ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Τεχνικές Αποθήκευσης Δεδομένων & Εξόρυξης Γνώσης για Βάσεις Κινούμενων Αντικειμένων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΓΕΡΑΣΙΜΟΣ Δ. ΜΑΡΚΕΤΟΣ

Πτυχίο Πληροφορικής, Πανεπιστήμιο Πειραιώς (2003)
MSc in Information Systems Engineering, UMIST (2004)

Πειραιάς, Δεκέμβριος 2009

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Συμβουλευτική Επιτροπή:

Επιβλέπων:

Ιωάννης Θεοδωρίδης
Αν. Καθηγητής Πανεπιστημίου Πειραιώς

Μέλη:

Γεώργιος Βασιλακόπουλος
Καθηγητής Πανεπιστημίου Πειραιώς

Δημήτριος Δεσπότης
Καθηγητής Πανεπιστημίου Πειραιώς

Διατριβή
για την απόκτηση Διδακτορικού
Διπλώματος του Τμήματος
Πληροφορικής

ΓΕΡΑΣΙΜΟΥ Δ. ΜΑΡΚΕΤΟΥ

"Τεχνικές Αποθήκευσης Δεδομένων & Εξόρυξης Γνώσης για Βάσεις Κινούμενων Αντικειμένων"

Εξεταστική Επιτροπή:

Ιωάννης Μανωλόπουλος
Καθηγητής Αριστοτέλειου
Πανεπιστημίου Θεσσαλονίκης

Θεμιστοκλής Παναγιωτόπουλος
Καθηγητής Πανεπιστημίου Πειραιώς

Ιωάννης Κωτίδης
Επ. Καθηγητής Οικονομικού
Πανεπιστημίου Αθηνών

Δέσποινα Πολέμη
Επ. Καθηγήτρια Πανεπιστημίου
Πειραιώς

.....
ΓΕΡΑΣΙΜΟΣ Δ. ΜΑΡΚΕΤΟΣ

Πρόλογος

Η ανάλυση δεδομένων κίνησης που συλλέγονται από συσκευές εντοπισμού θέσης προσφέρει δυνατότητες ανακάλυψης προτύπων συμπεριφοράς που μπορούν να αξιοποιηθούν σε ένα πλήθος εφαρμογών. Οι Τεχνολογίες Άμεσης Αναλυτικής Επεξεργασίας (OnLine Analytical Processing - OLAP) και οι τεχνικές Εξόρυξης Γνώσης μπορούν να χρησιμοποιηθούν για την μετατροπή αυτού του μεγάλου αριθμού πρωτογενών-ακατέργαστων δεδομένων σε πολύτιμη γνώση. Παρόλο που η εφαρμογή των τεχνικών αυτών σε συμβατικά δεδομένα έχει μελετηθεί σε μεγάλο βαθμό την τελευταία δεκαετία, ο μεγάλος όγκος των παραγόμενων δεδομένων κίνησης καθώς και η χωροχρονική τους φύση αποτελούν τις κύριες προκλήσεις για τη χρήση τέτοιων μεθόδων ανάλυσης. Στη διατριβή αυτή θα παρουσιάσουμε την πρόταση μας για ένα ολοκληρωμένο πλαίσιο για Αποθήκευση και Εξόρυξη Γνώσης από Δεδομένα Κίνησης (Mobility Data Warehousing and Mining) που αποτελείται από διάφορα συστατικά (στη πραγματικότητα, βήματα μιας Διαδικασίας Ανακάλυψης Γνώσης). Πιο συγκεκριμένα, προτείνουμε τεχνικές για Αποθήκευση Δεδομένων Τροχιών Κινούμενων Αντικειμένων δίνοντας έμφαση σε θέματα μοντελοποίησης, ETL διαδικασιών (ανακατασκευή τροχιών, τροφοδότηση κύβου) και σε OLAP λειτουργίες (συσσώρευση κτλ). Επίσης προτείνουμε τεχνικές Εξόρυξης Γνώσης που αξιοποιούν τα δεδομένα κίνησης και εξάγουν α) πρότυπα αλληλεπίδρασης που επιτρέπουν την χωροχρονική αναπαράσταση, σύνθεση και κατηγοριοποίηση των τροχιών κινούμενων αντικειμένων και β) πρότυπα κυκλοφορίας που μας βοηθούν να κατανοήσουμε με ποιον τρόπο διαχέεται η κυκλοφορία σε ένα δίκτυο.

Γεράσιμος Μαρκέτος

Δεκέμβριος 2009

Ευχαριστίες

Θα ήθελα πρώτα απ' όλα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Γιάννη Θεοδωρίδη για την υποστήριξη και καθοδήγηση που μου παρείχε κατά τη διάρκεια της εκπόνησης αυτής της διατριβής. Οι γνώσεις του, οι προτάσεις του, η συνεργασία μας σε ερευνητικά έργα και οι εποικοδομητικές συζητήσεις που είχαμε ήταν πολύτιμα στοιχεία για την εκπόνηση της. Ευχαριστώ ακόμη τους καθ. Γεώργιο Βασιλακόπουλο, Δημήτριο Δεσπότη, Ιωάννη Μανωλόπουλο, Θεμιστοκλή Παναγιωτόπουλο, Ιωάννη Κωτίδη και Δέσποινα Πολέμη που δέχτηκαν οι δυο πρώτοι να είναι μέλη 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 για τη 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 ΠΕΝΕΔ), από τα οποία είχα και οικονομική υποστήριξη. Ευχαριστώ τους υπευθύνους και τους διαχειριστές των παραπάνω προγραμμάτων.

Θα ήθελα επίσης να ευχαριστήσω τους γονείς μου και τον αδερφό μου για την ενθάρρυνση και υποστήριξη που μου παρείχαν στα χρόνια των σπουδών μου. Τέλος, θα ήθελα να ευχαριστήσω τη Γεωργία για τη συνεχή συμπαράσταση και την ανεξάντλητη υπομονή που έδειξε όλα αυτά τα χρόνια.

Πίνακας Περιεχομένων

Κατάλογος Πινάκων

Κατάλογος Εικόνων

1. Εισαγωγή

Στο κεφάλαιο αυτό αναλύεται το υπόβαθρο της διατριβής και περιγράφονται τα κύρια σημεία της δομής της. Στην Ενότητα 1.1 παρουσιάζονται βασικές θεωρητικές αρχές για την ανάλυση δεδομένων κινούμενων αντικειμένων, τα κίνητρα εκπόνησης της συγκεκριμένης διατριβής καθώς και ενδιαφέροντα παραδείγματα ανάπτυξης εφαρμογών. Στην Ενότητα 1.2 σκιαγραφείται η συμβολή της διατριβής στο χώρο των βάσεων δεδομένων και στην Ενότητα 1.3 περιγράφεται συνοπτικά η θεματολογία των επόμενων κεφαλαίων.

1.1. Αναλύοντας Δεδομένα Κίνησης

Οι καθημερινές μας δραστηριότητες, ο τρόπος που ζούμε και κινούμαστε, αφήνουν «ψηφιακά» ίχνη σε πληροφοριακά συστήματα. Αυτό συμβαίνει διότι χρησιμοποιούμε κινητά τηλεφώνα και άλλες συσκευές εντοπισμού θέσης για να επικοινωνούμε και να λαμβάνουμε πληροφορίες πλοήγησης. Στην πραγματικότητα, παίρνοντας πληροφορίες για ίχνη μέσω αυτών των συσκευών μπορούμε να αντιληφθούμε τις κινήσεις των αντικειμένων σε μια περιοχή, η ανάλυση των οποίων έχει μεγάλη αξία. Οι δυνατότητες για συλλογή μεγαλύτερου όγκου πληροφοριών, οι τεχνολογίες διάχυτου υπολογισμού (pervasive computing) και η αύξηση στον βαθμό ακρίβειας προσδιορισμού θέσεων αυτών των «ψηφιακών ιχνών» καθιστούν ακόμα σημαντικότερη την ανάλυση των δεδομένων αυτών.

Για το 2005, ο αριθμός χρηστών κινητών τηλεφώνων παγκόσμια ανερχόταν πάνω από 2 δισεκατομμύρια, που σημαίνει ότι ένα στα τρία άτομα κατείχε κινητό τηλέφωνο [Cia09]. Επιπλέον, τεχνολογίες προσδιορισμού τοποθεσίας, όπως GSM και UMTS, που χρησιμοποιούνται σήμερα από τους ασύρματους τηλεφωνικούς δέκτες καθιστούν ικανή την επίτευξη ακριβέστερων εκτιμήσεων για τον εντοπισμό της θέσης της συσκευής, ενώ παρατηρείται εξέλιξη των διαφόρων τεχνολογιών εντοπισμού θέσης [GPT08]: Συσκευές εξοπλισμένες με GPS δέκτες μπορούν να μεταδίδουν πληροφορίες θέσης σε κάποιον πάροχο υπηρεσίας. Επιπλέον, οι τελευταίες τεχνολογικές εξελίξεις, όπως Wi-Fi και Bluetooth συσκευές αποτελούν πηγή δεδομένων για εντοπισμό εσωτερικών θέσεων (indoor positioning) ενώ, το Wi-Max αποτελεί εναλλακτική για εντοπισμό εξωτερικών θέσεων (outdoor positioning) [GP07].

Σύγχρονες συσκευές επικοινωνίας είναι εκτενώς διαδεδομένες και μπορούν να μεταφέρονται παντού από άτομα και οχήματα. Αυτό έχει ως αποτέλεσμα να καθίσταται αντιληπτή η ανθρώπινη δραστηριότητα μέσα σε ένα χώρο – όχι απαραίτητα σκοπίμως, αλλά απλά ως δευτερεύον αποτέλεσμα των παρεχόμενων υπηρεσιών στους χρήστες αυτών των συσκευών. Το γεγονός αυτό μας επιτρέπει να θεωρήσουμε το δίκτυο ασύρματης τηλεφωνίας ως υποδομή για τη συλλογή δεδομένων κινούμενων

αντικειμένων με σκοπό την ανάλυση για την απόκτηση γνώσης των ανθρώπινων κινήσεων. Είναι ξεκάθαρο, ότι στις διάφορες υποθέσεις μας η ιδιωτικότητα είναι ένα θέμα που πρέπει με κάθε τρόπο να διασφαλιστεί [GP07]. Συγκεκριμένα, πώς θα μπορούσαν οι τροχιές των κινούμενων αντικειμένων να αποθηκεύονται και να αναλύονται χωρίς να παραβιάζονται θέματα ιδιωτικότητας; Πώς είναι δυνατό, από δεδομένα τροχιών να εξάγουμε πρότυπα κίνησης προστατεύοντας την ιδιωτικότητα των χρηστών;

Η χρήση συσκευών με δυνατότητα εντοπισμού θέσης, όπως τα κινητά τηλέφωνα και οι συσκευές που είναι εξοπλισμένες με GPS δέκτες, μας επιτρέπουν την πρόσβαση σε μεγάλο όγκο χωροχρονικών δεδομένων. Η χωρική και χρονική φύση των δεδομένων συνεπάγεται τη δημιουργία μεγάλου όγκου χωροχρονικών δεδομένων και θέτει νέες προκλήσεις για τα αναλυτικά εργαλεία που μπορούν να χρησιμοποιηθούν για να μετατρέψουν τα πρωτογενή δεδομένα σε γνώση. Μέρος αυτής της διατριβής είναι και η διερεύνηση δυνατοτήτων προέκτασης των παραδοσιακών μεθόδων ανάλυσης ώστε να είναι δυνατή η εφαρμογή σε δεδομένα κινούμενων αντικειμένων.

Η ανάλυση τέτοιων δεδομένων κίνησης εγείρει προοπτικές για την ανακάλυψη προτύπων συμπεριφοράς με σημαντική εφαρμογή στο χώρο της «κινητής διαφήμισης» (mobile marketing) και στη διαχείριση της κυκλοφοριακής κίνησης. Τεχνολογίες Άμεσης Αναλυτικής Επεξεργασίας (OnLine Analytical Processing - OLAP) και τεχνικές εξόρυξης γνώσης (Data Mining – DM) μπορούν να χρησιμοποιηθούν για τη μετατροπή αυτού του μεγάλου αριθμού πρωτογενών δεδομένων σε πολύτιμη γνώση. Η εφαρμογή των τεχνικών αυτών σε συμβατικά δεδομένα έχει μελετηθεί σε μεγάλο βαθμό την τελευταία δεκαετία. Ο μεγάλος όγκος των παραγόμενων δεδομένων κίνησης αποτελεί πρόκληση για τη χρήση τέτοιων μεθόδων ανάλυσης. Ωστόσο για την ορθή εφαρμογή των μεθόδων αυτών, πρέπει να ληφθεί υπόψη η πολύπλοκη φύση των χωροχρονικών δεδομένων την οποία πρέπει να αξιοποιούν αποτελεσματικά όποιες νέες μέθοδοι αναπτυχθούν.

1.1.1. Κίνητρα και Σενάρια Εφαρμογών

Οι ερευνητικοί στόχοι που παρουσιάζονται σε αυτή τη διατριβή έχουν προκύψει από σενάρια που αφορούν ανθρώπινη μετακίνηση και διαχείριση μεταφορικών μέσων, θέματα εμπνευσμένα από την έρευνα που έγινε στα πλαίσια του ερευνητικού έργου Geographic Privacy-aware Knowledge Discovery and Delivery (GeoPKDD) [Geo06]. Τα διαφορετικά χαρακτηριστικά τους αποδεικνύουν τη χρησιμότητα της έρευνας για την αξιοποίηση δεδομένων γεωγραφικής θέσης. Το πρώτο σενάριο εφαρμογής αφορά ένα εκπαιδευτικό παιχνίδι μετακινήσεων σε μια πόλη και το δεύτερο μία αποτελεσματική υπηρεσία για τη διαχείριση της κυκλοφοριακής κίνησης στο δίκτυο μιας πόλης.

Όσον αφορά τον πρώτο τομέα επιρροής, ας υποθέσουμε ότι μια εταιρία διαφήμισης ενδιαφέρεται για την ανάλυση δεδομένων κινούμενων αντικειμένων σε διαφορετικές περιοχές μιας πόλης για να μπορέσει να πάρει απόφαση σχετικά με διαφημίσεις που προορίζονται να τοποθετηθούν σε δρόμους (τοποθετημένες σε πίνακες ανάρτησης διαφημίσεων). Συγκεκριμένα, ενδιαφέρεται για δεδομένα που αφορούν τα δημογραφικά χαρακτηριστικά των ατόμων που επισκέπτονται διαφορετικές περιοχές της πόλης σε διαφορετικές ζώνες ωρών της ημέρας για να μπορέσουν να αποφασίσουν για την ορθή διαδοχή των διαφημίσεων στις διαφορετικές ζώνες ωρών. Η γνώση αυτή κάνει δυνατή την εκτέλεση πιο εστιασμένων εκστρατειών μάρκετινγκ και την εφαρμογή μιας πιο αποτελεσματικής στρατηγικής.

Ακόμη μια ενδιαφέρουσα εφαρμογή αυτού του τομέα επιρροής θα μπορούσε να αποτελέσει ένα σενάριο σχεδιασμού ψυχαγωγικής δραστηριότητας (recreational planning) που μπορεί να παρουσιαστεί μέσα από ένα παιχνίδι. Ένα τυπικό εκπαιδευτικό παιχνίδι είναι το «κυνήγι θησαυρού» (παιχνίδι ιχνηλασίας), ένα παλιό παιδικό παιχνίδι όπου βάση ερωτήσεων και γρίφων γραμμένων πάνω σε ένα χαρτί εξερευνάται μια περιοχή. Κάθε ομάδα προσπαθεί να βρει τις τοποθεσίες που υποδεικνύονται στο ερωτηματολόγιο. Όταν το μέρος βρεθεί, προσπαθούν να απαντήσουν σωστά και γρήγορα στις ερωτήσεις που είναι γραμμένες στο χαρτί, να σημειώσουν την απάντηση και να προχωρήσουν στο επόμενο σημείο ενδιαφέροντος. Το παιχνίδι αποτελείται κυρίως από μια σειρά σημείων ελέγχου που είναι συνδεδεμένα με γρίφους πολυμέσων.

Η συσκευή του παίκτη περιλαμβάνει ένα δέκτη GPS (Global Positioning System) όπου καταγράφει συνεχώς την τρέχουσα θέση του. Καθώς τα σημεία ελέγχου είναι εφοδιασμένα με στοιχεία εγγύτητας, ο εξυπηρετητής δικτύου (server) ενεργοποιείται κάθε φορά που συμβαίνει ένα γεγονός και ο παίκτης πλησιάζει ένα από τα εικονικά σημεία ελέγχου. Τότε, ο εξυπηρετητής δικτύου «απαντά» στέλνοντας πληροφορίες για τον αντίστοιχο γρίφο, ο οποίος παρουσιάζεται μέσω πολυμέσων στον χρήστη.

Κάθε γρίφος συνδέεται με μέσα όπως εικόνες ή άλλες πληροφορίες που απαιτούνται για την επίλυση του και επιτρέπει την αλληλεπίδραση από τον χρήστη. Ο παίκτης προσπαθεί όχι μόνο να βρει τη σωστή λύση στο γρίφο αλλά και να απαντήσει το γρηγορότερο δυνατόν, γιατί ο απαιτούμενος χρόνος για την επίλυση όλων των γρίφων συγκεντρώνεται και προστίθεται στην τελική βαθμολογία. Η απάντηση στον γρίφο διαβιβάζεται στον εξυπηρετητή του παιχνιδιού (game server). Είναι εφικτό ο παίκτης να αλληλεπιδρά με το σύστημα αλλά επίσης και με άλλα άτομα, για παράδειγμα ζητώντας βοήθεια για την επίλυση του γρίφου.

Λόγω της γενικότερης έλλειψης γνώσης γύρω από θέματα προτύπων συμπεριφοράς σε ένα χώρο αναψυχής (recreational site), επιλέγουμε να εστιάσουμε σε δυο κύριες προϋποθέσεις που πρέπει να καλύπτονται:

- *Σημεία Ενδιαφέροντος* σε ένα χώρο αναψυχής: Η παραδοχή είναι ότι οι δραστηριότητες του παιχνιδιού δεν εκτελούνται πάντα σε προσχεδιασμένες ζώνες ψυχαγωγίας.
- *Παρόμοιες αλληλεπιδράσεις χρηστών* σε ένα χώρο αναψυχής: Η παραδοχή είναι ότι η συμπεριφορά των ομάδων χρηστών εξαρτάται μερικώς από την αλληλεπίδραση με τους άλλους. Για παράδειγμα, μπορεί να διαπιστωθεί ότι ορισμένοι παίκτες έχοντας διαφορετική συμπεριφορά συγκρούονται μεταξύ τους (για ομάδες που διαγωνίζονται μεταξύ τους).

Πιο συγκεκριμένα, οι σχεδιαστές ψυχαγωγικών δραστηριοτήτων πρέπει να:

- κατανοούν τις πραγματικές τροχιές που ακολουθούν οι παίκτες, όχι μόνο τα σημεία εκκίνησης και τερματισμού,
- εκτιμούν την ροή μεταξύ των διαφόρων σημείων ενδιαφέροντος,
- αντιλαμβάνονται την αλληλεπίδραση και τις κινήσεις των παικτών και να προβλέπουν πιθανούς νέους γρίφους,

- αναγνωρίζουν τις πιθανές αλληλεπιδράσεις των παικτών και τον τρόπο απόκρισης όταν συμβεί κάποιο αναπάντεχο γεγονός,
- ανακαλύψουν εναλλακτικές τοποθεσίες για το ψυχαγωγικό παιχνίδι,
- διαπιστώσουν επικίνδυνη και ύποπτη συμπεριφορά σε ένα ψυχαγωγικό παιχνίδι.

Όσον αφορά το σενάριο για τη *διαχείριση της κυκλοφορίας μέσων μεταφοράς,* η κατανόηση, διαχείριση και πρόβλεψη του κυκλοφοριακού φαινομένου σε μια πόλη είναι ταυτόχρονα ενδιαφέρουσα και χρήσιμη. Για παράδειγμα, οι τοπικές αρχές θα μπορούσαν, μελετώντας την ροή της κίνησης, να βελτιώσουν τις υπάρχουσες κυκλοφοριακές συνθήκες, να αντιδράσουν αποτελεσματικά σε περιπτώσεις μη ομαλής ροής της κυκλοφορίας και να προγραμματίσουν τη διάνοιξη νέων δρόμων, την επέκταση και βελτίωση υφιστάμενων και την τοποθέτηση φωτεινών σηματοδοτών. Επιπλέον, η μελέτη των αλληλεπιδράσεων μεταξύ των αντικειμένων σε χωροχρονικά γειτνιάζουσες περιοχές είναι ουσιαστική για την απόκτηση πολύτιμης γνώσης σχετικά με τις συμπεριφορές κίνησης.

Ένας ενδιαφέρων τελικός στόχος θα μπορούσε να είναι η ανάπτυξη ενός εργαλείου υποστήριξης αποφάσεων για τη διαχείριση της κυκλοφοριακής κίνησης, αναλύοντας προηγούμενες κινήσεις και συμπεριφορές ατόμων με τη χρήση δεδομένων που έρχονται από τις κινητές συσκευές τους. Αυτή η υπηρεσία απευθύνεται σε τμήματα αστικού σχεδιασμού, κινητούς πράκτορες (mobility agents), διαχειριστές κυκλοφορίας, αλλά επίσης και στη δημόσια διοίκηση που είναι υπεύθυνη για τη κυκλοφορία τόσο σε αστικό όσο και σε περιφερειακό επίπεδο. Οι ανάγκες που αυτή η υπηρεσία στοχεύει να ικανοποιήσει είναι:

- η αναγνώριση και παρατήρηση των μεταβολών της ροής του χρήστη σε γεωγραφικές περιοχές σύμφωνα με αλλαγές στο αστικό περιβάλλον σε διαφορετικά χρονικά διαστήματα,
- η γνώση του πραγματικού μέσου χρόνου για τη μετακίνηση μεταξύ διαφορετικών περιοχών,
- η ταυτοποίηση των πιο δημοφιλών αντιπροσωπευτικών τροχιών.

Τα δεδομένα που θα καταχωρηθούν για αυτή την υπηρεσία μπορεί να διαφέρουν πολύ και πολύ διαφορετικές λειτουργίες μπορούν να υλοποιηθούν ανάλογα με τις διαθέσιμες πληροφορίες. Σε αυτή την περίπτωση, τόσο τύπου GSM όσο και τύπου GPS δεδομένα μπορούν να χρησιμοποιηθούν, μαζί με τα ήδη διαθέσιμα δεδομένα από τις κάμερες διαχείρισης κυκλοφορίας και τους αισθητήρες κίνησης.

Η λειτουργικότητα ενός τέτοιου εργαλείου περιλαμβάνει:

- την αυτόματη δόμηση και κατασκευή μιας μήτρας προέλευσης-προορισμού (source-destination matrix), για την εκτίμηση της ροής της κυκλοφορίας από την μια περιοχή στην άλλη, τόσο σε αστικό όσο και σε περιφερειακό επίπεδο,
- τον υπολογισμό ενός μέσου χρόνου ταξιδιού από τη μία ζώνη στην άλλη, τόσο σε αστική όσο και σε περιφερειακή κλίμακα,
- την προσομοίωση της ροής της κυκλοφορίας σε περίπτωση παρουσίας έκτακτων γεγονότων, όπως ποδοσφαιρικών αγώνων, απεργιών ή συναυλιών.

Τα παραπάνω σενάρια μπορούν να πραγματοποιηθούν χρησιμοποιώντας καινοτόμες τεχνικές ανάλυσης που θα μπορούν να αξιοποιήσουν την πλούσια σημασιολογία των κινούμενων αντικειμένων. Ενδεικτικά, μια Αποθήκη Δεδομένων Τροχιών Κινούμενων Αντικειμένων (Trajectory Data Warehouse – TDW) μπορεί να συντελέσει στην ανάλυση διάφορων μέτρων όπως ο αριθμός των κινούμενων αντικειμένων (άτομα, οχήματα) σε διαφορετικές αστικές περιοχές, η μέση ταχύτητα των οχημάτων (ή ατόμων), οι αυξομειώσεις στην ταχύτητα των οχημάτων καθώς επίσης και χρήσιμη μετα-πληροφορία, όπως για παράδειγμα οι πιο συνηθισμένες διαδρομές. Επιπλέον, τεχνικές εξόρυξης γνώσης που βασίζονται σε δεδομένα τροχιών μπορούν να χρησιμοποιηθούν για την ανακάλυψη προτύπων. Τα πρότυπα αυτά μπορούν να εκφραστούν π.χ. μέσω σχέσεων μεταξύ των ακμών ενός οδικού δικτύου μιας πόλης. Με άλλα λόγια, τα τμήματα των δρόμων που συνεισφέρον στη ροή και ο τρόπος με τον οποίο αυτό γίνεται. Τελικά, η αναπαράσταση της κίνησης σαν μια σειρά από περιγραφικά μέσα, μάς δίνει χρήσιμες γνώσεις σχετικά με τη συμπεριφορά των αντικειμένων καθώς και την αλληλεπίδραση με τα γειτονικά τους αντικείμενα.

1.2. Συνεισφορά της Διατριβής

Αυτή η διατριβή προτείνει καινοτόμες τεχνικές ανάλυσης με στόχο την εξαγωγή χρήσιμων προτύπων από χωροχρονικά δεδομένα. Αρχικά, συζητάει τη διαφορά μεταξύ των δυο τύπων χωροχρονικών δεδομένων: δεδομένων κίνησης και στατικών δεδομένων. Προκειμένου να αποσαφηνίσουμε τη διαφορά μεταξύ τους, επιλέγουμε τα σεισμολογικά δεδομένα ως μια τυπική περίπτωση στατικών (χωροχρονικών) δεδομένων και παρουσιάζουμε ένα *Σύστημα Διαχείρισης και Εξόρυξης Γνώσης Σεισμολογικών Δεδομένων* για γρήγορη και εύκολη συλλογή, επεξεργασία και οπτικοποίηση. Το προτεινόμενο πρωτόλειο σύστημα περιλαμβάνει, μεταξύ άλλων, μια σεισμολογική βάση δεδομένων για αποτελεσματική και αποδοτική απάντηση ερωτημάτων και μια σεισμολογική αποθήκη δεδομένων για OLAP ανάλυση και εξόρυξη γνώσης. Παρέχουμε κάποιες βασικές αρχές για αυτά τα δυο συστατικά καθώς και παραδείγματα λειτουργικότητας προκειμένου να υποστηριχθούν αποφάσεις. Τα αποτελέσματα της έρευνας μας σε αυτό το χώρο παρουσιάζονται στο Κεφάλαιο 2 και δημοσιεύονται στην [MTK08].

Το κυρίως μέρος της έρευνας μας εστιάζει σε τεχνικές αποθήκευσης δεδομένων και εξόρυξης γνώσης που μπορούν να εφαρμοστούν σε βάσεις δεδομένων κινούμενων αντικειμένων. Πιο συγκεκριμένα, αυτή η διατριβή προτείνει ένα ολοκληρωμένο πλαίσιο για Αποθήκευση και Εξόρυξη Δεδομένων Κίνησης (Mobility Data Warehousing and Mining) που αποτελείται από διάφορα συστατικά (στη πραγματικότητα, βήματα μιας Διαδικασίας Ανακάλυψης Γνώσης).

Αμέσως μετά, συζητάμε τις συνεισφορές της παρούσας διατριβής ομαδοποιημένες κατά θεματική περιοχή. Θα πρέπει να σημειώσουμε ότι η καινοτομία της κάθε μας προσέγγισης τεκμηριώνεται σε κάθε κεφάλαιο παρουσιάζοντας τις σχετικές εργασίες.

Αποθήκες Δεδομένων Τροχιών Κινούμενων Αντικειμένων: Διερευνάται η προσαρμογή του παραδοσιακού μοντέλου κύβου δεδομένων στα πλαίσια μιας αποθήκης δεδομένων τροχιών κινούμενων αντικειμένων, με στόχο τη μετατροπή πρωτογενών δεδομένων θέσης σε χρήσιμες

πληροφορίες. Πιο συγκεκριμένα εστιάζουμε στα παρακάτω σημεία που είναι κρίσιμα για την αποθήκευση δεδομένων τροχιών κινούμενων αντικειμένων:

- την ETL διαδικασία που τροφοδοτεί μια αποθήκη δεδομένων με συγκεντρωτικά δεδομένα τροχιών. Για αυτόν τον σκοπό, προτείνουμε δυο εναλλακτικές ETL διαδικασίες: μια (βασισμένη σε ευρετήρια) που είναι *προσανατολισμένη στα κελιά* του κύβου και μια (χωρίς ευρετήριο) που είναι *προσανατολισμένη στις τροχιές*. Όπως θα παρουσιάσουμε στην πειραματική μας μελέτη, η επιλογή μεταξύ των δυο μεθόδων είναι συνάρτηση μεταξύ του επιλεγμένου βαθμού κλιμάκωσης (granularity) και του αριθμού των τροχιών,
- τον προϋπολογισμό των μέτρων του κύβου για σκοπούς OLAP ανάλυσης. Η πρόκληση που πρέπει να αντιμετωπιστεί είναι ότι μια τροχιά μπορεί να διατρέχει πολλά κελιά στον κύβο προκαλώντας εμπόδια στις διαδικασίες συνάθροισης. Επεκτείνουμε το μοντέλο του κύβου δεδομένων προσθέτοντας κάποια βοηθητικά μέτρα που θα μας βοηθήσουν να διορθώσουμε τα σφάλματα που προκαλούνται από τις διπλομετρήσεις κατά τη διάρκεια της συσσώρευσης (roll-up). Πρόκειται για μια προσεγγιστική μέθοδο η οποία όμως φαίνεται να είναι αποδοτική.

Και στις δυο παραπάνω περιπτώσεις, παρέχουμε σχεδιαστικές λύσεις και ελέγχουμε το κατά πόσο είναι εφαρμόσιμες και αποδοτικές σε πραγματικές συνθήκες. Τα αποτελέσματα των ερευνών μας σε αυτό το χώρο παρουσιάζονται στο Κεφάλαιο 3 και δημοσιεύονται στις [MFN+08a], [MFN+08b] και [MT09b].

Ειδικό (ad-hoc) OLAP σε δεδομένα τροχιών: Παρουσιάζουμε μια νέα προσέγγιση στο σχεδιασμό ενός κύβου τροχιών με στόχο την παροχή απαντήσεων σε εξειδικευμένα ερωτήματα λαμβάνοντας υπόψη διαφορετικούς ορισμούς της έννοιας της τροχιάς (trajectory). Ένας ευέλικτος κύβος τροχιών που παρέχει εξειδικευμένη ανάλυση μπορεί να εξυπηρετεί μια σειρά από εφαρμογές ακόμη και αν αυτές απαιτούν διαφορετικούς ορισμούς της έννοιας της τροχιάς. Πιο αναλυτικά:

- Επεκτείνουμε το OLAP μοντέλο δεδομένων ώστε να περιλαμβάνει έναν ευέλικτο πίνακα συμβάντων (fact table) που μπορεί να απαντήσει ερωτήματα λαμβάνοντας υπόψη διαφορετικούς ορισμούς της έννοιας της τροχιάς και παρέχοντας την επιλογή να οριστεί η κατάλληλη σημασιολογία κατά τη διάρκεια υποβολής ενός ερωτήματος πάνω σε δεδομένα τροχιών.
- Αναπτύσσουμε μια ETL τεχνική που μετασχηματίζει κατάλληλα τα δεδομένα ώστε να τα φορτώσει στον πίνακα συμβάντων. Αυτή η τεχνική αξιοποιεί το νέο OLAP μοντέλο δεδομένων και τροφοδοτεί μια κατάλληλα σχεδιασμένη Αποθήκη Δεδομένων για Τροχιές Κινούμενων Αντικειμένων. Όπως αποδεικνύουμε στο πειραματικό σκέλος αυτής της έρευνας, η απόδοση της συγκεκριμένης τεχνικής είναι καλύτερη από αυτή του συμβατικού ETL.
- Εμπλουτίζουμε τις OLAP τεχνικές ώστε να αξιοποιούν το νέο OLAP μοντέλο δεδομένων. Για αυτό τον σκοπό, προτείνουμε έναν ανταγωνιστικό αλγόριθμο που μπορεί να απαντήσει εξειδικευμένα (ad-hoc) ερωτήματα συνάθροισης. Επίσης, συζητούμε θέματα προϋπολογισμού που βελτιώνουν την απόδοση του αλγορίθμου καθώς συντομεύουν τη διαδικασία υπολογισμού.

Τα προκαταρκτικά αποτελέσματα αποδεικνύουν την ορθότητα και αποδοτικότητα της προσέγγισης μας. Τα αποτελέσματα της έρευνας μας σε αυτό το χώρο παρουσιάζονται στο Κεφάλαιο 3 και έχουν υποβληθεί για αξιολόγηση [MT09a].

Εξόρυξη Προτύπων Αλληλεπίδρασης: Προτείνουμε ένα ανταγωνιστικό πλαίσιο για την εξόρυξη προτύπων αλληλεπίδρασης επιτρέποντας την χωροχρονική αναπαράσταση, σύνθεση και κατηγοριοποίηση τροχιών κινούμενων αντικειμένων. Αυτή η προσέγγιση εισάγει δυο βασικές ιδέες:

- προκειμένου να κατανοήσουμε τι συμβαίνει σε ένα συγκεκριμένο αντικείμενο θα πρέπει να κοιτάξουμε όχι μόνο την τροχιά του αλλά και το περιβάλλον μέσα στο οποίο κινείται·
- αυτό το περιβάλλον ορίζεται όχι μόνο από το γεωγραφικό χώρο αλλά και από την παρουσία άλλων αντικειμένων και την αλληλεπίδραση τους με το αντικείμενο που μας ενδιαφέρει.

Μοντελοποιούμε την αλληλεπίδραση ως ένα ομαδικό φαινόμενο στα πλαίσια του οποίου προσπαθούμε να κατανοήσουμε τη συμπεριφορά ενός αντικειμένου σε συνάρτηση με τη «γειτονιά» στην οποία βρίσκεται. Αυτό επιτυγχάνεται υπολογίζοντας ενδιαφέροντας περιγραφείς αλληλεπίδρασης (interaction descriptors) που μπορούν να μας βοηθήσουν να κατανοήσουμε την κίνηση σε διάφορα χωροχρονικά παράθυρα. Συγκρίνουμε εναλλακτικές διαδικασίες: από τη μια μεριά θεωρούμε για κάθε αντικείμενο μια δυναμικά οριζόμενη γειτονιά και από την άλλη, ένα σταθερό πλέγμα και υπολογίζουμε τους περιγραφείς για σταθερά χωροχρονικά παράθυρα. Τα πρώτα πειραματικά αποτελέσματα δείχνουν ότι η προσέγγιση μας είναι εφαρμόσιμη και αποτελεσματική. Τα μέχρι τώρα αποτελέσματα της έρευνας μας στο συγκεκριμένο χώρο έχουν υποβληθεί για αξιολόγηση [NMO09].

Εξόρυξη Προτύπων Κυκλοφορίας: Προτείνουμε ένα πλαίσιο εξόρυξης γνώσης από δεδομένα κυκλοφορίας που βρίσκονται αποθηκευμένα είτε σε μια Βάση Δεδομένων Κινούμενων Αντικειμένων είτε σε μια Αποθήκη Δεδομένων Τροχιών Κινούμενων Αντικειμένων. Παρουσιάζουμε μέτρα ομοιότητας μεταξύ των χρονοσειρών κυκλοφορίας και προτείνουμε έναν αλγόριθμο συσταδοποίησης που ομαδοποιεί τις ακμές του οδικού δικτύου με βάση αυτά τα μέτρα. Επίσης, ορίζουμε σχέσεις μεταξύ των ακμών που συνδυάζουν τις χρονικές πληροφορίες που παρέχονται από τις χρονοσειρές κυκλοφορίας με χωρικές πληροφορίες που προκύπτουν από το ίδιο το δίκτυο. Παρουσιάζουμε έναν αλγόριθμο ανακάλυψης τέτοιων σχέσεων μεταξύ των ακμών του οδικού δικτύου για συγκεκριμένες χρονικές περιόδους.

Συνοψίζοντας, οι προσεγγίσεις μας ανακαλύπτουν χωροχρονικά πρότυπα για την υποστήριξη αποφάσεων σχετικών με τη διαχείριση της κυκλοφορίας αλλά και για την κατανόηση της κίνησης των οχημάτων. Τα αποτελέσματα της έρευνας μας σε αυτό το χώρο παρουσιάζονται στο Κεφάλαιο 4 και έχουν δημοσιευτεί στις [NMM08] and [MT09b].

T-WAREHOUSE – ένα σύστημα για Οπτικοποίηση TDW: Βασισμένοι στο πλαίσιο που μετατρέπει το παραδοσιακό μοντέλο κύβου δεδομένων σε μια αποθήκη δεδομένων τροχιών κινούμενων αντικειμένων υλοποιήσαμε το T-WAREHOUSE, ένα σύστημα που ενσωματώνει όλα τα απαραίτητα σχετικά βήματα, από ανακατασκευή τροχιών και ETL επεξεργασία μέχρι Οπτική OLAP ανάλυση σε δεδομένα κίνησης. Αυτό το σύστημα βασίζεται στην ερεύνα που παρουσιάστηκε στο Κεφάλαιο 3 και παρουσιάζεται λεπτομερώς στο Κεφάλαιο 5. Σε ότι αφορά την ανακατασκευή τροχιών, παρουσιάζουμε

μια αποτελεσματική τεχνική που μετατρέπει ακολουθίες πρωτογενών δεδομένων θέσης σε τροχιές. Στη συνέχεια, αυτές οι ανακατασκευασμένες τροχιές αποθηκεύονται σε μια βάση δεδομένων κινούμενων αντικειμένων και είναι διαθέσιμες για ETL επεξεργασία ώστε να τροφοδοτηθεί η TDW με συγκεντρωτικά δεδομένα κίνησης.

Παρουσιάζουμε τις αρχιτεκτονικές προδιαγραφές του πλαισίου μας και διερευνούμε την αξία του, την ευελιξία και αποδοτικότητα του κατά την εφαρμογή OLAP ανάλυσης σε πραγματικά δεδομένα κίνησης. Τα πρώτα πειραματικά αποτελέσματα καταδεικνύουν την αποδοτικότητα της προσέγγισης μας. Τα μέχρι τώρα αποτελέσματα της έρευνας μας στο συγκεκριμένο χώρο έχουν δημοσιευτεί στις [MFN+08a], [MFN+08b] και [LMF+10].

1.3. Περίγραμμα της Διατριβής

Συνοπτικά, τα κύρια σημεία της διατριβής αυτής είναι τα εξής: Στο Κεφάλαιο 2, παρουσιάζονται οι βασικές αρχές των χωροχρονικών δεδομένων κάνοντας το διαχωρισμό μεταξύ των κινούμενων και μη κινούμενων οντοτήτων. Στο Κεφάλαιο 3 προτείνεται μια δομή για Αποθήκη Δεδομένων Τροχιών Κινούμενων Αντικειμένων, υποστηρίζοντας όλα τα βήματα από την εξαγωγή, επεξεργασία και φόρτωση δεδομένων (Extract-Transform-Load - ETL) μέχρι την Άμεση Αναλυτική Επεξεργασία. Επίσης, παρουσιάζεται μια κατάλληλα σχεδιασμένη παραλλαγή αυτού του μοντέλου αποθήκης δεδομένων ώστε να μπορεί να διαχειρίζεται διαφορετικούς ορισμούς της έννοιας της τροχιάς. Στο Κεφάλαιο 4 παρουσιάζονται δυο προσεγγίσεις για τεχνικές εξόρυξης γνώσης με τη χρήση δεδομένων τροχιών, η μεν πρώτη αφορά την ανακάλυψη προτύπων αλληλεπίδρασης από δεδομένα κινούμενων αντικειμένων και η δεύτερη την ανακάλυψης προτύπων κυκλοφοριακής κίνησης σε οδικά δίκτυα πόλεων. Στο Κεφάλαιο 5 παρουσιάζεται το T-WAREHOUSE, ένα πρωτόλειο σύστημα για την αποθήκευση δεδομένων κίνησης. Στο Κεφάλαιο 6 συνοψίζονται τα αποτελέσματα της έρευνας, καταγράφονται τόσο τα συμπεράσματα όσο και ενδιαφέροντα θέματα που προκύπτουν για μελλοντική έρευνα.

2. Βασικές Αρχές Χωροχρονικών Δεδομένων

Το παρόν κεφάλαιο εστιάζει στην έννοια των χωροχρονικών δεδομένων και, περιγράφει αναλυτικά τις δυο κατηγορίες τους που πρέπει να προσεγγίζονται με διαφορετικό τρόπο: κινούμενες και μη κινούμενες οντότητες. Η δομή του κεφαλαίου έχει ως εξής: Η Ενότητα 2.1 εισάγει τα θέματα που σχετίζονται με τα χωροχρονικά δεδομένα. Η Ενότητα 2.2 εξετάζει την περίπτωση των μη κινούμενων οντοτήτων ενώ η Ενότητα 2.3 παρουσιάζει την έννοια των κινούμενων οντοτήτων. Στην Ενότητα 2.4 υπογραμμίζεται η ανάγκη για ανάπτυξη καινοτόμων τεχνικών υποστήριξης αποφάσεων για την μία ή την άλλη κατηγορία χωροχρονικών δεδομένων. Τέλος, στην Ενότητα 2.5, συνοψίζονται τα συμπεράσματα του κεφαλαίου.

2.1. Εισαγωγή

Με τις τεχνολογικές εξελίξεις στους απομακρυσμένους αισθητήρες (remote sensors), τα δίκτυα αισθητήρων (sensor networks), και τον ταχύ πολλαπλασιασμό των συσκευών που είναι εφοδιασμένες με τεχνολογίες εντοπισμού θέσης και έχουν εισέλθει στην καθημερινή μας ζωή και επαγγελματική δραστηριότητα, έχει παρατηρηθεί τα τελευταία χρόνια ραγδαία αύξηση της δημιουργίας ποικίλων, δυναμικών και γεωγραφικά διασπαρμένων χωροχρονικών δεδομένων. Θα πρέπει να σημειωθεί ωστόσο από την αρχή αυτής της διατριβής ότι διαχωρίζουμε δυο διαφορετικά είδη για τα χωροχρονικά δεδομένα: από την μια πλευρά, αυτά που περιλαμβάνουν την έννοια της κίνησης και από την άλλη τα στατικά χωροχρονικά δεδομένα.

Ένα σημαντικό και ενδιαφέρον παράδειγμα της τελευταίας κατηγορίας είναι η περιοχή των επιστημονικών δεδομένων. Η σημαντική πρόοδος σε τεχνολογίες εδάφους, αέρα και διαστήματος σχετικά με τους αισθητήρες οδήγησαν σε μια χωρίς προηγούμενο πρόσβαση σε επιστημονικά δεδομένα για τους επιστήμονες διαφορετικών ειδικοτήτων, που ενδιαφέρονται να μελετήσουν τη συμπληρωματικής φύση διαφορετικών παραμέτρων. Τα επιτεύγματα αυτά οδηγούν σε ένα περιβάλλον με πλούσια δεδομένα (data-rich) αλλά με φτωχή πληροφορία/γνώση (information-poor). Ο ρυθμός με τον οποίο τα χωροχρονικά δεδομένα δημιουργούνται ξεπερνά τις δυνατότητές μας για να τα οργανώσουμε και τα αναλύσουμε ώστε να εξάγουμε συμπεράσματα για την τρέχουσα κατανόηση ενός κόσμου που αλλάζει δυναμικά. Η Επιστήμη των Υπολογιστών και η Γεω-Πληροφορική μπορούν να συνεισφέρουν ώστε να ανταποκριθούμε σε αυτές τις επιστημονικές και υπολογιστικές προκλήσεις και να παρέχουμε καινοτόμες και αποτελεσματικές λύσεις. Η Ενότητα 2.2 επικεντρώνεται στα σεισμολογικά δεδομένα, μια ενδιαφέρουσα κατηγορία επιστημονικών δεδομένων, και παρουσιάζεται

μια προσέγγιση για την εφαρμογή αναλυτικών τεχνικών με στόχο την υποστήριξη σεισμολόγων και γεωλόγων ώστε να αποκτήσουν πολύτιμη γνώση από τα πολυάριθμα σεισμικά δεδομένα.

Από την άλλη πλευρά, μια τυπική κατηγορία δεδομένων κίνησης είναι τα δεδομένα χρονοσημασμένης θέσης (time-stamped location) που μπορούν να συλλεχθούν από συσκευές με τεχνολογίες εντοπισμού θέσης. Η χρήση αυτών των συσκευών, όπως κινητά τηλέφωνα και συσκευές εφοδιασμένες με τεχνολογίες GPS, είναι εκτενώς διαδεδομένη σήμερα, επιτρέποντας την πρόσβαση σε μεγάλου όγκου δεδομένα που περιλαμβάνουν χρονοσημασμένες γεωγραφικές θέσεις. Ο κατάλληλος χειρισμός των πρωτογενών δεδομένων καταλήγει στη δημιουργία βάσεων δεδομένων τροχιών, κάτι το οποίο είναι γνωστό ως *ανακατασκευή των τροχιών* (trajectory reconstruction) [MFN+08a]. Για να ανταποκριθεί στις ανάγκες που ανακύπτουν, η παραδοσιακή τεχνολογία των βάσεων δεδομένων έχει εξελιχθεί στις επονομαζόμενες Βάσεις Κινούμενων Αντικειμένων (Moving Object Databases - MODs) που χειρίζονται από το σχεδιασμό/μοντελοποίηση έως τη δεικτοδότηση και την αποδοτική επεξεργασία ερωτήσεων για τροχιές [GS05]. Όπως συχνά συμβαίνει στο χώρο διαχείρισης των δεδομένων, η πρόκληση μετά την αποθήκευση των δεδομένων είναι η αξιοποίησή τους και η υλοποίηση των κατάλληλων μεθόδων για την εξαγωγή χρήσιμης γνώσης. Στην Ενότητα 2.3, παρουσιάζονται βασικές έννοιες για τις τροχιές και τις Βάσεις Κινούμενων Αντικειμένων καθώς επίσης και την επεξεργασία γεωγραφικών δεδομένων λαμβάνοντας υπόψη την ιδιωτικότητα (*Geographic Privacy-aware KDD process).*

2.2. Μη Κινούμενες Οντότητες: Η Περίπτωση των Σεισμολογικών Δεδομένων

Για αιώνες οι άνθρωποι βιώνουν, καταγράφουν και μελετούν το φαινόμενο του σεισμού. Λαμβάνοντας υπόψη ότι τουλάχιστον ένας σεισμός με μεγέθους $M < 3$ ($M > 3$) συμβαίνει κάθε ένα δευτερόλεπτο (κάθε 10 λεπτά, αντίστοιχα) παγκοσμίως, είναι αντιληπτό ότι η συλλογή των σεισμικών δεδομένων είναι τεράστια και αυξάνεται αλματωδώς. Οι επιστήμονες καταγράφουν αυτές τις πληροφορίες προκειμένου να περιγράψουν και να μελετήσουν την τεκτονική δραστηριότητα, η οποία περιγράφεται από τα καταχωρημένα γεωγραφικά χαρακτηριστικά (επίκεντρο και έκταση καταστροφών), η ώρα του συμβάντος, το μέγεθος, το εστιακό βάθος κλπ.

Από την άλλη πλευρά, για τους μηχανικούς υπολογιστών που είναι εξειδικευμένοι στην περιοχή της Διαχείρισης της Πληροφορίας και της Γνώσης, τα δεδομένα αυτά αποτελούν ένα «θησαυρό πληροφοριών», όπου θα μπορούσαν να επεξεργαστούν και να αναλύσουν για να συντελέσουν σε εξαγωγή πολύτιμης γνώσης. Ένας μεγάλος αριθμός εφαρμογών για τη διαχείριση και ανάλυση των σεισμολογικών ή, γενικότερα, των γεωφυσικών φαινομένων έχει προταθεί στην υπάρχουσα βιβλιογραφία [AA99], [KR00], [The03], [Yu05]. Ωστόσο, η συνεργασία μεταξύ των επιστημόνων που ασχολούνται με την εξόρυξη γνώσης και των επιστημόνων που ασχολούνται με τα φυσικά φαινόμενα έχει ξεκινήσει σχετικά πρόσφατα [BD00].

Τα σεισμικά φαινόμενα καταγράφονται την ώρα που πραγματοποιούνται από ένα πλήθος οργανισμών (για παράδειγμα τα Γεωδυναμικά Ινστιτούτα και τα εργαστήρια σχολών Φυσικών Επιστημών) σε παγκόσμια κλίμακα. Επομένως, είναι απαραίτητο ένα σύστημα που να συλλέγει και να αναλύει τα πιο ακριβή σεισμικά στοιχεία από διάφορες πηγές. Προφανώς, κάποιες πηγές παρέχουν δεδομένα για τα

ίδια τα σεισμικά φαινόμενα με κάποιες μικρές διαφορές (π.χ. ως προς το μέγεθος ή τη συγκεκριμένη χρονική στιγμή που καταγράφηκε η σεισμική δραστηριότητα). Αυτό που προκύπτει από τα παραπάνω είναι η ανάγκη για το σχεδιασμό της αρχιτεκτονικής ενός *Συστήματος Διαχείρισης Σεισμολογικών Δεδομένων και Εξόρυξης Γνώσης* (Seismic Data Management and Mining System - SDMMS) το οποίο να μπορεί να ενοποιήσει τα δεδομένα των απομακρυσμένων πηγών με κατάλληλο τρόπο διαχωρίζοντας και ομογενοποιώντας τα πρωτογενή δεδομένα [MTK08].

2.2.1. Αποθήκες Σεισμολογικών Δεδομένων και Εξόρυξη Γνώσης

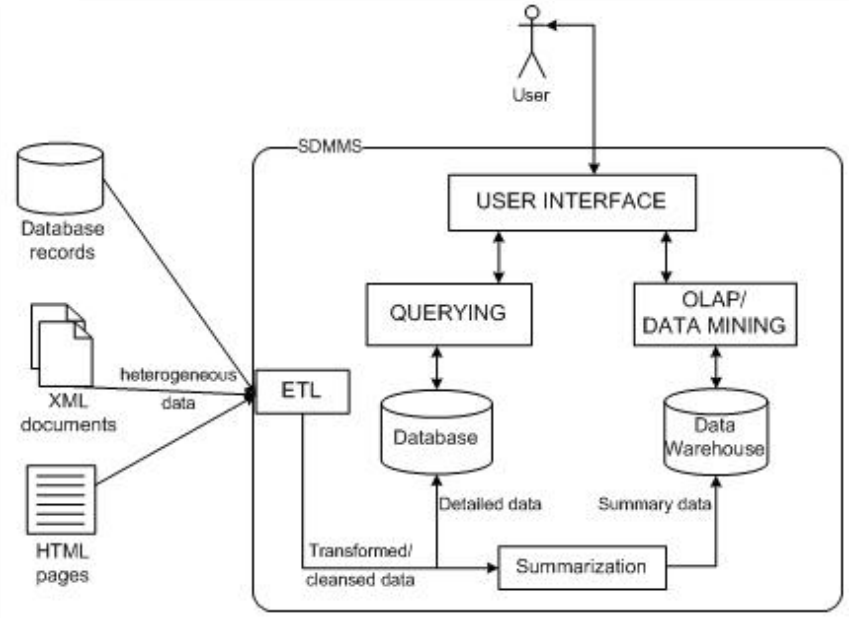
Τα επιθυμητά στοιχεία ενός SDMMS περιλαμβάνουν εργαλεία για γρήγορη και εύκολη προσπέλαση και επισκόπηση των δεδομένων, αλγορίθμους για τη δημιουργία ιστορικών προφίλ συγκεκριμένων γεωγραφικών περιοχών και χρονικών διαστημάτων, τεχνικές που να επιτρέπουν την σύνδεση των σεισμικών φαινομένων με άλλους επιθυμητούς γεωφυσικούς παράγοντες (π.χ. τοπολογία και κλίμα), χάρτες για την απεικόνιση των δεδομένων στον χρήστη και υποστήριξη αλληλεπίδρασης από έμπειρους χρήστες.

Συγκεντρωτικά, ταξινομούμε τα χαρακτηριστικά των χρηστών όπου ένα SDMMS θα μπορούσε να υποστηρίξει σε τρεις κατηγορίες:

- *Ερευνητές γεωφυσικών επιστημών,* με ενδιαφέροντα τη δημιουργία και απεικόνιση σεισμικών προφίλ συγκεκριμένων περιοχών σε συγκεκριμένα χρονικά διαστήματα ή στην ανακάλυψη περιοχών με παρόμοια σεισμική συμπεριφορά.
- *Στελέχη της Δημόσιας Διοίκησης,* που αναζητούν πληροφορίες όπως οι αποστάσεις μεταξύ των επικέντρων των σεισμών και σημείων ειδικού ενδιαφέροντος (σχολεία, νοσοκομεία, βιομηχανικές περιοχές, κτλ.)
- *Απλοί πολίτες* (που «σερφάρουν» στο διαδίκτυο) αναζητώντας πληροφορίες για τη σεισμική δραστηριότητα, και επομένως θέτουν ερωτήματα στο σύστημα για σεισμικές ιδιότητες γενικού ενδιαφέροντος, όπως για παράδειγμα το να βρεθούν όλα τα επίκεντρα των σεισμών σε απόσταση μέχρι 50 χλμ. από το αγαπημένο τους μέρος.

Η διαθεσιμότητα συστημάτων που ακολουθούν την προτεινόμενη αρχιτεκτονική SDMMS παρέχει στους χρήστες πληροφόρηση σχετικά με τους σεισμούς συντελώντας στην αφύπνιση και κατανόηση τους, δύο κρίσιμους παράγοντες για την ανάληψη αποφάσεων, είτε σε ατομικό είτε σε διοικητικό επίπεδο.

Τα δεδομένα που συλλέγονται μπορούν να αποθηκεύονται σε μια τοπική βάση δεδομένων (local database) και/η σε μια αποθήκη δεδομένων (για απλή εκτέλεση ερωτημάτων ή λογικής ανάλυσης για υποστήριξη ανάληψης αποφάσεων αντίστοιχα). Γενικά, τα στοιχεία μέσα στην βάση δεδομένων είναι δυναμικά και λεπτομερή, ενώ αυτά στην αποθήκη δεδομένων (ΑΔ) είναι στατικά και αθροιστικά (αυτό συμβαίνει διότι οι ενημερώσεις σε μια βάση δεδομένων είναι συνεχείς, ενώ μια αποθήκη δεδομένων υπόκειται σε περιοδικές αναθεωρήσεις).



Εικόνα 2-1: Μια προτεινόμενη SDMMS αρχιτεκτονική για διαχείριση σεισμολογικών δεδομένων.

Στην Εικόνα 2-1 απεικονίζεται η προτεινόμενη αρχιτεκτονική που εξυπηρετεί το έργο της συλλογής δεδομένων από διάφορες πηγές σε όλο τον κόσμο και της εναποθήκευσης σε μια τοπική αποθήκη (βάση δεδομένων ή/και αποθήκη δεδομένων). Ένας ενδιάμεσος μεσολαβητής είναι υπεύθυνος για τη διαχείριση της διαδικασίας από την εξαγωγή των δεδομένων από τις πηγές τους μέχρι την «φόρτωσή» τους στην τοπική αποθήκη, την αποκαλούμενη διαδικασία *Εξαγωγής – Μετατροπής – Τροφοδότησης* (Extract-Transform-Load - ETL). Τα αποθηκευμένα δεδομένα «καθαρίζονται» και μετατρέπονται ώστε να έχουν όμοια δομή για να μπορέσουν να αποθηκευτούν μετέπειτα στην βάση δεδομένων του SDMMS.

Τα παραδοσιακά Συστήματα Διαχείρισης Βάσεων Δεδομένων (DBMS) είναι γνωστά ως επιχειρησιακές βάσεις δεδομένων ή συστήματα άμεσης επεξεργασίας δοσοληψιών (On Line Transaction Processing - OLTP) αφού υποστηρίζουν την καθημερινή αποθήκευση και τις ανάγκες ανάκτησης των δεδομένων. Η εκτέλεση ερωτήσεων σε σεισμολογικές βάσεις δεδομένων εμπεριέχει χωροχρονικές έννοιες όπως στιγμιαίες απεικονίσεις, αλλαγές αντικειμένων και χαρτών, κίνηση και φαινόμενα [PT98], [The03]. Πιο συγκεκριμένα, ένα τέτοιο σύστημα SDMMS θα πρέπει να παρέχει στους χρήστες τουλάχιστον τις παρακάτω λειτουργίες:

- *Ανάκτηση χωρικής πληροφορίας σε συγκεκριμένη χρονική στιγμή*. Αυτό χρησιμοποιείται για παράδειγμα όταν επεξεργαζόμαστε εγγραφές που περιλαμβάνουν στοιχεία θέσης (γεωγραφικό πλάτος και μήκος του επίκεντρου του σεισμού) και χρονική στιγμή της σεισμικής δραστηριότητας μαζί με χαρακτηριστικά όπως μέγεθος, εστιακό βάθος και άλλα.
- *Ανάκτηση χωρικής πληροφορίας σε συγκεκριμένο χρονικό διάστημα.* Με αυτό τον τρόπο καταγράφεται χρονικά η εξέλιξη των αντικειμένων χωρικά (υποθέτοντας για παράδειγμα ότι επιθυμούμε να καταγράψουμε τη διάρκεια μιας σεισμικής δόνησης και τον τρόπο που

μεταβάλλονται συγκεκριμένες παράμετροι του φαινομένου κατά τη διάρκεια της εξέλιξής του).

- *Στρωματική Επικάλυψη χωρικής πληροφορίας σε συγκεκριμένη στιγμή ή συγκεκριμένο χρονικό διάστημα:* Ο συνδυασμός στρωμάτων και χρονικών πληροφοριών δίνει τη στιγμιαία απεικόνιση ενός στρώματος. Για παράδειγμα, αυτό το μοντέλο χρησιμοποιείται κατά τον σχηματισμό θεματικών χαρτών μεγέθους σεισμικών δονήσεων που πραγματοποιήθηκαν σε μια συγκεκριμένη μέρα και μέσα σε μια συγκεκριμένη περιοχή (χρονική στιγμή) ή κατά την μοντελοποίηση ολόκληρης της αλληλουχίας σεισμικών δονήσεων, περιλαμβάνοντας προ- και μετα-σεισμική δραστηριότητα (χρησιμοποιώντας την έννοια των στρωμάτων ως χρονικά διαστήματα).

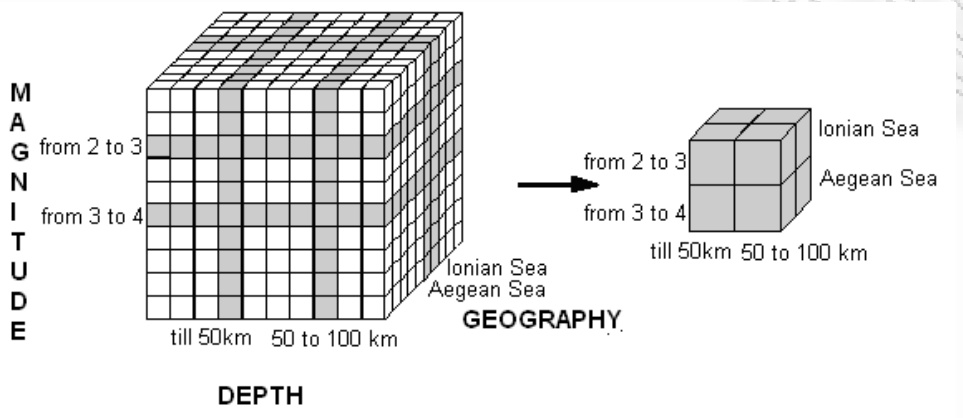
Χαρακτηριστικά παραδείγματα ερωτημάτων που μπορούν να εκτελεστούν συμπεριλαμβάνοντας την χωρική και χρονική διάσταση των σεισμολογικών δεδομένων αποτελούν τα παρακάτω [The03]:

- *Να βρεθούν τα δέκα επίκεντρα σεισμών που πραγματοποιήθηκαν κατά τη διάρκεια των τεσσάρων τελευταίων μηνών, και που βρίσκονται πιο κοντά σε μια συγκεκριμένη τοποθεσία.*
- *Να βρεθούν όλα τα σεισμικά επίκεντρα που βρίσκονται σε μια συγκεκριμένη περιοχή, με μέγεθος > 5, και που πραγματοποιήθηκαν τους τέσσερις τελευταίους μήνες.*
- *(Υποθέτοντας πολλαπλή διαστρωμάτωση πληροφοριών, για παράδειγμα αναφορικά με συντεταγμένες και πληθυσμό μεγάλων πόλεων) να βρεθούν οι πέντε πιο δυνατές σεισμικές δονήσεις που συνέβησαν σε απόσταση μικρότερη των 100 χλμ. από πόλεις με πληθυσμό άνω του 1 εκατομμυρίου κατά τον 20^{ο} αιώνα.*

Η διατήρηση συνοπτικών δεδομένων σε μια τοπική αποθήκη δεδομένων μπορεί να χρησιμοποιηθεί για αναλυτικούς σκοπούς. Δύο δημοφιλείς τεχνικές για ανάλυση δεδομένων και κατανόηση της σημασίας τους είναι το OLAP και η εξόρυξη γνώσης. Συνοπτικά δεδομένα και κρυμμένη γνώση που αποκτήθηκε από τα αποθηκευμένα δεδομένα, μπορεί να οδηγήσει σε ανάληψη ορθότερων αποφάσεων. Ομοίως, τα συνοπτικά σεισμολογικά δεδομένα ενδιαφέρουν πολύ τους γεω-επιστήμονες (earth science) διότι μπορούν να μελετήσουν τα φαινόμενα εμβαθύνοντας και να ερευνήσουν για κρυμμένη και μέχρι τώρα άγνωστη γνώση. Τα πλεονεκτήματα μιας τέτοιας προσέγγισης επεξηγούνται με δύο παραδείγματα εφαρμογών που υποστηρίζονται από χωρική αποθήκη δεδομένων και τεχνολογίες OLAP:

- Ένας χρήστης μπορεί να ζητήσει να δει μέρος του ιστορικού της σεισμικής δραστηριότητας, για παράδειγμα τους δέκα πιο καταστροφικούς σεισμούς τα τελευταία είκοσι χρόνια, και επιπλέον να δει την ίδια πληροφορία για την Ελλάδα (πιο λεπτομερή άποψη, τυπικά μια λειτουργία εμβάθυνσης (drill-down)) ή παγκοσμίως (πιο συγκεντρωτική άποψη, τυπικά μια λειτουργία συσσώρευσης (*roll-up*)).
- Με δεδομένη την ύπαρξη διαφόρων θεματικών χαρτών, για παράδειγμα ίσως έναν για τα μεγέθη σεισμικών δονήσεων και έναν για κάποια άλλη μη γεωφυσική παράμετρο, όπως το μέγεθος της ζημιάς που συντελέστηκε, θα μπορούσαν να χρησιμοποιηθούν για τη διερεύνηση πιθανών σχέσεων, όπως η εύρεση περιοχών με υψηλή σεισμικότητα, αλλά που δεν υπέστησαν ολική καταστροφή και αντίστροφα.

Πρόσθετα στις λειτουργίες συσσώρευσης (roll-up) και εμβάθυνσης (drill-down) που περιγράφηκαν παραπάνω, τυπικές λειτουργίες ενός κύβου δεδομένων είναι και ο *τεμαχισμός* (slice) και το *κομμάτιασμα* (dice), για επιλεγμένα μέρη ενός κύβου δεδομένων θέτοντας όρους/περιορισμούς σε μία ή περισσότερες κυβικές διαστάσεις, αντίστοιχα (Εικόνα 2-2), και η *νοητή περιστροφή* (pivot) που δίνει στο χρήστη τη δυνατότητα εναλλακτικών απόψεων του κύβου.



Εικόνα 2-2: Επιλεγμένα μέρη ενός κύβου φιλτράροντας μία (τεμαχισμός) ή περισσότερες διαστάσεις (κομμάτιασμα).

Η ενσωμάτωση τεχνικών ανάλυσης δεδομένων και εξόρυξης γνώσης σε ένα Σύστημα Διαχείρισης Σεισμολογικών Δεδομένων και Εξόρυξης Γνώσης αποσκοπεί τελικά στην ανακάλυψη ενδιαφέρουσας, κρυμμένης και προηγούμενα άγνωστης γνώσης. Παραδείγματα χρήσιμων προτύπων που βρέθηκαν μέσω της διαδικασίας KDD (Knowledge Discovery & Delivery) περιλαμβάνουν συσταδοποίηση πληροφοριών (π.χ. δονήσεις που συνέβησαν κοντά σε χώρο και/ή χρόνο), κατηγοριοποίηση φαινομένων με βάση την έκταση και το επίκεντρο, διερεύνηση της σημασιολογίας των φαινομένων με τη χρήση τεχνικών εξεύρεσης προτύπων (π.χ. χαρακτηρισμός των δονήσεων και πιθανών ισχυρών μετασεισμικών δονήσεων σε αλληλουχία, μέτρηση της ομοιότητας της ακολουθίας των σεισμικών δονήσεων σύμφωνα με το μέγεθος μέτρησης της ομοιότητας που ορίστηκε από τον εμπειρογνώμονα, κτλ).

2.3. Κινούμενες Οντότητες: Η Περίπτωση των Τροχιών Κινούμενων Αντικειμένων

Τα κινούμενα αντικείμενα είναι γεωμετρίες (δηλ. σημεία, γραμμές, περιοχές) που αλλάζουν με την πάροδο του χρόνου και η κίνηση τους αυτή περιγράφεται από τα δεδομένα των τροχιών τους. Η μετακίνηση υπονοεί δύο διαστάσεις: χωρική και χρονική. Πιο συγκεκριμένα, η μετακίνηση μπορεί να περιγραφεί ως συνεχής αλλαγή θέσης μεταξύ δύο διαφορετικών χρονικών στιγμών σε συγκεκριμένο γεωγραφικό χώρο [MVO+07].

Μια τροχιά T είναι μια συνεχής αναπαράσταση από το χρονικό $I \subseteq \mathrm{R}$ στο χωρικό πεδίο (γεωγραφικό χώρο $\mathbb{R}^2$, επίπεδο 2D) [MVO+07]:

$$I \subseteq \mathrm{R} \rightarrow \mathrm{R}^2 : \mathrm{t} \rightarrow \mathrm{a}(\mathrm{t}) = \left(\mathrm{a_x}(\mathrm{t}), \mathrm{a_y}(\mathrm{t})\right) \tag{2.1}$$

και,

$$T = \{\left(\mathrm{a_x}(\mathrm{t}), \mathrm{a_y}(\mathrm{t}), \mathrm{t}\right) \mid t \in I\} \subset \mathrm{R}^2 \times \mathrm{R} \tag{2.2}$$

όπου $\left(\mathrm{a_x}(\mathrm{t}), \mathrm{a_y}(\mathrm{t}), \mathrm{t}\right)$ είναι τα σημεία δειγματοληψίας που περιέχονται στο διαθέσιμο σύνολο δεδομένων.

Από πλευράς εφαρμογής, μια τροχιά είναι η συνάρτηση καταγραφής της κίνησης ενός αντικειμένου, με άλλα λόγια η καταγραφή των θέσεων ενός αντικειμένου σε συγκεκριμένα χρονικά διαστήματα. Ενώ η πραγματική τροχιά έχει μορφή καμπύλης, οι απαιτήσεις στην εφαρμογή υπονοούν ότι πρέπει να χτιστεί από ένα σύνολο δειγματοληπτικών δεδομένων. Έτσι, οι τροχιές κινούμενων σημείων συχνά ορίζονται ως ακολουθίες τριάδων (x, y, t) [GBE+00]:

$$T = \{(\mathrm{x_1}, \mathrm{y_1}, \mathrm{t_1}), (\mathrm{x_2}, \mathrm{y_2}, \mathrm{t_2}), \ldots, (\mathrm{x_n}, \mathrm{y_n}, \mathrm{t_n})\}, \tag{2.3}$$

όπου $x_i, y_i, t_i \in \mathrm{R}$, και $t_1 < t_2 < .. < t_n$,

Η χωροχρονική φύση των δεδομένων των κινούμενων αντικειμένων που συλλέγονται από σύγχρονες συσκευές αισθητήρων και ασύρματες τεχνολογίες δίνουν τη δυνατότητα για έρευνα σε νέες περιοχές συσχετιζόμενες με διαχείριση τους και υλοποίηση των κατάλληλων αναλυτικών μεθόδων για την εξαγωγή γνώσης.

Οι Βάσεις Κινούμενων Αντικειμένων είναι ένας σημαντικός ερευνητικός χώρος όπου έχει λάβει πολλή προσοχή κατά τη διάρκεια των τελευταίων ετών. Ο κύριος στόχος στο χώρο αυτό είναι η επέκταση της τεχνολογίας των βάσεων δεδομένων ώστε να συμπεριληφθούν οι κατάλληλες τεχνικές για την αναπαράσταση, εκτέλεση ερωτημάτων, δεικτοδότηση και μοντελοποίηση των τροχιών των κινούμενων αντικειμένων. Επιπλέον, η ανάλυση του μεγάλου όγκου των δεδομένων που συλλέχθηκαν αποτελεί σημαντικό ζήτημα. Αυτό οφείλεται στο γεγονός ότι οι παραδοσιακές αναλυτικές τεχνικές δεν μπορούν να εφαρμοστούν λόγω της χωροχρονικής διάστασης των δεδομένων. Αυτό που προκύπτει από τα παραπάνω είναι η ανάγκη ύπαρξης μια νέας διαδικασίας KDD όπου θα μπορεί να εφαρμοστεί σε δεδομένα τροχιών. Στην διατριβή αυτή γίνεται συζήτηση για διάφορα βήματα αυτής της διαδικασίας, περίληψη των οποίων παρουσιάζεται στην επόμενη υποενότητα.

2.3.1. Διαχείριση, Αποθήκευση και Εξόρυξη γνώσης από Δεδομένα Κινούμενων Αντικειμένων

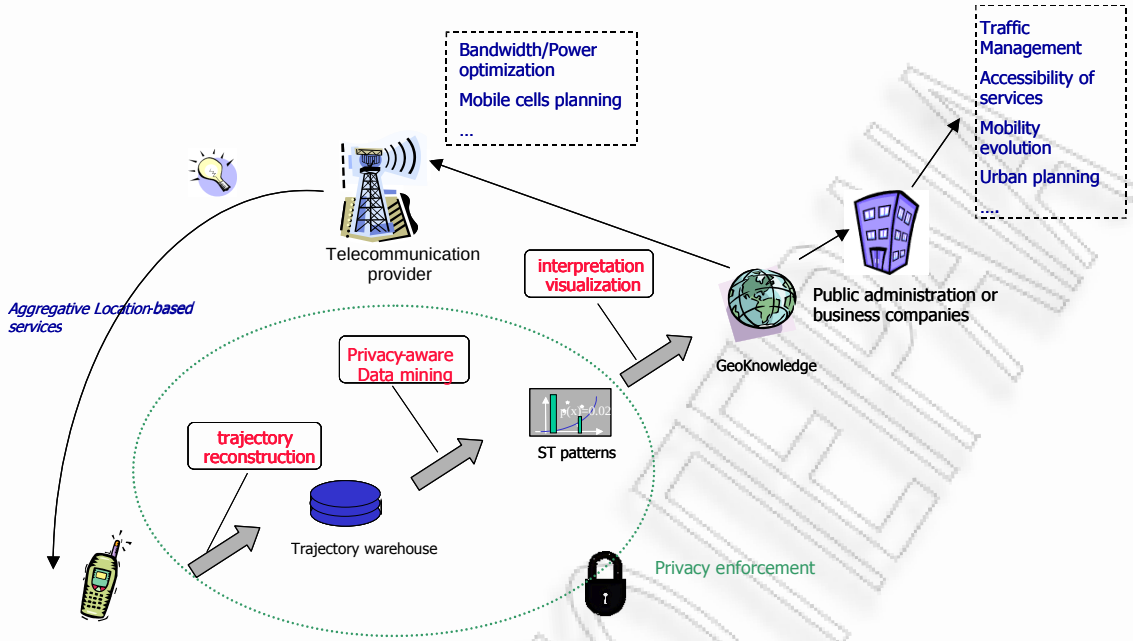
Ο ερευνητικός χώρος των Βάσεων Κινούμενων Αντικειμένων αντιμετωπίζει την ανάγκη αναπαράστασης της κίνησης των αντικειμένων (δηλαδή των τροχιών) στις βάσεις δεδομένων ώστε να γίνει δυνατή η εκτέλεση ειδικών (ad-hoc) ερωτημάτων και η ανάλυση τους. Κατά τη διάρκεια της τελευταίας δεκαετίας, πραγματοποιήθηκε σημαντική έρευνα που περιλαμβάνει από μοντελοποίηση των δεδομένων και ανάπτυξη γλωσσών για εκτέλεση ερωτημάτων μέχρι και θέματα υλοποίησης, όπως αποδοτική δεικτοδότηση, τεχνικές επεξεργασίας ερωτημάτων και τεχνικές βελτιστοποίησης. Τουλάχιστον δύο πρότυποι μηχανισμοί MOD έχουν προταθεί στην υπάρχουσα βιβλιογραφία, με ονομασίες SECONDO [AGB06] και HERMES [PFG+08].

Το σύστημα HERMES αποτελεί το βασικό μηχανισμό για βάσεις δεδομένων κινούμενων αντικειμένων που χρησιμοποιήθηκε για την ανάπτυξη των τεχνικών που προτείνονται σε αυτή τη διατριβή και το όποιο παρουσιάζεται συνοπτικά παρακάτω. Το HERMES είναι ο μηχανισμός μιας βάσης που παρέχει χωροχρονική λειτουργικότητα έτσι ώστε να γίνεται δυνατός ο χειρισμός αντικειμένων που αλλάζουν θέση, σχήμα και μέγεθος, με διακριτή ή συνεχή ροή στο χρόνο. Το σύστημα μπορεί να χρησιμοποιηθεί είτε ως αμιγώς χωρικό είτε ως αμιγώς χρονικό σύστημα, αλλά η κύρια λειτουργικότητα του είναι η υποστήριξη του σχεδιασμού και της εκτέλεσης ερωτημάτων για συνεχώς κινούμενα αντικείμενα. Μια τέτοια συλλογή τύπων δεδομένων και οι αντίστοιχες λειτουργίες ορίζονται, αναπτύσσονται και παρέχονται ως ένα Oracle Data Cartridge, που ονομάζεται HERMES Moving Data Cartridge (HERMES-MDC). Το Hermes Moving Data Cartridge (Hermes MDC) αποτελεί τον πυρήνα της αρχιτεκτονικής του συστήματος. Παρέχει λειτουργίες που διευκολύνουν την επεξεργασία ερωτημάτων για κινούμενα αντικείμενα καθώς επίσης και για ένα σύνολο μεταβαλλόμενων στον χρόνο γεωμετριών (κύκλος, πολύγωνο, σημείο κτλ). Μεταξύ άλλων, οι λειτουργίες του HERMES περιλαμβάνουν:

- επεξεργασία ερωτημάτων για στατικά αντικείμενα: για παράδειγμα ερωτήματα που αφορούν απόσταση από συγκεκριμένα σημεία ή γειτνίαση (π.χ. να βρεθούν αυτά που πέρασαν από τη συγκεκριμένη περιοχή κατά τη διάρκεια της τελευταίας ώρας),
- επεξεργασία ερωτημάτων για κινούμενα αντικείμενα: για παράδειγμα ερωτήματα που αφορούν απόσταση από συγκεκριμένα σημεία (π.χ. να βρεθούν αυτά που πέρασαν από κοντά μου σήμερα το απόγευμα) και ομοιότητα (π.χ. να βρεθούν οι τρεις πιο όμοιες τροχιές σε σχέση με αυτή που ακολούθησα χθες),
- επεξεργασία ερωτημάτων που περιλαμβάνουν τελεστές, όπως απόσταση που διανύθηκε ή ταχύτητα (π.χ. να βρεθεί η μέση ταχύτητα της τροχιάς που ακολούθησα το σαββατοκύριακο).

Όσον αφορά την ανάλυση των κινούμενων αντικειμένων, ο όρος *Διαδικασία Ανάλυσης Γεωγραφικών Δεδομένων λαμβάνοντας υπόψη την Ιδιωτικότητα* (Geographic Privacy-aware KDD process) προέκυψε από την έρευνα που διεξήχθη στα πλαίσια του ερευνητικού έργου GeoPKDD [Geo06]**,** μέσα από το οποίο προτάθηκαν οι βασικές θεωρίες για τη χρήση ιχνών και τροχιών των κινούμενων αντικειμένων σε εφαρμογές του πραγματικού κόσμου. Η διαδικασία αυτή αποτελείται από μια σειρά μεθόδων και τεχνολογιών που μπορούν να εφαρμοστούν σε κινούμενα αντικείμενα και είναι οργανωμένες σε συγκεκριμένα και σαφή βήματα με προκαθορισμένο στόχο: την εξαγωγή γνώσης για το χρήστη-καταναλωτή από μεγάλο αριθμό πρωτογενών δεδομένων αναφερόμενα σε χώρο και χρόνο, υπακούοντας στην αρχή της ιδιωτικότητας. Πιο συγκεκριμένα, οι κύριες λειτουργίες (περιγράφονται στην Εικόνα 2-3) της διαδικασίας GeoPKDD είναι:

- ανακατασκευή τροχιών από ρεύματα πρωτογενών δεδομένων κινούμενων αντικειμένων, και κατασκευή μιας βάσης δεδομένων, υπακούοντας στις αρχές της ιδιωτικότητας,
- χωροχρονικές μέθοδοι εξόρυξης γνώσης (που εξασφαλίζουν την ιδιωτικότητα) και αλγόριθμοι εξαγωγής γνώσης, ώστε να ανακαλύψουμε χωροχρονικά πρότυπα,
- κατανόηση και τεχνικές απεικόνισης της γνώσης από τα γεωγραφικά δεδομένα για να παρέχουν αξιοποιήσιμα πρότυπα στους τελικούς χρήστες.



Εικόνα 2-3: Γενική απεικόνιση διαχείρισης, αποθήκευσης δεδομένων κινούμενων αντικειμένων και έννοιες εξόρυξης γνώσης [Geo06].

Αυτή η KDD διαδικασία μπορεί να εφαρμοστεί σε ετερογενή δεδομένα κινούμενων αντικειμένων. Το κινητό τηλέφωνο που απεικονίζεται στην Εικόνα 2-3 αναπαριστά διάφορα σύνολα δεδομένων που έρχονται από διαφορετικές συσκευές:

- *GPS:* το λειτουργικό δορυφορικό σύστημα πλοήγησης που συγκεντρώνει πάνω από 24 δορυφόρους που εκπέμπουν ακριβή σήματα σε δέκτες GPS, επιτρέποντας τους να καθορίσουν με ακρίβεια την θέση τους (γεωγραφικό μήκος, πλάτος και ύψος) ανεξαρτήτως καιρού, ημέρας ή νύχτας, οπουδήποτε στη γη.
- *GSM*: Το πιο δημοφιλές πρότυπο για κινητά τηλέφωνα, χρησιμοποιείται από πάνω από 1.5 δισεκατομμύρια ανθρώπους σε πάνω από 210 χώρες στον κόσμο. Η καθολικότητα των προτύπων GSM καθιστά τη διεθνή περιαγωγή πολύ εύκολη μεταξύ των χρηστών κινητών τηλεφώνων, επιτρέποντας στους συνδρομητές να χρησιμοποιούν τα κινητά τηλέφωνά τους σε πολλά μέρη του κόσμου. Τα δίκτυα GSM αποτελούνται από μια σειρά σταθμών βάσεων, ο καθένας είναι υπεύθυνος για μια συγκεκριμένη γεωγραφική περιοχή (γνωστή ως κυψέλη). Ως εκ τούτου, για κάθε συσκευή εξοπλισμένη με τεχνολογία GSM είναι δυνατή η συλλογή δεδομένων για το από ποια βάση εξυπηρετήθηκαν οι σταθμοί σε συγκεκριμένες χρονικές στιγμές και άρα είναι δυνατή η υπόθεση της κίνησης.
- *Wi-Fi*: το πιο δημοφιλές πρότυπο για την ασύρματη επικοινωνία μεταξύ συσκευών. Μια συσκευή εξοπλισμένη με Wi-fi τεχνολογία όπως ένας φορητός ηλεκτρονικός υπολογιστής, ένα κινητό τηλέφωνο, ένα PDA κτλ, μπορούν να συνδεθούν με κάποια άλλη συσκευή όταν αυτή είναι μέσα στο πεδίο του ασύρματου δικτύου. Το ασύρματο δίκτυο ορίζεται ως το σύνολο των διασυνδεδεμένων στοιχείων πρόσβασης – αποκαλούμενων «θερμές ζώνες» (hotspots) – που μπορούν να καλύψουν μια περιοχή τόσο μικρή όσο ένα δωμάτιο ή τόσο

μεγάλη όσο μια ολόκληρη πόλη (WiMax). Όπως και στο GSM, μπορούμε να συλλέξουμε για την κάθε συσκευή τον κατάλογο των ζωνών που την εξυπηρέτησαν σε διαφορετικές χρονικές στιγμές.

Στον πραγματικό κόσμο, αυτοί οι διαφορετικοί τύποι δεδομένων μπορεί να είναι διαθέσιμοι και πρέπει να αποθηκευτούν, να υποστούν επεξεργασία και να αναλυθούν. Για να επιτευχθεί αυτός ο στόχος, είναι απαραίτητες κατάλληλες τεχνικές ανακατασκευής των τροχιών που να σέβονται τη διαφορετικότητα των δεδομένων. Προφανώς, η ουσία αυτών των τεχνικών είναι η ίδια: μετασχηματίζουν ακατέργαστα δεδομένα σε τροχιές. Ωστόσο, οι παραλλαγές στη φύση και την ακρίβεια των δεδομένων απαιτούν μια διαφορετική προσέγγιση για κάθε τύπο δεδομένων. Επιπλέον, αυτές οι τεχνικές μετασχηματισμού των τροχιών μπορούν να εφαρμόσουν κάποια βασική προεπεξεργασία των τροχιών. Αυτό περιλαμβάνει την παραμετρική συμπίεση τροχιάς (ώστε να απορριφθούν οι περιττές λεπτομέρειες και να κρατηθούν ταυτόχρονα οι αφαιρέσεις των τμημάτων των τροχιών που έχουν αποσταλεί μέχρι τώρα), καθώς επίσης και τις τεχνικές χειρισμού ελλιπών ή λανθασμένων τιμών.

Οι ανακατασκευασμένες τροχιές αποθηκεύονται σε μια TDW που εξυπηρετεί δυο κύριες ανάγκες: την παροχή της κατάλληλης υποδομής για προηγμένες ικανότητες υποβολής αναφορών (reporting) και τη διευκόλυνση της εφαρμογής αλγορίθμων εξόρυξης γνώσης από δεδομένα τροχιών πάνω σε συσσωρευμένα στοιχεία. Σύμφωνα με τις ανάγκες των τελικών χρηστών, οι τελευταίοι επιθυμούν να έχουν πρόσβαση είτε σε απλές εκθέσεις είτε σε εκθέσεις που προκύπτουν μετά από ανάλυση OLAP. Σενάρια τύπου *τι θα συνέβαινε εάν* (what - if) και πολυδιάστατη ανάλυση είναι χαρακτηριστικά παραδείγματα αναλυτικών μεθόδων που θα μπορούσαν να βασιστούν σε βάσεις δεδομένων τροχιών.

Στην ανωτέρω Εικόνα 2-3 η αποθήκη δεδομένων τροχιών μπορεί να περιλαμβάνει μια MOD για την αποθήκευση όλων των στοιχείων των δεδομένων τροχιάς, η οποία κατόπιν τροφοδοτεί την TDW με συγκεντρωτικά (aggregate) στοιχεία εφαρμόζοντας μια διαδικασία ETL με στόχο την παραγωγή ποιοτικών πληροφοριών (π.χ. τροχιές σε διαφορετικές κλιμακώσεις, συναθροίσεις, μετα-δεδομένα κίνησης κτλ.).

Για να αντιμετωπιστούν οι εφαρμογές κινούμενων αντικειμένων που περιορίζονται σε κάποιο δίκτυο, τόσο η MOD όσο και η TDW μπορεί να επιβάλλουν οι τροχιές να είναι αντιστοιχισμένες σε χάρτη (map matched). Με άλλα λόγια, μπορεί να χρειαστούν τα συγκεκριμένα σημεία και τμήματα των τροχιών να αντιστοιχούν σε έγκυρες διαδρομές του δικτύου. Αυτό μπορεί να περιλαμβάνει για παράδειγμα την εκτέλεση διαδικασιών προ-επεξεργασίας ή μετα-επεξεργασίας χωρίς να παραβιάζουν την εγκυρότητα των τροχιών σε όρους του πραγματικού υποκείμενου δικτύου.

Επιπροσθέτως, η γνώση της Γεωγραφίας (Geoknowledge) στην Εικόνα 2-3 μπορεί να περιγραφεί μέσω της ενσωμάτωσης GIS στρωμάτων που θα μπορούσαν να καταλήξουν σε ένα πλουσιότερο εννοιολογικό μοντέλο/πρότυπο παρέχοντας έτσι και πιο προηγμένες αναλυτικές δυνατότητες. Συνδυάζοντας τα δεδομένα τροχιών με θεματικά στρώματα (όπως τα γεωγραφικά, τοπογραφικά και δημογραφικά στρώματα) είναι δυνατή η ενίσχυση των αναλυτικών δυνατοτήτων πιθανών εφαρμογών.

Η *εξόρυξη γνώσης από δεδομένα τροχιών* (trajectory mining) αφορά την εφαρμογή τεχνικών εξόρυξης γνώσης σε δεδομένα τροχιών. Μέσω αυτής της εργασίας παράγονται πρότυπα τροχιών που περιγράφουν τη συμπεριφορά των τροχιών. Διαδοχικά και δημοφιλή πρότυπα μπορούν να ανακαλυφθούν χρησιμοποιώντας παραδοσιακές ή ειδικές (ad hoc) τεχνικές εξαγωγής προτύπων.

Αναφορικά με την ιδιωτικότητα, η διαχείριση τέτοιων ζητημάτων μπορεί να ενσωματωθεί μέσα στα εργαλεία αποθήκευσης δεδομένων και εξόρυξης γνώσης (όπως περιγράφεται στην Εικόνα 2-3) έτσι ώστε να διασφαλιστεί ότι τα δεδομένα των τροχιών του κάθε κινούμενου αντικειμένου θα αποθηκεύονται και θα αναλύονται χωρίς να παραβιάζονται οι αρχές προστασίας των προσωπικών δεδομένων. Μια πολύ απλή στρατηγική αποτελεί η αποφυγή της ανακάλυψης προτύπων που αφορούν περιορισμένο αριθμό χρηστών. Αυτό πρέπει να γίνει ώστε να αποφευχθεί η ταυτοποίηση αυτών των χρηστών.

Η γνώση που ανακαλύπτεται μπορεί να είναι χρήσιμη τόσο στην δημόσια διοίκηση όσο και σε ιδιωτικές εταιρίες. Συγκεκριμένα στη περίπτωση των παροχών τηλεπικοινωνίας, η γνώση αυτή μπορεί να είναι χρήσιμη σε πλήθος εφαρμογών: όπως για παράδειγμα στη βελτιστοποίηση των υπηρεσιών των δικτύων της κινητής τηλεφωνίας καθώς επίσης και στην ανάπτυξη καινοτόμων υπηρεσιών εντοπισμού θέσης που με τη σειρά τους θα προσφέρουν χρήσιμες εφαρμογές.

2.4. Η Ανάγκη Καινοτομίας στις Τεχνικές Υποστήριξης Αποφάσεων

Οι παραδοσιακές τεχνικές υποστήριξης αποφάσεων που αναπτύσσονται ως σύνολο εφαρμογών και τεχνολογιών για τη συλλογή, την αποθήκευση, την ανάλυση, και την παροχή πρόσβασης στα στοιχεία, π.χ. αποθήκευση δεδομένων, αναλυτική επεξεργασία (OLAP), εξόρυξη γνώσης και οπτικοποίηση. Αυτές οι τεχνικές ενσωματώνονται στα συστήματα υποστήριξης απόφασης για να υποστηρίξουν τις δραστηριότητες επιχειρησιακής (operational) και οργανωτικής (organizational) λήψης αποφάσεων. Τέτοια συστήματα βοηθούν τους ιθύνοντες να συνδυάζουν τα ακατέργαστα στοιχεία, τα έγγραφα, την προσωπική γνώση, τα επιχειρησιακά πρότυπα, κ.λπ. για να προσδιορίσουν, να αναλύσουν και να λύσουν τα προβλήματα καθώς επίσης και να λάβουν τις κατάλληλες αποφάσεις.

Οι τεχνικές υποστήριξης αποφάσεων αναπτύχθηκαν για να ικανοποιήσουν τις μεταβαλλόμενες και περίπλοκες ανάγκες της σημερινής επιχείρησης και του τεχνολογικού περιβάλλοντος. Προς αυτόν το στόχο, είναι απαραίτητη η συνεχής επέκταση τους ώστε να δοθούν οι κατάλληλες οι λύσεις στις νέες προκλήσεις που προκύπτουν. Συνήθως θεωρείται ότι υπάρχουν δύο προκλήσεις που οδηγούν την αλλαγή στον τομέα των τεχνικών υποστήριξης αποφάσεων: αφ' ενός υπάρχει η ανάγκη να επεκταθούν τα υπάρχοντα εργαλεία ώστε να μπορούν να χρησιμοποιήσουν σε πραγματικό χρόνο επιχειρησιακά (operational) δεδομένα και αφ' ενός οι τεχνολογικές πρόοδοι στον τομέα των επικοινωνιών που επιτρέπουν την πρόσβαση σε τεράστιους όγκους δεδομένων κίνησης.

Σε ότι αφορά το πρώτο, θεωρείται αναγκαία η επέκταση των υπαρχόντων εργαλείων ώστε να υποστηριχθούν και οι επιχειρησιακές πέρα από τις στρατηγικές και τακτικές αποφάσεις. Οι διαφορές μεταξύ αυτών των τύπων συσχετίζονται με το χρονικό διάστημα που κάθε απόφαση απαιτεί αλλά και με τη φύση τους. Η ανώτερη διοίκηση είναι αρμόδια για το στρατηγικό προγραμματισμό των οργανισμών, ενώ τα μεσαία στελέχη λαμβάνουν τις τακτικές αποφάσεις ακολουθώντας τα σχέδια της

ανώτατης διοίκησης. Ως εκ τούτου, υποτιμούνται οι επιχειρησιακές αποφάσεις, αρμόδιες για την καθημερινή δραστηριότητα του οργανισμού. Σήμερα, υπάρχει η τάση για ολοκληρωμένες μετρήσεις και διαχείριση απόδοσης [MT06] και αυτό οδηγεί σε νέα εργαλεία και τεχνικές που μπορούν να χρησιμοποιήσουν τα επιχειρησιακά στοιχεία [Kot06], [ADH+03], [CCD+04].

Αυτή η διατριβή εστιάζει στη δεύτερη πρόκληση και συζητά την επέκταση των παραδοσιακών τεχνικών ώστε να είναι δυνατή η ανάπτυξη νέων τεχνικών ανάλυσης που θα είναι κατάλληλα για δεδομένα κίνησης. Ο στόχος είναι να μπορέσουν να εξυπηρετηθούν αναδυόμενες εφαρμογές (π.χ. μάρκετινγκ που λαμβάνει υπόψη την κίνηση και διαχείριση κυκλοφορίας) που χρειάζονται τη μετατροπή των πρωτογενών δεδομένων θέσης σε χρήσιμη γνώση.

Η εφαρμογή τεχνικών ΑΔ και OLAP σε συμβατικά επιχειρηματικά δεδομένα έχει μελετηθεί εκτενώς στη βιβλιογραφία. Η προσαρμογή και εφαρμογή αυτών των τεχνικών σε δεδομένα κίνησης μπορεί να μας παρέχει σημαντική γνώση σχετικά με τις τροχιές κινούμενων αντικειμένων. Μια Αποθήκη Δεδομένων Τροχιών Κινούμενων Αντικειμένων μπορεί να βοηθήσει στον υπολογισμό συναθροίσεων πάνω σε δεδομένα τροχιών επιτρέποντας έτσι τη μελέτη τους από ένα υψηλότερο επίπεδο αφαίρεσης. Θεωρητικά, αυτή η προσέγγιση προστατεύει και την ιδιωτικότητα αφού δεν εστιάζει σε μεμονωμένες τροχιές. Επιπλέον, οδηγεί στη συλλογή και ομογενοποίηση δεδομένων από πολλαπλές πηγές εξυπηρετώντας έτσι και άλλες τεχνικές ανάλυσης που μπορούν να εφαρμοστούν σε συγκεντρωτικά αντί για πρωτογενή δεδομένα.

Οι τεχνικές εξόρυξης γνώσης χρησιμοποιούνται για την ανακάλυψη άγνωστων αλλά χρήσιμων προτύπων. Ο τεράστιος όγκος διαθέσιμων δεδομένων κίνησης απαιτεί την επέκταση των παραδοσιακών τεχνικών εξόρυξης γνώσης ώστε να είναι κατάλληλες για αυτόν τον νέο τύπο δεδομένων. Η ανακάλυψη χωροχρονικών συσχετίσεων, συστάδων και η πρόβλεψη ενεργειών κτλ μπορεί να οδηγήσει σε πρότυπα κίνησης που μας επιτρέπουν να κατασκευάσουμε χρήσιμες, συγκεντρωτικές αφαιρέσεις πρωτογενών δεδομένων ώστε να αποκτήσουμε γνώση για τις συμπεριφορές κίνησης.

2.5. Σύνοψη

Σε αυτό το κεφάλαιο, διαχωρίστηκαν οι δύο διαφορετικοί τύποι των χωροχρονικών δεδομένων: αυτών που περιλαμβάνουν την έννοια της κίνησης και των στατικών χωροχρονικών δεδομένων. Παρουσιάστηκαν επίσης διαφορετικές αρχιτεκτονικές για τη διαχείριση των χωροχρονικών δεδομένων που ανήκουν και στις δύο παραπάνω κατηγορίες. Σχετικά με τα στάσιμα χωροχρονικά δεδομένα, αναλύθηκε η περίπτωση των σεισμολογικών δεδομένων και παρουσιάστηκε ένα ολοκληρωμένο πλαίσιο για τη *Διαχείριση Σεισμολογικών Δεδομένων και Εξόρυξη Γνώσης - Seismic Data Management and Mining.* Από την άλλη μεριά, για τα δεδομένα κίνησης, παρουσιάστηκε ο χώρος των Βάσεων Δεδομένων Κινούμενων Αντικειμένων και αναλύθηκε η έννοια της *Διαδικασίας Ανάλυσης Γεωγραφικών Δεδομένων λαμβάνοντας υπόψη την Ιδιωτικότητα* (Geographic Privacy-aware KDD process).

3. Αποδοτικές Αποθήκες Δεδομένων Τροχιών Κινούμενων Αντικειμένων

Σε αυτό το κεφάλαιο εστιάζουμε σε τεχνικές αποθήκευσης δεδομένων και μελετούμε την εφαρμογή τους σε δεδομένων τροχιών. Παρουσιάζουμε δυο προτάσεις: ένα πλαίσιο για TDW που απαιτεί ένα κατάλληλο, μοναδικό και στατικό ορισμό της έννοιας της τροχιάς και ένα πλαίσιο για ειδική (ad-hoc) TDW που επιτρέπει πολλαπλούς ορισμούς της έννοιας της τροχιάς. Το κεφάλαιο διαμορφώνεται ως εξής: η Ενότητα 3.1 εισάγει τα θέματα που σχετίζονται με τις τεχνικές αποθήκευσης δεδομένων τροχιών κινούμενων ενώ, η Ενότητα 3.2 παρουσιάζει τα κίνητρα της έρευνας μας σε αυτό το χώρο. Η Ενότητα 3.3 εστιάζει σε θέματα αποθήκευσης δεδομένων θεωρώντας στατικούς σημασιολογικούς ορισμούς της έννοιας της τροχιάς ενώ, η Ενότητα 3.4 παρουσιάζει ένα πλαίσιο που λαμβάνει υπόψη του πολλαπλούς σημασιολογικούς ορισμούς της έννοιας τροχιάς. Η Ενότητα 3.5 εξετάζει τις σχετικές εργασίες και, τέλος, η Ενότητα 3.6 ολοκληρώνει το κεφάλαιο με την παρουσίαση των συμπερασμάτων.

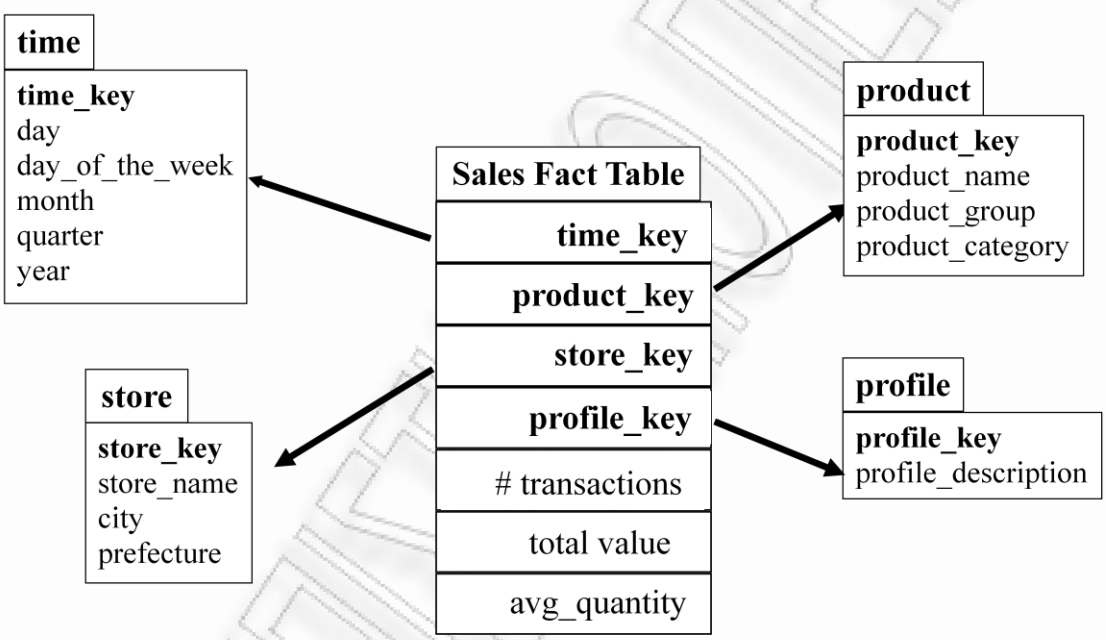
3.1. Εισαγωγή

Η αποθήκευση δεδομένων έχει συγκεντρώσει σημαντικό ενδιαφέρον από ερευνητές της κοινότητας των βάσεων δεδομένων ως μια τεχνολογία που μπορεί να ενσωματώσει δεδομένα συναλλαγών που μπορεί να βρίσκονται διασκορπισμένα μέσα σε οργανισμούς και των οποίων οι διάφορες εφαρμογές μπορεί να χρησιμοποιούν είτε παλαιάς τεχνολογίας (μη σχεσιακά) είτε προηγμένα συστήματα σχεσιακών βάσεων δεδομένων. Οι αποθήκες δεδομένων ορίζουν ένα τεχνολογικό πλαίσιο για υποστήριξη διαδικασιών λήψης αποφάσεων παρέχοντας πληροφοριακά δεδομένα. Μια αποθήκη δεδομένων ορίζεται ως μια θεματο-κεντρική, ενσωματωμένη, με χρονική διάσταση και μη ευμετάβλητη συλλογή δεδομένων για την υποστήριξη της διαχείρισης των διαδικασιών λήψης αποφάσεων [Inm96].

Σε μια αποθήκη δεδομένων τα δεδομένα οργανώνονται και υφίστατο χειρισμό σύμφωνα με τις έννοιες και τους τελεστές που παρέχονται από ένα πολυδιάστατο μοντέλο δεδομένων που θεωρεί ότι τα δεδομένα είναι οργανωμένα στη μορφή του κύβου δεδομένων [AAD+96]. Ένας κύβος δεδομένων επιτρέπει τη μοντελοποίηση και επισκόπηση των δεδομένων μέσα από πολλαπλές διαστάσεις όπου η κάθε μια αντιπροσωπεύει κάποια επιχειρηματική προοπτική και συνήθως υλοποιείται ως ένα μοντέλο με σχήμα αστέρα (ή χιονοστιβάδας). Με βάση αυτό το μοντέλο, μια αποθήκη δεδομένων αποτελείται

από έναν πίνακα συμβάντων (fact table) (σχηματικά, στο κέντρο του αστέρα) που είναι περικυκλωμένο από ένα σύνολο από πίνακες διαστάσεων που σχετίζονται με τον πίνακας συμβάντων, ο οποίος περιλαμβάνει τα κλειδιά των πινάκων διαστάσεων (dimension tables) καθώς και μέτρα (measures). Μια απλή καταχώρηση στον πίνακα συμβάντων αποτελεί το ελάχιστο επίπεδο ανάλυσης και καλείται *συμβάν* (fact).

Για παράδειγμα, ο σκοπός της παραδοσιακής αποθήκης δεδομένων που παρουσιάζεται στην Εικόνα 3-1 είναι η αποθήκευση συγκεντρωτικών δεδομένων σχετικά με συναλλαγές πωλήσεων που λαμβάνουν χώρα σε διάφορα καταστήματα της χώρας. Υπάρχει ο πίνακας συμβάντων *Sales Fact Table* που περιέχει κλειδιά για τους τέσσερις πίνακες διαστάσεων και τρία μέτρα: *#transactions* που μετρά τον αριθμό των συναλλαγών-πωλήσεων (με άλλα λόγια, τα καλάθια των πελατών), *total value* με τη συνολική αξία των πωλήσεων και *avg_quantity* που μετράει το μέσο αριθμό προϊόντων.



Εικόνα 3-1: Ένα απλό (συμβατικό) σχήμα κύβου δεδομένων.

Οι διαστάσεις αντιπροσωπεύουν τους άξονες ανάλυσης ενώ τα μέτρα τις μεταβλητές που αναλύονται με βάση τις διάφορες διαστάσεις. Για παράδειγμα στην Εικόνα 3-1, οι διαστάσεις είναι οι: *product, profile, store, time.* Σε ότι αφορά τη διάσταση του προφίλ (profile) αυτή αντιπροσωπεύει συγκεκριμένες ομάδες ανθρώπων με κοινά χαρακτηριστικά (ηλικία, φύλο, επάγγελμα κτλ). Έτσι σε αυτήν την περίπτωση, η αποθήκη δεδομένων αποθηκεύει τον αριθμό των συναλλαγών, το συνολικό ποσό από τις πωλήσεις και τη μέση ποσότητα για ένα συγκεκριμένο προϊόν που αγοράζεται από μια συγκεκριμένη ομάδα ανθρώπων, σε ένα συγκεκριμένο κατάστημα και σε μια συγκεκριμένη χρονική περίοδο.

Κάθε διάσταση οργανώνεται ως μια ιεραρχία (ή ακόμα και ως σύνολο ιεραρχιών) των επιπέδων της διάστασης και κάθε επίπεδο αντιστοιχεί σε ένα διαφορετικό επίπεδο κλιμάκωσης (granularity). Για παράδειγμα, το έτος είναι ένα επίπεδο στη διάσταση του χρόνου ενώ, η διάταξη <ημέρα, μήνας, έτος> ορίζει μια απλή ιεραρχία αυξανόμενης κλιμάκωσης στη διάσταση του χρόνου. Τέλος, τα μέλη ενός επιπέδου ιεραρχίας (π.χ. οι διαφορετικοί μήνες στη διάσταση του χρόνου) μπορούν να συσσωρευθούν

σε υψηλότερο επίπεδο κλιμάκωσης (π.χ. τα διαφορετικά έτη) και να αποτελέσουν μέλη αυτού του επιπέδου. Τα μέτρα επίσης συσσωρεύονται σε ακολουθώντας την ιεραρχία με τη βοήθεια μιας κατάλληλης συνάρτησης συσσώρευσης. Η ίδια προσέγγιση μπορεί να ακολουθηθεί και στις υπόλοιπες διαστάσεις.

Οι αποθήκες δεδομένων βελτιστοποιούνται για OLAP λειτουργίες. Τυπικές OLAP λειτουργίες περιλαμβάνουν τη συσσώρευση ή την εμβάθυνση/ εκλέπτυνση πληροφορίας σε μια διάσταση (*roll-up* και *drill-down*, αντίστοιχα), την επιλογή συγκεκριμένων τμημάτων του κύβου (*slicing* και *dicing*) και την αναδιάταξη της πολυδιάστατης όψης των δεδομένων στην οθόνη (*pivoting*) [Kim96].

Γενικά μιλώντας, οι αποθήκες δεδομένων και οι OLAP τεχνικές μπορούν να αξιοποιηθούν ώστε να μετατρέψουν μεγάλες ποσότητες πρωτογενών δεδομένων σε χρήσιμη γνώσης. Οι παραδοσιακές τεχνικές όμως δε σχεδιάστηκαν για να αναλύουν δεδομένα τροχιών. Έτσι, υπάρχει ανάγκη επέκτασης της τεχνολογίας των αποθηκών δεδομένων έτσι ώστε να είναι δυνατή η εφαρμογή του σε δεδομένα κίνησης. Σε αυτό το κεφάλαιο συζητούμε όλα τα απαραίτητα βήματα για την ανάπτυξη Αποθηκών Δεδομένων Τροχιών Κινούμενων Αντικειμένων. Ενδεικτικά, μια τέτοια αποθήκη θα μπορούσε να αναλύσει μέτρα όπως ο αριθμός των οχημάτων σε συγκεκριμένες χωρικές περιοχές, η μέση επιτάχυνση των οχημάτων, η μέγιστη και μέση ταχύτητα των οχημάτων. Αυτή η ανάλυση θα μπορούσε να επιτευχθεί μέσω κατάλληλων διαστάσεων (π.χ. μια χωρική και χρονική διάσταση) που θα μας επιτρέψει να εξερευνήσουμε συγκεντρωτικά δεδομένα κάτω από διαφορετικά επίπεδα κλιμάκωσης.

3.2. Κίνητρο

Το κίνητρο πίσω από μια TDW είναι η μετατροπή πρωτογενών τροχιών σε σημαντικές πληροφορίες που μπορούν να χρησιμοποιηθούν για σκοπούς υποστήριξης αποφάσεων σε κινητές εφαρμογές όπως LBS (Υπηρεσίες Θέσης/ Location-Based Services), διαχείριση κυκλοφορίας κτλ. Διαισθητικά, ο υψηλός όγκος πρωτογενών δεδομένων που παράγεται από τεχνολογίες αισθητήρων και τεχνολογίες εντοπισμού θέσης, η πολύπλοκη φύση των δεδομένων που αποθηκεύονται σε βάσεις τροχιών και οι εξειδικευμένες απαιτήσεις επεξεργασίας ερωτημάτων, καθιστούν την εξαγωγή πολύτιμης πληροφορίας από τα χωροχρονικά δεδομένα μια δύσκολη εργασία. Για το λόγο αυτό, μια ιδέα μπορεί να περιλαμβάνει την επέκταση των παραδοσιακών τεχνικών συσσώρευσης ώστε να παράγουν συνοπτικές πληροφορίες για τις τροχιές και την παροχή ανάλυση τύπου OLAP.

Η επέκταση των παραδοσιακών (π.χ. μη χωρικών), χωρικών ή χωροχρονικών μοντέλων ώστε να συμπεριλάβουν σημασιολογικά χαρακτηριστικά που καθοδηγούνται από τις τροχιές, εισάγει συγκεκριμένες απαιτήσεις αφού ο στόχος είναι διττός: να υποστηρίξει υψηλού επιπέδου OLAP ανάλυση και να διευκολύνει την εξόρυξη γνώσης από τις TDW. Έχοντας στο μυαλό ότι τα βασικά συστατικά ανάλυσης σε μια TDW (δηλαδή τα συμβάντα) είναι οι τροχιές, σε αυτό το κεφάλαιο εξειδικεύουμε τις απαιτήσεις στα θέματα της μοντελοποίησης, της ανάλυσης και διαχείρισης. Η πρώτη κατηγορία απαιτήσεων αφορά προκλήσεις αναφορικά με τα λογικά και τα εννοιολογικά συστατικά σε μια TDW, η δεύτερη αναφέρεται σε απαιτήσεις γύρω από την OLAP ανάλυση ενώ, η τρίτη εστιάζει σε πιο τεχνικά θέματα. Για κάθε μια περιοχή παρέχεται η σχετική βιβλιογραφία.

3.2.1. Θέματα Μοντελοποίησης Κύβου Δεδομένων

Στις ακόλουθες παραγράφους διερευνούμε τις απαιτήσεις και τους περιορισμούς που πρέπει να ληφθούν υπόψη κατά το σχεδιασμό μιας TDW τόσο για να καλυφθούν οι απαιτήσεις του χρήστη (εννοιολογικό μοντέλο) αλλά και για να είναι δυνατό να προκύψει ένα σύστημα που θα μπορεί να λειτουργεί ανεξάρτητα από συγκεκριμένες πλατφόρμες (λογικό μοντέλο).

Θεματικά, χωρικά, χρονικά μέτρα

Σε ότι αφορά τη μοντελοποίηση, μια τροχιά είναι ένα χωρικό αντικείμενο του οποίου η τοποθεσία μεταβάλλεται κατά τη διάρκεια του χρόνου (θυμίζουμε και τη σχετικά συζήτηση σχετικά με τη φύση των τροχιών που παραθέσαμε στην Ενότητα 2.3. Την ίδια στιγμή, οι τροχιές έχουν θεματικές ιδιότητες που συνήθως εξαρτώνται από το χώρο και το χρόνο. Αυτό υπονοεί ότι διαφορετικά χαρακτηριστικά των τροχιών θα πρέπει να καταγραφούν ώστε να μπορέσουν να αναλυθούν. Έτσι, κάνουμε τον εξής διαχωρισμό:

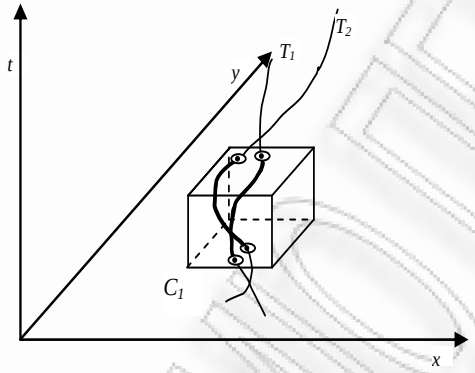
- αριθμητικά χαρακτηριστικά, όπως η μέση ταχύτητα της τροχιάς, η κατεύθυνσή της και η διάρκειά της,
- χωρικά χαρακτηριστικά, όπως το γεωμετρικό σχήμα της τροχιάς,
- χρονικά χαρακτηριστικά, όπως το χρονικό αποτύπωμα της κίνησης και
- χωροχρονικά χαρακτηριστικά, όπως η αντιπροσωπευτική τροχιά ή μια συστάδα τροχιών.

Επίσης, αν ενδιαφερόμαστε για θέματα αβεβαιότητας και ανακρίβειας, το μοντέλο TDW θα πρέπει να περιλαμβάνει μέτρα που εκφράζουν το βαθμό της αβεβαιότητα που ενσωματώνεται στην TDW εξαιτίας της ανακρίβειας των πρωτογενών δεδομένων. Η αβεβαιότητα θα πρέπει να μελετηθεί σε διαφορετικά επίπεδα κλιμάκωσης αφού υπάρχουν ειδικοί τελεστές συσσώρευσης που διαχέουν την αβεβαιότητα στα διάφορα επίπεδα. Λαμβάνοντας υπόψη τις απαιτήσεις των χρηστών και της εφαρμογής, διάφορα αριθμητικά μέτρα μπορούν να ληφθούν υπόψη:

- ο αριθμός των τροχιών που βρίσκονται μέσα στο ένα κελί (ή ξεκίνησαν/ολοκλήρωσαν τη διαδρομή τους μέσα στο κελί, ή πέρασαν/μπήκαν/έφυγαν από ένα κελί κτλ)
    - Για παράδειγμα, στην Εικόνα 3-2: καμία τροχιά δεν ξεκίνησε/ δεν ολοκληρώθηκε στο κελί C_1, δυο τροχιές πέρασαν και πέρασαν, μπήκαν και έφυγαν από το κελί αυτό.
- η {μέση/ελάχιστη/μέγιστη} απόσταση που διανύεται από τις τροχιές μέσα σε ένα κελί
    - για παράδειγμα, στην Εικόνα 3-2: η μέση απόσταση που διανύεται είναι το σύνολο του μήκους των τμημάτων των τροχιών που βρίσκονται μέσα στο C_1 (οι έντονες γραμμές) δια τον αριθμό των τροχιών που καταγράφηκαν μέσα σε αυτό το κελί (δυο τροχιές).
- ο {μέσος/ελάχιστος/μέγιστος} χρόνος που απαιτείται για την κάλυψη αυτής της απόστασης.
    - για παράδειγμα, στην Εικόνα 3-2: η μέση διάρκεια των τροχιών μέσα στο C_1 είναι το άθροισμα των διαρκειών των τμημάτων των τροχιών μέσα στο C_1 (οι έντονες

γραμμές) δια τον αριθμό των τροχιών που καταγράφηκαν μέσα σε αυτό το κελί (δυο τροχιές).

Άλλα μέτρα μπορεί να περιλαμβάνουν χαρακτηριστικά κίνησης των τροχιών π.χ. ταχύτητα και μεταβολή ταχύτητας (επιτάχυνσης), κατεύθυνση, και αλλαγή κατεύθυνσης (στροφή), χαρακτηριστικά που αφορούν το χωρικό πλαίσιο (π.χ. χρήση δικτύου, συχνότητα, πυκνότητα) και επίσης την αβεβαιότητα που σχετίζεται με τις θέσεις των αντικειμένων στη βάση. Προκειμένου να είναι δυνατός ο χειρισμός της αβεβαιότητας, η αποθήκη δεδομένων θα μπορούσε να περιέχει πληροφορίες σχετικά με την ποιότητα των πρωτογενών δεδομένων (π.χ. χωρική/χρονική ανοχή για τις εγγραφές).



Εικόνα 3-2: Τα τμήματα των τροχιών που βρίσκονται μέσα σε ένα κελί.

Ως τελευταία παρατήρηση για τα μέτρα, θα πρέπει να επισημάνουμε ότι ακόμα και στην περίπτωση των αριθμητικών μέτρων, η πολυπλοκότητα υπολογισμού τους μπορεί να διαφέρει σε κάθε περίπτωση. Μερικά μέτρα απαιτούν ελάχιστο προϋπολογισμό και μπορούν να ανανεώνονται στην αποθήκη δεδομένων ακόμα και όταν καταφτάνουν ελάχιστες καταγραφές τροχιών ενώ, κάποια άλλα απαιτούν ένα συγκεκριμένο όγκο τροχιών ώστε να είναι δυνατή η ανανέωση τους. Οι Braz κ.α. [BOO+07] προτείνουν την ακόλουθη κατηγοριοποίηση των μέτρων, λαμβάνοντας υπόψη το κόστος προϋπολογισμού:

i. κανένας προϋπολογισμός: το μέτρο μπορεί να ανανεώνεται χρησιμοποιώντας ακόμα και μια μοναδική καταγραφή,

ii. τοπικός προϋπολογισμός ανά τροχιά: το μέτρο ανανεώνεται αξιοποιώντας μόνο μερικές καταγραφές της ίδιας τροχιάς,

iii. συνολικός προϋπολογισμός ανά τροχιά: το μέτρο ανανεώνεται αξιοποιώντας όλες τις καταγραφές ανά τροχιά,

iv. συνολικός προϋπολογισμός ανά τροχιά: το μέτρο ανανεώνεται αξιοποιώντας όλες τις καταγραφές ανά τροχιά,

v. συνολικός προϋπολογισμός: το μέτρο ανανεώνεται αξιοποιώντας όλες τις καταγραφές όλων των τροχιών.

Για παράδειγμα:

- ο αριθμός των τροχιών που ξεκινάει/ ολοκληρώνεται μέσα σε ένα κελί είναι μέτρα τύπου i (όπως παρουσιάζεται πιο πάνω),
- αν σημειώσουμε το πρώτο/τελευταίο σημείο της τροχιάς, η απόσταση που καλύπτεται από τις τροχιές μέσα στο κελί και ο αριθμός των τροχιών που εισήλθαν, εξήλθαν από αυτό είναι μέτρα τύπου ii (στην Εικόνα 3-2 αυτά τα σημεία έχουν σημειωθεί για τις τροχιές T_1 και T_2),
- ο αριθμός των τροχιών που καλύπτουν μια συνολική απόσταση μεγαλύτερη από μια συγκεκριμένη τιμή *ν* είναι μέτρο τύπου c (π.χ. στην Εικόνα 3-2, η απόσταση που καλύπτεται από τα τμήματα των τροχιών μέσα στο κελί C_1 θα συγκριθεί με την τιμή *ν*),
- ο αριθμός των τροχιών που τέμνουν μια άλλη τροχιά μόνο μέσα σε ένα συγκεκριμένο κελί είναι μέτρο τύπου d (π.χ. στην Εικόνα 3-2 υπάρχει μια μόνο τομή στο κελί C_1)

Ο όγκος των προϋπολογισμών για κάθε τύπο μέτρου έχει ισχυρή επίπτωση στον όγκο της μνήμης που απαιτείται προκειμένου να αποθηκευτούν προσωρινά οι καταγραφές των τροχιών. Σημειώστε ότι από τη στιγμή που οι καταγραφές τροχιών μπορούν να φτάνουν ως ρεύματα δεδομένων μια διαφορετικές συχνότητες με ένα απρόβλεπτο και τυχαίο τρόπο, ο χαμηλός χρόνος επεξεργασίας και η περιορισμένη μνήμη είναι επίσης σημαντικοί περιορισμοί.

Παρόμοιες παρατηρήσεις μπορούν να βρεθούν στην [HSK98] όπου οι Han κ.α. παρουσιάζουν τρεις μεθόδους υπολογισμού χωρικών μέτρων στα πλαίσια κατασκευής ενός χωρικού κύβου δεδομένων. Η πρώτη περιλαμβάνει τη συλλογή και αποθήκευση των αντίστοιχων χωρικών δεδομένων χωρίς όμως να γίνεται κανένας προϋπολογισμός των χωρικών μέτρων. Η δεύτερη μέθοδος προϋπολογίζει και αποθηκεύει μερικές κατά προσέγγιση εκτιμήσεις/υπολογισμούς για τα χωρικά μέτρα ενός χωρικού κύβου. Για παράδειγμα, αν το μέτρο αποτελεί μια σύνθεση χωρικών αντικειμένων, θα μπορούσε να αποθηκευτεί το Ελάχιστο Περιβάλλον Ορθογώνιο (Minimum Bounding Rectangle) της σύνθεσης των αντικειμένων. Τέλος, κάποιος μπορεί επιλεκτικά να προϋπολογίσει μερικά χωρικά μέτρα. Στην τελευταία περίπτωση εγείρεται το θέμα της επιλογής ενός συνόλου χωρικών μέτρων για προϋπολογισμό. Κάποια κριτήρια για προϋπολογισμό ενός κυβοειδούς (cuboid) παρουσιάζονται στην [HSK98].

Θεματικές, χωρικές, χρονικές διαστάσεις

Μια τυπική TDW θα πρέπει να υποστηρίζει τη χωρική (π.χ.. συντεταγμένες, δρόμους, περιφέρειες, κελιά, πόλεις, περιοχές, χώρες) και τη χρονική (π.χ. δευτερόλεπτο, λεπτό, ώρα, ημέρα, μήνας και έτος) διάσταση και τις αντίστοιχες ιεραρχίες υποστηρίζοντας έτσι το υφιστάμενο χωροχρονικό πλαίσιο μέσα στο οποίο κινούνται οι τροχιές. Επιπλέον, είναι σημαντικό να επιτρέψουμε διαστάσεις που σχετίζονται με τον χώρο/χρόνο να αλληλεπιδρούν με θεματικές διαστάσεις περιγράφοντας έτσι επιπλέον πληροφορίες που σχετίζονται με τις τροχιές. Για παράδειγμα, τεχνογραφικά (π.χ. η κινητή συσκευή που χρησιμοποιήθηκε) ή δημογραφικά δεδομένα (π.χ. ηλικία και φύλο των χρηστών) [MFN+08a]. Αυτό επιτρέπει σε έναν αναλυτή να υποβάλλει στην TDW όχι μόνο ερωτήματα, για παράδειγμα, για τον αριθμό των αντικειμένων που πέρασαν από μια περιοχή ενδιαφέροντος αλλά και να μπορεί να αναγνωρίσει τα αντικείμενα που επιστρέφονται ως απάντηση.

Αυτό είναι ιδιαίτερα σημαντικό αφού στην πρώτη περίπτωση απλά παίρνουμε ποσοτικές πληροφορίες ενώ στη δεύτερη, λαμβάνουμε ποιοτικές πληροφορίες. Συνεπώς, ένα πλούσιο σχήμα TDW μπορεί να περιλαμβάνει τις επόμενες διαστάσεις:

- χρονική (χρόνος),
- γεωγραφική (τοποθεσία),
- δημογραφική (π.χ. φύλο, ηλικία, επάγγελμα, οικογενειακή κατάσταση, ταχυδρομικός κώδικας οικίας, ταχυδρομικός κώδικας εργασίας κτλ),
- τεχνογραφική (π.χ. κινητές συσκευές, GPRS δυνατότητες, συνδρομή σε ειδικές υπηρεσίες κτλ)

Σε ότι αφορά τις διαστάσεις δημογραφικών και τεχνογραφικών χαρακτηριστικών, η ιδέα πίσω από αυτές είναι ο εμπλουτισμός της αποθήκης δεδομένων με σημασιολογικές πληροφορίες. Αυτές οι διαστάσεις επιτρέπουν την ομαδοποίηση των τροχιών σύμφωνα με τα δημογραφικά τους χαρακτηριστικά ή/και βάσει των τεχνολογικών χαρακτηριστικών των συσκευών.

Ένα ανοικτό θέμα που αφορά τον ορισμό των διαστάσεων είναι το επιλεγμένο επίπεδο λεπτομέρειας σε κάθε περίπτωση. Ας θεωρήσουμε τη χωρική διάσταση: ως κατώτατο επίπεδο μπορούμε να θεωρήσουμε τις χωρικές συντεταγμένες από τη στιγμή που μια τροχιά αποτελεί ένα σύνολο δειγματικών θέσεων στο χρόνο για τις οποίες οι μεταξύ τους θέσεις υπολογίζονται με κάποιο τρόπο παρεμβολής. Επειδή όμως αυτό όμως έχει ως αποτέλεσμα μια τεράστια διακριτοποίηση της χωρικής διάστασης, επιλέγουμε να ακολουθήσουμε πιο γενικές προσεγγίσεις. Για παράδειγμα, μπορούν να ληφθούν υπόψη οι θέσεις των κελιών αντί για τις θέσεις των σημείων.

Ιεραρχίες στις διαστάσεις

Από τη στιγμή που έχουν οριστεί οι διαστάσεις, οι ιεραρχίες τους μπορούν να οριστούν αποκλειστικά από τους χρήστες ή να παραχθούν με ένα αυτόματο τρόπο εφαρμόζοντας τεχνικές συσταδοποίησης δεδομένων ή άλλες τεχνικές ανάλυσης. Μια γενική τεχνική που ακολουθείται για τον ορισμό ιεραρχιών περιλαμβάνει τη διακριτοποίηση του εύρους τιμών των διαστάσεων που οδηγεί σε μια ιεραρχία ομάδων. Με τον τρόπο αυτό μπορούμε να εξασφαλίσουμε μερική διάταξη μεταξύ αυτών των ομάδων τιμών. Ας αναλύσουμε τώρα τις διαφορετικές προτάσεις για δημιουργία ιεραρχιών για τις διαστάσεις που προτάθηκαν στο προηγούμενο κεφάλαιο καθώς και τις δυσκολίες που αυτές περιλαμβάνουν.

Ο ορισμός ιεραρχιών στη χρονική διάσταση είναι κάτι απλό από τη στιγμή που υπάρχει μια προφανής διάταξη μεταξύ των διαφορετικών επιπέδων της διάστασης. Για παράδειγμα, μια πιθανή ιεραρχία θα μπορούσε να είναι Έτος > Τρίμηνο > Μήνας > Ημέρα > Ώρα > Λεπτό > Δευτερόλεπτο. Άλλες ιεραρχίες στη διάσταση του χρόνου θα μπορούσαν να αφορούν τις εποχές, ζώνες ωρών, ώρες κυκλοφοριακής συμφόρησης κτλ.

Από την άλλη μεριά, η δημιουργία ιεραρχιών σε χωρικά δεδομένα απαιτεί μια πιο πολύπλοκη διαδικασία. Στη πραγματικότητα, είναι πιθανό να υπάρχουν μη σαφώς ορισμένες ιεραρχίες στα χωρικά δεδομένα. Για παράδειγμα, στην ιεραρχία Χώρα > Πόλη > Περιοχή > Κυψέλη > Οδός δεν υπάρχει

σχέση συμπερίληψης (inclusion relation) μεταξύ Περιοχής και Κυψέλης και μεταξύ Κυψέλης και Οδού αφού για παράδειγμα μια Οδός μπορεί να διασχίζει περισσότερα από μια Κυψέλες. Προκειμένου να λύσουν αυτό το πρόβλημα οι Jensen κ.α. [JKP+04] πρότειναν ένα εννοιολογικό μοντέλο που υποστηρίζει διαστάσεις με σχέσεις πλήρους ή μερικής συμμετοχής. Έτσι, όταν παρατηρείται μια σχέση μερικής συμμετοχής μεταξύ διαφορετικών επιπέδων μιας ιεραρχίας θα πρέπει να προσδιοριστεί ο βαθμός συμμετοχής, για παράδειγμα το 80% αυτής της Οδού καλύπτεται από αυτή τη Κυψέλη.

Ανάλογα με την εφαρμογή, εκτός από την τυπική σχέση Πόλη < Χώρα, μπορούν να οριστούν επιπλέον ιεραρχίες στη χωρική διάσταση, για παράδειγμα ιεραρχίες στις Περιοχές ανάλογα με το επίπεδο μόλυνσης σε αυτές.

Τέλος, σε ότι αφορά τις διαστάσεις δημογραφικών και τεχνογραφικών χαρακτηριστικών, η απλούστερη λύση η δημιουργία μιας ιεραρχίας για κάθε διάσταση. Αυτή η λύση όμως μπορεί να προκαλέσει προβλήματα πολυπλοκότητας ειδικά αν ο αριθμός των διαστάσεων είναι μεγάλος. Μια άλλη πιθανότητα είναι ο συνδυασμός χαρακτηριστικών από αυτές τις διαστάσεις δημιουργώντας ομάδες τιμών. Για παράδειγμα μπορούμε να θεωρήσουμε την ακόλουθη ομαδοποίηση: "φύλο = γυναίκα, ηλικία = 25 - 35, οικογενειακή κατάσταση = άγαμη". Αυτή η ομαδοποίηση μπορεί να πραγματοποιηθεί από κάποιον ειδικό ή εφαρμόζοντας στατιστική προεπεξεργασία στα δεδομένα. Αυτή η προσέγγιση ελαττώνει τον αριθμό των διαστάσεων επιτρέποντας έτσι ένα απλούστερο και πιο αποδοτικό σχήμα αποθήκης δεδομένων σε ότι αφορά θέματα χρόνου επεξεργασίας και απαιτήσεις αποθηκευτικού χώρου.

Τέλος, κάποιες προσεγγίσεις [JKP+04], [MZ04b] προτείνουν την υποστήριξη πολλαπλών ιεραρχιών για κάθε μια διάσταση. Αυτές οι εργασίες εστιάζουν στις ιεραρχίες που μπορούν να οριστούν στη χωρική διάσταση. Πιο συγκεκριμένα, η χωρική διάσταση μπορεί να περιλαμβάνει μη σαφώς ορισμένες ιεραρχίες. Έτσι, είναι δυνατό να υπάρξουν πολλαπλές διαδρομές συσσώρευσης που μπορούν να ληφθούν υπόψη κατά τη διάρκεια των OLAP πράξεων.

3.2.2. Απαιτήσεις σχετικά με το OLAP

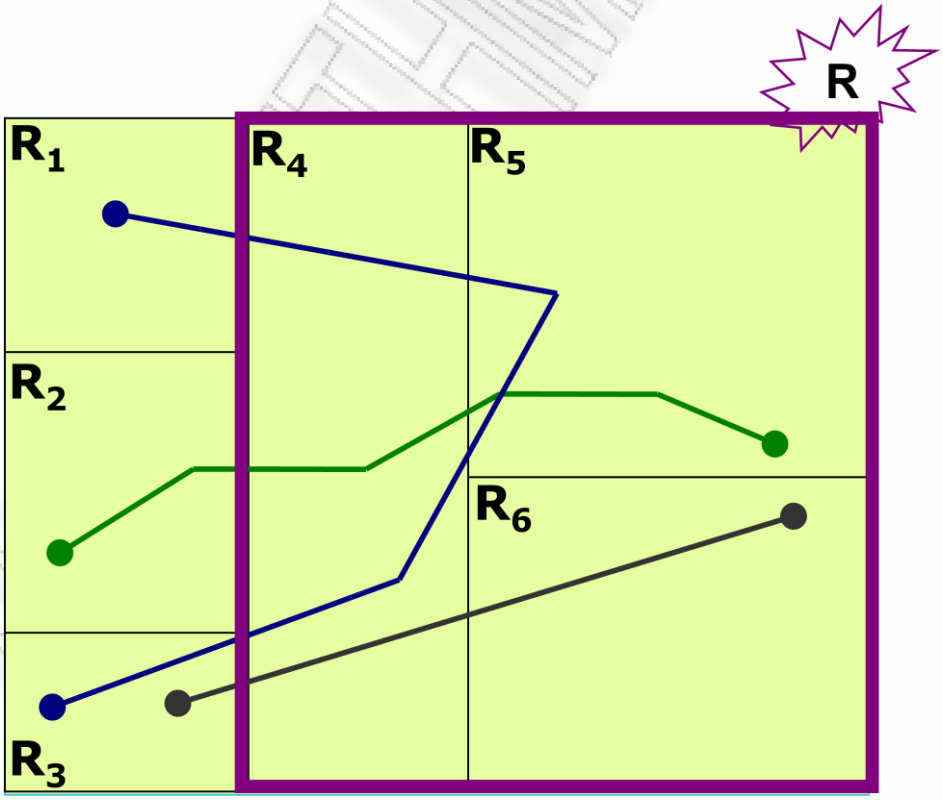
Στις παραδοσιακές αποθήκες δεδομένων, η ανάλυση δεδομένων είναι μια διαδραστική διαδικασία που πραγματοποιείται με την εφαρμογή OLAP λειτουργιών. Στις χωρικές αποθήκες δεδομένων, έχουν οριστεί ειδικές OLAP λειτουργίες για να αντιμετωπίσουν τις ιδιαιτερότητες του χώρου αυτού [PTK+02]. Αντιστοίχως, στην περίπτωση μας, αναζητούμε ένα σύνολο OLAP λειτουργιών με στόχο την ανάλυση δεδομένων τροχιών. Ένα τέτοια σύνολο θα πρέπει να περιλαμβάνει όχι μόνο τις παραδοσιακές λειτουργίες, όπως συσσώρευση, εμβάθυνση και επιλογή τροχιών αλλά και επιπρόσθετες λειτουργίες που λαμβάνουν υπόψη την ιδιαιτερότητα των χωροχρονικών δεδομένων. Παρακάτω παρουσιάζουμε αυτές τις λειτουργίες με περισσότερες λεπτομέρειες:

Συσσώρευση (Roll-up). Η λειτουργία συσσώρευσης μας επιτρέπει να πλοηγηθούμε από ένα λεπτομερές σε ένα πιο γενικό αφαιρετικό επίπεδο είτε «ανεβαίνοντας» στην ιεραρχία (π.χ. από το επίπεδο της «Πόλης» στο επίπεδο της «Χώρας») είτε εφαρμόζοντας μείωση διαστάσεων (π.χ. αγνοώντας τη διάσταση του χρόνου και εφαρμόζοντας συσσώρευση στη διάσταση της τοποθεσίας).

Όπως παρουσιάζεται στην [BOO+07], ανάλογα με το είδος των μέτρων που θέλουμε να χρησιμοποιήσουμε για ανάλυση, η λειτουργία συσσώρευσης σε μια TDW μπορεί να εισάγει κάποιο σφάλμα. Υποθέτοντας ότι οι προσδιοριστές των τροχιών ή των αντικειμένων δεν καταγράφονται κατά τη διάρκεια εφαρμογής κάποιας συνάρτησης συσσώρευσης τότε δεν είναι δυνατό να υπολογιστεί ο αριθμός μοναδικών τροχιών αφού υπάρχει μόνο συγκεντρωτική πληροφορία. Αυτή είναι μια ειδική περίπτωση του προβλήματος της μοναδικής προσμέτρησης (distinct count problem) [TKC+04] που παρουσιάζεται στην Υποενότητα 3.5.2.1.

Ας θεωρήσουμε ότι η Εικόνα 3-3 παρουσιάζει την χωρική προβολή μερικών κελιών του κύβου. Η TDW αποθηκεύει για παράδειγμα τον αριθμό των μοναδικών τροχιών σε κάθε κελί. Έτσι, έχουμε δυο μοναδικές τροχιές στο R_4, δύο στο R_5 και μια στο R_6. Αν προσπαθήσουμε να εφαρμόσουμε παραδοσιακή συσσώρευση τότε θα μετρήσουμε έξι μοναδικές τροχιές αντί για τρεις που είναι η σωστή απάντηση.

Ένα ακόμη ανοικτό θέμα είναι η εφαρμογή της λειτουργίας συσσώρευσης όταν παρουσιάζονται δεδομένα με αβεβαιότητα. Πράγματι, κατά τη διάρκεια της συσσώρευσης δυο παράγοντες θα πρέπει να λαμβάνονται υπόψη. Ο πρώτος αφορά την αβεβαιότητα που συσχετίζεται με τιμές των διαστάσεων και των μέτρων και η οποία μεταφέρεται από τη πηγή στην αποθήκη δεδομένων. Ο δεύτερος αναφέρεται στην αβεβαιότητα που παρουσιάζεται στην αποθήκη δεδομένων λόγω των μη σαφώς ορισμένων ιεραρχιών.



Εικόνα 3-3: Εφαρμόζοντας συσσώρευση στον κύβο.

Εμβάθυνση (Drill-down). Πρόκειται για την αντίστροφη λειτουργία της συσσώρευσης. Μας επιτρέπει να πλοηγηθούμε από ένα λιγότερο σε ένα περισσότερο λεπτομερές επίπεδο είτε κατεβαίνοντας στην

ιεραρχία (π.χ. από το επίπεδο της «Χώρας» στο επίπεδο της «Πόλης») είτε εισάγοντας επιπλέον διαστάσεις (π.χ. λαμβάνοντας υπόψη όχι μόνο τη χωρική αλλά και τη χρονική διάσταση). Αντίστοιχα με τη συσσώρευση, στην εμβάθυνση μπορούμε επίσης να συναντήσουμε το πρόβλημα της μοναδικής προσμέτρησης αλλά και αβεβαιότητα που μπορεί να σχετίζεται τόσο με τις τιμές όσο και με τις σχέσεις. Όπως ήδη αναφέραμε, στην Εικόνα 3-3, ο αριθμός των μοναδικών τροχιών στο R δεν είναι ίδιος με τον αριθμό των μοναδικών τροχιών στις περιοχές R_4, R_5 και R_6 στις οποίες έχουμε εμβαθύνει.

Τεμαχισμός (Slice), Κομμάτιασμα (Dice). Η λειτουργία του τεμαχισμού πραγματοποιεί μια επιλογή σε μια διάσταση (π.χ. «Πόλη = Αθήνα») ενώ, το κομμάτιασμα αφορά επιλογές σε δυο η περισσότερες διαστάσεις (π.χ. «Πόλη = Αθήνα και Έτος = 2006»). Οι συνθήκες μπορούν να αφορούν όχι μόνο αριθμητικές τιμές αλλά και πολύπλοκα κριτήρια όπως χωρικά ή/και χρονικά παράθυρα. Για να υποστηρίξουμε τέτοιες λειτουργίες, τα κριτήρια επιλογής πρέπει να ενσωματωθούν σε ένα ερώτημα προς την TDW το οποίο πρέπει να επεξεργαστεί με κατάλληλες μεθόδους. Συνοψίζοντας, οι παραδοσιακές OLAP λειτουργίες θα πρέπει να υποστηρίζονται και από μια TDW αφού μπορούν να μας παρέχουν σημαντικές πληροφορίες. Ένα άλλο κίνητρο μας της έρευνας μας είναι η αναζήτηση επιπλέον λειτουργιών που μπορούν να χρησιμοποιηθούν αποκλειστικά για τροχιές. Ενδεικτικά παραδείγματα:

- λειτουργίες που μεταβάλλουν δυναμικά τη χωροχρονική κλιμάκωση των μέτρων που αναπαριστούν τροχιές,
- λειτουργίες εύρεσης μέσου (medoid) που εφαρμόζουν προηγμένες μεθόδους συνάθροισης όπως η συσταδοποίηση τροχιών με στόχο την εξαγωγή των αντιπροσωπευτικών από ένα σύνολο τροχιών,
- λειτουργίες που διαδίδουν/ συναθροίζουν την αβεβαιότητα και την ανακρίβεια που μπορεί να παρουσιάζεται στα δεδομένα μια TDW.

Σε ότι αφορά το πρώτο θέμα, παρουσιάζουμε στην Ενότητα 3.4 ένα πλήρες πλαίσιο που επιτρέπει στα πλαίσια μιας TDW τον ορισμό της χωροχρονικής κλιμάκωσης των τροχιών με ένα δυναμικό τρόπο. Το δεύτερο θέμα συζητείται στην Ενότητα 6.2 όπου περιγράφεται το μέτρο της αντιπροσωπευτικής τροχιάς ως ανοικτό θέμα. Παρόλο που η αντιμετώπιση θεμάτων αβεβαιότητα είναι πέρα από τους σκοπούς της παρούσας διατριβής, συμπεριλάβαμε αυτό το αντικείμενο ως μια πιθανή απαίτηση στον πραγματικό κόσμο.

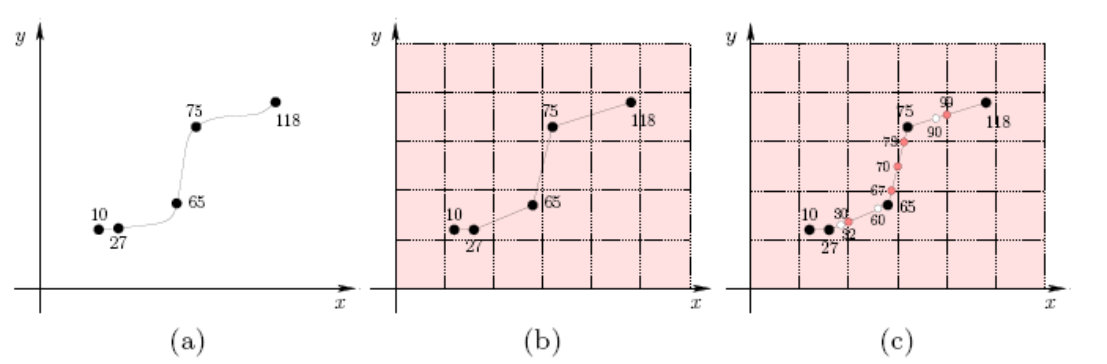
3.2.3. Απαιτήσεις που αφορούν τη Διαχείριση Δεδομένων: θέματα ETL, υποστήριξη συνεχών ρευμάτων δεδομένων και πολλαπλές χωρικές τοπολογίες

Στις προηγούμενες παραγράφους περιγράφηκαν, σε υψηλό επίπεδο, οι απαιτήσεις για μια TDW όπως αυτές προκύπτουν από την επέκταση των εννοιολογικών και λογικών μοντέλων αποθηκών δεδομένων. Σε αυτό το κεφάλαιο αναζητούμε τις απαιτήσεις διαχείρισης που έχει μια TDW χωρίς να περιορίζουμε τη συζήτηση σε συγκεκριμένα μοντέλα φυσικής σχεδίασης.

Έχοντας ως κύριο στόχο το χτίσιμο μια αποθήκης δεδομένων ειδικά για τροχιές και λαμβάνοντας υπόψη την πολυπλοκότητα και τον τεράστιο όγκο των τροχιών δεδομένων, θα πρέπει να διαφοροποιήσουμε τον αρχιτεκτονικό σχεδιασμό της TDW από αυτόν μιας παραδοσιακής ΑΔ. Η

κατάσταση περιπλέκεται αν λάβουμε υπόψη μας και το ότι ως πηγές δεδομένων μπορούμε να θεωρήσουμε και ρεύματα δεδομένων, όπως καταγραφές από συσκευές επικοινωνίας και εντοπισμού θέσης που μπορούν να καταφτάνουν ως σύνολο μη προκαθορισμένου όγκου. Έτσι, η αποτελεσματική και αποδοτική αποθήκευση των τροχιών σε μια αποθήκη δεδομένων θα πρέπει να μελετηθεί ώστε να είναι δυνατός ο χειρισμός ρευμάτων πρωτογενών δεδομένων. Επίσης η TDW θα πρέπει να μπορεί να εξυπηρετήσει εργασίες ανάλυσης και εξόρυξης γνώσης. Αυτό εισάγει επιπλέον προκλήσεις όπως τη δυνατότητα αυξητικής επεξεργασίας των ρευμάτων δεδομένων με αποδοτικό και ακριβή τρόπο καθώς και τον καθορισμό προσαρμοστικών στρατηγικών όπου θα επιτρέψουν στους κύβους δεδομένων να «εξελίσσονται» βάσει των ρευμάτων δεδομένων.

Επίσης, λόγω των ιδιαιτεροτήτων των τροχιών, μερικά προβλήματα μπορούν να προκύψουν στη φάση φόρτωσης του πίνακα συμβάντων. Για να δώσουμε μια διαισθητική ιδέα αυτών των ζητημάτων, θεωρήστε έναν κύβο δεδομένων όπου τα συμβάντα είναι οι τροχιές, αλλά η χωρική και η χρονική διάσταση έχει διακριτοποιηθεί σύμφωνα με ένα απλό πλέγμα, και ως μέτρο θεωρείται ο αριθμός των μοναδικών τροχιών στο χωροχρονικό κελί που προκύπτει από το πλέγμα.



Εικόνα 3-4: (a) Δειγματοληψία δισδιάστατης τροχιάς (b) Γραμμική παρεμβολή μιας τροχιάς (c) Η τροχιά που προέκυψε από την παρεμβολή με τα σημεία που ταιριάζουν με την ελάχιστη χωρική και χρονική κλιμάκωση.

Επιπλέον, υποθέστε ότι μια τροχιά μοντελοποιείται ως ένα πεπερασμένο σύνολο παρατηρήσεων, δηλ. ένα πεπερασμένο υποσύνολο των σημείων που λαμβάνονται από την πραγματική συνεχή τροχιά, αυτό που καλείται δειγματοληψία. Για παράδειγμα, η Εικόνα 3-4(a) παρουσιάζει τη δειγματοληψία μιας τροχιάς.

Τα κύρια θέματα που πρέπει να ληφθούν υπόψη είναι τα εξής:

- σε μια δειγματοληψία, οι τραχιές παρατηρήσεις δεν μπορούν να χρησιμοποιηθούν άμεσα για να υπολογίσουν τα μέτρα ενδιαφέροντος με σωστό τρόπο, και
- αυτές οι παρατηρήσεις δεν είναι ανεξάρτητα σημεία, το γεγονός ότι ανήκουν στην ίδια τροχιά πρέπει να χρησιμοποιηθεί κατά την υπολογισμό μερικών μέτρων.

Για παράδειγμα, η φόρτωση του πίνακα συμβάντων με τα σημεία που απεικονίζονται στην Εικόνα 3-4(b) οδηγεί σε ένα απλό πίνακα συμβάντων (Εικόνα 3-5). Να ληφθεί υπόψη ότι η πρώτη στήλη του πίνακα δεν ανήκει στον πίνακα συμβάντων αλλά χρησιμοποιείται για να διευκρινίσει ποιες

παρατηρήσεις εμπίπτουν στο συγκεκριμένο χωροχρονικό κελί. Είναι εμφανές ότι επιπλέον κελιά μπορούν διασχίζονται από την τροχιά (π.χ., το κελί [60; 90) x [60; 90) x [60; 90)), κάτι που συνεπάγεται ότι μερικές πληροφορίες ίσως και να λείπουν. Αφετέρου , το ίδιο κελί μπορεί να περιέχει περισσότερες από μια παρατηρήσεις οπότε ο υπολογισμός του μέτρου μπορεί να είναι λανθασμένος επειδή δεν αποθηκεύεται ο αριθμός των μοναδικών τροχιών (δείτε το κελί [30; 60) x [30; 60) x [0; 30)).

Time label	X Interval	Y Interval	T Interval	N Trajs
10,27	[30,60)	[30,60)	[0,30)	2
65	[60,90)	[30,60)	[60,90)	1
75	[90,120)	[90,120)	[60,90)	1
118	[120,150)	[90,120)	[60,120)	1

Εικόνα 3-5: Ένα απλός πίνακας συμβάντων για μια αποθήκη τροχιών.

Προκειμένου να λυθεί το πρώτο πρόβλημα, οι Braz κ.α.. [BOO+07] προτείνουν την προσθήκη επιπλέον ενδιάμεσων σημείων παρεμβάλλοντας γραμμικά την τροχιά. Τα σημεία που προστίθενται είναι αυτά που τέμνουν τα σύνορα του χωροχρονικού κελιού και στις τρεις διαστάσεις. Η Εικόνα 3-4(c) παρουσιάζει τα σημεία που παρεμβλήθηκαν με άσπρους και γκρίζους κύκλους. Σημειώστε ότι τα άσπρα σημεία, που συνδέονται με τις χρονικές ετικέτες 30, 60, και 90, έχουν προστεθεί για να ταιριάξουν με την κλιμάκωση της χρονικής διάστασης. Στην πραγματικότητα, αντιστοιχούν στα σημεία τομής των χρονικών συνόρων του τριδιάστατου κελιού. Επιπλέον, τα γκρίζα σημεία, που ονομάζονται ως 32, 67, 70, 73, και 99, έχουν εισαχθεί ώστε ταιριάξουν με τις χωρικές διαστάσεις. Αντιστοιχούν στα σημεία τομής των χωρικών συνόρων κάποιου τριδιάστατου κελιού, ή, ισοδύναμα, τα σημεία τομής των χωρικών διδιάστατων τετραγώνων που απεικονίζονται στην Εικόνα 3-4(c). Το δεύτερο πρόβλημα είναι πιο σύνθετο και αφορά τις διπλομετρήσεις. Μια προσέγγιση για την αντιμετώπιση του παρουσιάζεται στην Υποενότητα 3.3.2. Μια λεπτομερής συζήτηση για τα σφάλματα στον υπολογισμό των διαφορετικών μέτρων σχετικών με τα ζητήματα που παρουσιάστηκαν μπορεί να βρεθεί στην [BOO+07].

Ένας παράγοντας που χαρακτηρίζει μια TDW είναι η αλληλεξάρτηση μεταξύ της ανάπτυξης των τροχιών επάνω στις διάφορες πιθανές χωρικές τοπολογίες που αντιπροσωπεύονται από τις αντίστοιχες χωρικές διαστάσεις. Ο διαχωρισμός βασικών επιπέδων (base levels) μιας χωρικής τοπολογίας έχει άμεσες επιπτώσεις στην πολυδιάστατη ανάλυση των τροχιών. Οι πιθανές διαθέσιμες τοπολογίες μπορούν να είναι από απλά πλέγματα (π.χ. τεχνητός διαχωρισμός), έως σύνθετες πολύγωνες συγχωνεύσεις (π.χ. προάστια μιας πόλης), πραγματικά οδικά δίκτυα και δίκτυα κινητής τηλεφωνίας. Η πρώτη περίπτωση είναι η απλούστερη δεδομένου ότι το διάστημα διαιρείται στους ρητά καθορισμένους τομείς ενός πλέγματος και έτσι είναι εύκολο να διατεθούν τα σημεία τροχιάς στις συγκεκριμένες περιοχές. Εντούτοις, ο υπολογισμός του αριθμού αντικειμένων που πέρασε από μια περιοχή μπορεί να αποδειχθεί δύσκολος για μια TDW. Αυτό συμβαίνει επειδή η συχνότητα δειγματοληψίας μπορεί να μη βοηθάει στην αντιπροσώπευση της πραγματικής τροχιάς [BOO+07]. Κατά συνέπεια, μπορεί να είναι απαραίτητο να ανακατασκευαστεί η τροχιά (στα πλαίσια του ETL) για να προστεθούν τα ενδιάμεσα σημεία μεταξύ των στοιχείων δειγματοληψίας (δείτε την Εικόνα 3-4(c)).

Στην περίπτωση των οδικών δικτύων, οι τροχιές πρέπει να ανακατασκευαστούν ώστε να είναι περιορισμένες στο δίκτυο των δρόμων, ενώ η διαχείριση των τηλεπικοινωνιακών κυψελών είναι ένα πιο σύνθετο πρόβλημα μιας και οι περιοχές που καλύπτονται από τις κυψέλες μπορούν να αλλάξουν από στιγμή σε στιγμή ανάλογα με την ισχύ του σήματος που εκπέμπεται από τους σταθμούς βάσης του παρόχου. Οποιαδήποτε και να είναι τα βασικά επίπεδα της χωρικής διάστασης όλες οι χωρικές τοπολογίες υπόκεινται στο πρόβλημα της μοναδικής προσμέτρησης [TKC+04] που θα περιγραφεί αναλυτικά στην Υποενότητα 3.3.2.

Προφανώς, η ανακατασκευή των τροχιών και των πολλαπλών μετρήσεων ενός αντικειμένου που κινείται μέσα σε μια περιοχή εξαρτάται απευθείας από την παρεμβολή (π.χ. γραμμική, πολυωνυμική) που χρησιμοποιείται (ενδεχομένως) από το αντίστοιχο μοντέλο δεδομένων τροχιών. Η ανωτέρω συζήτηση υπονοεί ότι ένας αναλυτής έχει αρχικά τη δυνατότητα να αναλύσει μια δέσμη τροχιών σύμφωνα με έναν θεματικό χάρτη πληθυσμών και σε δεύτερο επίπεδο σύμφωνα με το οδικό δίκτυο της πιο πυκνοκατοικημένης περιοχής.

3.2.4. Η συνεισφορά μας

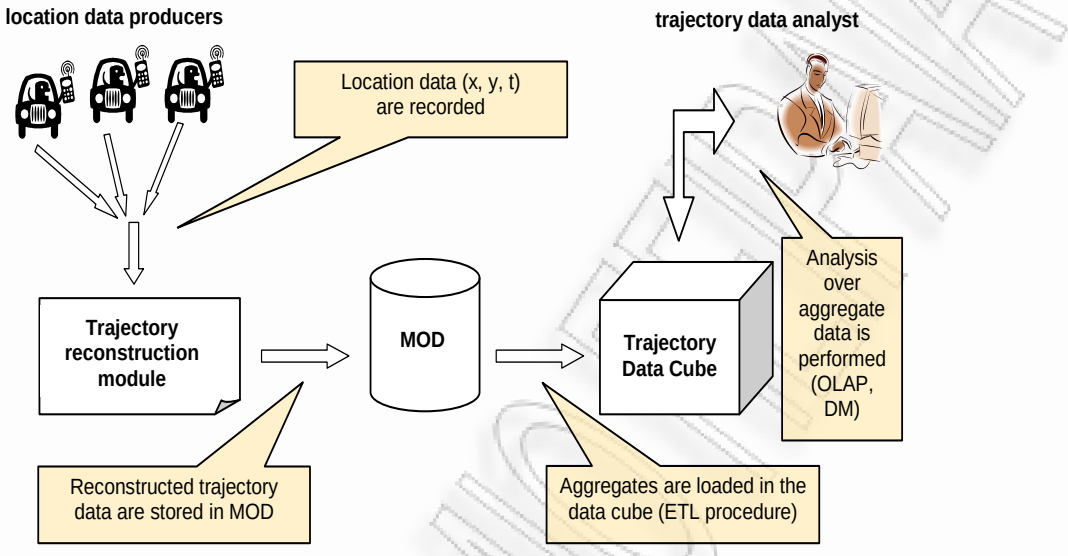
Προκειμένου να χτίσουμε μια TDW θα πρέπει να αντιμετωπιστούν μια σειρά ζητημάτων. Συνοψίζουμε αυτά τα ζητήματα συνοδεύοντας τα με τις συνεισφορές μας όπως παρουσιάζονται στις [MFN+08a], [MFN+08b], [MT09b], [MT09a]:

- Η TDW θα πρέπει να τροφοδοτείται με δεδομένα τροχιών. Για να το πετύχουμε αυτό προτείνουμε δυο εναλλακτικές λύσεις: μια ETL διαδικασία που *προσανατολίζεται στα κελιά* (βασισμένη σε ευρετήριο) και μια που *προσανατολίζεται στις τροχιές* (μη βασισμένη σε ευρετήριο).
- Αντιμετωπίζουμε θέματα συνάθροισης των μέτρων που παρέχονται για σκοπούς OLAP ανάλυσης (π.χ. πως μπορούμε να αξιοποιήσουμε μα μέτρηση στο χαμηλό επίπεδο μιας ιεραρχίας του κύβου ώστε να υπολογιστεί ένα μέτρο σε υψηλότερο επίπεδο). Η ιδιαιτερότητα των δεδομένων τροχιών είναι ότι μια τροχιά μπορεί να διασχίζει πολλαπλά κελιά βάσης (το προαναφερθέν *πρόβλημα μοναδικής προσμέτρησης)*. Αυτό προκαλεί προβλήματα συνάθροισης στις OLAP λειτουργίες. Παρέχουμε μια προσεγγιστική λύση για αυτό το πρόβλημα η οποία φαίνεται να είναι αποτελεσματική.
- Παρουσιάζουμε μια εναλλακτική, καινοτόμο οργάνωση του κύβου δεδομένων ώστε να είναι σε θέση να απαντήσει OLAP ερωτήματα λαμβάνοντας υπόψη διαφορετικούς ορισμούς της έννοιας της τροχιάς. Έτσι είναι δυνατή η ειδική ανάλυση των κύβων δεδομένων τροχιών που μπορεί να είναι χρήσιμη για την εξυπηρέτηση πολλών τροχιών.

3.3. Αποθήκευση Δεδομένων Τροχιών

Στη [MFN+08a], προτείναμε ένα πλαίσιο για TDW που λαμβάνει υπόψη την πλήρη ροή των εργασιών που απαιτούνται κατά τη διάρκεια της ανάπτυξης μιας TDW. Ο πλήρης κύκλος ζωής μιας TDW παρουσιάζεται στην Εικόνα 3-6 και αποτελείται από διάφορα βήματα. Μια διαδικασία ανακατασκευής τροχιάς εφαρμόζεται στα ακατέργαστα χρονοσημασμένα δεδομένα θέσης προκειμένου να παραχθούν

οι τροχιές, οι οποίες αποθηκεύονται έπειτα σε μια MOD. Κατόπιν, μια ETL διαδικασία ενεργοποιείται και τροφοδοτεί τον κύβο δεδομένων με τις συγκεντρωτικές πληροφορίες για τις τροχιές. Το τελικό βήμα της διαδικασίας προσφέρει δυνατότητες OLAP ανάλυσης και εξόρυξης γνώσης πάνω στις συναθροιστικές πληροφορίες που περιλαμβάνονται στο μοντέλο δεδομένων τροχιών.



Εικόνα 3-6: Η αρχιτεκτονική του πλαισίου μας.

Μια MOD συντηρεί θέσεις αντικειμένων σε διάφορες χρονικές στιγμές που καταγράφονται με τη μορφή τροχιών. Τυπικά, $D = \{T_1, T_2, \ldots, T_N\}$ είναι η MOD που αποτελείται από τις τροχιές ενός συνόλου κινούμενων αντικειμένων. Θεωρώντας γραμμική παρεμβολή μεταξύ των συνεχόμενων θέσεων που επιλέχτηκαν από τη δειγματοληψία, η τροχιά $T_i = <(x_{i_1}, y_{i_1}, t_{i_1}), \ldots, (x_{i_{n_i}}, y_{i_{n_i}}, t_{i_{n_i}})>$ αποτελείται από μια σειρά n_{i-1} ευθύγραμμων τμημάτων στον τριδιάστατο χώρο όπου το κάθε τμήμα αντιπροσωπεύει τη συνεχή «ανάπτυξη» του αντίστοιχου κινούμενου αντικειμένου μεταξύ δυο διαδοχικών θέσεων (x_{i_j}, y_{i_j}) και $(x_{i_{j+1}}, y_{i_{j+1}})$ που καταγράφτηκαν στις χρονικές στιγμές t_{i_j} και $t_{i_{j+1}}$. Κάνοντας προβολή το T_i στο χωρικό δισδιάστατο χώρο (χρονικό, μονοδιάστατο χώρο), παίρνουμε την *πορεία* r_i (αντίστοιχα, τη *διάρκεια* l_i) ενός κινούμενου αντικειμένου. Επιπρόσθετες παράμετροι κίνησης μπορούν να εξαχθούν όπως το μήκος *len* της πορείας r_i που διανύθηκε (αντίστοιχα, της *διάρκειας* l_i), η μέση ταχύτητα, επιτάχυνση κτλ.

Ας θεωρήσουμε μια MOD που αποθηκεύει τις ακατέργαστες θέσεις της κίνησης αντικειμένων (π.χ. ανθρώπων). Ένα τυπικό σχήμα που μπορεί να θεωρηθεί ως η ελάχιστη απαίτηση για μια τέτοια MOD παρουσιάζεται στην Εικόνα 3-7.

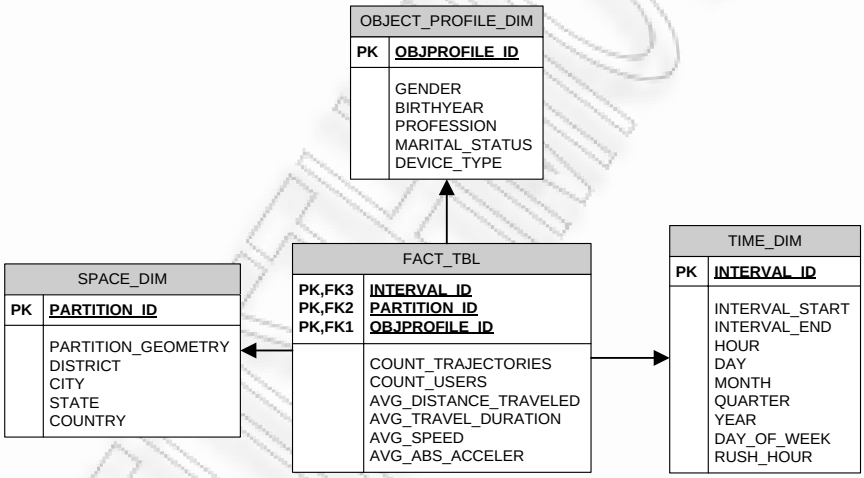
OBJECTS (id: *identifier*, description: *text*, gender: {M | F}, birth-date: *date*, profession: *text*, device-type: *text*)

RAW_LOCATIONS (object-id: *identifier*, timestamp: *datetime*, eastings-x: *numeric*, northings-y: *numeric*, altitude-z: *numeric*)

MOD_TRAJECTORIES (trajectory-id: *identifier*, object-id: *identifier*, trajectory: *3D geometry*)

Εικόνα 3-7: Παράδειγμα MOD.

Ο πίνακας *OBJECTS* περιλαμβάνει ένα μοναδικό προσδιοριστή (id), δημογραφικές πληροφορίες (π.χ. περιγραφή, γένος, ημερομηνία γέννησης, επάγγελμα) καθώς επίσης και τεχνογραφικές πληροφορίες σχετικές με τη συσκευή (π.χ. τύπος GPS). Ο πίνακας RAW_LOCATIONS αποθηκεύει τις θέσεις αντικειμένων στις διάφορες χρονικές στιγμές (δηλ., δείγματα), ενώ ο πίνακας MOD_TRAJECTORIES διατηρεί τις τροχιές των αντικειμένων, μετά από την εφαρμογή της (όποια) τεχνικής ανακατασκευής τροχιάς.



Εικόνα 3-8: Ένα παράδειγμα TDW.

Σίγουρα, ένα TDW πρέπει να περιλαμβάνει μια χωρική και χρονική διάσταση περιγράφοντας τη γεωγραφία και το χρόνο, αντίστοιχα. Θα μπορούσε επίσης να εξεταστεί μια άλλη διάσταση σχετικά με τις συμβατικές πληροφορίες για την κίνηση των αντικειμένων (συμπεριλαμβανομένων των δημογραφικών πληροφοριών, όπως το φύλο, η ηλικία, κ.λπ.).

Λαμβάνοντας υπόψη τα παραπάνω, θεωρούμε, απαραίτητες ως ελάχιστη απαίτηση, τις ακόλουθες διαστάσεις (Εικόνα 3-8):

- Γεωγραφία: η χωρική διάσταση (SPACE_DIM) μάς επιτρέπει να καθορίσουμε τις χωρικές ιεραρχίες. Στη συνέχεια αυτού του κεφαλαίου, υποθέτουμε ένα πλέγμα ισομεγεθών ορθογωνίων (PARTITION_GEOMETRY στην Εικόνα 3-8), το μέγεθος των οποίων είναι μια καθορισμένη από το χρήστη παράμετρος, (π.χ. 10×10 km^2).

- Χρόνος: η χρονική διάσταση (TIME_DIM) καθορίζει τις χρονικές ιεραρχίες. Η χρονική διάσταση έχει μελετηθεί εκτενώς στην βιβλιογραφία [AAD+96]. Στο πιο εκλεπτυσμένο επίπεδο κλιμάκωσης, υποθέτουμε τα καθορισμένα από το χρήστη χρονικά διαστήματα (π.χ. περίοδοι 1 ώρας).
- *Προφίλ Χρήστη*: η θεματική διάσταση (OBJECT_PROFILE_DIM) αναφέρεται σε δημογραφικά και τεχνογραφικά δεδομένα.

Εκτός από τα κλειδιά για τους πίνακες διάστασης, ο πίνακας συμβάντων περιέχει επίσης ένα σύνολο μέτρων με συγκεντρωτικές πληροφορίες. Τα μέτρα που εξετάζονται στο σχήμα TDW στην Εικόνα 3-8 περιλαμβάνουν τον αριθμό μοναδικών τροχιών (COUNT_TRAJECTORIES), τον αριθμό μοναδικών χρηστών (COUNT_USERS), τη μέση διανυμένη απόσταση (AVG_DISTANCE_TRAVELED), τη μέση διάρκεια ταξιδιού (AVG_TRAVEL_DURATION), τη μέση ταχύτητα (AVG_SPEED) και τη μέση επιτάχυνση σε απόλυτες τιμές (AVG_ABS_ACCELER), για μια συγκεκριμένη ομάδα ανθρώπων που κινούνται σε μια συγκεκριμένη χωρική περιοχή κατά τη διάρκεια ενός συγκεκριμένου χρονικού διαστήματος.

3.3.1. Θέματα ETL

Μόλις κατασκευαστούν οι τροχιές και αποθηκευτούν σε μια MOD, εκτελείται η διαδικασία ETL προκειμένου να τροφοδοτηθεί η TDW. Η φόρτωση δεδομένων στους πίνακες διάστασης είναι μια απλή διαδικασία αλλά δε συμβαίνει το ίδιο και στην περίπτωση του πίνακα συμβάντων όπου εκεί πρόκειται για μια σύνθετη διαδικασία. Με βάση την Εικόνα 3-8, ο κύριος στόχος είναι να συμπληρώσουμε τα μέτρα με τις κατάλληλες αριθμητικές τιμές για κάθε ένα από τα κελιά που προσδιορίζονται από τα τρία ξένα κλειδιά (PARTITION_ID, INTERVAL_ID, OBJPROFILE_ID) του πίνακα συμβάντων.

Το μέτρο COUNT_TRAJECTORIES για ένα κελί βάσης *bc* υπολογίζεται καταμετρώντας όλες τις μοναδικές τροχιές που περνούν από το *bc*. Το μέτρο COUNT_USERS για ένα κελί βάσης *bc* υπολογίζεται αντίστοιχα μετρώντας όλα τα μοναδικά αντικείμενα που πέρασαν από το *bc*.

Προκειμένου να υπολογίσουμε το μέτρο AVG_DISTANCE_TRAVELED για ένα κελί βάσης *bc*, ορίζουμε ένα βοηθητικό μέτρο που το ονομάζουμε SUM_DISTANCE και ορίζεται ως το άθροισμα των μηκών *len(TP)* κάθε τμήματος *TP* των τροχιών που βρέθηκε μέσα στο *bc*. Τυπικά:

$$SUM\_DISTANCE(bc) = \sum_{TP_i \in bc} len(TP_i) \tag{3.1}$$

Έτσι το μέτρο AVG_DISTANCE_TRAVELED υπολογίζεται διαιρώντας SUM_DISTANCE με το COUNT_TRAJECTORIES:

$$AVG\_DISTANCE\_TRAVELED(bc) = \frac{SUM\_DISTANCE(bc)}{COUNT\_TRAJECTORIES(bc)} \tag{3.2}$$

Αντίστοιχα για το μέτρο AVG_TRAVEL_DURATION:

$$AVG\_TRAVEL\_DURATION(bc) = \frac{SUM\_DURATION(bc)}{COUNT\_TRAJECTORIES(bc)} \tag{3.3}$$

όπου SUM_DURATION είναι επίσης ένα βοηθητικό μέτρο που ορίζεται στην εξίσωση (3.4) ως το άθροισμα των διαρκειών *lifespan(TP)* κάθε τμήματος *TP* των τροχιών μέσα στο *bc*.

$$SUM\_DURATION(bc) = \sum_{TP_i \in bc} lifespan(TP_i) \tag{3.4}$$

Με τον ίδιο τρόπο, το AVG_SPEED μέτρο υπολογίζεται διαιρώντας το βοηθητικό μέτρο SUM_SPEED (το άθροισμα των ταχυτήτων των τμημάτων *TP* μέσα στο *bc*) με το COUNT_TRAJECTORIES:

$$AVG\_SPEED(bc) = \frac{SUM\_SPEED(bc)}{COUNT\_TRAJECTORIES(bc)} \tag{3.5}$$

όπου $SUM\_SPEED(bc) = \sum_{TP_i \in bc} \frac{len(TP_i)}{lifespan(TP_i)}$ (3.6

Παρόμοια, το AVG_ABS_ACCELER:

$$AVG\_ABS\_ACCELER(bc) = \frac{SUM\_ABS\_ACCELER(bc)}{COUNT\_TRAJECTORIES(bc)} \tag{3.7}$$

όπου SUM_ABS_ACCELER είναι ένα βοηθητικό μέτρο που προσθέτει τις επιταχύνσεις των τμημάτων *TP* που βρίσκονται μέσα στο *bc*

$$SUM\_ABS\_ACCELER(bc) = \sum_{TP_i \in bc} \frac{\left| speed_{fin}(TP_i) - speed_{init}(TP_i) \right|}{lifespan(TP_i)} \tag{3.8}$$

και $speed_{fin}$ ($speed_{init}$) είναι η τελική (η αρχική αντίστοιχα) καταγεγραμμένη ταχύτητα των τμημάτων (TP_i) μέσα στο *bc*.

Είναι σημαντικό να σημειωθεί ότι όλα αυτά τα μέτρα υπολογίζονται με ακρίβεια χρησιμοποιώντας την MOD. Στην πραγματικότητα η MOD που έχουμε υιοθετήσει, η Hermes [PFG+08] παρέχει μια πλούσια παλέτα χωρικών και χρονικών λειτουργιών για το χειρισμό των τροχιών. Δυστυχώς, η εφαρμογή της συσσώρευσης πάνω σε αυτά τα μέτρα δεν είναι μια απλή διαδικασία δεδομένου του προβλήματος μοναδικής προσμέτρησης [TKC+04] και για αυτό το λόγο συζητούμε αναλυτικά αυτά τα θέματα στην επόμενο υποενότητα.

Όπως ήδη αναφέρθηκε, προκειμένου να υπολογιστούν τα μέτρα του κύβου δεδομένων, πρέπει να εξαγάγουμε τα τμήματα των τροχιών που βρίσκονται στα διάφορα κελιά βάσης του κύβου. Θεωρούμε μια MOD *U* χρηστών, *N* τροχιών, *M* χωρικών περιοχών και *K* χρονικών διαστημάτων. Προτείνουμε δύο εναλλακτικές προσεγγίσεις σε αυτό το πρόβλημα: (ι) μια προσανατολισμένη στα κελιά και (ii) μια προσανατολισμένη στις τροχιές.

Σύμφωνα με την προσέγγιση που είναι *προσανατολισμένη στα κελιά* (Cell Oriented Approach - COA), ψάχνουμε για τα τμήματα των τροχιών που βρίσκονται μέσα στα κελιά βάσης. Η αντίστοιχη ETL διαδικασία για την τροφοδοσία του πίνακα συμβάντων της TDW περιγράφεται από τον προτεινόμενο αλγόριθμο CELL-ORIENTED-ETL (Εικόνα 3-9). Αρχικά, ψάχνουμε για τα τμήματα των τροχιών που βρίσκονται μέσα σε ένα χωροχρονικό κελί C (γραμμή 4). Κατόπιν, ο αλγόριθμος συνεχίζει με τη ομαδοποίηση των τμημάτων με βάση τα προφίλ χρηστών (γραμμές 6-9).

```
Algorithm Cell-Oriented-ETL(D MODTrajectoryTable)
1. // For each pair <Region, Interval> forming a s-t cell Cj
2. FOR EACH cell Cj DO
3.    // Find the set of sub-trajectories inside the cell
4.    S = intersects(D, Cj);
5.    // Decompose S to subsets according to object profile
6.    FOR EACH subset S' of S DO
7.     // Compute the various measures
8.     Compute_Measures(S');
9.    END-FOR
10. END-FOR
```

Εικόνα 3-9: Ο αλγόριθμος CELL-ORIENTED-ETL.

Η αποδοτικότητα της *COA* διαδικασίας εξαρτάται από τον αποτελεσματικό υπολογισμό των τμημάτων των τροχιών των κινούμενων αντικειμένων που βρίσκονται μέσα στα χωροχρονικά κελιά (γραμμή 4). Αυτό το βήμα είναι στην πραγματικότητα μια χωροχρονική ερώτηση εύρους που επιστρέφει όχι μόνο τους προσδιοριστές των τροχιών αλλά και τα τμήματα των τροχιών που ικανοποιούν τους περιορισμούς εύρους. Για να υποστηρίξουμε αποτελεσματικά αυτές τις απαιτήσεις επεξεργασίας των τροχιών, χρησιμοποιούμε το TB-tree [PJT00], ένα ευρετήριο για τροχιές που μπορεί να υποστηρίξει αποδοτικά την επεξεργασία ερωτημάτων πάνω σε δεδομένα τροχιών.

Από την άλλη μεριά, η προσέγγιση που είναι *προσανατολισμένη στις τροχιές* (Trajectory Oriented Approach - TOA) περιγράφεται από τον προτεινόμενο αλγόριθμο TRAJECTORY-ORIENTED-ETL (Εικόνα 3-10). Με την προσέγγιση *TOA* ανακαλύπτουμε τα χωροχρονικά κελιά τα οποία τέμνει κάθε τροχιά (γραμμή 6). Προκειμένου να αποφύγουμε να ελέγξουμε όλα τα κελιά, χρησιμοποιούμε (γραμμή 4) μια τραχιά προσέγγιση της τροχιάς, το ελάχιστο περιβάλλον ορθογώνιό της (MBR), και εκμεταλλευόμαστε το γεγονός ότι η κλιμάκωση των κελιών είναι σταθερή ώστε να αναζητήσουμε (πιθανώς) εμπλεκόμενα κελιά. Κατόπιν, προσδιορίζουμε τα τμήματα της τροχιάς που αντιστοιχούν σε κάθε ένα από εκείνα τα κελιά (γραμμές 8-15).

```
Algorithm Trajectory-Oriented-ETL(D MODTrajectoryTable)
1. // For each Trajectory Ti
2. FOR EACH Trajectory Ti of D DO
3.    // Find the Minimum Bounding Rectangle of Ti
4.    MBRTi = Compute_MBR(Ti);
5.    // Find the set of s-t cells C that overlap with the MBR
6.    O = Overlap(C, MBRTi)
7.    // Find the portions (P) of trajectory Ti inside each cell
8.     FOR EACH O' of O DO
9.      P = singlet_intersects(Ti, O');
10.      //If the cell contains portions of the trajectory
11.      IF(P NOT NULL) THEN
12.           // Compute the various measures
13.          Compute_Measures(P);
14.      END-IF
15.     END-FOR
16. END-FOR
```

Εικόνα 3-10: Ο αλγόριθμος TRAJECTORY-ORIENTED-ETL.

3.3.2. OLAP Λειτουργίες: Αντιμετωπίζοντας το πρόβλημα της μοναδικής προσμέτρησης

Κατά τη διάρκεια της ETL διαδικασίας, τα μέτρα μπορούν να υπολογίζονται με ακρίβεια εκτελώντας ερωτήματα στην MOD που βασίζονται στους τύπους που αναφέρθηκαν στην προηγούμενη υποενότητα. Όταν όμως έχει τροφοδοτηθεί ο πίνακας συμβάντων, τα προσδιοριστικά τροχιάς και χρηστών δεν διατηρούνται και μόνο συγκεντρωτικές πληροφορίες αποθηκεύονται μέσα στην TDW.

Οι συναθροιστικές συναρτήσεις (aggregate functions) που υπολογίζουν τις υπερ-συσσωρεύσεις (super-aggregates) των μέτρων ταξινομούνται από τους Gray κ.α. [GCB+97] σε τρεις κατηγορίες βάσει της πολυπλοκότητας που απαιτείται για τον υπολογισμό τους, λαμβάνοντας υπόψη ένα σύνολο ήδη διαθέσιμων υπό-συσσωρεύσεων (sub-aggregates). Στην περίπτωσή μας, οι συναθροιστικές συναρτήσεις που εφαρμόζονται στα κύρια μέτρα που συζητούνται στην Υποενότητα 3.3.1 είναι ταξινομημένες ως ολιστικές και έτσι απαιτούν δεδομένα από την MOD για να υπολογίσουν τις υπερ-συσσωρεύσεις σε όλα τα επίπεδα των διαστάσεων. Αυτό οφείλεται στο γεγονός ότι τα μέτρα COUNT_USERS, COUNT_TRAJECTORIES και, κατά συνέπεια, τα υπόλοιπα μέτρα που ορίζονται βάσει του COUNT_TRAJECTORIES υπόκεινται στο πρόβλημα μοναδικής προσμέτρησης [TKC+04]: εάν ένα αντικείμενο παραμένει στην περιοχή ερώτησης για διάφορες πολλαπλές στιγμές κατά τη διάρκεια του διαστήματος ερώτησης, είναι πιθανό αυτό το αντικείμενο να μετρηθεί πολλαπλές φορές αντί για μία.

Σημειώστε ότι αν οριστεί μια τεχνική συσσώρευσης πάνω στο μέτρο COUNT_TRAJECTORIES, είναι απλό για να καθορίσει μια αντίστοιχη συνάρτηση για τα μέτρα AVG. Στην πραγματικότητα τα τελευταία μπορούν να θεωρηθούν ως το άθροισμα των αντίστοιχων βοηθητικών μέτρων που διαιρούνται με το αποτέλεσμα της συσσώρευσης στο μέτρο COUNT_TRAJECTORIES. Έτσι, ελαχιστοποιώντας τους υπολογισμούς στον αριθμητή, εστιάζουμε στον αριθμό (παρονομαστής) των μοναδικών τροχιών (COUNT_TRAJECTORIES). Το μέτρο COUNT_USERS αντιμετωπίζεται με παρόμοιο τρόπο.

Προκειμένου να υλοποιήσουμε μια λειτουργία συσσώρευσης πάνω σε αυτό το μέτρο, μια πρώτη λύση είναι να ορίζουμε μια διανεμητική συνάρτηση συνάθροισης(distributive aggregate function) που υπολογίζει την υπερ-συσσώρευση ενός κελιού C απλά αθροίζοντας τις τιμές του μέτρου COUNT_TRAJECTORIES στα κελιά που συνθέτου το C. Στη βιβλιογραφία, αυτό αποτελεί μια κοινή προσέγγιση για τη συσσώρευση χωροχρονικών δεδομένων αλλά, όπως θα παρουσιάσουμε στην Υποενότητα 3.3.2, παράγει μια πολύ τραχιά προσέγγιση. Ακλουθώντας την πρόταση της [OOR+07], μια εναλλακτική λύση είναι να καθοριστεί μια αλγεβρική συνάρτηση συνάθροισης (distributive aggregate function). Η ιδέα είναι να αποθηκευτούν στα κελιά βάσης βοηθητικά μέτρα που θα μας επιτρέψουν να διορθώσουμε τα λάθη που προκαλούνται λόγω διπλών καταμετρήσεων κατά τη συσσώρευση.

Πιο τυπικά, ας θεωρήσουμε ότι $C_{(x,y),t,p}$ είναι ένα κελί βάσης, που περιέχει μεταξύ άλλων, τα ακόλουθα μέτρα (είναι σημαντικό να σημειώσουμε ότι αυτά τα μέτρα φορτώνονται χωρίς να υπάρχει κάποιο σφάλμα αφού χρησιμοποιούμε τη λειτουργικότητα της MOD):

- $C_{(x,y),t,p}$.COUNT_TRAJECTORIES: ο αριθμός των μοναδικών τροχιών προφίλ p που τέμνουν το κελί ($C_{(x,y),t,p}.Traj$ για συντομία).

- $C_{(x,y),t,p}$.*cross-x* ο αριθμός των μοναδικών τροχιών προφίλ *p* που περνούν το χωρικό όριο μεταξύ των $C_{(x-1,y),t,p}$ και $C_{(x,y),t,p}$, όπου $C_{(x-1,y),t,p}$ είναι το παρακείμενο κελί (στα αριστερά) στο άξονα του x.
- $C_{(x,y),t,p}$.*cross-y*: ο αριθμός των μοναδικών τροχιών προφίλ *p* που περνούν το χωρικό όριο μεταξύ των $C_{(x,y-1),t,p}$ και $C_{(x,y),t,p}$, όπου $C_{(x,y-1),t,p}$ είναι το παρακείμενο κελί (κάτω) στο άξονα του y.
- $C_{(x,y),t,p}$.*cross-t*: ο αριθμός των μοναδικών τροχιών προφίλ *p* που περνούν το χρονικό όριο μεταξύ των $C_{(x,y),t-1,p}$ και $C_{(x,y),t,p}$, όπου $C_{(x,y),t-1,\mathrm{p}}$ είναι το παρακείμενο κελί (κάτω) στο άξονα του t.

Αν $C_{(x',y'),t',p'}$ είναι ένα κελί που αποτελείται από την ένωση δυο παρακείμενων κελιών με βάση τη χωρική/ χρονική διάσταση, για παράδειγμα $C_{(x',y'),t',p'} = C_{(x,y),t,p} \cup C_{(x+1,y),t,p}$ (όταν εφαρμόζουμε συσσώρευση στο x). Προκειμένου να υπολογίσουμε τις υπερ-συσσωρεύσεις στο $C_{(x',y'),t',p'}$, εφαρμόζουμε το εξής:

$$C_{(x',y'),t',p'}.Traj = C_{(x,y),t,p}.Traj + C_{(x+1,y),t,p}.Traj - C_{(x+1,y),t,p}.cross\text{-}x \qquad (3\text{-}9)$$

Τα υπόλοιπα μέτρα που συσχετίζονται με το $C_{(x',y'),t',p'}$ μπορούν να υπολογιστούν αντιστοίχως:

$C_{(x',y'),t',p'}.cross\text{-}x = C_{(x,y),t,p}.cross\text{-}x$

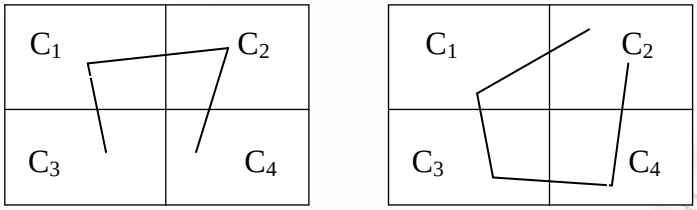
$C_{(x',y'),t',p'}.cross\text{-}y = C_{(x,y),t,p}.cross\text{-}y + C_{(x+1,y),t,p}.cross\text{-}y$

$C_{(x',y'),t',p'}.cross\text{-}t = C_{(x,y),t,p}.cross\text{-}t + C_{(x+1,y),t,p}.cross\text{-}t$

Ο υπολογισμός του $C_{(x',y'),t',p'}.Traj$ μπορεί να θεωρηθεί ως εφαρμογή της γνωστής αρχής Συνυπολογισμού/Αποκλεισμού για ομάδες: $|A \cup B| = |A| + |B| - |A \cap B|$. Ας σημειωθεί ότι σε μερικές περιπτώσεις, το $C_{(x+1,y),t,p}.cross\text{-}x$ δεν είναι όμοιο με το $|A \cap B|$, και αυτό έχει ως αποτέλεσμα την εισαγωγή σφάλματος στις τιμές που επιστρέφονται από την αλγεβρική συνάρτηση. Στη πραγματικότητα όταν όταν ένα κινούμενο αντικείμενο είναι γρήγορο και ευκίνητο, η τροχιά του μπορεί να βρεθεί τόσο στο $C_{(x,y),t,p}$ όσο και στο $C_{(x+1,y),t,p}$ χωρίς να περάσει από το *X* σύνορο (μπορεί να φτάσει στο $C_{(x+1,y),t,p}$ περνώντας το *Y* σύνορο $C_{(x,y),t,p}$ και $C_{(x+1,y),t,p}$).

Αξίζει να αναφερθεί ότι η ευκινησία μιας τροχιάς έχει επιπτώσεις κατά τη διάρκεια της συσσώρευσης Στην πραγματικότητα, μια τροχιά που επιστρέφει σε ένα κελί που έχει ήδη επισκεφτεί μπορεί να εισάγει σφάλμα. Στην ακόλουθη εικόνα επεξηγούμε τα δύο κύρια είδη σφάλματος που η αλγεβρική συνάρτηση συνάθροισης μπορεί να εισαγάγει.

Στην Εικόνα 3-11α, αν ομαδοποιήσουμε τα C_3 και C_4, βλέπουμε ότι ο αριθμός των μοναδικών τροχιών είναι $C_3.Traj + C_4.Traj - C_4.cross\text{-}x = 1+1-0 = 2$. Αυτό αποτελεί μια υπερεκτίμηση του αριθμού των μοναδικών τροχιών. Από την άλλη μεριά, στην Εικόνα 3-11β, αν ομαδοποιήσουμε τα C_1 και C_2 παίρνουμε ως αποτέλεσμα $C_1.Traj + C_2.Traj - C_2.cross\text{-}x = 1+1-1 = 1$, το ίδιο και στην περίπτωση των C_3 και C_4. Όμως, αν ομαδοποιήσουμε το $C_1 \cup C_2$ με το $C_3 \cup C_4$ παίρνουμε $C_1 \cup C_2.Traj + C_3 \cup C_4.Traj - C_1 \cup C_2.cross\text{-}y = 1+1-2 = 0$. Αυτό όμως αποτελεί μια υποεκτίμηση του αριθμού των μοναδικών τροχιών.



Εικόνα 3-11: α) Υπερεκτίμηση του *Traj*. β) Υποεκτίμηση του *Traj*.

Προκειμένου να δώσουμε ένα όριο σφάλματος στον μαθηματικό τύπο (3.10) ας εστιάσουμε σε μια τροχιά. Δεν πρόκειται για περιορισμό αφού τα μέτρα *Traj*, *cross-x*, *cross-y*, και *cross-t* μπορούν να υπολογιστούν αθροίζοντας τις τιμές που προκύπτουν από κάθε τροχιά ξεχωριστά. Από τη στιγμή που οι λειτουργίες συσσώρευσης είναι γραμμικές συναρτήσεις αυτή η ιδιότητα ισχύει και για τα συσσωρευμένα κελιά.

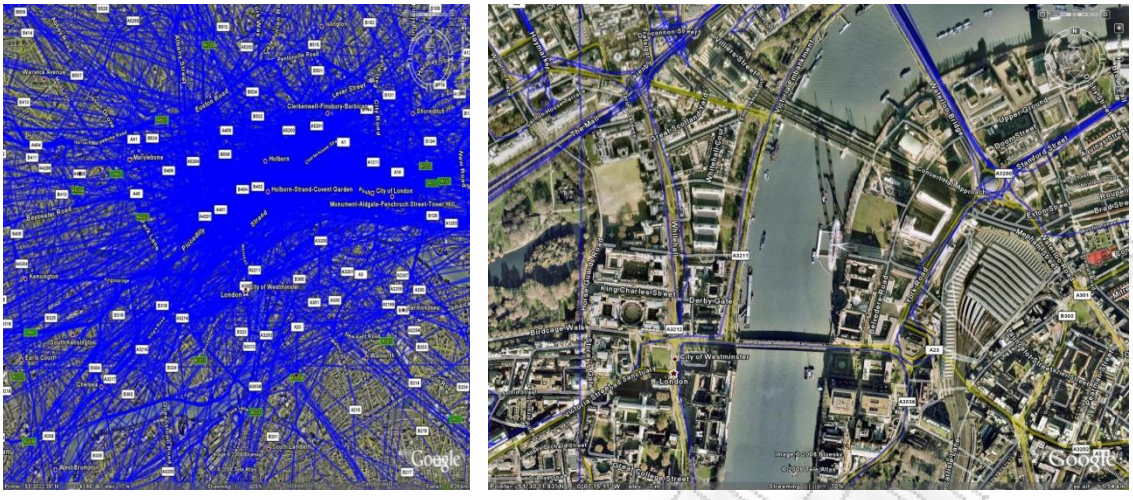
Πρώτα από όλα, ας εισάγουμε την έννοια της *ακολουθίας μονο-ογδοημορίων* (uni-octant sequence). Πρόκειται για μια ακολουθία συνδεδεμένων τμημάτων μια τροχιάς των οποίων οι κλίσεις εντοπίζονται στο ίδιο ογδοημόριο. Είναι εμφανές ότι μια τροχιά μπορεί με μοναδικό τρόπο να αποσυνθεθεί σε ακολουθίες μονο-ογδοημορίων.

Μια ακολουθία μονο-ογδοημορίων *us* μπορεί να διασχίζει ένα κελί *C* μόνο μια φορά, δηλ., αν το *us* ξεκινά από το *C* μπορεί να εξέρχεται μόνο από το *C*. Συνεπώς, όταν μια τροχιά αποτελείται από μια μοναδική ακολουθία μονο-ογδοημορίων δεν εισάγεται σφάλμα στην περίπτωση της συσσώρευσης για το μέτρο COUNT_TRAJECTORIES. Σφάλμα παρατηρείται μόνο όταν μια τροχιά επισκέπτεται ένα κελί τουλάχιστον δυο φορές.

Αυτό μπορεί να γενικευτεί στην περίπτωση μιας τροχιάς *T* που αποτελείται από ακολουθίες μονο-ογδοημορίων. Σε αυτήν την περίπτωση, η τιμή που υπολογίζεται για το μέτρο *Traj* σε ένα συσσωρευμένο κελί *C* περιορίζεται από τον αριθμό των ακολουθιών μονο-ογδοημορίων της *T* που τέμνουν το *C*. Αυτό είναι το άνω όριο που μπορεί που μπορεί να επιτευχθεί όπως φαίνεται στην Εικόνα 3-11a.

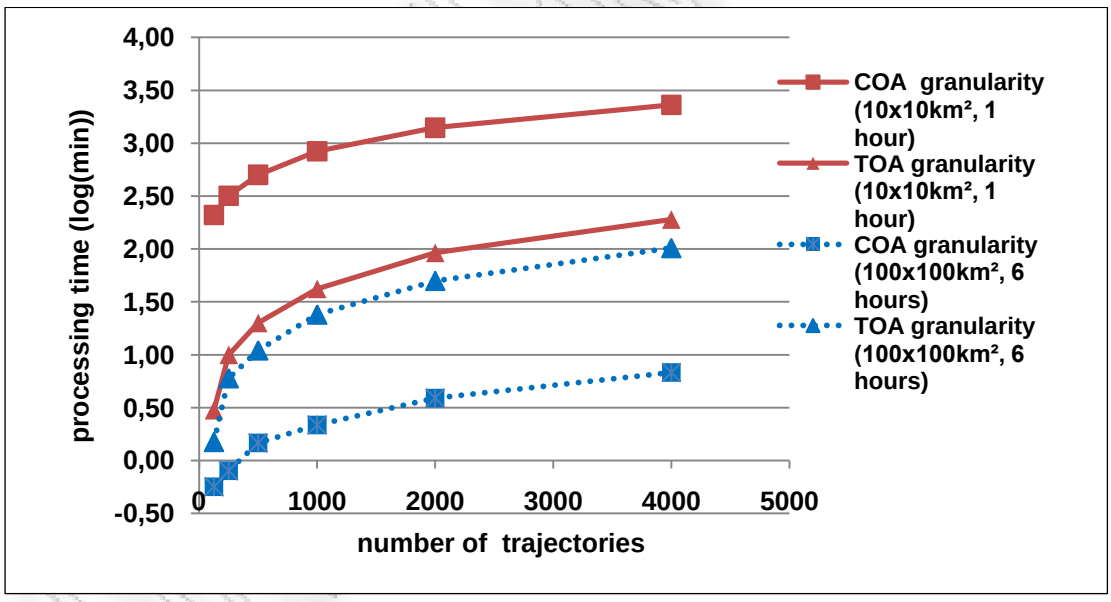
3.3.3. Πειραματική Μελέτη

Σε αυτή την υποενότητα, αξιολογούμε τις προτεινόμενους λύσεις για την υλοποίηση της αρχιτεκτονικής μιας TDW (Εικόνα 3-8) ως ένα πραγματικό σενάριο. Πιο συγκεκριμένα, χρησιμοποιήσαμε ένα μεγάλο, πραγματικό σύνολο δεδομένων: μέρος του συνόλου δεδομένων the e-Courier [Eco09] που αποτελείται από 6,67 εκατομμύρια ακατέργαστων εγγραφών θέσης (ένα αρχείο μεγέθους 504 Mb), που αναπαριστά την κίνηση 84 υπαλλήλων μια εταιρίας ταχυμεταφορών που κινούνται στην ευρύτερη περιοχή του Λονδίνου (συνολικά καλύπτεται μια έκταση 66,800 km^2) για μια περίοδο ενός μήνα (Ιούλιος 2007) και με ρυθμό δειγματοληψίας τα 10 δευτερόλεπτα. Η Εικόνα 3-12 παρουσιάζει μερικά στιγμιότυπα από το σύνολο δεδομένων. Για όλα τα πειράματα χρησιμοποιήσαμε ένα PC με 1 Gb RAM και P4 3 GHz CPU.



Εικόνα 3-12: a) Το σύνολο δεδομένων b) Εστιάζοντας στον Τάμεση

Για την αξιολόγηση της ETL διαδικασίας συγκρίναμε την απόδοση της TOA προσέγγισης με αυτήν της COA (βασισμένης σε ευρετήριο). Χρησιμοποιήσαμε δύο διαφορετικά επίπεδα κλιμάκωσης για τη χωρική και τη χρονική ιεραρχία: ένα χωρικό πλέγμα ισομεγεθών τετραγώνων 10×10 Km^2 (100×100 Km^2, αντίστοιχα) και ένα χρονικό διάστημα μιας ώρας (έξι ωρών, αντίστοιχα). Τα αποτελέσματα των τεσσάρων περιπτώσεων φαίνονται στην Εικόνα 3-13 όπου είναι σαφές ότι η επιλογή μιας συγκεκριμένης μεθόδου εξαρτάται από το επιλεγμένο επίπεδο κλιμάκωσης και τον αριθμό των τροχιών.

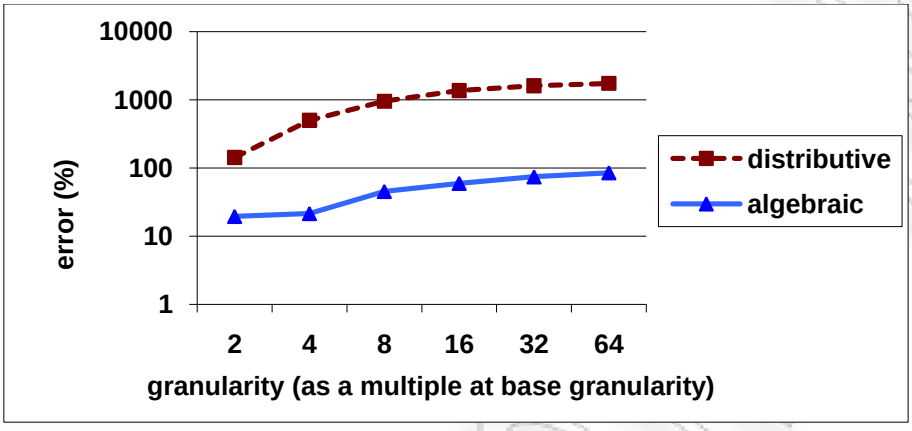


Εικόνα 3-13: Σύγκριση των εναλλακτικών ETL διαδικασιών.

Ολοκληρώνουμε την πειραματική μας μελέτη με αποτελέσματα στα θέματα συσσώρευσης (Εικόνα 3-14). Θέλουμε να αξιολογήσουμε την ακρίβεια των προσεγγίσεων του μέτρου COUNT_TRAJECTORIES κατά τη διάρκεια μια λειτουργίας συσσώρευσης χρησιμοποιώντας τόσο την διανεμητική όσο και την αλγεβρική συνάρτηση που παρουσιάστηκαν στην Υποενότητα 3.3.2. για το σκοπό αυτό χρησιμοποιούμε τον κανονικοποιημένο απόλυτο σφάλμα που προτάθηκε από τους Vitter κ.α. [VWI98]: για όλα τα OLAP ερωτήματα *q* ενός συνόλου *Q* ορίζουμε το σφάλμα ως:

$$Error = \frac{\sum_{q \in Q} |\overline{Mq} - Mq|}{\sum_{q \in Q} Mq} \quad (3.10)$$

όπου $\overline{Mq}$ είναι η προσεγγιστική τιμή του μέτρου που υπολογίζεται κατά το q, ενώ Mq είναι η ακριβής τιμή.



Εικόνα 3-14: Σύγκριση διανεμητικής και αλγεβρικής συνάρτησης συσσώρευσης (η βασική ανάλυση τίθεται στα 10×10 Km^2 στο χώρο και 1 ώρα περίοδος στο χρόνο).

Υποθέτουμε μια ανάλυση στα κελιά βάσης της τάξης των 10×10 Km^2 για το χωρικό πλέγμα και για 1 ώρα ως χρονική περίοδο. Επίσης παραθέτουμε αντίστοιχους υπολογισμούς για κλιμακώσεις $g' = n \times g$, με $n > 1$.

Η διανεμητική συνάρτηση συσσώρευσης παρουσιάζει σφάλμα που υπερβαίνει πάντα το 100% και αυξάνεται καθώς η κλιμάκωση αυξάνεται κατά την συσσώρευση. Αντίθετα, όπως αναμενόταν, οι υπολογισμοί που βασίζονται στην αλγεβρική συνάρτηση είναι πάντα ακριβέστεροι από εκείνους που βασίζονται στη διανεμητική και είναι ακριβείς για τις μικρές κλιμακώσεις. Το σφάλμα μεγαλώνει για τις μεγάλες κλιμακώσεις αλλά δεν υπερβαίνει ποτέ το 100%. Αν και τα αντίστοιχα πειράματα δεν αναφέρονται εδώ, αξίζει να αναφερθεί ότι ξεκινώντας από μικρότερες αναλύσεις και χρησιμοποιώντας την αλγεβρική συνάρτηση, πετυχαίνουμε καλύτερη ακρίβεια (σφάλμα κάτω από 10% για τα μικρά πολλαπλάσια του g).

3.4. Κατά περίπτωση (ad-hoc) OLAP σε Δεδομένα Τροχιών

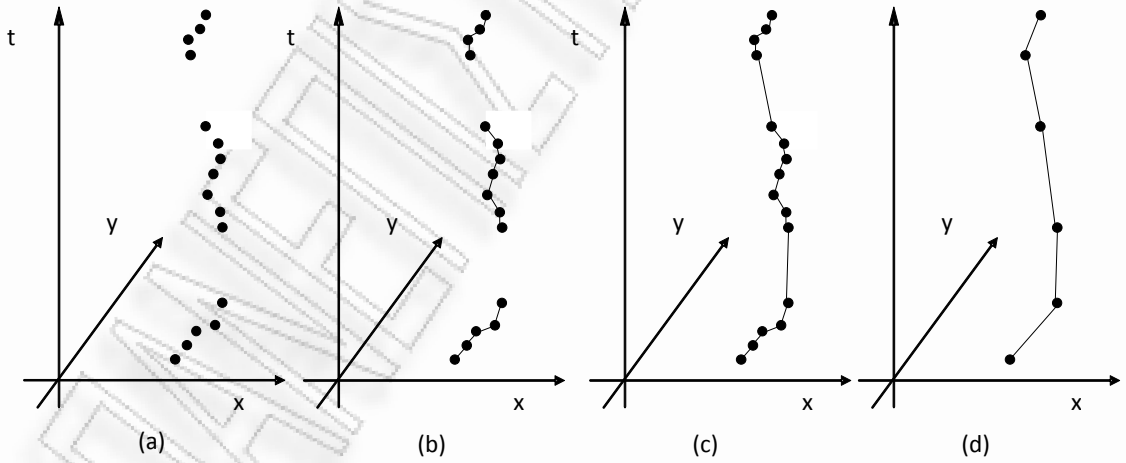
Σε αυτή την ενότητα, περιγράφουμε μια καινοτόμο οργάνωση ενός κύβου δεδομένων τροχιάς που μπορεί να απαντήσει σε OLAP ερωτήματα λαμβάνοντας υπόψη τις διαφορετικές ερμηνείες της έννοιας της τροχιάς. Η προσέγγιση αυτό παρουσιάστηκε στη [MT09a]. Κατά συνέπεια, μπορεί να επιτευχθεί κατά περίπτωση (ad-hoc) ανάλυση στους κύβους δεδομένων τροχιάς κάτι που είναι χρήσιμο για χρήση από διάφορες εφαρμογές. Τα προκαταρκτικά πειραματικά αποτελέσματα αποδεικνύουν τη δυνατότητα εφαρμογής και την αποδοτικότητα της προσέγγισής μας.

Η ανάλυση τροχιών βασίζεται στις συγκεκριμένες απαιτήσεις κάθε εφαρμογής. Για παράδειγμα, ένας αναλυτής κυκλοφορίας και, αφ' ενός, ένας διευθυντής εφοδιαστικής αλυσίδας (logistics) μπορεί να ορίζουν με ένα τελείως διαφορετικό σημασιολογικό τρόπο την τροχιά. Ας θεωρήσουμε ένα στόλο φορτηγών που κινούνται σε μια πόλη και που παραδίδουν αγαθά σε διάφορες θέσεις. Για κάθε

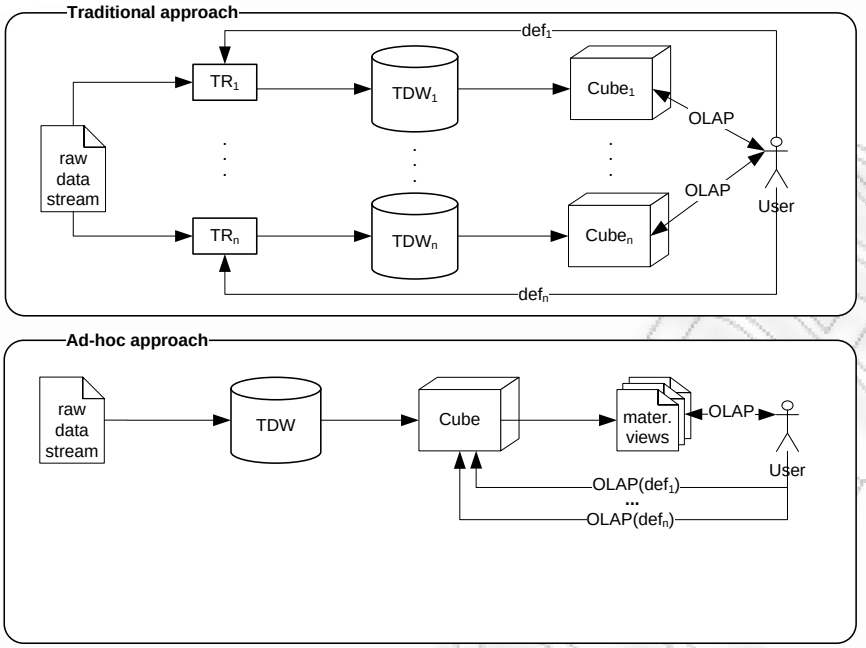
φορτηγό, ο διευθυντής logistics μπορεί να ενδιαφέρεται να εξετάσει ένα σύνολο τροχιών (π.χ. μεταξύ των διαφορετικών σημείων παράδοσης) ενώ ο αναλυτής κυκλοφορίας μπορεί απλά να ενδιαφέρεται μια ενιαία τροχιά για ολόκληρη την ημέρα. Κατά συνέπεια, προκειμένου να ικανοποιηθούν αυτές οι δύο, αρκετά διαφορετικές στη σημασιολογία, ανάγκες θα έπρεπε να ανακτήσουμε τα ακατέργαστα δεδομένα θέσης από μια κοινή αποθήκη και, κατόπιν, να εκτελέσουμε δύο διαφορετικές διαδικασίες ανακατασκευής ώστε να παραχθούν οι τροχιές που είναι σημασιολογικά συμβατές σε κάθε περίπτωση. Έπειτα, θα έπρεπε να χτίσουμε δύο διαφορετικούς κύβους δεδομένων τροχιάς προκειμένου να μπορέσουν οι χρήστες να εφαρμόσουν τις τεχνικές OLAP για να καλύψουν τις ανάγκες τους.

Για παράδειγμα, η Εικόνα 3-15a απεικονίζει ένα ακατέργαστο σύνολο χρονοσημασμένων δεδομένων θέσης. Διαφορετικές ανάγκες ανάλυσης είναι δυνατό να οδηγήσουν σε διαφορετικά σύνολα ανακατασκευασμένων τροχιών (Εικόνα 3-15b-d, αντίστοιχα). Θυμίζοντας ξανά το προηγούμενο παράδειγμα με τα φορτηγά, μπορούμε να θεωρήσουμε ότι η Εικόνα 3-15b και η Εικόνα 3-15c απεικονίζουν τις ανακατασκευασμένες τροχιές για τον διευθυντή logistics και για το διευθυντή κυκλοφορίας, αντίστοιχα. Ένα άλλο παράδειγμα της ανακατασκευασμένης τροχιάς παρουσιάζεται στην Εικόνα 3-15d που παρουσιάζει μια συμπιεσμένη τροχιά της κίνησης.

Αν ακολουθήσουμε τη συμβατική προσέγγιση (π.χ. [MFN+08a]) για το χτίσιμο μιας TDW τότε για κάθε σύνολο ανακατασκευασμένων τροχιών θα έπρεπε να εκτελέσουμε μια ETL διαδικασία ώστε να χτίσουμε διαφορετικούς κύβους δεδομένων τροχιών. Αυτό παρουσιάζεται με σαφήνεια στην Εικόνα 3-16 (traditional approach) όπου οι διαφορετικοί σημασιολογικοί ορισμοί $(def_1, \ldots, def_n)$ απαιτούν διαφορετικές εκτελέσεις της διαδικασίας ανακατασκευής $(TR_1, \ldots, TR_n)$ που καταλήγουν σε διαφορετικές TDW και κύβου δεδομένων τροχιών αντίστοιχα.



Εικόνα 3-15: Διαφορετικές προσεγγίσεις ανακατασκευής τροχιών (b, c, d) για ένα πρωτογενές σύνολο δεδομένων (a).



Εικόνα 3-16: Η παραδοσιακή (traditional) και η κατά περίπτωση (ad-hoc) προσέγγιση.

Στην πραγματικότητα, αυτό είναι ασυνήθιστο σε άλλα (συμβατικά ή όχι) σενάρια αποθήκευσης δεδομένων. Για παράδειγμα, η ίδια αποθήκη δεδομένων πωλήσεων μπορεί να χρησιμοποιηθεί και από τους πωλητές και από τα στελέχη του μάρκετινγκ επειδή και οι δύο συμφωνούν ότι μια πώληση έχει μερικά συγκεκριμένα και συνεπώς ευρέως αποδεκτά χαρακτηριστικά. Στις επιστημονικές βάσεις δεδομένων, η ίδια ΑΔ σε ένα SDMMS [MTK08] μπορεί να χρησιμοποιηθεί και από έναν επιστήμονα και από έναν υπάλληλο της δημόσιας διοίκησης προκειμένου να ερευνηθεί η σεισμική δραστηριότητα με βάση έναν αποδεκτό ορισμό της έννοιας του σεισμού. Ακόμη και στις πιο πρόσφατες εφαρμογές των τεχνικών αποθήκευσης δεδομένων σε δεδομένα ροών (workflow data) [Kot06], ο κύβος δεδομένων χτίζεται πάνω σε καλά ορισμένα συμβάντα (service provisions).

Από την άλλη μεριά, η χωρική-χρονική φύση των δεδομένων τροχιάς οδηγεί σε διαφορετικούς σημασιολογικούς ορισμούς των τροχιών. Διαισθητικά, θα πρέπει να ξανασκεφτούμε τις βασικές δομές (πίνακας συμβάντων, διαστάσεις, ETL, προϋπολογισμός κύβων) ενός DW, και να χτίσουμε πιο ευέλικτους μηχανισμούς ώστε να αντιμετωπίσουμε αυτήν την πρόκληση. Επιπλέον, είναι σημαντικό να επεκταθούν οι παραδοσιακές τεχνικές OLAP προκειμένου να είναι δυνατό να χειριστούν κατάλληλα τη δυναμική φύση των τροχιών. Ο στόχος μας είναι απεικονίζεται στην Εικόνα 3-16 (ad-hoc approach) όπου οι διαφορετικοί σημασιολογικοί ορισμοί των τροχιών trajectories (def_1, …, def_n) εφαρμόζονται στον ίδιο κύβο. Σε περίπτωση γνωστών σημασιολογικών ορισμών, ο κύβος μπορεί να προϋπολογιστεί και να παρέχει τις ίδιες ακριβώς λειτουργίες όπως ακριβώς και στην παραδοσιακή προσέγγιση.

Από όσο γνωρίζουμε, αυτή είναι η πρώτη προσέγγιση κατά περίπτωση ανάλυσης σε μια TDW. Η συνεισφορά μας σε αυτό το σημείο μπορεί να συνοψιστεί στα ακόλουθα:

- Επεκτείνουμε το OLAP μοντέλο δεδομένων για TDW με δύο τρόπους. Κατ' αρχάς, προτείνουμε έναν ευέλικτο πίνακα συμβάντων που θα είναι σε θέση να απαντήσει στα

ερωτήματα λαμβάνοντας υπόψη τους διαφορετικούς σημασιολογικούς ορισμούς των τροχιών. Δεύτερον, εισάγουμε μια παράμετρο ώστε να υποστηρίζεται μια συγκεκριμένη σημασιολογία για τις ερωτήσεις συνάθροισης σε δεδομένα τροχιών.

- Παρουσιάζουμε μια κατάλληλη ETL διαδικασία για τη φόρτωση ακατέργαστων δεδομένων θέσης στον ευέλικτο κύβο δεδομένων που προτείνουμε.
- Εμπλουτίζουμε τις OLAP τεχνικές ώστε να εκμεταλλεύονται τις νέες ad-hoc προσεγγίσεις. Προτείνεται ένας αποδοτικός αλγόριθμος που εκμεταλλεύεται το μοντέλο μας για να μπορέσει να απαντήσει ερωτήματα συσσώρευσης.
- Για να επιταχύνουμε τη διαδικασία υπολογισμού, συζητούμε θέματα προϋπολογισμού για γνωστούς σημασιολογικούς ορισμούς τροχιών.

3.4.1. Ορισμός Προβλήματος

Ακολουθώντας τη σχεσιακή αναπαράσταση [Kim96] του πολυδιάστατου μοντέλου [AAD+96] για αποθήκευση δεδομένων, ένα TDW, όπως έχει ήδη αναφερθεί, αποτελείται από έναν πίνακα συμβάντων που περιέχει τα κλειδιά για τους πίνακες διαστάσεων και διάφορα μέτρα. Οι πίνακες διαστάσεων μπορεί να περιέχουν πολλά χαρακτηριστικά πάνω στα οποία είναι δυνατό να οριστούν πολλαπλές ιεραρχίες ώστε να υποστηρίζεται η OLAP ανάλυση. Για κάθε διάσταση, καθορίζεται το πιο εκλεπτυσμένο επίπεδο κλιμάκωσης, και αφορά το επίπεδο λεπτομέρειας των στοιχείων που αποθηκεύονται στον πίνακα γεγονότων. Η πολυδιάστατη μορφή μιας αποθήκης δεδομένων μπορεί να θεωρηθεί ως μια πολυδιάστατη μήτρα αριθμητικών στοιχείων πάνω στην οποία εφαρμόζονται συναρτήσεις συσσώρευσης. Τυπικότερα:

Ορισμός 3-1 (Πίνακας διάστασης): Ένας *πίνακας διάστασης* είναι μια m-αδική σχέση $F \times A_1 \times A_2 \times \ldots \times A_v$, όπου:

i. F είναι το πρωτεύον κλειδί του πίνακα διάστασης,
ii. Κάθε στήλη A_j, $0 \leq j \leq v$ είναι ένα σύνολο τιμών των χαρακτηριστικών,
iii. $m = 1+v$. ■

Ορισμός 3-2 (Πίνακας συμβάντων): Ένας *πίνακας συμβάντων* είναι μια n-αδική σχέση $K \times M_1 \times M_2 \times \ldots \times M_r$, όπου:

i. K είναι το σύνολο των χαρακτηριστικών που αντιπροσωπεύει το πρωτεύον κλειδί του πίνακα συμβάντων που ορίζεται από $F_1 \times F_2 \times \ldots \times F_p$, όπου κάθε F_i, $1 \leq i \leq p$ είναι ένα ξένο κλειδί στους πίνακες συμβάντων,
ii. Κάθε στήλη M_k, $1 \leq k \leq r$ είναι ένα σύνολο μέτρων που μπορεί να υπολογιστεί εφαρμόζοντας συναρτήσεις συσσώρευσης στα χαρακτηριστικά των πρωτογενών τροχιών,
iii. $n = p+r$. ■

Η TDW που παρουσιάσαμε στην Ενότητα 3.3, ακολουθεί τους παραπάνω ορισμούς. Υλοποιώντας όμως μια TDW με βάση την προσέγγιση που δείξαμε εκεί, δε λαμβάνουμε υπόψη τους διαφορετικούς σημασιολογικούς ορισμούς των τροχιών. Υποθέτουμε ότι οι τροχιές έχουν ανακατασκευαστεί σε προηγούμενο στάδιο ακολουθώντας ένα συγκεκριμένο ορισμό τροχιάς και συνεπώς όλες οι μετρήσεις

έχουν βασιστεί σε αυτό τον ορισμό. Υιοθετούμε ως παράδειγμα τα μέτρα της TDW που αναλύσαμε στην Ενότητα 3.3, ένα υποσύνολο των οποίων συζητείται και στην [OOR+07] και συζητούμε με ποιον τρόπο μπορεί να μετασχηματιστεί η TDW ώστε να λαμβάνει υπόψη πολλαπλούς σημασιολογικούς ορισμούς (της τροχιάς).

Πριν προχωρήσουμε στο κυρίως μέρος αυτής νέας προσέγγισης, περιγράφουμε παρακάτω την έννοια των διαφορετικών σημασιολογικών ορισμών των τροχιών. Αν υποθέσουμε ότι μια ακολουθία χρονοσημασμένων θέσεων έχει καταγραφεί για ένα κινούμενο αντικείμενο (π.χ. χρησιμοποιώντας μια συσκευή GPS), η κίνηση αυτού του αντικειμένου μπορεί να οριστεί ως εξής:

$$M = ((x_1, y_1, t_1),...,(x_n, y_n, t_n)) \tag{3.11}$$

Μετά την ανακατασκευή τροχιάς αυτή η κίνηση μπορεί να καταλήξει ως ένα σύνολο τροχιών:

$$M' = (T_1,...,T_j) \tag{3.12}$$

όπου κάθε τροχιά T_i μπορεί να οριστεί ως [GBE+00]:

$$T_i = < (x_1, y_1, t_1),...,(x_k, y_k, t_k) > \tag{3.13}$$

όπου $1 \leq k$, $m \leq n$, δηλαδή τα σημεία της τροχιάς είναι ταυτόχρονα και σημεία που ανήκουν στο αντίστοιχο ακατέργαστο σύνολο δεδομένων, με άλλα λόγια, $\bigcup_i T_i \subset M$.

Δυο βασικές συναρτήσεις που συνήθως παρέχονται από τις MOD και μπορούν να εφαρμοστούν σε τροχιές είναι οι *len*(T_i) και *lifespan*(T_i) που επιστρέφουν το μήκος της διδιάστατης χωρικής προβολής της T_i και τη διάρκεια της στο χρόνο, αντίστοιχα.

Με βάση το προαναφερθέν μοντέλο, ένα σύνολο *M* των ακατέργαστων χρονο-σημασμένων σημείων θέσης μπορεί να οδηγήσει σε ένα σύνολο *M'* των τροχιών. Ο ακριβής αριθμός των ακατασκευασμένων τροχιών εξαρτάται από τους διαφορετικούς σημασιολογικούς ορισμούς που μπορούν να δοθούν σε μια τροχιά. Για παράδειγμα, ας θεωρήσουμε κάποιον που οδηγεί το πρωί από το σπίτι του στο γραφείο, εργάζεται για οκτώ ώρες, και επιστρέφει έπειτα το σπίτι μετά από μια σύντομη στάση για αγορές. Διαφορετικές εφαρμογές μπορούν να θεωρήσουν έναν διαφορετικό αριθμό τροχιών σε αυτήν την περίπτωση. Οι Spaccapietra κ.α. [SPD+08] υποστηρίζουν ότι οι διαφορετικές χρονικές κλιμακώσεις οδηγούν σε διαφορετικούς σημασιολογικούς ορισμούς των τροχιών. Έτσι, στο προηγούμενο παράδειγμα, δεν υπάρχει μια ενιαία απάντηση στην ερώτηση «πόσες τροχιές υπάρχουν;» δεδομένου ότι αυτή εξαρτάται από το επίπεδο χρονικής ανάλυσης για το οποίο ενδιαφερόμαστε. Κάποιος μπορεί να αναμένει ως απάντηση τη μια, δύο ή/και τρεις τροχιές (π.χ., που θέτουν τη χρονική ανάλυση στο επίπεδο ημέρας, ώρας, λεπτού, αντίστοιχα). Οι συγγραφείς στη [SPD+08] παραθέτουν μια συζήτηση σχετικά με τη σημασιολογία της τροχιάς και αναγνωρίζουν την ανάγκη εμπλουτισμού του χωροχρονικού μοντέλου δεδομένων ώστε να είναι ευέλικτο. Η ιδέα πίσω από αυτό είναι να μπορεί να προσαρμόζεται το μοντέλο σε μια συγκεκριμένη σημασιολογία όπως αυτή απαιτείται από μια συγκεκριμένη εφαρμογή. Η προσέγγιση που συζητείται σε αυτήν την υποενότητα περιλαμβάνει τον μετασχηματισμό των τροχιών σε ακολουθίες κινήσεων από μια στάση στην επόμενη (ή ως ακολουθία στάσεων που διαχωρίζουν τις κινήσεις). Αυτή η προσέγγιση βασίζεται στο γεγονός ότι οι τροχιές

μπορούν να τεμαχιστούν σημασιολογικά καθορίζοντας μιας χρονική ακολουθία υπο-διαστημάτων (κινήσεις) όπου η θέση ενός αντικειμένου είτε αλλάζει είτε παραμένει σταθερή (στάσεις). Με βάση αυτήν την προσέγγιση, οι Baglioni κ.α. [BMR+08] προτείνουν τον εμπλουτισμό των ακατέργαστων τροχιών με σημασιολογικές πληροφορίες και την εκμετάλλευση της γνώσης που μπορεί να υπάρχει για μια περιοχή (domain) και μπορεί να κωδικοποιείται σε μια οντολογία.

Το πρόβλημα που συζητείται σε αυτήν την ενότητα είναι αυτό της οικοδόμησης μια ευέλικτης TDW που προσφέρει δυνατότητες OLAP ανάλυσης στα στοιχεία κίνησης επιτρέποντας ένα συγκεκριμένο ορισμό της έννοιας της τροχιάς μόνο κατά τη διάρκεια εκτέλεσης ενός ερωτήματος συνάθροισης. Με άλλα λόγια, μελετούμε το πρόβλημα εύρεσης των *M'* τροχιών κατά τη διάρκεια της εκτέλεσης μιας ερώτησης συνάθροισης και τον υπολογισμό των αριθμητικών μέτρων για αυτό το σύνολο τροχιών. Αυτός ο στόχος είναι πολλαπλός δεδομένου ότι περιλαμβάνει διάφορα ζητήματα:

- *Μοντελοποίηση TDW*: ένα νέο ευέλικτο μοντέλο που εμπεριέχει την απαραίτητη ευελιξία σε ότι αφορά τους διαφορετικούς ορισμούς των τροχιών,
- *ETL επεξεργασία*: ένας κατάλληλος μηχανισμός που τροφοδοτεί την TDW και συμμορφώνεται με το ανώτερο μοντέλο.,
- *OLAP τεχνολογία*: θα πρέπει να εμπλουτιστεί ώστε να μπορεί να αξιοποιήσει το νέο μοντέλο TDW επιτρέποντας έτσι στους χρήστες να ορίσουν το επίπεδο της χρονικής κλιμάκωσης για το οποίο ενδιαφέρονται,
- *Προϋπολογισμός κύβου:* η απόδοση θα πρέπει να ληφθεί υπόψη ώστε να αυξηθεί ο χρόνος απόκρισης ερωτημάτων σε μια TDW.

Όλα αυτά τα θέματα αντιμετωπίζονται στις παραγράφους που ακολουθούν όπου και περιγράφεται με σαφήνεια η προτεινόμενη αρχιτεκτονική του κατά περίπτωση OLAP.

3.4.2. Ένα Πλαίσιο για κατά περίπτωση (ad-hoc) OLAP

Η πρόκληση της ανάπτυξης μιας TDW κατάλληλης για κατά περίπτωση ανάλυση εισάγει τρεις στόχους σχετικά με τον πίνακα συμβάντων: a) την υιοθέτηση μιας ευέλικτης δομής που θα επιτρέψει την κατά περίπτωσης ανάλυση β) τον υπολογισμό των μέτρων κατά τη διάρκεια της φάσης φόρτωσης του κύβου (ETL) και γ) την αντιμετώπιση των ζητημάτων συνάθροισης μετά τη φόρτωση του πίνακα συμβάντων. Στις ακόλουθες υποενότητες, παρουσιάζουμε τις λύσεις μας σχετικά με τα ανωτέρω ζητήματα, και παρουσιάζουμε επίσης μια προσέγγιση μοντελοποίησης για την αναπαράσταση διαφορετικών μορφών τροχιών σε μια TDW.

3.4.2.1. Το μοντέλο

Όπως αναφέραμε στη προηγούμενη υποενότητα, ένα πίνακας συμβάντων αποτελείται από διάφορα (προσανατολισμένα στην τροχιά) μέτρα θεωρώντας ότι ο συγκεκριμένος ορισμός της τροχιάς είναι εκ των προτέρων γνωστός. Σε αυτήν την υποενότητα, προτείνουμε μια πιο ευέλικτη δομή που θα επιτρέψει στο χρήστη να αποφασίσει για τα χαρακτηριστικά των τροχιών. Συγκεκριμένα, εστιάζουμε στα μέτρα των οποίων ο υπολογισμός τους απαιτεί τις χωρικές ή/και χρονικές αποστάσεις μεταξύ των σημείων κάθε τροχιάς, και προτείνουμε το μετασχηματισμό της μετακίνησης ως *ακολουθία*

χωροχρονικών αποστάσεων μεταξύ των διαδοχικών σημείων. Τυπικά, η μετακίνηση ενός κινούμενου αντικειμένου μπορεί να διατυπωθεί:

$$M_i^{"} = < (0,0),(sd_2, td_2),...,(sd_n, td_n) > \quad (3.14)$$

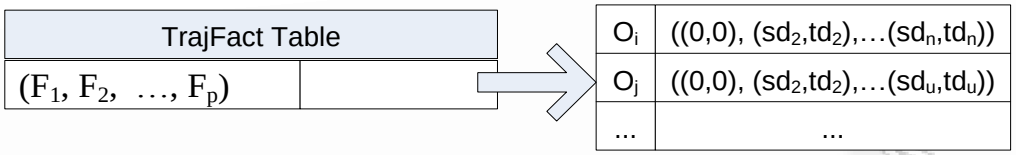
όπου (sd_i, td_i) αντιπροσωπεύει την (Ευκλείδεια) χωρική (*sd*) και χρονική απόσταση (*td*) μεταξύ του i^{ου} και του (i-1)^{ου} σημείου; το πρώτο ζευγάρι ορίζεται πάντα ως (0, 0) αφού δεν υπάρχει προηγούμενο στοιχείο για να συγκριθεί μαζί του. Προφανώς το $M_i^{"}$ αποτελεί μια αναπαράσταση στην οποία υπάρχουν απώλειες σε σχέση με το *M* ή το *M'*. Όμως, όπως συμβαίνει και στο χώρο της διαχείρισης παραδοσιακών δεδομένων, οι ΑΔ αποθηκεύουν μόνο την απαραίτητη πληροφορία ώστε να μπορούν να απαντήσουν ερωτήματα συσσώρευσης και δεν υποκαθιστούν τις βάσεις δεδομένων που συνεχίζουν να διατηρούν τις πλήρεις λεπτομέρειες. Αντίστοιχα, σε μια TDW, μπορούμε να αποθηκεύσουμε μια αναπαράσταση της κίνησης με απώλειες αποκλειστικά και μόνο για να προσφέρουμε OLAP ανάλυση σε δεδομένα κίνησης και τα *M* και *M'* να αποθηκεύονται σε μια MOD.

Με βάση αυτήν την διατύπωση, αντικαθιστάμε τον παραδοσιακό πίνακα συμβάντων όπως ορίστηκε στη προηγούμενη υποενότητα με την εισαγωγή της έννοιας του πίνακα *TrajFact*, ο οποίος περιλαμβάνει τα κλειδιά για τους πίνακες διαστάσεων καθώς επίσης και το κατώτατο επίπεδο πληροφοριών που μπορούν να χρησιμοποιηθούν για τον υπολογισμό των μέτρων χρησιμοποιώντας τις διαφορετικές ερμηνείες του όρου «τροχιά». Ας ληφθεί υπόψη ότι οι διαφορετικοί συνδυασμοί κλειδιών στον πίνακα συμβάντων καθορίζουν μια μοναδική θέση στο πολυδιάστατο χώρο του κύβου (επίσης καλούνται *κελιά βάσης*). Σε αυτήν τη θέση, αντί να αποθηκεύουμε τα μέτρα, για κάθε τροχιά που «συμβάλλει» σε αυτά, μπορούμε να διατηρούμε μια ακολουθία χρονικών και χωρικών αποστάσεων μεταξύ των σημείων της (δείτε Εικόνα 3-17). Τυπικά:

Ορισμός 3-3 (TrajFact πίνακας): ο πίνακας *TrajFact* είναι μια n-αδική σχέση $K \times DT$, όπου:

i. *K* είναι το σύνολο των χαρακτηριστικών που αντιπροσωπεύει το πρωτεύον κλειδί του πίνακα συμβάντων που ορίζεται από $F_1 \times F_2 \times ... \times F_p$, όπου κάθε F_i, $1 \leq i \leq$ είναι ένα ξένο κλειδί στους πίνακες συμβάντων;
ii. Ένας πίνακας αποστάσεων που αποτελείται από:
    - Το προσδιοριστικό O_j, για κάθε αντικείμενο που «συνεισφέρει» στη μοναδική θέση στον πολυδιάστατο χώρο του κύβου που ορίζεται από το πρωτεύον κλειδί,
    - Μια ακολουθία ζευγαριών (sd_i, td_i), όπου sd_i είναι η ευκλείδεια απόσταση και td_i είναι η χρονική απόσταση μεταξύ δυο διαδοχικών χρονοσημασμένων σημείων θέσης (x_i, y_i, t_i) και $(x_{i-1}, y_{i-1}, t_{i-1})$ του αντικειμένου O_j.
iii. $n = p+1$. ■

Το παραπάνω αναπαρίσταται γραφικά στην Εικόνα 3-17, $(F_1, ..., F_p)$ είναι ένας τυχαίος συνδυασμός κλειδιών διάστασης. Για αυτήν την γραμμή ορίζεται ένας πίνακας απόστασης:



Εικόνα 3-17: Ο πίνακας *TrajFact*.

Κάποιος θα μπορούσε να υποστηρίξει ότι η χωρική/χρονική απόσταση μεταξύ των διαδοχικών σημείων είναι πληροφορίες που μπορούν να εξαχθούν, κατά συνέπεια θα μπορούσαμε να αποθηκεύσουμε στον πίνακα συμβάντων τα ίδια τα σημεία (x_i, y_i, t_i). Αυτό ισχύει αλλά, τα δύο κύρια ενδιαφέροντα κατά το χτίσιμο μιας αποθήκης τροχιών δεδομένων είναι: α) το μέγεθος της ΑΔ και β) ο χρόνος απόκρισης ενός ερωτήματος. Αποθηκεύοντας μόνο τις αποστάσεις (sd_i, td_i), μπορούμε να εξοικονομήσουμε το 1/3 του χώρου (σε σχέση με το αν αποθηκεύαμε τα σημεία) ή σχεδόν τα 2/3 του χώρου (εάν αποθηκεύαμε τα τμήματα των τροχιών) και ταυτόχρονα έχουμε όλες εκείνες τις πληροφορίες που χρειαζόμαστε προκειμένου να υπολογίσουμε τα μέτρα.

Έπειτα, χρειαζόμαστε έναν τρόπο να μοντελοποιήσουμε τις διαφορετικές σημασιολογικές ερμηνείες των τροχιών. Προς αυτήν την κατεύθυνση, δεν προτείνουμε ακόμα ένα πρότυπο αλλά ακολουθούμε την [SPD+08], η οποία θεωρεί τις χρονικές αποστάσεις μεταξύ των διαδοχικών σημείων ως ένα τρόπο προσδιορισμού των τροχιών από ένα ακατέργαστο σύνολο δεδομένων. Ως εκ τούτου, το πρότυπό μας επιτρέπει στο χρήστη να επιλέξει το μέγιστο χρονικό διάστημα για τον καθορισμό μιας τροχιάς συγκεκριμένης μορφής:

sem_{time}: Η *μέγιστη επιτρεπτή χρονική διάρκεια* μεταξύ δυο διαδοχικών χρονοσημασμένων θέσεων του ίδιου κινούμενου αντικειμένου.

Με άλλα λόγια, για κάθε χρονοσημασμένη θέση ορίζουμε μια χρονική περίοδος μέσα στην οποία εξετάζουμε εάν η επόμενη χρονοσημασμένη θέση μπορεί να θεωρηθεί ως τμήμα της ίδιας τροχιάς ή όχι. Η τιμή αυτής της παραμέτρου επιλέγεται από το χρήστη ώστε να καθοριστεί το επίπεδο χρονικής κλιμάκωσης, και επομένως να περιγραφεί μια ιδιαίτερη ερμηνεία της έννοιας της τροχιάς. Η επιλογή γίνεται κατά τη διάρκεια των ερωτημάτων συνάθροισης και συνεπώς δεν επηρεάζει καθόλου την γενικότερη οργάνωση του κύβου δεδομένων.

Η τιμή της sem_{time} χρησιμοποιείται σε ερωτήματα συνάθροισης όπου απαιτείται ο προσδιορισμός των τροχιών των αντικειμένων που βρίσκονται σε κάθε κελί βάσης. Όπως συζητήσαμε στην Υποενότητα 3.4.2.3, αυτό μπορεί να επιτευχθεί εξετάζοντας τους πίνακες απόστασης των κελιών βάσης. Αυτή η διαδικασία οδηγεί στην ανακάλυψη του συνόλου τροχιών με έναν δυναμικό τρόπο. Ως εκ τούτου, οι συναρτήσεις *len(T)* και *lifespan(T)* μπορούν και πάλι να εφαρμοστούν στο σύνολο τροχιών που θα ανακαλυφθεί αλλά χρησιμοποιώντας μια διαφορετική τεχνική υπολογισμού αξιοποιώντας την προσέγγιση μας. Πιο συγκεκριμένα, οι δύο συναρτήσεις συνοψίζουν τις χωρικές και χρονικές αποστάσεις που αποθηκεύονται αντίστοιχα στους πίνακες απόστασης.

Το μοντέλο μας χρησιμοποιεί την έννοια του πίνακα διάστασης όπως καθορίστηκε στην προηγούμενη ενότητα. Ως εκ τούτου, το σχήμα ενός ειδικού TDW μπορεί να οριστεί ως:

Ορισμός 3-4 (Σχήμα μιας κατά περίπτωση TDW): Το σχήμα μιας κατά περίπτωση TDW (*adhocTDW*) μπορεί να οριστεί ως *adhocTDW = (DT, TFT)*, όπου *DT* είναι ένα μη κενό πεπερασμένο σύνολο πινάκων διαστάσεων (Ορισμός 3-1) και *TFT* είναι ο πίνακας *TrajFact* (Ορισμός 3-3). ■

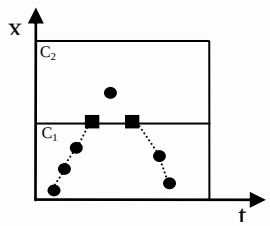
3.4.2.2. ETL επεξεργασία σε δεδομένα τροχιών

Μια διαδικασία ETL εκτελείται προκειμένου να τροφοδοτηθεί η ειδική TDW. Η φόρτωση δεδομένων στους πίνακες διάστασης είναι μια απλή διαδικασία. Αξίζει όμως να εστιάσουμε στη φόρτωση του πίνακα *TrajFact*. Αυτό επειδή, δεν εφαρμόζουμε εργασίες ανακατασκευής κατά τη διάρκεια της ETL φάσης (ή σε κάποια προηγούμενη φάση, κατά τη διάρκεια της φόρτωσης στην MOD) αλλά μετασχηματίζουμε κατάλληλα τα στοιχεία προκειμένου να τα φορτώσουμε στον πίνακα συμβάντων.

Χρησιμοποιώντας παλαιότερες προσεγγίσεις [MFN+08a], [OOR+07], ο υπολογισμός σχεδόν κάθε μέτρου (εκτός από το COUNT_USERS) υποθέτει ένα συγκεκριμένο σημασιολογικό ορισμό της τροχιάς. Για παράδειγμα, ο υπολογισμός του COUNT_TRAJECTORIES απαιτεί ένα συγκεκριμένο ορισμό της τροχιάς βάσει του οποίου θα μετρηθεί ο αριθμός των τροχιών.

Προτείνουμε μια πιο ευέλικτη ETL στρατηγική που βασίζεται στο μοντέλο που περιγράφεται στην προηγούμενη υποενότητα. Πιο συγκεκριμένα, θα πρέπει μόνο να υπολογίσουμε τις χρονικές/ χωρικές αποστάσεις μεταξύ των διαδοχικών θέσεων του ίδιου αντικειμένου. Όπως αναφέραμε ήδη, οι διαφορετικοί συνδυασμοί κλειδιών στον πίνακα συμβάντων καθορίζουν ένα μοναδικό κελί στο πολυδιάστατο χώρο του κύβου. Ως εκ τούτου, ο στόχος της ETL διαδικασίας είναι να βρει ποια σημεία ανήκουν σε αυτό το κελί και να υπολογίσει τους πίνακες αποστάσεων. Σε αυτήν την εργασία, δεν εξετάζουμε τα στοιχεία-θόρυβο και υποθέτουμε ότι εργαζόμαστε σε ένα καθαρό σύνολο δεδομένων, κατά συνέπεια υπολογίζουμε αποστάσεις που έχουν όντως νόημα.

Σημειώστε ότι προκειμένου να μελετηθεί η κίνηση των αντικειμένων μέσα στα κελιά, πρέπει να προσδιορίσουμε τα σημεία-περάσματα κάθε κελιού βάσης. Υποθέτουμε μια συνάρτηση γραμμικής παρεμβολής που θεωρεί ότι η κίνηση είναι ευθύγραμμη και με σταθερή ταχύτητα [PJT00]. Για παράδειγμα, στην Εικόνα 3-18, στο κελί C_1, τα τετράγωνα αντιπροσωπεύουν τα σημεία-περάσματα ενός συγκεκριμένου αντικειμένου. Σημειώστε ότι εξετάζουμε μόνο τις χωρικές/χρονικές αποστάσεις μεταξύ των σημείων που είναι μέσα στο C_1 και, προφανώς, δεν εισάγουμε τις αποστάσεις μεταξύ των σημείων-περασμάτων (για να είμαστε ακριβείς, θέτουμε τη χρονική και χωρική απόσταση μεταξύ των δύο τέτοιων σημείων ως μηδέν).



Εικόνα 3-18: Εφαρμόζοντας γραμμική παρεμβολή.

Η παραπάνω προσέγγιση έχει ενσωματωθεί στον αλγόριθμο LOAD-TDW-ETL και παρουσιάζεται στην Εικόνα 3-19. Ο αλγόριθμος αναζητά το σύνολο των σημείων που βρίσκεται σε κάθε κελί (γραμμή 4).

Τελικά για κάθε λίστα σημείων που ανήκει στο ίδιο αντικείμενο, υπολογίζεται η χωρική (Ευκλείδεια) και η χρονική απόσταση (γραμμή 8).

```
Algorithm Load-TDW-ETL(ListOfPoints LoP)
1. //We assume LoP is sorted by Object-id, timestamp
2. FOR EACH base cell bc_j DO
3. //Find the set of points inside the cell
4.  S = contains(LoP, bc_j);
5.    //Consider a list of points LP for
6.    //each object of S
7.    FOR EACH LP of S DO
8.     Compute_Distances(LP);
9.    END-FOR
10. END-FOR
```

Εικόνα 3-19: Ο αλγόριθμος Load-TDW-ETL για τη φόρτωση του πίνακα συμβάντων.

3.4.2.3. Κατά περίπτωση (ad-hoc) OLAP

Η προτεινόμενη οργάνωση της TDW χρειάζεται ένα νέο μηχανισμό OLAP που να αξιοποιεί την έννοια των πολλαπλών σημασιολογικών ορισμών των τροχιών. Σε αυτήν την υποενότητα, παρουσιάζουμε έναν αλγόριθμο που ενσωματώνει αυτήν την ιδιότητα και εκτελεί ερωτήσεις συνάθροισης με αποδοτικό τρόπο.

Ένα ερώτημα που τίθεται προς την TDW περιέχει διάφορα μέλη των διαστάσεων βάσει των οποίων τα κελιά βάσης φιλτράρονται και επιλέγεται ένα υποσύνολο αυτών καθώς και ένα μέτρο που υπονοεί μια συνάρτηση συνάθροισης (π.χ. SUM, MIN, MAX, AVG, COUNT, DISTINCT COUNT). Επιπλέον, περιέχει μια τιμή για την παράμετρο sem_{time}, η οποία θα χρησιμοποιηθεί για να προσδιορίσει τις διαφορετικές τροχιές. Το χρειαζόμαστε επειδή κάθε κελί βάσης δεν περιέχει καμία πληροφορία για τις τροχιές αλλά περιλαμβάνει έναν πίνακα που περιλαμβάνει τις χωρικές/ χρονικές αποστάσεις μεταξύ των διαδοχικών σημείων κάθε αντικειμένου. Τυπικά:

Ορισμός 3-5 (Ερώτημα συσσώρευσης): είναι ένα σύνολο $\{(bc_1, bc_2, \ldots, bc_i), tm, mtd\}$, όπου:

i. Κάθε bc_j, $1 \leq j \leq i$ είναι ένα κελί βάσης που φιλτράρεται με βάση τα επιλεγμένα μέλη μιας διάστασης (που τελικά οδηγούν σε συγκεκριμένες γραμμές του πίνακα *TrajFact*),
ii. *m* είναι το μέτρο,
iii. *mtd* είναι η επιλεγμένη τιμή για την παράμετρο sem_{time} ■

Στη συνέχεια, προτείνουμε τον αλγόριθμο AD-HOC-AGGREGATION που είναι κατάλληλα σχεδιασμένος ώστε να απαντάει ερωτήσεις συσσώρευσης και ενσωματώνει την ad-hoc προσέγγιση. Έτσι, ο χρήστης επιλέγει sem_{time} = *mtd* και το OLAP ερώτημα επιστρέφει ένα σύνολο *SOC* κελιών βάσης.

Λεπτομερώς, ο αλγόριθμος AD-HOC-AGGREGATION εξετάζει τον πίνακα που περιέχει τις χρονικές αποστάσεις σε κάθε κελί και για κάθε αντικείμενο (γραμμές 3-13), αναζητά τις χρονικές αποστάσεις με τιμές μικρότερες από *mtd* (γραμμή 5). Με αυτόν τον τρόπο, ο αλγόριθμος αναγνωρίζει ότι εξελίσσεται η ίδια τροχιά και έτσι ανανεώνει τα μέτρα (γραμμή 7) εφαρμόζοντας την αντίστοιχη συνάρτηση συνάθροισης *TM*. Αν βρεθεί χρονική απόσταση με τιμή μεγαλύτερη της *mtd* τότε προσδιορίζεται μια νέα τροχιά.

```
Algorithm General-Aggregation(SetOfCells SOC,
TargetMeasure TM, MaxTemporalDistance MTD)
1.  FOR EACH cell C of SOC DO
2.  //Process the distance table DT of the cell C
3.   FOR EACH object P of C DO
4.    FOR EACH TemporalDistance t of P in DT DO
5.     IF t < MTD THEN
6.      //the same trajectory evolves
7.      UpdateMeasure(TM);
8.     ELSE
9.      //a new trajectory is identified
10.      newTrajectory();
11.    END-IF
12.   END-FOR
13.  END-FOR
14. END-FOR
```

Εικόνα 3-20: Ο αλγόριθμος AD-HOC-AGGREGATION.

Όπως έχει ήδη αναφερθεί, οι συναρτήσεις συσσώρευσης που υπολογίζουν τις υπερ-συσσωρεύσεις των μέτρων κατηγοριοποιούνται από τους Gray κ.α. [GCB+97] σε τρεις κατηγορίες (διανεμητικές, αλγεβρικές και ολιστικές) βάση της πολυπλοκότητας που απαιτείται για τον υπολογισμό τους, ξεκινώντας από ένα σύνολο με ήδη διαθέσιμες υπο-συσσωρεύσεις. Η ίδια κατηγοριοποίηση ακολουθείται από τον OLAP μηχανισμό μας. Στις υπόλοιπες παραγράφους αυτής της υποενότητας, συζητάμε ζητήματα υπολογισμού, χρησιμοποιώντας το ad-hoc πλαίσιό μας, για τα μέτρα των [MFN+08a] και [OOR+07] και τα ταξινομούμε σε κατηγορίες σύμφωνα με την πολυπλοκότητά τους [GCB+97].

Το μέτρο COUNT_TRAJECTORIES κατηγοριοποιείται ως διανεμητικό, και υπολογίζεται με την εκτέλεση του αλγορίθμου AD-HOC-AGGREGATION στο σύνολο *SOC* των κελιών βάσης. Σε αυτό το σημείο, πρέπει να αναφέρουμε ότι, σε αυτήν την ενότητα, το μέτρο COUNT_TRAJECTORIES αναφέρεται στον αριθμό τροχιών και όχι στον αριθμό μοναδικών τροχιών που συζητείται στις [MFN+08a], [OOR+07]. Αυτό συμβαίνει επειδή στην προσέγγισή μας καμία τροχιά δεν έχει ανακατασκευαστεί σε προηγούμενη φάση, και κατά συνέπεια δεν έχει νόημα να αναζητήσουμε μοναδικές τροχιές στα κελιά βάσης, η οποία είναι η πηγή του προβλήματος μοναδικής προσμέτρησης [TKC+04]. Εντούτοις, συζητάμε μια διαφορετική έννοια του αριθμού μοναδικών τροχιών ως κομμάτι της μελλοντικής μας έρευνας.

Το μέτρο COUNT_USERS σε ένα κελί βάσης *bc*, αφορά τον *αριθμό των μοναδικών αντικειμένων*, και υπολογίζεται μετρώντας τον αριθμό των γραμμών στον πίνακα διάστασης *bc* αφού κάθε εγγραφή αντιστοιχεί σε ένα διαφορετικό αντικείμενο. Παρόλα αυτά, το μέτρο αυτό ταξινομείται ως ολιστικό αφού δε μπορούμε απλά να μετρήσουμε τα αποτελέσματα από όλα τα κελιά. Αυτό συμβαίνει επειδή ένα αντικείμενο μπορεί να εμφανίζεται σε περισσότερους από έναν πίνακες αποστάσεων (κελιά βάσης), οπότε θα πρέπει να αποφύγουμε να μετρήσουμε ένα αντικείμενο περισσότερες από μια φορές.

Τα υπόλοιπα μέτρα κατηγοριοποιούνται ως αλγεβρικά αφού υπολογίζονται στη βάση άλλων μέτρων. Για να υπολογίσουμε το μέτρο AVG_DISTANCE_TRAVELED μπορούμε να χρησιμοποιήσουμε τον τύπο (3.2). Όμως, το SUM_DISTANCE ορίζεται ως το άθροισμα των μηκών *len(T)* κάθε τροχιάς *T* που βρίσκεται μέσα σε ένα κελί βάσης *bc*. Τυπικότερα:

$$SUM\_DISTANCE(bc) = \sum_{T_i \in bc} len(T_i) \quad (3.15)$$

Έτσι, η συνάρτηση *len(T)* δεν εφαρμόζεται στα τμήματα των τροχιών που βρίσκονται μέσα στο *bc* αλλά στην τροχιά T που έχει ανακαλυφθεί με δυναμικό τρόπο όπως έχουμε ήδη συζητήσει.

Αντίστοιχα για το μέτρο AVG_TRAVEL_DURATION που μπορεί να υπολογιστεί χρησιμοποιώντας τον τύπο (3.3) αλλά το SUM_DURATION είναι επίσης ένα βοηθητικό μέτρο που ορίζεται ως το άθροισμα των διαρκειών *lifespan(T)* κάθε μιας τροχιάς *T* μέσα στο *bc*.

$$SUM\_DURATION(bc) = \sum_{T_i \in bc} lifespan(T_i) \quad (3.16)$$

Με τον ίδιο τρόπο, το μέτρο AVG_SPEED (τύπος (3.5)) υπολογίζεται διαιρώντας το βοηθητικό μετρό SUM_SPEED (δηλαδή το άθροισμα των ταχυτήτων κάθε τροχιάς *T* μέσα στο *bc*) με το COUNT_TRAJECTORIES:

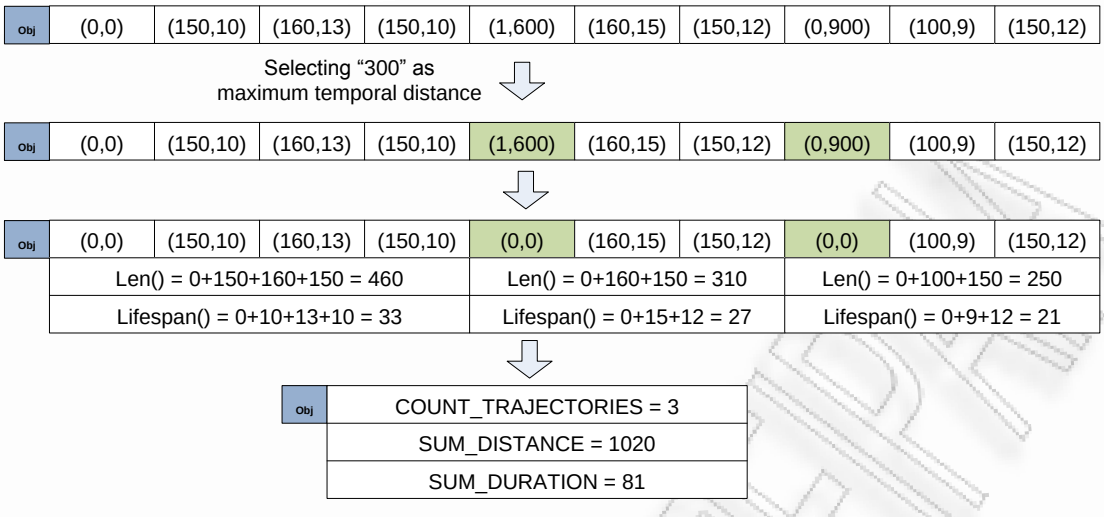
$$SUM\_SPEED(bc) = \sum_{T_i \in bc} \frac{len(T_i)}{lifespan(T_i)} \quad (3.17)$$

Αντίστοιχα, το μέτρο AVG_ABS_ACCELER υπολογίζεται βάσει του τύπου (3.7) όπου SUM_ABS_ACCELER είναι ένα βοηθητικό μέτρο που αθροίζει τις απόλυτες επιταχύνσεις όλων των τροχιών *T* που βρίσκονται στο *bc*

$$SUM\_ABS\_ACCELER(bc) = \sum_{T_i \in bc} \frac{\left| speed_{fin}(T_i) - speed_{init}(T_i) \right|}{lifespan(T_i)} \quad (3.18)$$

και $speed_{fin}$ ($speed_{init}$) είναι η τελική (αρχική, αντίστοιχα) καταγεγραμμένη ταχύτητα της τροχιάς T_i στο *bc*.

Ας παρουσιάσουμε τις ad-hoc OLAP λειτουργίες μας μέσα από ένα παράδειγμα. Θεωρήστε ότι ένας χρήστης ζητά τη συνολική απόσταση που καλύπτεται από τις τροχιές (SUM_DISTANCE), τη συνολική διάρκεια των τροχιών (SUM_DURATION) και τον αριθμό των τροχιών (COUNT_TRAJECTORIES) υπό συγκεκριμένους χωροχρονικούς περιορισμούς. Ο χρήστης καθορίζει τη μέγιστη χρονική απόσταση ως 300 δευτερόλεπτα. Η Εικόνα 3-21 επεξηγεί την πλήρη διαδικασία που εκτελείται για τον υπολογισμό των μέτρων SUM_DISTANCE και SUM_DURATION χρησιμοποιώντας τον πίνακα απόστασης ενός κελιού βάσης. Ο πρώτος πίνακας περιέχει τις χωρικές (σε μέτρα) και χρονικές αποστάσεις (σε δευτερόλεπτα) κάθε σημείου από ένα προηγούμενου του. Ως πρώτο βήμα, ο αλγόριθμος εντοπίζει τα σημεία με χρονική απόσταση μεγαλύτερη από εκείνου της απόστασης των 300 δευτερολέπτων. Στο παράδειγμα στην Εικόνα 3-21 υπάρχουν δύο τέτοια σημεία. Επομένως, αυτά τα σημεία θα είναι τα σημεία έναρξης των τροχιών (έτσι οι χρονικές και χωρικές αποστάσεις τους θέτονται μηδέν). Ως εκ τούτου, τρεις τροχιές προσδιορίζονται και στη συνέχεια υπολογίζονται τα μέτρα τους.



Εικόνα 3-21: Υπολογίζονται τα μέτρα COUNT_TRAJECTORIES, SUM_DISTANCE και SUM_DURATION.

3.4.2.4. Συζήτηση σχετικά με τον προϋπολογισμό του κύβου δεδομένων

Στις προηγούμενες υποενότητες, παρουσιάσαμε την αρχιτεκτονική μας για μια TDW που επιτρέπει ad-hoc ανάλυση OLAP στα στοιχεία κίνησης. Προφανώς, υπάρχει μια σχέση αλληλεπίδρασης (trade-off) μεταξύ της ευελιξίας και της απόδοσης. Μια TDW που στηρίζεται στις προκαθορισμένες και ανακατασκευασμένες τροχιές μπορεί να ξεπεράσει σε απόδοση μια κατά περίπτωση TDW. Εντούτοις, η πρώτη δεν προσφέρει καμία ευελιξία σχετικά με τους διαφορετικούς σημασιολογικούς ορισμούς των τροχιών. Σε μια παραδοσιακή ΑΔ, οι προϋπολογισμένες όψεις (materialized views) έχουν προταθεί για να επιταχύνουν την επεξεργασία των ερωτημάτων. Αυτές οι όψεις αναφέρονται ως *συνοπτικοί πίνακες* (summary tables) [CD97] που αποθηκεύουν τις πλεοναστικές, συγκεντρωτικές πληροφορίες. Το πλεονέκτημα των προϋπολογισμένων όψεων είναι το μικρό μέγεθός τους (έναντι των λεπτομερών αρχείων) που επιτρέπει την πολύ γρηγορότερη απάντηση των ερωτημάτων.

Ακολουθώντας την προαναφερθείσα προσέγγιση, ορίζουμε ένα κατάλληλο συνοπτικό πίνακα για την ad-hoc TDW με εγγραφές του τύπου:

Ορισμός 3-6 (Συνοπτικός πίνακας): Ένας συνοπτικός πίνακας *ST* είναι μια τετράδα *{bc, mtd, m, value}*, όπου:

i. *bc* είναι ένα συγκεκριμένο κελί βάσης,

ii. *mtd* είναι μια συγκεκριμένη μέγιστη χρονική απόσταση,

iii. *m* είναι ένα μέτρο; και

iv. *value* είναι η προϋπολογισμένη, συγκεντρωτική τιμή του μέτρου *m*. ■

Προφανώς, αυτή η στρατηγική αποδίδει καλά μόνο στην περίπτωση των διανεμητικών και των αλγεβρικών (που δεν βασίζονται σε ολιστικές) συναρτήσεις· αυτό συμβαίνει επειδή στην περίπτωση ενός ολιστικού μέτρου (π.χ. COUNT_USERS) οι υπερ-συσσωρεύσεις δεν μπορούν να υπολογιστούν από τις υπο-συσσωρεύσεις.

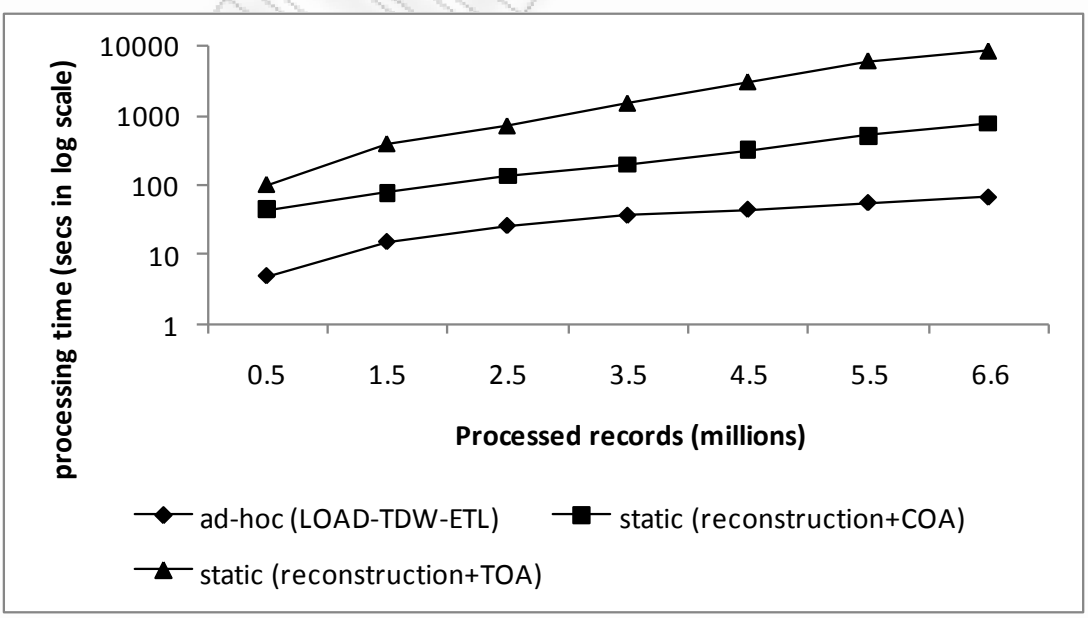
Ο πιο τετριμμένος τρόπος καθορισμού του *mtd* είναι να το γνωρίζουμε *εκ των προτέρων*. Πιο συγκεκριμένα, σε περίπτωση που μερικοί σημασιολογικοί ορισμοί των τροχιών είναι εκ των προτέρων γνωστοί, μπορούμε να προϋπολογίσουμε τα μέτρα ώστε να επιταχυνθεί η απάντηση του ερωτήματος.

Παρέχουμε αυτήν τη σύντομη συζήτηση προκειμένου να δώσουμε έμφαση στο γεγονός ότι προϋπολογίζοντας τον ευέλικτο κύβο δεδομένων μπορούμε να πετύχουμε τη λειτουργικότητα που παρέχεται από τους κύβους των [MFN+08a] και [OOR+07]. Αυτό συμβαίνει επειδή, σε αυτήν την περίπτωση, οι απαντήσεις έχουν ήδη υπολογιστεί όπως προτείνεται σε αυτές τις δυο εργασίες. Με άλλα λόγια, οι κύβοι δεδομένων που περιγράφονται σε αυτές τις δυο εργασίες μπορούν να θεωρηθούν ως συγκεκριμένοι προϋπολογισμένοι κύβοι της νέας αυτής προσέγγισης.

3.4.3. Πειραματική Μελέτη

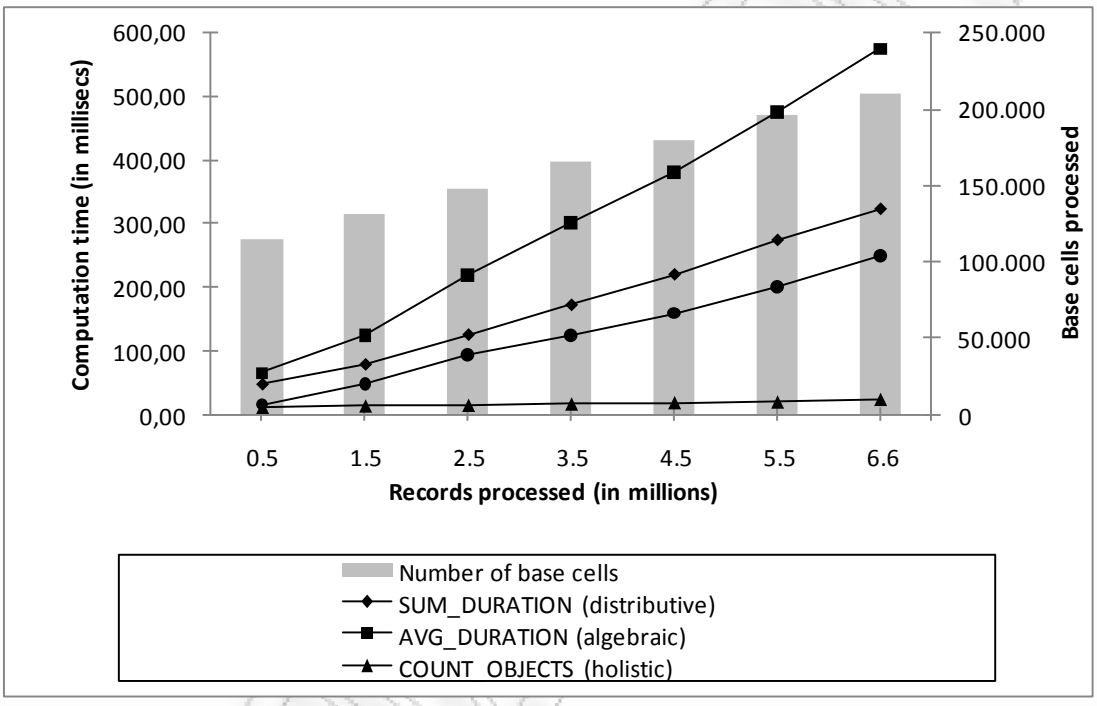
Σε αυτήν την ενότητα, αξιολογούμε τις προτεινόμενες λύσεις χρησιμοποιώντας το ίδιο σύνολο δεδομένων ([Eco09]) με αυτό της Υποενότητας 3.3.3. Περιέχει 6,67 εκατομμύρια πρωτογενών θέσεων που αντιπροσωπεύουν την κίνηση 84 υπαλλήλων εταιρία ταχυμεταφορών στην περιοχή του Λονδίνου για ένα μήνα (Ιούλιος 2007) με ρυθμό δειγματοληψίας τα 10 δευτερόλεπτα. Σε όλα τα πειράματα χρησιμοποιήσαμε ένα PC με 1 Gb RAM και P4 3 GHz CPU.

Παρακάτω, αξιολογούμε την απόδοση των βασικών συστατικών (ETL, μέθοδοι υπολογισμού, μέγεθος κύβου δεδομένων) της *ad-hoc προσέγγισής* που επιτρέπει την ανάπτυξη ευέλικτων κύβων δεδομένων τροχιάς. Επιπλέον, παρέχουμε σε μια σύγκριση με τα αντίστοιχα συστατικά της κλασικής TDW που προτείνεται στις [MFN+08a], [OOR+07] και περιγράφηκε στην Ενότητα 3.3 της Διατριβής. Σε αυτήν την υποενότητα, αναφερόμαστε στην τελευταία προσέγγιση ως *στατική* προσέγγιση. Για να επιτύχουμε αυτήν την σύγκριση χρησιμοποιούμε το μοντέλο TDW που παρουσιάζεται στο [MFN+08a] (το μοντέλο [OOR+07] μπορεί να θεωρηθεί ως υποσύνολο του) και αποτελείται από μια χωρική, χρονική, και μια διάσταση προφίλ του αντικειμένου, καθώς επίσης και τα μέτρα που έχουν ήδη συζητηθεί στην Υποενότητα 3.4.2.3.



Εικόνα 3-22: Απόδοση του αλγορίθμου Load-TDW-ETL και σύγκρισή του με το στατικό ETL.

Κατ' αρχάς, αξιολογούμε την αποτελεσματικότητα του αλγορίθμου LOAD-TDW-ETL (Εικόνα 3-22). Είναι σαφές ότι η απόδοση του είναι γραμμική συνάρτηση του μεγέθους του συνόλου δεδομένων (η επεξεργασία του πλήρους συνόλου δεδομένων επιτυγχάνεται σε περίπου 1 λεπτό). Το ETL βήμα της στατικής προσέγγισης περιλαμβάνει εργασίες ανακατασκευής τροχιών και τη φόρτωση του κύβου δεδομένων, και ακολουθεί μια προσέγγιση προσανατολισμένη είτε προς το κελί (διαδικασία *COA*) είτε μπρος την τροχιά (διαδικασία *TOA*). Προφανώς η προτεινόμενη ad-hoc προσέγγιση αποδίδει καλύτερα από τη στατική αφού δεν εκτελείται καμία εργασία ανακατασκευής τροχιών και επειδή η τροφοδότηση του κύβου δεδομένων δεν απαιτεί την ανακάλυψη των τμημάτων των τροχιών που βρίσκονται μέσα στα κελιά βάσης.

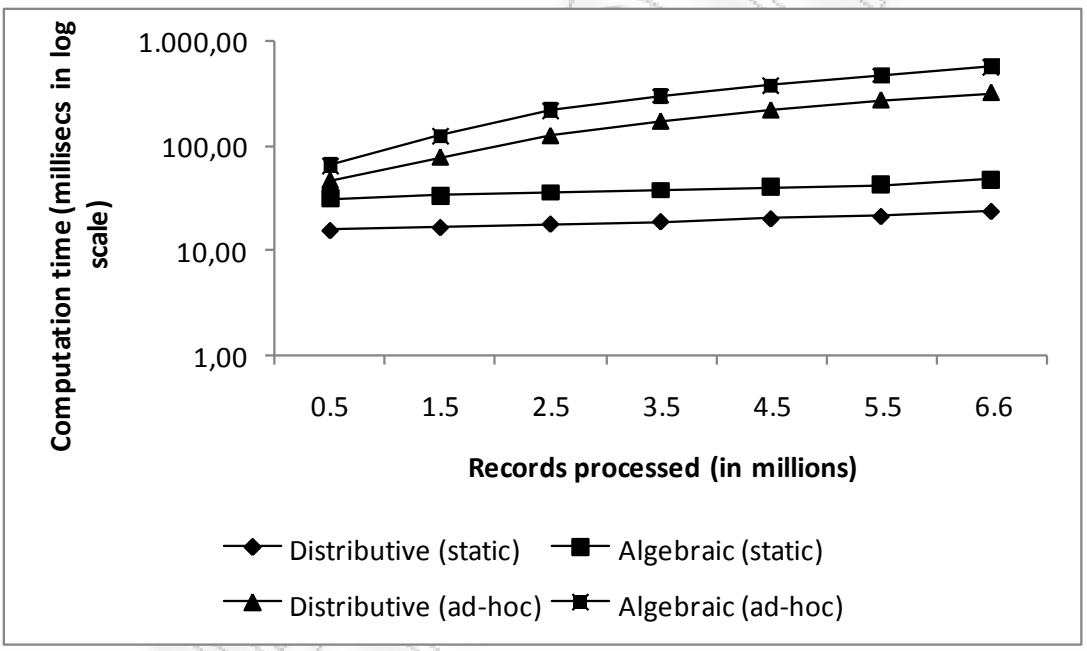


Εικόνα 3-23: Υπολογιστικοί χρόνοι στην κατά περίπτωση (ad-hoc) προσέγγιση.

Δεύτερον, αξιολογούμε το χρόνο που απαιτείται για τον υπολογισμό των διαφορετικών μέτρων (χρησιμοποιώντας τον αλγόριθμο AD-HOC-AGGREGATION). Πειραματιστήκαμε με διαφορετικά μεγέθη των συνόλων δεδομένων που οδηγούν σε έναν διαφορετικό αριθμό κελιών βάσης που θα πρέπει να υποστούν επεξεργασία. Η Εικόνα 3-23 παρουσιάζει τους χρόνους υπολογισμού για το πιο υψηλό επίπεδο κελιού του lattice (αριστερός κάθετος άξονας). Επιπλέον, παρουσιάζεται στο δεξί κάθετο άξονα ο αριθμός επεξεργασμένων κελιών βάσης για να δώσουμε μια εικόνα σχετικά με το μέγεθος του κύβου δεδομένων. Τα δύο διανεμητικά μέτρα (SUM_ DURATION και COUNT_TRAJECTORIES) παρουσιάζουν παρόμοια συμπεριφορά. Το τελευταίο έχει μια ελαφρώς καλύτερη απόδοση αφού σε αυτήν την περίπτωση ο αλγόριθμος προσδιορίζει μόνο τις διαφορετικές τροχιές και δεν υπολογίζει τίποτα άλλο (αντίθετα με ότι συμβαίνει στην περίπτωση του SUM_DURATION όπου οι αποστάσεις συνοψίζονται). Το AVG_DURATION είναι ένα αλγεβρικό μέτρο όπου η απόδοσή του είναι ευθέως ανάλογη με τα SUM_DURATION και COUNT_TRAJECTORIES. Αν και το μέτρο COUNT_USERS είναι ολιστικό, η απόδοση του είναι πολύ καλή δεδομένου ότι χρησιμοποιεί τη δομή που προτάθηκε στην

Υποενότητα 3.4.2.1. Επιπλέον, η απόδοσή του επηρεάζεται από τον αριθμό κελιών βάσης και όχι από τον αριθμό των εγγραφών. Αυτό οφείλεται στο γεγονός ότι ο υπολογισμός αυτού του μέτρου εξετάζει μόνο τα προσδιοριστικά των αντικειμένων από τους πίνακες απόστασης κάθε κελιού βάσης.

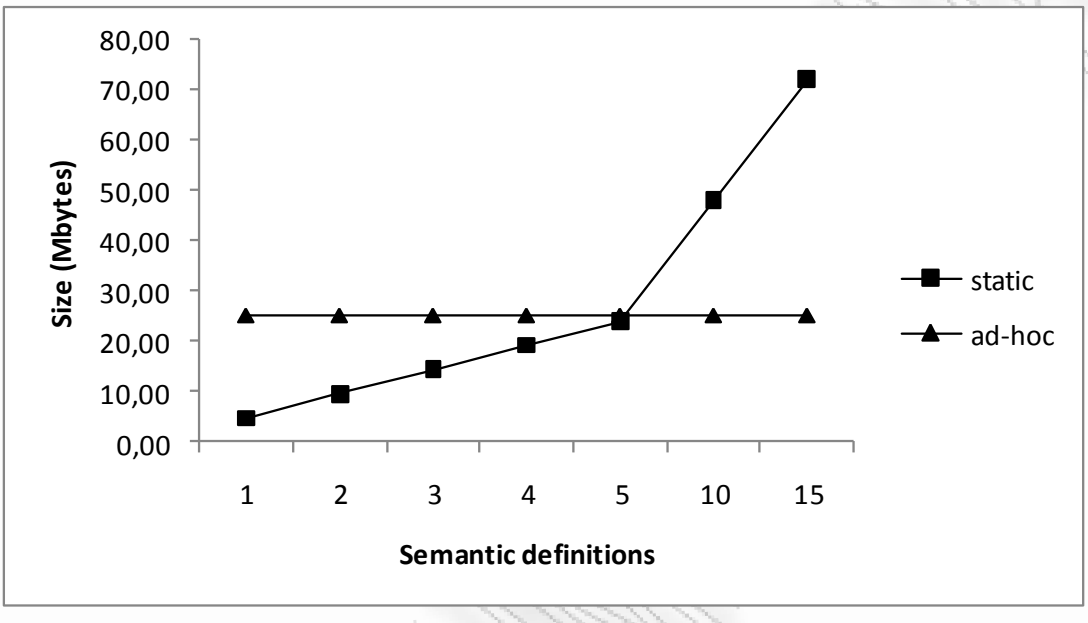
Η Εικόνα 3-24 αναλύει τον υπολογισμό ενός διανεμητικού και ενός αλγεβρικού μέτρου χρησιμοποιώντας τον ευέλικτο κύβο δεδομένων (χωρίς προϋπολογισμό) και έναν προϋπολογισμένο κύβο δεδομένων χρησιμοποιώντας είτε τους συνοπτικούς πίνακες όπως συζητήθηκε στην Υποενότητα 3.4.2.4 είτε τους *στατικούς* κύβους. Προφανώς, ο υλοποιημένος κύβος αποδίδει καλύτερα αφού οι τιμές των μέτρων είναι προϋπολογισμένες αλλά δεν παρέχει οποιαδήποτε ευελιξία σχετικά με τους διαφορετικούς ορισμούς των τροχιών (εδώ εντοπίζεται και η σχέση αλληλεπίδρασης μεταξύ της ευελιξίας και της απόδοσης). Δεν παρέχουμε σύγκριση αναφορικά με τα ολιστικά μέτρα επειδή αφενός δεν είναι δυνατό να χρησιμοποιηθούν οι συνοπτικοί πίνακες για τον υπολογισμό τους και αφετέρου οι κύβοι των [MFN+08a], [OOR+07] δεν υποστηρίζουν τέτοια μέτρα (υποστηρίζουν μόνο διανεμητικά και αλγεβρικά μέτρα).



Εικόνα 3-24: Συγκρίνοντας υπολογιστικούς χρόνους: ad-hoc και στατική προσέγγιση.

Στο τέταρτο πείραμά μας συγκρίνουμε τα μεγέθη των κύβων δεδομένων που υλοποιούνται χρησιμοποιώντας είτε την κατά περίπτωση είτε τη στατική προσέγγιση. Υπενθυμίζουμε ότι ο σκοπός μας είναι να χτίσουμε έναν κύβο δεδομένων που θα είναι αρκετά γενικός ώστε να εξυπηρετεί διάφορες εφαρμογές. Υποθέτουμε ότι αυτές οι εφαρμογές απαιτούν μερικούς διαφορετικούς ορισμούς της έννοιας της τροχιάς. Και οι δύο κύβοι αποτελούνται από 210.000 κελιά βάσης που περιέχουν το πλήρες σύνολο δεδομένων (6,6 εκατομμύρια εγγραφές). Προφανώς, το μέγεθος του ειδικού κύβου δεδομένων παραμένει το ίδιο ακόμα κι αν υπάρχει ένας μεγάλος αριθμός διαφορετικών ορισμών. Επιπλέον, ούτε ο αριθμός των μέτρων που υπολογίζονται από αυτόν τον κύβο δεδομένων δεν παίζει κάποιο ρόλο δεδομένου ότι οι απαντήσεις δεν είναι προϋπολογισμένες. Το μέγεθός του είναι ανάλογο με τον αριθμό σημείων θέσης αφού ένας ad-hoc κύβος δεδομένων αποθηκεύει τις αποστάσεις μεταξύ

αυτών των σημείων. Από την άλλη μεριά, το μέγεθος των στατικών κύβων είναι ανάλογο με τον αριθμό μέτρων που υπολογίζονται σε αυτόν τον κύβο στοιχείων και τον αριθμό των κελιών βάσης (για κάθε κελί βάσης θα πρέπει να υπολογιστούν οι διαφορετικές τιμές). Όπως βλέπουμε στην Εικόνα 3-25, η προσέγγισή μας αποδίδει σε ότι αφορά το μέγεθος του κύβων δεδομένων εάν έχουμε πέντε ή περισσότερους ορισμούς.



Εικόνα 3-25: Συγκρίνοντας μεγέθη των κύβων δεδομένων: ad-hoc και στατική προσέγγιση.

3.5. Σχετικές Εργασίες

Στη συνέχεια, εξετάζουμε τις σχετικές εργασίες στις περιοχές των ΑΔ για Χωρικά και Χωροχρονικά δεδομένα. Αυτές μπορούν να θεωρηθούν ως πρόγονοι των TDW.

3.5.1. Αποθήκες Χωρικών Δεδομένων

Η πρωτοπόρα εργασία των Han κ.α. [HSK98] εισάγει την έννοια της Αποθήκευσης Χωρικών Δεδομένων (ΑΧΔ). Οι συγγραφείς επεκτείνουν την ιδέα των διαστάσεων ώστε να περιληφθούν χωρικές και μη-χωρικές, και των μέτρων κύβων ώστε να αντιπροσωπεύουν γεωγραφικές περιοχές ή/και αριθμητικά δεδομένα. Μετά από αυτήν την εργασία, έχουν προταθεί στη βιβλιογραφία διάφορα μοντέλα στοχεύοντας στον εμπλουτισμό των κλασσικών μοντέλων αποθηκών δεδομένων με χωρικές έννοιες και των εργαλείων OLAP με τους χωρικούς τελεστές (Spatial OLAP - SOLAP). Εντούτοις, παρά την πολυπλοκότητα των χωρικών δεδομένα, οι σημερινές ΑΧΔ συνήθως περιέχουν αντικείμενα με απλές γεωμετρικές διαστάσεις. Επιπλέον, ενώ ένα μοντέλο ΑΧΔ υποτίθεται ότι αποτελείται από ένα σύνολο εννοιών αναπαράστασης και μια άλγεβρα των τελεστών SOLAP για τη διερεύνηση των δεδομένων, τη συνάθροιση και την απεικόνιση στοιχείων, οι προσεγγίσεις που προτείνονται στη βιβλιογραφία εστιάζουν είτε στις έννοιες είτε την άλγεβρα. Οι προσεγγίσεις που εξετάζουν και τα δύο είναι σπάνιες.

Η έρευνα στη μοντελοποίηση ΑΧΔ μπορεί να κατηγοριοποιηθεί με βάση το πώς αντιμετωπίζει τις απαιτήσεις των εφαρμογών είτε στο λογικό είτε το εννοιολογικό επίπεδο δεδομένων. Οι επικρατούσες

λύσεις στηρίζονται στο (σε λογικό επίπεδο) σχεσιακό μοντέλο δεδομένων [BMH01], [SHK00]. Σχετικά λίγες προσεγγίσεις εστιάζουν στις εννοιολογικές πτυχές των ΑΧΔ [JKP+04], [MZ04b], [BTM05], [TPG+01]. Η ανάλυση που παρουσιάζεται στην [Riz03] βεβαιώνει το μέτριο ενδιαφέρον της ερευνητικής κοινότητας για την εννοιολογική πολυδιάστατη μοντελοποίηση. Εντούτοις, ένα σημαντικό ποσοστό ΑΔ αποτυγχάνει να επιτύχει τους επιχειρησιακούς στόχους τους [Riz03]. Ένας σημαντικός λόγος για την αποτυχία είναι ο φτωχός ή ακατάλληλος σχεδιασμός, που οφείλεται κυρίως σε μια έλλειψη καθιερωμένων μεθόδων σχεδιασμού ΑΔ [RG00] και εννοιολογικών μοντέλων δεδομένων ΑΔ [RG00]. Ομοίως, οι συγγραφείς της [MZ06] δηλώνουν ότι τα προτεινόμενα μοντέλα είτε παρέχουν μια γραφική αναπαράσταση που βασίζεται στο E-R μοντέλο ή σε UML συμβολισμούς με ελάχιστους φορμαλιστικούς ορισμούς, ή παρέχουν τους τέτοιους ορισμούς χωρίς οποιαδήποτε γραφική υποστήριξη προσανατολισμένη προς το χρήστη.

Εστιάζοντας στη χωρική μοντελοποίηση, οι υπάρχουσες προσεγγίσεις δεν στηρίζονται σε τυποποιημένα μοντέλα δεδομένων για την αναπαράσταση των χωρικών πτυχών. Η χωρικότητα των συμβάντων αναπαρίσταται συνήθως μέσω ενός γεωμετρικού στοιχείου, αντί ενός OGC (Open Geospatial Consortium) χωρικού γνωρίσματος, δηλ., ενός αντικείμενου που να περιέχει και μια σημασιολογική αξία εκτός από το χωρικό χαρακτηρισμό του [Ope01].

Η επέκταση των κλασσικών μοντέλων ΑΔ ώστε να μπορούν να χειριστούν χωρικά δεδομένα απαιτεί τόσο οι διαστάσεις όσο και τα μέτρα να μπορούν να διατηρούν χωρικά και τοπολογικά χαρακτηριστικά. Πράγματι οι διαστάσεις και τα μέτρα πρέπει να επεκταθούν ώστε να είναι χωρικά προκειμένου να εμπλουτιστεί ο σχηματισμός του ερωτήματος και η οπτικοποίηση των αποτελεσμάτων. Εντούτοις η προσθήκη της χωρικότητας στις διαστάσεις και στα μέτρα δεν είναι από μόνη της αρκετή. Μια ΑΧΔ έχει επιπλέον συγκεκριμένες απαιτήσεις που έχουν αναφερθεί στην Ενότητα 3.2, όπως τα διαφορετικά είδη χωρικών διαστάσεων και μέτρων, οι πολλαπλές ιεραρχίες στις διαστάσεις, οι σχέσεις μερικής συμμετοχής μεταξύ των επιπέδων των διαστάσεων, οι μη-κανονοκιποιημένες ιεραρχίες, σχέσεις πολλά-προς-πολλά μεταξύ των μέτρων και των διαστάσεων και η διαμόρφωση των μέτρων ως σύνθετες οντότητες [BTM05], [BMH01], [JKP+04].

3.5.1.1. Χωρικές διαστάσεις

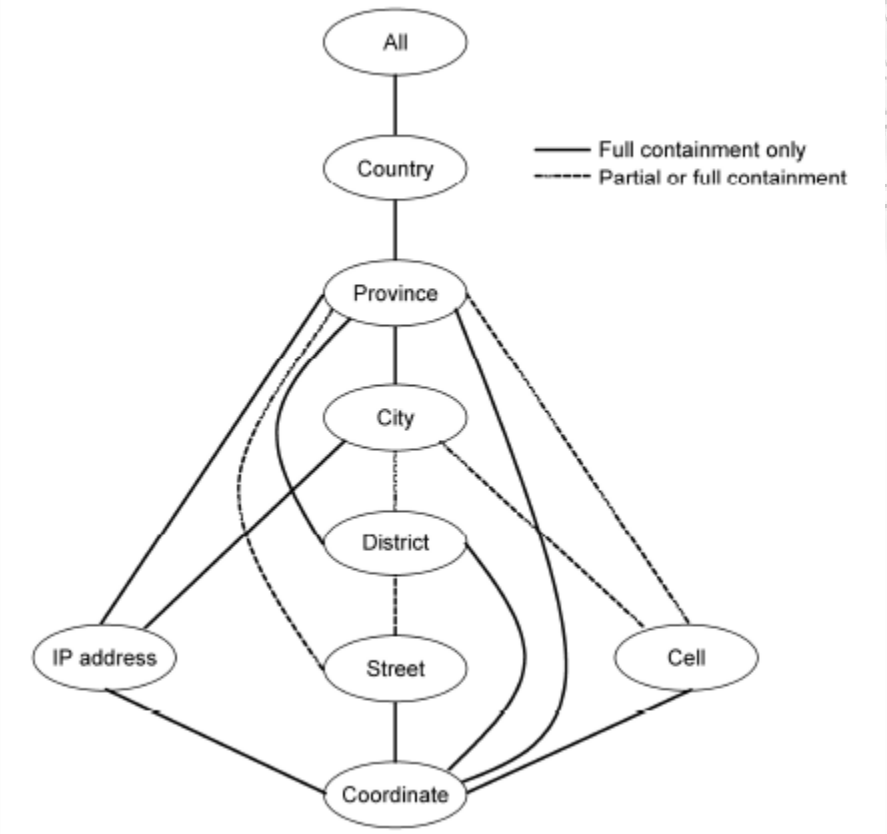
Για την προσθήκη της χωρικότητας στις διαστάσεις, οι περισσότερες προτάσεις ακολουθούν τις προσεγγίσεις των Stefanovic κ.α. [SHK00] και Bédard κ.α. [BMH01] που διακρίνουν τρεις τύπους ιεραρχιών διαστάσεων βάσει των χωρικών αναφορών των μελών της ιεραρχίας: μη-γεωμετρικές, γεωμετρικές-προς-μη-γεωμετρικές και πλήρως γεωμετρικές χωρικές διαστάσεις. Η μη-γεωμετρική χωρική διάσταση χρησιμοποιεί ονομαστικές (nominal) χωρικές αναφορές (π.χ. το όνομα των πόλεων και των χωρών) και αντιμετωπίζεται ως οποιαδήποτε άλλη περιγραφική διάσταση [RBM01], [RBP+05]. Οι δύο άλλοι τύποι αφορούν τις διαστάσεις όπου τα μέλη του χαμηλότερου ή και όλων των επιπέδων συνδέονται με μια γεωμετρία. Στην πλήρως γεωμετρική χωρική διάσταση, όλα τα μέλη όλων των επιπέδων είναι γεω-αναφερόμενα ενώ στη γεωμετρική-προς-μη-γεωμετρική χωρική διάσταση, τα μέλη είναι γεω-αναφερόμενα μέχρι ενός ορισμένου επιπέδου διάστασης και έπειτα γίνονται μη-γεωμετρικά.

Οι Malinowski κ.α. [MZ04b] επεκτείνουν αυτήν την κατηγοριοποίηση και θεωρούν ότι μια διάσταση μπορεί να είναι χωρική ακόμη και ελλείψει διάφορων σχετικών χωρικών επιπέδων. Στην πρότασή τους, ένα χωρικό επίπεδο ορίζεται ως ένα επίπεδο για το οποίο η εφαρμογή πρέπει να κρατήσει τα χωρικά χαρακτηριστικά της, δηλαδή τη γεωμετρία της όπως αυτή αντιπροσωπεύεται από τους τυποποιημένους χωρικούς τύπους (π.χ. σημεία, περιοχές). Αυτό επιτρέπει τη σύνδεση των χωρικών επιπέδων μιας διάστασης μέσω των τοπολογικών σχέσεων που υπάρχουν μεταξύ των χωρικών συστατικών των μελών τους (περιέχει, είναι ίσο με, επικαλύπτει, κ.λπ.). Με βάση αυτό, ορίζουν μια χωρική ιεραρχία ως ιεραρχία που περιλαμβάνει τουλάχιστον ένα χωρικό επίπεδο. Υπό αυτήν τη μορφή, μια χωρική διάσταση είναι μια διάσταση που περιέχει τουλάχιστον ένα χωρικό επίπεδο διαφορετικά θεωρείται μια θεματική διάσταση. Ένα πλεονέκτημα αυτής της μοντελοποίησης είναι ότι οι διαφορετικοί χωρικοί τύποι δεδομένων συνδέονται με τα επίπεδα μιας ιεραρχίας. Παραδείγματος χάριν, υποθέτοντας την ιεραρχία *χρήστης* < *πόλη* < *νομός*, ο τύπος σημείο συσχετίζεται με το *χρήστη*, η περιοχή με την *πόλη*, και το σύνολο περιοχών με το *νομό*.

Η διατήρηση των διαστάσεων και η οργάνωσή τους σε ιεραρχίες είναι πολύ απλή στις παραδοσιακές αποθήκες στοιχείων. Τα επίπεδα των παραδοσιακών μη-χωρικών διαστάσεων συνήθως οργανώνονται σε ιεραρχίες συμμετοχής όπως *περιοχή* < *πόλη* < *νομός* < *χώρα*. Εντούτοις όταν έχουμε να κάνουμε με χωρικά στοιχεία, τότε δεν ισχύει ότι δυο χωρικές τιμές μπορούν είτε να είναι ανεξάρτητες είτε η μια να περιέχεται στην άλλη, αφού μπορούν και να επικαλύπτονται. Για παράδειγμα, εάν προσθέτουμε την *κυψέλυ* ως επίπεδο διάστασης πριν από το επίπεδο *περιοχή*, τότε θα υπάρξει επικάλυψη αφού μια κυψέλη μπορεί να επικαλύπτει δυο περιοχές. Άλλες εργασίες εξερευνούν ένα μεγαλύτερο φάσμα πιθανών εφαρμογών προκειμένου να καλυφθούν οι απαιτήσεις από διάφορες εφαρμογές. Οι Jensen κ.α. [JKP+04] προτείνουν ένα εννοιολογικό μοντέλο που υποστηρίζει τις διαστάσεις με τις πλήρεις ή μερικές σχέσεις συμμετοχής (δείτε την Εικόνα 3-26). Οι ιεραρχίες μιας διάστασης μπορούν να περιέχουν επίπεδα που μπορούν να συνδέονται από τις πλήρεις ή μερικές σχέσεις συμμετοχής. Για τα μέλη ενός επιπέδου που συνδέονται με σχέση μερικής συμμετοχής με τα μέλη ενός άλλου επιπέδου, ο βαθμός συμμετοχής πρέπει να οριστεί (π.χ. 80% αυτού του κελιού περιλαμβάνεται σε αυτήν την περιοχή).

Η υποστήριξη πολλαπλών ιεραρχιών σε μια διάσταση είναι επίσης μια σημαντική απαίτηση που προτείνεται από τα μοντέλα των Jensen κ.α. [JKP+04] και Malinowski κ.α. [MZ06]. Αυτό συνεπάγεται ότι μπορούν να υπάρχουν πολλαπλές διαδρομές συνάθροισης σε μια διάσταση (π.χ. οι κυψέλες μπορούν να συναθροιστούν στις περιοχές ή άμεσα στους νομούς). Σύμφωνα με αυτά τα μοντέλα, οι πολλαπλές διαδρομές συνάθροισης επιτρέπουν καλύτερο χειρισμό της ανακρίβειας των ερωτήσεων που προκαλούνται από τις μερικές σχέσεις συμμετοχής. Βάζοντας αυτήν την ιδέα στο ανωτέρω παράδειγμα, υποστηρίζουν ότι το αποτέλεσμα της συνάθροισης των κελιών στο νομό μπορεί να δώσει καλύτερα αποτελέσματα σε σχέση με αυτά της συνάθροισης των κελιών στην περιοχή, έπειτα στην πόλη και έπειτα στο νομό. Τα μοντέλα των Jensen κ.α. [JKP+04] και Malinowski κ.α. [MZ06] υποστηρίζουν μη-κανονικοποιημένες ιεραρχίες δηλ., ιεραρχίες των οποίων τα μέλη μπορούν να έχουν περισσότερα από ένα αντίστοιχα μέλη στο υψηλότερο επίπεδο ή κανένα αντίστοιχο μέλος (π.χ. μια κυψέλη μπορεί να έχει σχέση με δύο περιοχές ενώ μια περιοχή μπορεί να μην περιέχει καμία κυψέλη). Τέλος, στο μοντέλο Malinowski κ.α. [MZ06], οι απλές ιεραρχίες χαρακτηρίζονται ως: συμμετρικές

(δηλ. όλα τα επίπεδα της ιεραρχίας είναι υποχρεωτικά), ασύμμετρες, γενικευμένες (δηλ. συμπεριλαμβανομένης μιας σχέσης γενίκευσης/ειδίκευσης μεταξύ των μελών της διάστασης), μη-αυστηρών (το ίδιο με τις μη-κανονικοποιημένες) και μη-καλυπτόμενες (δηλ. μερικά επίπεδα της ιεραρχίας μπορούν να αγνοηθούν κατά τη συσσώρευση).



Εικόνα 3-26: Ιεραρχία με πλήρη και μερική σχέση συμμετοχής (από [JKP+04]).

3.5.1.2. Χωρικά μέτρα

Ομοίως με τις χωρικές διαστάσεις, όταν προσθέτουμε χωρικότητα στα μέτρα, οι περισσότερες από τις προτάσεις διακρίνουν δύο τύπους χωρικών μέτρων [HSK98], [RBM01], [RBP+05]: χωρικά μέτρα που αναπαρίστανται από μια γεωμετρία και που συνδέονται με έναν γεωμετρικό τελεστή για να τα συναθροίσει βάσει των διαστάσεων, μια αριθμητική αξία που υπολογίζεται χρησιμοποιώντας έναν τοπολογικό ή μετρικό τελεστή.

Τα χωρικά μέτρα που αντιπροσωπεύονται από μια γεωμετρία αποτελούνται είτε από ένα σύνολο συντεταγμένων [BTM05], [MZ04b], [PT01], [RBM01], [RBP+05] ή από ένα σύνολο δεικτών στα γεωμετρικά αντικείμενα όπως στη [SHK00]. Τέλος, οι Bimonte κ.α. [BTM05] και Malinowski κ.α. [MZ04b] υποστηρίζουν τον καθορισμό των μέτρων ως σύνθετες οντότητες. Στην [BTM05], ένα μέτρο είναι ένα αντικείμενο που περιέχει διάφορες ιδιότητες (χωρικές ή μη) και διάφορες συναρτήσεις συνάθροισης (τελικά ειδικές λειτουργίες). Με παρόμοιο τρόπο ορίζουν οι Malinowski κ.α. [MZ04b] τα μέτρα ως ιδιότητες μιας n-αδικής σχέσης συμβάντων μεταξύ των διαστάσεων. Αυτή η σχέση συμβάντων μπορεί να είναι χωρική, εάν συνδέει τουλάχιστον δύο χωρικές διαστάσεις, και μπορεί να συνδέεται με έναν χωρικό περιορισμό όπως, για παράδειγμα, η χωρική συμμετοχή.

Ένα σημαντικό ζήτημα σχετικό με τα χωρικά μέτρα αφορά το επίπεδο λεπτομέρειας στο οποίο περιγράφονται. Πράγματι τα χωρικά στοιχεία είναι συχνά διαθέσιμα και περιγράφονται σε διάφορα επίπεδα λεπτομέρειας: για παράδειγμα, το ίδιο χωρικό αντικείμενο μπορεί να οριστεί ως μια περιοχή, σύμφωνα με ένα ακριβές επίπεδο λεπτομέρειας, αλλά και ως σημείο σύμφωνα με ένα λιγότερο λεπτομερές. Αυτό είναι ιδιαίτερα σημαντικό για τις τροχιές όπου η θέση των αντικειμένων μπορεί να μην είναι απόλυτα ακριβής. Οι Damiani κ.α. [DS06] προτείνουν ένα μοντέλο που επιτρέπει τον καθορισμό των χωρικών μέτρων σε διαφορετικές χωρικές κλιμακώσεις. Αυτό το μοντέλο, που καλείται MuSD, επιτρέπει την αναπαράσταση των χωρικών μέτρων και διαστάσεων με βάση τα χαρακτηριστικά του OGC. Ένα χωρικό μέτρο μπορεί να αντιπροσωπεύσει τη θέση ενός συμβάντος σε πολλαπλά επίπεδα χωρικής κλιμάκωσης. Τέτοια πολυ-κλιμακωτά χωρικά μέτρα μπορούν είτε να αποθηκευτούν είτε να υπολογιστούν δυναμικά με την εφαρμογή ενός συνόλου τελεστών εκτράχυνσης (coarsening). Στην [DS06] προτείνεται μια άλγεβρα SOLAP τελεστών συμπεριλαμβανομένων των ειδικών χειριστών που επιτρέπουν την αναρρίχηση (scaling up) των χωρικών μέτρων στις διαφορετικές κλιμακώσεις.

Μια άλλη απαίτηση που τονίζεται από τους Jensen κ.α. [JKP+04] και Bimonte κ.α. [BTM05] αφορά τις σχέσεις μεταξύ των μέτρων και των διαστάσεων. Πράγματι ενώ τα περισσότερα μοντέλα προτείνουν μόνο τον καθορισμό των 1:1 σχέσεων μεταξύ μέτρων και διαστάσεων, οι συγγραφείς υποστηρίζουν τον ορισμό πολλαπλών σχέσεων, οι οποίες θα επέτρεπαν τη συσχέτιση του ίδιου μέτρου με διάφορα μέλη μιας διάστασης.

3.5.2. Αποθήκες Χωροχρονικών Δεδομένων

Η έρευνα για την εξαγωγή σημασιολογικά πλούσιων πληροφοριών από τα πρωτογενή χωρικά/χρονικά δεδομένα έχει εστιάσει στις χωρικές και χωροχρονικές αποθήκες δεδομένων. Δεδομένου ότι θα επιθυμούσαμε να αντιμετωπίσουμε τις TDW ως κλάδο της χωροχρονικής αποθήκευσης [GKM+09], οι δύο επόμενες υποενότητες παρουσιάζουν σχετικές προσεγγίσεις σε αυτήν την περιοχή και κατηγοριοποιούν τις ερευνητικές προσπάθειες, αφ' ενός, στις εννοιολογικές και λογικές μεθοδολογίες μοντελοποίησης, και, αφετέρου, στα ζητήματα υλοποίησης στις τεχνικές συνάθροισης ως πεμπτουσία της έννοιας της αποθήκευσης δεδομένων.

3.5.2.1. Συναρτήσεις συσσώρευσης και ζητήματα υλοποίησης

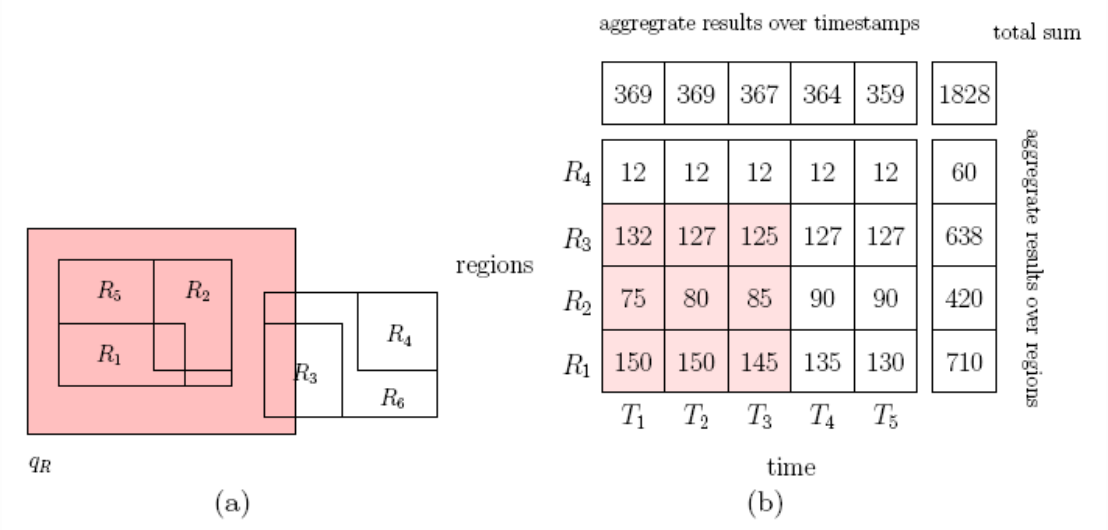
Ένα σχετικό ερευνητικό ζήτημα που έχει πρόσφατα συγκεντρώσει αυξανόμενο ενδιαφέρον και είναι σχετικό με την ανάπτυξη περιεκτικών μοντέλων δεδομένων ΑΧΔ αφορά την προδιαγραφή και την αποδοτική υλοποίηση των τελεστών για χωρική και χωροχρονική συνάθροιση.

Οι χωρικές διαδικασίες συνάθροισης συνοψίζουν τις γεωμετρικές ιδιότητες των αντικειμένων και υπό αυτήν τηη έννοια αποτελούν την ιδιαίτερη πτυχή μιας ΑΧΔ. Εντούτοις, ένα καθορισμένο σύνολο τελεστών (όπως για παράδειγμα των SQL χειριστών SUM, AVG, MIN) δεν έχει οριστεί ακόμα. Στην πραγματικότητα, κατά τον καθορισμό των χωρικών, χρονικών και χωροχρονικών συναθροίσεων θα πρέπει να αντιμετωπιστούν μερικά πρόσθετα προβλήματα, τα οποία δεν εμφανίζονται στην περίπτωση των παραδοσιακών δεδομένων.

Συγκεκριμένα, ενώ στις παραδοσιακές βάσεις δεδομένων ενδιαφερόμαστε μόνο για τα σαφή γνωρίσματα, η μοντελοποίηση της χωρικής και χρονικής έκτασης ενός αντικειμένου χρησιμοποιεί τις ερμηνευμένες ιδιότητες και ο καθορισμός των συναθροίσεων βασίζεται στις κλιμακώσεις.

Μια πρώτη περιεκτική κατηγοριοποίηση και διαμόρφωση των χωροχρονικών συναρτήσεων συσσώρευσης παρουσιάζονται από τους Lopez κ.α. [LT05]. Η λειτουργία της συνάθροισης ορίζεται ως μια συνάρτηση που εφαρμόζεται σε μια συλλογή εγγραφών και επιστρέφει μια ενιαία τιμή. Οι συγγραφείς διακρίνουν τρία είδη μεθόδων προκειμένου να παραχθεί η συλλογή εγγραφών στην οποία εφαρμόζεται η συνάρτηση: ομαδική σύνθεση (group composition), σύνθεση χωρισμάτων (partition composition) και σύνθεση παραθύρων ολίσθησης (sliding window).

Θυμίζουμε ότι μια (χρονική/ χωρική) κλιμάκωση δημιουργεί μια *διακριτή* εικόνα, από άποψη *κλιμάκωσης*, της (χρονικής/χωρικής) περιοχής. Λαμβάνοντας υπόψη μια χωρική κλιμάκωση G^S και μια χρονική κλιμάκωση G^T, μια *χωροχρονική ομαδική σύνθεση* διαμορφώνει τις ομάδες εγγραφών που μοιράζονται την ίδια χωρική και χρονική τιμή στην κοκκοποίηση G^S x G^T. Μια συνάρτηση συσσώρευσης μπορεί έπειτα να εφαρμοστεί σε κάθε ομάδα. Από την άλλη μεριά, η *χωροχρονική ομαδική σύνθεση* χρησιμοποιείται όταν απαιτείται ένα λεπτότερο επίπεδο συνάθροισης και περιλαμβάνει τουλάχιστον δύο κλιμακώσεις. Η πρώτη, που είναι περισσότερο τραχιά, καθορίζει τις συλλογές εγγραφών (τα χωρίσματα). Σε κάθε χώρισμα, εκτελείται μια *σύνθεση παραθύρων ολίσθησης*. Αντί της παραγωγής μιας ενιαίας συγκεντρωτικής τιμής για κάθε χώρισμα, υπολογίζεται μια συνολική αξία για κάθε εγγραφή στη συλλογή στη λεπτότερη κλιμάκωση. Προκειμένου να εφαρμοστεί αυτό σε όλες τις εγγραφές της συλλογής, χρησιμοποιείται ένα χωροχρονικό παράθυρο ολίσθησης.



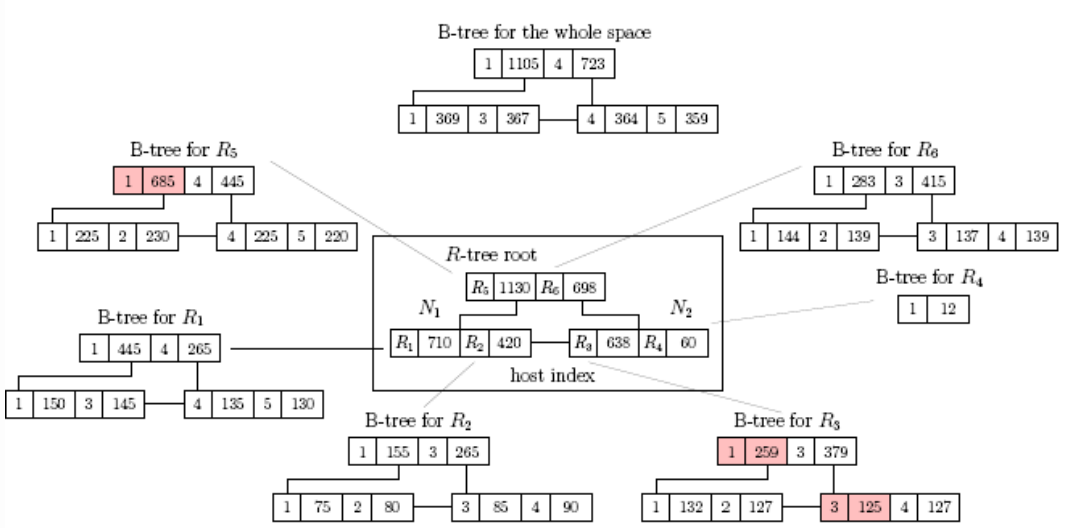
Εικόνα 3-27: (a) Περιοχές ενδιαφέροντας, (b) Ένα παράδειγμα κύβου δεδομένων.

Εκτός από τις εννοιολογικές πτυχές της χωροχρονικής συνάθροισης, ένα άλλο σημαντικό ζήτημα θεωρείται η ανάπτυξη μεθόδων για τον αποδοτικό υπολογισμό αυτού του είδους τελεστών προκειμένου να διαχειριστεί ο μεγάλος όγκος των χωροχρονικών δεδομένων. Συγκεκριμένα, έχουν αναπτυχθεί τεχνικές βασισμένες στη συνδυασμένη χρήση των εξειδικευμένων ευρετηρίων, του προϋπολογισμού των μέτρων συσσώρευσης και αλγορίθμων υπολογιστικής γεωμετρίας, για να υποστηρίξουν ειδικά τη συνάθροιση των δυναμικά υπολογισμένων συνόλων χωρικών αντικειμένων

[PTK+02], [TP05], [RZY+03], [ZT05]. Ενδεικτικά, οι Papadias κ.α. [PTK+02], [TP05] προτείνουν μια προσέγγιση βασισμένη σε δύο τύπους ευρετηρίων: ένα *ευρετήριο-οικοδεσπότης* (host index), που διαχειρίζεται τις χωρικές περιοχές και συσχετίζει με αυτές τις περιοχές συγκεντρωτικές πληροφορίες για όλες τις χρονικές στιγμές, και μερικά *ευρετήρια μέτρων* (ένας για κάθε είσοδο του ευρετηρίου οικοδεσπότη), τα οποία είναι συγκεντρωτικές χρονικές δομές που αποθηκεύουν ιστορικές τιμές των μέτρων. Για ένα σύνολο στατικών περιοχών, οι συγγραφείς ορίζουν το *συναθροιστικό* (aggregate) *R-B-tree* (aRB-δέντρο), το οποίο υιοθετεί ένα R-tree με τις συνοψισμένες πληροφορίες ως ερευτήριο-οικοδεσπότη, και ένα B-tree που περιέχει τα χρονικά μεταβαλλόμενα συγκεντρωτικά στοιχεία, ως ευρετήριο μέτρου.

Για να παρουσιάσουμε αυτήν την έννοια, ας θεωρήσουμε τις περιοχές R_1, R_2, R_3 και R_4 στην Εικόνα 3-27(a) ας υποθέσουμε ότι ο αριθμός των τηλεφωνικών κλήσεων που ξεκίνησαν μέσα στο χρονικό διάστημα [T_1; T_5] μέσα σε αυτές τις περιοχές καταγράφεται ως ένα μέτρο του πίνακα συμβάντων που παρατίθεται στην Εικόνα 3-27(b).

Έτσι, η Εικόνα 3-28 δείχνει το αντίστοιχο aRB-tree. Αυτή η δομή είναι κατάλληλη για την αποδοτική επεξεργασία των *παραθύρων συναθροιστικής ολίσθησης* (window aggregate queries) δλδ. για τον υπολογισμό των συσσωρευμένων μέτρων των περιοχών που τέμνονται με ένα συγκεκριμένο παράθυρο. Στην πραγματικότητα, για τους κόμβους που περιέχονται συνολικά μέσα στο παράθυρο ερωτήματος, το συγκεντρωτικό μέτρο είναι ήδη διαθέσιμο αποφεύγοντας έτσι την «κάθοδο» σε αυτούς τους κόμβους. Συμπερασματικά, η συγκεντρωτική επεξεργασία πραγματοποιείται πιο γρήγορα.



Εικόνα 3-28: Το aRB-tree.

Για παράδειγμα, ας υπολογίσουμε τον αριθμό κλήσεων μέσα στη σκιασμένη περιοχή στην Εικόνα 3-27(a) κατά τη διάρκεια του χρονικού διαστήματος [T_1; T_3]. Δεδομένου ότι η R_5 συμπεριλαμβάνεται ολοκληρωτικά στο παράθυρο της ερώτησης δεν υπάρχει καμία ανάγκη να εξερευνήσουμε περαιτέρω τις R_1 και R_2 όταν κάποιος ανατρέχει στο B-tree για την R_5. Η πρώτη καταχώρηση της ρίζας αυτού του B-tree περιέχει το μέτρο για το διάστημα [T_1; T_3], που είναι η τιμή για την οποία ενδιαφερόμαστε. Αντ' αυτού, προκειμένου να βρεθεί ο αριθμός των κλήσεων στο διάστημα [T_1; T_3] στην R_3 θα πρέπει να

υπάρξει τόσο η είσοδος της ρίζας του B-tree στην R_3 όσο και ένα φύλλο (οι χρωματισμένοι κόμβοι στην Εικόνα 3-28). Οι Tao κ.α. [TKC+04] έδειξε ότι το aRB-tree μπορεί να υποφέρει από το πρόβλημα μοναδικής προσμέτρησης, δηλ., εάν ένα αντικείμενο παραμένει στην περιοχή ερώτησης για διάφορες χρονικές στιγμές κατά τη διάρκεια του διαστήματος ερώτησης, θα έχει ως αποτέλεσμα την καταμέτρηση του πολλαπλές φορές. Για να αντιμετωπίσουν αυτό το πρόβλημα, οι συγγραφείς προτείνουν μια προσέγγιση που συνδυάζει τα χωροχρονικά ευρετήρια με τα σκίτσα (sketches), μια παραδοσιακή προσεγγιστική τεχνική μέτρησης βασισμένη στον πιθανοτικό υπολογισμό [FM85]. Η δομή των ευρετηρίων είναι παρόμοια με αυτή του aRB-tree: ένα R-tree δεικτοδοτεί τις περιοχές ενδιαφέροντος, ενώ τα B-trees καταγράφουν τα ιστορικά σκίτσα της αντίστοιχης περιοχής. Εντούτοις, αυτός ο δείκτης διαφέρει από τα aRB-trees στους αλγορίθμους ερωτημάτων δεδομένου ότι κάποιος μπορεί να εκμεταλλευτεί την ιδιότητα περικοπής (pruning) που προσφέρουν τα σκίτσα για να καθορίσει κάποιο ευριστικό τρόπο προκειμένου να μειώσει το χρόνο ερώτησης.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η ανάλυση των Gray κ.α. [GCB+97] σχετικά με τις συναρτήσεις συσσώρευσης. Πιο συγκεκριμένα, οι συντάκτες κατηγοριοποιούν τις συναρτήσεις συνάθροισης σε τρεις κατηγορίες:

- *διανεμητικές*, των οποίων οι τιμές μπορούν να υπολογιστούν από τις τιμές του αμέσως κατώτερου επιπέδου ιεραρχίας,
- *αλγεβρικές*, των οποίων οι τιμές μπορούν να υπολογιστούν από ένα σύνολο συσσωρεύσεων από το αμέσως κατώτερο επίπεδο ιεραρχίας,
- *ολιστικές*, που χρειάζονται τα δεδομένα της βάσης του κύβου για να υπολογίσουν αποτελέσματα στα διάφορα επίπεδα των διαστάσεων.

Τέλος, αξίζει να αναφέρουμε την εργασία των Shekhar κ.α. [SLC+01], όπου οι συγγραφείς προτείνουν ένα μοντέλο αποθηκών δεδομένων κυκλοφορίας για τη μητροπολιτική περιοχή Twin-Cities. Αν και το χτίσιμο μιας ΑΔ για την κυκλοφοριακή διαχείριση, είναι ευκολότερο από το χτίσιμο μιας ΑΔ για τροχιές (θυμίζοντας ότι η κύρια δυσκολία είναι ότι οι τροχιές μπορούν να επεκταθούν σε περισσότερα από ένα κελιά), σε αυτήν την εργασία αναλύονται διάφορα ενδιαφέροντα ζητήματα. Επίσης, για κάθε κατηγορία συναρτήσεων από την [GCB+97], οι συγγραφείς παρουσιάζουν αντιπροσωπευτικούς τελεστές συσσώρευσης για την περιοχή των ΓΠΣ (GIS) (Πίνακας 3-1), που στάθηκαν χρήσιμοι και στην έρευνα μας.

Πίνακας 3-1: Τελεστές συσσώρευσης (από [SLC+01])

	Aggregation Function		
Data Type	**Distributive Function**	**Algebraic Function**	**Holistic Function**
Set of numbers	Count, Min, Max, Sum	Average, MaxN, MinN, Standard, Deviation	Median, Rank, MostFrequent
Set of points, lines, polygons	Minimal Orthogonal Bounding Box, Geometric Union, Geometric Intersection	Centroid, Center of mass, Center of gravity	Equi-partition, Nearest neighbor index

3.5.2.2. Αποθήκες Δεδομένων Τροχιών Κινούμενων Αντικειμένων

Το κίνητρο εδώ είναι να μετασχηματιστούν οι πρωτογενείς τροχιές σε πολύτιμες πληροφορίες που μπορούν να αξιοποιηθούν για υποστήριξη αποφάσεων σε εφαρμογές με ενδιαφέρον για την κίνηση, όπως το κινητό μάρκετινγκ, τις υπηρεσίες θέσης στο τη διαχείριση ελέγχου της κυκλοφορίας. Οι αποθήκες δεδομένων για τροχιές [PRD+08] είναι ακόμη στα σπάργανα αλλά μπορούμε να διακρίνουμε τρεις σημαντικές ερευνητικές κατευθύνσεις σε αυτόν τον τομέα: μοντελοποίηση, συνάθροιση και δεικτοδότηση.

Από την προοπτική της μοντελοποίησης, ο ορισμός των ιεραρχιών στη χωρική διάσταση εισάγει ζητήματα που πρέπει να αντιμετωπιστούν. Η χωρική διάσταση μπορεί να περιλαμβάνει μη σαφώς ορισμένες ιεραρχίες [JKP+04] που επιτρέπουν πολλαπλές διαδρομές συνάθροισης οι οποίες θα πρέπει να ληφθούν υπόψη κατά τη διάρκεια των OLAP λειτουργιών. Οι Tao και Papadias [TP05] προτείνουν την ολοκλήρωση των χωρικών και χρονικών διαστάσεων και παρουσιάζουν κατάλληλες δομές δεδομένων που ενοποιούν τη χωροχρονική δεικτοδότηση και την προ-συνάθροιση. Οι Choi κ.α. [CKL06] προσπαθούν να υπερνικήσουν τους περιορισμούς των πολυ-δεντρικών δομών με την εισαγωγή μιας νέας δομής ευρετηρίου που συνδυάζει τα πλεονεκτήματα των τετραδικών δένδρων (Quadtrees) και των αρχείων πλέγματος (Grid files). Εντούτοις, τα ανωτέρω πλαίσια εστιάζουν στον υπολογισμό απλών μέτρων (π.χ. πλήθος πελατών).

Πρόσφατα, μια προσπάθεια να διαμορφωθεί και να συντηρηθεί μια TDW παρουσιάζεται στην [OOR+07] όπου ορίζεται ένας απλός κύβος δεδομένων που αποτελείται από τις χωρικές/χρονικές διαστάσεις και τα αριθμητικά μέτρα σχετικά με τις τροχιές. Οι συγγραφείς εστιάζουν στη φόρτωση και τον υπολογισμό του μέτρου presence, που το καθορίζουν ως το πλήθος των μοναδικών τροχιών που βρίσκονται σε ένα κελί. Μια επέκταση της [OOR+07] μπορεί να βρεθεί στην [LOR+09] που συζητά ζητήματα αποθήκευσης και συνάθροισης για τα συχνά χωροχρονικά πρότυπα που εξάγονται από τις τροχιές των κινούμενων αντικειμένων και λαμβάνουν χώρα σε μια συγκεκριμένη χωρική ζώνη και κατά τη διάρκεια ενός δεδομένου χρονικού διαστήματος.

Από όσο γνωρίζουμε, οι τεχνικές προϋπολογισμού δεν έχουν μελετηθεί στα πλαίσια μιας TDW. Εντούτοις, υπάρχουν πολλές εργασίες που συζητούν την ιδέα της εκμετάλλευσης των όψεων για να επιταχύνουν τους υπολογισμούς στις παραδοσιακές αποθήκες δεδομένων. Μια επισκόπηση αυτών των τεχνικών μπορεί να βρεθεί στη [Kot02].

3.6. Σύνοψη

Σε αυτήν την ενότητα, μελετήσαμε την εφαρμογή των τεχνικών αποθήκευσης δεδομένων σε δεδομένα τροχιών. Παρουσιάσαμε αναλυτικά δυο πλαίσια για TDW: ένα που επιτρέπει τη μοναδική και στατική ερμηνεία της έννοιας της τροχιάς και ένα για ad-hoc TDW που επιτρέπει πολλαπλούς σημασιολογικούς ορισμούς αναφορικά με την τροχιά.

Σε ότι αφορά την πρώτη προσέγγιση, παρουσιάσαμε εναλλακτικές ETL διαδικασίες για την τροφοδότηση της TDW με συγκεντρωτικά δεδομένα τροχιών: μια (βασισμένη σε ευρετήριο) *προσανατολισμένη στα κελιά* και μια (μη βασισμένη σε ευρετήριο) *προσανατολισμένη στις τροχιές*. Επιπλέον, παρείχαμε μια προσεγγιστική λύση για την επίλυση του λεγόμενου *προβλήματος της*

μοναδικής προσμέτρησης (distinct count problem) που παρουσιάζεται κατά τη διάρκεια των OLAP λειτουργιών.

Σε ότι αφορά τη δεύτερη προσέγγιση, επεκτείναμε το OLAP μοντέλο δεδομένων ώστε να περιλαμβάνει ένα πίνακα συμβάντων που μπορεί να απαντήσει σε ερωτήματα λαμβάνοντας υπόψη διαφορετικούς σημασιολογικούς ορισμούς των τροχιών και που υποστηρίζει την επιλογή της σημασιολογίας στα ερωτήματα συσσώρευσης των δεδομένων τροχιών. Επιπλέον, εμπλουτίσαμε τις OLAP τεχνικές παρουσιάζοντας έναν αποδοτικό αλγόριθμο που αξιοποιεί το OLAP μοντέλο προκειμένου να απαντήσει σε OLAP ερωτήματα. Τέλος, συζητήσαμε θέματα προϋπολογισμού των αποτελεσμάτων στην περίπτωση που είναι γνωστοί κάποιοι σημασιολογικοί ορισμοί των τροχιών.

4. Εξόρυξη Γνώσης Βασισμένη σε Τροχιές

Στο προηγούμενο κεφάλαιο, παρουσιάσαμε δυο μοντέλα για ΑΔ και λειτουργίες OLAP που είναι κατάλληλα για δεδομένα τροχιών. Όπως όμως συμβαίνει και στο χώρο της διαχείρισης παραδοσιακών δεδομένων, το επόμενο βήμα μετά την υλοποίηση της ΑΔ είναι η εξόρυξη γνώσης που μπορεί να εφαρμοστεί είτε στην ΑΔ είτε και κατευθείαν στην βάση δεδομένων. Με τέτοια θέματα θα ασχοληθούμε σε αυτό το κεφάλαιο, αναπτύσσοντας τεχνικές εξόρυξης γνώσης που χρειάζονται, στη μια περίπτωση, τα λεπτομερή δεδομένα που αποθηκεύονται στη MOD αλλά και τεχνικές που μπορούν να εφαρμοστούν στα συγκεντρωτικά δεδομένα που διατηρούνται στη TDW. Η δομή του κεφαλαίου περιλαμβάνει τα εξής: Η Ενότητα 4.1 εισάγει έννοιες που αφορούν την εξαγωγή προτύπων (pattern mining) από δεδομένα κινούμενων αντικειμένων ενώ, η Ενότητα 4.2 παρουσιάζει τα κίνητρα της έρευνας μας. Στην Ενότητα 4.3 παρουσιάζεται ένα προτεινόμενο πλαίσιο για την ανακάλυψη προτύπων αλληλεπίδρασης που μπορούν να χρησιμοποιηθούν για σκοπούς αναπαράστασης, σύνθεσης και κατηγοριοποίησης και στην Ενότητα 4.4 συζητιούνται δύο προσεγγίσεις για εξόρυξη κυκλοφοριακών προτύπων. Στο τέλος, στην Ενότητα 4.5 καταγράφεται η σχετική έρευνα που έχει γίνει στη περιοχή και στην Ενότητα 4.6 συνοψίζονται τα συμπεράσματα του κεφαλαίου.

4.1. Εισαγωγή

Διανύουμε τον αιώνα της πληροφορίας και συνεπώς το ενδιαφέρον μας είναι επικεντρωμένο στη συλλογή και ανάλυση δεδομένων, εξαγωγή πληροφοριών και γνώσης από αυτά. Εταιρίες, οργανισμοί, επιστήμονες, ακόμα και μεμονωμένα άτομα δεν ενδιαφέρονται μόνο για τη συλλογή δεδομένων αλλά επίσης και για την επεξήγηση της έννοιας και της σημασίας τους για να μπορέσουν να τα χρησιμοποιήσουν για τη λήψη καλύτερων αποφάσεων, την επίλυση προβλημάτων και γενικά για τη καλύτερη κατανόηση των φαινομένων που αφορούν το χώρο τους.

Η εξόρυξη γνώσης αποτελεί μέρος της διαδικασίας KDD και συμπεριλαμβάνει την εφαρμογή αλγορίθμων για την ανακάλυψη χρήσιμων προτύπων από πρωτογενή δεδομένα. Σχετική έρευνα διεξάγεται για περισσότερο από είκοσι έτη, η οποία συνδυάζει τεχνικές από τρεις βασικούς επιστημονικούς τομείς. Στην κλασσική στατιστική, περί τα 1960, ο όρος εξόρυξη γνώσης εισάγεται για πρώτη φορά. Η κλασσική στατιστική περιλαμβάνει έννοιες όπως η γραμμική παλινδρόμηση, η κανονική κατανομή, η τυπική απόκλιση, η διακύμανση, η διαχωριστή ανάλυση, η ανάλυση συστάδων και τα διαστήματα εμπιστοσύνης. Όλα αυτά είναι τεχνικές που χρησιμοποιούνται για τη διερεύνηση δεδομένων και των μεταξύ τους συσχετίσεων και γι' αυτό το λόγο θεωρούνται πολύ σημαντικά ακόμη και σήμερα. Ο δεύτερος επιστημονικός τομέας αφορά την τεχνητή νοημοσύνη (artificial intelligence -

AI), η οποία βασίζεται σε heuristics και, σε αντίθεση με τη στατιστική, προσπαθεί να εφαρμόσει μεθόδους επεξεργασίας στατιστικών προβλημάτων που λαμβάνουν περισσότερο υπόψη τον παράγοντα άνθρωπο. Ο τρίτος επιστημονικός τομέας αφορά τη μηχανική μάθηση (machine learning) που θεωρείται η συνένωση των δύο προηγούμενων επιστημών (στατιστικής και τεχνητής νοημοσύνης). Η μηχανική μάθηση εισήγαγε την ιδέα για την ανάπτυξη «έξυπνων» υπολογιστικών συστημάτων που να κατανοούν τα δεδομένα που διαχειρίζονται.

Η εξόρυξη γνώσης μπορεί να κατηγοριοποιηθεί περαιτέρω σε τεχνικές συσταδοποίησης (clustering), κανόνων συσχέτισης (association rule) και σε τεχνικές κατηγοριοποίησης (classification). Η πρώτη κατηγορία [KR90], [JMF99] είναι η μη επιτηρουμένη διαδικασία συσταδοποίησης αντικειμένων σε κλάσεις, που επίσης ονομάζονται συστάδες, βάση ενός μεγέθους ομοιότητας. Συνεπώς, ερευνάται η συμπεριφορά των ομάδων παρά των μεμονωμένων καταγραφών. Η κατάτμηση μιας βάσης δεδομένων σε τμήματα δίνει μια ολοκληρωμένη εικόνα της βάσης και βοηθά στην καλύτερη κατανόηση των δεδομένων.

Η εξόρυξη σχεσιακών κανόνων έχει ως στόχο την ανακάλυψη συσχετίσεων μεταξύ των χαρακτηριστικών μιας βάσης δεδομένων [AIS93]. Οι κανόνες συσχέτισης είναι της μορφής $A \Rightarrow B$ [*s*, *c*], $A \subset J$, $B \subset J$ όπου *A*, *B* και *J* είναι ομάδες πραγμάτων (π.χ. χαρακτηριστικά), όπου χαρακτηρίζονται από δύο μεγέθη: υποστήριξη (*support* (*s*)) και εμπιστοσύνη (*confidence* (*c*)). Η υποστήριξη ενός κανόνα $A \Rightarrow B$ εκφράζει την πιθανότητα να περιέχει ένα συμβάν μαζί τα *A* και *B*, ενώ η εμπιστοσύνη του κανόνα εκφράζει την υπό όρους πιθανότητα ένα συμβάν που περιέχει το *A* να να περιέχει επίσης και το *B*.

Η κατηγοριοποίηση αποτελεί μια από τις πιο επιτηρούμενες τεχνικές μάθησης. Αντικείμενο της κατηγοριοποίησης είναι σε πρώτη φάση η ανάλυση ενός συνόλου εκπαίδευσης (ήδη χαρακτηρισμένα-κατηγοριοποιημένα και, μέσω αυτής της διαδικασίας να φτιάξει ένα μοντέλο για χαρακτηρισμό των νέων δεδομένων που εισέρχονται [HK00]. Συγκεκριμένα, στο πρώτο βήμα δημιουργείται ένα μοντέλο κατηγοριοποίησης χρησιμοποιώντας ένα *σύνολο δεδομένων εκπαίδευσης* (training data set) αποτελούμενο από εγγραφές που είναι γνωστό ότι ανήκουν σε μια συγκεκριμένη κλάση και μια κατάλληλη μέθοδο επιτηρούμενης εκμάθησης, π.χ. δέντρα αποφάσεων ή νευρωνικά δίκτυα. Στη περίπτωση των δέντρων αποφάσεων, για παράδειγμα, το μοντέλο αποτελείται από ένα δέντρο με «αν» καταστάσεις που οδηγούν σε μια ετικέτα που υποδηλώνει την κατηγορία που ανήκει η εγγραφή. Σε δεύτερη φάση, το μοντέλο που δημιουργήθηκε χρησιμοποιείται για την κατηγοριοποίηση των εγγραφών που δεν περιλαμβάνονταν στο σύνολο εκπαίδευσης. Πολλές μέθοδοι έχουν αναπτυχθεί για την κατηγοριοποίηση, περιλαμβάνοντας την επαγωγική μέθοδο για τα δέντρα απόφασης, τα νευρωνικά και Bayesian δίκτυα [FPS+96].

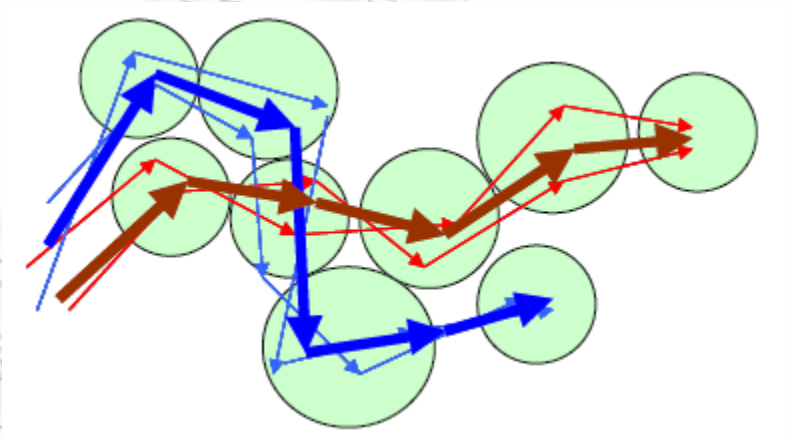
Η εφαρμογή των παραπάνω τεχνικών σε συμβατικά δεδομένα έχει μελετηθεί εκτενώς τις τελευταίες δεκαετίες. Στις μέρες μας, ο μεγάλος όγκος των δεδομένων κινούμενων αντικειμένων που συλλέγονται αποτελεί σπουδαία πρόκληση για την εφαρμογή κατάλληλων μεθόδων εξόρυξης γνώσης σε αυτά. Οι παραδοσιακές τεχνικές πρέπει να εξελιχθούν ώστε να λαμβάνουν υπόψη την πολύπλοκη φύση των χωροχρονικών δεδομένων και να είναι δυνατή η διαχείριση τους με αποτελεσματικό τρόπο. Πρόσφατα, η γρήγορη ανάπτυξη και εκτενή εξάπλωση τεχνολογιών με δυνατότητα χωρικής

τοποθέτησης οδήγησε στην ανάπτυξη ενός νέου πεδίου, της *εξόρυξης κίνησης (mobility mining)* [GP07], που στοχεύει στη συλλογή και ανάλυση δεδομένων κινούμενων αντικειμένων για το σχεδιασμό και κατανόηση των συμπεριφορών μεγάλων πληθυσμών κινούμενων αντικειμένων, όπως οχήματα σε εφαρμογές διαχείρισης κυκλοφορίας ή ατόμων σε υπηρεσίες βασιζόμενες σε στοιχεία θέσης/τοποθεσίας για κινητά τηλέφωνα.

4.2. Κίνητρο

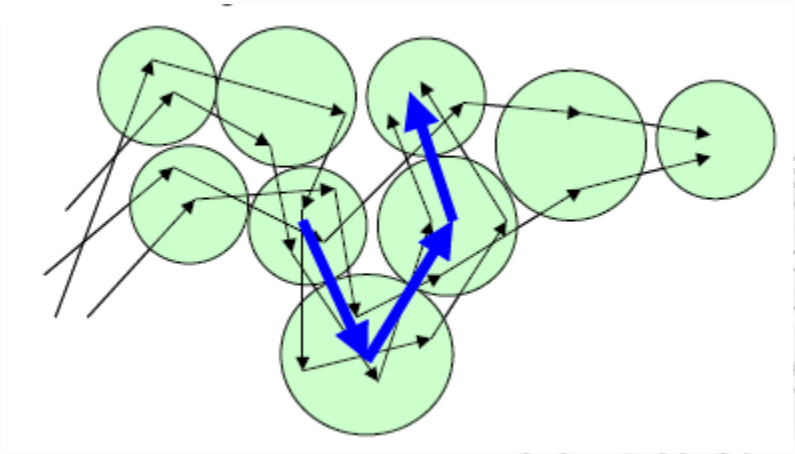
Κίνητρο μας για εξόρυξη βασισμένη σε δεδομένα τροχιών είναι η ανάπτυξη μεθόδων που μπορούν να εξάγουν χρήσιμα πρότυπα από δεδομένα κινούμενων αντικειμένων. Γενικά μιλώντας, η εξόρυξη γνώσης από δεδομένα κινούμενων αντικειμένων είναι σε πρώιμα στάδια, και ακόμα οι περισσότερες βασικές ερωτήσεις στο χώρο αυτό παραμένουν αναπάντητες: τι είδη προτύπων μπορούν να εξαχθούν από τροχιές; Ποιες μέθοδοι και αλγόριθμοι μπορούν να εφαρμοστούν για την εξαγωγή τους; Πώς αυτά τα πρότυπα μπορούν αποτελεσματικά να βελτιώσουν την κατανόηση του πεδίου εφαρμογής και να προσφέρουν καλύτερες υπηρεσίες; Τα παρακάτω απλά παραδείγματα δίνουν μια σύντομη ματιά στη μεγάλη ποικιλία προτύπων και πιθανών εφαρμογών που αναμένεται να εξυπηρετήσουν:

- *Συσταδοποίηση (Clustering),* η ανακάλυψη ομάδων με «παρόμοιες» τροχιές, μαζί με περίληψη για τη κάθε ομάδα. Γνωρίζοντας ποιες είναι οι κύριες διαδρομές (αναπαριστάμενες ως συστάδες) που ακολουθούν τα άτομα κατά τη διάρκεια της ημέρας μπορεί να αποτελέσει σημαντική πληροφορία για τη βελτίωση πολλών διαφορετικών υπηρεσιών για τους πολίτες. Για παράδειγμα, συστάδες τροχιών μπορούν να προβάλλουν την παρουσία σημαντικών δρόμων που δεν καλύπτονται αποτελεσματικά από την υπηρεσία της δημοτικής συγκοινωνίας.



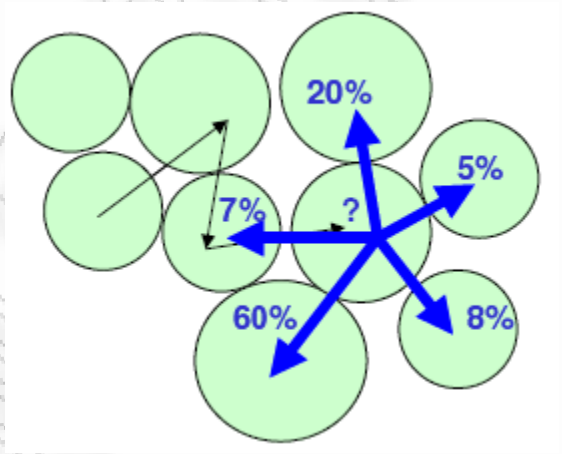
Εικόνα 4-1: Ένα παράδειγμα συσταδοποίησης τροχιών.

- *Συχνά πρότυπα (frequent patterns),* η ανακάλυψη των συχνά ακολουθούμενων υπο-διαδρομών. Τέτοιες πληροφορίες είναι χρήσιμες για την αστική ανάπτυξη π.χ. ο εντοπισμός μη επαρκών συχνά ακολουθούμενων διαδρομών οχημάτων που είναι αποτέλεσμα λανθασμένου σχεδιασμού των δρόμων.



Εικόνα 4-2: Ένα παράδειγμα συχνών προτύπων τροχιών.

- *Κατηγοριοποίηση (Classification),* η ανακάλυψη κανόνων συμπεριφοράς, στοχεύοντας στην επεξήγηση της συμπεριφοράς των τρεχόντων χρηστών και προβλέποντας τον αριθμό των μελλοντικών. Η εξομοίωση της αστικής κυκλοφορίας αποτελεί ένα απλό παράδειγμα εφαρμογής για αυτού του είδους τη γνώση, αφού ένα μοντέλο κατηγοριοποίησης μπορεί να αναπαραστήσει μια εξελιγμένη εναλλακτική των ειδικών (ad-hoc) κανόνων συμπεριφοράς, που παρέχονται από εμπειρογνώμονες του χώρου, πάνω στους οποίους βασίζονται οι εξομοιωτές.



Εικόνα 4-3: Ένα παράδειγμα κατηγοριοποίησης τροχιών.

Τα χωροχρονικά δεδομένα εισάγουν νέες δυνατότητες και αντίστοιχα, καινοτόμους τρόπους εκτέλεσης των παραπάνω εργασιών. Η συσταδοποίηση τροχιών κινούμενων αντικειμένων, για παράδειγμα, απαιτεί την εύρεση τόσο του κατάλληλου επιπέδου κλιμάκωσης (ισχυρά εξαρτώμενο από την εφαρμογή) όσο και μια χρονικής υποδομής (π.χ., οι ώρες αιχμής μπορούν να δίνουν πληροφορίες για τον καθορισμό μιας δομής συσταδοποιημένων δεδομένων κυκλοφορίας, ενώ άλλες χρονικές περίοδοι μπορεί απλά να προσθέτουν θόρυβο στη διαδικασία συσταδοποίησης).

Όποιο είδος προτύπου και αν εξάγεται, πρέπει να δίνεται ιδιαίτερη προσοχή στον κίνδυνο παραβίασης της ιδιωτικότητας. Μέχρι τώρα, η υπάρχουσα έρευνα στο χώρο προστασίας της ιδιωτικότητας για τις τεχνικές εξόρυξης γνώσης είναι ουσιαστικά περιορισμένη σε δύο κατευθύνσεις: η πρώτη στοχεύει στην εξαγωγή προτύπων/μοντέλων με μέλημα τη μη διάδοση πρωτογενών, ευαίσθητων δεδομένων, συχνά αποκτώμενων από ανακατάταξη δεδομένων ή υπολογιστικές πολυμερών τεχνικών, η δεύτερη πραγματεύεται δεδομένα που γνωστοποιούνται μετά από χειρισμό *εκκαθάρισης* (sanitation) που στοχεύει στην απόκρυψη ενός προκαθορισμένου συνόλου ευαίσθητων προτύπων. Η πρώτη προσέγγιση είναι μη επαρκής για την περιοχή που μας ενδιαφέρει, αφού χωρίς καμία διαδικασία ελέγχου πάνω στη δημιουργία προτύπου υπάρχει πάντα ο κίνδυνος εξαγωγής προτύπων που παραβιάζουν σε κάποιο βαθμό την ιδιωτικότητα ενός ατόμου ή μιας κοινότητας. Η δεύτερη προσέγγιση αντιμετωπίζει το πρόβλημα με πιο άμεσο τρόπο.

Ωστόσο, μπορεί να εφαρμοστεί μόνο σε απλοποιημένο πλαίσιο, και αυτό είναι ξεκάθαρα μη ικανοποιητικό για τους σκοπούς μας: γενικά, δεν είναι εφικτό (ή καθόλου πιθανό) να καθοριστεί προκαταβολικά ποια πρότυπα είναι ευαίσθητα. Επιπλέον, αυτή η προσέγγιση φαίνεται να μην είναι επεκτάσιμη σε άλλες τεχνικές εξόρυξης γνώσης που δε βασίζονται σε πρότυπα, όπως η συσταδοποίηση και η κατηγοριοποίηση. Ως παράδειγμα θεμάτων ιδιωτικότητας στην τελευταία κατηγορία εργασιών: αν έχουμε ως αποτέλεσμα μια εξαιρετικά ομογενοποιημένη συστάδα, γνωρίζοντας την «περίληψή» της (πχ. αντιπροσωπευτική τροχιά) μπορεί να διαθέτουμε επαρκή στοιχεία για την ανακατασκευή πληροφοριών σχετικά με τα άτομα που περιέχει (π.χ. οι τροχιές τους θα συμπίπτουν σχεδόν πλήρως με αυτήν της αντιπροσωπευτικής τροχιάς της συστάδας). Με τον ίδιο τρόπο, ένας κατηγοριοποιητής μπορεί να περιέχει κανόνες που παράγονται στη βάση ενός μόνου ατόμου ή πολύ λίγων ατόμων αποκαλύπτοντας συνεπώς κάποιες (πιθανά ευαίσθητες) πληροφορίες των χαρακτηριστικών τους.

Για να αντιμετωπίσουμε αποτελεσματικά το θέμα της προστασίας των προσωπικών δεδομένων, τα πρότυπα που δημιουργούνται πρέπει να χαρακτηρίζονται ανάλογα με το αν σέβονται την ιδιωτικότητα (privacy-preserving) ή μη, με την έννοια ότι θα είναι διαθέσιμο ένα σύνολο περιορισμών για τα χωροχρονικά δεδομένα και πρότυπα ώστε να εκφραστούν οι προϋποθέσεις ιδιωτικότητας – π.χ. πρότυπα που δεν αποκαλύπτουν καμία ευαίσθητη πληροφορία σχετικά με τις τροχιές των ατόμων από όπου προέρχονται.

4.2.1. Η συνεισφορά μας

Αναπτύξαμε δύο τεχνικές εξόρυξης γνώσης από δεδομένα τροχιών που είναι ικανές να εξάγουν σχεσιακά πρότυπα από δεδομένα κινούμενων αντικειμένων. Οι λειτουργίες τους συνοψίζονται παρακάτω, όπως περιγράφηκαν στα [NMO09], [NMM08], [MT09b]:

- Η αυξανόμενη διάδοση των τεχνολογιών με δυνατότητα εντοπισμού θέσης (GPS, GSM δίκτυα κτλ.) συνεισφέρει στη συλλογή μεγάλων συνόλων χωροχρονικών δεδομένων και προσφέρει την ευκαιρία για ανακάλυψη χρήσιμης γνώσης σχετικά με τη συμπεριφορά κίνησής τους, το οποίο ενισχύει την ανάπτυξη καινοτόμων εφαρμογών και υπηρεσιών. Στα πλαίσια αυτής της έρευνας, προχωρήσαμε προς αυτή τη κατεύθυνση και επεκτείναμε το επιτηρούμενο παράδειγμα εκμάθησης (supervised learning paradigm) που αναλύει τις

τροχιές κινούμενων αντικειμένων. Εισάγαμε *πρότυπα αλληλεπίδρασης* (interaction patterns) τροχιών ως περιεκτικούς περιγραφείς περιοχών, αναφορικά με διαστήματα (π.χ. διαστήματα περιοχών που προσπελάστηκαν κατά τη διάρκεια κίνησης) και σχέσεις ομοιότητας μεταξύ τροχιών, για την εξαγωγή σημασιολογίας από αυτά. Καταθέσαμε μια ολοκληρωμένη αρχιτεκτονική σχετικά με τη νέα προσέγγιση της εξόρυξης γνώσης και μετά μελετήσαμε διάφορες περιπτώσεις με διαφορετική πολυπλοκότητα.

- Η ροή των δεδομένων που λαμβάνονται από τις σύγχρονες συσκευές αισθητήρων δίνει τη δυνατότητα για ανάπτυξη καινοτόμων ερευνητικών τεχνικών σχετικά με τη διαχείριση των δεδομένων και την εξόρυξη γνώσης. Σε αυτή την έρευνα, ασχοληθήκαμε με την ανάλυση του κυκλοφοριακού προβλήματος σε ένα οδικό δίκτυο έτσι ώστε να προσφέρουμε βοήθεια στις Αρχές της πόλης να βελτιώσουν τη ροή της κυκλοφορίας. Χρησιμοποιούμε ένα γράφο με πληροφορίες για τη ροή της κίνησης στο οδικό δίκτυο και τον αξιοποιούμε με στόχο την ανακάλυψη σχέσεων κυκλοφορίας (traffic relationships) όπως μετάδοση (propagation), διάσπαση (split) και συγχώνευση (merge) της κυκλοφορίας στα διάφορα οδικά τμήματα.

Στον παραδοσιακό κόσμο της διαχείρισης δεδομένων, οι εργασίες εξόρυξης γνώσης μπορούν να εφαρμοστούν είτε σε βάσεις δεδομένων είτε σε αποθήκες δεδομένων. Το ίδιο ισχύει εδώ και για τις δύο μας τεχνικές εξόρυξης γνώσης από τροχιές. Για την πρώτη, είναι απαραίτητες οι λεπτομέρειες των πρωτογενών δεδομένων οπότε εφαρμόζεται στην MOD, ενώ η δεύτερη που χρησιμοποιεί συγκεντρωτικά δεδομένα μπορεί να εφαρμοστεί στην TDW.

4.3. Εξόρυξη Προτύπων Αλληλεπίδρασης για Χωροχρονική Αναπαράσταση, Σύνθεση και Κατηγοριοποίηση

Στις ακόλουθες παραγράφους, περιγράφουμε την έρευνά μας που παρουσιάζεται στην [NMO09] και που εισάγει δύο βασικές ιδέες: (ι) προκειμένου να κατανοήσουμε τι συμβαίνει σε ένα κινούμενο αντικείμενο απαιτείται να εξετάσουμε όχι μόνο στην τροχιά του, αλλά και το γενικότερο πλαίσιο στο οποίο κινείται, (ii) αυτό το πλαίσιο καθορίζεται όχι μόνο το γεωγραφικό χώρο, αλλά και από την παρουσία άλλων αντικειμένων και την αλληλεπίδραση που υπάρχει με το κινούμενο αντικείμενο που μας ενδιαφέρει να αναλύσουμε.

Προτείνουμε ένα πλαίσιο για την εξαγωγή χρήσιμων γνωρισμάτων που μπορούν να χρησιμοποιηθούν για να αναπαραστήσουν της κίνησης ενός αντικειμένου που μας παρέχει γνώση σχετικά με τα χαρακτηριστικά αυτής την κίνηση. Επιπλέον, μπορούμε να αξιοποιήσουμε το σύνολο αυτών των γνωρισμάτων εφαρμόζοντας μοντέλα κατηγοριοποίησης που θα μας βοηθήσουν να ανακαλύψουμε τις πιθανές σχέσεις μεταξύ των χαρακτηριστικών της μετακίνησης και των συμπεριφορών.

4.3.1. Ορισμός Προβλήματος

Το πρόβλημα που αντιμετωπίζεται σε αυτήν την ενότητα είναι μια εφαρμογή του κλασσικού προβλήματος κατηγοριοποίησης που εξειδικεύεται όμως σε δεδομένα τροχιών. Στη συνέχεια, παρέχουμε τους βασικούς ορισμούς για το μοντέλο δεδομένων τροχιάς που υιοθετούμε σε αυτήν την

έρευνα και για το συγκεκριμένο πρόβλημα κατηγοριοποίησης που μελετάμε. Το πρόβλημα κατηγοριοποίησης μπορεί να οριστεί γενικά ως:

Ορισμός 4-1 (Κατηγοριοποίηση Τροχιών): Δεδομένου ενός συνόλου ετικετών L και ενός συνόλου δεδομένων D με n τροχιές με ετικέτες, $D = <(T_1; l_1),\ldots,(T_n; l_n)>$, όπου $\forall\ 1 \leq i \leq n : l_i \in L$ και $T_i \in T$, το πρόβλημα της κατηγοριοποίησης τροχιών εξάγει μια συνάρτηση $C : T \rightarrow L$ όπου εντοπίζει όσο γίνεται με πιο ακρίβεια την ετικέτα που τέθηκε από το D. ■

Ένας τέτοιος γενικός ορισμός μπορεί να βρει εφαρμογή σε διάφορες συγκεκριμένες προσεγγίσεις, η πιο κοινή των οποίων είναι αυτή που βασίζεται σε γνωρίσματα αντικειμένων, με στόχο την εξαγωγή ενός συνόλου τιμών που περιγράφουν μερικές από τις ιδιότητές τους και μπορούν να αντιπροσωπευθούν εύκολα ως διάνυσμα σταθερού μεγέθους. Αυτό μας επιτρέπει να εφαρμόσουμε οποιαδήποτε γνωστή τεχνική ταξινόμησης που έχει αναπτυχθεί για σχεσιακά δεδομένα, και θα υιοθετηθεί σε αυτήν την εργασία. Άλλες εναλλακτικές λύσεις ίσως χρειάζονται το σχήμα κατηγοριοποίησης να προσαρμόζεται βάση του συγκεκριμένου τύπου δεδομένων, για παράδειγμα προσεγγίσεις βασισμένες σε μοντέλα, ή ένα γενικό σχήμα όπως οι κ-Κοντινότεροι Γείτονες, οι οποίοι μπορούν να εφαρμοστούν στο συγκεκριμένο τύπο δεδομένων με τον καθορισμό μιας συνάρτησης ομοιότητας μεταξύ των αντικειμένων (π.χ., μέτρα ομοιότητας μεταξύ των τροχιών).

Επομένως, σε αυτήν την ερευνητική προσπάθεια, θα συμβάλουμε στο κυρίως βήμα της κατηγοριοποίησης βάση γνωρισμάτων, δηλ., την εξαγωγή των γνωρισμάτων, η οποία καθορίζεται τυπικά κατωτέρω.

Ορισμός 4-2 (Εξαγωγή Γνωρισμάτων): Δεδομένου ενός συνόλου ετικετών L και ένα σύνολο δεδομένων D με n τροχιές με ετικέτες, $D = <(T_1; l_1) ,\ldots, (T_n; l_n)>$, όπου $\forall\ 1 \leq i \leq n : l_i \in L$ και $Ti \in T$, η εξαγωγή γνωρισμάτων ορίζει ένα θετικό ακέραιο n και μια συνάρτηση $F : T \rightarrow D_1\ x\ .\ x\ D_n$ που συσχετίζει κάθε τροχιά με ένα διάνυσμα n γνωρισμάτων κατάλληλου τύπου $(D_i, i = 1,\ldots, n)$. ■

4.3.2. Πρότυπα αλληλεπίδρασης

Τα αντικείμενα μπορούν γενικά να αναπαρασταθούν με τη διάρκεια ζωής τους συν το υποσύνολο (των μεταβαλλόμενων στο χρόνο) παρατηρούμενων ιδιοτήτων τους που είναι σχετικές με το πλαίσιο ανάλυσης. Για παράδειγμα, ένα κινούμενο αντικείμενο θα μπορούσε να αναπαρασταθεί ως ένα χρονικό διάστημα (στο οποίο παρακολουθείτο το αντικείμενο) συν τις μεταβαλλόμενες στο χρόνο χωρικές συντεταγμένες, την ταχύτητα και την επιτάχυνσή του.

Ορισμός 4-3 (Παρατηρούμενο Σύνολο): Κάθε αντικείμενο O αναπαριστάται ως ένα ζευγάρι που αποτελείται από ένα χρονικό διάστημα I και μια ακολουθία συναρτήσεων που συσχετίζει κάθε χρονική στιγμή με μια πραγματική τιμή:

$$F_O = (I, (f_O^1, \ldots, f_O^n)), \forall\ 1 \leq i \leq n\ f_O^i : I \rightarrow \mathbb{R}$$

Το F_O καλείται το παρατηρούμενο σετ του αντικειμένου O. ■

Ένα υποσύνολο των ιδιοτήτων που περιγράφουν ένα αντικείμενο καθορίζει συνήθως το χώρο (φυσικό ή εικονικό) όπου τα αντικείμενα υπάρχουν, εξελίσσονται και, εάν είναι αρκετά κοντινά, ενδεχομένως

αλληλεπιδρούν. Για παράδειγμα, οι χωρικές συντεταγμένες της κίνησης των αντικειμένων καθορίζουν πού βρίσκονται τα αντικείμενα, και τα πιο κοντινά αντικείμενα είναι πιθανότερο να αλληλεπιδρούν.

Ορισμός 4-4 (Χαρακτηριστικά Θέσης): Δεδομένου ενός παρατηρούμενου συνόλου F_O = *(I, S)* του αντικειμένου *O*, ορίζουμε τα χαρακτηριστικά θέσης του αντικειμένου *O* ως ένα υποσύνολο Loc_O του *S*: $Loc_O \subseteq S$ ■

Ομοίως, ένα υποσύνολο των ιδιοτήτων του αντικειμένου που περιγράφουν τις σχετικές δραστηριότητές του μπορεί να χρησιμοποιηθεί για να χαρακτηρίσει την πιθανή αλληλεπίδραση μεταξύ των αντικειμένων. Για παράδειγμα, μια σχετική ιδιότητα για ένα κινούμενο αντικείμενο μπορεί να είναι η ταχύτητα. Οι σχέσεις μεταξύ των τιμών ταχύτητας των διαφορετικών αντικειμένων που αλληλεπιδρούν μπορούν να μας βοηθήσουν να περιγράψουμε την αλληλεπίδραση.

Ορισμός 4-5 (Χαρακτηριστικά Αλληλεπίδρασης): Δεδομένου ενός παρατηρούμενου συνόλου F_O = *(I, S)* του αντικειμένου *O*, ορίζουμε τα χαρακτηριστικά αλληλεπίδρασης του *O* ως ένα υποσύνολο $Intc_O$ του *S*: $Int_O \subseteq S$ ■

Τα Χαρακτηριστικά Θέσης και *Αλληλεπίδρασης* είναι δυνατό να επικαλύπτονται, δηλ., στη γενική περίπτωση, $Loc_O \cap Int_O \neq \emptyset$.

Η αλληλεπίδραση μεταξύ των αντικειμένων είναι μια σχέση που υπονοεί μια κοινή τοποθεσία, επομένως θα πρέπει να οριστεί η έννοια της εγγύτητας - ή, με άλλα λόγια, της περιοχής επιρροής κάθε αντικειμένου. Μια τέτοια έννοια πρέπει να βασίζεται στα χαρακτηριστικά θέσης των αντικειμένων, δεδομένου ότι περιγράφουν τη θέση τους στο χώρο. Επιπλέον, τέτοιες σχέσεις έχουν συνήθως μια περιορισμένη διάρκεια και μπορούν να μεταβληθούν κατά τη διάρκεια του χρόνου επομένως πρέπει να μετρηθούν μέσα στα χρονικά παράθυρα περιορισμένου εύρους. Για παράδειγμα, για τα κινούμενα αντικείμενα πρέπει να οριστεί ένα σύνολο ενδιαφερόντων χωρικών περιοχών και να μετρηθεί η αλληλεπίδραση μεταξύ των αντικειμένων θα μπορούσε να μετρηθεί χωριστά για κάθε περιοχή κάθε ώρα.

Προκειμένου να χαρακτηριστεί η αλληλεπίδραση μεταξύ των κοντινών αντικειμένων, ψάχνουμε τις σχέσεις μεταξύ των χαρακτηριστικών αλληλεπίδρασης των αντικειμένων που αλληλεπιδρούν. Αυτό μπορεί να γίνει με πολλούς διαφορετικούς τρόπους, όπως η αναζήτηση στατιστικών συσχετίσεων ή προκαθορισμένων προτύπων. Σε αυτήν την εργασία, επιλέγουμε να συγκρίνουμε για κάθε χαρακτηριστικό τις τιμές όλων των αντικειμένων μέσα σε κάθε περιοχή και διάστημα χρόνου, κατόπιν δίνοντας περισσότερη έμφαση στα αντικείμενα που παρουσιάζουν σημαντική απόκλιση από τα άλλα. Για παράδειγμα, θα μπορούσαμε να συγκρίνουμε την ταχύτητα όλων των αντικειμένων μέσα σε μια περιοχή και ένα χρονικό διάστημα, και να εντοπίσουμε εκείνα των οποίων η ταχύτητα διαφέρει σημαντικά από τα υπόλοιπα.

Ορισμός 4-6 (Περιγραφείς Αλληλεπίδρασης): Δεδομένου ενός πλήθους αντικειμένων, έτσι ώστε το αντικείμενο *O* να έχει τα ίδια χαρακτηριστικά θέσης και αλληλεπίδρασης *Loc* και *Int* αντίστοιχα, ορίζουμε τον περιγραφέα θέσης για ένα αντικείμενο *O=(I, S)* στην περιοχή *r* και στο χρονικό διάστημα *T*, και για το χαρακτηριστικό αλληλεπίδρασης *f* , ως μια συνάρτηση *ID(O, r, T, f)*:

$$ID(O,r,T,f) = \begin{cases} AVG_{t \in T \wedge O' : Loc_{O'}(t) \in r}, & (f_O(t) - f_{O'}(t)) if\ t \in I \cap I' \\ απροσδιόριστο, & διαφορετικά \end{cases}$$

όπου *I'* δηλώνει το διάστημα ορισμού του αντικειμένου *O'*. ■

Οι περιγραφείς αλληλεπίδρασης (ΠΑ) περιγράφουν πώς κάθε αντικείμενο συσχετίζεται με άλλα σε κάθε περιοχή και κάθε χρονικό διάστημα, αυτό για κάθε διαφορετικό χαρακτηριστικό αλληλεπίδρασης δίνει μια διαφορετική προοπτική. Προφανώς, κάθε αντικείμενο θα διασχίσει μόνο ένα υποσύνολο των πιθανών περιοχών, και θα το κάνει αυτό με συγκεκριμένη σειρά. Επιπλέον, σε πραγματικές καταστάσεις, οι ιδιότητες κάθε αντικειμένου δεν είναι διαθέσιμες συνεχώς, αλλά μόνο σε συγκεκριμένα σημεία, με ένα δεδομένο ρυθμό δειγματοληψίας. Το τελικό αποτέλεσμα είναι ο χαρακτηρισμός κάθε αντικειμένου ως μια συλλογή χρονοσειρών, *ID*(*O; f*), μια για κάθε χαρακτηριστικό αλληλεπίδρασης *f*, που περιγράφει πώς, από κάθε άποψη, μεταβλήθηκε η αλληλεπίδραση μεταξύ του αντικειμένου *O* και των υπολοίπων κατά τη διάρκεια του χρόνου. Ή, ακριβέστερα, πόσο διαφορετικά συμπεριφέρθηκε το αντικείμενο από τα υπόλοιπα σε μια τέτοια αλληλεπίδραση.

Ενώ οι τυχαίες αλληλεπιδράσεις μπορούν να παραμεληθούν ως πλαστά φαινόμενα, οι επαναλαμβανόμενες συμπεριφορές μπορούν να θεωρηθούν ως ενδεχομένως σημαντικές για τη σκιαγράφηση του προφίλ κάθε αντικειμένου. Για παράδειγμα, εάν μόνο ένας οδηγός σε μια ομάδα οδηγών ταξί οδηγεί πάντα πολύ γρηγορότερα από τους άλλους γύρω του (που οδηγεί σε ακολουθία περιγραφέων αλληλεπίδρασης ταχύτητας με υψηλές τιμές), αυτός μπορεί να αντιμετωπιστεί ως ακραία περίπτωση (outlier). Αλλά εάν μια τέτοια συμπεριφορά είναι κοινή σε πολλές περιπτώσεις, είναι χρήσιμο να εξεταστεί η ύπαρξη μια κατηγορίας γρήγορων οδηγών.

Ορισμός 4-7 (Πρότυπα Αλληλεπίδρασης): Δεδομένου ενός πλήθους αντικειμένων *Objs*, έτσι ώστε το αντικείμενο $O \in Objs$ να έχει τα ίδια χαρακτηριστικά θέσης και αλληλεπίδρασης *Loc* και *Int* αντίστοιχα, ορίζουμε ένα *πρότυπο αλληλεπίδρασης* για το χαρακτηριστικό αλληλεπίδρασης $f \in Int$ ως μια ακολουθία τιμών $V = (v_1, \ldots, v_n) \in R^n$ έτσι ώστε, δεδομένου ενός κατωφλιού υποστήριξης $\sigma \in [0, 1]$ και μιας ανοχής σφάλματος $\varepsilon \in R^+$:

$$\frac{|\{O \in Objs \mid match\ (V, ID(O,f))\}|}{|Objs|} \geq \sigma$$

όπου $match\ (V, V') \Leftrightarrow \exists\ i_1 < \cdots < i_n\ \forall\ 1 \leq j \leq n \left|V[j] - V'[i_j]\right| \leq \varepsilon$ ■

Μια προσέγγιση των ανωτέρω καθορισμένων προτύπων αλληλεπίδρασης μπορεί να θεωρηθεί η διακριτοποίηση των τιμών κάθε περιγραφέα αλληλεπίδρασης, και το ταίριασμα δύο τιμών εάν πέφτουν στο ίδιο διακριτό διάστημα.

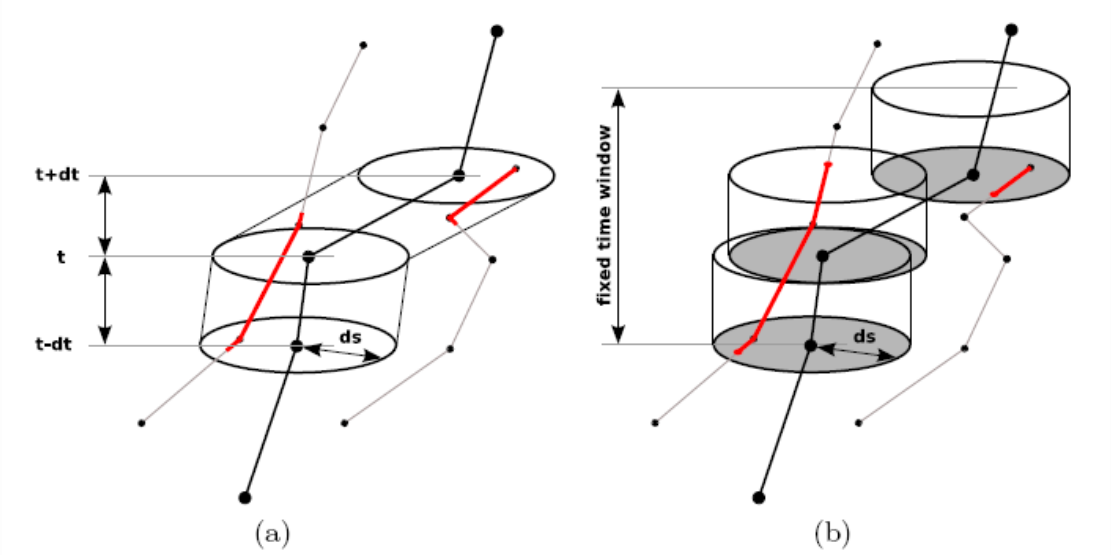
4.3.3. Υπολογισμός ΠΑ και γνωρισμάτων τροχιών

4.3.3.1. Επιλογή των τμημάτων των τροχιών που αλληλεπιδρούν

Σε γενικές γραμμές, η αλληλεπίδραση ενός αντικειμένου με άλλα είναι ιδιαίτερα δυναμική και αλλάζει συνεχώς κατά τη διάρκεια του χρόνου. Επομένως, πρέπει να επαν-υπολογίζεται για κάθε στιγμή *t* που περιλαμβάνεται στην τροχιά του αντικειμένου. Προκειμένου να γίνει αυτό, μια ακριβής λύση θα

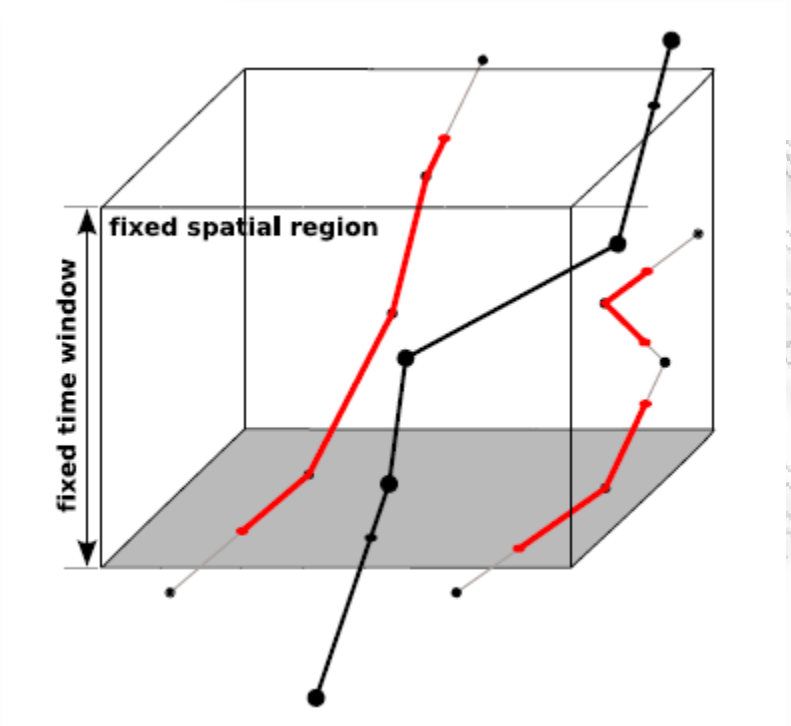
απαιτούσε να ανακαλύπτει τα αντικείμενα που περνούν κοντά (δηλ., εντός απόστασης *ds*) στο αντικείμενο αναφοράς μέσα σε ένα χρονικό παράθυρο *[t-dt, t+dt]*, και στη συνέχεια να συγκρίνει το τμήμα της τροχιάς του αντικειμένου αναφοράς που βρίσκεται στο χρονικό παράθυρο με τα αντίστοιχα τμήματα των γειτονικών αντικειμένων. Ένα παράδειγμα αυτής της διαδικασίας παρουσιάζεται στην Εικόνα 4-4(a), όπου τα κόκκινα τμήματα γραμμών προσδιορίζουν τα μέρη των τροχιών που μπορούν να συμβάλουν στους περιγραφείς αλληλεπίδρασης για την κεντρική τροχιά στο χρόνο *t*.

Ένας ελαφρώς λιγότερο ακριβής υπολογισμός των περιγραφέων αλληλεπίδρασης, που ακολουθείται σε αυτό το κεφάλαιο, απαιτεί τον χωρισμό του χρονικού άξονα σε διαστήματα ίσου πλάτους, και τον υπολογισμό της τιμής των περιγραφέων για κάθε χρονικό διάστημα. Τα τμήματα των γειτονικών αντικειμένων που εξετάζονται προσδιορίζονται με τρόπο παρόμοιο με τη γενική λύση που περιγράφηκε πιο πάνω, με αποτέλεσμα η χωροχρονική γειτονιά που υιοθετείται να απλοποιείται τελικά σε ένα σύνολο κυλίνδρων που έχουν τα κέντρα τους σε κάθε σημείο της τροχιάς αναφοράς ύψος τόσο ώστε να να καλύπτει το χρονικό χάσμα μεταξύ του σημείου και του επόμενου στην τροχιά. Αυτό απεικονίζεται σε Εικόνα 4-4(b), όπου, πάλι, τα κόκκινα τμήματα προσδιορίζουν τα σχετικά μέρη των γειτονικών αντικειμένων. Ας σημειωθεί ότι το χρονικό διάστημα της ανάλυσης είναι σταθερό, ενώ η χωρική γειτονιά μετακινείται με το αντικείμενο αναφοράς.



Εικόνα 4-4: (a) Ιδεατή γειτονιά γύρω από ένα αντικείμενο (b) γειτονιά σε ένα χρονικό παράθυρο.

Τέλος, ένα τραχύτερο, προσεγγιστικό αποτέλεσμα μπορεί να επιτευχθεί καθορίζοντας ένα χωρικό πλέγμα, έτσι ώστε οι περιγραφείς να υπολογίζονται για κάθε τμήμα τροχιάς που περιλαμβάνεται στο χωροχρονικό κελί που προκύπτει (ένα χωρικό κελί σε ένα δεδομένο χρονικό διάστημα), λαμβάνοντας υπόψη όλα τα τμήματα τροχιάς των γειτονικών αντικειμένων που εμπίπτουν επίσης στο κελί. Η Εικόνα 4-5 παρουσιάζει ένα γραφικό παράδειγμα, με τα ίδια χρώματα που χρησιμοποιήθηκαν και πριν. Σημειώστε ότι αυτή τη φορά και το χρονικό διάστημα και η χωρική γειτονιά είναι σταθερά.

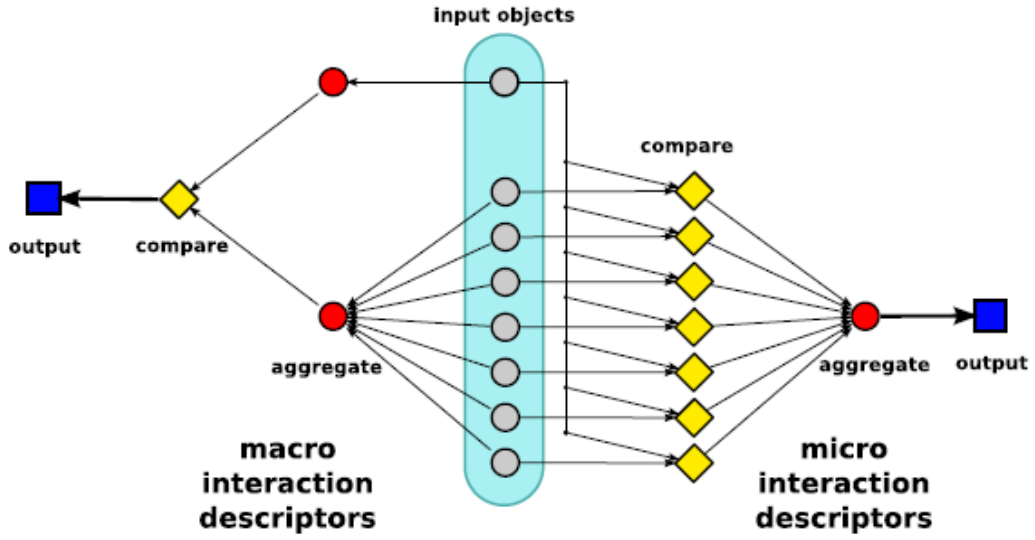


Εικόνα 4-5: Γειτονιά σταθερού πλέγματος

4.3.3.2. Υπολογισμός Περιγραφέων Αλληλεπίδρασης

Σε αυτό το κεφάλαιο εξετάζουμε δυο οικογένειες περιγραφέων ανάλογα με την υπολογιστική διαδικασία που απαιτούν. Η Εικόνα 4-6 συνοψίζει με γραφικό τρόπο τις δυο προσεγγίσεις:

- *Μακρο-* περιγραφείς αλληλεπίδρασης: όλα τα άλλα αντικείμενα συναθροίζονται σε μια ενιαία τιμή ή σε ένα σύνθετο αντικείμενο, κατόπιν το αντικείμενο αναφοράς που μας ενδιαφέρει να αναλύσουμε συγκρίνεται με το αποτέλεσμα μιας τέτοιας συνάθροισης και παράγει μια ενιαία τιμή. Π.χ., μια συνάθροιση μπορεί να επιστρέψει την ολική μέση ταχύτητα, και να υπολογιστεί η διαφορά μεταξύ της μέσης ταχύτητας του αντικειμένου αναφοράς και της ολικής μέσης ταχύτητα των γειτόνων της.
- *Μίκρο-* περιγραφείς αλληλεπίδρασης: το αντικείμενο αναφοράς συγκρίνεται με κάθε ένα από τα γειτονικά αντικείμενα, κατόπιν το σύνολο των αποτελεσμάτων συναθροίζεται σε μια ενιαία τιμή. Π.χ., για κάθε γειτονικό αντικείμενο, υπολογίζεται η ελάχιστη απόσταση από το αντικείμενο αναφοράς, και επιστρέφεται ο μέσος όρος όλων των τιμών που προέκυψαν.



Εικόνα 4-6: Μάκρο και μίκρο περιγραφείς αλληλεπίδρασης

Ας θεωρήσουμε μια τροχιά *T* και ένα σύνολο τροχιών *ST* που βρίσκονται σε μια συγκεκριμένη περιοχή *R* (που μπορεί να υπολογιστεί χρησιμοποιώντας είτε τη δυναμική γειτονιά Εικόνα 4-4(b) είτε την προσέγγιση σταθερού πλέγματος Εικόνα 4-5). Ο στόχος μας είναι να υπολογίσουμε την αλληλεπίδραση μεταξύ των *T* και *ST* που μπορεί να ποσοτικοποιηθεί μέσω διάφορων περιγραφέων αλληλεπίδρασης.

Για παράδειγμα, εξετάζουμε: τον περιγραφέα AVG_PERC_DISTANCE που μας επιτρέπει να συγκρίνουμε το μήκος της *T* με το μέσο μήκος των *ST*. Το ίδιο ισχύει και για τον περιγραφέα AVG_PERC_DURATION όπου η διάρκεια ζωής του *T* συγκρίνεται με τη μέση διάρκεια ζωής του *T*. Με τον ίδιο τρόπο, ο περιγραφέας AVG_PERC_SPEED υπολογίζεται διαιρώντας τη μέση ταχύτητα της *T* με τη μέση ταχύτητα κάθε τροχιάς *T'* του συνόλου *ST*. Παρομοίως, ο περιγραφέας AVG_PERC_ABS_ACCELER υπολογίζεται διαιρώντας τη μέση επιτάχυνση της *T* με τη μέση επιτάχυνση κάθε τροχιάς *T'* του συνόλου *ST*. Τέλος, ο περιγραφέας VAR_PERC_ACCELER υπολογίζεται ως το ποσοστό της διακύμανσης της επιτάχυνσης της τροχιάς *T* προς τη διακύμανση των επιταχύνσεων των *ST*. Οι περιγραφείς αυτοί μπορούν να υπολογιστούν χρησιμοποιώντας είτε την *μάκρο* είτε την *μίκρο* προσέγγιση.

Όπως είναι σαφές από την περιγραφή του υπολογισμού των διάφορων περιγραφέων, πρέπει να εργαστούμε με τις τροχιές αντί με τα ακατέργαστα σημεία. Τα πρωτογενή δεδομένα που συλλέγονται αντιπροσωπεύουν τις χρονοσημασμένες γεωγραφικές θέσεις από τις οποίες μπορούμε να πάρουμε πολύτιμες πληροφορίες για τις τροχιές. Το κύριο ενδιαφέρον μας είναι να ανακατασκευάσουμε τις τροχιές προκειμένου να μελετηθεί η κίνηση των αντικειμένων. Αυτό επιτυγχάνεται αποφασίζοντας και καθορίζοντας, σε κάθε συγκεκριμένη περίπτωση, ποιο σύνολο σημείων θέσης ορίζει μια συγκεκριμένη τροχιά. Για το σκοπό αυτό, υιοθετούμε τη μέθοδο εντοπισμού διαφορετικών τροχιών που προτείνεται στη [MFN+08a] και παρουσιάζεται λεπτομερώς στο Κεφάλαιο 5.

Προκειμένου να υπολογιστούν οι διαφορετικοί περιγραφείς, προτείνουμε δύο αλγορίθμους: COMP-DESCRIPTORS-DYNAMICNEIGHBORHOOD και COMP-DESCRIPTORS-FIXEDGRID. Ο πρώτος

χρησιμοποιείται για να υπολογίσει τους περιγραφείς χρησιμοποιώντας την προσέγγιση της δυναμικής γειτονιάς. Παίρνει ως είσοδο α) τη λίστα διαθέσιμων σημείων (lop), β) μια λίστα προκαθορισμένων χρονικών περιόδων (lotp) κατά τη διάρκεια των οποίων επιθυμούμε να υπολογίσουμε τους περιγραφείς, και γ) μια μέγιστη χωρική απόσταση (msd) που χρησιμοποιείται για να καθορίσει τη γειτονιά γύρω από κάθε σημείο. Αρχικά, καθορίζεται (γραμμή 4) το υποσύνολο των σημείων μέσα σε κάθε χρονική περίοδο. Μετά από αυτό, για κάθε διαφορετικό αντικείμενο, ανακατασκευάζονται μια ή περισσότερες τροχιές (γραμμή 6) και για κάθε τροχιά υπολογίζεται η γειτονιά της (γραμμή 10).

```
Algorithm Comp-Descriptors-DynamicNeighborhood (ListOfPoints
lop, ListOfTemporalPeriods lotp, MaxSpatialDistance msd)
 1. FOR EACH period p of lotp DO
 2.  //choose the subset of points that are temporally
 3.  //restricted inside period p
 4.  lop' = GetPointsForPeriod(lop, p);
 5.   FOR EACH object o of lop' DO
 6.    lot = ReconstructTrajectory(o, lop');
 7.     FOR EACH trajectory T in lot DO
 8.      //lont is a list of trajectories that are considered
 9.      //as the neighborhood of T
10.      lont = ComputeNeighborhood(T, msd);
11.      //compare T with lont set of trajectories
12.      ComputeDescriptors(T, lont);
13.     END-FOR
14.   END-FOR
15. END-FOR
```

Εικόνα 4-7: Ο αλγόριθμος COMP-DESCRIPTORS-DYNAMICNEIGHBORHOOD

Ο αλγόριθμος COMP-DESCRIPTORS-FIXEDGRID χρησιμοποιείται για να υπολογίσει περιγραφείς σε ένα σταθερό χωροχρονικό πλέγμα. Παίρνει ως είσοδο μια λίστα με τα διαθέσιμα σημεία (lop) και μια λίστα με προκαθορισμένα χωροχρονικά κελιά (lostc). Αρχικά, ορίζεται το υποσύνολο των σημείων που βρίσκεται σε κάθε χωροχρονικό κελί (γραμμή 4). Εδώ δεν υπάρχει η έννοια της γειτονιάς και έτσι το σύνολο των τροχιών είναι σταθερό επομένως μπορεί να ανακατασκευαστεί μια μόνο φορά (γραμμή 5). Έπειτα, υπολογίζονται οι περιγραφείς για κάθε τροχιά (γραμμή 9).

```
Algorithm Comp-Descriptors-FixedGrid
(ListOfPoints lop, ListOfSpatiotemporalCells lostc)
 1. FOR EACH spatiotemporal cell c of lostc DO
 2.  //choose the subset of points that are spatiotemporally
 3.  //restricted inside cell c
 4.  lop' = GetPointsForSTCell(lop, c);
 5.  lot = ReconstructTrajectories(lop');
 6.   FOR EACH trajectory T in lot DO
 7.    lont = lot.Exclude(T);
 8.     //compare T with lont set of trajectories
 9.     ComputeDescriptors(T,lont);
10.   END-FOR
11. END-FOR
```

Εικόνα 4-8: Ο αλγόριθμος COMP-DESCRIPTORS-FIXEDGRID

4.3.3.3. Εξαγωγή προτύπων αλληλεπίδρασης

Η γενική μέθοδος εξαγωγής γνωρισμάτων που προτείνεται ακολουθεί τρία κύρια βήματα που συνοψίζονται κατωτέρω:

Εξαγωγή περιοχής: επιλέγουμε μερικές περιοχές στο χώρο και στο χρόνο που θα χρησιμοποιηθούν ως θέσεις αναφοράς. Σε αυτό το βήμα επιλέγεται είτε η προσέγγιση της δυναμικής γειτονιάς (Εικόνα 4-4(b)) είτε του σταθερού πλέγματος (Εικόνα 4-5).

Περιγραφείς αλληλεπίδρασης: σε κάθε περιοχή *R*, εντοπίζουμε τα τμήματα των τροχιών που βρίσκονται μέσα στο *R* και ακολουθούμε είτε την *μάκρο-* είτε την *μίκρο-* προσέγγιση ώστε να υπολογίσουμε τους διάφορους περιγραφείς.

Πρότυπα αλληλεπίδρασης: αναλύονται οι ακολουθίες των περιοχών (π.χ. εκείνες που επισκέπτονται από κάθε τροχιά), αναζητώντας συχνές εξελίξεις στους περιγραφείς. Τα πρότυπα αλληλεπίδρασης μπορούν να εξαχθούν χρησιμοποιώντας ένα συμβατικό αλγόριθμο εξόρυξης διαδοχικών (sequential) προτύπων.

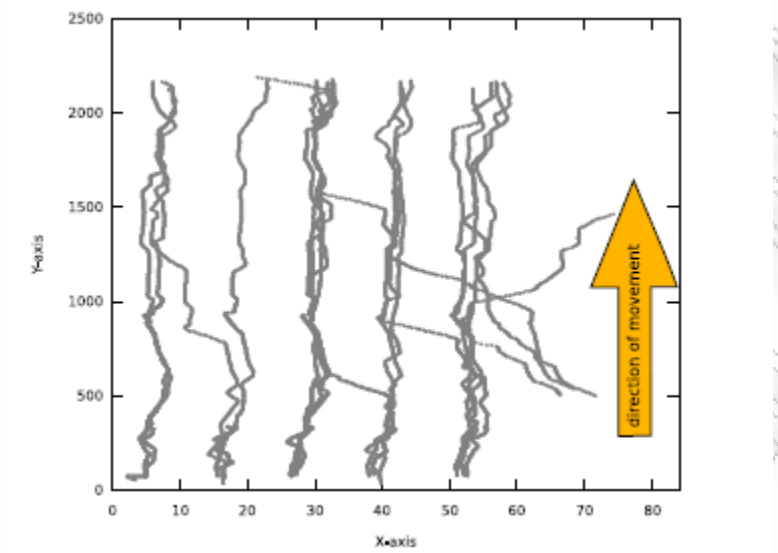
Τα πρότυπα αλληλεπίδρασης περιγράφουν περιληπτικά μερικές ενδιαφέρουσες πτυχές της δυναμικής του ευρύτερου συστήματος, όπως η επιτάχυνση/ επιβράδυνση ομάδων σε κάποια περιοχή, σε κάποιο χρονικό διάστημα, μια αλλαγή στην αμοιβαία απόσταση μεταξύ των ατόμων, κ.λπ. Γενικά, ένα πρότυπο αλληλεπίδρασης περιγράφει μια σύνδεση ή μια ακολουθία τέτοιων γεγονότων, κατά συνέπεια μοντελοποιεί ενδεχομένως σύνθετες, όμως στατιστικά σημαντικές, συμπεριφορές. Συνεπώς, κάθε πρότυπο αλληλεπίδρασης μπορεί να χρησιμοποιηθεί ως ένα περιγραφικό γνώρισμα, χαρακτηρίζοντας κάθε τροχιά ανάλογα με αν περιλαμβάνεται σε ένα τουλάχιστον επεισόδιο του. Στη παρούσα μελέτη, αυτά τα γνωρίσματα αντιμετωπίζονται ως ανεξάρτητες μεταβλητές κατά τη διαδικασία κατηγοριοποίησης τροχιών.

4.3.4. Πειραματική Μελέτη

Τα πειράματα πραγματοποιήθηκαν με το σύνολο δεδομένων NGSIM U.S. 101, μια ελεύθερα διαθέσιμη συλλογή ανακατασκευασμένων, από εικόνες, τροχιών που παρέχονται από την Ομοσπονδιακή Διοίκηση Εθνικών Οδών (Federal Highway Administration) και το πρόγραμμα Generation Simulation[1]. Τα στοιχεία περιγράφουν την κυκλοφορία των οχημάτων σε 2100 πόδια-μήκος και 6 λωρίδες-πλάτος (5 κύριες συν μια βοηθητική) ενός τμήματος της Εθνικού οδού των ΗΠΑ 101 (αυτοκινητόδρομος Hollywood) που βρίσκεται στο Λος Άντζελες της Καλιφόρνιας των ΗΠΑ και που καλύπτει ένα χρονικό διάστημα 45 λεπτών. Τα δεδομένα περιέχουν 6101 μοναδικές τροχιές, που αντιστοιχούν σε περίπου 4 εκατομμύρια σημεία, κάθε ένα από τα οποία δίνεται ως μια τετράδα *(ID, t, x, y)*, συν κάποιες επιπλέον πληροφορίες, όπως η ταχύτητα και η απόσταση από το επόμενο όχημα στην λωρίδα, που επίσης χρησιμοποιούνται στα πειράματά μας. Η Εικόνα 4-9 απεικονίζει τη χωρική προβολή ενός δείγματος του συνόλου δεδομένων που χρησιμοποιείται. Η κατεύθυνση της κίνησης των αυτοκινήτων είναι κατά μήκος του Y-άξονα - συγκεκριμένα η ύπαρξη διάφορων λωρίδων κατά μήκος του Y-άξονα (συμπεριλαμβανομένης της βοηθητικής για την είσοδο/την έξοδο από την εθνική οδό)

[1] http://www.ngsim.fhwa.dot.gov

είναι ορατή. Παρατηρήστε τις διαφορετικές κλίμακες που χρησιμοποιούνται για τους δύο άξονες, δεδομένου ότι το μήκος της περιοχής που μας ενδιαφέρει είναι πολύ μεγαλύτερο από το πλάτος του.



Εικόνα 4-9: Δείγμα του συνόλου δεδομένων U.S.101.

Η εργασία κατηγοριοποίησης εστίασε στον προσδιορισμό των αντικειμένων όπου στην άκρη του τμήματος της εθνικής οδού που αναλύουμε, θα παρουσίαζε τελικά κάποια επικίνδυνη συμπεριφορά. Μια τέτοια κατάσταση έχει τυποποιηθεί μέσω ενός απλού κανόνα που ονομάζει το τμήμα μιας τροχιάς ως προφίλ επικίνδυνης οδήγησης εάν συμβαίνει κάτι από τα παρακάτω σε κάποια χρονική στιγμή:

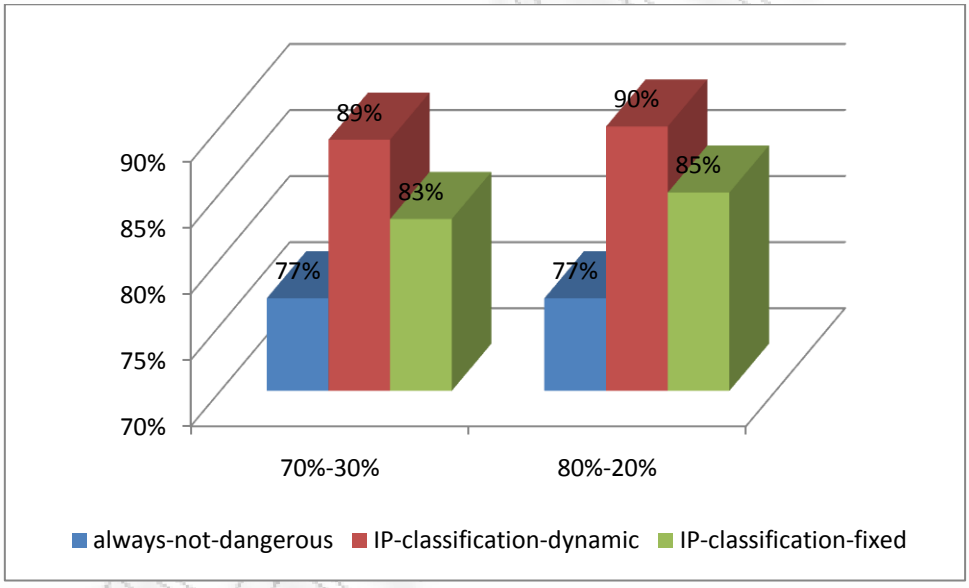
- το όχημα αλλάζει 2 ή περισσότερες λωρίδες, ή
- το όχημα κινείται τουλάχιστον στο 20% των χρονικών περιόδων πάνω από την μέση ταχύτητα, ή
- το όχημα κινείται το τουλάχιστον στο 10% των χρονικών περιόδων πάνω από το όριο ταχύτητας.

Αυτή η διαδικασία απόδοσης ετικέτας είχε ως αποτέλεσμα το 23% να θεωρηθούν ως προφίλ επικίνδυνης οδήγησης και 77% ως μη επικίνδυνων. Ο μηχανισμός κατηγοριοποίησης χρησιμοποιεί τις διακριτές τιμές των διάφορων περιγραφέων. Αυτές οι τιμές χρησιμοποιούνται ως «στοιχεία» για να παράγουν μια ακολουθία στοιχειοσυνόλων για κάθε αντικείμενο. Κατόπιν, οι συχνές υπο--ακολουθίες (τα πρότυπα αλληλεπίδρασής μας) εξάγονται. Κάθε σχέδιο χρησιμοποιείται έπειτα ως γνώρισμα, θέτοντας 1 εάν το αντικείμενο το ακολούθησε ειδάλλως 0. Επομένως, οι κατηγοριοποιητές (classifiers) που χτίζουμε προσπαθούν να καταλάβουν εάν το αντικείμενο είναι ένας επικίνδυνος οδηγός (ή, ακριβέστερα, θα είναι επικίνδυνος οδηγός στη «ουρά» της τροχιάς του) βάσει του ΠΑ που ακολούθησε στο πρώτο μέρος της κίνησής του.

Ο σκοπός του πρώτου πειράματος είναι να αποδείξει ότι χρησιμοποιώντας το πλαίσιο μας μπορούμε να προβλέψουμε με ακρίβεια τα επικίνδυνα και τα μη επικίνδυνα προφίλ χρησιμοποιώντας τους περιγραφείς αλληλεπίδρασης που περιγράφηκαν στην Υποενότητα 4.3.3.2. Για αυτόν τον σκοπό, το σύνολο δεδομένων χωρίστηκε σε δύο μέρη, στο70% και το 30% (δοκιμάσαμε επίσης το χωρισμό 80%-20%) του συνόλου δεδομένων, με το πρώτο να χρησιμοποιείται ως σύνολο δεδομένων εκπαίδευσης

και το δεύτερο για λόγους επικύρωσης. Σε ότι αφορά τη προσέγγιση της δυναμικής γειτονιάς, θέτουμε το χρονικό παράθυρο στο 1,5 λεπτό και τη χωρική ανωχή στα 5 μέτρα. Σε ότι αφορά τη στατική προσέγγιση, δημιουργούμε χωροχρονικά κελιά των 10 μέτρων (πλάτος) x 64 μέτρων (ύψος) x 2 λεπτά.

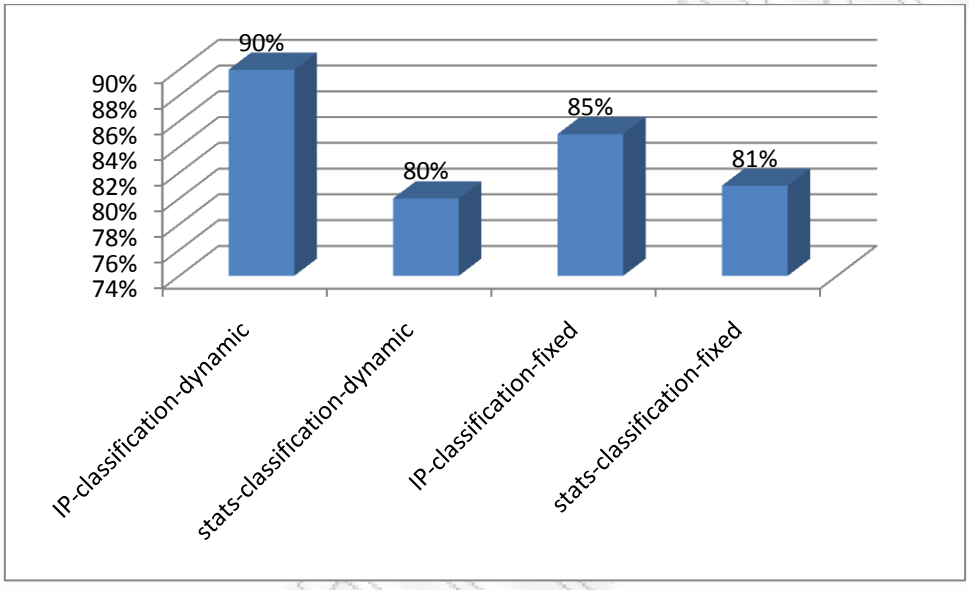
Στην Εικόνα 4-10, συγκρίνουμε τρεις κατηγοριοποιητές: έναν αφελή που κατηγοριοποιεί τους οδηγούς πάντα ως «μη επικίνδυνους» και επομένως έχει ένα σταθερό ποσοστό ακρίβειας (77%), και δύο άλλοι αξιοποιούν τη μέθοδο κατηγοριοποίησης που βασίζεται στους ΠΑ με εφαρμογή των προσεγγίσεων δυναμικής γειτονιάς και σταθερού πλέγματος. Τρέχουμε το πείραμα χρησιμοποιώντας τόσο το 70%-30% όσο και το 80%-20% του συνόλου δεδομένων για την εκπαίδευση-επικύρωση. Η διαφορά μεταξύ της δυναμικής-ταξινόμησης-ΠΑ (IP-classification-dynamic) και σταθερής-ταξινόμησης-ΠΑ (IP-classification-fixed) μπορεί να εξηγηθεί εύκολα αφού στην πρώτη περίπτωση οι περιγραφείς αλληλεπίδρασης που υπολογίζονται είναι μεταξύ της κίνησης των αντικειμένων που ανήκουν στην ίδια γειτονιά και πιθανότατα αλληλεπιδρούν. Αφετέρου, η σταθερή-ταξινόμηση-ΠΑ, μας επιτρέπει να υπολογίσουμε περιγραφείς μη λαμβάνοντας υπόψη έναν αυστηρό ορισμό της έννοιας της γειτονιάς (αλλά τα αποτελέσματα παραμένουν καλύτερα από την αφελή προσέγγιση).



Εικόνα 4-10: Μετρώντας την ακρίβεια του πλαισίου μας: ΠΑ κατηγοριοποίηση και αφελής κατηγοριοποιητής.

Στο δεύτερο πείραμά μας, χρησιμοποιούμε απλά στατιστικά που δεν λαμβάνουν υπόψη τις αλληλεπιδράσεις μεταξύ των αντικειμένων και τις συγκρίνουμε με την ΠΑ-κατηγοριοποίηση που προτείνουμε (Εικόνα 4-11). Σε αυτήν την περίπτωση, καθορίζουμε το ποσοστό εκπαίδευσης-επικύρωσης σε 80%-20%. Αυτό που θέλουμε να αποδείξουμε είναι ότι η μέτρηση των αλληλεπιδράσεων μεταξύ των αντικειμένων είναι σημαντική και μπορεί να μας δώσει μια πιο ολοκληρωμένη εικόνα σε ότι αφορά την κίνηση σε σχέση με εκείνη που παίρνουμε από τον υπολογισμό μερικών στατιστικών στοιχείων για τις απομονωμένες κινήσεις. Πιο συγκεκριμένα, η δυναμική-ταξινόμηση-ΠΑ και η σταθερή-ταξινόμηση-ΠΑ χρησιμοποιεί τους περιγραφείς αλληλεπίδρασης που παρουσιάζονται στην υποενότητα 4.3.3.2.: AVG_PERC_DISTANCE, AVG_PERC_DURATION, AVG_PERC_SPEED, AVG_PERC_ABS_ACCELER, και VAR_PERC_ACCELER. Από

την άλλη μεριά, οι τεχνικές δυναμική-ταξινόμηση- βάσει-στατιστικών και σταθερή-ταξινόμηση-βάσει-στατιστικών χρησιμοποιούν απλά στατιστικά, δηλ. AVG_SPEED_SINGLETRAJ (η μέση ταχύτητα της τροχιάς), AVG_ACCELERATION_SINGLETRAJ (η μέση επιτάχυνση της τροχιάς), VAR_SPEED_SINGLETRAJ (η διακύμανση της ταχύτητας της τροχιάς), VAR_ACCELERATION_SINGLETRAJ (η διακύμανση της επιτάχυνσης της τροχιάς) και COUNT_LINESCHANGED_SINGLETRAJ (ο αριθμός λωρίδων που άλλαξε το αντικείμενο). Όπως βλέπουμε στην Εικόνα 4-11, τα απλά στατιστικά δεν μπορούμε να μας βοηθήσουν να διαφοροποιήσουμε τους επικίνδυνους από τους μη επικίνδυνους οδηγούς με την ίδια ακρίβεια που πετυχαίνουμε χρησιμοποιώντας τα πρότυπα αλληλεπίδρασης.



Εικόνα 4-11: Μετρώντας την ακρίβεια του πλαισίου μας: κατηγοριοποίηση ΠΑ και κατηγοριοποιητής απλών στατιστικών.

4.4. Εξόρυξη Προτύπων Κυκλοφορίας

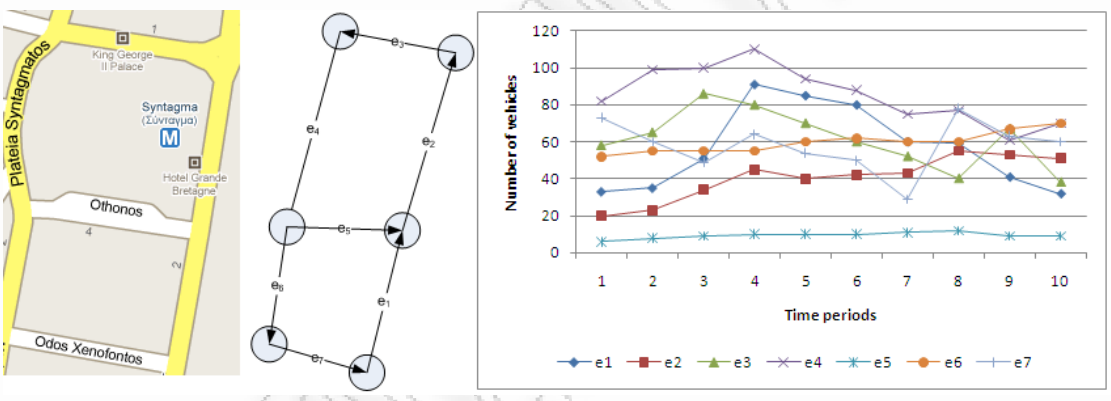
Η προσπάθεια κατανόησης, διαχείρισης και πρόβλεψης του φαινομένου της κυκλοφορίας είναι και ενδιαφέρουσα και χρήσιμη. Για παράδειγμα, οι αρχές των πόλεων αναζητούν βελτιώσεις της κυκλοφοριακής ροής και επιθυμούν να αντιδρούν αποτελεσματικά σε περίπτωση κυκλοφοριακού προβλήματος, όπως ένα αυτοκινητιστικό ατύχημα ή οδικά έργα. Επιπλέον, μελετώντας την κυκλοφοριακή ροή, οι αρχές των πόλεων θα μπορούσαν καλύτερα να προγραμματίσουν την κατασκευή νέων δρόμων, την επέκταση των υπαρχόντων, την τοποθέτηση των φαναριών κυκλοφορίας κτλ..

Τα Ευφυή Συστήματα Μεταφορών (ΕΣΜ/ Intelligent Transportation Systems - ITS) έχουν αναπτυχθεί για να επιτρέπουν την καλύτερη παρακολούθηση και τον έλεγχο της κυκλοφορίας προκειμένου να βελτιστοποιηθεί η κυκλοφοριακή ροή. Τα ΕΣΜ συλλέγουν δεδομένα αξιοποιώντας διάφορες τεχνολογίες όπως οι αισθητήρες, κάμερες κυκλοφορίας και κινητά τηλέφωνα και στη συνέχεια, προτείνουν εναλλακτικές διαδρομές ώστε να αποφευχθεί η συμφόρηση. Πέρα από την εφαρμογή στις εφαρμογές, η πληροφορική παρέχει ποικίλες εφαρμογές και τεχνολογίες για τη συλλογή, αποθήκευση, ανάλυση και παροχή πρόσβασης σε δεδομένα κυκλοφορίας. Ο στόχος μας είναι να παρουσιάσουμε ένα

ευφυές σύστημα υποστήριξης αποφάσεων διαχείρισης κυκλοφορίας που ενσωματώνει τεχνικές εξόρυξης γνώσης για την εξαγωγή προτύπων κυκλοφορίας.

Προκειμένου να εξυπηρετηθούν καλύτερα οι ανωτέρω σκοποί, στοχεύουμε στην ανάλυση των δεδομένων κυκλοφορίας ώστε να μελετηθεί η κυκλοφοριακή ροή και έτσι να ανακαλύψουμε πρότυπα κυκλοφορίας. Αυτά τα δεδομένα μπορούν να προκύψουν είτε από αισθητήρες που τοποθετούνται κατά μήκος του δικτύου και αποστέλλουν πληροφορίες για την κυκλοφορία σε διαδοχικές χρονικές περιόδους ή από GPS δεδομένα που αντιστοιχίζονται σε χάρτη (map match) και συναθροίζονται σε συγκεκριμένες χρονικές κλιμακώσεις. Τα πρότυπα που σχετίζονται με την κυκλοφορία μπορούν να εκφραστούν ως σχέσεις μεταξύ των οδικών τμημάτων του δικτύου πόλεων. Με άλλα λόγια, στοχεύουμε να ανακαλύψουμε πώς η κυκλοφορία ρέει σε αυτό το δίκτυο, ποια από τα οδικά τμήματα συμβάλλουν στη ροή και πώς αυτό συμβαίνει.

Μοντελοποιούμε το οδικό δίκτυο (Εικόνα 4-12α) ως ένα κατευθυνόμενο γράφο (Εικόνα 4-12β). Υποθέτουμε ότι συλλέγονται δεδομένα κίνησης με τη μορφή χρονοσειρών (Εικόνα 4-12γ) που μπορούν να αναλυθούν επιπλέον ώστε να ανακαλύψουμε σχέσεις μεταξύ των ακμών/ οδικών τμημάτων του δικτύου.



(α) (β) (γ)

Εικόνα 4-12: (α) ένα πραγματικό οδικό δίκτυο, (β) ο γράφους του αντίστοιχου δικτύου και (γ) οι χρονοσειρές των ακμών του δικτύου.

Ορίζουμε διαφορετικές σχέσεις μεταξύ των ακμών του γράφου του δικτύου: τη *διάδοση* (propagation) κυκλοφορίας από μια ακμή σε κάποια άλλη ακμή, τη *διάσπαση* (split) της κυκλοφορίας από μια ακμή σε πολλαπλές ακμές και τη *συγχώνευση* (merge) της κυκλοφορίας από τις πολλαπλές ακμές σε μια ακμή. Αυτές οι σχέσεις μάς επιτρέπουν να συνδυάσουμε τις χρονικές πληροφορίες που παρέχονται από τις χρονοσειρές κυκλοφορίας και τη χωρική σημασιολογία που προκύπτει από το οδικό δίκτυο. Με αυτόν τον τρόπο, είμαστε σε θέση να ανακαλύψουμε χωροχρονικά πρότυπα και να υποστηρίξουμε αποφάσεις διαχείρισης κυκλοφορίας. Αντίθετα με τις υπάρχουσες προσεγγίσεις, δεν στηριζόμαστε στις διακριτές τροχιές κινούμενων αντικειμένων αλλά σε συσσωρευτική πληροφορία που δίνει εικόνα της κυκλοφορίας μέσα στο δίκτυο.

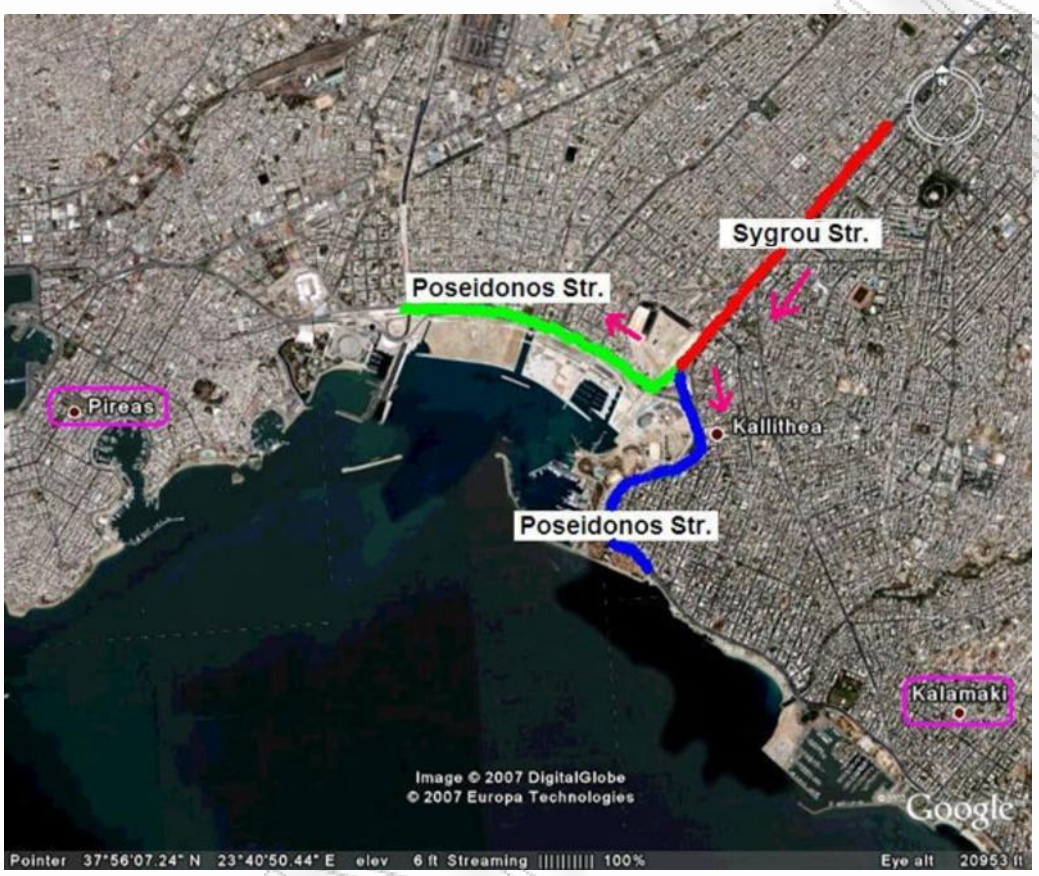
Στην Εικόνα 4-13, Εικόνα 4-14 και Εικόνα 4-15 παρουσιάζουμε πραγματικά παραδείγματα σχέσεων κυκλοφορίας στο οδικό δίκτυο της Αθήνας (τα τόξα στις εικόνες δείχνουν την κατεύθυνση της κίνησης) . Στην Εικόνα 4-13, απεικονίζεται ένα παράδειγμα διάδοσης κίνησης στην Λεωφόρο Κηφισίας: τα αντικείμενα συνεχίζουν να κινούνται πάνω στην Λεωφόρο Κηφισίας. Στην Εικόνα 4-14, παρουσιάζεται ένα παράδειγμα διάσπασης κίνησης: τα αντικείμενα στο τέλος της Λεωφόρου Συγγρού μπορούν με δυο τρόπους να συνεχίσουν στην Λεωφόρο Ποσειδώνος, είτε στρίβοντας αριστερά κατευθυνόμενα προς την περιοχή Καλαμάκι, είτε στρίβοντας δεξιά για Πειραιά. Στην Εικόνα 4-15 δείχνουμε ένα παράδειγμα συγχώνευσης κίνησης: τα αντικείμενα εισέρχονται στην Λεωφόρο Βασιλίσσης Σοφίας από τη Λεωφόρο Κηφισίας και την Λεωφόρο Μεσογείων.

Εικόνα 4-13: Παράδειγμα διάδοσης κυκλοφορίας στο οδικό δίκτυο της Αθήνας.

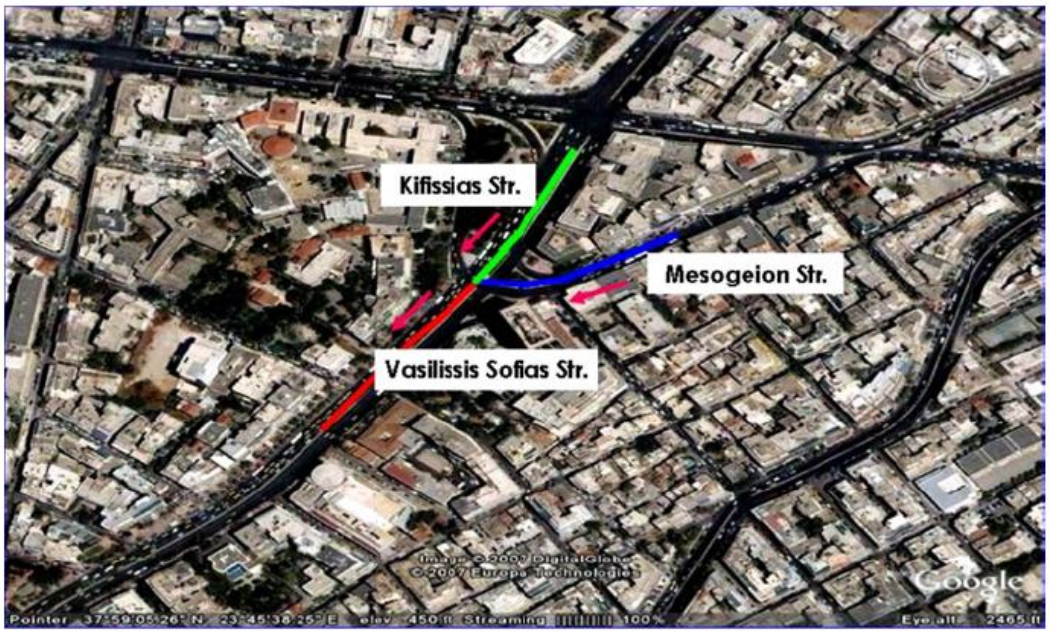
Αυτά τα παραδείγματα δείχνουν ότι η ανακάλυψης κυκλοφοριακών σχέσεων μεταξύ των ακμών του οδικού δικτύου μιας πόλης είναι ένα ενδιαφέρον θέμα και ταυτόχρονα ότι η κατανόησης αυτών των σχέσεων υα μπορούσε θα βοηθήσει τις Αρχές μιας πόλης να βελτιώσουν την κυκλοφοριακή ροή και να είναι σε θέση να αντιδράσουν αποτελεσματικά σε περίπτωση κυκλοφοριακού προβλήματος, όπως κάποιο ατύχημα ή επισκευές στους δρόμους. Παρέχουμε αποτελεσματικές μεθόδους για την ανακάλυψη τέτοιων σχέσεων κυκλοφορίας που βασίζονται σε κατάλληλα μέτρα απόστασης που ορίζουμε μεταξύ των χρονοσειρών των ακμών.

Θα πρέπει να σημειώσουμε ότι η αρχιτεκτονική TDW που παρουσιάστηκε αναλυτικά στο Κεφάλαιο 3 μπορεί να αξιοποιηθεί ως η αποθήκη δεδομένων αυτών των συνόλων κυκλοφοριακών δεδομένων. Πιο συγκεκριμένα, οι δυο κύριες διαστάσεις μιας TDW είναι η χωρική και η χρονική. Αυτές διαχωρίζουν τον χώρο και τον χρόνο σε συγκεκριμένες ελάχιστες κλίμακες (granularities) και μερικές ιεραρχίες μπορούν να στηθούν πάνω σε αυτές τις κλίμακες. Υπενθυμίζοντας την Εικόνα 3-1, μπορούμε να θέσουμε τις ακμές του δικτύου και τα συγκεκριμένα χρονικά διαστήματα ως τις ελάχιστες κλιμακώσεις της χωρικής και της χρονικής ιεραρχίας αντίστοιχα. Επιπλέον, μπορούμε να ορίσουμε ως μέτρο της

TDW τον αριθμό των οχημάτων. Έτσι, αυτή η TDW αποθηκεύει συγκεντρωτικά δεδομένα που μπορούν να χρησιμοποιηθούν για την ανάλυση δεδομένων κυκλοφορίας. Με άλλα λόγια, όπως συμβαίνει και στην παραδοσιακή KDD διαδικασία, η TDW τροφοδοτεί τους αλγορίθμους εξόρυξης γνώσης με τα κατάλληλα δεδομένα.



Εικόνα 4-14: Παράδειγμα διάσπασης κυκλοφορίας στο οδικό δίκτυο της Αθήνας.



Εικόνα 4-15: Παράδειγμα συγχώνευσης κυκλοφορίας στο οδικό δίκτυο της Αθήνας.

4.4.1. Μοντελοποίηση κυκλοφορίας

Στο μοντέλο μας, εξετάζουμε ένα προκαθορισμένο δίκτυο, όπως αυτό που απεικονίζεται στην Εικόνα 4-12α, που αποτελείται από ένα σύνολο μη-επικαλυπτόμενων περιοχών. Οι περιοχές μπορεί να είναι διασταυρώσεις ή σημεία ενδιαφέροντος όπως τράπεζες και εμπορικά κέντρα.

Ένα Δίκτυο Κυκλοφορίας (ΔK) αποτελείται από ένα σύνολο ανεξάρτητων (disjoint) περιοχών, $\{r_1, r_2, \ldots, r_m\} : (r_i \cap r_j) = \emptyset,\ 1 \le i, j \le m.$ ■

Τα αντικείμενα κινούνται σε αυτές τις περιοχές με βάση τη τοπολογία του δικτύου. Θεωρούμε ότι ένα οδικό δίκτυο μοντελοποιείται ως ένας κατευθυνόμενο γράφο $G = (V, E)$ όπου το σύνολο V των κόμβων αναπαριστά τοποθεσίες (π.χ. εμπορικά κέντρα, εργασιακούς χώρος, διασταυρώσεις) και το σύνολο E των ακμών αντιστοιχεί σε απευθείας συνδέσεις (δηλ., οδικά τμήματα) μεταξύ τους. Πιο συγκεκριμένα, μια ακμή e προστίθεται μεταξύ δύο περιοχών αν υπάρχει ένα οδικό δίκτυο που τις συνδέει. Υποθέτουμε ότι οι κινήσεις των αντικειμένων λαμβάνουν χώρα μόνο επί των ακμών και ότι από τη στιγμή που ένα αντικείμενο εισέλθει σε μια ακμή τότε τη διασχίζει όλη. Το μοντέλο του γράφου για το δείγμα δικτύου που δείξαμε στην Εικόνα 4-12α, παρουσιάζεται στην Εικόνα 4-12b.

Συμβολίζουμε με $start(e)$ τον κόμβο εκκίνησης της ακμής e και $end(e)$ τον κόμβο κατάληξης της e. Ονομάζουμε ως *εξερχόμενες ακμές της e*, συμβολίζοντας τις ως $out(e)$, εκείνες τις ακμές που ξεκινούν από τον κόμβο $end(e)$. Επίσης, ονομάζουμε ως *εισερχόμενες ακμές της e*, συμβολίζοντας τις ως $in(e)$, εκείνες τις ακμές που καταλήγουν στον κόμβο $start(e)$.

Θεωρούμε ότι υπάρχουν διαθέσιμα συγκεντρωτικά δεδομένα και πιο συγκεκριμένα: για κάθε ακμή, ο κυκλοφοριακός όγκος που περνάει από αυτήν σε διαδοχικές περιόδους. Έτσι, για κάθε ακμή $e \in E$ θεωρούμε μια σειρά χρονοσημασμένων μετρήσεων (v, t), όπου v είναι ο αριθμός των οχημάτων που πέρασαν από την ακμή κατά τη διάρκεια της χρονικής περιόδου $[t, t+\Delta t)$ όπου Δt αντιστοιχεί σε ένα χρονικό διάστημα για το οποίο μας ενδιαφέρει να συναθροίσουμε δεδομένα κυκλοφορίας. Οι χρονοσειρές των ακμών που δείχνουμε στην Εικόνα 4-12β, παρουσιάζονται στην Εικόνα 4-12γ.

Η *χρονοσειρά κυκλοφορίας* σε μια ακμή του δικτύου e κατά τη διάρκεια μιας χρονικής περιόδου $[t_s, t_e]$ ορίζεται ως ο αριθμός των οχημάτων που πέρασαν από την e κατά τη διάρκεια αυτής της περιόδου, καταγραμμένα αν Δt διαστήματα και ταξινομημένα στο χρόνο: $TS = (v_i, t_i)$, όπου v_i είναι ο αριθμός των οχημάτων που πέρασαν από την e κατά τη διάρκεια $[t_s, t_e]$, $t_s \le t_i \le t_e$ και $\Delta t = t_i - t_{i-1}$. ■

Ένα δείγμα χρονοσειράς κυκλοφορίας μοιάζει κάπως έτσι: $TS=\{(150, 10:00), (100,10:15), (80, 10:30), \ldots\}$.

4.4.2. Συσταδοποίηση κυκλοφοριακών ακμών

Σε αυτήν την υποενότητα, παρουσιάζουμε την προσέγγιση μας για συσταδοποίηση των ακμών ενός οδικού δικτύου αναλύοντας τις χρονοσειρές κυκλοφορίας χρησιμοποιώντας ένα σύνολο μέτρων ομοιότητας.

4.4.2.1. Μέτρα ομοιότητας μεταξύ κυκλοφοριακών ακμών

Αν θεωρήσουμε e_1, e_2 ως δυο ακμές του δικτύου και $TS_1=\{(v_{1i},t_i)\}$, $TS_2 =\{(v_{2i},t_i)\}$ $t_i \in [t_s, t_e]$ τις αντίστοιχες χρονοσειρές κυκλοφορίας, Στην εργασία [NMM08] προτείναμε μια απόσταση μεταξύ των

ακμών του δικτύου e_1, e_2 ως ένα συνδυασμό (βασισμένο σε βάρη) των αντίστοιχων αποστάσεων βάσει τιμής, σχήματος και δομής:

$$dis(e_1, e_2) = a * dis_{shape}(e_1, e_2) + b * dis_{struct}(e_1, e_2) + c * dis_{value}(e_1, e_2) \quad (4.1)$$

όπου $dis_{value}(e_1, e_2)$ είναι η *απόσταση βάση τιμής* (value based distance) μεταξύ των e_1 και e_2 που δίνεται από την Ευκλείδεια απόσταση των αντίστοιχων χρονοσειρών τους (TS_1, TS_2):

$$dis_{value}(e_1, e_2) = dis_{value}(TS_1, TS_2) = \sqrt{\sum (v_1[t_i] - v_2[t_i])^2}, t_s \le t_i \le t_e \quad (4.2)$$

Η απόσταση βάσει τιμής αναζητά χρονοσειρές κυκλοφορίας με ίδιες (ή σχεδόν ίδιες) τιμές.

Εκτός από ακμές με ίδιες τιμές κυκλοφορίας, ενδιαφερόμαστε επίσης να εντοπίσουμε ακμές με παρόμοιο σχήμα κυκλοφορίας, δηλ., ακμές των οποίων η κυκλοφορία αυξάνεται και ελαττώνεται με τον ίδιο ρυθμό. Η Ευκλείδεια απόσταση δεν είναι κατάλληλη σε αυτήν την περίπτωση αφού δεν επιτρέπει διαφορετικές τάξεις μεγέθους στις χρονοσειρές (π.χ., μια χρονοσειρά κυμαίνεται γύρω στο 100 ενώ μια άλλη γύρω στο 30, ή διαφορετικές κλίμακες (π.χ., μια χρονοσειρά έχει εύρος μεταξύ 90 και 100 ενώ μια άλλη λίγο μεγαλύτερο μεταξύ 20 και 40). Για να ανακαλύψουμε χρονοσειρές κυκλοφορίας με παρόμοιο σχήμα (άλλα όχι αναγκαστικά παρόμοιες τιμές), εφαρμόζουμε έναν μετασχηματισμό κανονικοποίησης [DG03] πάνω στις αρχικές χρονοσειρές. Αν $\mu(TS)$, $\sigma(TS)$ είναι ο μέσος και η τυπική απόκλιση μιας χρονοσειράς $TS = \{v_i, t_i\}, 1 \le i \le n$. Η χρονοσειρά TS αντικαθίσταται από την *κανονικοποιημένη* $TS' = \{v_i', t_i\}, 1 \le i \le n$, όπου:

$$v_i' = \frac{v_i - \mu}{\sigma}, \text{ όπου } \mu = \frac{1}{n}\sum_{i=1}^{n} v_i \text{ και } \sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(v_i - \mu)^2} \quad (4.3)$$

Έτσι η Ευκλείδεια απόσταση μεταξύ των αντίστοιχων ψ (TS_1', TS_2') ορίζεται ως εξής:

$$EuclideanDistance(TS_1', TS_2') = \sqrt{\sum (v_1'[t_i] - v_2'[t_i])^2}, t_s \le t_i \le t_e \quad (4.4)$$

Ορίζουμε με $dis_{shape}(e_1, e_2)$ την *απόσταση βάσει σχήματος* μεταξύ των e_1 και e_2 που δίνεται από την Ευκλείδεια απόσταση των αντίστοιχων κανονικοποιημένων (ώστε να αποφύγουμε διαφορές σε εύροι τιμών, κλίμακες κτλ.) χρονοσειρών κυκλοφορίας (TS_1', TS_2'):

$$dis_{shape}(e_1, e_2) = dis_{shape}(TS_1, TS_2) = EuclideanDistance(TS_1', TS_2') \quad (4.5)$$

Τέλος, $dis_{struct}(e_1, e_2)$ είναι η *απόσταση βάσει δομής* μεταξύ δυο ακμών e_1 και e_2 και ισούται με τον ελάχιστον αριθμό ακμών μεταξύ των $end(e_1)$ και $start(e_2)$.

Για κάθε εφαρμογή, μπορούμε να αρχικοποιήσουμε τα βάρη a, b, c σύμφωνα με το μέτρο (μέτρα) στο οποίο θέλουμε να δώσουμε έμφαση. Αν δούμε τα τρία μέτρα ξεχωριστά, παρατηρούμε ότι κάθε μέτρο φιλτράρει περαιτέρω το αρχικό σύνολο των ακμών κυκλοφορίας. Πιο συγκεκριμένα, η απόσταση βάση *σχήματος* επιστρέφει ένα σύνολο ακμών με παρόμοιο σχήμα κυκλοφορίας, η απόσταση βάσει δομής, αναζητά γειτονικές ακμές και τέλος, η απόσταση βάσει τιμής περιορίζει το τελικό σύνολο αναζητώντας ακμές με παρόμοιες τιμές.

4.4.2.2. Ο αλγόριθμος Traffic-Clustering

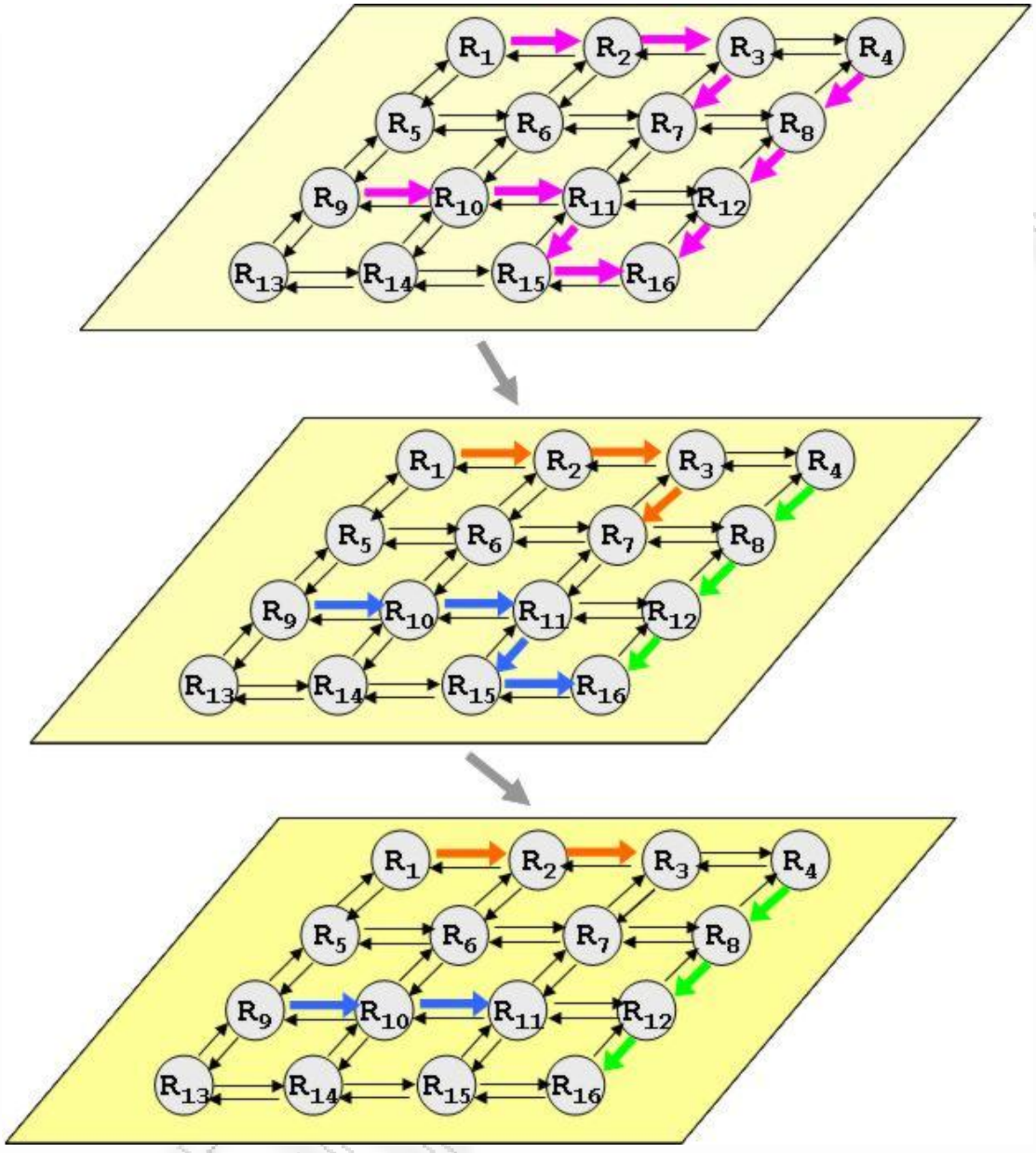
Εισαγάγαμε τρεις τρόπους για να μετρήσουμε την ομοιότητα μεταξύ δύο ακμών κυκλοφορίας βασισμένων είτε στις τιμές τους είτε στην εγγύτητά τους στο δικτυακό γράφο. Κάθε μέτρο αποκαλύπτει διαφορετικές πτυχές του προβλήματος κυκλοφορίας: το μέτρο που είναι βασισμένο στις τιμές ανακαλύπτει χρονοσειρές κυκλοφορίας με παρόμοιες τιμές, το μέτρο που είναι βασισμένο στο σχήμα ανακαλύπτει χρονοσειρές κυκλοφορίας παρόμοιου σχήματος, ενώ η απόσταση που βασίζεται στη δομή ανακαλύπτει τις ακμές που είναι η μια κοντά στην άλλη όσον αφορά την τοπολογία του γράφου.

Τα βάρη *a, b, c* της εξίσωσης (4.1) αρχικοποιούνται σύμφωνα με τις συγκεκριμένες απαιτήσεις κάθε εφαρμογής. Αυτό οδηγεί σε μια λογική (π.χ. $a >> b >> c$) που παρέχει μια ιεραρχία κυκλοφοριακής ροής που οργανώνεται σε διαφορετικά επίπεδα. Για παράδειγμα, αν η λογική είναι $a >> b >> c$ τότε η ιεραρχία σχηματίζεται ως: το επίπεδο των ακμών *κοινού κυκλοφοριακού σχήματος* (*L*1), το επίπεδο των ακμών *κοινού κυκλοφοριακού σχήματος* που βρίσκονται επίσης *κοντά* στο γράφο (*L*2) και τέλος, το επίπεδο των *ακμών με κοινές κυκλοφοριακές τιμές* (*L*3). Ένα τέτοιο παράδειγμα ιεραρχίας παρουσιάζεται στην Εικόνα 4-17.

Για να ανιχνεύσουμε μια τέτοια ιεραρχία υιοθετούμε τον TRAFFIC-CLUSTERING, ένα διαιρετικό, ιεραρχικό αλγόριθμο συσταδοποίησης που παρουσιάζεται στην Εικόνα 4-16. Το μέτρο απόστασης που χρησιμοποιεί είναι αυτό της εξίσωσης (4.1), η οποία συνδυάζει τις τρεις έννοιες της απόστασης μεταξύ των ακμών κυκλοφορίας. Ο αλγόριθμος λειτουργεί ως εξής: Αρχικά όλες οι ακμές τοποθετούνται σε μια συστάδα. Σε κάθε βήμα του αλγορίθμου, μια συστάδα χωρίζεται περαιτέρω σε υποσυστάδες (γραμμές 7-9) σύμφωνα με τα μέτρα ομοιότητας dis_{value}, dis_{struct}, dis_{shape}. Η σειρά με την οποία εφαρμόζονται εξαρτάται από τις τιμές των βαρών *a*, *b* και *c*.

```
Algorithm TrafficClustering (ListOfEdgesTS loe, weight a, weight
b, weight c)
1. //similarity measures (dis_value, dis_struct, dis_shape) are applied
2. //in the order that is defined by the weights a, b and c.
3. //we define dis1, dis2 and dis3 as the first, the second and
4. //the last measure respectively
5. //each step is completed when a split is caused by the next
6. //distance measure
7.  clusters = splitClusters(loe, dis1)
8.  clusters = splitClusters(clusters, dis2)
9.  clusters = splitClusters(clusters, dis3)
```

Εικόνα 4-16: Ο αλγόριθμος TRAFFIC-CLUSTERING



Εικόνα 4-17: Ιεραρχία των ακμών – οι ακμές με κοινό χρώμα (μωβ, πράσινο, πορτοκαλί, μπλε) ανήκουν στην ίδια συστάδα.

Για παράδειγμα, ορίζοντας $a >> b >> c$ δείχνουμε ότι ενδιαφερόμαστε πρώτα για τις ακμές με παρόμοιο σχήμα κυκλοφορίας, κατόπιν για τις ακμές που βρίσκονται η μια κοντά στην άλλη και τέλος, για τις ακμές που έχουν παρόμοιες τιμές. Με αυτόν τον τρόπο, ο αλγόριθμος «ευνοεί» πρώτα τις τιμές με παρόμοιο σχήμα κυκλοφορίας (βάρος *a*), έπειτα ακμές που είναι επίσης γειτονικές (βάρος *b*) και τέλος, ακμές που είναι επίσης παρόμοιες σε ότι αφορά τις τιμές τους (βάρος *c*). Σε αυτήν την περίπτωση, ο αλγόριθμος χωρίζει την αρχική ενιαία συστάδα σε υποσυστάδες σύμφωνα με τα ακόλουθα τρία βήματα:

- Βήμα 1 [Ακμές παρόμοιου σχήματος]: Μια συστάδα χωρίζεται σε υποσυστάδες βάσει της ομοιότητας σχήματος των ακμών-μελών της. Αυτή η διαδικασία συνεχίζεται μέχρι να υπάρξει

διαχωρισμός από το επόμενο μέτρο απόστασης, την απόσταση βάση δομής. Στο τέλος αυτού του βήματος, οι συστάδες περιλαμβάνουν ακμές με παρόμοιο κυκλοφοριακά σχήμα.

- Βήμα 2 [Γειτονικές ακμές]: Οι συστάδες που προέκυψαν από το προηγούμενο βήμα διαχωρίζονται επιπλέον βάσει της δομικής τους απόστασης μέχρι να προκληθεί διαχωρισμός λόγω της απόσταση βάσει των κυκλοφοριακών τιμών. Σε αυτό το σημείο, οι συστάδες περιλαμβάνουν γειτονικές ακμές κυκλοφορίας με παρόμοιο σχήμα.
- Βήμα 3 [Ακμές παρόμοιων τιμών]: Οι συστάδες που προέκυψαν από το προηγούμενο βήμα διαχωρίζονται επιπλέον βάσει της απόστασης των τιμών τους. Στο τέλος της εκτέλεσης, οι συστάδες περιλαμβάνουν γειτονικές ακμές με παρόμοιες τιμές και παρόμοια σχήματα.

Η ανακάλυψη μιας τέτοιας ιεραρχίας της κυκλοφορίας σε ότι αφορά τις ακμές του δικτύου είναι χρήσιμη για τον εμπειρογνώμονα (π.χ., Αρχές πόλεων), δεδομένου ότι είναι σε να εμβαθύνει στην ιεραρχία από κάποιες γενικές ομάδες ακμών με παρόμοιο σχήμα, σε ομάδες γειτονικών ακμών και στη συνέχεια σε ομάδες ακμών με παρόμοιες τιμές.

4.4.3. Ανακαλύπτοντας σχέσεις κυκλοφορίας εστιασμένες στο χρόνο

Σε αυτήν την υποενότητα, παρουσιάζουμε την προσέγγισή μας για την ανακάλυψη των σχέσεων κυκλοφορίας μεταξύ των ακμών του οδικού δικτύου. Η λογική πίσω από τη μεθοδολογία μας είναι να αναλύσουμε τη χρονοσειρά κυκλοφορίας και να προσπαθήσουμε να ανακαλύψουμε τις ομοιότητες μεταξύ τους σε συγκεκριμένες περιόδους.

4.4.3.1. Μέτρα ομοιότητας μεταξύ κυκλοφοριακών ακμών

Αν θεωρήσουμε ότι e_1, e_2 είναι δυο ακμές του δικτύου και $TS_1=\{(v_{1i},t_i)\}$, $TS_2=\{(v_{2i},t_i)\}$, $t_i \in [t_s, t_e]$ είναι οι αντίστοιχες χρονοσειρές τους. Ορίζουμε ως *απόσταση* βάσει *τιμής μεταξύ των* e_1, e_2 κατά τη διάρκεια των περιόδων p και p' ως την απόλυτη απόσταση μεταξύ των αντίστοιχων χρονοσειρών TS_1, TS_2 σε εκείνες τις περιόδους:

$$dis_{value}\left(e_1^p, e_2^{p'}\right) = dis_{value}\left(TS_1^p, TS_2^{p'}\right) \quad (4.6)$$

Όπως αναφέραμε στην Υποενότητα 4.4.2.1, ενδιαφερόμαστε επίσης για την ανίχνευση των ακμών με κυκλοφορία παρόμοιου σχήματος, δηλ., ακμές των οποίων η κυκλοφορία αυξάνεται και μειώνεται με τον ίδιο ρυθμό. Για να ανακαλύψουμε χρονοσειρές κυκλοφορίας με παρόμοιο σχήμα (αλλά όχι απαραίτητα με παρόμοιες τιμές), χρησιμοποιούμε τρεις γνωστές τεχνικές: Ευκλείδεια απόσταση και Dynamic Time Warping (DTW) [SC78] σε κανονικοποιημένες χρονοσειρές και Correlation-Coefficient. Το μέτρο ομοιότητας σχήματος δεν εφαρμόζεται στο πλήρες σύνολο δεδομένων, έτσι ορίζουμε ένα παράθυρο της μορφής $(p\text{-}w_b, p+w_a)$ στο οποίο ψάχνουμε την ομοιότητα.

Όσο αφορά τα δυο πρώτα, εφαρμόζουμε το μετασχηματισμούς (4.3) και (4.5) στην αρχική χρονοσειρά κυκλοφορίας. Με αυτόν τον τρόπο, είμαστε σε θέση να πιάσουμε τις αλλαγές στους όγκους σε διαφορετικές χρονοσειρές.

Ορίζουμε ως $dis_{shapeEu}\left(e_1^p, e_2^p\right)$ την *απόσταση βάσει σχήματος* μεταξύ των e_1 και e_2 την περίοδο p την *απόσταση βάση τιμής* (4.2) των τμημάτων των αντίστοιχων κανονικοποιημένων χρονοσειρών τους $(TS_1^{'}, TS_2^{'})$ στο χρονικό παράθυρο *(p-*w_b*, p+*w_a*)*:

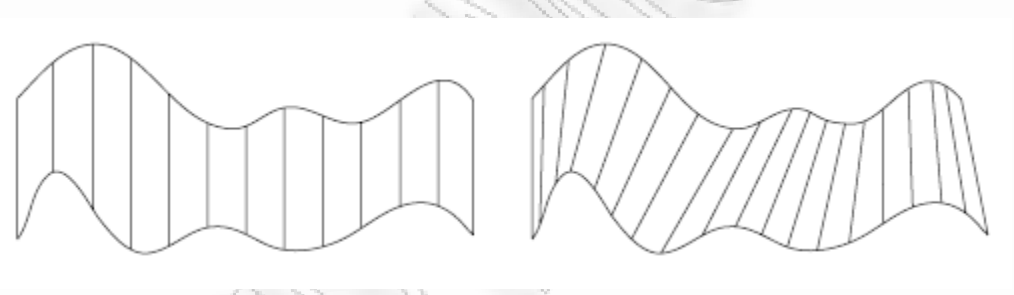
$$dis_{shapeEu}\left(e_1^p, e_2^p\right) = dis_{value}\left(TS_1^{'}, TS_2^{'}\right), \text{ όπου για } TS_1^{'}, TS_2^{'}, t_s = p - w_b \text{ και } t_e = p + w_a \quad (4.7)$$

Επίσης ορίζουμε ως $dis_{shapeDTW}\left(e_1^p, e_2^p\right)$ την *απόσταση βάσει σχήματος* μεταξύ των e_1 και e_2 την περίοδο p την *απόσταση* μεταξύ των τμημάτων των αντίστοιχων κανονικοποιημένων χρονοσειρών τους $(TS_1^{'}, TS_2^{'})$ στο χρονικό παράθυρο *(p-*w_b*, p+*w_a*)* έτσι ώστε να ελαχιστοποιείται το distance warping path:

$$dis_{shapeDTW}\left(e_1^p, e_2^p\right) = \sqrt{\sum\left(v_1[t_i] - v_2[t_j]\right)^2}, \text{όπου } p - w_b \le t_i, t_j \le p + w_a,$$
$$\text{και } \sum\left|v_1[t_i] - v_2[t_j]\right| = min_w\left[\sum_{k=1}^{K} d(w_k)\right] \quad (4.8)$$

όπου $d(w_k)$ είναι η απόσταση μεταξύ των *K* στοιχείων της χρονοσειρά.

Με άλλα λόγια, αποφασίζουμε να μην υπολογίσουμε τις αποστάσεις μεταξύ των ευθυγραμμισμένων τιμών χρησιμοποιώντας τη Ευκλείδεια απόσταση αλλά να υπολογίσουμε τις αποστάσεις μεταξύ αυτών των ζευγαριών ώστε να ελαχιστοποιείται το distance warping path (Εικόνα 4-18).



Εικόνα 4-18: Οι γραμμές μεταξύ των χρονοσειρών δείχνουν ευθυγραμμίσεις τιμών όπως τις αντιμετωπίζει η Ευκλείδεια απόσταση (αριστέρα) και το Dynamic Time Warping μέτρο ομοιότητας (δεξιά).

Τέλος, ορίζουμε ως $dis_{shapeCC}\left(e_1^p, e_2^p\right)$ την *απόσταση βάσει σχήματος* μεταξύ των e_1 και e_2 την περίοδο p ως το Correlation-Coefficient (r) μεταξύ των τμημάτων των χρονοσειρών τους στο χρονικό παράθυρο *(p-*w_b*, p+*w_a*)*:

$$dis_{shapeCC}\left(e_1^p, e_2^p\right) = r\left(TS_1^{(p-w_b, p+w_a)}, TS_2^{(p-w_b, p+w_a)}\right) \quad (4.9)$$

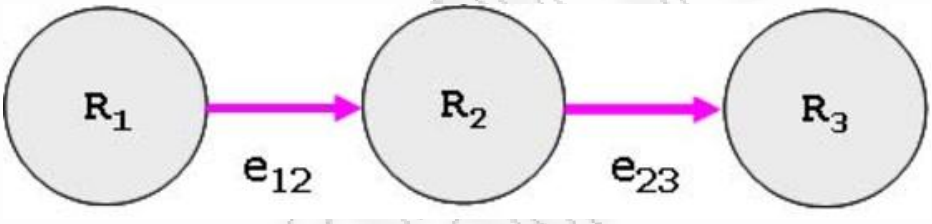
Αυτό το μέτρο είναι ανεξάρτητο από την κλίμακα της χρονοσειράς και μπορεί να μας δώσει ένα μέτρο ομοιότητας μεταξύ τους.

Θα πρέπει να σημειώσουμε ότι η Ευκλείδεια και η DTW απόσταση δεν είναι κανονικοποιημένες. Αυτό καθιστά δύσκολη την σύγκριση μεταξύ των χρονοσειρών. Για να το αντιμετωπίσουμε αυτό ορίζουμε μια τιμή που αναπαριστά το μέγιστο όγκο οχημάτων που μπορούν να βρίσκονται σε μια ακμή ανά χρονική περίοδο. Με τον τρόπο αυτό μπορούμε να δημιουργήσουμε με τεχνητό τρόπο μια μέγιστη απόσταση μεταξύ των χρονοσειρών και έτσι να κανονικοποιήσουμε τις αποστάσεις.

4.4.3.2. Σχέσεις κυκλοφορίας

Η ανίχνευση των σχέσεων κυκλοφορίας μεταξύ των διαφορετικών οδικών τμημάτων είναι ένα ενδιαφέρον πρόβλημα. Εντός ενός συγκεκριμένου χρονικού διαστήματος, η κυκλοφορία μιας ακμή μπορεί (διαζευκτικά) να: i) διαδίδεται «όπως είναι» σε κάποια εξερχόμενη ακμή (που δείχνει π.χ., ότι τα αντικείμενα συνεχίζουν να κινούνται πάνω σε μια κεντρική αρτηρία), ii) διασπάται σε πολλαπλές εξερχόμενες ακμές (που δείχνει π.χ. ότι τα αντικείμενο φεύγουν από μια κεντρική αρτηρία και ακολουθούν διαφορετικές κατευθύνσεις προς τον προορισμό τους), iii) είναι το αποτέλεσμα της συγχώνευσης από μερικές από τις εισερχόμενες ακμές της (που δείχνουν π.χ. αντικείμενα που μπαίνουν σε μια κεντρική αρτηρία μέσω διαφορετικών κατευθύνσεων). Επίσης, μια ακμή μπορεί να λειτουργεί ως iv) ακμή-προορισμός (sink) (που δείχνει π.χ. έναν χώρο εργασίας, όπου οι άνθρωποι σταθμεύουν το πρωί) ή ως v) ακμή-πηγή (source) (που δείχνει π.χ. τον ίδιο χώρο εργασίας το απόγευμα, όπου οι άνθρωποι αφήνουν τα γραφεία και επιστρέφουν σπίτι τους).

Παρακάτω ορίζουμε αυτές τις σχέσεις με τυπικό τρόπο. Με $p = [p_s, p_e]$ συμβολίζουμε τη χρονική περίοδος κατά τη διάρκεια της οποία θέλουμε να βρούμε πώς σχετίζονται οι ακμές και με TS_e^p τα τμήματα των χρονοσειρών της ακμής *e* κατά τη διάρκεια εκείνης της περιόδους.



Εικόνα 4-19: Η ακμή e_{12} διαδίδει την κυκλοφορία της στην ακμή e_{23}.

Χρησιμοποιούμε δυο κατώφλια απόστασης: τ_v ως την ελάχιστη επιτρεπόμενη ομοιότητα τιμής και τ_s ως την ελάχιστη επιτρεπόμενη ομοιότητα σχήματος. Επιπλέον, χρησιμοποιούμε δυο χρονικά διαστήματα w_b και w_a ώστε να ορίσουμε ένα χρονικό παράθυρο μέσα στο οποίο ψάχνουμε για ομοιότητα σχήματος μεταξύ των δυο χρονοσειρών.

Στους παρακάτω ορισμούς, dis_{shape} αναπαριστά ένα από τα μέτρα ομοιότητας σχήματος που συζητήθηκαν στην Υποενότητα 4.4.3.1 ($dis_{shapeEu}$, $dis_{shapeDTW}$, $dis_{shapeCC}$).

Ορισμός 4-8 (Διάδοση κυκλοφορίας): Μια ακμή e_1 διαδίδει την κυκλοφορία της κατά τη διάρκεια της *p* σε μια ακμή e_2 κατά τη διάρκεια της *p'* ανν:

i) $end(e_1) = start(e_2)$,

ii) $dis_{value}\left(TS_{e_1}^{p}, TS_{e_2}^{p'}\right) \geq t_v$ και

iii) $dis_{shape}\left(e_1^p, e_2^p\right) \geq t_s$. ■

Διαισθητικά, η κυκλοφορία της ακμής e_1 κατά τη διάρκειας της *p* διαδίδεται (Εικόνα 4-19) στην e_2 κατά τη διάρκεια της *p'* αν η περισσότερη κυκλοφορία της e_1 κατευθύνεται στην e_2 κατά τη διάρκεια αυτής της περιόδου (ο απαιτούμενος όγκος που πρέπει να μεταφέρεται διασφαλίζεται μέσω του κατωφλιού t_v) αλλά και το σχήμα των χρονοσειρών τους, μέσα στο παράθυρο είναι παρόμοιο (όπως διασφαλίζεται από το κατώφλι t_s).

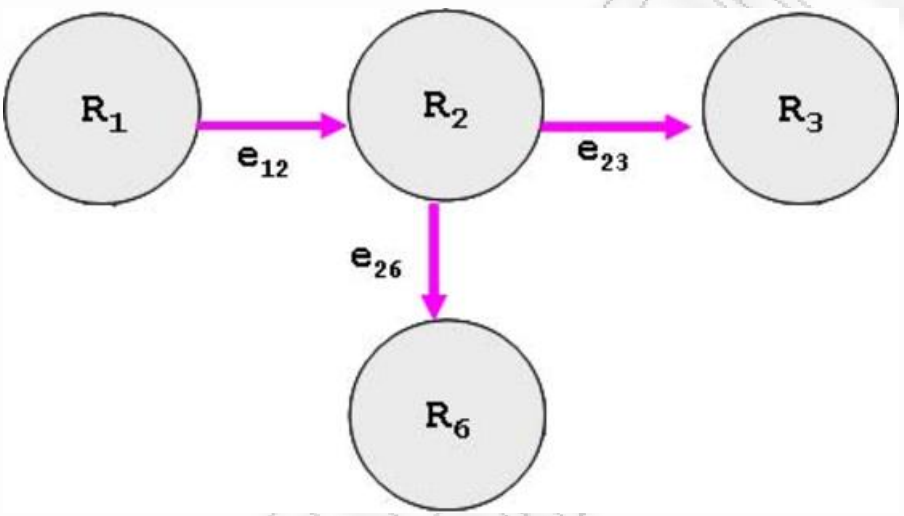
Ορισμός 4-9 (Διάσπαση κυκλοφορίας): Μια ακμή *e* διασπά την κυκλοφορία της κατά τη διάρκεια της *p* στις ακμές $e_i (i = 1 : k)$ κατά τη διάρκεια της *p'* ανν:

i) $end(e) = start(e_i)$,

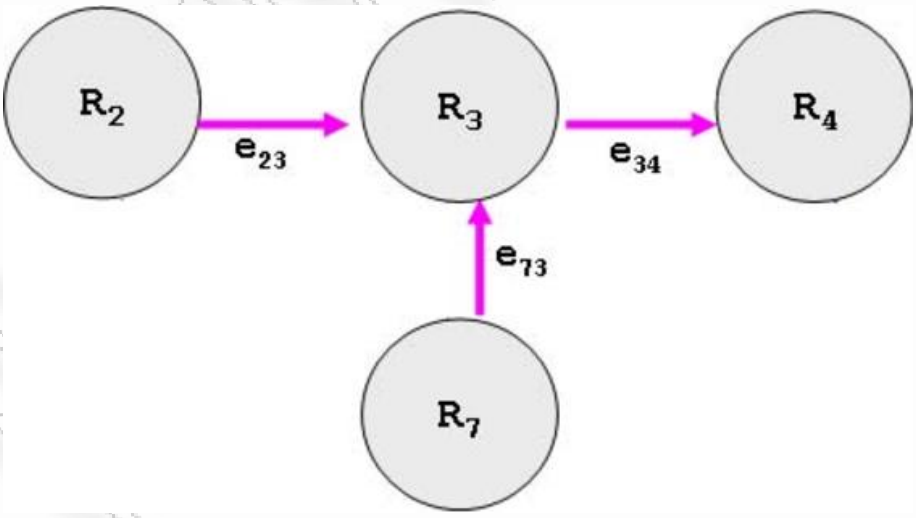
ii) $dis_{value}\left(TS_e^p, \sum_i (TS_{e_i}^{p'})\right) \geq t_v$ και

iii) $\forall\, 1 \leq i \leq k : dis_{shape}\left(TS_e^p, TS_{e_i}^p\right) \geq t_s$. ■

δηλ., η κυκλοφορία στις ακμές $e_1, \ldots, e_k$ κατά τη διάρκεια της *p'* ως σύνολο «σχηματίζει» την κίνηση της ακμής *e* κατά τη διάρκειας της *p* αλλά και το σχήμα των χρονοσειρών τους μέσα στο παράθυρο, είναι παρόμοιο (Εικόνα 4-20).



Εικόνα 4-20: Η κυκλοφορία της ακμής e_{12} διασπάται στις ακμές e_{23} και e_{26}.



Εικόνα 4-21: Η κυκλοφορία της ακμής e_{34} είναι το αποτέλεσμα της εισερχομένης κυκλοφορίας από τις ακμές e_{23} και e_{73}.

Ορισμός 4-10 (Συγχώνευση κυκλοφορίας): Η κυκλοφορία μια ακμής *e* κατά τη διάρκεια της περιόδου *p* είναι το αποτέλεσμα συγχώνευσης κυκλοφορίας από τις ακμές $e_i (i = 1 : k)$ στην *p'* ανν:

i) $end(e_i) = start(e)$,

ii) $dis_{value}\left(\sum_i (TS_{e_i}^p), TS_e^{p'}\right) \geq t_v$ και

iii) $\forall\, 1 \leq i \leq k : dis_{shape}\left(TS_{e_i}^p, TS_e^p\right) \geq t_s$. ■

δηλ., κατά τη διάρκεια της p οι ακμές $e_1, \ldots, e_k$ «σχηματίζουν» την κυκλοφορία της e.

Εναλλακτικά, μπορούμε να πούμε ότι η κυκλοφορία των e_i κατά τη διάρκεια της p (Εικόνα 4-21) συγχωνεύεται στην κυκλοφορία της e κατά τη p' και το σχήμα των χρονοσειρών, μέσα στο χρονικό παράθυρο, είναι παρόμοιο. Η σχέση της συγχώνευσης είναι η αντίθετη της σχέσης της διάσπασης.

Ορισμός 4-11 (Προορισμός κυκλοφορίας): Μια ακμή e χαρακτηρίζεται ως προορισμός κυκλοφορίας κατά τη διάρκεια της p ανν δε σχετίζεται με κάποια σχέση διάδοσης ή διάσπασης με τις εξερχόμενες ακμές της κατά τη διάρκεια αυτής της περιόδου. ■

Δηλαδή η κυκλοφορία της δεν μεταφέρεται σε κάποια από τις εξερχόμενες ακμές κατά τη διάρκεια αυτής της περιόδου.

Ορισμός 4-12 (Πηγή κυκλοφορίας): Μια ακμή e χαρακτηρίζεται ως πηγή κυκλοφορίας κατά τη διάρκεια της p ανν κατά τη διάρκεια αυτής της περιόδους δεν σχετίζεται με κάποια σχέση διάδοσης ή συγχώνευσης με τις εισερχόμενες ακμές. ■

Δηλαδή, η κυκλοφορία της δεν «δικαιολογείται» από τις εισερχόμενες ακμές της κατά τη διάρκεια αυτής της περιόδου.

Στους παραπάνω ορισμούς, ορίσαμε μια σχέση μεταξύ των ακμών e_1, e_2 κατά τη διάρκεια μια συγκεκριμένης περιόδου p, δηλ. με ποιον τρόπο σχετίζεται η e_1 με την e_2 κατά τη διάρκεια της p. Σημειώστε όμως ότι λόγω της καθυστέρησης λόγω της πορείας από την e_1 στην e_2 μια τέτοια σχέση μπορεί να ισχύει για διαφορετικές περιόδους. Η πιο γενική περίπτωση είναι ότι η κυκλοφορία της e_1 κατά τη διάρκειας της p_i σχετίζεται με την κυκλοφορία της e_2 κατά τη διάρκεια της p_j , όπου p_i, p_j πρέπει να είναι κοντινές χρονικές περίοδοι με $j - i \leq w$, όπου w είναι το μέγιστο επιτρεπόμενο κενό. Στην έρευνά μας θεωρούμε ότι $w = 0$ (δηλ. $p_j = p_i$, οπότε μιλάμε για την ίδια χρονική περίοδο) ή $w = 1$ (δηλ. $p_j = p_i+1$, αντιστοιχώντας σε διαδοχικές χρονικές περιόδους).

Από τους ανωτέρω ορισμούς, είναι σαφές ότι στην κυκλοφορία μιας ακμής μπορεί να αποδοθεί ένας από τους ακόλουθους χαρακτηρισμούς: διαδεδομένη, διασπασμένη σε, συγχωνευμένη προς, προορισμός, πηγή. Εντούτοις, υπάρχουν σύνθετες περιπτώσεις όπου δεν μπορούμε να αποφασίσουμε πραγματικά για την ακριβή σχέση μεταξύ δύο ακρών.

4.4.3.3. Ανακαλύπτοντας σχέσεις

Σε αυτήν την υποενότητα, περιγράφουμε την προσέγγισή μας για την ανακάλυψη των σχέσεων μεταξύ των ακμών του γράφου. Ο αλγόριθμος μας TRAFFIC-RELATIONSHIP-DETECTOR που απεικονίζεται στην Εικόνα 4-22 και ως είσοδο χρειάζεται το δίκτυο κυκλοφορίας $G = (V, E)$ και τα κατώφλια για τα μέτρα ομοιότητας τιμής και σχήματος καθώς επίσης και το μέγεθος του χρονικού παραθύρου μέσα στο οποίο αναζητούμε ομοιότητα σχήματος χρονοσειρών.

Ο αλγόριθμος αρχίζει με ένα τυχαίο κόμβο του δικτύου και ανακαλύπτει πώς σχετίζονται οι εισερχόμενες και εξερχόμενες ακμές της από την άποψη της κυκλοφορία εντός μιας περιόδου παρατήρησης. Στην αρχή, ψάχνει τις περιπτώσεις όπου μόνο είτε σχέση διάδοσης, διάσπασης είτε συγχώνευσης μπορεί να βρεθεί (γραμμές 4-15). Οι συνάρτησεις defineOutSetEdges (γραμμή 9) και defineInSetEdges (γραμμή 14) επιλέγουν το υποσύνολο του εξερχόμενων και εισερχόμενων ακμών που θα ελεγχθούν για τις σχέσεις διάσπασης και συγχώνευσης αντίστοιχα. Χρησιμοποιούμε μια ευριστική προσέγγιση που επιλέγει ένα υποσύνολο των εξερχόμενων (εισερχόμενων) ακμών ώστε να εξασφαλίσουμε τουλάχιστον την ομοιότητα τιμής με την εισερχόμενη (εξερχόμενη) ακμή. Όσον αφορά στις συναρτήσεις CHECKFORPROPAGATE, CHECKFORSPLIT και CHECKFORMERGE είναι βασισμένοι στους ορισμούς που δίνονται στην προηγούμενη υποενότητα. Ας σημειωθεί ότι κάθε μια από αυτές αναζητά σχέσεις είτε στην ίδια είτε σε διαδοχική περίοδο.

Σε πιο σύνθετες περιπτώσεις (γραμμές 16-27), ο αλγόριθμος δημιουργεί τα πιθανά ζευγάρια των εισερχόμενων και εξερχόμενων ακμών και ψάχνει όλους τους τύπους σχέσεων. Εάν βρεθούν περισσότεροι της μιας δε μπορεί να αποφασίσει και χαρακτηρίζει αυτό το ζευγάρι ως *σύνθετο*. Τέλος, ο αλγόριθμος αναζητά για πηγές και προορισμούς ακολουθώντας τη λογική που περιγράψαμε στην προηγούμενη υποενότητα: εάν για μια συγκεκριμένη περίοδο μια ακμή συμμετέχει ως δεξί μέρος σε μια σχέση (δηλ. η κυκλοφορία διαδίδεται, συγχωνεύεται, διασπάται σε αυτήν) αλλά όχι στο αριστερό μέρος μιας άλλης σχέσης τότε θεωρείται ως προορισμός (γραμμές 31-33). Εάν συμβαίνει το αντίθεταο θεωρείται ως *πηγή* (γραμμές 34-37).

```
Algorithm Traffic-Relationship-Detector(ListOfVertices lov,
ValueSimThreshold t_v, ShapeSimThreshold t_s, TimeInterval w_b, TimeInterval
w_a)
1. FOR EACH Vertice v IN lov
2.   FOR EACH Period p
3.     //check incoming and outgoing edges for v
4.     IF incomingEdges(v) = 1 AND outgoingEdges(v) = 1 THEN
5.       checkForPropagate(v.e_in, v.e_out, p, t_v, t_s, w)
6.     ELSE IF incomingEdges(v) = 1 AND outgoingEdges(v) > 1 THEN
7.       //it is defined the set of outgoing edges to look for a
8.       //relationship between the set and the incoming edge
9.       le = defineOutEdges(v)
10.      checkForSplit(v.e_in, le, p, t_v, t_s, w)
11.    ELSE IF incomingEdges(v) > 1 AND outgoingEdges(v) = 1 THEN
12.      //it is defined the set of incoming edges to look for a
13.      //relationship between the set and the outgoing edge
14.      le = defineInSetEdges(v)
15.      checkForMerge(le, v.e_out, p, t_v, t_s, w)
16.    ELSE
17.      FOR EACH pair pr IN v
18.        rels = checkForPropagate(pr.e, pr.e', p, t_v, t_s, wb)
19.        rels += checkForSplit(pr.e, pr.e_i, p, t_v, t_s, w)
20.        rels += checkForMerge(pr.e_i, pr.e, p, t_v, t_s, w)
21.        //rels stores the number of total relationships found
22.        IF rels >1 THEN
23.        //a complex relationship is found
24.        complexRel(pr, p)
```

```
25.      END IF
26.     END FOR
27.    END IF
28.   END FOR
29.    //consider as r the set of discovered relationships
30.    FOR EACH Edge e IN v
31.      IF v.e IN rightPart(r) AND NOT IN leftPart(r) THEN
32.       //the edge keeps its traffic at period p
33.       sinkEdge(e, p)
34.      ELSE IF v.e in leftPart(r) AND NOT IN rightPart(r) THEN
35.       //the edge keeps its traffic at period p
36.       sourceEdge(e, p)
37.      END IF
38.     END FOR
39. END FOR

Function checkForPropagate (Edge e, Edge e', Period p,
ValueSimThreshold t_v, ShapeSimThreshold t_s, TimeInterval w_b, TimeInterval
w_a)
1.  //let TS_e^p be the part of the timeseries of edge e during p
2.  IF dis_shape(TS_e^p, TS_e'^p) ≥ t_s THEN
3.   IF dis_value(TS_e^p, TS_e'^p) ≥ t_v THEN
4.    //a propagation relationship is found
5.    propagateRel(e, e', p)
6.   ELSE IF dis_value(TS_e^p, TS_e'^(p+1)) ≥ t_v THEN
7.    //a propagation relationship is found
8.    propagateRel(e, e', p+1)
9.   END IF
10. END IF

Function checkForSplit (Edge e, ListOutEdges e_i, Period p,
ValueSimThreshold t_v, ShapeSimThreshold t_s, TimeInterval w_b, TimeInterval
w_a)
1.  //let TS_e^p be the part of the timeseries of edge e during p
2.  // e_i(i = 1 : k) where k is the number of outgoing edges
3.  IF ∀ 1 ≤ i ≤ k : dis_shape(TS_e^p, TS_ei^p) ≥ t_s THEN
4.   IF dis_value(TS_e^p, ∑_i(TS_ei^p)) ≥ t_v AND THEN
5.    //a split relationship is found
6.    splitRel(e, e_i, p)
7.   ELSE IF dis_value(TS_e^p, ∑_i(TS_ei^(p+1))) ≥ t_v THEN
8.    //a split relationship is found
9.    splitRel (e, e_i, p+1)
10.  END IF
11. END IF

Function checkForMerge (ListInEdges e_i, Edge e, Period p,
ValueSimThreshold t_v, ShapeSimThreshold t_s, TimeInterval w_b, TimeInterval
w_a)
1.  //let TS_e^p be the part of the timeseries of edge e during p
2.  //e_i(i = 1 : k) where k is the number of incoming edges
3.  IF ∀ 1 ≤ i ≤ k : dis_shape(TS_ei^p, TS_e^p) ≥ t_s THEN
```

```
4.    IF dis_value(∑_i(TS_{e_i}^p), TS_e^p) ≥ t_v AND  THEN
5.     //a merge relationship is found
6.     mergeRel (e_i, e, p)
7.    ELSE IF dis_value(∑_i(TS_{e_i}^{p+1}), TS_e^p) ≥ t_v THEN
8.     //a merge relationship is found
9.     mergeRel (e_i, e, p+1)
10.   END IF
11. END IF
```

Εικόνα 4-22: Ο αλγόριθμος TRAFFIC-RELATIONSHIP-DETECTOR.

4.4.4. Πειραματική Μελέτη

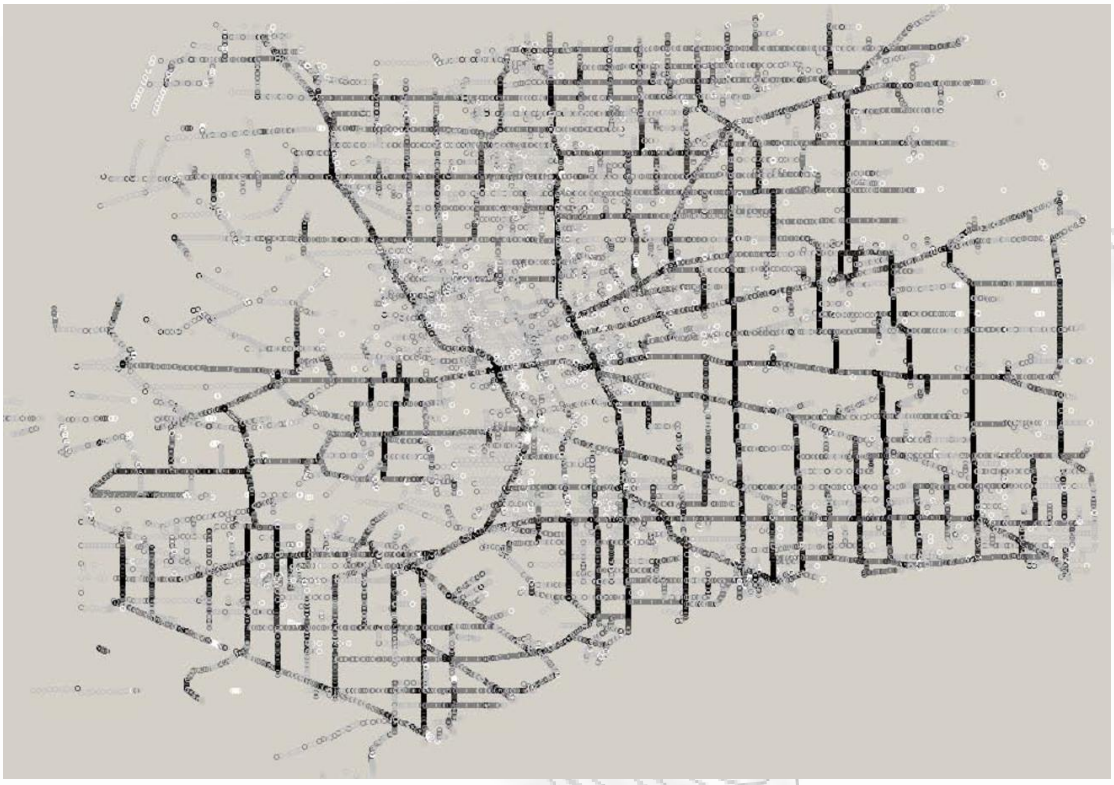
Τα πειράματα που παρουσιάζονται σε αυτήν την υποενότητα στοχεύουν να επιδείξουν τη δυνατότητα εφαρμογής και την χρησιμότητα των προσεγγίσεών μας. Πιο συγκεκριμένα, πειραματιζόμαστε με τους αλγορίθμους TRAFFIC-CLUSTERING και TRAFFIC-RELATIONSHIP-DETECTOR.

4.4.4.1. Συσταδοποίηση ακμών

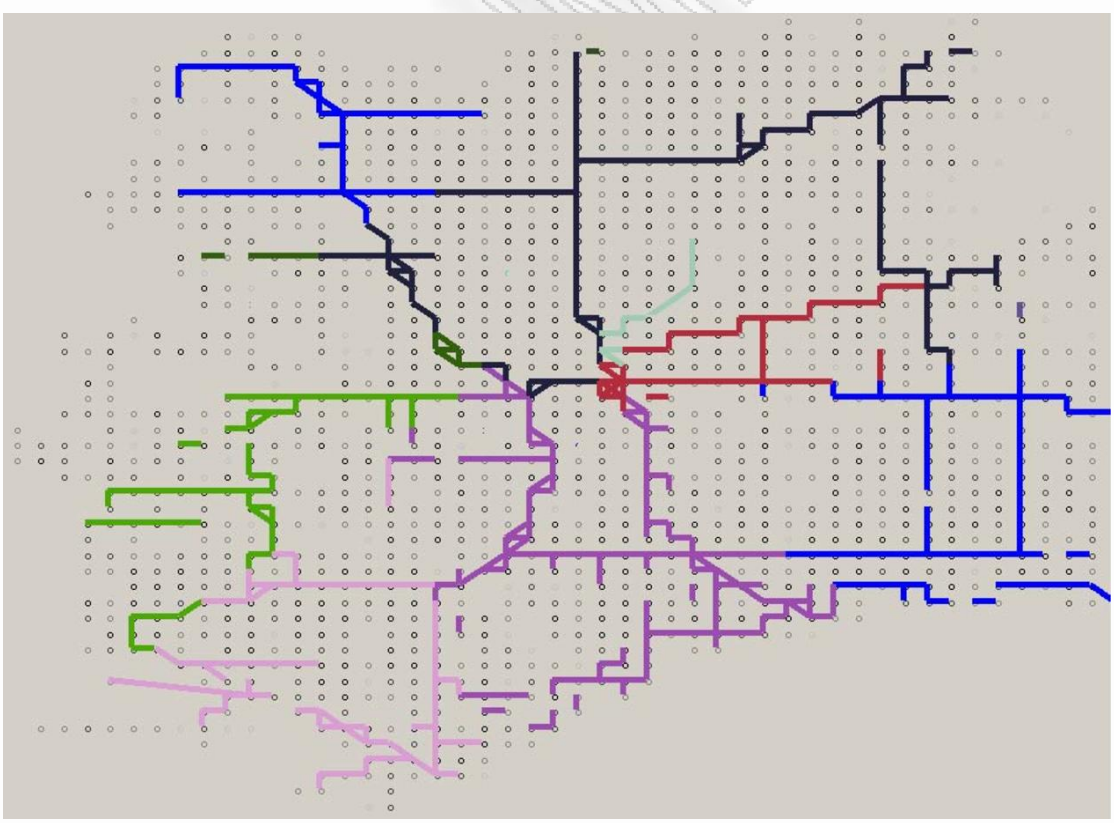
Χρησιμοποιήσαμε συνθετικά δεδομένα που παρήχθησαν από γεννήτρια δεδομένων βασισμένων σε δίκτυο του Brinkhoff [Bri02]. Συγκεκριμένα, παραγάγαμε 2000 κινούμενα αντικείμενα με 400 θέσεις δειγματοληψίας για κάθε αντικείμενο. Το δίκτυο που χρησιμοποιήθηκε στα πειράματα μας απεικονίζεται στην Εικόνα 4-23.

Προτού συνεχίσουμε με τα πειραματικά αποτελέσματα, πρέπει να σημειώσουμε ότι δεν εξετάζουμε την καθυστέρηση της μετάβασης από μια ακμή *e* στις εξερχόμενες ακμές της. Υποθέτουμε ότι ο χρόνος που απαιτείται για να διασχίσει ένα αντικείμενο την ακμή *e* είναι μικρότερος από τη διάρκεια της χρονικής περιόδου μέσα στην οποία εντοπίζουμε τις σχέσεις. Επίσης, χρησιμοποιήσαμε ένα κατώφλι ώστε να αποκλειστούν από τη διαδικασία συσταδοποίησης οι ακμές τις οποίες επισκέφτηκαν σπάνια κατά τη διάρκεια της περιόδου παρατήρησης τα κινούμενα αντικείμενα . Στην Εικόνα 4-24, παρουσιάζουμε τις συστάδες που συλλέχθηκαν στο πρώτο βήμα του αλγορίθμου συσταδοποίησης, όπου η ομαδοποίηση των ακμών είναι βασισμένη στην ομοιότητα σχήματος κυκλοφορίας τους. Οι ακμές με το ίδιο χρώμα δείχνουν τις περιοχές δικτύων που μοιράζονται παρόμοιο σχήμα κυκλοφορίας. Όπως μπορούμε να παρατηρήσουμε σε αυτήν την εικόνα, υπάρχουν περιοχές στο δίκτυο που μοιράζονται το ίδιο σχήμα κυκλοφορίας, ακόμα κι αν δε συνδέονται. Παραδείγματος χάριν, η συστάδα που απεικονίζεται με το μπλε χρώμα αποτελείται από τρεις υποσυστάδες που βρίσκονται στις διαφορετικές θέσεις του δικτύου. Αυτό σημαίνει ότι υπάρχουν τρεις μη συνδεμένες περιοχές στο δίκτυο που τις μοιράζονται το ίδιο σχήμα κυκλοφορίας και έτσι, όλες μαζί, διαμορφώνουν μια συστάδα (μπλε).

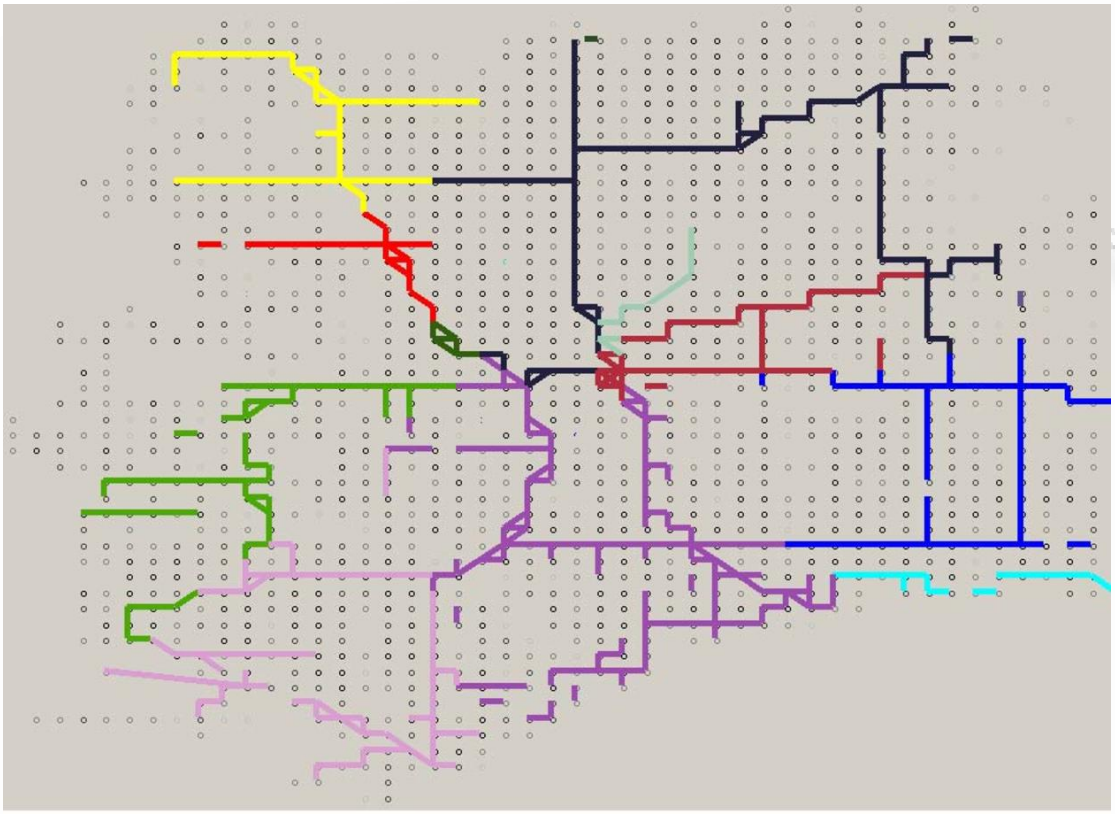
Στην Εικόνα 4-25, παρουσιάζουμε τις συστάδες που παρήχθησαν από το δεύτερο βήμα του αλγορίθμου συσταδοποίησης, όπου η ομαδοποίηση των ακμών εξετάζει επίσης την εγγύτητά τους στη δομή του γράφου του δικτύου. Συγκρίνοντας με το πρώτο βήμα του αλγορίθμου (βλ. Εικόνα 4-24), μπορούμε να παρατηρήσουμε ότι παλαιές συστάδες διασπάστηκαν σε νέες συστάδες, λόγω του μέτρου απόστασης βασισμένου στην εγγύτητα που εφαρμόζεται σε αυτό το βήμα.



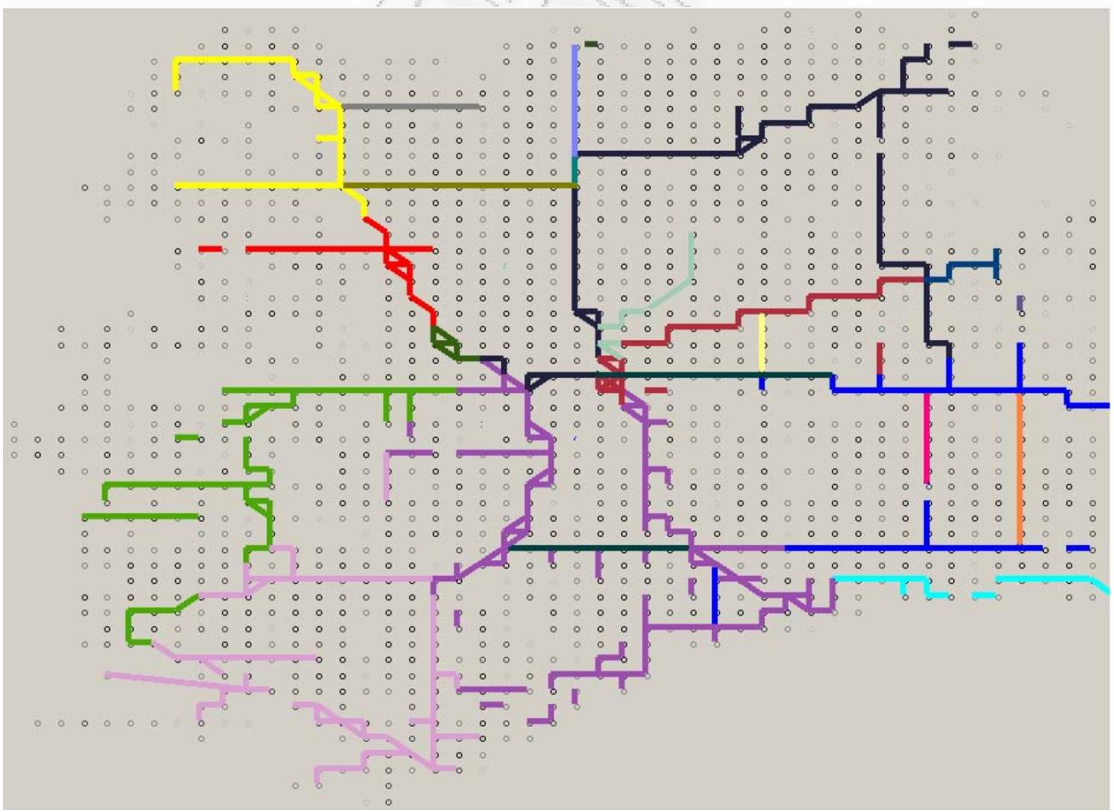
Εικόνα 4-23: Το αρχικό δίκτυο κυκλοφορίας.



Εικόνα 4-24: Αποτελέσματα του Επιπέδου 1 που βασίζονται στις ομαδοποιήσεις των ακμών με παρόμοιο σχήμα κυκλοφορίας (οι συστάδες απεικονίζονται με διαφορετικά χρώματα).



Εικόνα 4-25: Αποτελέσματα του Επιπέδου 1 που βασίζονται σε ομαδοποίησης βάση της εγγύτητας (οι συστάδες απεικονίζονται με διαφορετικά χρώματα).



Εικόνα 4-26: Αποτελέσματα του Επιπέδου 3 που βασίζονται στην ομαδοποίηση των ακμών με παρόμοιες τιμές κυκλοφορίας (οι συστάδες απεικονίζονται με διαφορετικά χρώματα

Για παράδειγμα, η μπλε συστάδα που περιγράφηκε στην πρώτη φάση, αντικαταστάθηκε από τρεις νέες συστάδες (κίτρινη, μπλε και τυρκουάζ μπλε). Προφανώς, αυτή η διάσπαση πραγματοποιήθηκε επειδή, όπως εξηγείται νωρίτερα, η αρχική συστάδα που περιελάμβανε ασύνδετες περιοχές, οι οποίες τοποθετήθηκαν από το μέτρο απόστασης που εφαρμόστηκε σε αυτό το βήμα σε διαφορετικές συστάδες.

Τέλος, στην Εικόνα 4-26, παρουσιάζουμε το αποτέλεσμα του τρίτου βήματος του αλγορίθμου συσταδοποίησης, ο οποίος κοιτάζει περαιτέρω για ακμές κυκλοφορίας με παρόμοιες τιμές. Μπορούμε να δούμε ότι πολλές νέες συστάδες έχουν προκύψει. Οι νέες συστάδες δείχνουν τις περιοχές που μοιράζονται τις ίδιες τιμές κυκλοφορίας.

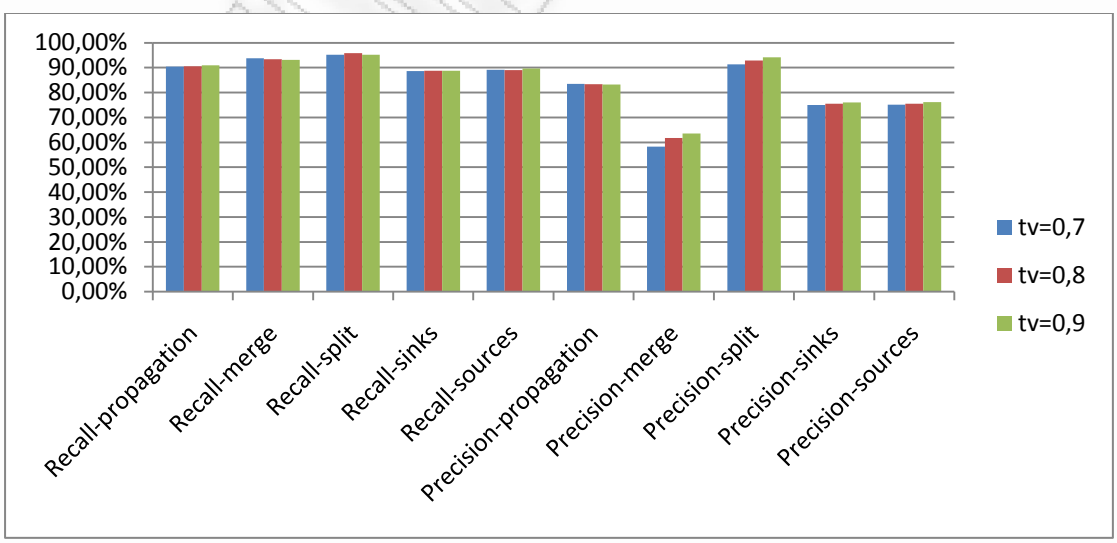
Όπως δείξαμε στα πειράματα, η προσέγγισή μας μπορεί να παρέχει τις χρήσιμες ιδέες στο πρόβλημα κυκλοφορίας, ενώ η τριών επιπέδων αρχιτεκτονική διευκολύνει τον τελικό χρήση και του επιτρέπει να παρακολουθήσει την κυκλοφορία σε διαφορετικά αφαιρετικά επίπεδα.

4.4.4.2. Ανακαλύπτοντας σχέσεις προσδιορισμένες στο χρόνο

Χρησιμοποιήσαμε ένα πραγματικό σύνολο δεδομένων που αποτελείται από 59263 θέσεις που ανήκουν σε 990 αντικείμενο που κινούνται στην ευρύτερη περιοχή του Μιλάνου και αντιστοιχίζονται σε 9256 ακμές. Πειραματιζόμαστε με διάφορα μέτρα ομοιότητας και υπολογίζουμε την ακρίβεια (precision) και την ανάκληση (recall):

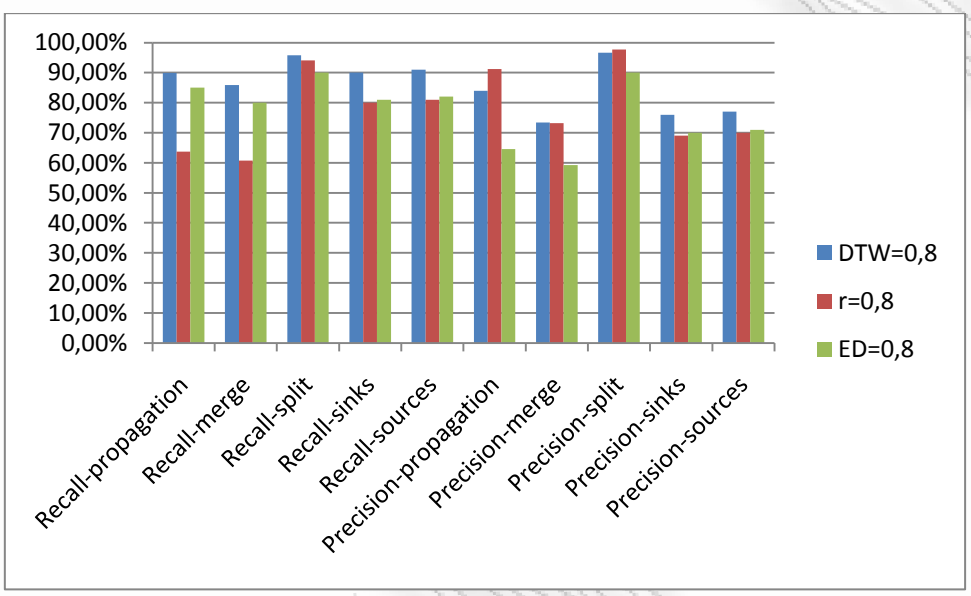
- Ακρίβεια = αληθή θετικά / (αληθή θετικά + ψευδή θετικά)
- Ανάκληση = αληθή θετικά / (αληθή θετικά + ψευδή αρνητικά)

Στην Εικόνα 4-27, εξετάζουμε διαφορετικές τιμές του κατωφλιού ομοιότητας τιμής και υπολογίζουμε ακρίβεια/ανάκληση για κάθε μια από τις πιθανές σχέσεις (διάδοση, συγχώνευση, διάσπαση, πηγή, προορισμός). Παρατηρούμε ότι η ακρίβεια στον εντοπισμό των διαφόρων σχέσεων είναι μεγάλη ακόμα και όταν χρησιμοποιούμε μόνο το φίλτρο της ομοιότητας τιμής.



Εικόνα 4-27: Ακρίβεια/ανάκληση κατά την εφαρμογή μόνο του φίλτρου ομοιότητας τιμής.

Στην Εικόνα 4-28, εφαρμόζουμε διαφορετικά μέτρα ομοιότητας σχήματος μέσα στο χρονικό παράθυρο *(p-5,p+5)*. Αυτό που προκύπτει από αυτό το πείραμα είναι όχι η χρήση του DTW φίλτρου μπορεί να μας δώσει καλύτερα αποτελέσματα σε σύγκριση με το Correlation-Coefficient φίλτρο ακαι φυσικά πολύ καλύτερα από το φίλτρο της απλής Ευκλείδειας απόστασης. Παρατηρούμε μια μικρή βελτίωση στην ακρίβεια εντοπισμού των σχέσεων συγχώνευσης που ήταν χαμηλά όταν χρησιμοποιούσαμε μόνο το φίλτρο ομοιότητας τιμής.



Εικόνα 4-28: Ακρίβεια/ανάκληση κατά την εφαρμογή ομοιότητας τιμής και σχήματος στο χρονικό παράθυρο *(p-5,p+5)*.

4.5. Σχετικές Εργασίες

Πρόσφατα, έχουν υπάρξει αρκετές ερευνητικές προσπάθειες για την επέκταση των παραδοσιακών τεχνικών εξόρυξης γνώσης στο πλαίσιο των τροχιών (δείτε [KNK+07] για μια περιεκτική επισκόπηση).

4.5.1. Συσταδοποίηση Τροχιών

Ένα ερευνητικό αντικείμενο σχετικό με την έρευνα μας είναι η *χωροχρονική συσταδοποίηση* ή *συσταδοποίηση τροχιών* που στοχεύει στην ομαδοποίηση παρόμοιων τροχιών κινούμενων αντικειμένων.

Σχετικά με την προσέγγισή μας είναι η εργασία [LHL+07] για την ανακάλυψη των συνηθισμένων διαδρομών (hot routes) σε ένα οδικό δίκτυο. Οι συνηθισμένες διαδρομές αντιστοιχούν σε ακολουθίες οδικών τμημάτων με μεγάλο όγκο κυκλοφορίας. Τα οδικά τμήματα που συμμετέχουν στο σχηματισμό μιας συνηθισμένης διαδρομής πρέπει να μοιράζονται κάποια κοινή κυκλοφορία και πρέπει επίσης να είναι κοντινά. Οι συγγραφείς προτείνουν έναν αλγόριθμο βασισμένο στην πυκνότητα, τον αποκαλούμενο FlowScan, για να ανακαλύψουν τις συνηθισμένες διαδρομές, ο οποίος συσταδοποιεί τα οδικά τμήματα βάση της πυκνότητας της κοινής κυκλοφορίας που μοιράζονται. Ο αλγόριθμος, εντούτοις, απαιτεί τις τροχιές των αντικειμένων που κινούνται στο δίκτυο, κατά συνέπεια δεν μπορεί να εφαρμοστεί στις συνθήκες τους προβλήματός μας.

Οι Lee κ.α. [LHW07] προτείνουν ένα πλαίσιο για διαχωρισμό-και-ομαδοποίηση για συσταδόποιηση τροχιών το οποίο δεν εξετάζει τη συσταδοποίηση ολόκληρων των τροχιών, αλλά συγκεντρώνει τις κοινές υποτροχιές. Στο βήμα του διαχωρισμού, μια τροχιά είναι χωρίζεται σε ένα σύνολο τμημάτων γραμμών που χρησιμοποιούν την αρχή της Minimum Description Length (MDL). Στο βήμα ομαδοποίησης, συσταδοποιούνται τα παρόμοια τμήματα γραμμών χρησιμοποιώντας μια μέθοδοο συσταδοποίησης που βασίζεται στην πυκνότητα. Για κάθε συστάδα, ανακαλύπτεται η αντιπροσωπευτική τροχιά που ορίζεται ως η τροχιά που περιγράφει τη γενική κίνηση των τμημάτων τω τροχιών που ανήκουν στην ίδια συστάδα. Σε αυτήν την εργασία, εντούτοις, απαιτούνται οι τροχιές των κινούμενων αντικειμένων, κατά συνέπεια η λύση αυτή δεν μπορεί να μεταφερθεί στις συνθήκες του προβλήματός μας. Επιπλέον, αυτή η εργασία θεωρεί ελεύθερη μετακίνηση και συνεπώς όχι κάποιο προκαθορισμένο δίκτυο όπως, στην περίπτωσή μας, το οδικό δίκτυο.

Κάτι παρόμοιο προτείνουν οι Giannotti κ.α. [GNP+07] όπου εισάγουν την έννοια των προτύπων τροχιάς (Trajectory Patterns - T-patterns) και προτείνουν, για την ανακάλυψη τους, κατάλληλους αλγορίθμους εξόρυξης γνώσης από τροχιές. Τα πρότυπα τροχιάς δεν αναπαριστούν πραγματικές τροχιές αλλά ακολουθίες χωρικών περιοχών που συσχετίζονται χρονικά. Τέτοιες περιοχές ενδιαφέροντος μπορούν να προκαθοριστούν από τον χρήση ή μπορούν να ανακαλυφθούν με δυναμικό τρόπο χρησιμοποιώντας κάποιον αλγόριθμο που βασίζεται στην πυκνότητα.

Οι Kalnis κ.α. [KMB05] εισάγουν την έννοια των κινούμενων συστάδων για την ανακάλυψη ομάδων αντικειμένων που κινούνται κοντά το ένα με το άλλο για μακρύ χρονικό διάστημα. Σε κάθε χρονική στιγμή, τα αντικείμενα ομαδοποιούνται χρησιμοποιώντας έναν παραδοσιακό αλγόριθμο συσταδοποίησης και στη συνέχεια ανακαλύπτονται οι κινούμενες συστάδες ανιχνεύοντας τις καταγραφές κοινών δεδομένων μεταξύ συστάδων διαδοχικών στιγμών. Εντούτοις η μέθοδός τους απαιτεί τα IDs των αντικειμένων, και συνεπώς δεν ταιριάζει στην περίπτωσή μας όπου είναι διαθέσιμος μόνο ο αριθμός αντικειμένων που περνά μέσω κάποια ακμή/οδικό τμήμα. Επιπλέον, η μέθοδός δεν εξετάζει την περιορισμένη σε δίκτυο μετακίνηση, ενώ η έρευνα μας αφορά αντικείμενα που κινούνται σε ένα προκαθορισμένο οδικό δίκτυο. Επιπλέον, οι συγγραφείς εξετάζουν μόνο την ανακάλυψη μιας συστάδας των αντικειμένων την επόμενη χρονική στιγμή ενώ εμείς ανακαλύπτουμε επιπλέον σχέσεις όπως η διάσπαση και η συγχώνευση.

Επίσης, σχετική με την προσέγγιση μας είναι η έρευνα που έχει γίνει σε θέματα *ανίχνευσης μεταβολής* (change detection). Για παράδειγμα, στην Spiliopoulou κ.α. [SNT+06] προτείνουν το πλαίσιο MONIC για την μοντελοποίηση και ανακάλυψη μεταβάσεων μεταξύ συστάδων που ανακαλύπονται σε διαδοχικές χρονικές στιγμές. Το MONIC εξετάζει τόσο εξωτερικές (π.χ., επιβίωση, διάσπαση, απορρόφηση) όσο και εσωτερικές μεταβάσεις (π.χ., αλλαγή στο μέγεθος, αλλαγή στη τοποθεσία). Όμως, η μέθοδος τους βασίζεται στα μέλη των συστάδων οπότε δε μπορεί να εφαρμοστεί στο δικό μας πρόβλημα αφού εμείς διαθέτουμε μόνο τον αριθμό των αντικειμένων που πέρασαν από κάποια ακμή του δικτύου και όχι τα IDs αυτών των αντικειμένων.

Οι Li κ.α. [LHK06] εξετάζουν το πρόβλημα της ανίχνευσης ανωμαλιών στην κίνηση των αντικειμένων. Αντί να εστιάζουν σε τροχιές, προτείνουν μεθόδους δημιουργίας εκφράσεων μοτίβου (motif expressions) με σκοπό να λειάνουν τις χωρικές και χρονικές κλιμακώσεις και συνδυάζουν αυτήν

την πληροφορία με άλλα συμβατικά ή χωροχρονικά γνωρίσματα έτσι ώστε να διευκολύνουν την διαδικασία εξόρυξης γνώσης. Εντούτοις οι συγγραφείς στηρίζονται σε πληροφορίες που εξάγονται από τα δεδομένα τροχιάς, κατά συνέπεια το εξεταζόμενο πρόβλημα είναι διαφορετικό από το δικό μας.

4.5.2. Κατηγοριοποίηση Τροχιών

Οι Lee κ.α.. [LHL+08] πρότειναν τον *TraClass* για κατηγοριοποίηση τροχιών δείχνοντας ότι είναι απαραίτητο και σημαντικό να εφαρμόζεται η εξόρυξη γνώσης σε τμήματα των τροχιών και όχι στο σύνολο τους. Για αυτόν τον σκοπό χρησιμοποιούν αλγορίθμους που βασίζονται σε τροχιές και σε περιοχές ώστε να πετύχουν το διαχωρισμό.

Οι περισσότερες από τις προσεγγίσεις (π.χ., [Doc06]) εφαρμόζουν παραλλαγές των Markov μοντέλων για να περιγράψουν τη κίνηση των τροχιών, και συγκρίνονται με κάθε συστατικό του μοντέλου για να το κατατάξουν σε μια από τις κατηγορίες. Οι Keogh και Pazzani [KP98] χρησιμοποιούν μια τμηματική γραμμική αναπαράσταση της χρονοσειράς και σταθμίζουν κάθε τμήμα σύμφωνα με τη σημασία του. Αυτή η αναπαράσταση χρησιμοποιείται για κατηγοριοποίηση, συσταδοποίηση και εύρεση συσχετίσεων. Στην [Geu01] οι χρονοσειρές κατηγοριοποιούνται θεωρώντας τα προτυπα ως κριτήρια ελέγχου στα δέντρα απόφασης. Με αυτόν τον τρόπο, κάθε πρότυπο αντιστοιχεί σε ένα χρονικά περιορισμένο, σταθερό μοντέλο που μπορεί, για παράδειγμα, να αντιπροσωπεύει την ταχύτητα ενός αντικειμένου.

Γενικά, οι τροχιές μπορούν να κατηγοριοποιηθούν χρησιμοποιώντας τους αλγορίθμους κοντινότερου γείτονα υπό τον όρο ότι δίνεται μια κατάλληλη συνάρτηση απόστασης. Εντούτοις, ο καθορισμός μιας τέτοιας συνάρτησης εξαρτάται από το στόχο κατηγοριοποίησης και δεν είναι εύκολος αφού θα πρέπει να ληφθούν υπόψη οι διαφορετικές κλίμακες, ο θόρυβος κτλ.

Όλες οι προσεγγίσεις που αντιμετωπίζουν αυτό το πρόβλημα χρησιμοποιούν περίπλοκες τεχνικές που όμως δεν λαμβάνουν υπόψη τους παράγοντες που σχετίζονται με τη βαθύτερη φύση τέτοιων δυναμικών συστημάτων όπως οι γνωστικές πτυχές (cognitive aspects) καθώς και οι περιορισμοί και οι αλληλεξαρτήσεις μεταξύ των οντοτήτων και των αλληλεπιδράσεων τους.

Οι στενή σχέση μεταξύ των τοπικών προτύπων (local patterns) και των σφαιρικών μοντέλων (global models) έχουν αποτελέσει το αντικείμενο πρόσφατων ερευνών. Για παράδειγμα, τα συχνά τοπικά πρότυπα έχουν εφαρμοστεί στα προβλήματα κατηγοριοποίησης σε γράφους [DKW+05], όπου χρησιμοποιούνται αλγόριθμοι ανακάλυψης συχνών υπογράφων (subgraphs) προκειμένου να βρεθούν όλες οι τοπολογικές και γεωμετρικές υποδομές που παρουσιάζονται στο σύνολο δεδομένων, οι οποίες χρησιμοποιούνται στη συνέχεια ως γνωρίσματα στα πλαίσια μιας διαδικασία κατηγοριοποίησης. Η χρησιμότητα των συχνών προτύπων ως βασικά γνωρίσματα για την κατηγοριοποίηση έχει ερευνηθεί πρόσφατα στην [CYH+07] για την περίπτωση των συχνών στοιχειοσυνόλων, που δείχνουν ότι η συχνότητα είναι ένας αποτελεσματικός τρόπος χαρακτηρισμού των γνωρισμάτων που βασίζονται σε πρότυπα.

4.5.3. Ανάλυση Κυκλοφορίας

Οι Shekhar κ.α. [SLC+01] συζητούν την εφαρμογή παραδοσιακών τεχνικών αποθήκευσης δεδομένων και εξόρυξης γνώσης στις μετρήσεις ενός δικτύου αισθητήρων που συλλέγονται από

αυτοκινητόδρομους. Επιπλέον, προτείνουν ερευνητικές κατευθύνσεις στην ανίχνευση ακραίων τιμών (outliers) κυκλοφοριακής ροής, ανακάλυψη των χωροχρονικών κανόνων συσχέτισης και των ακολουθιακών προτύπων (sequential patterns). Εντούτοις, αυτή η εργασία δεν εξετάζει λεπτομερώς το πρόβλημα κυκλοφορίας. Οι συγγραφείς παρουσιάζουν απλώς μια περιπτωσιολογική μελέτη ανάλυσης δεδομένων κυκλοφορίας χρησιμοποιώντας παραδοσιακές τεχνικές.

Πιο κοντά στην έρευνά μας είναι η εργασία [LCZ+06], όπου προτείνεται ένα κατανεμημένο σύστημα εξόρυξης ρευμάτων κυκλοφορίας: ο κεντρικός υπολογιστής εκτελεί τις εργασίες εξόρυξης γνώσης και στέλνει τα πρότυπα που ανακαλύπτονται πίσω στους έχουν εξαχθεί από τα ιστορικά στοιχεία. Όλοι οι αισθητήρες που έχουν ένα κοινό πρότυπο ομαδοποιούνται στην ίδια συστάδα. Εάν εμφανιστεί κάποια παραβίαση σε αισθητήρα, τότε αυτός στέλνει ένα συναγερμό στον κεντρικό υπολογιστή, ο οποίος ειδοποιεί στη συνέχεια όλα τα μέλη της ίδιας συστάδας. Αυτό το πλαίσιο αρχικοποιείται για πρότυπα συχνών επεισοδίων (frequent episode patterns). Εντούτοις, αυτή η εργασία δίνει έμφαση στην περιγραφή του κατανεμημένου συστήματος ρευμάτων κυκλοφορίας παρά στην ανακάλυψη προτύπων σχετικών με την κυκλοφορία.

Οι Nakata και Takeuchi [NT04] χρησιμοποιούν ένα αυτοκίνητο-μοντέλο (probe-car) για να συλλέξουν πληροφορίες για την κίνηση σε περιοχές πολύ μεγαλύτερες από αυτές που καλύπτουν οι παραδοσιακοί στατικοί αισθητήρες. Χρησιμοποιούν χρονοσειρές κυκλοφορίας και εφαρμόζουν ένα αυτοπαλινδρούμενο μοντέλο (Auto Regression Model) αφού έχουμε αφαιρέσει τα περιοδικά πρότυπα. Εντούτοις, σε αυτήν την εργασία η χωρική διάσταση δε λαμβάνεται υπόψη.

4.6. Σύνοψη

Σε αυτό το κεφάλαιο, εισάγαμε τις έννοιες των *προτύπων αλληλεπίδρασης* με σκοπό την εξαγωγή σημασιολογίας της κίνησης καθώς και *προτύπων κίνησης* ως σχέσεις μεταξύ των οδικών τμημάτων ενός δικτύου πόλης.

Τα *πρότυπα αλληλεπίδρασης* είναι συνοπτικοί περιγραφείς των περιοχών, από άποψη γεωγραφικού χώρου και σχέσεων ομοιότητας μεταξύ των τροχιών που μπορούν να μας βοηθήσουν να καταλάβουμε τις συμπεριφορές κίνησης. Προς αυτόν τον σκοπό, προτείναμε μια προσέγγιση που μπορεί να ανακαλύψει τι συμβαίνει σε ένα αντικείμενο εξετάζοντας όχι μόνο στην τροχιά του, αλλά και το πλαίσιο στο όπου κινείται (που καθορίζεται από την παρουσία άλλων αντικειμένων και την αλληλεπίδρασή τους με το κινούμενο αντικείμενο που αναλύεται).

Επίσης μελετήσαμε το πρόβλημα της ανάλυσης κίνησης σε ένα οδικό δίκτυο λαμβάνοντας υπόψη τις χρονοσειρές κυκλοφορίας σε κάθε οδικό τμήμα και αξιοποιώντας μέτρα ομοιότητας. Προτείναμε δυο τεχνικές εξόρυξης γνώσης από δεδομένα τροχιών: μια για την συσταδοποίηση των ακμών βάση της κυκλοφορίας τους και μια για εύρεση σχέσεων κυκλοφορίας όπως η διάδοση, διάσπαση και συγχώνευση της κυκλοφορίας μεταξύ αυτών των οδικών τμημάτων καθώς επίσης και η πηγή και ο προορισμός κυκλοφορίας.

5. Ένα Πλαίσιο Υλοποίησης Αποθηκών Δεδομένων Τροχιών Κινούμενων Αντικειμένων

Αφού μελετήσαμε στα προηγούμενα κεφάλαια τεχνικές αποθήκευσης δεδομένων, OLAP ανάλυσης και εξόρυξης γνώσης πάνω σε τροχιές, το κεφάλαιο αυτό εστιάζει στο σχεδιασμό και ανάπτυξη ενός συστήματος που ενσωματώνει τις τεχνικές που μελετήθηκαν στα προηγούμενα κεφάλαια. Πιο συγκεκριμένα, υλοποιείται το T-WAREHOUSE, ένα σύστημα που περιλαμβάνει όλα τα απαιτούμενα βήματα για την Οπτική Υποστήριξη Αποθήκης Δεδομένων Τροχιών (Visual Trajectory Data Warehousing), από την ανακατασκευή τροχιών και την επεξεργασία μέσω της διαδικασίας Εξαγωγή – Μεταφορά - Φόρτωση (ETL) στην οπτική OLAP ανάλυση πάνω σε δεδομένα κινούμενων αντικειμένων. Παρουσιάζεται, επίσης, η προσέγγιση μας για την ανακατασκευή τροχιών. Τα κύρια σημεία αυτού του κεφαλαίου είναι: Στην Ενότητα 5.1 αναπτύσσονται θέματα που σχετίζονται με τεχνικές αποθήκευσης δεδομένων και πώς αυτά εφαρμόζονται στην περίπτωση δεδομένων κινούμενων αντικειμένων, ενώ στην Ενότητα 5.2 αναφέρονται τα κίνητρα που μας οδήγησαν στην υλοποίηση αυτής της έρευνας. Στην Ενότητα 5.3 παρουσιάζεται η αρχιτεκτονική του συστήματος T-WAREHOUSE και περιγράφονται τα διάφορα μέρη και λειτουργίες του, ενώ την Ενότητα 5.4 παρουσιάζονται τα επιμέρους χαρακτηριστικά του. Τέλος, στην Ενότητα 5.5 αναλύεται η πειραματική μεθοδολογία που ακολουθήθηκε και στην Ενότητα 5.6 καταγράφονται τα συμπεράσματα του κεφαλαίου.

5.1. Εισαγωγή

Όπως έχει προηγούμενα αναφερθεί, το κίνητρο πίσω από μια TDW είναι ο μετασχηματισμός των ακατέργαστων δεδομένων τροχιών σε χρήσιμες πληροφορίες που μπορούν να χρησιμοποιηθούν υποστηρικτικά για τη λήψη αποφάσεων σε πολυάριθμες εφαρμογές, όπως Υπηρεσίες βασισμένες σε Θέση (LBS), η διαχείριση της κυκλοφορίας, κτλ. Διαισθητικά, η μεγάλη ποσότητα των πρωτογενών δεδομένων που παράγονται από τις τεχνολογίες αισθητήρων και προσδιορισμού θέσης, ο σύνθετος χαρακτήρας των δεδομένων που αποθηκεύονται σε βάσεις δεδομένων τροχιών και οι εξειδικευμένες απαιτήσεις για την εκτέλεση ερωτημάτων καθιστούν την εξαγωγή της πολύτιμης γνώσης από τέτοια χωροχρονικά στοιχεία ένα δύσκολο στόχο. Για το λόγο αυτό, προτείνεται να επεκταθούν οι

παραδοσιακές τεχνικές συσσώρευσης έτσι ώστε να παράγουν περιληπτικές πληροφορίες για τις τροχιές και να παρέχουν ανάλυση τύπου OLAP.

Αυτή η ανάλυση μπορεί να γίνει αποτελεσματικά μέσω μιας TDW. Ωστόσο, αρκετά θέματα πρέπει να ληφθούν υπόψη, όπως:

- η παρουσία μιας φάσης προ-επεξεργασίας που έχει να κάνει αποκλειστικά με την κατασκευή των τροχιών που έπειτα αποθηκεύονται στην MOD που παρέχει ισχυρές και αποδοτικές λειτουργίες για το χειρισμό τους,
- η υλοποίηση μιας αποδοτικής διαδικασίας ETL προσανατολισμένης σε τροχιές,
- η ενσωμάτωση των κατάλληλων συσσωρευτικών μηχανισμών που θα ακολουθήσουν το μοντέλο κύβου προσανατολισμό σε τροχιές,
- Ο σχεδιασμός μιας οπτικής OLAP διεπαφής που επιτρέπει την πολυδιάστατη και διαδραστική ανάλυση.

5.2. Κίνητρα

Μεγάλος αριθμός εφαρμογών θα μπορούσε να ωφεληθεί από την προαναφερθείσα προσέγγιση. Για παράδειγμα, μια διαφημιστική εταιρία που ενδιαφέρεται για την ανάλυση δεδομένων κινούμενων αντικειμένων σε διαφορετικές περιοχές μιας πόλης ώστε να αποφασίσει τι στρατηγική θα ακολουθήσει για τις οδικές διαφημίσεις (που αναρτώνται σε ειδικούς πίνακες στο δρόμο). Ενδιαφέρεται, επίσης, να αναλύσει τα δημογραφικά προφίλ των ανθρώπων που επισκέπτονται τις διαφορετικές περιοχές της πόλης σε διαφορετικά χρονικά διαστήματα της ημέρας ώστε να μπορέσει να αποφασίσει για την καταλληλότερη διαδοχή των διαφημίσεων στους πίνακες στους διάφορους δρόμους της πόλης σε διαφορετικά χρονικά διαστήματα.

Αναφέροντας περισσότερα κίνητρα ως σενάρια χρήσης του T-WAREHOUSE, παραθέτουμε παρακάτω κάποιες ενδιαφέρουσες ερωτήσεις που ένας αναλυτής θα μπορούσε αμφίδρομα να προσπαθήσει να απαντήσει μέσω των λειτουργιών που προσφέρονται από το T-WAREHOUSE:

- Πού εμφανίζεται το μεγαλύτερο πρόβλημα κυκλοφοριακής συμφόρησης; Ποια ώρα;
- Τι συμβαίνει ακριβώς σε επίπεδο οδικού δικτύου;
- Πώς εξελίσσεται η κίνηση από μέρος σε μέρος;

5.2.1. Η συνεισφορά μας

Βασιζόμενοι στα πρόσφατα ερευνητικά μας αποτελέσματα στο χώρο [MFN+08a], που αντιμετωπίζουν το πρόβλημα της Αποθήκευσης Δεδομένων Τροχιών σε όλες του τις διαστάσεις, αναπτύξαμε το T-WAREHOUSE, ένα σύστημα για Οπτική Αποθήκη Δεδομένων Τροχιών. Παρακάτω παρουσιάζονται συνοπτικά τα διάφορα θέματα που μας απασχόλησαν μαζί με την ερευνητική συνεισφορά μας στις [MFN+08a], [MFN+08b], [MT09b], [LMF+10]:

- Τα δειγματοληπτικά δεδομένα θέσης που λαμβάνονται από συσκευές εφοδιασμένες με τεχνολογία GPS πρέπει να μετασχηματιστούν σε δεδομένα τροχιών και να αποθηκευτούν σε μια MOD: για να επιτευχθεί αυτός ο σκοπός, προτείνουμε μια τεχνική *ανακατασκευής τροχιών* που μετατρέπει τις σειρές των πρωτογενών δειγματοληπτικών δεδομένων σε αξιοποιήσιμες τροχιές.
- Περιγράφουμε τις αρχιτεκτονικές προδιαγραφές του πλαισίου καθώς και τα διάφορα ερευνητικά θέματα που αντιμετωπίζονται,
- Διερευνούμε την ισχύ, προσαρμοστικότητα και αποδοτικότητα του πλαισίου για την εφαρμογή αναλυτικών μεθόδων OLAP σε πραγματικά δεδομένα κινούμενων αντικειμένων.

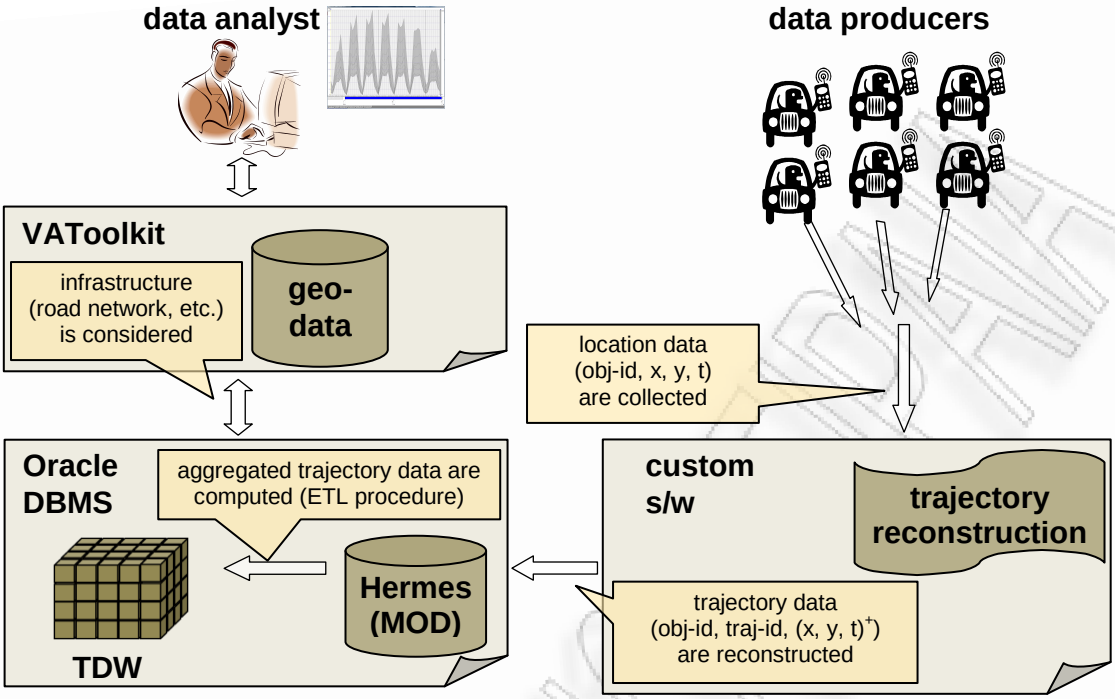
5.3. Αρχιτεκτονική Συστήματος

Η όλη αρχιτεκτονική του συστήματος T-WAREHOUSE απεικονίζεται στην Εικόνα 5-1. Πιο συγκεκριμένα, οι κινητές συσκευές στέλνουν περιοδικά δεδομένα για το τελευταίο μέρος της τροχιάς τους. Αυτή η μεγάλη ποσότητα δεδομένων που συλλέγεται από όλους τους συνδρομητές προωθείται για επεξεργασία μέσω του λογισμικού ανακατασκευής τροχιών, του οποίου βασική λειτουργία είναι κάποια βασική προ-επεξεργασία των τροχιών. Αυτό μπορεί να περιλαμβάνει ανακατασκευή τροχιών (έτσι ώστε να ληφθεί απόφαση για το ποια σημεία αποτελούν μέρος ποιων τροχιών), καθώς επίσης και τεχνικών για την αντιμετώπιση τιμών θορύβου. Αυτές οι ανακατασκευασμένες τροχιές αποθηκεύονται σε μια MOD απ' όπου με κατάλληλη εκτέλεση ερωτημάτων και μια ETL διαδικασία μπορούν να εφαρμοστούν (πιθανά λαμβάνοντας υπόψη τους διαφορετικούς τύπους των γεω-δεδομένων) για την ανάκτηση πληροφοριών σχετικά με τις τροχιές (π.χ. περιεχόμενο τροχιάς σε διαφορετική κλιμάκωση, συσσωρεύσεις, μετα-δεδομένα κίνησης κτλ) για να τροφοδοτηθεί η TDW.

Μια TDW εξυπηρετεί δύο κύριες ανάγκες: την παροχή της κατάλληλης υποδομής για δημιουργία προηγμένων αναφορών (reporting) και τη διευκόλυνση της εφαρμογής αλγορίθμων εξόρυξης γνώσης από συσσωρευμένα δεδομένα. Σύμφωνα με τις ανάγκες των τελικών χρηστών, οι τελευταίοι έχουν ανάγκη από πρόσβαση είτε σε βασικές αναφορές είτε σε αναφορές με δυνατότητες OLAP ανάλυσης. Τα Σενάρια «τι θα συνέβαινε – εάν» (what - if) και η πολυδιάστατη ανάλυση είναι χαρακτηριστικά παραδείγματα αναλυτικών μεθόδων που θα μπορούσαν να βασιστούν σε βάσεις δεδομένων τροχιών.

Επιπροσθέτως, η ενσωμάτωση GIS στρωμάτων με γεωγραφική πληροφορία μπορεί να οδηγήσει σε ένα πλουσιότερο εννοιολογικό μοντέλο/πρότυπο παρέχοντας άρα και πιο προηγμένες αναλυτικές δυνατότητες. Συνδυάζοντας τα δεδομένα τροχιών με θεματικά στρώματα (όπως τα γεωγραφικά, τοπογραφικά και δημογραφικά στρώματα) είναι δυνατή η ενίσχυση των αναλυτικών δυνατοτήτων πιθανών εφαρμογών. Τελικά, πάνω σε αυτή τη λειτουργία, έχουμε αναπτύξει περαιτέρω ένα προηγμένο μοντέλο για την εκτέλεση διαδικασιών OLAP με οπτικό τρόπο.

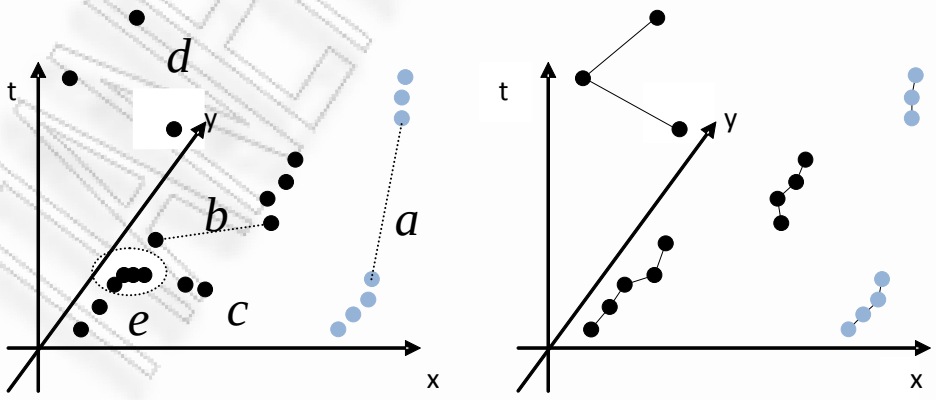
Στις επόμενες παραγράφους, παρουσιάζεται η δομή του πλαισίου και περιγράφονται οι διάφορες λειτουργίες του μαζί με την ερευνητική συνεισφορά μας.



Εικόνα 5-1: Η αρχιτεκτονική του T-WAREHOUSE.

5.3.1. Ανακατασκευή τροχιών

Όπως έχει και προηγούμενα αναφερθεί, τα πρωτογενή δεδομένα που συλλέγονται αναπαριστούν γεωγραφικές τοποθεσίες με χρονική ένδειξη (Εικόνα 5-2α). Εκτός από την αποθήκευση αυτών των δεδομένων σε μια MOD, αυτό που είναι ενδιαφέρον είναι η ανακατασκευή των τροχιών τους (Εικόνα 5-2β). Η εργασία αυτή που αποκαλείται *ανακατασκευή τροχιών* δεν είναι μια απλή διαδικασία. Λαμβάνοντας υπόψη ότι τα δεδομένα φτάνουν μαζικά, είναι επιτακτική η ανάγκη ύπαρξης ενός φίλτρου που θα αποφασίζει αν το νέο σύνολο δεδομένων πρέπει να επισυναφθεί σε μια υπάρχουσα τροχιά ή όχι. Στο [MFN+08a], παρουσιάζουμε έναν αλγόριθμο για την ανακατασκευή των τροχιών.



Εικόνα 5-2: α) πρωτογενείς θέσεις, β) ανακατασκευασμένες τροχιές.

Σε αυτή την έρευνα, υποθέτουμε ότι το φίλτρο είναι μέρος του διαχειριστή της ανακατασκευής της τροχιάς, μαζί με μια απλή μέθοδο για τον καθορισμό διαφορετικών τροχιών, που εφαρμόζεται στα

ακατέργαστα σημεία. Με δεδομένο ότι η έννοια της τροχιάς δεν μπορεί να ορίζεται με τον ίδιο τρόπο σε κάθε εφαρμογή λόγω διαφορετικών απαιτήσεων και σημασιολογίας, ορίζουμε τις ακόλουθες γενικές παραμέτρους για την ανακατασκευή τροχιάς:

- *Χρονικό κενό μεταξύ τροχιών* gap_{time}: το μέγιστο επιτρεπόμενο χρονικό διάστημα μεταξύ δύο συνεχόμενων χρονοσημασμένων σημείων της ίδιας τροχιάς για ένα κινούμενο αντικείμενο. Με αυτό τον τρόπο, κάθε χρονοσημασμένο σημείο του αντικειμένου o_i, που λαμβάνεται μετά από το ορισμένο gap_{time} από την τελευταία θέση του, θα οδηγήσει στην δημιουργία μιας νέας τροχιάς του ίδιου αντικειμένου (περίπτωση *a* στην Εικόνα 5-2α) .
- *Χωρικό κενό μεταξύ τροχιών* gap_{space}: η μέγιστη επιτρεπόμενη απόσταση σε δισδιάστατο επίπεδο μεταξύ δύο συνεχόμενων χρονοσημασμένων σημείων της ίδιας τροχιάς. Με αυτό τον τρόπο, κάθε χρονοσημασμένο σημείο του αντικειμένου o_i, με απόσταση από την προηγούμενη θέση του μεγαλύτερη από την ορισμένη gap_{space}, θα οδηγήσει στη δημιουργία μίας νέας τροχιάς για το αντικείμενο o_i (περίπτωση *b* στην Εικόνα 5-2α).
- *Μέγιστη ταχύτητα* V_{max}: η μέγιστη επιτρεπόμενη ταχύτητα ενός κινούμενου αντικειμένου. Χρησιμοποιείται για να καθοριστεί πότε ένα χρονοσημασμένο σημείο πρέπει να θεωρηθεί θόρυβος και συνεπώς να αποβάλλεται από την σχηματιζόμενη τροχιά. Κάθε φορά που λαμβάνονται στοιχεία για μια νέα χρονοσημασμένη θέση του αντικειμένου o_i, ελέγχονται με βάση την τελευταία γνωστή θέση του αντικειμένου, και υπολογίζεται η αντίστοιχη στιγμιαία ταχύτητα. Εάν αυτή ξεπερνά την ορισμένη ως V_{max}, τότε αυτή η θέση θεωρείται θόρυβος και (προσωρινά) δεν συμπεριλαμβάνεται στην διαδικασία ανακατασκευής τροχιάς (ωστόσο, κρατείται ξεχωριστά σε περίπτωση που κριθεί χρήσιμη μετέπειτα – δείτε την παράμετρο που ακολουθεί) (περίπτωση *c* στην Εικόνα 5-2α).
- *Μέγιστη διάρκεια θορύβου* $noise_{max}$: είναι η μέγιστη διάρκεια θορύβου ενός τμήματος τροχιάς. Κάθε ακολουθία από χρονοσημασμένα σημεία θορύβου του ίδιου αντικειμένου καταλήγουν σε μία νέα τροχιά δεδομένου ότι η διάρκεια της ξεπερνά την ορισμένη $noise_{max}$. Για παράδειγμα, ας θεωρήσουμε μια εφαρμογή που καταγράφει τις θέσεις των πεζών, όπου η μέγιστη ταχύτητα για ένα πεζό έχει οριστεί V_{max} = 3 m/sec. Όταν επιλέξει να χρησιμοποιήσει ένα μεταφορικό μέσο (π.χ. λεωφορείο), η στιγμιαία ταχύτητα του που θα καταγραφεί θα ξεπεράσει την V_{max}, θέτοντας (πρόσκαιρα όπως θα προκύψει στη συνέχεια) τις θέσεις του λεωφορείου ως θόρυβο. Το μήκος της παραμέτρου μέγιστου θορύβου υποστηρίζει αυτό το σενάριο: όταν η διάρκεια αυτής της ακολουθίας «θορύβου» ξεπερνά το ορισμένο $noise_{max}$, μια νέα τροχιά δημιουργείται περιλαμβάνοντας όλες αυτές τις θέσεις (περίπτωση *d* στην Εικόνα 5-2α).
- *Απόσταση ανοχής* D_{tol}: η ανοχή των χρονοσημασμένων θέσεων που αποστέλλονται. Με άλλα λόγια, είναι η μέγιστη απόσταση μεταξύ δύο συνεχόμενων χρονοσημασμένων θέσεων του ίδιου αντικειμένου ώστε το αντικείμενο να θεωρείται στάσιμο. Κάθε φορά που στοιχεία για μια νέα χρονοσημασμένη θέση του αντικειμένου o_i λαμβάνονται, ελέγχονται με βάση την τελευταία γνωστή θέση του αντικειμένου, και αν η απόσταση είναι μικρότερη από την ορισμένη D_{tol} τότε χαρακτηρίζεται ως άχρηστο και στη συνέχεια αποβάλλεται/απορρίπτεται (περίπτωση *e* στην Εικόνα 5-2α).

Ο προτεινόμενος αλγόριθμος TRAJECTORY-RECONSTRUCTION που υιοθετεί τις παραπάνω παραμέτρους απεικονίζεται στην Εικόνα 5-3. Ως δεδομένα εισαγωγής χρειάζεται τις θέσεις μαζί με στοιχεία της ταυτότητας του αντικειμένου (object-id) και ένα κατάλογο που περιλαμβάνει τις μερικώς επεξεργασμένες τροχιές μέχρι εκείνη τη στιγμή από τον *διαχειριστή της ανακατασκευής των τροχιών*. Αυτές οι μερικώς επεξεργασμένες τροχιές αποτελούνται από μερικά από τα πιο πρόσφατα σημεία τροχιών, ανάλογα με τις τιμές των παραμέτρων που έχουν τεθεί στον αλγόριθμο.

Ως πρώτο βήμα (γραμμές 1-6), ο αλγόριθμος ελέγχει αν το αντικείμενο έχει υποστεί κάποια επεξεργασία μέχρι εκείνη τη στιγμή, και αν ναι, ανακτά την μερικώς ανακατασκευασμένη τροχιά από την αντίστοιχη λίστα, ενώ, στην αντίθετη περίπτωση, δημιουργεί μια νέα τροχιά και την προσθέτει στην λίστα. Μετά (γραμμές 7-31), συγκρίνει το εισερχόμενο σημείο P με την ουρά της μερικώς ανακατασκευασμένης τροχιάς (LastPoint) εφαρμόζοντας τις παραπάνω παραμέτρους ανακατασκευής τροχιάς:

- απορρίπτει το P αν είναι πιο κοντά στο ορισμένο D_{tol} από το LastPoint (γραμμές 7-12) ή
- απορρίπτει το P όταν υπολογίζεται μια ταχύτητα μεγαλύτερη από την ορισμένη $V_{max,}$ εκτός εάν προκαλείται η περίπτωση της $noise_{max}$
- δημιουργεί μια νέα τροχιά αν η χρονική διάρκεια μεταξύ P και LastPoint είναι μεγαλύτερη από την ορισμένη gap_{time} (γραμμές 8-12 και 24-27) ή η χωρική τους απόσταση είναι μεγαλύτερη από την ορισμένη gap_{space} (γραμμές 19-22).

ενώ σε κάθε άλλη περίπτωση, χαρακτηρίζει το LastPoint ως μέρος της μερικώς ανακατασκευασμένης τροχιά και αντικαθιστά το P.

Η παραπάνω διαδικασία υποστηρίζει τη λογική απαίτηση για την ανίχνευση μιας τροχιάς κατά τη διάρκεια ενός ταξιδιού: ας υποθέσουμε την περίπτωση ενός χρήστη που διανύει απόσταση από το σπίτι στο χώρο εργασίας του το πρωί και από την εργασία του στο σπίτι του το απόγευμα, κρατώντας μια συσκευή με δυνατότητα εντοπισμού θέσης (π.χ. GPS) πάντοτε ενεργή. Σε αυτή τη περίπτωση, κατά τη διάρκεια που ο χρήστης σταθμεύει το αυτοκίνητο δεν υπάρχουν χωρικά κενά, ούτε προβλήματα μέγιστης ταχύτητας, που να μπορούν να προκαλέσουν δημιουργία νέας τροχιάς. Επιπλέον, τα παραγόμενα δεδομένα από τη συσκευή με GPS δίνουν τη θέση του χρήστη ανά δευτερόλεπτο, οπότε αρχικά δεν υπάρχουν και χρονικά κενά. Ωστόσο, από τη στιγμή που το αυτοκίνητο αρχίσει να κινείται, ο αλγόριθμος αποβάλλει όλα τις καταγεγραμμένες θέσεις που σημειώθηκαν τη διάρκεια της μη-κίνησης, και δημιουργεί ένα τεχνητό χρονικό κενό (π.χ. μόνο η πρώτη θέση μετά την στάθμευση και η τελευταία πριν ξεκινήσει ξανά υπάρχουν στον αλγόριθμο για την ανακατασκευή της τροχιάς). Ως αποτέλεσμα, ο αλγόριθμος ανιχνεύει το χρονικό κενό και δημιουργεί νέες τροχιές, όπου χρειάζεται, βασιζόμενος μόνο στην πληροφορία ότι το εντοπιζόμενο αντικείμενο σταμάτησε να κινείται για ένα αρκετά μεγάλο χρονικό διάστημα (δηλαδή μεγαλύτερο από το ορισμένο gap_{time}).

```
Algorithm Trajectory-Reconstruction (PartialTrajectories List, P
Point, OId ObjectId)
 1.   IF NOT PartialTrajectories.Contains(OId) THEN
 2.     CTrajectory=New Trajectory;
 3.     CTrajectory.AddPoint(P);
 4.     PartialTrajectories.Add(CTrajectory);
 5.   ELSE
 6.     CTrajectory=PartialTrajectories(OId);
 7.     IF Distance(CTrajectory.LastPoint,P)<= D_TOL THEN
 8.       IF P.T – CTrajectory.LastPoint.T > gap_Time THEN
 9.         Report Ctrajectory.LastPoint;
10.         CTrajectory.Id=CTrajectory.Id+1;
11.         CTrajectory.AddPoint(P);
12.       ENDIF
13.     ELSEIF Speed(CTrajectory.LastPoint,P)> V_max THEN
14.       IF P.T – CTrajectory.LastPoint.T > noise_max THEN
15.         Report CTrajectory.Noise;
16.       ELSE
17.         CTrajectory.AddNoise(P);
18.       ENDIF
19.     ELSEIF Distance(CTrajectory.LastPoint,P)> gap_space THEN
20.       Report Ctrajectory.LastPoint;
21.       CTrajectory.Id=CTrajectory.Id+1;
22.       CTrajectory.AddPoint(P);
23.     ELSE
24.       IF P.T – CTrajectory.LastPoint.T > gap_Time THEN
25.         Report Ctrajectory.LastPoint;
26.         CTrajectory.Id=CTrajectory.Id+1;
27.         CTrajectory.AddPoint(P);
28.       ELSE
29.         CTrajectory.AddPoint(P);
30.       ENDIF
31.     ENDIF
32.   ENDIF
```

Εικόνα 5-3: Ο αλγόριθμος TRAJECTORY-RECONSTRUCTION.

5.3.2. Η MOD

Όπως έχουμε ήδη αναφέρει, το HERMES [PTV+06] είναι ένα σύστημα με ισχυρές δυνατότητες που μπορεί να υποστηρίξει έναν προγραμματιστή χωροχρονικών βάσεων δεδομένων στο σχεδιασμό, την ανάπτυξη και την εκτέλεση ερωτημάτων σε βάσεις δεδομένων με δυναμικά αντικείμενα που αλλάζουν θέση, σχήμα και μέγεθος, είτε στιγμιαία είτε συνεχώς στο χρόνο. Το HERMES παρέχει χωροχρονική λειτουργικότητα σε προηγμένης τεχνολογίας αντικειμενοσχεσιακά συστήματα διαχείρισης βάσεων δεδομένων (Object-Relational DBMS - ORDBMS). Ολόκληρη η λειτουργία παρέχεται ως data cartridge μέσω της επεκτάσιμης διεπαφής της Oracle.

Το σύστημα HERMES χρησιμοποιείται ως βάση δεδομένων των τροχιών που χρησιμοποιούνται από το εργαλείο T-WAREHOUSE. Επιπλέον, παρέχει την λειτουργικότητα που είναι απαραίτητη κατά τη διάρκεια της φάσης τροφοδοσίας της TDW. Μερικές χαρακτηριστικές ερωτήσεις που μπορούν να εκτελεστούν κατά τη διάρκεια της ETL παρουσιάζονται παρακάτω:

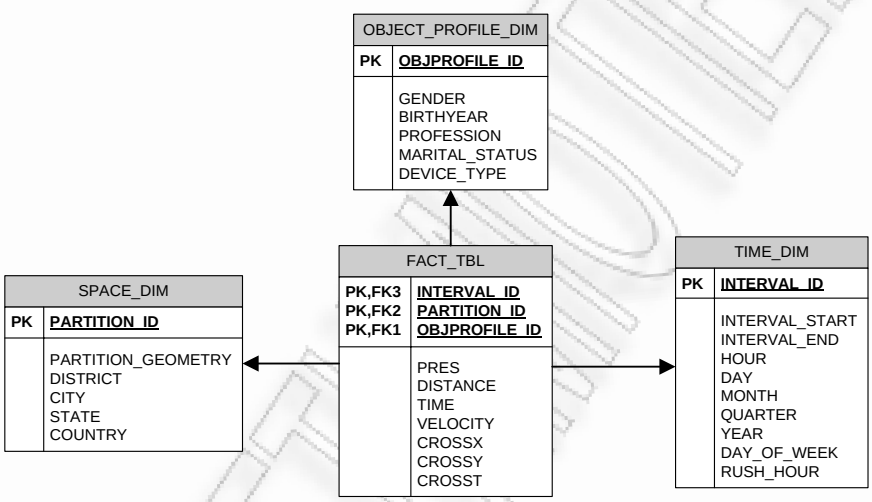
- "*View RELtrajectories*": ανακτά τροχιές σε σχεσιακή μορφή που ορίζονται από ένα αντικείμενο και ένα προσδιοριστικό τροχιάς.
- "*View MODtrajectories*": ανακτά τροχιές σε αντικειμενοσχεσιακή μορφή που ορίζονται από ένα προσδιοριστικό τροχιάς. Ωστόσο πρέπει να σημειωθεί ότι όποτε το αποτέλεσμα της εκτέλεσης ενός ερωτήματος είναι μια τροχιά, δηλαδή ένα πολύπλοκο αντικείμενο, η αντίστοιχη ανακτώμενη στήλη μετατρέπεται σε μορφή γραμμής έτσι ώστε να μπορεί ο περιηγητής (browser) να το οπτικοποιήσει.
- "*(Spatial) Intersection operator*": περιορίζει μια τροχιά μέσα σε μία δεδομένη γεωμετρία, η οποία προσδιορίζεται ως ένα χωρικό αντικείμενο τύπου Oracle's Spatial sdo_geometry (π.χ. ορθογώνιο, πολυγωνική περιοχή κτλ.).
- "*(Temporal) Intersection operator*": περιορίζει μια τροχιά μέσα σε μια δεδομένη χρονική περίοδο.
- "*(Multi-Temporal) Intersection operator"*: περιορίζει μια τροχιά μέσα σε μια σειρά (πιθανά διεσπαρμένων) χρονικών περιόδων.
- "*Topological operator*": ελέγχει (τελεστής συσχέτισης) αν ένα αντικείμενο τύπου sdo_geometry ικανοποιεί μια συγκεκριμένη τοπολογική σχέση (προσδιορισμένη από μια μάσκα, π.χ. 'ANYINTERACT') με την πρόβλεψη μιας τροχιάς σε μια δεδομένη χρονική στιγμή. Το αποτέλεσμα της εκτέλεσης του ερωτήματος είναι μια boolean τιμή.
- "*within_distance operator*": ελέγχει αν μια τροχιά σε δεδομένη χρονική στιγμή είναι μέσα στα πλαίσια δεδομένης απόστασης από ένα αντικείμενο τύπου sdo_geometry.
- "*distance operator*": επιστρέφει την τιμή της απόστασης μιας τροχιάς σε μια δεδομένη χρονική στιγμή από άλλη τροχιά που προβλέπεται για άλλη δεδομένη χρονική στιγμή.
- "*enter operator*": αυτή η ερώτηση δείχνει την δομή της τροχιάς την συγκεκριμένη ώρα που εκτελείται και για τη συγκεκριμένη τροχιά επιστρέφει τη χρονική στιγμή κατά την οποία η τροχιά εισέρχεται σε μια δεδομένη περιοχή (sdo_geometry).
- "*leave operator*": αυτή η ερώτηση δείχνει τη δομή μιας τροχιάς την συγκεκριμένη ώρα που εκτελείται και επιστρέφει τη χρονική στιγμή κατά την οποία η τροχιά εξέρχεται από μια περιοχή (sdo_geometry).
- "*directional operator*": επιστρέφει την τιμή της κατεύθυνσης μιας γραμμής που σχηματίζεται μεταξύ δύο σημείων. Το πρώτο σημείο είναι η πρόβλεψη μιας τροχιάς σε μια δεδομένη χρονική στιγμή και το δεύτερο σημείο δίνεται ως σημείο sdo_geometry.
- "*speed operator*": δίνει την τιμή της ταχύτητας μιας τροχιάς σε μια δεδομένη χρονική στιγμή.

5.3.3. Τροφοδοσία της TDW

Ας θεωρήσουμε ως παράδειγμα, την σχηματική απεικόνιση μιας TDW, όπως αυτή περιγράφεται στην Εικόνα 5-4, που περιέχει μια *χωρική* (SPACE_DIM) και μια *χρονική* (TIME_DIM) *διάσταση* περιγράφοντας τη γεωγραφία και το χρόνο, αντίστοιχα. Μη-χωροχρονικές διαστάσεις μπορούν επίσης να συμπεριληφθούν. Για παράδειγμα, η σχηματική απεικόνιση στην Εικόνα 5-4 περιέχει την διάσταση

OBJECT_PROFILE_DIM που συλλέγει δημογραφικές πληροφορίες, όπως γένος, ηλικία, εργασία των κινούμενων αντικειμένων.

Εκτός από τα κλειδιά στους πίνακες με τις διάφορες διαστάσεις, υπάρχει ο πίνακας που περιέχει μία σειρά μέτρων που αντιπροσωπεύουν συσσωρευμένες πληροφορίες (aggregated information). Τα μέτρα αυτά που συμπεριλήφθηκαν για το παράδειγμα στην Εικόνα 5-4 περιλαμβάνουν τον *αριθμό των διακριτών τροχιών* (PRES), τη *μέση διανυόμενη απόσταση* (DISTANCE), τη *μέση διάρκεια μετακίνησης* (TIME) , τη *μέση ταχύτητα* (VELOCITY) καθώς και κάποια βοηθητικά μέτρα (π.χ. CROSSX, CROSSY, CROSST) για μια συγκεκριμένη ομάδα ανθρώπων (με συγκεκριμένα χαρακτηριστικά) που κινούνται σε μια συγκεκριμένη περιοχή – οριοθετημένη - σε μια συγκεκριμένη χρονική περίοδο. Η σχηματική αναπαράσταση είναι παρόμοια με αυτή που παρουσιάστηκε στην Εικόνα 3-8, και υπάρχει μια μικρή διαφοροποίηση στην ονομασία (όχι στην έννοια) των μέτρων.



Εικόνα 5-4: Δείγμα TDW που μπορεί να χρησιμοποιηθεί από το T-WAREHOUSE.

Η TDW τροφοδοτείται με συσσωρευμένα δεδομένα τροχιών. Για να επιτευχθεί αυτό εκτελείται μια αποτελεσματική ETL διαδικασία έτσι ώστε να δώσει στα μέτρα της TDW τις κατάλληλες αριθμητικές τιμές για κάθε κελί βάσης. Η προτεινόμενη ETL διαδικασία, που παρουσιάζεται εκτενώς στην Υποενότητα 3.3.1, ανιχνεύει τα τμήματα της τροχιάς που βρίσκονται μέσα στα όρια των κελιών βάσης. Αυτό το βήμα αντιστοιχεί στην πραγματικότητα σε χωροχρονικές ερωτήσεις που δίνουν όχι μόνο τα αναγνωριστικά αλλά επίσης και τα τμήματα των τροχιών που ικανοποιούν τους περιορισμούς. Για την αποτελεσματική υποστήριξη της εναποθήκευσης που περιγράφεται παραπάνω καθώς επίσης και των απαιτήσεων για την εκτέλεση ερωτημάτων σχετικά με τροχιές, χρησιμοποιούμε την MOD HERMES (όπως περιγράψαμε στην Υποενότητα 5.3.2).

5.3.4. Συσσώρευση

Οι δυνατότητες συσσώρευσης δεδομένων πάνω στα μέτρα προσφέρονται για τους σκοπούς της OLAP ανάλυσης (π.χ. πώς τα μεγέθη στο χαμηλότερο επίπεδο της ιεραρχίας μπορούν να αξιοποιηθούν ώστε να υπολογίσουν τα μέτρα σε κάποιο υψηλότερο επίπεδο ιεραρχίας). Μια δυσκολία με τα δεδομένα τροχιών είναι ότι μια τροχιά μπορεί να εκτείνεται σε πολλά κελιά βάσης. Οπότε στη φάση συσσώρευσης θα έχουμε να κάνουμε με το επονομαζόμενο *πρόβλημα μοναδικής προσμέτρησης*. Αυτό

είναι ένα ζήτημα, γιατί από τη στιγμή που «φορτώνεται» η TDW, τα αναγνωριστικά των τροχιών χάνονται. Με αυτό τον τρόπο δημιουργούνται εμπόδια στην συσσώρευση των δεδομένων στις διεργασίες OLAP, για παράδειγμα στο υπολογισμό του μεγέθους PRES που μπορεί να μας δώσει τον αριθμό των *διακριτών* τροχιών ενός συγκεκριμένου χαρακτηριστικού που διασχίζει ένα χωροχρονικό κελί. Αυτόν επηρεάζει επίσης και άλλα μεγέθη, όπως τον μέσο όρο (*average*) που ορίστηκε πάνω από το PRES. Για να αντιμετωπιστεί αυτό το πρόβλημα, χρησιμοποιούμε μια προσεγγιστική λύση, που περιγράφηκε ήδη στην Υποενότητα 3.3.2, η οποία αποδεικνύεται να λειτουργεί αποτελεσματικά. Υποθέτοντας ότι η PARTITION_GEOMETRY στην Εικόνα 5-4 είναι ένα κανονικό πλέγμα, αποθηκεύουμε αντίστοιχα στα βοηθητικά μεγέθη CROSSX, CROSSY και CROSST τον αριθμό των *διακριτών* τροχιών ενός συγκεκριμένου χαρακτηριστικού που ξεπερνά το χωροχρονικό δύο γειτονικών κελιών κατά μήκος του άξονα x/y/t. Η γνώση του αριθμού των τροχιών που ξεπερνούν το όριο μεταξύ κελιών είναι χρήσιμη για τη διόρθωση λαθών εξαιτίας διπλοκαταχωρήσεων κατά την συσσώρευση τέτοιων κελιών (Υποενότητα 3.3.2 για περισσότερες λεπτομέρειες).

5.3.5. OLAP λειτουργίες και Οπτικοποίηση

Για να ξεπεραστούν αυτοί οι περιορισμοί, αναπτύξαμε οπτικές OLAP λειτουργίες, χρησιμοποιώντας το εργαλείο ανάλυσης (VAToolkit) [AAW07], ένα διαδραστικό γεωγραφικό πληροφοριακό σύστημα στηριγμένο σε τεχνολογία Java. Το εργαλείο αυτό επιτρέπει στο χρήστη να δει γεω-αναφερόμενα δεδομένα πάνω σε ένα χάρτη και, επίσης λειτουργίες για το χειρισμό χρονικών δεδομένων, χρησιμοποιώντας γραφήματα ή κινούμενα σχέδια, ανάλογα με το είδος των δεδομένων προς ανάλυση.

Τα πλεονεκτήματα του συστήματος μας είναι πολλαπλά. Πρώτα, ο χρήστης μπορεί να οπτικοποιήσει το τμήμα του χωρικού πεδίου πάνω στο χάρτη στο οποίο τα χωρικά δεδομένα αναφέρονται. Επιπλέον, ο χρήστης μπορεί να επιλέξει μια περιοχή στο γράφημα και να εφαρμόσει λειτουργίες συσσώρευσης (roll-up) και εμβάθυνσης (drill-down) με στόχο να αποκτήσει, αντίστοιχα, μια πιο περιληπτική ή λεπτομερή άποψη της περιοχής. Πάνω σε αυτές τις απεικονίσεις ο χρήστης μπορεί να εφαρμόσει εξειδικευμένες τεχνικές οπτικοποίησης, που του προσφέρουν καλύτερη κατανόηση των μεγεθών που περιέχονται στην TDW.

Συνοψίζοντας, η οπτική διεπαφή που υλοποιήσαμε επιτρέπει στο χρήστη να πλοηγείται εύκολα / έχει εύκολη πρόσβαση στα αποθηκευμένα δεδομένα μέσα στην TDW στα διαφορετικά επίπεδα ιεράρχησης, να έχει μια ολοκληρωμένη άποψη των δεδομένων σε χρόνο και σε χώρο ή να εστιάζει σε συγκεκριμένα μεγέθη, χωρικές περιοχές ή χρονικά διαστήματα.

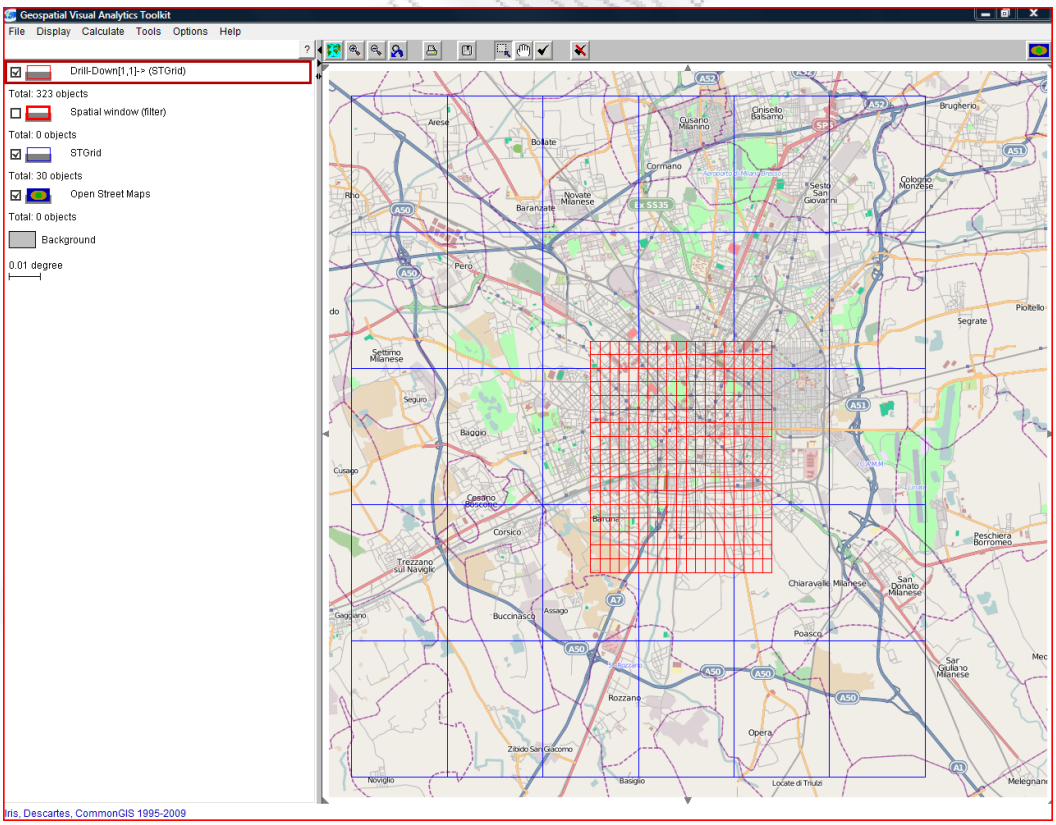
5.4. Επίδειξη του Συστήματος

Σε αυτή την υποενότητα, επιδεικνύουμε την λειτουργικότητα του συστήματος με χρήση ενός μεγάλου όγκο πραγματικών δεδομένων ενός συνόλου αυτοκινήτων που κινούνται στη μητροπολιτική περιοχή του Μιλάνου (Ιταλία). Το σύνολο των δεδομένων αποτελείται από δυο εκατομμύρια πρωτογενείς εγγραφές που αναπαριστούν τη μετακίνηση 17.000 αντικειμένων (περίπου 200.000 τροχιές) κατά τη διάρκεια μιας εβδομάδας, από Κυριακή σε Σάββατο.

Πριν δείξουμε τη λειτουργία του T-WAREHOUSE, θα περιγράψουμε τα επιμέρους χαρακτηριστικά μοντέλου της TDW που θα χρησιμοποιηθεί για αυτήν την εφαρμογή.

Ο χρήστης μπορεί να επιλέξει το επίπεδο κλιμάκωσης τόσο για την χωρική όσο και για την χρονική διάσταση και τις αντίστοιχες ιεραρχίες. Θέτουμε ένα ορθογώνιο πλέγμα (PARTITION_GEOMETRY στην Εικόνα 5-4), το μέγεθος του οποίου είναι 300×400 m^2, και χρονικά διαστήματα της 1 ώρας, ως βασική κλίμακα. Η χωρική ιεράρχηση αποτελείται από ένα σύνολο πλεγμάτων συσσωρευμένων ομάδων γειτονικών χωρικών κελιών βάσης, ενώ η χρονική ιεράρχηση είναι ωριαία – 3-ωρη – ημερήσια – εβδομαδιαία.

Γραφική Διεπαφή Χρήστη (GUI) και Οπτική ανάλυση (Visual analytics). Παρουσιάζουμε παρακάτω τις λειτουργίες που παρέχονται στον αναλυτή από το T-WAREHOUSE. Χρησιμοποιώντας το σύστημά μας, διευκολύνεται ο χειρισμός και η οπτικοποίηση των χωροχρονικών πλεγμάτων της TDW σε διάφορα επίπεδα κλιμάκωσης. Αν οι λειτουργίες συσσώρευσης (roll-up) περιέχουν την χωρική διάσταση, οπτικά αυτό επιδρά στην κλιμάκωση του πλέγματος, το οποίο γίνεται μεγαλύτερο. Αντίθετος χειρισμός, με λειτουργία εμβάθυνσης (drill-down), αυξάνει το επίπεδο παροχής λεπτομερειών των δεδομένων. Επιτρέπει έτσι στο χρήστη καθοδική εμβάθυνση στην ιεράρχηση. Στην Εικόνα 5-5, παρουσιάζεται ένα παράδειγμα λειτουργίας εμβάθυνσης (drill-down) για να αποκτήσουμε δεδομένα εστιασμένα στο κέντρο του Μιλάνου.

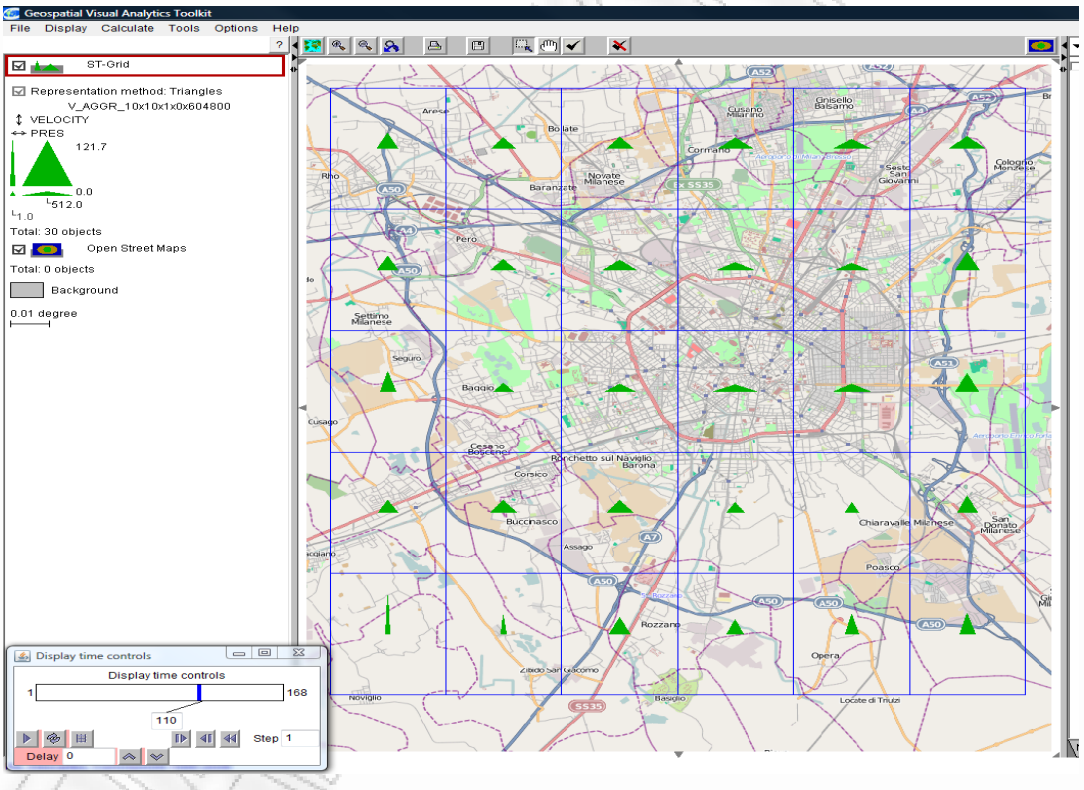


Εικόνα 5-5: Λειτουργία εμβάθυνσης στο T-WAREHOUSE.

Ξεκινώντας από την οπτικοποίηση του διαστήματος, είναι δυνατό να προβληθούν κάποια μεγέθη, τα οποία μπορούν να οπτικοποιηθούν με πολλές μεθόδους.

Στον *Τριγωνικό* (*Triangle*) τύπο οπτικοποίησης, ένα τρίγωνο σχεδιάζεται σε κάθε κελί του πλέγματος σε δεδομένο επίπεδο της ιεράρχησης της TDW. Η βάση και το ύψος αυτού του τριγώνου αντιστοιχούν στις τιμές των δύο επιλεγμένων μεγεθών που ο χρήστης επιθυμεί να αναλύσει.

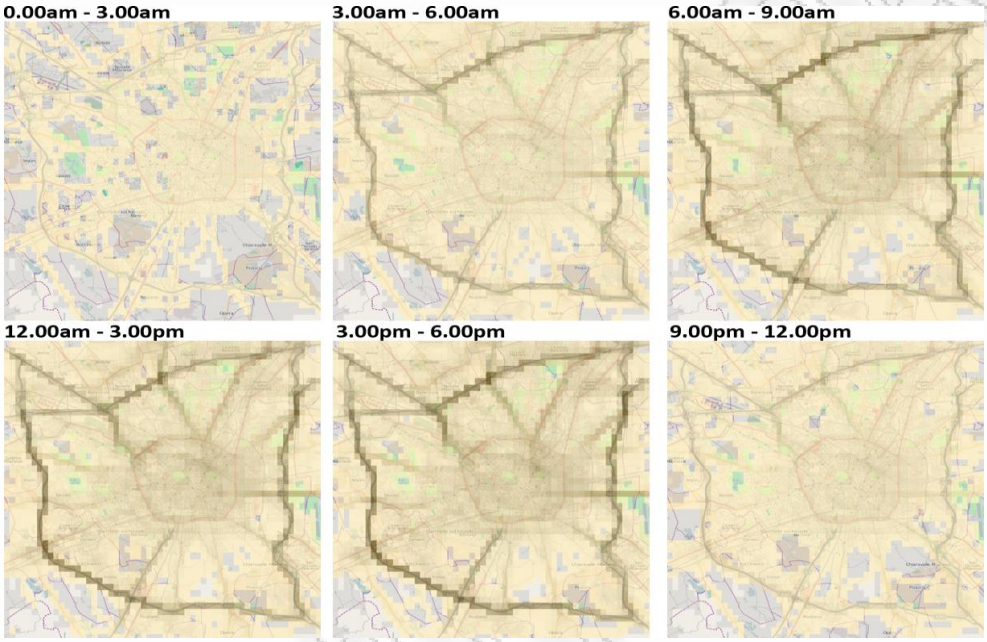
Για παράδειγμα, η Εικόνα 5-6 δείχνει τη στιγμιαία απεικόνιση της κίνησης των αντικειμένων που επιδεικνύει την μεταβλητότητα της ταχύτητας και της παρουσίας του ανά μία ώρα μέσα σε όλη την εβδομάδα, χρησιμοποιώντας τη τριγωνική μορφή. Το ύψος του τριγώνου αναπαριστά την Ταχύτητα (Velocity), ενώ η βάση την Παρουσία (Presence). Παρατηρήστε τον υποκείμενο χάρτη του Μιλάνου, με τον οποίο μπορούμε να έχουμε καλύτερη κατανόηση του φαινομένου της κυκλοφοριακής κίνησης. Η παρουσία των αυτοκινήτων είναι εντονότερη στο κέντρο λόγω του ότι η ταχύτητα είναι πολύ χαμηλή. Το αντίθετο ισχύει κατά μήκος των περιφερειακών δρόμων όπου η ταχύτητα είναι υψηλότερη, εκτός από τη βόρειο-ανατολική ζώνη, όπου ο μεγάλος αριθμός αυτοκινήτων επιβραδύνει την κυκλοφοριακή κίνηση.



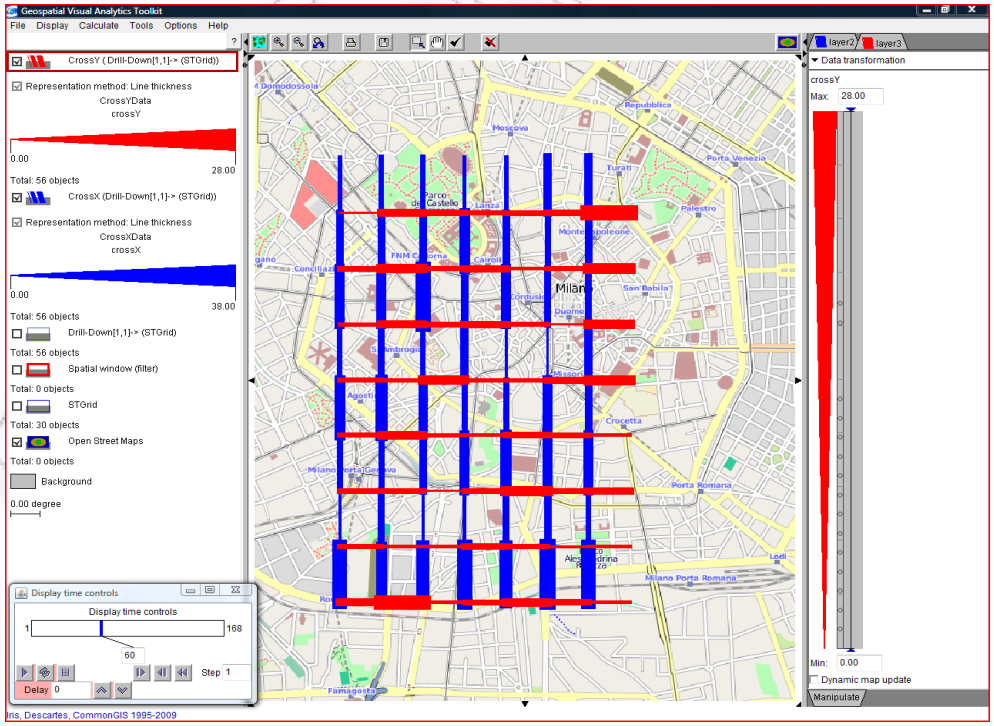
Εικόνα 5-6: Σχέση μεταξύ των μέτρων Presence και Velocity.

Ο *μη κατηγοριοποιημένος χωροπληθικός χάρτης* (unclassified choropleth map) είναι ένας τύπος οπτικοποίησης, στον οποίο όλα τα κελιά του πλέγματος έχουν σκίαση, σύμφωνα με την κατηγοριοποίηση της τιμής του επιλεγμένου μεγέθους της TDW. Η γραφική απεικόνιση αυτού του τύπου δίνεται στην Εικόνα 5-7, όπου περιγράφονται 6 στιγμιαίες απεικονίσεις αντίστοιχα στα διαστήματα από 0-3 πμ., 3-6 πμ. 6-9 πμ., 12πμ.-3μμ., 3-6μμ., 9-12μμ. της Τρίτης (μιας εργάσιμης ημέρας). Οι εικόνες μάς δίνουν μια ποιοτική άποψη του μέτρου PRES: όσο πυκνότερη είναι η κυκλοφοριακή κίνηση, τόσο σκουρότερο είναι το χρώμα του κελιού. Σε σύγκριση με το πλέγμα στην

Εικόνα 5-6, αυτό το χρωματισμένο επίπεδο της χωρικής κλιμάκωσης δίνει έμφαση στο οδικό δίκτυο: μερικοί δρόμοι-δακτυλίδια γύρω από το κέντρο, και κάποιες ακτινωτές οδοί χρησιμοποιούνται για την είσοδο/έξοδο προς/από το κέντρο. Κατά τη διάρκεια των ωρών αιχμής, η κυκλοφοριακή κίνηση αυξάνεται στο κέντρο της πόλης, καθώς επίσης και στις κεντρικές οδούς. Στο διάστημα 0-3π.μ. κυκλοφορούν λίγα αυτοκίνητα αφού δεν υπάρχουν πυκνές περιοχές, μετά η κυκλοφορία στην εξωτερική περιφερειακή οδό της πόλης γίνεται πυκνότερη και αργότερα, πυκνώνει και στις εσωτερικές περιφερειακές οδού και στις ακτινωτές οδούς γύρω από το κέντρο.



Εικόνα 5-7: Το μέτρο Presence την Τρίτη στο χαμηλότερο επίπεδο κλιμάκωσης.

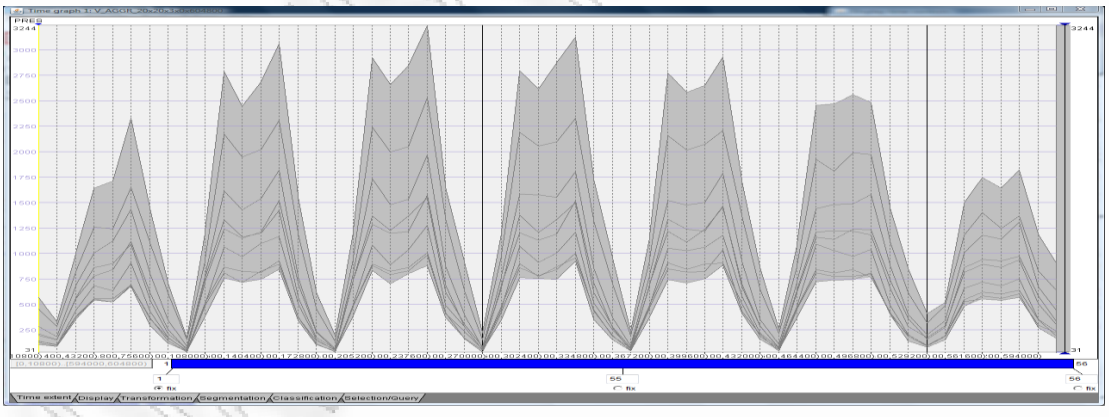


Εικόνα 5-8: Οπτικοποίηση των CROSSX και CROSSY.

Ο τύπος οπτικοποίησης *Πάχος Γραμμής* (*Line thickness*), αντίθετα, επιτρέπει το σχεδιασμό γραμμών των οποίων το πάχος είναι ανάλογο με την τιμή που έχει δοθεί στο μέγεθος της TDW. Στη στιγμιαία απεικόνιση της Εικόνα 5-8, οι γραμμές αυτές χρησιμοποιούνται για να οπτικοποιήσουν τα cross μεγέθη (που υποδηλώνουν το πλήθος των αντικειμένων που περνούν από ένα κελί βάσης σε ένα γειτονικό του). Το μέγεθος CROSSX (τομή στον άξονα X) αναπαρίσταται από τις οριζόντιες γραμμές, ενώ το μέγεθος CROSSY (τομή στον άξονα Y) από τις κάθετες γραμμές.
Η μέθοδος οπτικοποίησης του περιγράφηκε μπορεί να παράγει επίσης κινούμενα σχέδια, στην οποία κάθε πλαίσιο μπορεί να αναπαριστά ένα επιλεγμένο μέγεθος στο χρονικό διάστημα της περιόδου που μας ενδιαφέρει. Με αυτό τον τρόπο, ο χρήστης μπορεί να αποκτήσει μια οπτική αναπαράσταση της μεταβλητότητας του κάθε μεγέθους (ων) σε διαφορετικές ζώνες του επιλεγμένου χώρου και κατά τη διάρκεια διαφορετικών χρονικών διαστημάτων.

Άλλος τύπος οπτικοποίησης είναι η *Χρονική Γραφική Απεικόνιση* (*Time Graph)* που παράγει ένα γράφημα που δείχνει τη χρονική εξέλιξη ενός επιλεγμένου μεγέθους.

Για παράδειγμα, στην Εικόνα 5-9 αναπαριστάται με γράφημα ο χρόνος για την εξέλιξη του μεγέθους PRES, κατά τη διάρκεια μιας εβδομάδας, από Κυριακή μέχρι Σάββατο, σε μια κλίμακα 6 x 8 km^2 για τη χωρική διάσταση και ενός 3-ωρου χρονικού διαστήματος για τη χρονική διάσταση. Φαίνεται ξεκάθαρα ότι μέσα στις μέρες της εβδομάδας: η κυκλοφοριακή κίνηση μεγαλώνει κατά τη διάρκεια της ημέρας και ελαττώνεται αργότερα την ίδια μέρα. Επιπλέον, κατά τη διάρκεια του σαββατοκύριακου η παρουσία είναι λιγότερο έντονη από αυτήν κατά τις εργάσιμες ημέρες της εβδομάδας. Αξιοσημείωτο είναι ότι κάθε καμπύλη του γραφήματος σχετίζεται με ένα κελί του πλέγματος και αυτή η αντιστοιχία αναδεικνύεται όταν επιλέγουμε την καμπύλη.



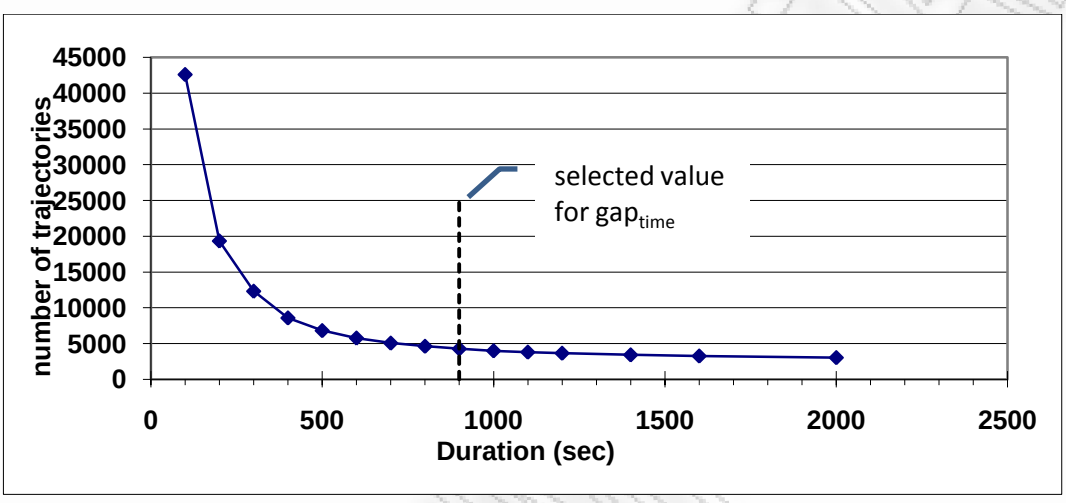
Εικόνα 5-9: Η εξέλιξη του μέτρου Presence κατά τη διάρκεια της εβδομάδας.

5.5. Πειραματική Μελέτη

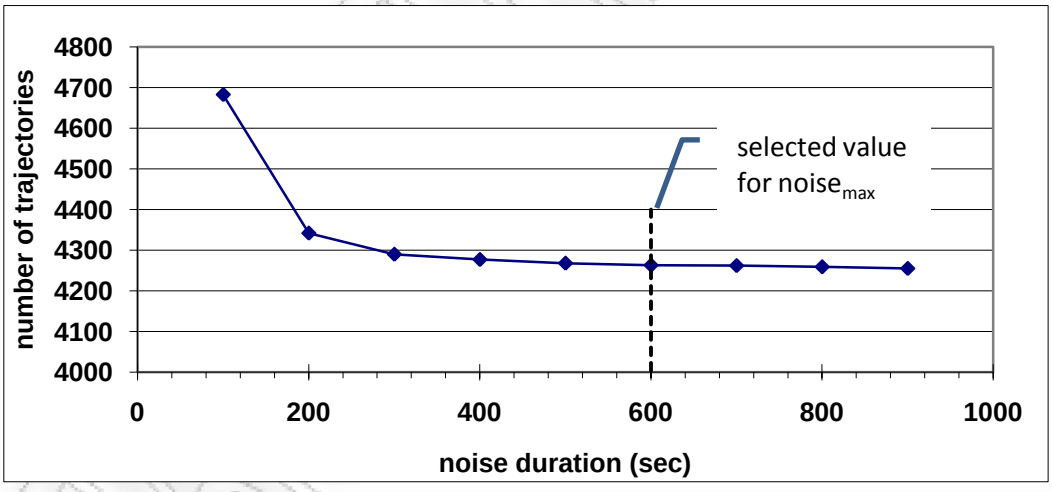
Σε αυτό το κεφάλαιο, ασχολούμαστε με την αξιολόγηση του αλγορίθμου ανακατασκευής τροχιών. Όπως και στο Κεφάλαιο 3, χρησιμοποιήσαμε το σύνολο δεδομένων E-Courier [Eco09] αποτελούμενο από 6,67 εκατομμύρια πρωτογενείς καταγραφές (ένα αρχείο 504 Mb, συνολικά) που αντιπροσωπεύουν την κίνηση 84 υπαλλήλων μιας εταιρίας ταχυμεταφορών στην ευρύτερη περιοχή του Λονδίνου

(περιοχή κάλυψης 66,800 km^2) για περίοδο ενός μηνός (Ιούλιο 2007) με συχνότητα λήψης δείγματος 10 δευτερολέπτων. Για όλα τα πειράματα χρησιμοποιήσαμε ένα PC με 1 Gb RAM και P4 3 GHz CPU.

Οι τιμές των παραμέτρων για την ανακατασκευή τροχιών που επιλέχθηκαν είναι οι εξής: gap_{time} = 900 sec, gap_{space} = 5 Km, V_{max} = 50 m/s, $noise_{max}$ = 600 sec and D_{tol} = 20 m, και το μέγεθος του συνόλου των πρωτογενών δεδομένων ποικίλλει από 0.5 εκατομμύρια καταγραφές μέχρι το μέγιστο διαθέσιμο όγκο. Αυτή η ρύθμιση έδωσε αποτελέσματα 4263 τροχιών (που αντιστοιχεί σε ένα μέσο μέγεθος 1.64 τροχιών ανά ταχυμεταφορέα την ημέρα).



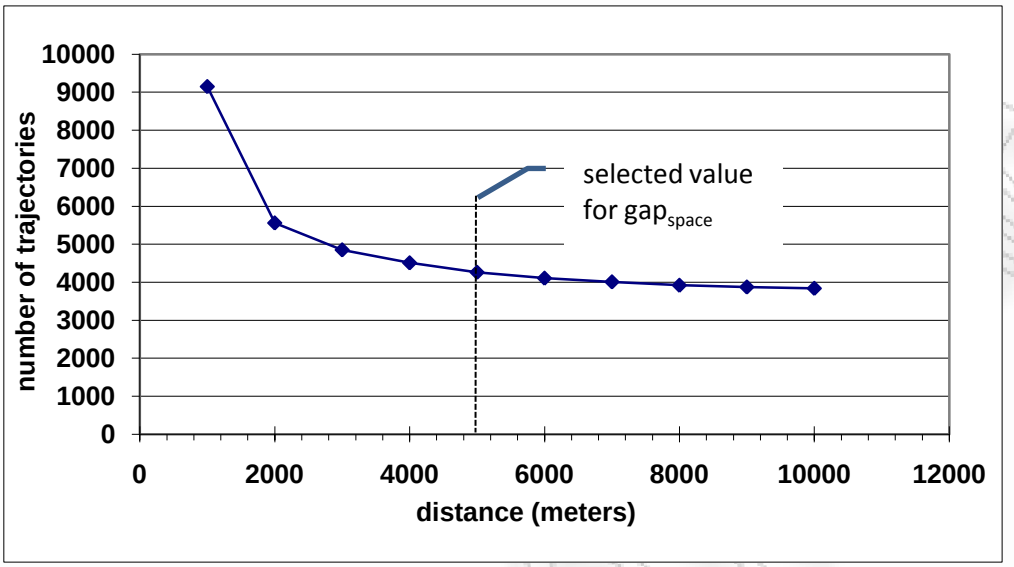
Εικόνα 5-10: Η επίδραση της παραμέτρου για το χρονικό κενό.



Εικόνα 5-11: Η επίδραση της παραμέτρου για τη διάρκεια θορύβου.

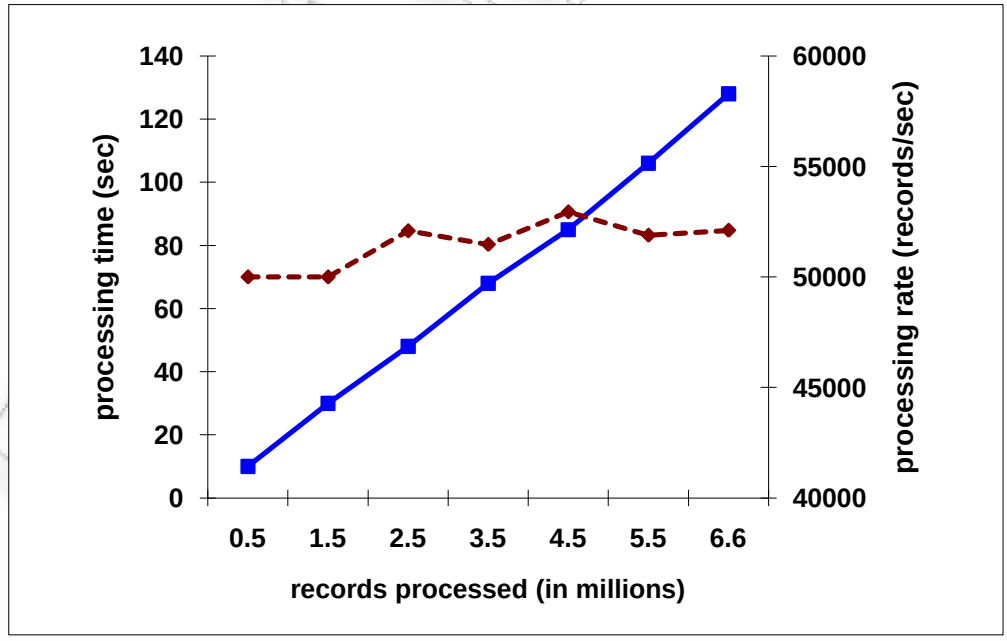
Καταλήξαμε στις παραπάνω αναφερόμενες τιμές μετά από ανάλυση της συμπεριφοράς διαφορετικών υποψήφιων τιμών και εκτιμώντας τον αριθμό των παραγόμενων τροχιών. Αρχικά θεωρήσαμε ότι η επιλεγμένη τιμή για τη μέγιστη ταχύτητα των οχημάτων είναι λογική. Μετά, διαπιστώσαμε ότι διαφορετικές τιμές της απόσταση ανοχής δεν έχουν επίδραση στον αριθμό των παραγόμενων τροχιών. Παρακάτω, απεικονίζονται τα αποτελέσματα της ανάλυσης των παραμέτρων: το χρονικό κενό (Εικόνα 5-10), τη μέγιστη διάρκεια θορύβου (Εικόνα 5-11) και το χωρικό κενό (Εικόνα 5-12). Όπως προκύπτει για κάθε παράμετρο, ο αριθμός των παραγόμενων

τροχιών φαίνεται να παραμένει σταθερός από το επιλεγμένο σημείο και μετά.



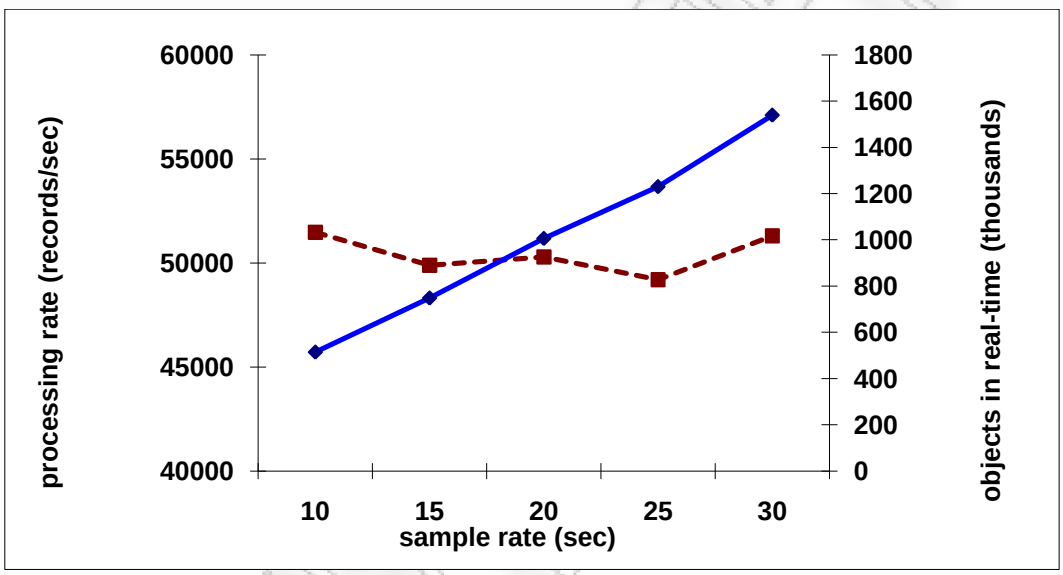
Εικόνα 5-12: Η επίδραση της παραμέτρου για το χωρικό κενό.

Το επόμενο πείραμα, παρουσιάζεται στην Εικόνα 5-13, αφορά την αποδοτικότητα του αλγορίθμου TRAJECTORY-RECONSTRUCTION που προτάθηκε στην Υποενότητα 5.3. Φαίνεται ξεκάθαρα ότι η επίδοση του αλγορίθμου TRAJECTORY-RECONSTRUCTION εξελίσσεται γραμμικά με το μέγεθος των εισαγόμενων δεδομένων (και επιτρέπει την επεξεργασία όλου του συνόλου των δεδομένων σε περίπου 2 λεπτά). Επιπλέον, ο μέσος ρυθμός επεξεργασίας είναι σχεδόν σταθερός (~50K καταγραφές/δευτερόλεπτο) .



Εικόνα 5-13: Η απόδοση του αλγορίθμου ανακατασκευής τροχιών
(συνεχής γραμμή: χρόνος επεξεργασίας, διακεκομμένη γραμμή: ρυθμός επεξεργασίας).

Στη συνέχεια των πειραμάτων, αξιολογούμε την αποδοτικότητα του αλγορίθμου TRAJECTORY-RECONSTRUCTION στην περίπτωση μεγάλου αριθμού χρηστών και με το στόχο την επίτευξη επεξεργασίας δεδομένων σε πραγματικό χρόνο. Πιο συγκεκριμένα, μετράμε το χρόνο επεξεργασίας για διαφορετικές τιμές δειγμάτων, όπως απεικονίζεται στην Εικόνα 5-14. Σύμφωνα με το πείραμα, επιλέγοντας π.χ. ένα δείγμα με συχνότητα δειγματοληψίας τα 20 δευτερόλεπτα ο αλγόριθμος μπορεί να διαχειριστεί σε πραγματικό χρόνο μέχρι 1000K αντικείμενα, το οποίο θεωρείται μεγάλος αριθμός για εφαρμογές σε πραγματικές συνθήκες (τουλάχιστον με τα σημερινά δεδομένα). Το συμπέρασμα από την εκτέλεση του συγκεκριμένου πειράματος είναι ότι ο προτεινόμενος αλγόριθμος TRAJECTORY-RECONSTRUCTION είναι αποδοτικός για επεξεργασία δεδομένων σε πραγματικό χρόνο, διατηρώντας ισορροπία μεταξύ του αριθμού των χρηστών που θα υποστηριχθούν και του ρυθμού δειγματοληψίας που τέθηκε για αποστολή της θέσης τους.



Εικόνα 5-14: Η επίδραση του ρυθμού δειγματοληψίας στην επεξεργασία σε πραγματικό χρόνο (συνεχής γραμμή: πλήθος αντικείμενων σε πραγματικό χρόνο, διακεκομμένη γραμμή: ρυθμός επεξεργασίας).

5.6. Σύνοψη

Σε αυτό το κεφάλαιο, παρουσιάστηκε το T-WAREHOUSE, ένα σύστημα για Οπτική Αποθήκευση των Δεδομένων Τροχιών, που ενσωματώνει όλα τα απαραίτητα βήματα, από την ανακατασκευή τροχιών και επεξεργασία μέσω μιας ETL διαδικασίας σε Οπτική OLAP Ανάλυση. Συγκεκριμένα για το πρώτο βήμα, προτείναμε μια τεχνική *ανακατασκευής τροχιών* που μετατρέπει τις σειρές των πρωτογενών δεδομένων θέσης σε τροχιές. Περιγράψαμε, στη συνέχεια, την αρχιτεκτονική του πλαισίου μας και στην τελευταία ενότητα, εξετάσαμε τις δυνατότητες, την προσαρμοστικότητα και αποδοτικότητα του πλαισίου μας για εφαρμογή αναλυτικών OLAP μεθόδων σε πραγματικά δεδομένα κινούμενων αντικειμένων.

6. Επίλογος

6.1. Συμπεράσματα

Στη διατριβή αυτή μελετήσαμε διάφορες τεχνικές για αποτελεσματική *Αποθήκευση και Εξόρυξη Δεδομένων Κίνησης* (Mobility Data Warehousing & Mining) που βρίσκονται σε Βάσεις Κινούμενων Αντικειμένων. Ειδικότερα, ο στόχος μας ήταν να συνεισφέρουμε σε επιμέρους βήματα μιας διαδικασίας *Ανακάλυψης Γνώσης από Γεωγραφικά Δεδομένα*. Στις επόμενες παραγράφους συνοψίζουμε τις κύριες συνεισφορές αυτής της διατριβής.

Στο Κεφάλαιο 2, ξεκαθαρίσαμε τη διαφορά μεταξύ των δυο κατηγοριών χωροχρονικών δεδομένων: στατικών και κίνησης. Συζητήσαμε σχετικά με τα σεισμολογικά δεδομένα ως ένα παράδειγμα της πρώτης κατηγορίας και προτείναμε ένα *Σύστημα Διαχείρισης και Εξόρυξης Γνώσης Σεισμολογικών Δεδομένων* που υποστηρίζει όλα τα απαραίτητα βήματα: από τη συλλογή δεδομένων μέχρι την εξόρυξη προτύπων. Σε ότι αφορά τα δεδομένα κίνησης, παρουσιάσαμε κάποιες βασικές αρχές δεδομένων τροχιών και θέσαμε τις προδιαγραφές για ένα πλαίσιο *Αποθήκευσης και Εξόρυξης Δεδομένων Κίνησης*.

Στο Κεφάλαιο 3, προτείναμε λύσεις για αποτελεσματική και αποδοτική ανάπτυξη Αποθηκών Δεδομένων Τροχιών Κινούμενων Αντικειμένων. Πιο συγκεκριμένα, προτείναμε τεχνικές για την υποστήριξη διαδικασιών ETL σε δεδομένα τροχιών με στόχο την τροφοδοσία των κύβων τροχιών καθώς και για την επίλυση θεμάτων συσσώρευσης δεδομένων εστιάζοντας ιδιαίτερα στο πρόβλημα της *μοναδικής προσμέτρησης*. Σε ότι αφορά την ETL διαδικασία, παρουσιάσαμε δυο διαφορετικές μεθοδολογίες: μια (βασισμένη σε ευρετήρια) *προσανατολισμένη στα κελιά* του κύβου και μια (χωρίς ευρετήριο) *προσανατολισμένη στις τροχιές*. Αναφορικά με το πρόβλημα της *μοναδικής προσμέτρησης*, προτείναμε την ενσωμάτωση κάποιων βοηθητικών μέτρων στο μοντέλο του κύβου δεδομένων που θα επιτρέψουν τη διόρθωση των ανακριβειών που προκύπτουν όταν εκτελείται λειτουργία συσσώρευσης (roll-up). Οι προτεινόμενες προσεγγίσεις έχουν ελεγχθεί πειραματικά και τα αποτελέσματα αποδεικνύουν την αποδοτικότητα τους. Το συμπέρασμα που προκύπτει είναι ότι η επιλογή της κατάλληλης ETL διαδικασίας αποτελεί συνάρτηση του επιθυμητού βαθμού κλιμάκωσης και τον αριθμό των τροχιών. Αναφορικά με την προσεγγιστική μας μέθοδο για το *πρόβλημα της μοναδικής προσμέτρησης*, τα αποτελέσματα δείχνουν ότι αποτελεί μια αποτελεσματική λύση λαμβάνοντας υπόψη την απόδοση μιας διανεμητικής (distributive) συνάρτησης συγκέντρωσης.

Επιπλέον στο Κεφάλαιο 3, προτείναμε μια καινοτόμο οργάνωση ενός κύβου δεδομένων τροχιών που λαμβάνει υπόψη του τις διαφορετικές ερμηνείες που μπορούν να δοθούν στην έννοια της τροχιάς

εξυπηρετώντας έτσι όχι μια αλλά περισσότερες εφαρμογές. Πιο συγκεκριμένα, επεκτείναμε το OLAP μοντέλο δεδομένων ώστε να είναι αρκετά ευέλικτο με τις διαφορετικές ερμηνείες της τροχιάς, προτείναμε κατάλληλους ETL και OLAP μηχανισμούς που εκμεταλλεύονται το νέο μοντέλο και συζητήσαμε θέματα προϋπολογισμού (materialization) σε ένα κύβο δεδομένων. Στη πειραματικής μας μελέτη χρησιμοποιήσαμε το ίδιο σύνολο δεδομένων με αυτό του προηγούμενη πλαισίου και έτσι προχωρήσαμε σε σύγκριση μεταξύ των δυο μεθόδων. Από αυτήν την πειραματική μελέτη προκύπτει ξεκάθαρα ότι κατά περίπτωση (ad-hoc) προσέγγιση που προτείναμε είναι χρήσιμη σε περίπτωση που ένα κύβος δεδομένων τροχιών πρέπει να εξυπηρετεί πολλές εφαρμογές.

Στο Κεφάλαιο 4 παρουσιάστηκε ένα πλαίσιο για εξόρυξη προτύπων αλληλεπίδρασης που προσφέρει χωροχρονική αναπαράσταση, σύνθεση και κατηγοριοποίηση των τροχιών κινούμενων αντικειμένων επιτρέποντας έτσι την ολοκληρωμένη κατανόηση του τι συμβαίνει στο ευρύτερο περιβάλλον των κινούμενων αντικειμένων (ο γεωγραφικός χώρος, η αλληλεπίδραση με άλλα αντικείμενα κτλ). Το πλαίσιο μας υποστηρίζει τον υπολογισμό διάφορων περιγραφέων αλληλεπίδρασης (interaction descriptors) είτε σε μια γειτονιά που μπορεί να ορίζεται με δυναμικό τρόπο είτε λαμβάνοντας υπόψη ένα σταθερό πλέγμα. Αυτοί οι περιγραφείς μάς παρέχουν γνώση σχετικά με τις τάσεις της κίνησης σε διάφορα χωροχρονικά παράθυρα δίνοντας μας έτσι πληροφορίες σχετικά με την αλληλεπίδραση των κινούμενων αντικειμένων μέσα σε αυτά τα παράθυρα. Τα πρώτα πειραματικά αποτελέσματα αποδεικνύουν ότι η προσέγγισή μας είναι εφαρμόσιμη και αποτελεσματική και περιλαμβάνουν σύγκριση μεταξύ των δυο μεθόδων (δυναμική και σταθερή) καθώς και ανάλυση ακρίβειας με χρήση διαφορετικών περιγραφέων κτλ.

Στο Κεφάλαιο 4 επίσης συζητήσαμε το πρόβλημα της διαχείρισης κυκλοφορίας σε ένα οδικό δίκτυο λαμβάνοντας υπόψη τη χρονοσειρά κίνησης για κάθε ακμή του δικτύου. Ορίσαμε μέτρα ομοιότητας μεταξύ των χρονοσειρών κίνησης και προτείναμε ένα αλγόριθμο συσταδοποίησης που ομαδοποιεί τις ακμές με βάση αυτά τα μέτρα. Πιο συγκεκριμένα, η προσέγγισή μας παρέχει μια ιεραρχία ομάδων παρόμοιων ακμών. Η ιεραρχία αυτή μπορεί να είναι τριών επιπέδων και να περιλαμβάνει ομάδες με παρόμοιο σχήμα χρονοσειρών, ομάδες ακμών που βρίσκονται κοντά και ακμές με παρόμοιες τιμές κίνησης. Επίσης, ορίσαμε διαφορετικές σχέσεις κυκλοφορίας (διάδοση, διάσπαση, συγχώνευση, πηγή, προορισμός) μεταξύ των ακμών του δικτύου ώστε να συλλάβουμε τον τρόπο με τον οποίο η κυκλοφορία διαχέεται μέσα στο δίκτυο και προτείναμε έναν αλγόριθμο ανακάλυψης αυτών των σχέσεων. Τα δεδομένα που απαιτούνται για αυτήν την ανάλυση μπορούν να συλλεχθούν είτε χρησιμοποιώντας αισθητήριες συσκευές που παρακολουθούν την κίνηση σε κάθε ακμή ή συναθροίζοντας τις θέσεις αντικειμένων σε χρονοσειρές κίνησης. Κάποιος θα μπορούσε να ισχυριστεί ότι το συγκεκριμένο πλαίσιο αποτελεί μια διαφορετική υποδομή σε σχέση με αυτή που παρουσιάστηκε στα προηγούμενα κεφάλαια της παρούσας διατριβής. Στην πραγματικότητα η αρχιτεκτονική της αποθήκης δεδομένων για τροχιές κινούμενων αντικειμένων που παρουσιάστηκε στο Κεφάλαιο 3 μπορεί να χρησιμοποιηθεί για τη φύλαξη των δεδομένων κυκλοφορίας. Το οδικό δίκτυο μπορεί να μοντελοποιηθεί ως η χωρική διάσταση της αποθήκης, οι διαφορετικές χρονικές περίοδοι ως η χρονική διάσταση και τα μέτρα καταγράφουν συγκεντρωτικές πληροφορίες για την ανάλυση των δεδομένων κυκλοφορίας. Τα πειραματικά αποτελέσματα κατάδειξαν ότι οι δυο μέθοδοι που παρουσιάστηκαν μπορούν να μας δώσουν γνώση για τη διάχυση της κίνησης σε ένα οδικό δίκτυο. Για παράδειγμα,

δείξαμε πώς μπορούν να εντοπιστούν ομάδες ακμές με παρόμοιο σχήμα χρονοσειρών και πως αυτές οι ομάδες διαμορφώνονται όταν ληφθούν υπόψη και τα μέτρα της εγγύτητας (μεταξύ των ακμών) ή/και της ομοιότητας των χρονοσειρών με βάση τις τιμές τους. Επίσης, παρουσιάσαμε μια μελέτη για τον βαθμό ακρίβειας της μεθόδου εντοπισμού σχέσεων μεταξύ των ακμών του οδικού δικτύου που αποδεικνύει την επίδραση που έχουν τα διαφορετικά μέτρα ομοιότητας.

Τέλος, στο Κεφάλαιο 5, παρουσιάσαμε το T-WAREHOUSE, ένα σύστημα που ενσωματώνει όλα τα απαραίτητα βήματα για οπτική ανάλυση σε μια αποθήκη δεδομένων τροχιών κινούμενων αντικειμένων, από την *ανακατασκευή τροχιών* και την ETL επεξεργασία μέχρι την οπτική OLAP ανάλυση. Ειδικότερα, περιγράφηκαν οι αρχιτεκτονικές αρχές του πλαισίου και εξετάστηκαν οι δυνατότητές του, ο βαθμός ευελιξίας και η αποτελεσματικότητά του κατά την εφαρμογή OLAP ανάλυσης σε πραγματικά δεδομένα κίνησης. Επίσης, παρουσιάσαμε έναν αποδοτικό αλγόριθμο μετατροπής μιας αλληλουχίας πρωτογενών δεδομένων θέσης σε τροχιές και εξετάσαμε την απόδοση χρησιμοποιώντας ένα πραγματικό, μεγάλο σύνολο δεδομένων. Η πειραματική μελέτη κατέδειξε ότι το προτεινόμενο πλαίσιο είναι αποδοτικό ακόμα και για επεξεργασία πραγματικού χρόνου.

6.2. Ανοικτά Θέματα

Διάφορα ερευνητικά θέματα παραμένουν ανοικτά στο χώρο της Αποθήκευσης Δεδομένων & Εξόρυξης Γνώσης από δεδομένα κίνησης αντικειμένων. Στις ακόλουθες παραγράφους παρουσιάζουμε το μελλοντικό ερευνητικό έργο που πηγάζει από την πρόοδό μας σε αυτήν την διατριβή.

Σε ότι αφορά τις TDW, μπορούν να εξεταστούν νέα μέτρα που είναι κατάλληλα για τροχιές. Ένα παράδειγμα τέτοιου μέτρου η επονομαζόμενη *τυπική τροχιά* (π.χ. [GNP+07], [LHW07]) που περιγράφει την τάση της κίνησης μέσα σε ένα κελί. Η πρόκληση σε αυτό το πρόβλημα είναι μεγάλη αφού δε μπορεί να εξαχθεί εύκολα η αντιπροσωπευτική τροχιά ενός κυβοειδούς βάσει των αντιπροσωπευτικών τροχιών των υπο- κυβοειδών του. Η αντιπροσωπευτική τροχιά ενός κυβοειδούς μπορεί να θεωρηθεί ως μια ζώνη (buffer) τροχιάς που αναπαριστά το πρότυπο της κίνησης μέσα σε ένα κελί. Ας θεωρήσουμε ότι $D = \{T_1, T_2, \ldots, T_n\}$ είναι ένα σύνολο τροχιών που ορίζονται μεταξύ των χρονικών στιγμών t_1 και t_m. Ας θεωρήσουμε επίσης ότι τ_s είναι μια *ανοχή* στο *χωρικό* επίπεδο. Σε μια συγκεκριμένη χρονική στιγμή t_i, δυο τροχιές T_1, T_2 έχουν τ_s ανοχή αν οι αντίστοιχες χωρικές συντεταγμένες τους διαφέρουν λιγότερο από τ_s. Αν αυτό ισχύει για όλες τις στιγμές t_1 και t_m τότε οι T_1, T_2 περιλαμβάνονται στο ίδιο buffer τροχιάς. Αν αυτό συμβαίνει για κάθε ζευγάρι μέσα στο *D*, τότε δηλώνουμε ότι τροχιές του *D* σχηματίζουν ένα buffer τροχιάς *TB*. Ο αριθμός των τροχιών που περιλαμβάνεται στο buffer καλείται ως *υποστήριξη* του buffer. Η ανακάλυψη τέτοιων αντιπροσωπευτικών τροχιών είναι ένα διττό πρόβλημα: ανακάλυψη αντιπροσωπευτικών τροχιών από πρωτογενή δεδομένα και από άλλες αντιπροσωπευτικές τροχιές. Το πρώτο είναι ένα πρόβλημα συσταδοποίησης ενώ το δεύτερο είναι πρόβλημα συσσώρευσης στα πλαίσια μιας TDW αφού η αντιπροσωπευτική τροχιά ενός κυβοειδούς θα πρέπει να υπολογίζεται από τα κυβοειδή κατώτερου επιπέδου (προκειμένου να αποφύγουμε να εκτελέσουμε την εργασία συσταδοποίησης στα δεδομένα του συσσωρευμένου κυβοειδούς).

Σε μια παρόμοια γραμμή έρευνας, άλλα ενδιαφέροντα μέτρα μπορούν να ανακαλυφθούν όπως η *μέση κατεύθυνση* των τροχιών μέσα σε ένα κελί. Επίσης, αξίζει να μελετηθούν θέματα προϋπολογισμού (materialization) σε Αποθήκες Δεδομένων Τροχιών Κινούμενων Αντικειμένων. Πιο συγκεκριμένα, να αναπτυχθούν τεχνικές που να μπορούν να αποφασίσουν ποια κυβοειδή πρέπει να προϋπολογιστούν και πως αυτά μπορούν να αξιοποιηθούν για να απαντηθούν OLAP ερωτήματα. Στις παραδοσιακές ΑΔ, μια κοινή προσέγγιση είναι να χρησιμοποιηθεί ο μικρότερος γονέας (smallest parent) ώστε να απαντηθεί ένα ερώτημα. Αυτό όμως στην περίπτωση μιας TDW ίσως είναι σοφότερο να επιλοεγεί το κυβοειδές που ελαττώνει το σφάλμα (εξαιτίας του προβλήματος μοναδικής προσμέτρησης).

Η προσέγγιση για κατά περίπτωση (ad-hoc) ανάλυση OLAP πάνω σε TDW που παρουσιάστηκε σε αυτήν την διατριβή μπορεί επίσης να επεκταθεί. Αξίζει να μελετηθούν μέτρα που εμπλέκουν τις μεταβάσεις από το ένα κελί σε κάποιο άλλο χρησιμοποιώντας το μοντέλο αποθήκευσης δεδομένων που προτάθηκε. Ένα παράδειγμα τέτοιου μέτρου είναι ο αριθμός των διακριτών τροχιών σε ένα σύνολο κελιών. Σε αυτήν την περίπτωση χρειαζόμαστε μια τεχνική που μπορεί να προσδιορίσει τροχιές που εξελίσσονται σε διαδοχικά κελιά. Επίσης, τα θέματα προϋπολογισμού ενός τέτοιου κύβου δεδομένων είναι ενδιαφέροντα. Χρειαζόμαστε τεχνικές που θα «σαρώνουν» τα δεδομένα και θα προτείνουν μέγιστες χρονικές αποστάσεις βάση των οποίων θα πραγματοποιείται ο προϋπολογισμός του κύβου.

Σε ότι αφορά την μέθοδο που προτείναμε για εξόρυξη προτύπων αλληλεπίδρασης, μια ερευνητική κατεύθυνση είναι αυτή της σύγκρισης της προτεινόμενης μεθόδους με επιπλέον κλασικές προσεγγίσεις χωροχρονικής συσταδοποίησης και κατηγοριοποίησης. Για παράδειγμα μπορεί να συγκριθεί με τα αποτελέσματα ενός αλγόριθμου συσταδοποίησης που βασίζεται στην πυκνότητα ώστε να αποκαλυφθούν οι ακραίες τιμές. Επίσης, οι περιγραφείς που υπολογίζονται από το πλαίσιο μας μπορούν να χρησιμοποιηθούν για OLAP λειτουργίες. Για παράδειγμα, αντί να εφαρμοστεί μια λειτουργία συσσώρευσης με σκοπό να επιτευχθεί συσσώρευση ενός υποσυνόλου των κελιών του κύβου, θα μπορούσαν να συναθροιστούν μόνο τα χωροχρονικά κελιά που περιλαμβάνουν συγκεκριμένα πρότυπα αλληλεπίδρασης (π.χ. «επικίνδυνη οδήγηση»). Επίσης, ένα ενδιαφέρον θέμα είναι η ανάπτυξη μιας έξυπνης υπολογιστικής μεθόδου που θα επιταχύνει τη φάση της συσσώρευσης.

Η μέθοδός μας για εξόρυξη προτύπων κυκλοφορίας μπορεί να εμπλουτιστεί ώστε να είναι δυνατή η ανακάλυψη πιο περίπλοκων σχέσεων μεταξύ των ακμών. Μέχρι τώρα, μελετήσαμε την διάδοση, διάσπαση και συγχώνευση της κυκλοφορίας. Μπορούν όμως να υπάρξουν και πιο πολύπλοκες περιπτώσεις όπως συνδυασμοί αυτών των σχέσεων π.χ. μια διάδοση που ακολουθείται από μια διάσπαση και στη συνέχεια από μια συγχώνευση κυκλοφορίας. Επιπλέον, ένα ενδιαφέρον θέμα είναι η συνεχής παρακολούθηση του οδικού δικτύου. Στη μέχρι τώρα έρευνά μας αναλύουμε την κυκλοφορία σε ένα οδικό δίκτυο για συγκεκριμένες χρονικές περιόδους. Εντούτοις, τα δεδομένα κυκλοφορίας είναι από τη φύση τους δυναμικά συνεπώς αξίζει να παρακολουθούμε τη ροή κυκλοφορίας σε πραγματικό χρόνο ώστε να μπορέσουμε να ανακαλύψουμε ανωμαλίες λαμβάνοντας υπόψη την τυπική συμπεριφορά (που προκύπτει από την ιστορική ανάλυση της κυκλοφορίας).

Τέλος, στο θέμα της *ανακατασκευής τροχιών*, μπορούμε να προτείνουμε ως επόμενα βήματα την ανακάλυψη ευφυών μεθόδων που θα ανακαλύπτουν δυναμικά τις τιμές των παραμέτρων που

χρησιμοποιούνται από τον προτεινόμενο αλγόριθμο (λαμβάνοντας υπόψη τα χαρακτηριστικά του συνόλου δεδομένων) και την επέκταση της συγκεκριμένης τεχνικής ώστε να είναι δυνατός ο εντοπισμός διαφορετικών τύπων κίνησης (πεζός, ποδήλατο, μοτοποδήλατο, αυτοκίνητο, φορτηγό κτλ) ώστε να προσαρμόζεται ανάλογα η διαδικασία ανακατασκευής.

7. Αναφορές

[ADH+03] van der Aalst, W. M., van Dongen, B. F., Herbst, J., Maruster, L., Schimm, G., and Weijters, A. J. Workflow mining: a survey of issues and approaches. *DKE* 47(2):237-267, 2003.

[AAD+96] Agarwal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J., Ramakrishnan, R., and Sarawagi. S. On the computation of multidimensional aggregates. *Proceedings of VLDB*, 1996.

[AIS93] Agrawal, R., Imielinski, T., and Swami, A. Mining Association Rules between Sets of Items in Large Databases. *Proceedings of ACM SIGMOD,* 2003.

[AGB06] Almeida, V. T., Güting, R. H., and Behr, T. Querying moving objects in secondo. *Proceedings of MDM,* 2006.

[AA99] Andrienko, G., Andrienko N., Knowledge-based visualization to support spatial data mining. *Proceedings of IDA*, 1999.

[AAW07] Andrienko, G., Andrienko N., and Wrobel, S. Visual Analytics Tools for Analysis of Movement Data. *ACM SIGKDD Explorations* 9(2): 28-46, 2007.

[BMR+08] Baglioni, M., Macedo, J., Renso, C., and Wachowicz, M. An Ontology-Based Approach for the Semantic Modeling and Reasoning on Trajectories. *Proceedings ER Workshops*, 2008.

[BMH01] Bédard, Y., Merrett, T., and Han, J. Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery. *Geographic Data Mining and Knowledge Discovery*, pp. 53-73. Taylor & Francis, 2001.

[BD00] Behnke, J., Dobinson, E., NASA Workshop on Issues in the Application of Data Mining to Scientific Data. *ACM SIGKDD Explorations* 2(1):70-79, 2000..

[BTM05] Bimonte, S., Tchounikine, A., and Miquel, M. Towards a spatial multidimensional model. *Proceedings* of *DOLAP*, 2005.

[BOO+07] Braz, F., Orlando, S., Orsini, R., Raffaetta, A., Roncato, A., and Silvestri, C. Approximate Aggregations in Trajectory Data Warehouses. *Proceedings of ICDE Workshop on Spatio-Temporal Data Mining*, 2007.

[Bri02] Brinkhoff, T. A Framework for Generating Network-Based Moving Objects. *GeoInformatica* 6(2):153-180, 2002.

[CCD+04] Castellanos, M., Casati, F., Dayal, U., Shan, M.C. A Comprehensive and Automated Approach to Intelligent Business Processes Execution Analysis. *Int. J. Distributed and Parallel Databases* 16:239-273, 2004.

[CD97] Chaudhuri, S., and Dayal, U. An overview of Data Warehousing and OLAP Technology. *SIGMOD Record* 26(1):65-74, 1997.

[Cia09] CIA. The World Factbook. 2009

[CYH+07] Cheng, H., Yan, X., Han, J. and Hsu, C.W. Discriminative Frequent Pattern Analysis for Effective Classification. *Proceedings of ICDE*, 2007.

[CKL06] Choi, W., Kwon, D., and Lee, S. Spatio-temporal data warehouses using an adaptive cell-based approach. *DKE* 59(1):189-207, 2006.

[DG03] Das, G. and Gunopulos, D. Time series similarity and indexing. *Chapter in Ye, N.(Ed.): The Handbook of Data Mining*, pp. 279–302, Lawrence Erlbaum Associates, 2003.

[DS06] Damiani, M.-L. and Spaccapietra, S. Spatial Data Warehouse Modelling. *Chapter in Processing and Managing Complex Data for Decision Support*, pp. 12-27, Idea Group Publishing, 2006.

[DKW+05] Deshpande, M., Kuramochi, M., Wale, N. and Karypis, G. Frequent Substructure-Based Approaches for Classifying Chemical Compounds. *TKDE* 17(8):1036-105, 2005.

[Doc06] Dockstader, S.L. Motion Trajectory Classification for Visual Surveillance and Tracking. *Proceedings of AVSS*, 2006.

[Eco09] eCourier.co.uk dataset, http://api.ecourier.co.uk/. (URL valid on December 14, 2009).

[FPS+96] Fayad, U., Piatetsky-Shapiro, G., Smith, P., and Uthurusami, R. *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996.

[FM85] Flajolet, P. and Martin, G. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.* 31(2):182-209, 1985.

[Geo06] GeoPKDD project, http://www.geopkdd.eu/. (URL valid on December 14, 2009).

[Geu01] Geurts, P. Pattern Extraction for Time Series Classification. *Proceedings of PKDD*, 2001.

[GNP+07] Giannotti, F., Nanni, M., Pinelli, F., and Pedreschi, D. Trajectory pattern mining. *Proceedings of KDD*, 2007.

[GP07] Giannotti, F., and Pedreschi, D. (Eds). Mobility, Data Mining and Privacy. Springer, 2007.

[GPT08] Giannotti, F., Pedreschi, D., and Turini, F. Mobility, Data Mining and Privacy the Experience of the GeoPKDD Project. *ACM SIGKDD PinKDD*, 2008.

[GKM+09] Gómez, L., Kuijpers, B., Moelans, B., Vaisman, A. A Survey of Spatio-Temporal Data Warehousing. *IJDWM* 5(3):28-55, 2009.

[GCB+97] Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. Data cube: A relational aggregation operator generalizing groub-by, cross-tab and sub-totals. *DMKD* 1(1):29-54, 1997.

[GBE+00] Güting, R. H., Böhlen, M. H., Erwig, M., Jensen, C. S., Lorentzos, N. A., Schneider, M., and Vazirgiannis, M. A foundation for representing and quering moving objects. *ACM Transactions on Database System* 25(1):1-42, 2000.

[GS05] Güting, R.H., and Schneider, M. Moving Object Databases, Morgan Kaufman Publishers. 2005.

[HK00] Han, J., and Kamber, M. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

[HSK98] Han, J., Stefanovic, N., and Koperski, K. Selective Materialization: An Efficient Method for Spatial Data Cube Construction. *Proceedings of PAKDD*, 1998.

[Inm96] Inmon, W. Building the Data Warehouse. John Wiley & Sons, 1996.

[JMF99] Jain, A., Murty, M., and Flynn, P. Data Clustering: A Review. *ACM Computing Surveys* 31(3):264–323, 1999.

[JKP+04] Jensen, C.S., Kligys, A., Pedersen, T.B., Dyreson, C.E., and Timko, I. Multidimensional data modeling for location-based services. *The VLDB Journal* 13(1):1–21, 2004.

[KMB05] Kalnis, P., Mamoulis, N. and Bakiras, S. On discovering moving clusters in spatio-temporal data. *Proceedings of SSTD*, 2005.

[KR90] Kaufman, L., and Rousseeuw, P.J. Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 1990.

[KP98] Keogh, E. and Pazzani, M. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. *Proceedings of KDD*, 1998.

[Kim96] Kimball, R. The Data Warehouse Toolkit. John Wiley & Sons, 1996..

[Kot02] Kotidis, Y. Aggregate View Management in Data Warehouses. *Chapter in Handbook of Massive Data Sets*, pp. 711-742, Kluwer Academic Publishers, 2002.

[Kot06] Kotidis, Y. Extending the data warehouse for service provisioning data. *DKE* 59(3):700-724, 2006.

[KR00] Kretschmer, U., Roccatagliata, E., CommonGIS: A European Project for an Easy Access to Geo-data. *Proceedings of EUGISES*, 2000.

[KNK+07] Kuijpers, B., Nanni, M., Korner, C., May, M. and Pedreschi, D. Spatiotemporal data mining. *Chapter in Giannotti, F. and Pedreschi, D. (Eds.): Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*, pp.275–300, Springer, 2008.

[LHL+08] Lee, J., Han, J., Li, X., and Gonzalez, H. TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering. *VLDB Endowment* 1(1), pp. 1081-1094, 2008.

[LHW07] Lee, J., Han, J., and Whang, K. Trajectory Clustering: A Partition-and-Group Framework. *Proceedings of ACM SIGMOD*, 2007.

[LMF+10] Leonardi, L., Marketos, G., Frentzos, E., Giatrakos, N., Orlando, S., Pelekis, N., Raffaetà, A., Roncato, A., Silvestri, C., Theodoridis, Y. T-Warehouse: Visual OLAP Analysis on Trajectory Data. *Proceedings of ICDE*, 2010.

[LOR+09] Leonardi, L., Orlando, R., Raffaetà, A., Roncato, A., and Silvestri, C Frequent spatio-temporal patterns in trajectory data warehouses. *Proceedings of SAC*, 2009.

[LHK06] Li, X., Han, J. and Kim, S. Motion-alert: automatic anomaly detection in massive moving objects. *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, 2006.

[LHL+07] Li, X., Han, J., Lee, J-G. and Gonzalez, H. Traffic density-based discovery of hot routes in road networks. *Proceedings of SSTD*, 2007.

[LCZ+06] Liu, Y., Choudhary, A.N., Zhou, J. and Khokhar, A.A. A scalable distributed stream mining system for highway traffic data. *Proceedings of ECML/PKDD*, 2006.

[LT05] Lopez, I., Snodgrass, R., and Moon, B. Spatiotemporal Aggregate Computation: A Survey. *TKDE* 2(17):271-286, 2005.

[MVO+07] Macedo, J., Vangenot, C., Othman, W., Pelekis, N., Frentzos, E., Kuijpers, B., Ntoutsi, I., Spaccapietra, S. and Theodoridis, Y. Trajectory data models. *Chapter in F. Giannotti and D. Pedreschi (eds), Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*. Springer, 2008.

[MZ04a] Malinowski, E. and Zimányi, E. OLAP Hierarchies: A Conceptual Perspective. *Proceedings of CAiSE*, 2004.

[MZ04b] Malinowski, E. and Zimányi, E. Representing Spatiality in a Conceptual Multi-dimensional Model. *Proceedings of ACM-GIS*, 2004.

[MZ06] Malinowski, E. and Zimányi, E. Hierarchies in a multidimensional model: From conceptual modeling to logical representation. *DKE* 59(2):348-377, 2006.

[MFN+08a]Marketos, G., Frentzos, E., Ntoutsi, I., Pelekis, N., Raffaeta, A., and Theodoridis, Y., Building Real-World Trajectory Warehouses. *Proceedings of MobiDE*, 2008.

[MFN+08b] Marketos, G., Frentzos, E., Giatrakos, N., Ntoutsi, I., Pelekis, N., Raffaeta, A., and Theodoridis, Y., A Framework for Trajectory Data Warehousing. *Proceedings of HDMS*, 2008.

[MT09a] Marketos, G., Theodoridis, Y., Ad-hoc OLAP on Trajectory Data. *Submitted*.

[MT06] Marketos, G., Theodoridis, Y., Measuring Performance in the retail industry. *Proceedings of BPI*, 2006.

[MT09b] Marketos, G., Theodoridis, Y., Mobility Data Warehousing & Mining. *Proceedings of VLDB PhD Workshop*, 2009.

[MTK08] Marketos, G., Theodoridis, Y., and Kalogeras, I., Seismological Data Warehousing and Mining – A Survey. *IJDWM* 4(1):1-16, 2008.

[NT04] Nakata, T. and Takeuchi, J. Mining traffic data from probe-car system for travel time prediction. *Proceedings of ACM SIGKDD*, 2004.

[NMO09] Nanni, M., Marketos, G., Quattrociocchi. W. Mining Interaction Patterns for Spatiotemporal Representation, Synthesis and Classification. *Manuscript prepared*.

[NMM08] Ntoutsi, I., Mitsou, N., Marketos, G., Traffic mining in a road-network: How does the traffic flow?. *IJBIDM* 3(1), 2008.

[Ope01] OpenGIS Consortium. Abstract Specification, Topic 1: Feature Geometry (ISO 19107 Spatial Schema), 2001. http://www.opengeospatial.org. (URL valid on December 14, 2009).

[OOR+07] Orlando, S., Orsini, R., Raffaetà, A., Roncato, A., and Silvestri, C. Trajectory Data Warehouses: Design and Implementation Issues. *JCSE* 1(2):240-261, 2007.

[PKZ+01] Papadias, D., Kalnis, P., Zhang, J., and Tao, Y. Efficient OLAP Operations in Spatial Data Warehouses. *Proceedings of SSTD*, 2001.

[PTK+02] Papadias, D., Tao, Y., Kalnis, P., and Zhang, J. Indexing Spatio-Temporal Data Warehouses. *Proceedings of ICDE*, 2002.

[PT01] Pedersen, T. and Tryfona, N. Pre-aggregation in Spatial Data Warehouses. *Proceedings of SSTD*, 2001.

[PFG+08] Pelekis, N., Frentzos, E., Giatrakos, N. and Theodoridis, Y. HERMES: Aggregative LBS via a Trajectory DB Engine. *Proceedings of ACM SIGMOD*, 2008.

[PKM+07] Pelekis, N., Kopanakis, I., Marketos, G., Ntoutsi, I., Andrienko, G., and Theodoridis, Y., Similarity Search in Trajectory Databases. *Proceedings of TIME*, 2007

[PRD+08] Pelekis, N., Raffaetà, A., Damiani, M.-L., Vangenot, C., Marketos, G., Frentzos, E., Ntoutsi, I., and Theodoridis, Y. Towards Trajectory Data Warehouses. *Chapter in Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*. Springer-Verlag. 2008.

[PTV+06] Pelekis, N., Theodoridis, Y., Vosinakis, S. and Panayiotopoulos, T. Hermes - A Framework for Location-Based Data Management. *Proceedings of EDBT*, 2006.

[PJT00] Pfoser, D., Jensen, C.S., and Theodoridis, Y. Novel Approaches to the Indexing of Moving Object Trajectories, *Proceedings of VLDB*, 2000.

[PT98] Pfoser, D., and Tryfona, N. Requirements, Definitions and Notations for Spatiotemporal Application Environments. *Proceedings of ACM-GIS*, 1998.

[RZY+03] Rao, F., Zhang, L., Yu, X., Li, Y., and Chen, Y. Spatial Hierarchy and OLAP-Favored Search in Spatial Data Warehouse. *Proceedings of DOLAP*, 2003.

[RBM01] Rivest, S., Bédard, Y., and Marchand, P. Towards Better Support for Spatial Decision Making: Defining the Characteristics of Spatial On-Line Analytical processing (SOLAP). *Geoinformatica* 55(4):539-555, 2001.

[RBP+05] Rivest, S., Bédard, Y., Proulx, M., Nadeau, M., Hubert, F., and Pastor, J. SOLAP: Merging Business Intelligence with Geospatial Technology for Interactive Spatio-Temporal Exploration and Analysis of Data. *Journal of International Society for Photogrammetry & Remote Sensing* 60(1):17-33, 2005.

[Riz03] Rizzi, S. Open problems in data warehousing: eight years later. *Proceedings of Design and Management of Data Warehouses*, 2003.

[RG00] Rizzi, S.and Golfarelli, M. Data warehouse design. *Proceedings of Enterprise Information Systems*, 2000.

[SC78] Sakoe, H., and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* 26(1):43–49, 1978.

[SLC+01] Shekhar, S., Lu, C-T., Chawla, S. and Zhang, P. Data Mining and Visualization of Twin-cities Traffic Data, *Technical Report* (TR 01-015), 2001.

[SPD+08] Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., and Vangenot, C. A conceptual view on trajectories. *DKE* 65(1):126-146, 2008.

[SNT+06] Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y. and Schult, R. Monic: modeling and monitoring cluster transitions. *Proceedings of ACM SIGKDD*, 2006.

[SHK00] Stefanovic, N., Han, J., and Koperski, K. Object-based selective materialization for efficient implementation of spatial data cubes. *TKDE* 12(6):938-958, 2000.

[TKC+04] Tao, Y., Kollios, G., Considine, J., Li, F., and Papadias, D. Spatio-Temporal Aggregation Using Sketches. *Proceedings of ICDE*, 2004.

[TP05] Tao, T., and Papadias, D. Historical Spatio-Temporal Aggregation. *Proceedings of ACM TODS* 23(1):61-102, 2005.

[The03] Theodoridis, Y. Seismo-Surfer: A Prototype for Collecting, Querying and Mining Seismic Data. *Proceedings of PCI*, 2003.

[TPG+01] Trujillo, J., Palomar, M., Gomez, J. and Song, I. Designing Data Warehouses with OO Conceptual Models. *IEEE Computer* 34(12):66-75, 2001.

[VS99] Vassiliadis, P. and Sellis, T. A survey of logical models for OLAP databases. *SIGMOD Record* 28(4):64-69, 1999.

[VWI98] Vitter, J.S., Wang, M., and Iyer, B. Data Cube Approximation and Histograms via Wavelets. *Proceedings of CIKM*, 1998.

[Yu05] Yu, B., Mining Earth Science Data for Geophysical Structure: A Case Study in Cloud Detection. *Proceedings of SIAM*, 2005.

[ZT05] Zhang D. Tsotras, V. Optimizing spatial Min/Max aggregations. *The VLDB Journal* 14(2):170-181, 2005.