τους για το σημείο των αρχικών δεδομένων τροχιάς που θα διαγράψουν, στην τιμή της *SED*· ωστόσο, επειδή $SED_i(t)=\sqrt{\delta x_i(t)^2+\delta y_i(t)^2}$, ο αντίστοιχος αλγόριθμος θα πρέπει πρώτα να αξιολογήσει τις $\delta x_i(t)$ και $\delta y_i(t)$ στα χρονικά αποτυπώματα $t_{i,k}$ λαμβάνοντας έτσι τις $\delta x_{i,k}$ και $\delta y_{i,k}$, αντίστοιχα. Κατά συνέπεια, κάθε αλγόριθμος συμπίεσης τροχιάς που χρησιμοποιεί την *SED* ως κριτήριο βάσει του οποίου θα αποφασίσει ποια σημεία τροχιάς θα διαγράψει, υπολογίζει επίσης και τις $\delta x_{i,k}$ και $\delta y_{i,k}$. Άρα, η Εξ.(6.1) μπορεί να υπολογιστεί κατά την εκτέλεση του αλγορίθμου, με ελάχιστη επιπλέον επιβάρυνση στον αρχικό αλγόριθμο· η παραπάνω παρατήρηση επιβεβαιώνεται περαιτέρω στην πειραματική μας μελέτη που παρουσιάζεται στην επόμενη ενότητα.

Επιπλέον, επειδή η Εξ.(6.1) περιλαμβάνει τις διαστάσεις επερώτησης a και b, προκύπτει ότι διαφορετικές τιμές των a και b θα οδηγήσουν σε διαφορετικούς υπολογισμούς για το μέσο σφάλμα. Ωστόσο, μια τέτοια προσέγγιση (δηλαδή, η αξιολόγηση της Εξ.(6.1) από την αρχή για κάθε διαφορετικό μέγεθος επερώτησης), θα οδηγήσει σε υψηλό υπολογιστικό κόστος δεδομένου ότι απαιτεί επιπλέον $O(n \cdot m)$ χρόνο. Για να ξεπεράσουμε αυτό το μειονέκτημα, η Εξ.(6.1) μπορεί να ξαναδιατυπωθεί ως εξής:

$$E_N(R_{a\times b}) = E_P(R_{a\times b}) = \frac{A\cdot a + B\cdot b + C}{(1+a)\cdot(1+b)} \tag{6.12}$$

όπου $A=\sum_{i=1}^{n}\sum_{k=1}^{m_i-1}(t_{i,k+1}-t_{i,k})\cdot\frac{|\delta y_{i,k}|+|\delta y_{i,k+1}|}{2}$, $B=\sum_{i=1}^{n}\sum_{k=1}^{m_i-1}(t_{i,k+1}-t_{i,k})\cdot\frac{|\delta x_{i,k}|+|\delta x_{i,k+1}|}{2}$ και

$C = -\sum_{i=1}^{n}\sum_{k=1}^{m_i-1}\left(t_{i,k+1} - t_{i,k}\right)\cdot\frac{e}{6}$. Συνεπώς, στην περίπτωση όπου το μέσο σφάλμα πρέπει να προσδιοριστεί για διάφορα μεγέθη επερώτησης (ήτοι, διάφορα μεγέθη *a* και *b*), αντί να υπολογίσουμε απευθείας την Εξ.(6.1) για κάθε διαφορετικό μέγεθος επερώτησης, θα μπορούσαν να υπολογιστούν πρώτα τα *A*, *B* και *C* και κατόπιν να χρησιμοποιηθούν στην Εξ.(6.12)· μια προσέγγιση που μειώνει κατά πολύ το κόστος υπολογισμού σε χρόνο *O*(1).

6.4. Πειραματική Μελέτη

Στην ενότητα αυτή παρουσιάζουμε μια σειρά πειραμάτων χρησιμοποιώντας συνθετικά και πραγματικά σύνολα δεδομένων. Ο στόχος της πειραματικής μας μελέτης είναι διττός:

- πρώτον, να παρουσιάσουμε την επιβάρυνση που εισάγεται στον υπολογισμό ενός αλγορίθμου συμπίεσης όταν υπολογίζουμε κατά τη διάρκειά του τους παράγοντες *A*, *B* και *C* της Εξ.(6.12) και,
- δεύτερον, να παρουσιάσουμε την ακρίβεια της εκτίμησης που δίνεται από το αναλυτικό μας μοντέλο για τον αριθμό των λανθασμένων αρνητικών και λανθασμένων θετικών σε συνθετικά και πραγματικά σύνολα δεδομένων τροχιάς.

6.4.1. Πειραματικό Πλαίσιο

Για άλλη μια φορά, πειραματιστήκαμε με σύνολα δεδομένων από τον πραγματικό κόσμο, το στόλο των *φορτηγών (trucks)* (Ενότητα 1.5.1). Χρησιμοποιήσαμε επίσης το συνθετικό σύνολο δεδομένων NG 2000 (Ενότητα 1.5.3). Όλα τα σύνολα δεδομένων κανονικοποιήθηκαν στο χώρο [0,1]. Για να ελέγξουμε την ακρίβεια του μοντέλου μας και να παραχθούν συμπιεσμένα σύνολα δεδομένων, εφαρμόσαμε τον αλγόριθμο TD-TR που προτείνεται από την [MB04]. Κατόπιν τον εκτελέσαμε με όλα τα (πραγματικά και συνθετικά) σύνολα δεδομένων, με κατώφλια μεταξύ 0.001 και 0.02 του συνολικού χώρου, παράγοντας έτσι, τα αντίστοιχα συμπιεσμένα σύνολα δεδομένων. Τέλος χρησιμοποιήσαμε τα αρχικά και συμπιεσμένα σύνολα δεδομένων και δημιουργήσαμε αρκετά 3D R-δέντρα [TVS96] για να επιταχύνουμε τη διαδικασία επερωτήσεων που χρησιμοποιήθηκαν κατά την εκτέλεση των πειραμάτων ποιότητας. Ο Πίνακας 6.2 παρουσιάζει συνοπτικές πληροφορίες για τα (αρχικά και συμπιεσμένα) σύνολα δεδομένων που χρησιμοποιούνται. Τα πειράματα εκτελέστηκαν σε ένα PC με Microsoft Windows XP με επεξεργαστή AMD Athlon 64 3GHz, 1 GB RAM και αρκετά GB μέγεθος δίσκου.

Πίνακας 6.2: Συνοπτικές Πληροφορίες Συνόλων Δεδομένων

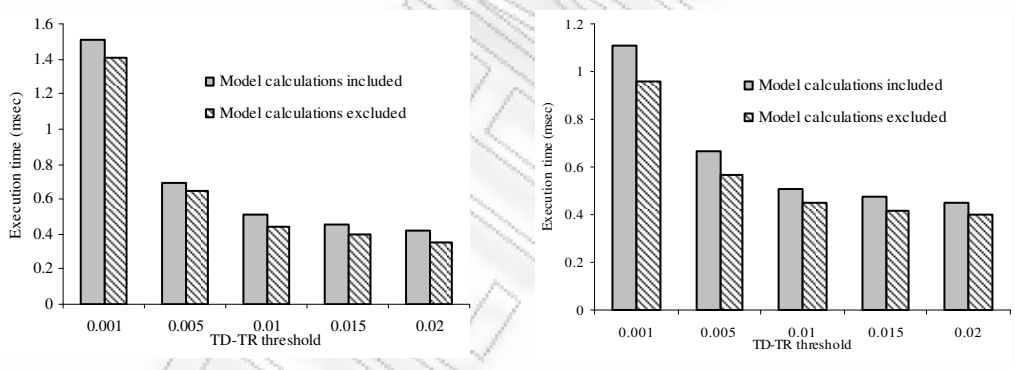
	Αρχικά Σύνολα Δεδομένων		*Συμπιεσμένα Σύνολα Δεδομένων (#εγγραφών)*				
			τιμή κατωφλίου του αλγορίθμου TD-TR				
	# τροχιών	*# εγγραφών*	*0.001*	*0.005*	*0.010*	*0.015*	*0.020*
Trucks	273	112,203	62,067	20,935	12,636	9,274	7,571
NG 2000	2,000	800,000	229,167	120,437	88,565	74,638	65,410

6.4.2. Πειράματα ως προς την Απόδοση

Για να αποδείξουμε ότι η πρότασή μας εφαρμόζεται σε δεδομένα τροχιών και να εκτιμήσουμε την επιβάρυνση που εισάγεται σε έναν αλγόριθμο συμπίεσης όταν υπολογίζουμε τις τιμές των παραγόντων *A*, *B* και *C* που εισάγονται στην Εξ.(6.12), τρέξαμε τον αλγόριθμο συμπίεσης TD-TR στα πραγματικά

δεδομένα και μετρήσαμε το μέσο χρόνο εκτέλεσης που απαιτείται για κάθε τροχιά, κλιμακώνοντας επίσης και το κατώφλι του αλγορίθμου. Κατόπιν τροποποιήσαμε τον αλγόριθμο για να υπολογίσουμε τις παραμέτρους του μοντέλου (δηλαδή τις τιμές των *A*, *B* και *C* στην Εξ.(6.12)) εντός της εκτέλεσης του και τον τρέξαμε ξανά για το ίδιο σύνολο δεδομένων με τις ίδιες παραμέτρους. Τα σχετικά αποτελέσματα φαίνονται στο Σχήμα 6.8.

Πιο συγκεκριμένα, το Σχήμα 6.8(a) και το Σχήμα 6.8(b) παριστάνουν το χρόνο εκτέλεσης του αλγορίθμου TD-TR ανά συμπιεσμένη τροχιά (σε milliseconds), με και χωρίς την αξιολόγηση των παραμέτρων του μοντέλου, στο σύνολα δεδομένων trucks και στο σύνολο NG 2000. Ένα πρώτο συμπέρασμα είναι ότι ο χρόνος εκτέλεσης του αλγορίθμου μειώνεται όσο η τιμή του κατωφλίου TD-TR αυξάνεται· αυτό είναι αναμενόμενο αποτέλεσμα, δεδομένου ότι τυπικά, ο αριθμός των επαναλήψεων του αλγορίθμου αυξάνεται, όσο μειώνεται η τιμή του κατωφλίου. Ωστόσο, το κύριο αποτέλεσμα που προκύπτει από το Σχήμα 6.8 είναι ότι η επιβάρυνση που εισάγεται στην εκτέλεση του αλγορίθμου, είναι τυπικά μικρή (ήτοι, η διαφορά μεταξύ των δύο στηλών). Σε όλες τις περιπτώσεις, η επιβάρυνση που εισάγεται στον αλγόριθμο είναι μεταξύ του 7% και του 19% του αρχικά απαιτούμενου χρόνου εκτέλεσης· επιπλέον σε απόλυτους χρόνους, η επιβάρυνση που εισάγεται ποτέ δεν υπερβαίνει τα 0.2 milliseconds ανά τροχιά. Κατά συνέπεια, η συζήτηση που παρουσιάστηκε στην Ενότητα 3.2 επιβεβαιώνεται και το μοντέλο μας μπορεί να αξιολογηθεί ως επέκταση της εκτέλεσης του αλγορίθμου συμπίεσης, εισάγοντας μία μικρή / ίσως και αμελητέα επιβάρυνση.



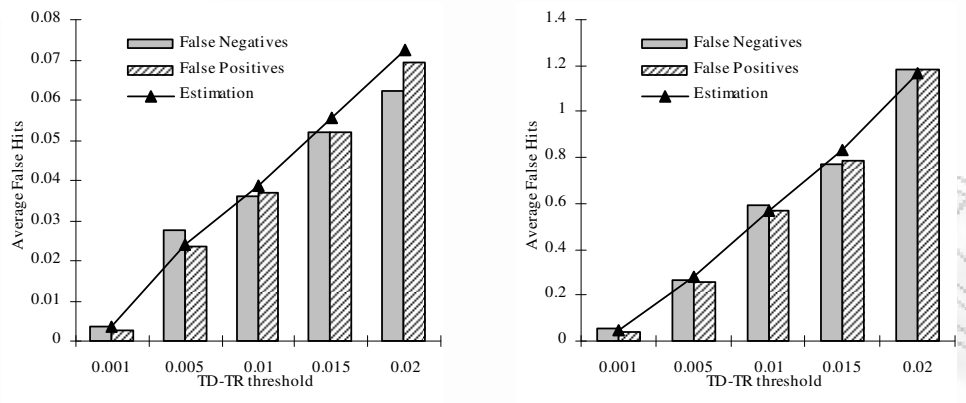
Σχήμα 6.8: Χρόνος εκτέλεσης για τον αλγόριθμο TD-TR με και χωρίς τον υπολογισμό των παραμέτρων του μοντέλου (a) στο σύνολο δεδομένων trucks και (b) στο σύνολο συνθετικών δεδομένων, κλιμακώνοντας την τιμή του κατωφλίου TD-TR.

6.4.3. Πειράματα ως προς την Ποιότητα

Η μετρική που χρησιμοποιείται για να δείξουμε την ποιότητα των εκτιμήσεών μας, είναι ο αναφερόμενος *μέσος αριθμός των λανθασμένων αρνητικών και λανθασμένων θετικών*, $\overline{E_N}$ και $\overline{E_P}$, αντίστοιχα. Επισήμως, οι μετρικές αυτές ορίζονται ως:

$$\overline{E_N} = \frac{1}{n}\sum_{i=1..n} E_{N,i}\ ,\ \overline{E_P} = \frac{1}{n}\sum_{i=1..n} E_{P,i}$$

όπου, *n* είναι ο αριθμός των εκτελεσθέντων επερωτήσεων και $E_{N,i}$ ($E_{P,i}$) ο πραγματικός αριθμός των λανθασμένων αρνητικών (λανθασμένων θετικών, αντίστοιχα) στην *i*-στη επερώτηση. Στα επόμενα πειράματα, το *n* τίθεται στις 10000 επερωτήσεις χρονικής στιγμής.



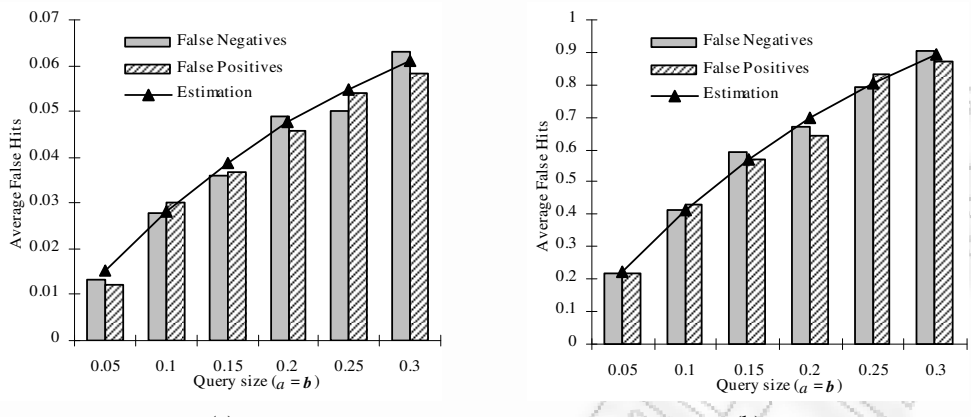
(a) (b)

Σχήμα 6.9: Ακρίβεια του μοντέλου κλιμακώνοντας τη τιμή του κατωφλίου του TD-TR (a) στο σύνολο δεδομένων trucks και (b) στο σύνολο των συνθετικών δεδομένων

Η πρώτη σειρά πειραμάτων εκτελέστηκε και με τα πραγματικά και τα συνθετικά σύνολα δεδομένων. Πιο συγκεκριμένα, εκτελέσαμε 10000 τετράγωνες επερωτήσεις χρονικής στιγμής μεγέθους 0.10×0.10 (ήτοι, που καλύπτουν το 1% του μοναδιαίου χώρου) που κατανέμονται τυχαία εντός του μοναδιαίου χώρου, και στα αρχικά και στα συμπιεσμένα σύνολα δεδομένων και μετά, χρησιμοποιώντας τα αποτελέσματα κάθε μίας επερώτησης στα δύο σύνολα δεδομένων, υπολογίσαμε τον πραγματικό αριθμό των λανθασμένων αρνητικών και των λανθασμένων θετικών, $E_{N,i}$ και $E_{P,i}$, αντίστοιχα. Το Σχήμα 6.9 παρουσιάζει τα αποτελέσματα αυτού του πειράματος κλιμακώνοντας την τιμή του κατωφλίου συμπίεσης για το σύνολο δεδομένων trucks και για το συνθετικό σύνολο δεδομένων. Ένα πρώτο συμπέρασμα είναι ότι ο μέσος αριθμός λανθασμένων αποτελεσμάτων (αρνητικών και θετικών) είναι γραμμικός με την τιμή του κατωφλίου συμπίεσης του TD-TR. Επιπλέον, οι εκτιμήσεις, $\overline{E_N}$ και $\overline{E_P}$, του μοντέλου μας είναι πολύ κοντά στις πραγματικές τιμής των μέσων λανθασμένων αρνητικών και λανθασμένων θετικών που αναφέρθηκαν από τα πειράματα, ανεξάρτητα από την τιμή του κατωφλίου συμπίεσης. Πιο συγκεκριμένα, το μέσο σφάλμα στην εκτίμηση του συνθετικού συνόλου δεδομένων είναι περίπου 6% και κυμαίνεται μεταξύ του 0.2% και του 14%· όσον αφορά το σύνολο δεδομένων trucks, το μέσο σφάλμα αυξάνεται γύρω στο 10.6%, κυρίως λόγω του σφάλματος που εισάγεται σε μικρές τιμές του κατωφλίου TD-TR.

Στο δεύτερό μας πείραμα χρησιμοποιήσαμε τις ίδιες πειραματικές ρυθμίσεις (δηλαδή σύνολα δεδομένων, αριθμός επερωτήσεων), αλλά καθορίσαμε το κατώφλι TD-TR στο 0.01 και κλιμακώσαμε το μέγεθος του παραθύρου επερώτησης χρονικής στιγμής μεταξύ 0.05×0.05 και 0.30×0.30 (με αποτέλεσμα 0.25% και 9% μοναδιαίου χώρου, αντίστοιχα). Τα αντίστοιχα αποτελέσματα παρουσιάζονται στο Σχήμα 6.10(a) και Σχήμα 6.10(b) για το σύνολο δεδομένων trucks και για το συνθετικό σύνολο δεδομένων, αντίστοιχα. Και πάλι, είναι σαφές ότι το μοντέλο μας είναι υψηλής ακρίβειας, δίνοντας εκτιμήσεις του $\overline{E_N}$ και $\overline{E_P}$ με σφάλματα για τα συνθετικά σύνολα δεδομένων μεταξύ του 0.2% και του 8.7% και ο μέσος όρος είναι περίπου 2.9% (ενώ το αντίστοιχο μέσο σφάλμα για το σύνολο δεδομένων trucks είναι 7.5%.) Άλλο αξιοσημείωτο συμπέρασμα είναι ότι ο μέσος αριθμός των λανθασμένων θετικών και των λανθασμένων αρνητικών είναι υπογραμμικός με το μέγεθος επερώτησης· αναμενόμενο αποτέλεσμα το οποίο προκύπτει απευθείας από τον τρόπο με τον

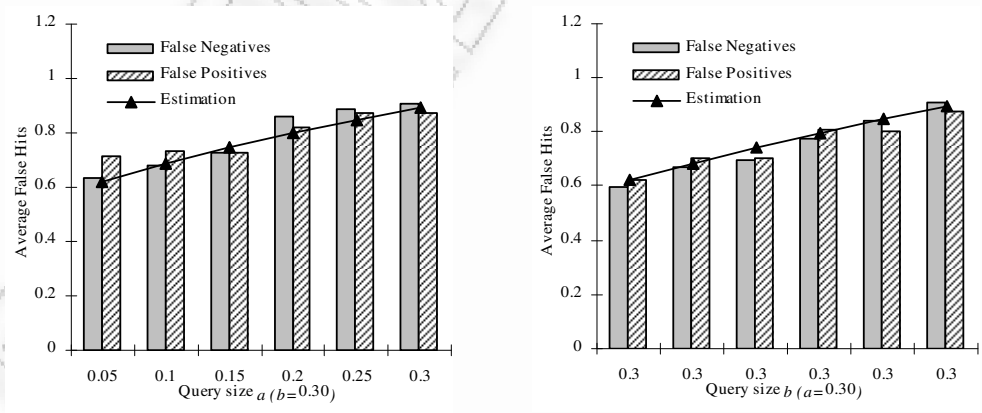
οποίο η Εξ.(6.12) εμπλέκει τα μήκη *a* και *b* των πλευρών επερωτήσεων στον υπολογισμό του σφάλματος.



(a) (b)

Σχήμα 6.10: Ακρίβεια του μοντέλου κλιμακώνοντας το μέγεθος της τετράγωνης επερώτησης (a) στο σύνολο δεδομένων trucks και (b) στο σύνολο των συνθετικών δεδομένων

Στο τελευταίο πείραμα επαληθεύσαμε το αποτέλεσμα της χρήσης μη τετραγώνων επερωτήσεων χρονικής στιγμής (δηλαδή, $a \neq b$) στο συνθετικό σύνολο δεδομένων (ενώ τα πειράματα με το σύνολο δεδομένων trucks μας έδωσε παρόμοια αποτελέσματα). Πιο συγκεκριμένα, χρησιμοποιήσαμε παράθυρα επερώτησης χρονικής στιγμής από 0.05×0.30 (όπου $a<<b$) μέχρι 0.30×0.30 (όπου $a=b$)· επίσης κλιμακώσαμε το μέγεθος επερώτησης προς την άλλη κατεύθυνση (από 0.30×0.05 έως 0.30×0.30). Το αποτέλεσμα αυτού του πειράματος, που παρουσιάζεται στο Σχήμα 6.11(a) και (b) αντίστοιχα, που είχαν παρόμοια αποτελέσματα με αυτά που παρουσιάστηκαν στην προηγούμενη παράγραφο σχετικά με τις τετράγωνες (δηλαδή, $a=b$) επερωτήσεις χρονικής στιγμής. Πιο συγκεκριμένα, το μοντέλο μας είναι για άλλη μια φορά πολύ ακριβές, παράγοντας εκτιμήσεις με σφάλμα μεταξύ 0.6% και 7.2%, ενώ το μέσο σφάλμα είναι 3.5%.



(a) (b)

Σχήμα 6.11: Ακρίβεια του μοντέλου που κλιμακώνοντας το μέγεθος της επερώτησης προς τον (a) x-άξονα και τον (b) y- άξονα, στο συνθετικό σύνολο δεδομένων.

6.5. Συμπεράσματα

Οι σχετικές εργασίες στο θέμα της συμπίεσης τροχιάς έχουν εστιάσει στην ανάπτυξη αλγορίθμων συμπίεσης τονίζοντας επίσης και το σφάλμα που εισάγεται στη θέση του κάθε αντικειμένου από τη συμπίεση. Στη διατριβή αυτή, αναγνωρίζοντας ότι οι χρήστες κατά πάσα πιθανότητα ενδιαφέρονται για το σφάλμα που εισάγεται από τη συμπίεση στα χωροχρονικά *αποτελέσματα επερωτήσεων*, παρουσιάσαμε το πρώτο θεωρητικό μοντέλο που εκτιμά το σφάλμα αυτό στα αποτελέσματα επερωτήσεων χρονικής στιγμής. Δώσαμε έναν κλειστό τύπο του μέσου αριθμού λανθασμένων αποτελεσμάτων (λανθασμένων αρνητικών και λανθασμένων θετικών) που καλύπτουν την περίπτωση αυθαίρετα κατανεμημένων τροχιών με διάφορες ταχύτητες, κατευθύνσεις κλπ. Υπό διάφορα συνθετικά και πραγματικά σύνολα δεδομένων τροχιών, αποδείξαμε πρώτα ότι το μοντέλο μας μπορεί να εφαρμοσθεί σε πραγματικές συνθήκες – προκύπτει ότι η εκτίμηση των παραμέτρων του μοντέλου εισάγει μία ελάχιστη επιβάρυνση στον αλγόριθμο συμπίεσης τροχιάς – και κατόπιν παρουσιάσαμε την ακρίβεια των εκτιμήσεων μας, με μέσο σφάλμα της τάξης περίπου του 6%.

7. Επίλογος

7.1. Συμπεράσματα

Στη διατριβή αυτή παρουσιάσαμε αρκετές τεχνικές απαραίτητες για την αποτελεσματική Διαχείριση Δεδομένων Τροχιών Κινούμενων Αντικειμένων. Πιο συγκεκριμένα, παρέχουμε αποτελεσματικούς μηχανισμούς που επιτρέπουν στις Βάσεις Δεδομένων Κινούμενων Αντικειμένων την αποτελεσματική αποθήκευση και πραγματοποίηση επερωτήσεων σε ιστορικές τροχιές, προάγοντας τους τομείς της δεικτοδότησης, της επεξεργασίας επερωτήσεων, της υποστήριξη της αβεβαιότητας και της συμπίεσης τροχιών. Στη συνέχεια, παρουσιάζουμε τις συγκεκριμένες συνεισφορές της διατριβής μας.

Στο Κεφάλαιο 2, παρουσιάσαμε δύο πρωτότυπες τεχνικές δεικτοδότησης· εξ' αυτών, η πρώτη βελτιώνει μία ήδη υπάρχουσα λύση (ήτοι το TB*-δέντρο), ενώ η δεύτερη εκμεταλλεύεται τους περιορισμούς που τίθενται στη κίνηση των αντικειμένων από οδικά ή άλλα δίκτυα δίνοντας μία λύση που υπερέχει σε σχέση με τις γενικές μεθόδους δεικτοδότησης. Πιο συγκεκριμένα, για την περίπτωση όπου τα αντικείμενα κινούνται ελεύθερα στο χώρο, αναγνωρίζοντας τα βασικά πλεονεκτήματα του TB-δέντρου, προχωρήσαμε ένα βήμα παραπέρα προτείνοντας ένα καινοτόμο ευρετήριο, που ονομάζεται TB*-δέντρο. Το προτεινόμενο ευρετήριο ξεπερνά τα κύρια μειονεκτήματα του προκατόχου του ενώ ταυτόχρονα διατηρεί όλες τις «επιθυμητές» του ιδιότητες· έτσι, υποστηρίζει εισαγωγές και διαγραφές, συμπίεση τροχιών, ενώ οι επερωτήσεις εκτελούνται χρησιμοποιώντας τους αλγόριθμους που παρέχονται στην [PJT00]. Στη δεύτερη περίπτωση των αντικειμένων που είναι περιορισμένα να κινούνται σε δίκτυο, παρουσιάζουμε το R-δέντρο Σταθερού Δικτύου (Fixed Network R-tree), που είναι ένα δάσος αρκετών 1D R-δέντρων [Gut84] πάνω από ένα απλό 2D R-δέντρο [Gut84]. Το 2D R-δέντρο χρησιμοποιείται για τη δεικτοδότηση των χωρικών δεδομένων του διαγράμματος του δικτύου, ενώ τα 1D R-δέντρα χρησιμοποιούνται για τη δεικτοδότηση του χρονικού διαστήματος της κίνησης του κάθε αντικειμένου σε ένα δεδομένο τμήμα του δικτύου. Επιπλέον, οι τα φύλλα όλων των 1D R-δέντρων δεικτοδοτούνται από ένα άλλο 1D R-δέντρο που χρησιμοποιείται για την απάντηση επερωτήσεων χωρίς χωρική έκταση.

Συγκρίναμε πειραματικά το FNR-δέντρο με το TB*-δέντρο και τα παραδοσιακά 3D R- [TVS96] και TB-δέντρα [PJT00]. Σε διάφορα σύνολα δεδομένων και επερωτήσεις εύρους, το FNR-δέντρο αποδείχτηκε ότι υπερέχει όλων των ανταγωνιστών του στην συντριπτική πλειοψηφία των ρυθμίσεων. Το FNR-δέντρο έχει υψηλή χρησιμοποίηση χώρου, μικρότερο μέγεθος ανά κινούμενο αντικείμενο και υποστηρίζει επερωτήσεις εύρους πολύ πιο αποτελεσματικά. Σε γενικές γραμμές, θεωρούμε ότι το FNR-δέντρο είναι μία μέθοδος προσπέλασης ιδανική για εφαρμογές διαχείρισης στόλου. Ωστόσο, το

FNR-δέντρο μπορεί να χρησιμοποιηθεί μόνο βάσει του σεναρίου που προβλέπει περιορισμό της κίνησης επάνω σε ένα δίκτυο· όταν τα αντικείμενα κινούνται ελεύθερα στο χώρο, αποδεικνύεται ότι το TB*-δέντρο υπερέχει του αρχικού TB-δέντρου στην συντριπτική πλειοψηφία των ρυθμίσεων, όσον αφορά την εισαγωγή δεδομένων και την υποστήριξη επερωτήσεων. Επιπλέον, το TB*-δέντρο είναι πιο συμπαγές από τους ανταγωνιστές του, συμπεριφέρεται καλά σε μη χρονολογικές εισαγωγές τροχιών που εμφανίζονται στον πραγματικό κόσμο και υποστηρίζει διαδικασίες διαγραφής και συμπίεσης τροχιάς αποτελεσματικά.

Στο Κεφάλαιο 3, μελετήσαμε το πρόβλημα της εκτέλεσης επερωτήσεων πλησιέστερου γείτονα (Nearest Neighbor (NN)) σε ιστορικές τροχιές. Οι σχετικές εργασίες επί του θέματος μέχρι τώρα, ασχολούνται κυρίως είτε με σταθερά είτε με κινούμενα σημεία επερωτήσεων σε σταθερά σύνολα δεδομένων ή μελλοντικές (προβλεπόμενες) θέσεις σε ένα σύνολο συνεχώς κινουμένων σημείων. Στην παρούσα διατριβή, παρουσιάσαμε την πρώτη πλήρη αντιμετώπιση ιστορικών επερωτήσεων NN σε τροχιές κινούμενων αντικειμένων που αποθηκεύονται σε δομές τύπου R-δέντρου· άρα, οι λύσεις που παρουσιάζουμε μπορούν να εφαρμοσθούν σε μία ποικιλία ευρετηρίων, όπως το 3D R-δέντρο [TVS96], το TB-δέντρο [PJT00], ή και το πρωτότυπο TB*-δέντρο. Δίνουμε μία σειρά πρωτότυπων μετρικών και βελτιώνουμε τον υφιστάμενο τρόπο υπολογισμό της μετρικής *MINDIST* μεταξύ γραμμικών τμημάτων και ορθογωνίων. Οι μετρικές που προτείνουμε υποστηρίζουν τις στρατηγικές μας για τη διάταξη και το κλάδεμα των δεδομένων και χρησιμοποιούνται σε ένα σύνολο αλγορίθμων που απαντούν σε επερωτήσεις πλησιέστερου γείτονα και ιστορικές επερωτήσεις συνεχούς πλησιέστερου γείτονα για σταθερά ή κινούμενα σημεία επερώτησης. Οι παρουσιαζόμενοι αλγόριθμοι, ακολουθούν τα παραδείγματα «πρώτα στο βαθύτερο» (depth-first) [RKV95] και «πρώτα στο καλύτερο» (best-first) [HS99] και κατόπιν γενικεύονται για την αναζήτηση των k πλησιέστερων γειτόνων.

Για να μετρήσουμε την απόδοση των εισαγόμενων αλγορίθμων διεξήχθη μία εκτεταμένη πειραματική μελέτη βασισμένη σε συνθετικά και πραγματικά σύνολα δεδομένων. Όσον αφορά τους ιστορικά μη συνέχειες αλγορίθμους, δείχθηκε ότι ενώ η αυξητική (πρώτα στον καλύτερο) προσέγγιση είναι πάντα λιγότερο δαπανηρή από τη μη αυξητική (πρώτα στο βαθύτερο) σε όρους προσβάσεων κόμβων, ο πραγματικός χρόνος εκτέλεσης εξαρτάται σε μεγάλο βαθμό από το μήκος της χρησιμοποιούμενης ουράς. Σε γενικές γραμμές η προσέγγιση πρώτα στον καλύτερο υπερέχει του ανταγωνιστή της μόνο για επερωτήσεις NN σημείου με μικρή χρονική έκταση (μικρότερη του 2-4% ανάλογα με το ευρετήριο που χρησιμοποιείται και υπό οποιοδήποτε k), ενώ σε όλες τις υπόλοιπες περιπτώσεις η προσέγγιση πρώτα στο βαθύτερο εκτελείται σε μικρότερο χρόνο. Το μειονέκτημα αυτό των αυξητικών αλγορίθμων οφείλεται κυρίως στο μήκος της ουράς που μπορεί να γίνει τεράστιο, στην περίπτωση του TB-δέντρου και του TB*-δέντρου. Επιπλέον, δείξαμε ότι η βελτίωση που προτείνουμε για τον υπολογισμό της *MINDIST* μπορεί να αυξήσει σε αρκετά μεγάλο βαθμό την απόδοση των προτεινόμενων αλγορίθμων. Τέλος, η πειραματική μελέτη καταδεικνύει ότι η πλειονότητα των αλγορίθμων που παρουσιάστηκαν είναι γραμμικοί ή υπογραμμικοί με τις κύριες παραμέτρους του προβλήματος (σε όρους προσπελάσεων κόμβων): το πλήθος του συνόλου δεδομένων, τη χρονική έκταση της επερώτησης και τον αριθμό των k.

Στο Κεφάλαιο 4, επεκτείνοντας την εργασία μας επάνω στην αναζήτηση NN εξετάσαμε το πρόβλημα της αναζήτησης ομοιότερης τροχιάς (Most Similar Trajectory (MST)). Πιο συγκεκριμένα, οι

υπάρχουσες σχετική εργασίες στον τομέα των επερωτήσεων ομοιότητας μεταξύ τροχιών είτε αγνοούν τη χρονική τους διάσταση, είτε εξετάζουν τροχιές με τον ίδιο ρυθμό δειγματοληψίας. Από την άλλη πάλι, στην παρούσα διατριβή αποδεσμευτήκαμε από αυτές τις υποθέσεις ορίζοντας μία νέα μετρική που βασίζεται στη μέση Ευκλείδεια απόσταση μεταξύ τροχιών, που καλείται *DISSIM*, ενώ στη συνέχεια παρουσιάσαμε μία πλήρη αντιμετώπιση των ιστορικών επερωτήσεων MST σε τροχιές κινούμενων αντικειμένων που αποθηκεύονται σε δομές τύπου R-δέντρου και αποφεύγουν τα μειονεκτήματα των υπαρχόντων μεθόδων.

Προτείναμε ένα σύνολο μετρικών, βασισμένες σε απλές έννοιες των τροχιών, όπως η μέγιστη ταχύτητα του συνόλου δεδομένων και την κάθε μία από τις μετρικές ακολούθησε ένα λήμμα για την υποστήριξη των στρατηγικών διάταξης και κλαδέματος των αλγορίθμων μας· κατόπιν παρουσιάσαμε δύο αλγορίθμους αναζήτησης MST εν ονόματι `BFMSTSearch` και `DFMSTSearch`, που ακολουθούν το παράδειγμα «πρώτα στο καλύτερο» [HS99] και «πρώτα στο βαθύτερο» [RKV95], αντίστοιχα. Υπό διάφορα συνθετικά και πραγματικά σύνολα δεδομένων τροχιών, αποδείξαμε την υπεροχή της προτεινόμενης μετρικής *DISSIM* ως προς άλλες σχετικές προτάσεις [VKG02], [COO05], από άποψη ποιότητας, ενώ οι αλγόριθμοι μας παρουσιάζουν υψηλή ισχύ κλαδέματος κατά την επεξεργασία επερωτήσεων MST, που επαληθεύτηκε επίσης στην περίπτωση επερωτήσεων *k*-MST. Μεταξύ των αλγορίθμων που προτάθηκαν, ο `BFMSTSearch` φαίνεται πιο πολλά υποσχόμενος διότι παρουσιάζει καλύτερη απόδοση σε σχέση με τον ανταγωνιστή του `DFMSTSearch`· πιο συγκεκριμένα, παρουσιάζει γραμμική συμπεριφορά όσον αφορά το χρόνο εκτέλεσης και τις προσπελάσεις κόμβων, ενώ η ισχύς κλαδέματος που παρέχει, παραμένει άνω του 90% σε όλες τις περιπτώσεις που εξετάσαμε κατά την πειραματική μας μελέτη (ενώ η ισχύς κλαδέματος του `DFMSTSearch` υποβαθμίζεται σε πολύ χαμηλές τιμές όσο αυξάνεται το μήκος της τροχιάς επερώτησης).

Θα πρέπει εδώ να επισημάνουμε ότι κανένας από τους αλγορίθμους που προτείνονται για επερωτήσεις πλησιέστερου γείτονα και ομοιότερης τροχιάς δεν απαιτεί κάποια ειδική δομή ευρετηρίου· το αντίθετο, όλοι οι προτεινόμενοι αλγόριθμοι μπορούν να εφαρμοσθούν απευθείας σε οποιοδήποτε μέλος της οικογένειας των R-δέντρων χρησιμοποιείται για τη δεικτοδότηση τροχιών, όπως το 3D R-δέντρο [TVS96], το TB-δέντρο [PJT00] και το TB*-δέντρο που προτείνεται στην παρούσα διατριβή. Εξ' όσων γνωρίζουμε, η εφαρμογή της πρότασης της διατριβής μας δίνει για πρώτη φορά τη δυνατότητα σε ένα χωροχρονικό ευρετήριο να υποστηρίζει επερωτήσεις εύρους, χρονικής στιγμής, τοπολογικές, πλησιέστερου γείτονα και ομοιότερης τροχιάς. Επιπροσθέτως, κάποιες από τις προτεινόμενες τεχνικές, της διατριβής μας, έχουν υλοποιηθεί στο Αντικειμενο-Σχεσιακό Σύστημα Διαχείρισης Βάσης Δεδομένων (Object – Relational DBMS) ORACLE και ενσωματωθεί στην μηχανή HERMES [PFGT08]. Πιο συγκεκριμένα, η μηχανή HERMES, μέχρι τούδε, έχει επεκταθεί ώστε να συμπεριλαμβάνει το TB-δέντρο [PJT00], μαζί με τους αλγορίθμους πλησιέστερου γείτονα σημείου και τροχιάς που παρουσιάστηκαν στο Κεφάλαιο 3.

Όσον αφορά τη διαχείριση της αβεβαιότητας θέσης των χωροχρονικών τροχιών, στο Κεφάλαιο 5 επιχειρηματολογούμε υπέρ της άποψης ότι υπάρχουν περιπτώσεις όπου ο χρήστης θα προτιμούσε να γνωρίζει την επίδραση της αβεβαιότητας στα αποτελέσματα της επερώτησης, χωρίς να εκτελεί στην πράξη την επερώτηση. Τέτοιες περιπτώσεις περιλαμβάνουν τις διαδραστικές επερωτήσεις βάσεων δεδομένων, ρυθμίσεις ανακρίβειας, λειτουργίες αποθηκών δεδομένων και επερωτήσεις κάτω από το

σενάριο των ανοικτών αγορών [Ioa07] όπως συζητήθηκε εκτενώς. Για το σκοπό αυτό, δόθηκε ένα θεωρητικό μοντέλο που εκτιμά το σφάλμα που εισάγεται από την αβεβαιότητα θέσης του κάθε αντικειμένου στα αποτελέσματα των χωροχρονικών επερωτήσεων χρονικής στιγμής, καθώς επίσης και σε απλές επερωτήσεις εύρους σε σταθερά χωρικά δεδομένα. Το μοντέλο που προτείνεται αποτελείται από ένα κλειστό τύπο που υπολογίζει το μέσο αριθμό λανθασμένων απαντήσεων, που ταξινομούνται ως λανθασμένες θετικές και λανθασμένες αρνητικές, με βάση τρεις αρχικές υποθέσεις: ομοιόμορφη αβεβαιότητα θέσης (ακολουθώντας το προτεινόμενο μοντέλο από την [TWHC04] προκειμένου να περιγράψουμε την αβέβαιη θέση των τροχιών), ομοιόμορφα κατανεμημένα δεδομένα, και, σταθερή τιμή του κατωφλίου αβεβαιότητας [TWHC04] (ακτίνα του δίσκου αβεβαιότητας). Στη συνέχεια, αποδεσμεύσαμε τις υποθέσεις αυτές προς πιο ρεαλιστικές ρυθμίσεις, χρησιμοποιώντας τη διμεταβλητή κανονική κατανομή για την περιγραφή της αβεβαιότητας θέσης και τα ιστογράμματα *MinSkew* ώστε να υποστηρίξουμε αυθαίρετα κατανεμημένα δεδομένα και ακτίνες αβεβαιότητας.

Η ακρίβεια του προτεινόμενου μοντέλου σε χωροχρονικά δεδομένα τροχιάς αξιολογήθηκε μέσω πειραμάτων με διάφορα σύνολα χωροχρονικών δεδομένων. Πιο συγκεκριμένα, το μοντέλο μας αποδεικνύεται ότι παρέχει υψηλή ακρίβεια με το μέσο σφάλμα των $\overline{E_P}$ και $\overline{E_N}$ να μην υπερβαίνει ποτέ το 6% για χωροχρονικά και σταθερά χωρικά δεδομένα. Όσον αφορά την εφαρμογή του μοντέλου μας σε δεδομένα τροχιάς, το μοντέλο παρουσίασε μέτριες τιμές των $\overline{ES_P}$ και $\overline{ES_N}$, δηλαδή μέσο απόλυτο σφάλμα σε κάθε επερώτηση, της τάξης του 40%. Ενώ εκ πρώτης όψεως το σφάλμα αυτό φαίνεται υψηλό, στην πράξη, οφείλεται στο μικρό αριθμό ανεξάρτητων τροχιών που χρησιμοποιούνται (βάση των τύπων, για τυπικά μεγέθη επερωτήσεων και ακτίνες αβεβαιότητας αναμένουμε $0.0004 \times N$ και $0.0025 \times N$ ως λανθασμένες θετικές / αρνητικές ανά επερώτηση). Λαμβάνοντας υπ' όψιν αυτές τις τιμές, είναι σαφές ότι για τυπικά μεγέθη επερωτήσεων και ακτίνες αβεβαιότητας, ο πληθυσμός του συνόλου δεδομένων θα πρέπει να είναι αρκετά μεγάλος ώστε να μας δώσει ένα σημαντικό αριθμό λανθασμένων απαντήσεων, κατάλληλο για την μέτρηση και τη σύγκριση σε σχέση με τα αποτελέσματα του προτεινόμενου μοντέλου. Συνεπώς, οι λεπτομέρειες του μοντέλου που αναπτύχθηκε εξετάσθηκαν περαιτέρω χρησιμοποιώντας μία σειρά συνθετικών και πραγματικών συνόλων χωρικών δεδομένων κατάλληλου πλήθους.

Όσον αφορά την εφαρμογή του μοντέλου σε συνθετικά (τυχαία) σύνολα χωρικών δεδομένων, η εκτίμηση του αριθμού των λανθασμένων απαντήσεων είναι ακριβής ανεξάρτητα της τιμής του μεγέθους επερώτησης και της ακτίνας του d του κύκλου αβεβαιότητας, ή του σ στην περίπτωση δεδομένων με κανονικά κατανεμημένη αβεβαιότητα. Τα πειράματα σε πραγματικά χωρικά δεδομένα αποδεικνύουν ακρίβεια ακόμα και υψηλότερη από αυτή που αναφέρεται για τα συνθετικά δεδομένα, με πολύ χαμηλά σφάλματα $\overline{ES_P}$ και $\overline{ES_N}$, που υποδηλώνει το πλεονέκτημα που προκύπτει από τη χρήση των ιστογραμμάτων, ακόμα και στην περίπτωση της μεταβλητής σ. Τα αποτελέσματα της εφαρμογής του προτεινόμενου μοντέλου σε χωρικούς κύβους δεδομένων και χωρικές λειτουργίες OLAP είναι επίσης πολλά υποσχόμενες. Τέλος, η εφαρμογή των προτεινόμενων λύσεων σε περιβάλλοντα του πραγματικού κόσμου (PostgreSQL [Post08a] με χωρική επέκταση PostGIS [Post08b]) απέδειξε την αποτελεσματικότητα της πρότασής μας όταν χρησιμοποιείται ως εκτιμητής (estimator), δεδομένου ότι ο χρόνος εκτέλεσης είναι τυπικά μόνο ελάχιστα milliseconds. Το

προτεινόμενο μοντέλο, πέρα από τις εφαρμογές του σε MODs, μπορεί να χρησιμοποιηθεί απευθείας σε υπάρχοντα συστήματα SDBMS για να δώσει στους χρήστες την ακρίβεια των αποτελεσμάτων χωρικών επερωτήσεων βάσει μόνο γνωστών χαρακτηριστικών του συνόλου δεδομένων και των επερωτήσεων· εξ' άλλου τα συνηθισμένα ιστογράμματα που ήδη χρησιμοποιούνται σε χωρικές βάσεις δεδομένων για σκοπούς βελτιστοποίησης επερωτήσεων, μπορούν να εξυπηρετήσουν το μοντέλο μας χωρίς καμία επιπλέον προσαρμογή.

Το τελευταίο θέμα αυτής της διατριβής είναι η διαχείριση του αποτελέσματος των αλγορίθμων συμπίεσης τροχιάς σε χωροχρονικές επερωτήσεις. Οι σχετικές εργασίες στον τομέα αυτό μέχρι τώρα, έχουν εστιάσει στην ανάπτυξη αλγορίθμων συμπίεσης τονίζοντας επιπλέον και το σφάλμα που εισάγεται στη θέση κάθε αντικειμένου λόγω συμπίεσης. Εμείς πάλι, στο Κεφάλαιο 6, αναγνωρίζοντας ότι οι χρήστες ενδιαφέρονται περισσότερο για το σφάλμα που εισάγεται λόγω της συμπίεσης σε χωροχρονικά *αποτελέσματα επερωτήσεων*, παρουσιάσαμε το πρώτο θεωρητικό μοντέλο που εκτιμά το σφάλμα αυτό στα αποτελέσματα επερωτήσεων χρονικής στιγμής. Δώσαμε ένα κλειστό τύπο του μέσου αριθμού λανθασμένων αποτελεσμάτων (λανθασμένα αρνητικά και λανθασμένα θετικά) που καλύπτει την περίπτωση των αυθαίρετα κατανεμημένων δεδομένων τροχιάς με διάφορες ταχύτητες, κατευθύνσεις κλπ. Προκύπτει ότι το σφάλμα εξαρτάται από το άθροισμα των απόλυτων τιμών της $\delta x_{i,k}$ και της $\delta y_{i,k}$ (δηλαδή της διαφοράς μεταξύ της συμπιεσμένης και της αρχικής τροχιάς, κατά μήκος των αξόνων *x*- και *y*- , αντίστοιχα) σε κάθε χρονικό αποτύπωμα t_k που η αρχική τροχιά έχει καταγράψει τη θέση της.

Επιπλέον, εκμεταλλευόμενοι τον τύπο που αναπτύχθηκε, στην παρούσα διατριβή, δίνουμε το έναυσμα για μία πρωτότυπη προσέγγιση που μπορεί να βελτιώσει την αποτελεσματικότητα των υφιστάμενων αλγορίθμων συμπίεσης τροχιάς. Δεδομένου ότι σύμφωνα με το μοντέλο, το σφάλμα εξαρτάται από τις απόλυτες τιμές της $\delta y_{i,k}$ και $\delta x_{i,k}$, η ελαχιστοποίηση του θα πρέπει να περιλαμβάνει την ελαχιστοποίηση της τιμής $|\delta x_{i,k}|+|\delta y_{i,k}|$, αντί της ελαχιστοποίησης της $SED_i(t_k)=\sqrt{\delta x_{i,k}^2+\delta y_{i,k}^2}$ που θεωρείται το κριτήριο βελτιστοποίησης στην πλειονότητα των υφιστάμενων αλγορίθμων συμπίεσης τροχιάς.

Στη συνέχεια αποδείξαμε ότι το μοντέλο μας εφαρμόζεται σε συνθήκες πραγματικής λειτουργίας – προκύπτει ότι η εκτίμηση των παραμέτρων του μοντέλου επιβαρύνει ελάχιστα τον αλγόριθμο συμπίεσης τροχιών – και κατόπιν παρουσιάσαμε την ακρίβεια των εκτιμήσεών μας, με ένα μέσο σφάλμα της τάξης του 6% βάσει διαφόρων συνόλων συνθετικών και πραγματικών δεδομένων. Συνεπώς αποδείχτηκε ότι το μοντέλο μας μπορεί να χρησιμοποιηθεί αμέσως μετά τη συμπίεση ενός συνόλου δεδομένων τροχιών για να δώσει στο χρήστη το μέσο σφάλμα που εισάγεται στα αποτελέσματα των χωροχρονικών επερωτήσεων διαφόρων μεγεθών (με ελάχιστη επιβάρυνση). Κατόπιν ο / η χρήστης θα μπορούσε να το χρησιμοποιήσει ως ένα επιπλέον κριτήριο για να αποφασίσει κατά πόσον τα συμπιεσμένα δεδομένα είναι κατάλληλα για τις ανάγκες του / της και ίσως να επιλέξει άλλους βαθμούς συμπίεσης κ.ο.κ.

7.2. Ανοικτά Θέματα

Αρκετοί ερευνητικοί τομείς παραμένουν ανοιχτοί στον τομέα της διαχείρισης δεδομένων τροχιάς. Στις επόμενες παραγράφους παρουσιάζουμε το μελλοντικό ερευνητικό έργο που πηγάζει απευθείας από την πρόοδό μας σε αυτή τη διατριβή.

Στον τομέα της δεικτοδότησης τροχιών, η τεχνολογία βάσεων δεδομένων έχει προχωρήσει τα τελευταία χρόνια, αναπτύσσοντας ευρετήρια που υπερβαίνουν την αποτελεσματικότητα των προτάσεων μας. Και τα δύο ευρετήρια που προτείνονται στην παρούσα διατριβή εισήχθησαν στα αρχικά της στάδια· εντωμεταξύ, άλλες δομές προτάθηκαν στην βιβλιογραφία που αποδείχτηκαν ότι είναι πιο αποδοτικές. Επί του παρόντος, ως τεχνολογία αιχμής για τη δεικτοδότηση τροχιών κινούμενων αντικειμένων θεωρούνται το PA-δέντρο [NR07] και το MON-δέντρο [AG05], για τροχιές που κινούνται σε απεριόριστο και περιορισμένο σε δίκτυο χώρο, αντιστοίχως. Από την άλλη πλευρά, σύμφωνα με τα αποτελέσματα της αντίστοιχης πειραματικής μας μελέτης στο Κεφάλαιο 2, καθώς και με τα αντίστοιχα αποτελέσματα που δημοσιεύτηκαν στην [AG05], οι δομές που εκμεταλλεύονται τον περιορισμό της κίνησης στο δίκτυο είναι πολύ πιο αποτελεσματικές από αυτές που δεικτοδοτούν αντικείμενα στον απεριόριστο χώρο· στην πράξη, η πρώτη συνήθως υπερέχει της δεύτερης κατά τάξεις μεγέθους. Ωστόσο, καμία από τις προτεινόμενες δομές δεικτοδότησης με περιορισμούς στο δίκτυο δεν είναι σχεδιασμένη για τη διατήρηση τροχιών: και το FNR και το MON-δέντρο εξ' ορισμού δεν διαθέτουν ένα μηχανισμό για την ανάκτηση τροχιών και ασχολούνται μόνο με την επεξεργασία των επερωτήσεων που βασίζονται στις συντεταγμένες. Ακόμα και το SETI, που είναι ένα από τα πιο αποτελεσματικά σχήματα δεικτοδότησης σε απεριόριστο χώρο σχετικά με επερωτήσεις που βασίζονται στις συντεταγμένες, πάσχει από το ίδιο μειονέκτημα. Ωστόσο, όπως επισημάναμε στο Κεφάλαιο 2, η διατήρηση της τροχιάς αποτελεί προϋπόθεση για την επεξεργασία επερωτήσεων βασισμένες σε αυτή. Άρα, η πρώτη ερευνητική κατεύθυνση που προκύπτει από το θέμα της δεικτοδότησης τροχιών είναι η ανάπτυξη μεθόδων προσπέλασης που υποστηρίζουν αποτελεσματικά τις επερωτήσεις βασισμένες στη τροχιά και στον περιορισμένο και στον απεριόριστο χώρο.

Όσον αφορά την προηγμένη επεξεργασία επερωτήσεων, η πρότασή μας δίνει τη δυνατότητα σε δομές τύπου R-δέντρου να υποστηρίξουν αποτελεσματικά αλγορίθμους αναζήτησης NN· αφ' ετέρου, κανένα από τα προτεινόμενα χωροχρονικά ευρετήρια, πέρα από τις δομές τύπου R-δέντρου, δεν εξετάζει τους αλγορίθμους αναζήτησης NN. Ωστόσο, για κάποια από αυτά (π.χ. το FNR-δέντρο), η επερώτηση NN μπορεί κατά πάσα πιθανότητα να υποστηριχτεί. Μία πρώτη ιδέα σε αυτό το θέμα είναι ότι επειδή στο FNR-δέντρο το δίκτυο δεικτοδοτείται από ένα συμβατικό R-δέντρο, ο αλγόριθμος «πρώτα στον καλύτερο» που περιγράφεται στην [HS99] μπορεί να χρησιμοποιηθεί για να βρεθεί ο χωρικός πλησιέστερος γείτονας· έτσι, δεδομένου ότι τα γραμμικά τμήματα του δικτύου (δηλαδή τα χωρικά στοιχεία των τμημάτων τροχιάς) αναφέρονται με αύξουσα σειρά της απόστασής τους από το αντικείμενο της επερώτησης, ο αλγόριθμος θα πρέπει να αναφέρει τέτοια πλησιέστερα τμήματα μέχρι να ανακτήσει το πρώτο που αλληλεπικαλύπτεται με την επερώτηση στη χρονική διάσταση· μία παρόμοια προσέγγιση μπορεί να χρησιμοποιηθεί στο MON-δέντρο.

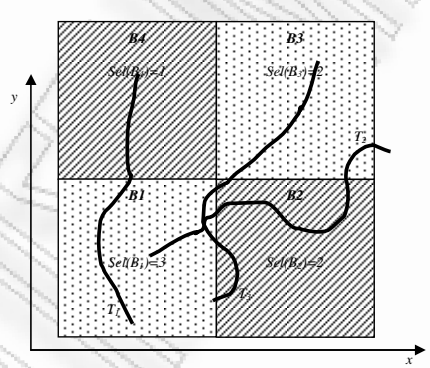
Οι μελλοντικές εργασίες στην προηγμένη επεξεργασία επερωτήσεων περιλαμβάνει επίσης την ανάπτυξη αλγορίθμων για την υποστήριξη επερωτήσεων χωρικής σύνδεσης («*βρείτε ζεύγη αντικειμένων που πέρασαν εγγύτερα το ένα στο άλλο (ή εντός κάποιας απόστασης d μεταξύ των) κατά τη*

διάρκεια ενός χρονικού διαστήματος και / ή ενός συγκεκριμένου χωρικού περιορισμού») και επερωτήσεις *MST ανεξάρτητες του χρόνου* (*Time-Relaxed MST*) σε τροχιές χρησιμοποιώντας την προτεινόμενη μετρική *DISSIM*. Αυτός ο τύπος επερώτησης υπολογίζει τη μέγιστη ανομοιότητα μεταξύ τροχιών ανεξάρτητα από τη χρονική στιγμή στην οποία ξεκινά το αντικείμενο της επερώτησης. Οι αλγόριθμοι αυτοί θα πρέπει να εξετάζουν τροχιές που δεικτοδοτούνται από δομές τύπου R-δέντρου, που είναι και τα πιο ευρέως διαδεδομένα ευρετήρια τροχιάς. Όμως, η πιο πολλά υποσχόμενη μελλοντική εργασία λαμβάνει υπ' όψιν τη χρήση της μετρικής *DISSIM* μαζί με τις τεχνικές διάταξης και κλαδέματος που αναπτύχθηκαν σε αυτή τη διατριβή, ώστε να υποστηριχτεί αποτελεσματικά η αναζήτηση εύρους ομοιότητας, ένας τύπος επερώτησης που έχει εξαιρετικές εφαρμογές στο πεδίο της εξόρυξης δεδομένων. Πιο συγκεκριμένα, δεδομένου ότι η εφαρμογή του γενικού, βασισμένου στη πυκνότητα, αλγορίθμου συσταδοποίησης OPTICS [ABKS99] βάσει της μετρικής *DISSM* όπως προτείνεται στην [NP06] απαιτεί την εύρεση, για κάθε τροχιά στο σύνολο δεδομένων, του αριθμού των τροχιών που είναι εγγύτερα (ήτοι, ομοιότερα) από μία δεδομένη τιμή της απόστασης (ομοιότητας), η εξαντλητική αναζήτηση που χρησιμοποιείται στην [NP06] αποδεικνύεται ότι είναι μία πάρα πολύ δαπανηρή διαδικασία. Ωστόσο, υπό αυτές τις συνθήκες, μία μέθοδος που βασίζεται στα R-δέντρα για επερωτήσεις εύρους ομοιότητας, όπως αυτή που παρουσιάστηκε στη παρούσα διατριβή θα βελτιώσει σημαντικά την απόδοση σε σχέση με εναλλακτικές στρατηγικές δεικτοδότησης και επερωτήσεων.

Τέλος, οι μελλοντικές εργασίες στο τομέα της προηγμένης επεξεργασίας επερωτήσεων θα πρέπει να περιλαμβάνουν την ανάπτυξη μοντέλων κόστους για επερωτήσεις πλησιέστερου γείτονα [TZPM04] και πιο όμοιας τροχιάς σε βάσεις δεδομένων ιστορικών τροχιών. Κατά τον ίδιο τρόπο, θα πρέπει να αναπτυχθούν και οι μηχανισμοί για την εκτίμηση της επιλεκτικότητας επερωτήσεων για σκοπούς βελτιστοποίησης επερωτήσεων, επενδύοντας στα όσα παρουσιάστηκαν στην [TSP03] για χωροχρονικές επερωτήσεις πρόβλεψης.

Ένα παράπλευρο αποτέλεσμα της παρούσας διατριβής, που παρουσιάζεται στην Ενότητα 5.4.2 είναι η ανάπτυξη ενός χωροχρονικού ιστογράμματος που βασίζεται σε υπάρχουσες προσεγγίσεις των χωρικών βάσεων δεδομένων [APR99], για την υποστήριξη της εκτίμησης επιλεκτικότητας για επερωτήσεις χρονικής στιγμής. Αφ' ετέρου, η εκτίμηση του αριθμού των διακριτών τροχιών, για γενικές επερωτήσεις εύρους (δηλαδή με χρονική έκταση $\neq 0$), δεν είναι καθόλου εύκολη, δεδομένου ότι εμπλέκει το γνωστό *πρόβλημα διακριτής μέτρησης πλήθους (distinct-counting problem)* [TKC+04]. Το πρόβλημα της διακριτής μέτρησης πλήθους εμφανίζεται όταν ένα αντικείμενο καταγράφει τη θέση του σε διάφορα χρονικά αποτυπώματα εντός ενός δεδομένου παραθύρου επερώτησης, που σημαίνει ότι υπολογίζεται πολλαπλά στο αποτέλεσμα της επερώτησης. Η [TKC+04] παρέχει μία λύση για το προαναφερθέν πρόβλημα ολοκληρώνοντας χωροχρονικά ευρετήρια με σκίτσα (sketches), τα οποία χρησιμοποιούνται παραδοσιακά για την προσεγγιστική επεξεργασία επερωτήσεων. Ωστόσο, η πρότασή της μειώνει τις απαιτήσεις χώρου μόνο λίγες φορές (τυπικά, περίπου στο 40% του αρχικού μεγέθους της βάσης δεδομένων), ενώ η αντίστοιχη δομή ευρετηρίου διατηρείται στο δίσκο. Σαφώς, μια τέτοια προσέγγιση δεν μπορεί να χρησιμοποιηθεί αντί των ιστογραμμάτων (που έχουν τυπικό μέγεθος μερικών KB [APR99]), δεδομένου ότι εισάγει σημαντική επιβάρυνση τόσο όσον αφορά την απαιτούμενη μνήμη όσο και όσον αφορά τις απαιτήσεις σε χρόνο επεξεργασίας.

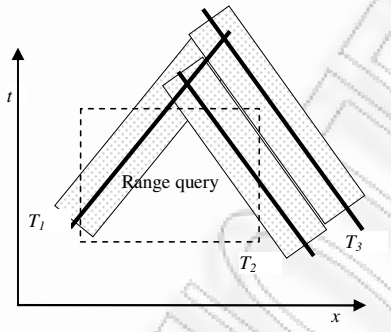
Κατά τον ίδιο τρόπο, ένα χωροχρονικό ιστόγραμμα για τον αριθμό των διακριτών τροχιών, θα πρέπει να διαμερίσει κάθε τροχιά σε αρκετούς χωροχρονικούς κάδους, μετρώντας τον αριθμό των διακριτών τροχιών εντός του κάθε κάδου. Ωστόσο, όταν προσπαθούμε να έχουμε μια εκτίμηση της επιλεκτικότητας ενός παραθύρου επερώτησης που περιλαμβάνει περισσότερους το ενός κάδους, η εκτίμηση αυτή δεν μπορεί να υπολογιστεί ως το άθροισμα του πλήθους των κάδων διότι οι τροχιές μπορούν να υπολογιστούν αρκετές φορές ανάλογα με τον αριθμό των κάδων με τους οποίους αλληλεπικαλύπτονται. Το Σχήμα 7.1 δίνει ένα παράδειγμα του προβλήματος, παρουσιάζοντας τέσσερις κάδους ενός ιστογράμματος (B_1, B_2, B_3, B_4) μαζί και με την επιλεκτικότητά τους $Sel(B_i)$: η συνολική επιλεκτικότητα και των τεσσάρων κάδων $Sel\left(\bigcup B_i\right) = 3$ απέχει πολύ από το να είναι το άθροισμα του $\sum Sel\left(B_i\right) = 7$ διότι οι τροχιές T_1, T_2, T_3 θα μετρηθούν όσες φορές, όσες και οι κάδοι με τους οποίους αλληλεπικαλύπτεται η κάθε μία τους. Επιπλέον, το ίδιο πρόβλημα ανακύπτει κατά την κατασκευή ιστογραμμάτων που ακολουθούν τη μεθοδολογία που εισάγεται στην [APR99] για απλά χωρικά ιστογράμματα: ο αλγόριθμος κατασκευής αρχικά υπολογίζει τον αριθμό διακριτών αντικειμένων εντός του κάθε κελιού που παράγεται από ένα πυκνό χωρικό πλέγμα και κατόπιν σε κάθε επανάληψη συγκεντρώνει ομάδες κελιών για να σχηματίσει ευρύτερους κάδους βάσει της ευριστικής *MinSkew*. Ωστόσο, κατά τη συγκέντρωση αυτή, ο αριθμός των τροχιών εντός του κάθε κάδου που προκύπτει πρέπει να υπολογιστεί, σαφώς, όχι ως το άθροισμα των τροχιών που περιέχεται εντός κάθε θεμελιώδους κελιού.



Σχήμα 7.1. Το πρόβλημα της διακριτής συνάθροισης τροχιών σε ιστογράμματα τροχιάς

Όσον αφορά στο θέμα της διαχείρισης αβεβαιότητας, υπάρχουν αρκετές ενδιαφέρουσες ερευνητικές κατευθύνσεις που προκύπτουν από τα όσα παρουσιάστηκαν σε αυτή τη διατριβή, συμπεριλαμβανομένης και της εφαρμογής του μοντέλου μας σε χώρους δεδομένων υψηλότερης διάστασης και της επέκτασής του για να υποστηρίξει τις γενικές χωροχρονικές επερωτήσεις εύρους (ήτοι με χρονική έκταση ≠ 0), μη σημειακά σύνολα δεδομένων και μη ορθογώνια παράθυρα επερωτήσεων καθώς και επερωτήσεις πλησιέστερου γείτονα. Η πλειονότητα των προαναφερθέντων ερευνητικών κατευθύνσεων απαιτεί σημαντική προσπάθεια. Μεταξύ αυτών το πρώτο που πρέπει να εξεταστεί στα πλαίσια των χωροχρονικών βάσεων δεδομένων είναι η επέκτασή τους στην περίπτωση των γενικών επερωτήσεων εύρους. Κάτι τέτοιο δεν είναι καθόλου εύκολο· ωστόσο στη συνέχεια παρέχουμε στοιχεία προς αυτή την κατεύθυνση. Έστω για παράδειγμα το Σχήμα 7.2 που παρουσιάζει τροχιές των τριών κινούμενων αντικειμένων κατά μήκος των περιοχών αβεβαιότητας τους (ήτοι οι διάστικτες επιφάνειες) στο χώρο *x*-*t*, κατά μήκος μιας επερώτησης εύρους. Για λόγους απλότητας όλες

οι τροχιές παρουσιάζονται ως γραμμικά τμήμα χωρίς να χάνεται η γενικότητα. Οι Τροχιές T_1 και T_2 δεν μπορούν καν να δοθούν ως λανθασμένο αποτέλεσμα σχετικά με το παράθυρο επερώτησης λόγω του γεγονότος ότι για τουλάχιστο μία χρονική στιγμή η περιοχή αβεβαιότητάς τους βρίσκεται πλήρως εντός του παραθύρου. Αφ' ετέρου, η τροχιά T_3 μπορεί να επιστραφεί ως λανθασμένο αποτέλεσμα επειδή δεν βρίσκεται εντός του παραθύρου επερώτησης· ωστόσο, η περιοχή αβεβαιότητας της το διασχίζει. Γενικεύοντας την παραπάνω παρατήρηση, μπορούμε να δηλώσουμε ότι μόνο τα αντικείμενα των οποίων η περιοχή αβεβαιότητας διασχίζει το παράθυρο επερώτησης χωρίς να βρίσκεται πλήρως εντός του σε οποιαδήποτε χρονική στιγμή, μπορεί να συνεισφέρει στον αριθμό των λανθασμένων απαντήσεων στα αποτελέσματα της επερώτησης.



Σχήμα 7.2. Το αποτέλεσμα της αβεβαιότητας σε γενικές επερωτήσεις εύρους

Το τελευταίο θέμα που εξετάστηκε σε αυτή τη διατριβή, ήτοι η συμπίεση τροχιάς, μας δίνει επίσης αρκετές ενδιαφέρουσες κατευθύνσεις, συμπεριλαμβανομένης και της ανάπτυξης των αντιστοίχων του παρουσιασθέντος μοντέλου για επερωτήσεις πλησιέστερου γείτονα ή ακόμα περισσότερο, γενικών χωροχρονικών επερωτήσεων εύρους. Πιο συγκεκριμένα, η επέκταση της προσέγγισης μας προς τη δεύτερη κατεύθυνση, απαιτεί να προσδιορίσουμε το σχήμα του χωροχρονικού διαστήματος εντός του οποίου η κάτω αριστερά γωνία της επερώτησης εύρους (ήτοι, το ελάχιστο σημείο της επερώτησης εύρους) πρέπει να βρεθεί προκειμένου η συμπιεσμένη τροχιά να ανακτηθεί ως λανθασμένο αποτέλεσμα (αρνητικό ή θετικό), σύμφωνα με το Σχήμα 6.6, το Σχήμα 6.7 και στη συνέχεια να καθορίσουμε τον όγκο του βάσει της Εξ.(6.4). Παρόλο που αυτός ο όγκος μπορεί να υπολογιστεί όταν οι δx_i και δy_i εκφράζονται ως απλές συναρτήσεις (δηλαδή μεταξύ διαδοχικών χρονικών αποτυπωμάτων), στη γενική περίπτωση όπου οι δx_i και δy_i εκφράζονται ως πολλαπλές συναρτήσεις (δηλαδή διαφορετικές συναρτήσεις σε διαφορετικά γραμμικά τμήματα αρχικής τροχιάς), ο αντίστοιχος όγκος είναι πολύ δύσκολο να προσδιορισθεί. Παρόλα αυτά, δεν παύει ν' αποτελεί εξαιρετική πρόκληση για το μέλλον.

Τέλος, σκοπεύουμε να εφαρμόσουμε τα όσα αντιληφθήκαμε σχετικά με το κρίσιμο κριτήριο βελτιστοποίησης των αλγορίθμων συμπίεσης τροχιάς, ώστε να δώσουμε μια νέα προσέγγιση που να βελτιώνει την αποτελεσματικότητα των υφιστάμενων λύσεων. Αυτή η βελτίωση θα μετρηθεί σε όρους βαθμού συμπίεσης ως προς τον αριθμό των λανθασμένων απαντήσεων που εισάγονται σε χωροχρονικές επερωτήσεις λόγω της συμπίεσης, αντίθετα προς τις υπάρχουσες προσεγγίσεις που τη μετρούν σε όρους μέσου σφάλματος που εισάγεται στη θέση της κάθε τροχιάς [MB04].

8. Αναφορές

[ABKS99] Ankerst, M., Breunig, M., Kriegel, H.,P., and Sander, J.: OPTICS: Ordering Points To Identify the Clustering Structure. *Proceedings of ACM SIGMOD*, 1999

[ACNV99] Arcieri, F., Cammino, C., Nardelli, E., and Venza, A. The Italian Cadastral Information System: a Real-Life Spatio-Temporal DBMS. *Proceedings of STDM* 1999

[AFH02] Agarwal, P. K., Flato, E., Halperin, D., Polygon decomposition for efficient construction of Minkowski sums, *Computational Geometry*, 21(1-2): 39-61 (2002)

[AFS93] Agrawal, R., Faloutsos, C., and Swami, A., Efficient Similarity Search in Sequence Databases. *Proceedings of FODO*, 1993.

[AG05] Guting, R., H., Almeida, V., T., Indexing the Trajectories of Moving Objects in Networks. *GeoInformatica* 9(1):33-60, 2005.

[AGB06] Almeida, V., T., Guting, R., H., and Behr, T. Querying Moving Objects in SECONDO. *Proceedings of MDM*, 2006

[APR99] Acharya, S., Poosala, V., and Ramaswamy, S., Selectivity Estimation in Spatial Databases. *Proceedings of ACM SIGMOD*, 1999.

[BC96] Berndt, J. and Clifford, J., Finding patterns in time series: A dynamic programming approach. Advances in Knowledge Discovery and Data Mining. *AAAI/MIT Press*, 1996

[BJKS02] Benetis, R., Jensen, C., Karciauskas, G., and Saltenis, S., Nearest Neighbor and Reverse Nearest Neighbor Queries for Moving Objects. *Proceedings of IDEAS*, 2002.

[BKSS90] Beckmann, N., Kriegel, H.P., Schneider, R., and Seeger, B. The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. *Proceedings of ACM SIGMOD*, 1990.

[BS03] Beresford, A. R., and Stajano, F. Location Privacy in Pervasive Computing. *IEEE Pervasive Computing*, 2(1):46-55, 2003.

[Bri02] Brinkhoff, T.: A Framework for Generating Network-Based Moving Objects, *Geoinformatica*, 6(2):153-180, 2002

[BW01] Babu, S., and Widom, J., Continuous Queries over Data Streams, *SIGMOD Record*, 30(3):109-120, 2001.

[CC02] Choi, Y.-J., and Chung, C.-W., Selectivity estimation for spatio-temporal queries to moving objects. *Proceedings of ACM SIGMOD*, 2002

[CC07] Chen, J. and Cheng, R., Efficient Evaluation of Imprecise Location-Dependent Queries. *Proceedings of IEEE ICDE*, 2007

[CEP03] Chakka, V.P., Everspaugh, A. and Patel, J., Indexing Large Trajectory Data Sets with SETI. *Proceedings of CIDR*, 2003.

[CF98] Cheung, K.L., and Fu, A.,W., Enhanced Nearest Neighbour Search on the R-tree. *SIGMOD Record*, 27(3):16-21, 1998

[CF99] Chan, K.P., and Fu, A.W-C., Efficient time series matching by Wavelets. *Proceedings of ICDE*, 1999.

[CKP04] Cheng, R., Kalashnikov, D., and Prabhakar, S., Querying Imprecise Data in Moving Object Environments. *IEEE TKDE* 16(9):1112-1127, 2004.

[CN04] Cai, Y., and Ng, R., Indexing spatio-temporal trajectories with Chebyshev polynomials. *Proceedings of ACM SIGMOD*, 2004.

[COO05] Chen, L., Tamer Özsu, M., and Oria, V., Robust and Fast Similarity Search for Moving Object Trajectories. *Proceedings of ACM SIGMOD*, 2005.

[CPZ97] Ciaccia, P., Patella, M., and Zezula, P. M-tree: An efficient access method for similarity search in metric spaces. *Proceedings of VLDB*, 1997

[CR99] Chomicki, J. and Revesz, P., A Geometric Framework for Specifying Spatio-temporal Objects. *Proceedings of TIME*, 1999.

[CWT03] Cao, H., Wolfson, O., and Trajcevski, G., Spatio-temporal Data Reduction with Deterministic Error Bounds. *Proceedings of DIALM–POMC*, 2003.

[CXP+04] Cheng, R., Xia, Y., Prabhakar, S., Shah, R., and Vitter, J.S., Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data. *Proceedings of VLDB*, 2004

[CZBP06] Cheng, R., Zhang, Y., Bertino, E., and Prabhakar, S. Preserving user location privacy in mobile data management infrastructures. *Proceedings of the 6th Workshop on Privacy Enhancing Technologies*, 2006.

[DP73] Douglas, D. H., Peucker, T. K., Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer* 10 (1973) 112–122.

[DYM+05] Dai, X., Yiu, M.L., Mamoulis, N., Tao, Y., and Vaitis, M., Probabilistic Spatial Queries on Existentially Uncertain Data. *Proceedings of SSTD*, 2005.

[EGSV99] Erwig, M. Güting, R. H., Schneider, M., and Varzigiannis, M., Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. *GeoInformatica* 3(3): 265-291, 1999

[FGNS00] L. Forlizzi, R. H. Güting, E. Nardelli, and M. Schneider, A Data Model and Data Structures for Moving Objects Databases. *Proceedings of ACM SIGMOD*, 2000.

[FGPT05] Frentzos, E., Gratsias, K., Pelekis, N., and Theodoridis, Y., Nearest Neighbor Search on Moving Object Trajectories. *Proceedings of SSTD,* 2005.

[FGPT07] Frentzos, E., Gratsias, K., Pelekis, N., and Theodoridis, Y., Algorithms for Nearest Neighbor Search on Moving Object Trajectories. *Geoinformatica* 11(2): 159-193 (2007)

[FGT07] Frentzos, E., Gratsias, K., and Thedoridis, Y., Index-based Most Similar Trajectory Search. *Proceedings of ICDE*, 2007

[FGT08] Frentzos, E., Gratsias, K., and Thedoridis, Y., On the Effect of Uncertainty in Spatial Querying, *IEEE TKDE, accepted*, 2008

[Fre02] Frentzos, E., Spatio-temporal Indexing Techniques. *MSc thesis, National Technical Univesity of Athens*, 2003. Available at http://isl.cs.unipi.gr/db/people/efrentzo (στα Ελληνικά).

[Fre03] Frentzos, E., Indexing objects moving on fixed networks. *Proceedings of SSTD*, 2003.

[FT06] Frentzos, E. and Theodoridis, Y., The TB*-tree: Indexing Moving Object Trajectories in Real-World Environments. *UNIPI-ISL-TR-2006-02, Technical Report Series, University of Piraeus*, 2006. Available at http://isl.cs.unipi.gr/db/people/efrentzo.

[FT07] Frentzos, E., and Theodoridis, Y., On the Effect of Trajectory Compression in Spatio-temporal Querying. *Proceedings of ADBIS*, 2007.

[GBE+00] Guting, R., H., Bohlen, M., H., Erwig, M., Jensen, C., S., Lorentzos, N., A., Schneider, M., and Vazirgiannis, M., A Foundation for Representing and Querying Moving Objects. *ACM TODS*, 25(1): 1-42, 2000.

[GL05] Gedik, B., and Liu, L. A customizable k-anonymity model for protecting location privacy. *Proceedings of ICDCS*, 2005.

[GS05] Güting, R.H., and Schneider, M., *Moving Objects Databases*. Morgan Kaufmann Publishers, 2005.

[Gut84] Guttman, A.: R-Trees: A Dynamic Index Structure for Spatial Indexing. *Proceedings of ACM SIGMOD*, 1984.

[HKT03] Hadjieleftheriou, M., Kollios, G., and Tsotras, V., Performance Evaluation of Spatio-temporal Selectivity Estimation Techniques. *Proceedings of SSDBM*, 2003

[HKTG02] Hadjieleftheriou, M., Kollios, G., Tsotras, V. J., and Gunopulos, D., Efficient Indexing of Spatio-temporal Objects. *Proceedings of EDBT*, 2002.

[HKTG06] M., Hadjieleftheriou, G., Kollios, V., Tsotras, and D., Gunopulos, Indexing Spatio-temporal Archives. *The VLDB Journal, to appear*

[HS92] Hershberger, J., Snoeyink, J.: Speeding up the Douglas-Peucker line-simplification algorithm. *Proceeedings of SDH,* 1992.

[HS99] Hjaltason, G., and Samet, H., Distance Browsing in Spatial Databases, *ACM TODS*, 24(2): 265-318, 1999.

[HXL05] Hu, H., Xu, J., and Lee, D.L., A Generic Framework for Monitoring Continuous Spatial Queries over Moving Objects. *Proceedings of ACM SIGMOD*, 2005.

[Ioa93] Ioannidis, Y., Universality of Serial Histograms. *Proceedings of VLDB*, 1993.

[Ioa07] Ioannidis, Y., Emerging Open Agoras of Data and Information. *Proceedings of ICDE,* 2007

[IP95] Ioannidis, Y. and Poosala, V., Balancing histogram optimality and practicality for query result size estimation. *Proceedings of ACM SIGMOD*, 1995.

[ISS03] Iwerks, G.S., Samet, H., and Smith, K., Continuous K-Nearest Neighbor Queries for Continuously Moving Points with Updates. *Proceedings of VLDB*, 2003.

[Keo02] Keogh, E., Exact indexing of dynamic time warping. *Proceedings of VLDB*, 2002.

[KF93] Kamel, I., and Faloutsos, C.: On Packing R-trees. *Proceedings of CIKM*, 1993.

[KGT99] Kollios, G., Gunopulos, D., and Tsotras, V. On Indexing Mobile Objects. *Proceedings of ACM PODS*, 1999.

[KWX+06] Keogh, E., Wei, L., Xi, X., Lee, S.H., and Vlachos, M., LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures. *Proceedings of VLDB*, 2006.

[Lei95] A. Leick, *GPS satellite surveying*, John Wiley and Sons, New York, 1995.

[LS05] Lin, B., and Su, J., Shapes Based Trajectory Queries for Moving Objects. *Proceedings of ACM-GIS*, 2005.

[MB04] Meratnia, N., By, R., Spatio-temporal Compression Techniques for Moving Point Objects. *Proceedings of EDBT*, 2004.

[MFN+08] Marketos, G., Frentzos, E., Ntoutsi, I., Pelekis, N., Raffaeta, A., and Theodoridis, Y., Building Real-World Trajectory Warehouses. *Proceedings of MobiDE*, 2008

[MHP05] Mouratidis K., Hadjieleftheriou M., Papadias, D., Conceptual Partitioning: An Efficient Method for Continuous Nearest Neighbor Monitoring. *Proceedings of ACM SIGMOD*, 2005.

[MNPT05] Manolopoulos, Y., Nanopoulos, A., Papadopoulos, A.N., and Theodoridis, Y., *R-trees: Theory and Applications.* Springer-Verlag, 2005

[MXA04] Mokbel, M.F., Xiong, X., and Aref, W.G., SINA: Scalable Incremental Processing of Continuous Queries in Spatio-temporal Databases. *Proceedings of ACM SIGMOD*, 2004.

[NP06] Nanni, M., and Pedreschi, D., Time-focused density-based clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 27(3):267–289, 2006.

[NR07] Ni, Y., and Ravishankar, C., Indexing Spatio-temporal Trajectories with Efficient Polynomial Approximations, *IEEE TKDE*, 19(5): 663-678, 2007.

[NRB03] Ni, J., Ravishankar, C.V., and Bhanu, B., Probabilistic Spatial Database Operations. *Proceedings of SSTD*, 2003.

[NST99] Nascimento, M., Silva, J.R.O., and Theodoridis, Y. Evaluation of Access Structures for Discretely Moving Points. *Proceedings of STDM*, 1999

[PFGT08] Pelekis, N., Frentzos, E., Giatrakos, N. and Theodoridis, Y., Aggregative LBS via a Trajectory DB Engine. *Proceedings of ACM SIGMOD*, 2008 (to appear)

[Pfo02] Pfoser, D, Indexing the Trajectories of Moving Objects. *IEEE DE Bulletin*, 25(2):2-9, 2002.

[PKM+07] Pelekis, N., Kopanakis, I., Marketos, G., Ntoutsi, I., Andrienko, G., and Theodoridis, Y., Similarity Search in Trajectory Databases. *Proceedings of TIME*, 2007

[PJ99] Pfoser, D. and Jensen, C.S., Capturing the Uncertainty of Moving-Object Representations, *Proceedings of SSD*, 1999

[PJ01] Pfoser, D., and Jensen, C.S., Querying the trajectories of on-line mobile objects. *Proceedings of MobiDE*, 2001

[PJ03] Pfoser, D., and Jensen, C.S., Indexing of network constrained moving objects. *Proceedings of ACM-GIS*, 2003

[PJT00] Pfoser D., Jensen C. S., and Theodoridis, Y., Novel Approaches to the Indexing of Moving Object Trajectories. *Proceedings of VLDB*, 2000.

[Post08a] PostGIS, URL: http://postgis.refractions.net (accessed 15 May 2008)

[Post08b] PostgreSQL, URL: http://www.postgresql.org (accessed 15 May 2008)

[PPS06] Potamias, M., Patroumpas, K. and Sellis, T., Sampling Trajectory Streams with Spatio-temporal Criteria. *Proceedings of SSDBM*, 2006.

[PPS06a] Potamias, M., Patroumpas, K. and Sellis, T., Amnesic Online Synopses for Moving Objects. *Proceedings of CIKM*, 2006.

[PPS07] Potamias, M., Patroumpas, K. and Sellis, T., Online Amnesic Summarization of Streaming Locations. *Proceedings of SSTD*, 2007.

[PT06] Pelekis N., Theodoridis Y. Boosting Location-Based Services with a Moving Object Database Engine. *Proceedings of MobiDE*, 2006.

[PTJ05] Pfoser, D., Tryfona, N., and Jensen, C.S., Indeterminacy and Spatio-temporal Data: Basic Definitions and Case Study, *GeoInformatica* 9(3): 211-236, 2005.

[PTKZ02] Papadias, D., Tao, Y., Kalnis., P., and Zhang, J.: Indexing Spatio-Temporal Data Warehouses. *Proceedings ICDE*, 2002.

[PTVP06] Pelekis N., Theodoridis Y., Vosinakis S., and Panayiotopoulos T.. Hermes - A Framework for Location-Based Data Management. *Proceedings of EDBT*, 2006.

[RKV95] Roussopoulos, N., Kelley, S., and Vincent, F., Nearest Neighbor Queries. *Proceedings of ACM SIGMOD,* 1995.

[SJ02] Saltenis, S. and Jensen, C. S., Indexing of Moving Objects for Location-Based Services. *Proceedings of ICDE*, 2002.

[SJLL00] Saltenis, S., Jensen, C. S., Leutenegger, S. and Lopez, M., Indexing the Positions of Continuously Moving Objects. *Proceedings of ACM SIGMOD*, 2000.

[SKS03] Shahabi, C., Kolahdouzan, M., and Sharifzadeh, M., A Road Network Embedding Technique for K-Nearest Neighbor Search in Moving Object Databases, *GeoInformatica*, 7(3): 255-273, 2003.

[SR01] Song, Z., and Roussopoulos, N., K-Nearest Neighbor Search for Moving Query Point. *Proceedings of SSTD*, 2001.

[SYF05] Sakurai, Y., Yoshikawa, M., and Faloutsos, C., FTW: Fast Similarity Search under the Time Warping Distance. *Proceedings of PODS*, 2005.

[TCX+05] Tao, Y., Cheng, R., Xiao, X., Ngai, W.K., Kao, B., and Prahbakar, S., Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions. *Proceedings of VLDB*, 2005.

[The03] Theodoridis, Y., Ten Benchmark Database Queries for Location-based Services. The *Computer Journal* 46(6): 713-725, 2003.

[TKC+04] Tao, Y., Kollios, G., Considine, J., Li, F., and Papadias, D., Spatio-Temporal Aggregation Using Sketches. *Proceedings of ICDE*, 2004

[TP01] Tao, Y., and Papadias, D., MV3R-Tree: A Spatio-Temporal Access Method for Timestamp and Interval Queries. *Proceedings of VLDB*, 2001

[TP02] Tao, Y., and Papadias, D., Time Parameterized Queries in Spatio-Temporal Databases, *Proceedings of ACM SIGMOD*, 2002.

[TPS02] Tao, Y., Papadias, D., and Shen, Q., Continuous Nearest Neighbor Search. *Proceedings of VLDB*, 2002.

[TPS03] Tao, Y., Papadias, D., and Sun, J., An optimized Spatio-temporal Access Method for Predictive Queries. *Proceedings of VLDB*, 2003

[Tra03] Trajcevski, G., Probabilistic Range Queries in Moving Objects Databases with Uncertainty. *Proceedings of MobiDE,* 2003.

[TS96] Theodoridis, Y., and Sellis, T., A Model for the Prediction of R-tree Performance. *Proceedings of ACM PODS*, 1996.

[TSN99] Theodoridis, Y., Silva, J. R. O., and Nascimento, M. A., On the Generation of Spatio-temporal Datasets. *Proceedings of SSD,* 1999.

[TSP03] Tao, Y., Sun, J., and Papadias, D., Analysis of predictive spatio-temporal queries. *ACM TODS*, 28(4):295-336, 2003

[TVS96] Theodoridis, Y., Vazirgiannis, M., and Sellis, T., Spatio-temporal Indexing for Large Multimedia Applications. *Proceedings of ICMCS*, 1996.

[TWZC02] Trajcevski, G., Wolfson, O., Zhang, F., and Chamberlain, S., The geometry of uncertainty in moving objects databases. *Proceedings of EDBT*, 2002.

[TWHC04] Trajcevski, G., Wolfson, O., Hinrichs, K. and Chamberlain, S. Managing uncertainty in moving objects databases, *ACM Trans., Database Systems,* 29(3), 463-507, 2004.

[TZPM04] Tao, Y., Zhang, J., Papadias, D., and Mamoulis, N., An Efficient Cost Model for Optimization of Nearest Neighbor Search in Low and Medium Dimensional Spaces, *IEEE TKDE* 16(10):1169-1184, 2004

[VGD04] Vlachos, M., Gunopulos, D., and Das, G., Rotation Invariant Distance Measures for Trajectories. *Proceedings of SIGKDD*, 2004.

[VKG02] Vlachos, M., Kollios, G., and Gunopulos, D., Discovering Similar Multidimensional Trajectories. *Proceedings of ICDE*, 2002.

[WD04] Worboys, M., and Duckham, K., *GIS: A Computing Perspective*. CRC Press, 2004

[WSCY99] Wolfson, O., Sistla, P.A., Chamberlain, S., and Yesha, Y., Updating and Querying Databases that Track Mobile Units. *Distributed and Parallel Databases*, 7(3):257-387, 1999.

[XMA05] Xiong, X., Mokbel, M., and Aref, W., SEA-CNN: Scalable Processing of Continuous K-Nearest Neighbor Queries in Spatio-temporal Databases. *Proceedings of ICDE*, 2005.

[XP03] Yuni Xia, Sunil Prabhakar: Q+Rtree: Efficient Indexing for Moving Object Database. *Proceedings of DASFAA*, 2003

[YAS03] Yanagisawa, Y., Akahani, J., and Satoh, T., Shape-Based Similarity Query for Trajectory of mobile Objects. *Proceedings of MDM*, 2003.

[YM03] Yu, X., and Mehrotra, S., Capturing Uncertainty in Spatial Queries over Imprecise Data. *Proceedings of DEXA*, 2003

[YPK05] Yu, X., Pu, K., and Koudas, N., Monitoring k-Nearest Neighbor Queries Over Moving Objects. *Proceedings of ICDE*, 2005.

[ZG02] J. Zhang and M. Goodchild. *Uncertainty in Geographical Information*. Taylor & Francis, 2002.

[ZSI02] Zhu, J, Su, J. and Ibarra, O., Trajectory queries and octagons in moving object databases. *Proceedings of CIKM*, 2002.