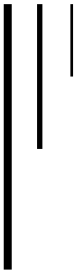


Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Σύγκριση Τεχνικών Προστασίας Ιδιωτικότητας για Χωροχρονικά Δεδομένα** **Comparison of Privacy Preservation Techniques for Spatio-temporal Data**
Ονοματεπώνυμο Φοιτητή	**Χαρίλαος Κεραμυδάς**
Πατρώνυμο	**Ελευθέριος**
Αριθμός Μητρώου	**ΜΠΣΠ/12030**
Επιβλέπων	**Πελέκης Νίκος, Επίκουρος Καθηγητής**

Ημερομηνία Παράδοσης **Δεκέμβριος 2015**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Νίκος Πελέκης Επ. Καθηγητής	Ιωάννης Θεοδωρίδης Καθηγητής	Ιωάννης Σίσκος Καθηγητής

Περίληψη

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος με τίτλο «Προηγμένα Συστήματα Πληροφορικής» που έλαβε χώρα στο Πανεπιστήμιο του Πειραιά για το ακαδημαϊκό έτος 2012-2013.

Η εργασία αυτή συγκρίνει και αξιολογεί μεθόδους ανωνυμοποίησης που έχουν προταθεί στην βιβλιογραφία για τροχιές κινούμενων αντικειμένων. Η σύγκριση που πραγματοποιείται γίνεται ανάμεσα στις υπάρχουσες μεθόδους ανωνυμοποίησης οι οποίες εφαρμόζονται σε διαφορετικά σύνολα δεδομένων για όλες τις παραλλαγές τους και σε διαφορετικά επίπεδα ανωνυμοποίησης. Από την ανωνυμοποίηση των δεδομένων και την εφαρμογή των κριτηρίων αξιολόγησης προκύπτουν αποτελέσματα ικανά να οδηγήσουν στην εξαγωγή κρίσιμων συμπερασμάτων. Τα οποία μας δεικνύουν τα πλεονεκτήματα και τα μειονεκτήματα των μεθόδων ανωνυμοποίησης και βοηθούν στην καλύτερη δυνατή αξιοποίηση τους ανάλογα με τα χαρακτηριστικά του συνόλου δεδομένων που πρόκειται να επεξεργαστούν.

Αρχικά παρουσιάζεται το πρόβλημα της προστασίας της ιδιωτικότητας και οι απειλές που μπορούν να προκύψουν όταν γίνεται κακόβουλη χρήση των προσωπικών δεδομένων. Παρουσιάζονται οι τεχνικές που έχουν προταθεί στην βιβλιογραφία για σχεσιακά δεδομένα αλλά κυρίως για δεδομένα κινούμενων αντικειμένων των οποίων η διαφύλαξη είναι ο στόχος της παρούσας εργασίας. Εν συνεχεία παρουσιάζεται βήμα προς βήμα όλη η διαδικασία από την οποία προκύπτουν οι αξιολογήσεις των μεθόδων. Τέλος γίνεται η ανάλυση των αποτελεσμάτων και παρουσιάζονται τα συμπεράσματα που προέκυψαν.

Abstract

The current thesis compares and evaluates well-known anonymization methods that have been proposed in the literature for moving object trajectories. The comparison is conducted between the existing anonymization methods that are applied to different data sets for all their variations and different levels of anonymity. Through data anonymization and evaluation criteria application, various some results are derived that are capable to lead to the extraction of critical conclusions which give the advantages and disadvantages of anonymization methods to help defining the optimal use of them depending on the characteristics of the datasets that are going to be processed.

Initially, the problem of privacy protection and the potential threats that may arise due to malicious use of personal data are presented. Well-known techniques that have been proposed in the literature for relational but mainly for moving object data whose preservation is the goal of this work are introduced. Then, step by step the whole process that generates the assessment of the methods is presented. Finally the analysis of the results and the conclusions of the thesis are presented.

Περιεχόμενα

Περίληψη 3
Abstract 4
Περιεχόμενα 5
Κατάλογος Εικόνων 7
Κατάλογος Πινάκων 8
Κατάλογος Ερωτημάτων 9
Κατάλογος Γραφημάτων 10
1. Εισαγωγή 11
2. Σχετικές Εργασίες 14
    2.1. *k*-anonymity 14
    2.2. *l*-diversity 14
    2.3. *t*-closeness 15
    2.4. Προστασία χωρικής ιδιωτικότητας μέσω σύγχυσης τροχιών (Protecting Location Privacy Through Path Confusion) 16
    2.5. Διαφύλαξη της ιδιωτικότητας κατά τη δημοσίευση των τροχιών (Privacy Preservation in the Publication of Trajectories) 17
    2.6. Η Αβεβαιότητα για την Ανωνυμία των Βάσεων Δεδομένων Κινούμενων Αντικειμένων (Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases) 18
    2.7. Ανωνυμοποίηση βάσεων δεδομένων κινούμενων αντικειμένων μέσω ομαδοποίησης και σύγχυσης (Anonymization of moving objects databases by clustering and perturbation) 19
    2.8. Ανωνυμοποίηση Τροχιών Κινούμενων Αντικειμένων: μια προσέγγιση βασισμένη στην γενίκευση (Towards Trajectory Anonymization: a Generalization-Based Approach) 20
    2.9. Ανωνυμία Δεδομένων Κίνησης μέσω Γενίκευσης - Movement Data Anonymity through Generalization (Voronoys) 21
    2.10. Μια προσέγγιση βασιζόμενη στη συσταδοποίηση για εξατομικευμένη προστασία ιδιωτικότητας κατά την δημοσίευση των δεδομένων τροχιών κινούμενων αντικειμένων (A Clustering-Based Approach for Personalized Privacy Preserving Publication of Moving Object Trajectory Data) 23
3. Αντικείμενο εργασίας και κριτήρια αποτίμησης 25
4. Περιγραφή υλοποίησης της εργασίας σύγκρισης τεχνικών προστασίας ιδιωτικότητας 27
    4.1. Βήμα 1ο: Προεπεξεργασία δεδομένων 27
    4.2. Βήμα 2ο: Ανωνυμοποίηση δεδομένων 31
    4.3. Βήμα 3ο: Αξιολόγηση αποτελεσμάτων 35
5. Πειραματική Μελέτη 40
    5.1. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν 40
    5.2. Παράμετροι πειράματος 41

5.3. Παρουσίαση αποτελεσμάτων ανά κριτήριο 41
5.3.1. Distortion 42
5.3.2. F-measure 46
5.3.3. Runtime 48
5.3.4. Removed Points 50
6. Συμπεράσματα 52
7. Βιβλιογραφικές αναφορές 53

Κατάλογος Εικόνων

Εικόνα 1: Βασικά τμήματα μιας location-based υπηρεσίας και ροή πληροφορίας [12] 11

Εικόνα 2: Σύγχυση τροχιών 16

Εικόνα 3: Βάση τροχιών....17

Εικόνα 4: Δυο τροχιές με τους τόμους αβεβαιότητάς τους ακτίνας *δ*, και ένας κεντρικός κυλινδρικός τόμος ακτίνας *δ*/2 ο οποίος τις εμπεριέχει....18

Εικόνα 5: Χωρική τροποποίηση τριών διαφορετικών τροχιών 18

Εικόνα 6: Παράδειγμα EDR αντιστοίχισης και χωροχρονικής επεξεργασίας....19

Εικόνα 7: Διαδικασία Ανωνυμοποίησης AWO 20

Εικόνα 8: A) Τροχιές B) Χαρακτηριστικά σημεία μιας τροχιάς C) Όλα τα χαρακτηριστικά σημεία των τροχιών D) Ψηφιδοποίηση περιοχής E) Γενικευμένες τροχιές F) Συνοπτική αναπαράσταση των τροχιών 21

Εικόνα 9: Παράδειγμα ανωνυμοποίησης με τους αλγορίθμους της μεθόδου Generalization 21

Εικόνα 11: Προθεματικό δέντρο με εφαρμογή του KAM_CUT 22

Εικόνα 10: Προθεματικό δέντρο 22

Εικόνα 12: Προθεματικό δέντρο με εφαρμογή του KAM_REC 22

Εικόνα 13: Συσταδοποίηση των τροχιών με την χρήση της πρώτης στρατηγικής 23

Εικόνα 14: Συσταδοποίηση των τροχιών με την χρήση της δεύτερης στρατηγικής 24

Εικόνα 15: Παράμετροι της NWA 31

Εικόνα 16: Παράμετροι της W4M EDR 32

Εικόνα 18: Παράμετροι της AWO 33

Εικόνα 19: Παράμετροι για την γενίκευση των τροχιών 33

Εικόνα 20: Παράμετροι της Generalization 34

Κατάλογος Πινάκων

Πίνακας 1: Παράδειγμα 4-ανωνυμίας....14

Πίνακας 2: Παράδειγμα 3-diversity....15

Πίνακας 3: Χαρακτηριστικά των συνόλων δεδομένων....40

Κατάλογος Ερωτημάτων

Ερώτημα 1: Εύρεση των σημείων που πρέπει να διασπαστούν οι τροχιές 27

Ερώτημα 2: Διάσπαση των τροχιών στο πρώτο σημείο που παρουσιάζεται μεγάλο χρονικό κενό 28

Ερώτημα 3: Διαγραφή των τροχιών με μικρή χρονική διάρκεια 29

Ερώτημα 4: Διαγραφή των τροχιών που διένυσαν μικρή απόσταση 29

Ερώτημα 5: Βασικά χαρακτηριστικά του συνόλου δεδομένων 30

Ερώτημα 7: Δείκτης σφάλματος DAI 36

Ερώτημα 6: Δημιουργία τυχαίων χωροχρονικών περιοχών 36

Ερώτημα 8: Δείκτης σφάλματος PSI 37

Ερώτημα 9: Πλήθος διαγεγραμμένων σημείων 37

Ερώτημα 10: Υπολογισμός F-Measure 39

Κατάλογος Γραφημάτων

Γράφημα 1: DAI error ratio για το σύνολο δεδομένων Trucks....42

Γράφημα 2: DAI error ratio για το σύνολο δεδομένων Bus....43

Γράφημα 3: DAI error ratio για το σύνολο δεδομένων Milan....43

Γράφημα 4: PSI error ratio για το σύνολο δεδομένων Trucks....44

Γράφημα 5: PSI error ratio για το σύνολο δεδομένων Bus....45

Γράφημα 6: PSI error ratio για το σύνολο δεδομένων Milan....45

Γράφημα 7: F-Measure για το σύνολο δεδομένων Trucks....46

Γράφημα 8: F-Measure για το σύνολο δεδομένων Bus....47

Γράφημα 9: F-Measure για το σύνολο δεδομένων Milan....47

Γράφημα 10: Runtime για το σύνολο δεδομένων Trucks....48

Γράφημα 11: Runtime για το σύνολο δεδομένων Bus....49

Γράφημα 12: Runtime για το σύνολο δεδομένων Milan....49

Γράφημα 13: Deleted Points για το σύνολο δεδομένων Trucks....50

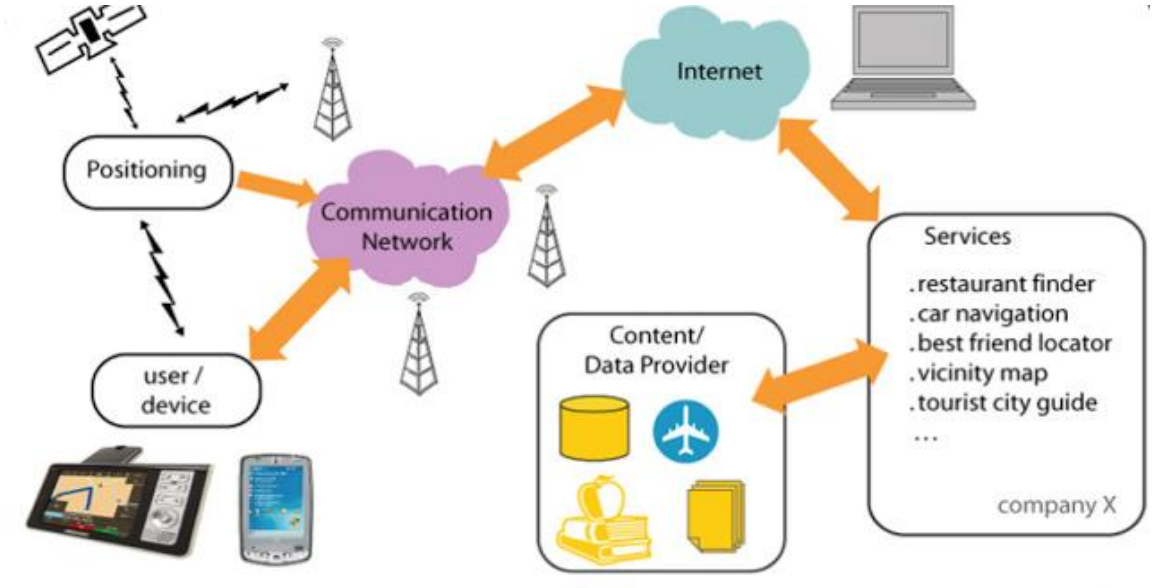
Γράφημα 14: Deleted Points για το σύνολο δεδομένων Bus....51

Γράφημα 15: Deleted Points για το σύνολο δεδομένων Milan....51

1. Εισαγωγή

Από τις αρχές του 21ου αιώνα είναι ευρέως διαδεδομένη η χρήση συσκευών εντοπισμού θέσης. Από τις πιο διαδεδομένες συσκευές εντοπισμού θέσης είναι τα κινητά τηλέφωνα, οι συσκευές GPS και τα συστήματα ταυτοποίησης μέσω ραδιοσυχνοτήτων (RFID) τα οποία χρησιμοποιούνται από υπηρεσίες εντοπισμού θέσης. Οι υπηρεσίες που βασίζονται στη θέση (*Location-Based Services – LBS*) είναι υπηρεσίες που αναπτύσσονται και διανέμονται από ασύρματους φορείς και παρέχουν πληροφορίες που σχετίζονται ή εξαρτώνται από την θέση της συσκευής. Πλέον οι χρήστες που χρησιμοποιούν τέτοιες συσκευές δεν χρειάζεται να παρέχουν οι ίδιοι πληροφορίες σχετικά με την θέση τους (π.χ. τον ταχυδρομικό τους κώδικα) ώστε να χρησιμοποιήσουν location-based services. Οποιαδήποτε υπηρεσία λοιπόν χρησιμοποιεί την θέση του χρήστη μπορεί να χαρακτηριστεί ως location-based service. Οι υπηρεσίες αυτές καταγράφουν τις θέσεις των αντικειμένων ανά χρονικά διαστήματα, η ένωση των οποίων αποτελεί την συνολική τροχιά του αντικειμένου. Άρα οι τροχιές είναι χωροχρονικά ίχνη που μπορεί να συλλέγονται για μεγάλα χρονικά διαστήματα. Τα αντικείμενα που καταγράφονται μπορεί να αφορούν οχήματα, ανθρώπους ακόμα και ζώα. Για παράδειγμα από την καταγραφή των τροχιών των αυτοκινήτων μιας πόλης μπορούμε μετά από εξόρυξη γνώσης στα δεδομένα (*data mining*) να εξάγουμε πολλά ενδιαφέροντα συμπεράσματα. Τέτοιου είδους συμπεράσματα μπορεί να είναι σχετικά με το κυκλοφοριακό σχεδιασμό της πόλης, την συμπεριφορά των οδηγών σε ώρες αιχμής κ.α. Όμως από αυτή την καταγραφή δεδομένων προκύπτουν και προβλήματα ιδιωτικότητας λόγω της ακρίβειας του εντοπισμού της θέσης με απόκλιση μερικών μέτρων που προσφέρουν σήμερα οι συσκευές GPS. Έτσι όταν αυτά δημοσιοποιούνται θα μπορεί κάποιος κακόβουλος χρήστης με την βοήθεια εξωτερικής γνώσης να καταλήξει σε ταυτοποίηση, να βγάλει συμπεράσματα για τον τρόπο ζωής και να αποκαλύψει ευαίσθητες προσωπικές πληροφορίες. Έτσι τα δεδομένα τροχιών δεν μπορούν να δημοσιεύονται εκτός και αν είναι σωστά ανωνυμοποιημένα και για το λόγω αυτό έχουν προταθεί διάφορες τεχνικές προστασίας της ιδιωτικότητας, οι επικρατέστερες των οποίων είναι οι NWA[1], W4M[2], AWO[3] και Generalazation (Voronoys) [4] οι οποίες θα περιγραφούν αναλυτικά στην ενότητα 2.

Ένας εύκολος τρόπος που μπορούμε να εφαρμόσουμε για να προστατέψουμε την ιδιωτικότητα σε ένα σύνολο τροχιών είναι να αφαιρέσουμε βασικές πληροφορίες όπως είναι ο αριθμός αναγνώρισης του αντικειμένου της τροχιάς. Όμως με αυτό τον τρόπο δεν εξασφαλίζεται η προστασία της ιδιωτικότητας γιατί



Εικόνα 1: Βασικά τμήματα μιας location-based υπηρεσίας και ροή πληροφορίας [12]

με έναν συνδυασμό από εξωτερικές πληροφορίες που υπάρχουν δημοσιευμένες μπορούν να μας οδηγήσουν στην αποκάλυψη της ταυτότητας του αντικειμένου. Αυτός ο προβληματισμός παρουσιάζεται στην μέθοδο της *k*-ανωνυμίας [5] για σχεσιακούς πίνακες δεδομένων στην οποία παρουσιάζεται μια μελέτη όπου το 87% του πληθυσμού των Ηνωμένων Πολιτειών μπορεί να ταυτοποιηθεί χρησιμοποιώντας φαινομενικά ακίνδυνα χαρακτηριστικά όπως φύλο, ημερομηνία γέννησης και πέντε ψηφία από τον ταχυδρομικό κώδικα.

Οι χωροχρονικές τροχιές για κάθε θέση μπορούν να περιέχουν πληροφορίες όπως γεωγραφικό ύψος και πλάτος, χρονική στιγμή, υψόμετρο, ταχύτητα κ.α. Τέτοια χαρακτηριστικά θεωρούνται ισχυρά αναγνωριστικά τα οποία μπορούν να συνδεθούν με κάποια εξωτερική γνώση και να οδηγήσουν σε ταυτοποίηση κάποιου ατόμου. Έτσι θα μπορούσε κάποιος να αντιστοιχίσει την αρχή και το τέλος μιας τροχιάς που κάνει ένα όχημα κάθε μέρα με το σπίτι του ιδιοκτήτη και τον χώρο εργασίας του. Αυτό θα έχει ως αποτέλεσμα την ταυτοποίηση του ιδιοκτήτη κάνοντας χρήση μόνο ενός τηλεφωνικού καταλόγου σε συνδυασμό με την γεωγραφική του θέση.

Για την αποτροπή τέτοιου είδους εξωτερικών επιθέσεων η λύση είναι η σωστή ανωνυμοποίηση. Παρόλα αυτά η χρήση της *k*-ανωνυμίας είναι η πιο διαδεδομένη μεταξύ των μεθόδων ανωνυμοποίησης για δεδομένα κίνησης και χρησιμοποιείται ως βασική ιδέα παρά την φυσική πολυπλοκότητα των χωροχρονικών τροχιών και την εξάρτηση των διαδοχικών σημείων σε μία τροχιά που κάνει το πρόβλημα της ανωνυμίας δυσκολότερο.

Σε αυτή την εργασία ο σκοπός είναι η ανάλυση των αποτελεσμάτων από τη σύγκριση των τεχνικών ανωνυμοποίησης για δεδομένα κίνησης. Η σύγκριση αυτή των μεθόδων γίνεται με χρήση κοινών κριτηρίων αξιολόγησης, για τα ίδια σύνολα τροχιών και εφαρμόζοντας την ανωνυμοποίηση στο ίδιο λειτουργικό περιβάλλον. Θέτονται λοιπόν θέματα όπως η ταχύτητα εκτέλεσης των μεθόδων ανωνυμοποίησης καθώς και η παραμόρφωση των δεδομένων μετά την ανωνυμοποίηση ως προς το επίπεδο *k*-ανωνυμίας πάνω σε διαφορετικού πλήθους και διαφορετικού μήκους σύνολα χωροχρονικών δεδομένων. Έτσι θα μπορέσουμε να εξάγουμε χρήσιμα συμπεράσματα για την ορθή χρήση των μεθόδων ανωνυμοποίησης, ανάλογα με το μέγεθος του συνόλου των δεδομένων και την εφαρμογή από την οποία μπορούν να χρησιμοποιηθούν. Οι μέθοδοι ανωνυμοποίησης πάνω στις οποίες θα πραγματοποιηθούν οι συγκρίσεις είναι οι NWA[1], W4M[2], AWO[3] και Generalazation (Voronoys) [4].

Στο δεύτερο μέρος παρουσιάζονται συνοπτικά σχετικές εργασίες για μεθόδους ανωνυμοποίησης. Αρχικά παρουσιάζονται οι μέθοδοι ανωνυμοποίησης που εφαρμόζονται σε σχεσιακούς πίνακες και εν συνεχεία παρουσιάζονται οι μέθοδοι ανωνυμοποίησης για δεδομένα κίνησης. Οι μέθοδοι πάνω στις οποίες θα πραγματοποιηθούν οι συγκρίσεις παρουσιάζονται αναλυτικότερα.

Στο τρίτο μέρος παρουσιάζεται αναλυτικά το αντικείμενο με το οποίο πρόκειται να ασχοληθεί η εργασία. Περιγράφονται αναλυτικά τα τέσσερα κριτήρια πάνω στα οποία πρόκειται να γίνει η σύγκριση εξηγώντας την επιλογή τους, σε τι μας χρησιμεύουν και για ποιες περιπτώσεις αυτά θα χρησιμοποιηθούν.

Στο τέταρτο μέρος περιγράφεται με τεχνικό τρόπο η διαδικασία που ακολουθήθηκε για να καταλήξουμε στα μετρήσιμα αποτελέσματα. Παρουσιάζεται δηλαδή βήμα προς βήμα όλη η διαδικασία από την φάση της προεπεξεργασίας, την φάση της ανωνυμοποίησης έως την τελική φάση από την οποία θα βγουν τα τελικά αποτελέσματα.

Τέλος, στο πέμπτο μέρος παρουσιάζονται λεπτομερώς τα σύνολα δεδομένων που χρησιμοποιήθηκαν για τις μεθόδους ανωνυμοποίησης με όλα τα χαρακτηριστικά τους και πως αυτά μπορεί να επηρέασαν το

αποτέλεσμα των μεθόδων ανωνυμοποίησης. Επίσης αναλύονται οι παράμετροι που τέθηκαν σε κάθε μέθοδο για να πραγματοποιήσουν τις ανωνυμοποιήσεις και τέλος παρουσιάζονται τα αποτελέσματα ανά κριτήριο σε μορφή διαγράμματος για το κάθε σύνολο δεδομένων.

2. Σχετικές Εργασίες

2.1. *k*-anonymity

Η μέθοδος της *k*-ανωνυμίας [5] είναι μια μέθοδος για σχεσιακούς πίνακες δεδομένων στην οποία χωρίζονται τα πεδία του πινάκα σε *quasi-identifiers* (αναγνωριστικά πεδία) και σε *sensitive-attributes* (ευαίσθητα χαρακτηριστικά) και χρησιμοποίει τεχνικές γενίκευσης (generalization) και καταστολής (suppression) πληροφοριών στα αναγνωριστικά πεδία ούτως ώστε μια εγγραφή να έχει άλλες *k*-1 έγγραφες με κοινά χαρακτηριστικά πεδία. Ως γενίκευση (generalization) ορίζεται η διαδικασία κατά την οποία η αρχική τιμή που εμφανίζεται στα δεδομένα αντικαθίστανται με μία πιο γενική τιμή και καταστολή (suppression) ορίζεται η διαδικασία κατά την οποία μια μεμονωμένη τιμή αποσιωπάται με αστερίσκο. Με την χρήση αυτών των τεχνικών χάνεται μέρος της χρήσιμης πληροφορίας που εμφανίζεται στα αρχικά δεδομένα με τέτοιο τρόπο ώστε για κάθε συνδυασμό που δημιουργείται να υπάρχουν τουλάχιστον *k* εγγραφές που μοιράζονται τα αποτελέσματα αυτά όπως φαίνεται στον πίνακα 1β που έχουμε επιτύχει 4-ανωνυμία.. Για αυτό το λόγο απαιτείται να αφαιρεθεί όσο το δυνατό λιγότερη πληροφορία από τα αρχικά δεδομένα διατηρώντας την χρηστικότητα τους για εκείνους που θέλουν να τα αξιοποιήσουν.

Πίνακας 1: Παράδειγμα 4-ανωνυμίας

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

α) Πραγματικά δεδομένα

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

β) Ανωνυμοποιημένα δεδομένα 4-anonymity

2.2. *l*-diversity

Η *l*-diversity [6] είναι μια μέθοδος προστασίας ιδιωτικότητας που λειτουργεί ουσιαστικά ως συνέχεια της μεθόδου *k*-anonymity και εφαρμόζεται σε σχεσιακά δεδομένα. Αυτό που περιγράφει η μέθοδος αυτή είναι ότι θα πρέπει να υπάρχει ποικιλότητα των ευαίσθητων δεδομένων μεγαλύτερη ή ίση του *l* για κάθε υποομάδα που δημιουργείται από την *k*-ανωνυμία έτσι ώστε να αποφευχθούν εξωτερικές επιθέσεις. Για παράδειγμα αν σε μια υποομάδα που έχει δημιουργηθεί και τηρεί την k-ανωνυμία, γνωρίζουμε ότι ο Α ανήκει σίγουρα σε αυτή και τυγχάνει να έχουν όλοι την ίδια ασθένεια, τότε μπορούμε να συμπεράνουμε ότι πάσχει από την συγκεκριμένη ασθένεια. Το σενάριο αυτό φαίνεται στο παράδειγμα του πίνακα 1β για αυτούς που ανήκουν στην τρίτη ομάδα που δημιουργείται από την ανωνυμοποίηση των εγγραφών. Αυτού του τύπου η επίθεση ονομάζεται *ομοιογενής* (*homogeneity attack*). Άλλου τύπου επίθεση είναι αυτή του *γνωστικού υπόβαθρου* (*background knowledge attack*). Σε αυτή την περίπτωση αν κάποιος γνωρίζει πληροφορίες που παρουσιάζουν κάποια ιδιαιτερότητα για αυτόν που ψάχνει τότε μπορεί να περιορίσει τις επιλογές του και να καταλήξει στην πιο πιθανή ευαίσθητη πληροφορία. Τέτοια περίπτωση θα μπορούσε

να είναι για κάποιον που ανήκει στην πρώτη ανωνυμοποιημένη ομάδα του πίνακα 1 για τον οποίο γνωρίζουμε ότι έχει πολύ χαμηλές πιθανότητες για κάποια καρδιακή πάθηση όποτε καταλήγουμε να γνωρίζουμε ότι υπέστη κάποια λοίμωξη. Έτσι με την εφαρμογή της *l*-diversity η πιθανότητα του επιτιθέμενου στο να ανακαλύψει τα σωστά ευαίσθητα δεδομένα μειώνεται στο 1// όπως γίνεται στις ανωνυμοποιημένες ομάδες του πίνακα 2 στον οποίο εφαρμόστηκε η μέθοδος *l*-diversity.

2.3. *t*-closeness

Η μέθοδος *t*-closeness [7] έρχεται και αυτή με την σειρά της να καλύψει τα κενά που αφήνει η μέθοδος *l*-diversity από *ασύμμετρες επιθέσεις* (*skewness attack*) και *επιθέσεις ομοιότητας* (*similarity attack*). Για παράδειγμα, ασύμμετρη επίθεση μπορούμε να έχουμε σε ένα ευαίσθητο χαρακτηριστικό με αληθές ή ψευδές αποτέλεσμα μιας ανωνυμοποιημένης βάσης που τηρεί την k-ανωνυμία και 2-diversity. Όταν εμφανίζεται το αρνητικό σε κάποια υποομάδα σε ποσοστό 50% ενώ σε ολόκληρη την βάση το χαρακτηριστικό αυτό είναι μόλις 1% τότε αυξάνονται δραματικά οι πιθανότητες πρόβλεψης αυτού του ευαίσθητου χαρακτηριστικού από το 1% στο 50% για αυτή την υποομάδα. Το παράδειγμα της επίθεσης ομοιότητας φαίνεται στον πίνακα 2 όπου μπορούμε να συμπεράνουμε για κάποιον που ανήκει στην πρώτη ομάδα ότι είναι χαμηλόμισθος και ότι πάσχει από μια ασθένεια του στομάχου. Αυτό συμβαίνει γιατί η *l*-diversity αφενός μπορεί να εξασφαλίζει την διαφορετικότητα των τιμών σε κάθε υποομάδα αφετέρου δεν λαμβάνει όμως καθόλου υπόψη τη σημασιολογική εγγύτητα των τιμών αυτών.

Πίνακας 2: Παράδειγμα 3-diversity

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

α)Πραγματικά δεδομένα

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

β)Ανωνυμοποιημένα δεδομένα 3-diversity

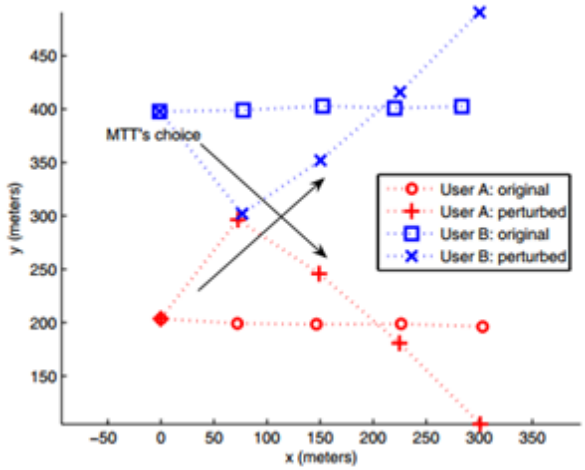
Η μέθοδος της *t*-closeness εξασφαλίζει ότι, η κατανομή ενός ευαίσθητου πεδίου σε κάθε κλάση ισοδυναμίας διαφέρει από την κατανομή του συγκεκριμένου πεδίου σε όλη την βάση δεδομένων το πολύ κατά ένα κατώφλι *t*. Όσο πιο μικρή είναι η τιμή του *t*, τόσο πιο κοντά βρίσκονται οι δύο κατανομές. Έτσι η μέθοδος *t*-closeness προστατεύει από επιθέσεις που σκοπό έχουν την αποκάλυψη ευαίσθητων γνωρισμάτων στο σύνολο εγγραφών σε συνδυασμό με τις μεθόδους *k*-anonymity και *l*-diversity.

Όμως ο τρόπος χρήσης της μεθόδου αυτής για δεδομένα κίνησης είναι διαφορετικός λόγω του ότι, στους πίνακες με δεδομένα δεν γνωρίζουμε ποια σημεία της τροχιάς των χρηστών περιέχουν ευαίσθητη πληροφορία και ποια όχι. Δεν μπορούν να διαχωρίσουν δηλαδή τα αναγνωριστικά πεδία από αυτά που περιέχουν ευαίσθητη προσωπική πληροφορία. Άρα κάθε σημείο της τροχιά μπορεί να είναι ευαίσθητο. Στην συνέχεια παρουσιάζονται τεχνικές προστασίας ιδιωτικότητας σε τροχιές κινούμενων αντικειμένων.

2.4. Προστασία χωρικής ιδιωτικότητας μέσω σύγχυσης τροχιών (Protecting Location Privacy Through Path Confusion)

Η κύρια ιδέα του αλγορίθμου προστασίας ιδιωτικότητας [8] βασίζεται στην έννοια σύγχυσης των μονοπατιών. Για κάθε δυο χρήστες των οποίων οι τροχιές διασταυρώνονται ή βρίσκονται πολύ κοντά δημιουργείται μια πιθανότητα για τον επιτιθέμενο να χάσει την σωστή τροχιά και να ακολουθήσει τον λάθος χρηστή. Έτσι ο αλγόριθμος εκμεταλλεύεται τέτοιου είδους περιοχές "συνάντησης" για να αυξήσει τις πιθανότητες σύγχυσης του επιτιθέμενου.

Ο αλγόριθμος αυτός επιστρέφει συγκεχυμένες τις αρχικές τροχιές ενός ζεύγους χρηστών. Με τον τρόπο αυτό αυξάνεται το επίπεδο ιδιωτικότητας σε κάθε βήμα στο οποίο διαμορφώνεται ένα ζεύγος δειγμάτων τροχιών που βρίσκονται εντός μιας μέγιστης ακτίνας διατάραξης *R*. Όσο μεγαλύτερη είναι η ακτίνα *R*, τόσο μεγαλύτερο είναι το επίπεδο της ιδιωτικότητας. Και όσο μικρότερη είναι η ακτίνα *R*, τόσο μεγαλύτερη είναι η ποιότητα των δεδομένων και μικρότερη η προστασία ιδιωτικότητας.



Εικόνα 2: Σύγχυση τροχιών

Ένα απλό σενάριο εφαρμογής του αλγορίθμου σύγχυσης είναι αυτό όπου δυο χρήστες ταξιδεύουν σχεδόν παράλληλα. Δημιουργώντας δυο τροχιές οι οποίες χωρίζονται σε ένα πεπερασμένο αριθμό δειγμάτων όπως παρουσιάζεται στην εικόνα 2. Οι κόκκινοι κύκλοι είναι περιοδικά σημεία πάνω στη τροχιά του χρηστή Α και τα μπλε τετράγωνα είναι τα αντίστοιχα περιοδικά σημεία πάνω στη τροχιά του χρήστη Β. Η κίνηση και των δυο είναι από τα αριστερά στα δεξιά με μια μέση ταχύτητα 15m/s, η απόσταση αναμεταξύ τους είναι περίπου 200 μέτρα και η μέγιστη ακτίνα σύγχυσης είναι 100 μέτρα. Αρχικά ο αλγόριθμος ξεκινά με μια υπόθεση ίση με 1 για το αρχικό βήμα και στην συνέχεια δημιουργεί δυο υποθέσεις για κάθε υπόθεση γονέα. Έτσι και για τους δυο χρήστες ο αλγόριθμος πρέπει να έχει δημιουργήσει 2*k*-1 υποθέσεις για το αντίστοιχο βήμα *k*. Στην επόμενη φάση του ο αλγόριθμος προσπαθεί να μεγιστοποιήσει την πιθανότητά του αναμενόμενου λάθους και οδηγείται στο να μετατρέψει τις σχεδόν παράλληλες τροχιές να μοιάζουν με διασταύρωση.

Η απόδοση του αλγορίθμου μειώνεται σημαντικά όταν εφαρμόζεται για πολλούς χρήστες με μεγάλες τροχιές εξαιτίας του εξαιρετικά μεγάλου αριθμού υποθέσεων που δημιουργούνται. Για το λόγο αυτό χρειάζεται να γίνει προεπεξεργασία των τροχιών και να βρεθούν εκείνα τα τμήματα ανά ζευγάρι τροχιών που μπορεί να γίνει εφαρμογή του αλγορίθμου. Τα τμήματα των τροχιών που επιλέγονται θα πρέπει ανά ζεύγος όλα τα αντίστοιχα δείγματα τους να έχουν μέγιστη απόσταση κάτω από ένα κατώφλι *D*.

2.5. Διαφύλαξη της ιδιωτικότητας κατά τη δημοσίευση των τροχιών (Privacy Preservation in the Publication of Trajectories)

Σε αυτή τη μέθοδο προστασίας ιδιωτικότητας [9], όπου οι τροχιές είναι μια αλληλουχία σημείων μετατρέπεται μια βάση τροχιών T (εικόνα 3α) σε μια νέα βάση T' (εικόνα 3γ), τέτοια ώστε κάθε πιθανός επιτιθέμενος με γνώση T_A (εικόνα 3γ) να μην μπορεί να συμπεράνει οποιαδήποτε θέση αρχικής τροχιάς με πιθανότητα μεγαλύτερη από ένα κατώφλι.

id	trajectory
t_1	$a_1 \rightarrow b_1 \rightarrow a_2$
t_2	$a_1 \rightarrow b_1 \rightarrow a_2 \rightarrow b_3$
t_3	$a_1 \rightarrow b_2 \rightarrow a_2$
t_4	$a_1 \rightarrow a_2 \rightarrow b_2$
t_5	$a_1 \rightarrow a_3 \rightarrow b_1$
t_6	$a_3 \rightarrow b_1$
t_7	$a_3 \rightarrow b_2$
t_8	$a_3 \rightarrow b_2 \rightarrow b_3$

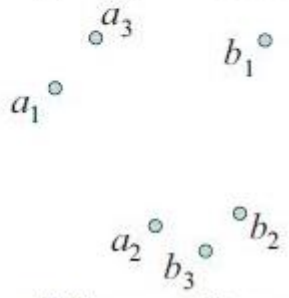
(α) Ακριβή δεδομένα (*T*)

id	trajectory
t_1^A	$a_1 \rightarrow a_2$
t_2^A	$a_1 \rightarrow a_2$
t_3^A	$a_1 \rightarrow a_2$
t_4^A	$a_1 \rightarrow a_2$
t_5^A	$a_1 \rightarrow a_3$
t_6^A	a_3
t_7^A	a_3
t_8^A	a_3

(β) Γνώση του A (T_A)

id	trajectory
t'_1	$a_1 \rightarrow b_1 \rightarrow a_2$
t'_2	$a_1 \rightarrow b_1 \rightarrow a_2$
t'_3	$a_1 \rightarrow b_2 \rightarrow a_2$
t'_4	$a_1 \rightarrow a_2 \rightarrow b_2$
t'_5	$a_3 \rightarrow b_1$
t'_6	$a_3 \rightarrow b_1$
t'_7	$a_3 \rightarrow b_2$
t'_8	$a_3 \rightarrow b_2$

(γ) Μετασχηματισμένη βάση (*T'*)



(δ) Χάρτης των θέσεων

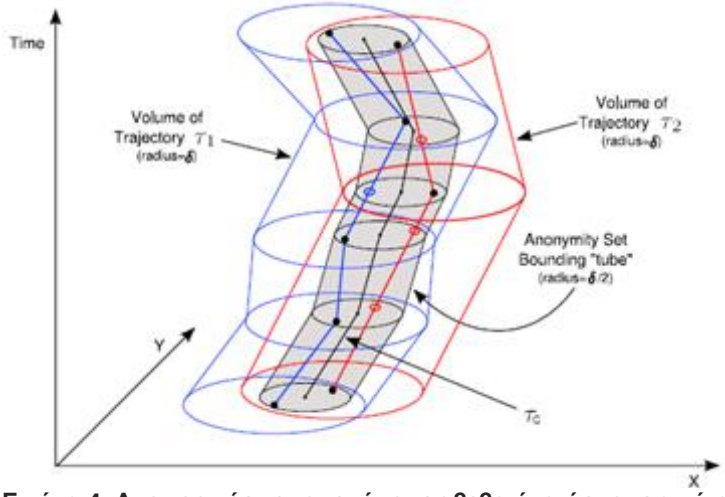
Εικόνα 3: Βάση τροχιών

Η παραπάνω μετατροπή πραγματοποιείται με την εφαρμογή ενός "άπληστου" αλγορίθμου [9] ο οποίος καταργεί σημεία των τροχιών έως ότου επιτευχθεί το επιθυμητό επίπεδο προστασίας. Σε πρώτη φάση εξάγονται οι προβολές για κάθε ένα από τους επιτιθεμένους και ύστερα ο αλγόριθμος αναγνωρίζει τις προβολές οι οποίες μπορούν να προκαλέσουν παραβίαση ιδιωτικότητας. Από την στιγμή που έχουν αναγνωριστεί όλες οι απειλές της ιδιωτικότητας ο αλγόριθμος μπαίνει σε μια επαναληπτική διαδικασία στην οποία ενοποιεί ζευγάρια προβολών του ίδιου επιτιθέμενου στα οποία τουλάχιστον το ένα είναι προβληματικό. Επίσης ενοποιήσεις που μπορούν να γίνουν είναι μόνο ανάμεσα σε προβολές όπου η μια είναι υποτροχιά της άλλης. Σε κάθε επανάληψη γίνεται μια ενοποίηση ζεύγους προβολών, αυτού με το μικρότερο κόστος χαμένης πληροφορίας ανάμεσα στις τροχιές πριν και μετά την ενοποίησή τους.

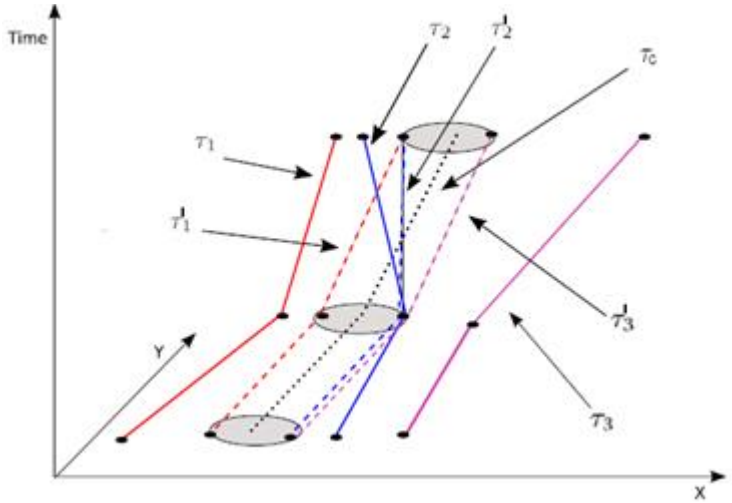
Η βελτίωση της παραπάνω διαδικασίας επιτυγχάνεται με την ενοποίηση πολλαπλών ζευγαριών ανά επανάληψη. Η παράλληλη ενοποίηση έχει νόημα να πραγματοποιηθεί από την στιγμή που μια ενοποίηση δεν επηρεάζει όλες τις προβολές και τα προβλήματα τα οποία καλούνται να επιλύσουν. Έτσι κάθε προβολή μπορεί να συμμετέχει σε περισσότερες από μια ενοποιήσεις.

2.6. Η Αβεβαιότητα για την Ανωνυμία των Βάσεων Δεδομένων Κινούμενων Αντικειμένων (Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases)

Ο αλγόριθμος NWA(Never Walk Alone) [1] αποτελεί μια μέθοδο προστασίας ιδιωτικότητας η οποία δημιουργεί συστάδες μεγέθους *k* έως 2*k*-1 τροχιών τηρώντας την (*k*,*δ*)-ανωνυμία. Η (*k*,*δ*) ανωνυμία είναι ουσιαστικά ένα σύνολο τουλάχιστον *k* τροχιών όπου η απόσταση αναμεταξύ τους σε κάθε χρονικό σημείο



Εικόνα 4: Δυο τροχιές με τους τόμους αβεβαιότητάς τους ακτίνας *δ*, και ένας κεντρικός κυλινδρικός τόμος ακτίνας *δ*/2 ο οποίος τις εμπεριέχει



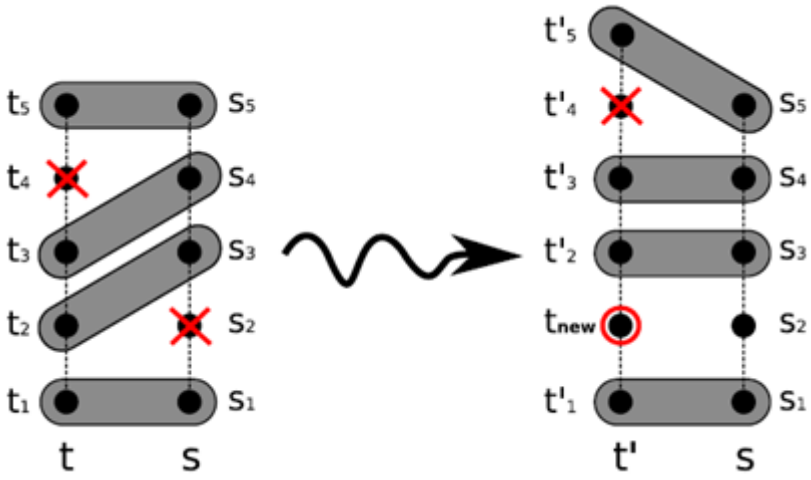
Εικόνα 5: Χωρική τροποποίηση τριών διαφορετικών τροχιών

είναι έως *δ* (Εικόνα 4). Το τελικό αποτέλεσμα του αλγορίθμου προκύπτει ύστερα από 3 στάδια: την προεπεξεργασία, την συσταδοποίηση και τέλος τη χωρική μετατόπιση κάποιων σημείων.

- Στην προεπεξεργασία χωρίζονται τα εισαγόμενα δεδομένα σε ισοδύναμες κλάσεις ως προς την αρχή και το τέλος της κάθε τροχιάς. Αυτό γίνεται γιατί η συγκεκριμένη μέθοδος χρησιμοποίει την συνάρτηση ευκλείδειας απόστασης για την μέτρηση αποστάσεων μεταξύ των τροχιών και το αρνητικό της συνάρτησης αυτής είναι ότι εφαρμόζεται σε ίδιου μήκους τροχιές.
- Στη φάση της συσταδοποίησης ο αλγόριθμος επιλεγεί μια σειρά από τροχιές τις οποίες θεώρει κέντρο της κάθε συστάδας. Για κάθε κέντρο συστάδας αντιστοιχεί σε αυτή τις *k*-1 κοντινότερες τροχιές. Όσες τροχιές περισσέψουν αντιστοιχούνται στις κοντινότερες συστάδες αλλά θα πρέπει να τηρούν μια μέγιστη απόσταση *δ* από τις υπόλοιπες τροχιές που έχουν προκαθοριστεί από την αρχή. Όσα κέντρα συστάδας δεν μπορούν να δημιουργήσουν μια πλήρη συστάδα απενεργοποιούνται και μπορούν να συμπεριληφθούν σε κάποια άλλη γειτονική συστάδα. Τέλος όσες τροχιές δεν μπορούν να συμπεριληφθούν σε καμία από τις υπάρχουσες συστάδες θεωρούνται θόρυβος και απλά διαγράφονται.
- Στο τελευταίο στάδιο της χωρικής μετατόπισης τροποποιούνται τα σημεία εκείνα των τροχιών που είναι σε απόσταση μεγαλύτερη από την ακτίνα (*δ*/2) από το κέντρο συστάδας τους. Η τροποποίηση γίνεται έτσι ώστε στο σύνολο της τροχιάς όλα τα σημεία να είναι πλέον εντός της ακτίνας αυτής (Εικόνα 5).

2.7. Ανωνυμοποίηση βάσεων δεδομένων κινούμενων αντικειμένων μέσω ομαδοποίησης και σύγχυσης (Anonymization of moving objects databases by clustering and perturbation)

Ο αλγόριθμος W4M (Wait for Me) [2] αποτελεί μια εξέλιξη της προηγούμενης μεθόδου NWA η οποία καταργεί ουσιαστικά την Ευκλείδεια μέθοδο μέτρησης αποστάσεων και την αντικαθιστά με πιο ανεκτικές ως προς τον χρόνο μεθόδους διατηρώντας τη φάση της συσταδοποίησης ως έχει. Αυτό έχει σαν συνέπεια να καταργείται τελείως η φάση της προεπεξεργασίας που χώριζε τις τροχιές σε ισοδύναμες ως προς τον χρόνο κλάσεις. Έτσι στην φάση της συσταδοποίησης αντιμετωπίζεται όλο το σύνολο δεδομένων ως μια κλάση. Η κύρια μέθοδος που χρησιμοποιείται από τον W4M για την μέτρηση αποστάσεων ανάμεσα στις τροχιές είναι η EDR, η οποία επιλέγει μια αλληλουχία ζευγών σημείων από κάθε τροχιά ώστε να



Εικόνα 6: Παράδειγμα EDR αντιστοίχισης και χωροχρονικής επεξεργασίας

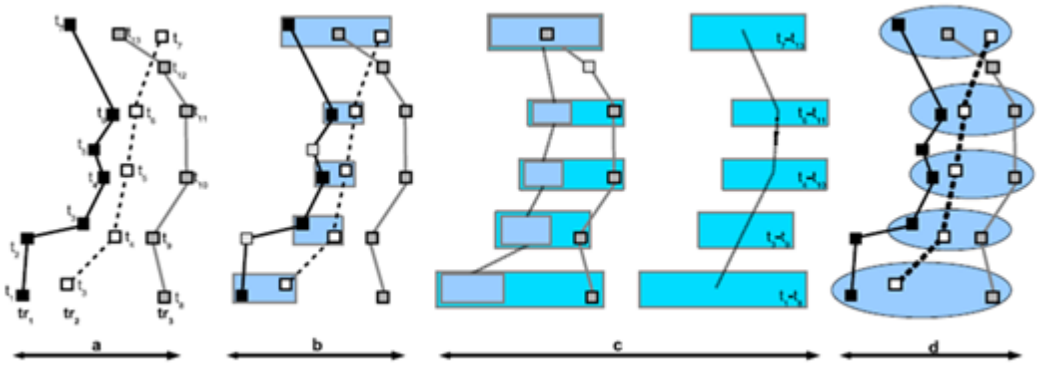
ελαχιστοποιηθεί η ακολουθούμενη επεξεργασία (Εικόνα 6 αριστερά). Σαν αποτέλεσμα κάποια σημεία διαγράφονται και αλλά αντιστοιχούνται. Ωστόσο κάποια διαγεγραμμένα σημεία της κεντρικής τροχιάς *s* μπορούν να αντιγραφούν στις άλλες τροχιές *t* όπως φαίνεται στην Εικόνα 6 (δεξιά). Παρόλα αυτά η εφαρμογή του αλγορίθμου σε βάσεις με πάρα πολλές τροχιές ή πολύ μεγάλες τροχιές τον κάνει ασύμφορο. Για αυτό το λόγο σε αυτές τις περιπτώσεις χρησιμοποιείται η μέθοδος μέτρησης χωροχρονικής απόστασης LSTD της όποιας η πολυπλοκότητα είναι γραμμική σε σχέση με το πλήθος των τροχιών. Άλλη μια παραλλαγή του αλγορίθμου είναι με την εφαρμογή του chunking στην βάση. Σπάζοντας τη βάση ουσιαστικά σε μικρότερες και χρησιμοποιώντας ως μέθοδο μέτρησης χωροχρονικής απόστασης την LSTD.

Συνοπτικά ο αλγόριθμος W4M έχει τις παρακάτω παραλλαγές:

- W4M, με την χρήση της EDR ακολουθούμενη από χωροχρονική επεξεργασία.
- W4M(L), με την χρήση της πιο αποτελεσματικής LSTD αντί της EDR.
- W4M(LC) μια εξέλιξη της W4M(L) που επιταχύνει την ανωνυμοποίηση σε μεγάλες βάσεις σπάζοντας ουσιαστικά τη βάση σε μικρότερες(chunking) και εφαρμόζοντας την W4M(L) σε αυτές.

2.8. Ανωνυμοποίηση Τροχιών Κινούμενων Αντικειμένων: μια προσέγγιση βασισμένη στην γενίκευση (Towards Trajectory Anonymization: a Generalization-Based Approach)

Σε αυτή τη μέθοδο [3] όπως και στην προηγούμενη έχουμε δυο στάδια, αυτό της ομαδοποίησης των τροχιών και αυτό της ανωνυμοποίησης των ομάδων που δημιουργούνται.



Εικόνα 7: Διαδικασία Ανωνυμοποίησης AWO

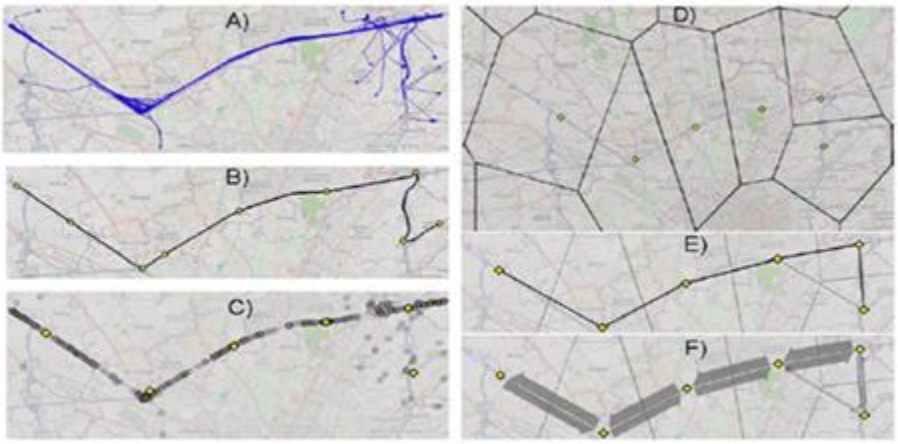
Στο στάδιο της ομαδοποίησης χρησιμοποιείται ο αλγόριθμος multi TGA όπου σε κάθε επανάληψη δημιουργείται μια άδεια ομάδα για την οποία επιλέγεται μια τυχαία αντιπροσωπευτική τροχιά. Ύστερα βρίσκει την κοντινότερη τροχιά, την προσθέτει στην ομάδα και στην περίπτωση του αλγορίθμου multi TGA προχωράει σε τροποποίηση της αντιπροσωπευτικής τροχιάς. Η νέα αντιπροσωπευτική τροχιά δημιουργείται από την ανωνυμοποίηση της προϋπάρχουσας αντιπροσωπευτικής τροχιάς με την καινούρια τροχιά που προστέθηκε στην ομάδα. Η διαδικασία αυτή τελειώνει έως ότου συμπληρωθούν *k* τροχιές για να τηρηθεί η *k*-ανωνυμία και εν συνεχεία αυτές αφαιρούνται από το σύνολο των τροχιών για να μην ξαναχρησιμοποιηθούν. Οι επαναλήψεις σταματάνε όταν μια ομάδα δεν μπορεί να συμπληρώσει *k* τροχιές. Επειδή το κόστος αυτής της αναζήτησης είναι υψηλό χρησιμοποιείται και ο αλγόριθμος fast TGA όπου αγνοεί το βήμα της ανανέωσης του αντιπροσώπου της ομάδας σε κάθε επανάληψη.

Αφού έχουν ολοκληρωθεί οι ομαδοποιήσεις των τροχιών, κάθε τροχιά μέσα στις ομάδες θα πρέπει να ανωνυμοποιηθεί. Σε κάθε ομάδα ο αλγόριθμος ξεκινά βρίσκοντας τη τροχιά που έχει την μικρότερη

απόσταση από τις υπόλοιπες και σε κάθε βήμα αντιστοιχεί τα σημεία της επόμενης τροχιάς που δεν έχει χρησιμοποιηθεί με αυτά της ανωνυμοποιημένης τροχιάς που έχει γίνει μέχρι εκείνη τη στιγμή. Στη συνέχεια συνδέει τα ζεύγη των σημείων των τροχιών αντικαθιστώντας τα με ένα ελάχιστο πλαίσιο οριοθέτησης. Στη διαδικασία αυτή κάθε σημείο που δεν έχει αντιστοιχηθεί αφαιρείται και έτσι δημιουργείται ένα ελάχιστο πλαίσιο οριοθέτησης για κάθε ομάδα που καλύπτει όλα τα σημεία της ομάδας (εικόνα 7d). Σημειώνεται ότι οι σύνδεσμοι των σημείων αυτών που δημιουργήθηκαν είναι διατεταγμένοι, δεν επικαλύπτονται και μπορεί να υπάρχουν μη αντιστοιχισμένα σημεία από οποιεσδήποτε τροχιές.

2.9. Ανωνυμία Δεδομένων Κίνησης μέσω Γενίκευσης - Movement Data Anonymity through Generalization (Voronoys)

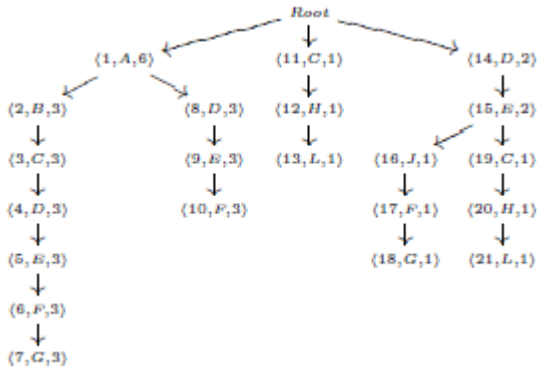
Η μέθοδος αυτή της γενίκευσης [4] εφαρμόζεται σε μια προψηφιδοποιημένη γεωγραφική περιοχή που καλύπτει όλες τις τροχιές ενός συνόλου δεδομένων. Η ψηφιδοποίηση της περιοχής έχει προκύψει από την ακόλουθη διαδικασία. Αρχικά εξάγονται τα χαρακτηριστικά σημεία των τροχιών. Στην συνέχεια ομαδοποιούνται τα γειτονικά σημεία που εξήχθησαν σε συστάδες και από αυτές τις συστάδες που προκύπτουν εξάγονται τα κέντρα τους τα οποία χρησιμοποιούνται για τη δημιουργία της ψηφίδωσης. Έτσι κάθε τροχιά μετατρέπεται σε μια γενικευμένη συνέχεια επισκέψεων μέσα από τα κέντρα των συστάδων όπως φαίνεται στο παράδειγμα της Εικόνας 8.



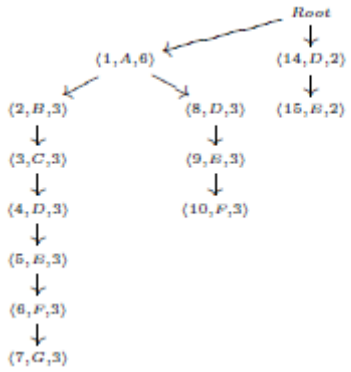
Εικόνα 8: A) Τροχιές B) Χαρακτηριστικά σημεία μιας τροχιάς C) Όλα τα χαρακτηριστικά σημεία των τροχιών D) Ψηφιδοποίηση περιοχής E) Γενικευμένες τροχιές F) Συνοπτική αναπαράσταση των τροχιών

Εικόνα 9: Παράδειγμα ανωνυμοποίησης με τους αλγορίθμους της μεθόδου Generalization

Έχοντας πλέον γενικεύσει τις τροχιές η μέθοδος προχωράει σε ανωνυμοποίηση τους η οποία αποτελείται από 3 βήματα.



Εικόνα 11: Προθεματικό δέντρο



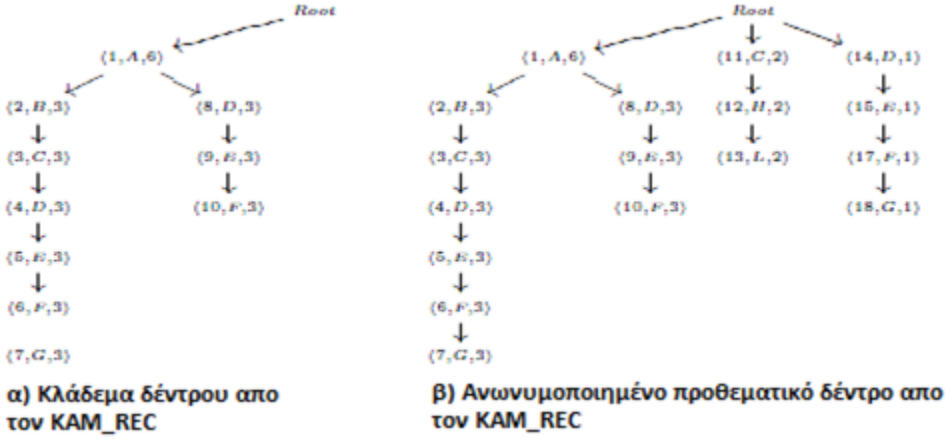
Εικόνα 10: Προθεματικό δέντρο με εφαρμογή του KAM_CUT

1. Την δημιουργία ενός προθεματικού δέντρου που αποτυπώνει όλες τις τροχιές του συνόλου δεδομένων.
2. Τον αλγόριθμο που κόβει τις τροχιές ή υποτροχιές του δέντρου που δεν τηρούν την *k*-ανωνυμία.
3. Και τον αλγόριθμο που αναλύει ξανά το δέντρο στις πλέον ανωνυμοποιημένες τροχιές όπως φαίνεται στο παράδειγμα Εικόνα 9β και 9γ.

Το δέντρο ορίζεται από έναν πεπερασμένο αριθμό κόμβων, από τα άκρα του και από τον κόμβο-ρίζα του δέντρου. Κάθε κόμβος έχει έναν πατέρα κόμβο και μπορούμε να καταλήξουμε σε αυτόν ξεκινώντας από την ρίζα του δέντρου. Ο κάθε κόμβος ορίζεται από τον αναγνωριστικό αριθμό του, το σημείο της τροχιάς που αντιπροσωπεύει, το πόσες φόρες εξυπηρετεί μια τροχιά από την ρίζα του δέντρου και από τους κόμβους παιδιά του όπως αυτά φαίνονται στην Εικόνα 10 που είναι βασισμένη στο παράδειγμα της Εικόνας 9α.

Για την φάση του κλαδέματος του δέντρου υπάρχουν δυο αλγόριθμοι που χρησιμοποιούνται. Ο KAM_CUT ο οποίος χρησιμοποιείται κυρίως για πυκνά σύνολα δεδομένων και ο KAM_REC ο οποίος προσπαθεί να διατηρήσει υποτροχιές που τηρούν την *k*-ανωνυμία.

Ο KAM_CUT αλγόριθμος ψάχνει για τους κόμβους πιο κοντά στη ρίζα που εξυπηρετούν τροχιές λιγότερες από την *k*-ανωνυμία και αφαιρεί τα υποδέντρα τους όπως φαίνεται στην Εικόνα 11 όπου το επιτρεπτό επίπεδο της *k*-ανωνυμίας είναι από 3 και πάνω. Έτσι το δέντρο αναλύεται ξανά στις



Εικόνα 12: Προθεματικό δέντρο με εφαρμογή του KAM_REC

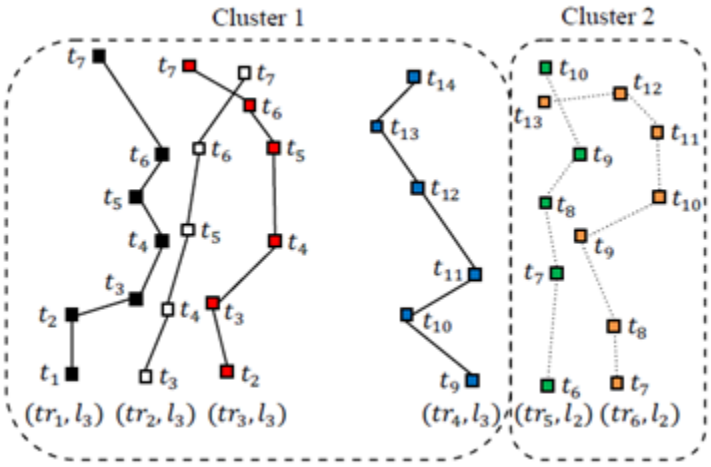
ανωνυμοποιημένες πλέον τροχιές.

Στον αλγόριθμο KAM_REC χρησιμοποιείται μια επιπλέον παράμετρος, αυτή του ποσοστού της υποτροχιάς που καλύπτει την αρχική τροχιά. Η παράμετρος αυτή προκαθορίζεται όπως η *k*-ανωνυμία και στο παρόν παράδειγμα είναι το 40% της αρχικής τροχιάς. Έτσι ξεκινώντας ο αλγόριθμος ψάχνει για τους κόμβους πιο κοντά στη ρίζα που εξυπηρετούν τροχιές λιγότερες από την *k*-ανωνυμία που έχει συμφωνηθεί. Αναγνωρίζει ολόκληρες τις τροχιές τους, τις αφαιρεί από το δέντρο και τις εισάγει σε μια προσωρινή λίστα. Πλέον το δέντρο παίρνει την μορφή της Εικόνας 12α. Από την λίστα αυτή τώρα ψάχνει για αυτές τις υποτροχιές που τηρούν την *k*-ανωνυμία και αποτελούν τουλάχιστον το προσυμφωνημένο ποσοστό της αρχικής τροχιάς(40%). Όσες υποτροχιές τηρούν αυτές τις προϋποθέσεις επανέρχονται στο δέντρο κάτω από την ρίζα και από εκεί αναλύονται ξανά στις ανωνυμοποιημένες πλέον τροχιές όπως και στον αλγόριθμο KAM_CUT(Εικόνα 12β).

2.10. Μια προσέγγιση βασιζόμενη στη συσταδοποίηση για εξατομικευμένη προστασία ιδιωτικότητας κατά την δημοσίευση των δεδομένων τροχιών κινούμενων αντικειμένων (A Clustering-Based Approach for Personalized Privacy Preserving Publication of Moving Object Trajectory Data)

Σε αυτή την μέθοδο ανωνυμοποίησης που βασίζεται στην συσταδοποίηση βάση του επιπέδου ιδιωτικότητας εφαρμόζονται δυο κύριες φάσεις:

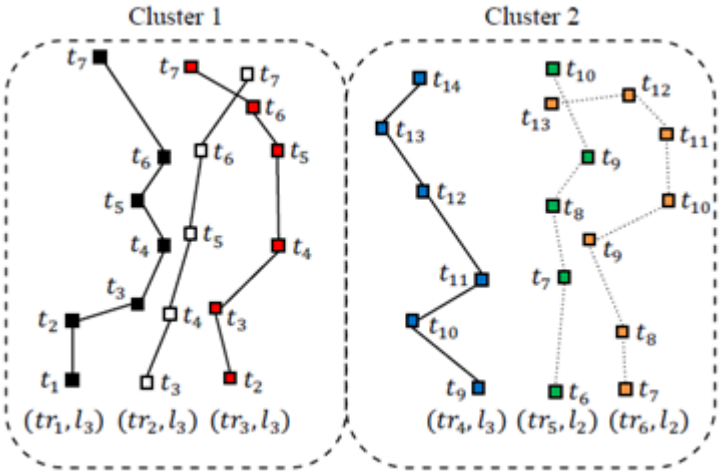
- Η φάση της συσταδοποίησης βάση του επιπέδου ιδιωτικότητας της κάθε τροχιάς.
- Η φάση της ανωνυμοποίησης των τροχιών εντός της κάθε συστάδας.



Εικόνα 13: Συσταδοποίηση των τροχιών με την χρήση της πρώτης στρατηγικής

Στη φάση της συσταδοποίησης προτείνονται δυο διαφορετικές στρατηγικές. Στην πρώτη στρατηγική οι τροχιές κατηγοριοποιούνται σε ομάδες βάση του επιπέδου ιδιωτικότητας και εν συνεχεία κάποιες συστάδες δημιουργούνται για κάθε επίπεδο ιδιωτικότητας όπως φαίνεται στην εικόνα 13, όπου τροχιές (tr) του ίδιου επιπέδου (l) μπαίνουν στις ίδιες συστάδες . Όπως μπορούμε να παρατηρήσουμε και από το παράδειγμα αυτή η στρατηγική δεν είναι η πιο αποτελεσματική, διότι μπορεί να υπάρξουν τροχιές που είναι πιο μακριά από τις υπόλοιπες της ίδιας συστάδας και να είναι πιο κοντά σε άλλες τροχιές άλλου επιπέδου ιδιωτικότητας. Στη δεύτερη στρατηγική οι τροχιές κατηγοριοποιούνται πάλι βάση του επιπέδου

ιδιωτικότητας των τροχιών και εν συνεχεία η συσταδοποίηση ξεκίνα από την ομάδα με το υψηλότερο



Εικόνα 14: Συσταδοποίηση των τροχιών με την χρήση της δεύτερης στρατηγικής

επίπεδο ιδιωτικότητας με φθίνουσα σειρά καθώς οι τροχιές αυτές έχουν προτεραιότητα. Η δημιουργία των συστάδων γίνεται έτσι ώστε το μέγεθος της κάθε συστάδας να είναι ανάλογο του επιπέδου ιδιωτικότητας του κεντροειδούς της συστάδας αυτής και η απόσταση της κάθε τροχιάς από το κεντροειδές να είναι μικρότερη από ένα κατώφλι *δ*. Έτσι με την χρήση των ίδιων τροχιών του παραδείγματος της εικόνας 13 οι συστάδες που θα δημιουργηθούν θα είναι όπως φαίνονται στην εικόνα 14 όπου πλέον η τροχιά tr_4 ανήκει στη δεύτερη συστάδα με τις τροχιές που βρίσκεται πιο κοντά. Η εφαρμογή αυτής της στρατηγικής γίνεται με την χρήση του αλγορίθμου CTR που παίρνει σαν είσοδο το σύνολο των τροχιών TR, την αρχική ακτίνα των συστάδων δ_{init}, την μεγίστη ακτίνα που μπορούν να πάρουν οι συστάδες δ_{max} και σαν έξοδο επιστρέφει ένα σύνολο συστάδων *C*. Για την δημιουργία της κάθε συστάδας ο CTR χρησιμοποίει τον αλγόριθμο PTR ο οποίος παίρνει σαν είσοδο ένα σύνολο τροχιών TR_r, την ακτίνα της συστάδας *d*, το επίπεδο ιδιωτικότητας *l* και επιστρέφει μια συστάδα *C(δ,l)*.

Μόλις ολοκληρώνεται η διαδικασία της συσταδοποίησης ξεκινά η διαδικασία της ανωνυμοποίησης με την χρήση του αλγορίθμου ATR. Ο αλγόριθμος αυτός παίρνει σαν είσοδο το σύνολο των συστάδων, ένα χωρικό κατώφλι *δ* και επιστρέφει ένα σύνολο ανωνυμοποιημένων τροχιών.

3. Αντικείμενο εργασίας και κριτήρια αποτίμησης

Το αντικείμενο της εργασίας αυτής είναι η ενιαία πλέον σύγκριση των μεθόδων ανωνυμοποίησης με χρήση κοινών μέτρων αξιολόγησης, τα οποία μπορούν να εφαρμοστούν και για τις τέσσερις μεθόδους ανωνυμοποίησης NWA, W4M, AWO, Generalization και τις παραλλαγές τους. Αυτά τα μέτρα αξιολόγησης θα εφαρμοστούν σε κοινά σύνολα δεδομένων κίνησης για να καταδείξουν την συμπεριφορά των ως προς σύγκριση μεθόδων σε διαφορετικού μεγέθους σύνολα δεδομένων. Τα μέτρα αυτά βάση των οποίων θα γίνει η αξιολόγηση είναι τα ακόλουθα:

- *Distortion*, το οποίο θα καταδείξει το επίπεδο παραμόρφωσης των αρχικών δεδομένων ύστερα από την ανωνυμοποίηση τους. Κάνοντας χρήση ερωτημάτων τύπου *Possibly_Sometime_Inside (P_S_I)* [2] και *Definitely_Always_Inside (D_A_I)* [2] για όλες τις μεθόδους δοκιμάζοντας τα σε διαφορετικές χωροχρονικές περιοχές για πιο αντικειμενικά αποτελέσματα.
- *F-measure* [11], το οποίο αποτελεί έναν αρμονικό μέσο όρο της ακρίβειας (α) και της ανάκλησης (β) ως (2αβ/(α+β)) των σημείων των αρχικών τροχιών (S) πάνω στις ανωνυμοποιημένες (S'). Όπου η ακρίβεια ορίζεται ως α= |S'∩S|/|S'| και η ανάκληση ως β=|S'∩S|/|S|.
- *Runtime Performance*, το οποίο είναι ο συνολικός χρόνος εκτέλεσης της διαδικασίας ανωνυμοποίησης για την κάθε μέθοδο.
- *Removed Points*, το οποίο είναι ο συνολικός αριθμός των σημείων που διεγράφησαν από το σύνολο δεδομένων. Τα σημεία αυτά μπορεί να προκύπτουν είτε από τροχιές που θεωρούνται θόρυβος είτε από σημεία τα οποία κατά την ανωνυμοποίηση δεν μπόρεσαν να αντιστοιχηθούν.

Με την επιλογή αυτών των κριτηρίων καλύπτουμε τα περισσότερα χαρακτηριστικά αξιολόγησης μιας μεθόδου ανωνυμοποίησης. Με το κριτήριο του Distortion μπορούμε να εξάγουμε συμπεράσματα για την ποιότητα των δεδομένων μετά την ανωνυμοποίηση τους. Βασισμένοι στο μοντέλο Trajcevski [13] εστιάζουμε σε δυο ακραία ερωτήματα για το εάν ένα κινούμενο αντικείμενο βρισκόταν κάποιες φόρες ή πάντα εντός μιας περιοχής για ένα συγκεκριμένο χρονικό διάστημα. Από τα ερωτήματα αυτά καταλήγουμε στους δυο τύπους ερωτημάτων *Possibly_Sometime_Inside (P_S_I)* [2] και *Definitely_Always_Inside (D_A_I)* [2] που εφαρμόζονται σε ένα μεγάλο αριθμό τυχαίων χωροχρονικών περιοχών για τα αυθεντικά και ανωνυμοποιημένα δεδομένα ώστε να μας δώσουν το πλήθος των επιτυχημένων απαντήσεων ανά περίπτωση. Με τον τρόπο αυτό έγινε το μέτρο της παραμόρφωσης μετρήσιμο εφαρμόζοντας την συνάρτηση |Q(D)-Q(D')|/Q(D) όπου Q(D) είναι το πλήθος των επιτυχημένων απαντήσεων στην ερώτηση που κάναμε για τα αυθεντικά δεδομένα και Q(D') αντίστοιχα για τα ανωνυμοποιημένα.

Το κριτήριο *F-measure* [11] αποτελεί και αυτό ένα επιπλέον μέτρο αξιολόγησης της παραμόρφωσης των ανωνυμοποιημένων δεδομένων συγκριτικά με τα αρχικά δεδομένα. Η διαφορά του σε σχέση με το προηγούμενο κριτήριο πέρα από τον τρόπο εφαρμογής του είναι ότι το κριτήριο αυτό δεν λαμβάνει υπόψη του την κίνηση των αντικειμένων ως προς το χρόνο και έτσι εστιάζει μόνο στα χωρικά δεδομένα των τροχιών. Ο τρόπος εφαρμογής του κριτηρίου ανάμεσα στις αρχικές και ανωνυμοποιημένες τροχιές αυτού γίνεται ως εξής. Αρχικά χωρίζεται ο γεωγραφικός χώρος που καλύπτουν οι τροχιές σε ένα πίνακα διατάσεων n x n. Με αυτό τον τρόπο καταγράφουμε μια τροχιά από ποιες τετραγωνικές περιοχές του πίνακα περνάει και έτσι μετατρέπουμε την κάθε τροχιά σε μια αλληλουχία τετραγωνικών περιοχών πάνω στον πίνακα. Αυτή η μετατροπή εφαρμόζεται σε όλες τις τροχιές στα αρχικά και στα ανωνυμοποιημένα δεδομένα. Μετά την ολοκλήρωση της μετατροπής αυτής έχουμε μετρήσιμο πλέον τον αριθμό των σημείων των τροχιών που πέρασαν από την κάθε περιοχή και σαν ελάχιστο όριο διατηρούμε εκείνες τις περιοχές στις οποίες το πλήθος των σημείων από τις τροχιές που πέρασαν είναι μεγαλύτερο από το 1% επί του συνόλου των σημείων του συνόλου δεδομένων. Έχοντας πλέον προπαρασκευάσει τα δεδομένα μπορούμε να εφαρμόσουμε την μέθοδο F-measure [[10]] πάνω σε αυτά. Το πλήθος των κοινών περιοχών ανάμεσα

στα αρχικά και ανωνυμοποιημένα δεδομένα ως προς το πλήθος των περιοχών των ανωνυμοποιημένων δεδομένων ορίζει την ακρίβεια α=|S'∩S|/|S'|. Αντίστοιχα το πλήθος των κοινών περιοχών ανάμεσα στα αρχικά και στα ανωνυμοποιημένα δεδομένα ως προς το πλήθος των περιοχών των αρχικών δεδομένων ορίζει την ανάκληση ως β=|S'∩S|/|S|. Το τελικό μετρήσιμο αποτέλεσμα του κριτηρίου αυτού ονομάζεται αρμονικός μέσος όρος, προκύπτει από τον συνδυασμό της ακρίβειας *α* και της ανάκλησης *β* και ορίζεται ως 2αβ/(α+β).

Το κριτήριο Runtime Performance που χρησιμοποιούμε δεν μας δίνει στοιχεία για τα χωροχρονικά δεδομένα μετά την ανωνυμοποίηση αλλά εστιάζει μόνο στον χρόνο που διαρκεί η ανωνυμοποίηση της κάθε μεθόδου. Αυτό προϋποθέτει ότι όλες οι μέθοδοι θα εφαρμοστούν σε κοινό περιβάλλον ώστε να μπορούν να εκμεταλλευτούν τους ίδιους πόρους συστήματος. Άρα εκτός από τα τεχνικά χαρακτηριστικά του συστήματος απαιτείται η εφαρμογή των μεθόδων να γίνει και σε κοινό λειτουργικό σύστημα.

Το τελευταίο κριτήριο το οποίο πρόκειται να εφαρμοστεί, Removed Points, επιστρέφει ουσιαστικά το πλήθος των σημείων που διεγράφησαν κατά την ανωνυμοποίηση. Παρουσιάζει δηλαδή τον τρόπο με τον οποίο αντιμετωπίζουν οι μέθοδοι την ανωνυμοποίηση των τροχιών δίνοντας περισσότερη ή λιγότερη πληροφορία. Είναι ένα κριτήριο που από μόνο του δεν έχει μεγάλη αξία αλλά όταν αυτό συνδυάζεται με τα παραπάνω κριτήρια αξιολόγησης της παραμόρφωσης μπορεί να μας δώσει πολύ χρήσιμα συμπεράσματα για τις μεθόδους αυτές.

Όπως αναφέραμε και στο τελευταίο κριτήριο δεν έχει τόση μεγάλη σημασία να δούμε τα κριτήρια αυτά μεμονωμένα όσο να τα δούμε συνδυαστικά και να παρατηρήσουμε πως οι μέθοδοι αυτές λειτουργούν σε ειδικές περιπτώσεις σε ένα μεγάλο φάσμα επιπέδων ανωνυμοποίησης.

Η εφαρμογή των παραπάνω μέτρων αξιολόγησης θα γίνει για διαφορετικά μεγέθη *k*-ανωνυμίας ξεκινώντας με επίπεδο ανωνυμοποίησης *k*=3 και φτάνοντας έως και *k*=100 στα σύνολα δεδομένων που μας το επιτρέπουν λόγω του μεγάλου όγκου πληροφοριών. Επίσης θα δοκιμαστούν για διαφορετικά μεγέθη ακτίνας *δ* για τις μεθόδους ανωνυμοποίησης NWA και W4M, όπου αυτό απαιτείται, όχι για λόγο σύγκρισης της ίδιας της μεθόδου αλλά για να δούμε την συμπεριφορά τους ως προς τις υπόλοιπες μεθόδους. Επιπλέον θα εφαρμοστούν σε τρία σύνολα δεδομένων, τα οποία περιγράφονται αναλυτικά στην πειραματική μελέτη, ώστε να είμαστε σίγουροι ότι τα αποτελέσματα που θα προκύψουν δεν θα ευνοήσουν κάποια μέθοδο και θα αναδείξουν πλεονεκτήματα ή μειονεκτήματα ανάλογα με τις ιδιαιτερότητες των συνόλων. Τα αποτελέσματα που θα προκύψουν από αυτή την έρευνα θα αποτελέσουν τα μέτρα αποτίμησης για τις μεθόδους ώστε να καταλήξουμε στην καλύτερη δυνατή εκμετάλλευση τους ανάλογα με το σύνολο δεδομένων και το επίπεδο της *k*-ανωνυμίας το οποίο θα εφαρμόσουμε.

4. Περιγραφή υλοποίησης της εργασίας σύγκρισης τεχνικών προστασίας ιδιωτικότητας

Για την υλοποίηση της εργασίας αυτής, έχοντας ως στόχο τον προσδιορισμό μετρήσιμων μεγεθών για την σύγκριση των μεθόδων ανωνυμοποίησης χρειάστηκε να πραγματοποιηθεί μια σειρά από βήματα. Η σειρά αυτών των βημάτων εφαρμόστηκε με τον ίδιο ακριβώς τρόπο και στα τρία σύνολα δεδομένων που επιλέχτηκαν να εφαρμοστούν οι μέθοδοι ανωνυμοποίησης. Στο πρώτο βήμα προεπεξεργαζόμαστε το σύνολο των δεδομένων να είναι έτοιμο προς επεξεργασία από την μέθοδο ανωνυμοποίησης, εν συνεχεία στο δεύτερο βήμα ανωνυμοποιούμε το σύνολο δεδομένων για κάθε μια από τις μεθόδους για όλα τα επίπεδα ανωνυμίας. Τέλος, στο τελευταίο βήμα εισάγουμε τα αποτελέσματα των μεθόδων ανωνυμοποίησης σε μια βάση δεδομένων πάνω στην οποία εφαρμόζουμε SQL ερωτήματα. Τα ερωτήματα αυτά χρησιμοποιούν τον εκάστοτε πίνακα με τα ανωνυμοποιημένα δεδομένα μαζί με τον πίνακα που περιέχει τα αρχικά δεδομένα για να μας δώσουν τα τελικά αποτελέσματα. Αναλυτικά τα βήματα αυτά περιγράφονται στην συνέχεια.

4.1. Βήμα 1ο: Προεπεξεργασία δεδομένων

Κατά την πρώτη φάση της προεπεξεργασίας επιλέγουμε από το σύνολο των δεδομένων εκείνες τις στήλες που είναι απαραίτητες. Αυτές οι στήλες είναι οι εξής: το ID της τροχιάς, η στήλη με την ημερομηνία και την ώρα και τέλος οι στήλες που προσδιορίζουν γεωγραφικά το σημείο (longitude και latitude). Σε κάποια σύνολα δεδομένων υπήρχε διαθέσιμη η πληροφορία object_id και trajectory_id δηλαδή ότι ένα αντικείμενο μπορεί να έχει περισσότερες από μια τροχιές. Σε αυτή την περίπτωση ενοποιήσαμε τα δεδομένα αυτά σε μια στήλη στην οποία η κάθε τροχιά είχε μοναδικό αριθμό.

Μετά από αυτό κομμάτι της προεπεξεργασίας ανεβάζουμε το σύνολο δεδομένων σε έναν πίνακα μιας βάσης δεδομένων για να επεξεργαστούμε αυτή την φορά τις τροχιές που περιέχει το σύνολο δεδομένων. Από το ακόλουθο ερώτημα 1 εξάγουμε την πληροφορία για το ποιες τροχιές και σε ποια σημεία της τροχιάς παρουσιάζουν χρονικά κενά πάνω από 12 ώρες.

```
select a.rn, a.tr_id,
to_timestamp('020' || a.datetime, 'YYMMDDHH24MISS'),
to_timestamp('020' || b.datetime, 'YYMMDDHH24MISS'),
substring(to_char(to_timestamp('020' || a.datetime, 'YYMMDDHH24MISS'),'YYMMDDHH24MISS')
from 4 for 9)::integer date1,
substring(to_char(to_timestamp('020' || b.datetime, 'YYMMDDHH24MISS'), 'YYMMDDHH24MISS')
from 4 for 9)::integer date2
from
        (select row_number() over(order by tr_id,datetime) rn,a.tr_id,a.datetime from
base a) a,
        (select row_number() over(order by tr_id,datetime) rn,a.tr_id,a.datetime from
base a) b
where a.rn-1=b.rn
and a.tr_id=b.tr_id
and to_timestamp('020' || a.datetime, 'YYMMDDHH24MISS')>
to_timestamp('020' || b.datetime, 'YYMMDDHH24MISS') + '1 hour'::INTERVAL *12
```

Ερώτημα 1: Εύρεση των σημείων που πρέπει να διασπαστούν οι τροχιές

Αυτές οι τροχιές θα πρέπει να διασπαστούν. Την λειτουργία αυτή την αναλαμβάνει το ακόλουθο ερώτημα 2 το οποίο ανεβάζει το ID της τροχιάς κατά ένα από το πρώτο σημείο που υπάρχει το χρονικό κενό μέχρι το τέλος της τροχιάς.

Εδώ όμως προκύπτουν τα εξής προβλήματα. Εάν η τροχιά έχει περισσότερα από ένα χρονικά κενά αυτά δεν διασπώνται με μια εκτέλεση του ερωτήματος 2 και αν υπάρχει ήδη κάποια άλλη τροχιά με το ID+1 αυτές θα ενοποιηθούν. Οπότε με την χρήση του 1ου ερωτήματος κάνουμε μια ομαδοποίηση ως προς τις τροχιές για να δούμε τον μέγιστο αριθμό διασπάσεων που μπορεί να χρειαστεί μια τροχιά. Έτσι κάνουμε μια ενημέρωση πρώτα του trajectory_ID πολλαπλασιάζοντας το με 10 ή 100 ανάλογα με το αν ο αριθμός των διασπάσεων είναι μονοψήφιος ή διψήφιος. Με τον τρόπο αυτό εξασφαλίζουμε ότι καμιά άλλη τροχιά δεν θα επηρεαστεί από την κάθε διάσπαση. Έχοντας πλέον εξασφαλίσει ότι οι τροχιές δεν πρόκειται να επηρεαστούν μπορούμε να εισάγουμε το 2ο ερώτημα σε μια συνάρτηση που θα το επαναλάβει τόσες φορές όσες τα περισσότερα χωροχρονικά κενά σε μια τροχιά. Η διαδικασία αυτή επαναλήφθηκε με αντίστοιχα ερωτήματα ώστε να διασπάσουμε τις τροχιές εκείνες που είχαν μεγάλα χωρικά κενά. Όμως στα σύνολα δεδομένων που χρησιμοποιήσαμε δεν βρέθηκε κάτι το αξιόλογο.

```
update base set tr_id= base.tr_id + 1 from
        (
        select c.* from(
        select a.rn, a.tr_id,
        to_timestamp('020' || a.datetime, 'YYMMDDHH24MISS'),
        to_timestamp('020' || b.datetime, 'YYMMDDHH24MISS') from
        (select row_number() over(order by tr_id,datetime) rn,a.tr_id,a.datetime from
base a ) a,
        (select row_number() over(order by tr_id,datetime) rn,a.tr_id,a.datetime from
base a ) b
        where a.rn-1=b.rn
        and a.tr_id=b.tr_id
        and to_timestamp('020' || a.datetime, 'YYMMDDHH24MISS') >
        to_timestamp('020' || b.datetime, 'YYMMDDHH24MISS') + '1 hour'::INTERVAL *12
        ) a,
        (
        select max(a.rn) rn, a.tr_id from (
                select row_number() over(order by tr_id,datetime) rn, * from base
                ) a
        group by a.tr_id
        ) b,
        (
        select row_number() over(order by tr_id,datetime) rn, * from base
        ) c
        where c.rn>=a.rn
        and c.rn<=b.rn
        and c.tr_id=a.tr_id
        and c.tr_id=b.tr_id
        order by 1
        ) a
where base.tr_id=a.tr_id
and base.datetime=a.datetime
```

Ερώτημα 2: Διάσπαση των τροχιών στο πρώτο σημείο που παρουσιάζεται μεγάλο χρονικό κενό

Μετά από την διαδικασία της διάσπασης των τροχιών χρειάστηκε να αφαιρεθούν όλες εκείνες οι τροχιές οι οποίες είχαν πολύ μικρή διάρκεια ή η απόσταση την οποία διένυσαν ήταν πολύ μικρή. Για την διαγραφή των τροχιών με διάρκεια κάτω από τρεις ώρες εφαρμόστηκε το παρακάτω ερώτημα 3. Ενώ για την διαγραφή των τροχιών που διένυσαν κάτω από 5 km εφαρμόστηκε το ακόλουθο ερώτημα 4.

```
delete from base where tr_id in (
        select a.tr_id from (
                select * from (
                select tr_id, to_timestamp('020' || min(a.datetime) , 'YYMMDDHH24MISS')
min
                , to_timestamp('020' || max(a.datetime), 'YYMMDDHH24MISS') max
                 from base a group by tr_id) a
                where
                a.max <
                a.min + '1 hour'::INTERVAL *3
        )a
)
```

Ερώτημα 3: Διαγραφή των τροχιών με μικρή χρονική διάρκεια

```
delete from base where tr_id in (
select a.tr_id from (
        select * from (
                select a.tr_id, sum(a.dist) total_dist from (
                        select
                        asin(
                        sqrt(
                        pow(sin((a.lat*pi()/180-
b.lat*pi()/180)/2),2)*pow(cos((a.lon*pi()/180-b.lon*pi()/180)/2),2)
                        + pow(sin((a.lon*pi()/180-
b.lon*pi()/180)/2),2)*pow(cos((a.lat*pi()/180+b.lat*pi()/180)/2),2)
                        )
                        )* 6371000 dist,
                        a.lat,a.lon,b.lat,b.lon,a.tr_id
                        from
                        (select row_number() over(order by tr_id,datetime) rn,a.* from
base a) a,
                        (select row_number() over(order by tr_id,datetime) rn,a.* from
base a) b
                        where a.rn-1=b.rn
                        and a.tr_id=b.tr_id
                ) a
                group by a.tr_id
        ) a
        where total_dist < 5000
)a
)
```

Ερώτημα 4: Διαγραφή των τροχιών που διένυσαν μικρή απόσταση

Το ερώτημα 5 εφαρμόστηκε ώστε να ορίσουμε τα χαρακτηριστικά των συνόλων δεδομένων που χρησιμοποιήθηκαν. Τα χαρακτηριστικά αυτά είναι η ακτίνα, το πλήθος των τροχιών, το πλήθος των σημείων, το μέγιστο πλήθος σημείων σε μια τροχιά και η μέση απόσταση από τα διαδοχικά σημεία του συνόλου δεδομένων. Η παρουσίαση των χαρακτηριστικών αυτών για τα σύνολα δεδομένων που χρησιμοποιήθηκαν ακολουθεί στο πέμπτο κεφάλαιο.

```
--radius(D)
select
(asin(
sqrt(
pow(sin((lat1*pi()/180-lat2*pi()/180)/2),2)*pow(cos((lon1*pi()/180-lon2*pi()/180)/2),2)
+ pow(sin((lon1*pi()/180-
lon2*pi()/180)/2),2)*pow(cos((lat1*pi()/180+lat2*pi()/180)/2),2)
)
)* 6371000)/2 dist_meters
from (
        select max(lat) lat1,max(lon) lon1,min(lat) lat2,min(lon) lon2 from base
) a

--|D|
select count(distinct(tr_id)) from base

--Points
select count(*) from base

--Max length
select count(tr_id) from base group by tr_id order by 1 desc

--AVG step
select  avg(
        asin(
        sqrt(
        pow(sin((a.lat*pi()/180-b.lat*pi()/180)/2),2)*pow(cos((a.lon*pi()/180-
b.lon*pi()/180)/2),2)
        + pow(sin((a.lon*pi()/180-
b.lon*pi()/180)/2),2)*pow(cos((a.lat*pi()/180+b.lat*pi()/180)/2),2)
        )
        )* 6371000
        )
from
        (select row_number() over(order by tr_id,datetime) rn,a.tr_id,a.lon, a.lat from
base a) a,
        (select row_number() over(order by tr_id,datetime) rn,a.tr_id,a.lon, a.lat from
base a) b
where a.rn-1=b.rn
and a.tr_id=b.tr_id
```

Ερώτημα 5: Βασικά χαρακτηριστικά του συνόλου δεδομένων

Μετά την ολοκλήρωση της επεξεργασίας των τροχιών το σύνολο δεδομένων εξάγεται και το προετοιμάζουμε για την κάθε μέθοδο. Για τις μεθόδους NWA, W4M και AWO η τροποποίηση είναι απλή αφαιρώντας μόνο τις επικεφαλίδες και χρησιμοποιώντας για διαχωριστικό χαρακτήρα το TAB. Για την μέθοδο Generalization οι τροχιές θα πρέπει να περάσουν το στάδιο της γενίκευσης πρώτα και από κει οι γενικευμένες τροχιές θα ανωνυμοποιηθούν. Για την διαδικασία αυτή το σύνολο δεδομένων πρέπει να αποκτήσει την ίδια μορφή όπως στις άλλες μεθόδους. Δηλαδή τα σύνολα δεδομένων πρέπει να έχουν την μορφή object_id, datetime, longitude και latitude με διαχωριστικό χαρακτήρα το TAB χωρίς κεφαλίδες και την μορφή της ημερομηνίας ως εξής “YYYYMMDD24hhmmss. Επίσης θα μπορούσαμε αντί να βάλουμε την πλήρη ημερομηνία να διασπάσουμε την χρονική περίοδο του συνόλου δεδομένων σε 100 ή και περισσότερες ενδιάμεσες αριθμημένες χρονικές στιγμές, αφού όλες οι μέθοδοι αντιμετωπίζουν τη datetime στήλη σαν αριθμό.

4.2. Βήμα 2ο: Ανωνυμοποίηση δεδομένων

Έχοντας πλέον τα προεπεξεργασμένα σύνολα δεδομένων μπορούμε να τα χρησιμοποιήσουμε στις μεθόδους ανωνυμοποίησης. Όλες οι μέθοδοι έχουν στηθεί σε ένα σύστημα linux που έχει στηθεί για αυτή

```
+-----------------------------------------------------+
| Never Walk Alone v1.0 (c) 2007                      |
| M. Nanni, KDDlab, ISTI-CNR, Pisa                    |
| Mirco.Nanni@isti.cnr.it                             |
+-----------------------------------------------------+
| Reference:                                          |
|    O. Abul, F. Bonchi, M. Nanni.                    |
|    "Never Walk Alone: Uncertainty for Anonymity     |
|    in Moving Objects Databases". ICDE 2008.         |
+-----------------------------------------------------+

Syntax:

  ./nwa1.0  infile K delta outfile [-pi N] [-delta_max N] [-trash_max N] [-stats]

Mandatory parameters:
  infile  = input trajectories
  K       = anonymity level
  delta   = uncertainty
  outfile = anonymyzed trajectories (translated & trash-less)

Optional parameters:
  -pi N (default: 5)
       trajectories start/end at timestamps that are multiples of N
  -delta_max N (default: 0.01)
       initial max_radius for NWA (fraction of dataset radius)
  -trash_max N  (default: 10)
       global max trash size, expressed as % of dataset size
  -stats  (default: OFF)
       print statistics on STDERR at the end of the computation
```

Εικόνα 15: Παράμετροι της NWA

την διαδικασία.

Η μέθοδος NWA εκτελείται μέσω τερματικού σε γραμμή εντολών. Για την εκτέλεση της ανωνυμοποίησης χρησιμοποιούμε το αρχείο nwa1.0 με παραμέτρους αυτές που φαίνονται στην εικόνα 15. Αναλυτικά οι πρώτες τέσσερεις παράμετροι είναι το όνομα του προεπεξεργασμένου αρχείου με τις τροχιές, το επίπεδο της ανωνυμοποίησης *K*, η ακτίνα *δ* της αβεβαιότητας και το όνομα του αρχείου που θα δημιουργήσει με τις ανωνυμοποιημένες τροχιές. Αυτές οι παράμετροι είναι απαραίτητες για να λειτουργήσει η μέθοδος, υπάρχουν όμως και οι προαιρετικές. Με την παράμετρο *pi* ορίζουμε ότι η στήλη της ημερομηνίας θα πρέπει να είναι πολλαπλάσια ενός αριθμού *N*. Με τη παράμετρο *delta_max* ορίζουμε την αρχική μέγιστη ακτίνα αβεβαιότητας που απαιτείται για την συσταδοποίηση. Με την παράμετρο *trash_max* ορίζουμε το μέγιστο ποσοστό των τροχιών που θα καταλήξουν στον κάδο για την μέθοδο και τέλος η παράμετρος *stats* εμφανίζει περισσότερα στατιστικά στοιχεία στο τέλος της διαδικασίας ανωνυμοποίησης.

Η μέθοδος ανωνυμοποίησης W4M εκτελείται και αυτή μέσω τερματικού με την χρήση των εκτελέσιμων αρχείων *w4m_EDR* και *w4m_LST* που αντιστοιχούν στις μεθόδους μέτρησης απόστασης μεταξύ τροχιών *EDR* και *LST*. Για την εκτέλεση του αρχείου w4m_EDR χρησιμοποιούνται εννέα συνολικά παράμετροι που φαίνονται στην εικόνα 16 που ακολουθεί. Σαν πρώτη παράμετρο παίρνει το όνομα του προεπεξεργασμένου αρχείου με τις αρχικές τροχιές. Σαν δεύτερη παράμετρο παίρνει το πρόθεμα του αρχείου που θα δημιουργήσει γιατί το υπόλοιπο μέρος του ονόματος το συμπληρώνει με στοιχεία της ανωνυμοποίησης. Τρίτη παράμετρος *K* που παίρνει είναι το επίπεδο της ανωνυμοποίησης που θα χρησιμοποιήσει η μέθοδος. Η τέταρτη παράμετρος delta είναι η ακτίνα *δ* της αβεβαιότητας. Με την πέμπτη παράμετρο *radius_max* ορίζουμε την αρχική μέγιστη ακτίνα αβεβαιότητας που απαιτείται για την συσταδοποίηση. Με την έκτη παράμετρο *trash_max* ορίζεται το μέγιστο ποσοστό των τροχιών που θα καταλήξουν στον κάδο. Οι υπόλοιπες τρεις παράμετροι *delta_match_x, delta_match_y* και *delta_match_t* χρησιμοποιούνται για την αντιστοίχιση των τροχιών που γίνεται από την μέθοδο EDR. Για την εκτέλεση της

```
+-----------------------------------------------------+
| Wait 4 Me v1.1 (c) 2010                             |
| O. Abul, Dept Comp. Eng., TOBBS Univ., Ankara       |
|         OsmanAbul@etu.edu.tr                        |
| M. Nanni, KDDlab, ISTI-CNR, Pisa                    |
|         Mirco.Nanni@isti.cnr.it                     |
+-----------------------------------------------------+
| Reference:                                          |
|   O. Abul, F. Bonchi, M. Nanni.                     |
|   "Anonymization of moving objects databases by     |
|   clustering and perturbation".                     |
|   Information Systems J., 35(8), 2010, pp. 884-910. |
+-----------------------------------------------------+
Parsing parameters...

Syntax:

  W4M(4) infile outfileprefix K delta radius_max trash_max edr_dx edr_dy edr_dt

-------------------------------------------------------------------------------
  infile        = input trajectories file name
  outfileprefix = anonymyzed trajectories (translated & trash-less)
  K             = anonymity level
  delta         = uncertainty
  radius_max    = initial maximum radius used in clustering
  trash_max     = global maximum trash size
  delta_match_x = delta_x used for EDR match
  delta_match_y = delta_y used for EDR match
  delta_match_t = delta_t used for EDR match
-------------------------------------------------------------------------------
```

Εικόνα 16: Παράμετροι της W4M EDR

έκδοσης της W4M που χρησιμοποιεί την LST μέθοδο χρησιμοποιείται το εκτελέσιμο αρχείο w4m_LST το οποίο παίρνει επτά παραμέτρους όπως φαίνεται στην εικόνα 17 με τις έξι πρώτες να είναι ίδιες με αυτές που παίρνει το αρχείο w4m_EDR. Στην έβδομη παράμετρο ορίζουμε σαν *M* την μέγιστη απόσταση που χρησιμοποίει η μέθοδος LST και προτρέπει ότι αν δεν γνωρίζουμε τι ακριβώς τιμή πρέπει να του δώσουμε τότε αυτή η τιμή να είναι το 10.

Για την εκτέλεση της μεθόδου ανωνυμοποίησης AWO χρειάστηκε να εκτελέσουμε τον java κώδικα της εφαρμογής “by Nergiz anonymizer2” μέσα από το πρόγραμμα σύνταξης java εφαρμογών Intellij IDEA στο οποίο τροποποιούμε την συνάρτηση εκτέλεσης της μεθόδου που βρίσκεται στο αρχείο Run.java. Για την

εκκίνηση της εφαρμογής Intellij IDEA εκτελούμε από το φάκελο εγκατάστασης το script /bin/idea.sh. Οι παράμετροι που παίρνει η συνάρτηση εμφανίζονται στην εικόνα 18 όπου η πρώτη παράμετρος είναι το επίπεδο ανωνυμοποίησης *k* και δεύτερη παράμετρος είναι false ή true ανάλογα με το αν ο αλγόριθμος που θα χρησιμοποιηθεί θα είναι ο fastTGA ή ο multiTGA αντίστοιχα. Οι επόμενες δυο παράμετροι ορίζουν την χωροχρονική ευαισθησία. Η πέμπτη παράμετρος παίρνει τιμές LOG ή RADIUS για να χρησιμοποιήσει τις συναρτήσεις κόστους LCM ή RCM αντίστοιχα. Οι επόμενες τρεις παράμετροι χρησιμοποιούνται για την καταστολή του κόστους των x, y, t. Όσο μεγαλύτερος είναι ο αριθμός των παραμέτρων αυτών τόσο μεγαλύτερο είναι το πλήθος των σημείων ενώ στην αντίθετη περίπτωση θα καταστείλει περισσότερα σημεία. Σαν αρχείο εισαγωγής χρησιμοποιεί το αρχείο με όνομα DatasetResample.txt και εξάγει τα αποτελέσματα στο αρχείο output.txt και για να αλλάξουμε τα ονόματα των αρχείων εισαγωγής και εξαγωγής ανατρέχουμε στο java αρχείο Anonymizer.java όπου αυτά ορίζονται.

```
//new Anonymizer(2,true);

//first parameter; k
//second parameter false for fastTGA. haven't try the multiTGA version
//third and fourth space and time sensitivity
//fifth parameter is for the cost function, LOG for LCM, RADIUS for RCM
//the rest of the parameters are suppression costs for x,y,t.
// Using higher values will create larger points
// whereas using lower values will suppress more points.
// May need to run the algorithm several times to find optimal values
// maximizing utility. Suggested values, 0.5, 1, 1.5, 2...
new Anonymizer(3,true,1,0,Anonymizer.RADIUS,1,1,1);
```

Εικόνα 17: Παράμετροι της AWO

Για την εφαρμογή της μεθόδου Generalization θα πρέπει πρώτα οι τροχιές που περιέχει το προεπεξεργασμένο αρχείο να απλοποιηθούν. Η διαδικασία αυτή πραγματοποιείται από την java εφαρμογή "TrackSimplifier" που είναι βασισμένη στην ερευνητική εργασία [14]. Η εφαρμογή αυτή παίρνει δυο παραμέτρους, το όνομα του αρχείου από το οποίο θα εισάγει τα δεδομένα και το όνομα του αρχείου στο οποίο θα εξάγει τα αποτελέσματα. Επίσης υπάρχει και ένα αρχείο με τις παραμέτρους που θα χρησιμοποιηθούν "params.cfg" του οποίου τα περιεχόμενα φαίνονται στην εικόνα 18. Σημειώνουμε ότι για τις παραμέτρους η γωνία μετριέται σε μοίρες και οι αποστάσεις σε μέτρα. Μετά την πραγματοποίηση της απλοποίησης τα δεδομένα που εξάγονται παρουσιάζουν την κάθε τροχιά σε μια γραμμή ως εξής: trajectory_id:longitude1,latitude1,date1 longitude2,latitude2,date2 …

```
1  minAngle = 20
2  minDistance = 100
3  maxDistance = 10000
4  minStopDuration = 300000
5  maxRad = 500
```

Εικόνα 18: Παράμετροι για την γενίκευση των τροχιών

Το αρχείο αυτό που εξάχθηκε από την εφαρμογή "TrackSimplifier" δεν χρειάζεται τροποποιήσεις και είναι έτοιμο προς ανωνυμοποίηση με την χρήση του jar αρχείου TR-Anonymity.jar που το εκτελούμε από

```
haris@ubuntu:~/Diplomatiki/generalization$ java -jar TR-Anonymity.jar
- Number of arguments is uncorrect
- Input file sequences with id:item1 item2 item3...
- Output file of the anonymous sequences
- Value of k for anonymity
- Anonymization method: 1=append; 2=cut with/without append; 3=cut with/without segmentation
- For a random sorting of the sequences you must insert the option -r otherwise -n
 -For Method 1: -append; For method 2: -noappend or -append, For method 3: -segmentationLcs or -nosegmentationLcs
```

Εικόνα 19: Παράμετροι της Generalization

γραμμή εντολών. Όπως και στις προηγούμενες μεθόδους έτσι και σε αυτή για να γίνει σωστά η ανωνυμοποίηση θα πρέπει να βάλουμε και τις σωστές παραμέτρους οι οποίες παρουσιάζονται στην εικόνα 20. Σαν πρώτη παράμετρο δέχεται το όνομα του αρχείου με τις απλοποιημένες τροχιές. Ως δεύτερη παράμετρο δέχεται το όνομα του αρχείου το οποίο θα δημιουργήσει και θα περιέχει τα ανωνυμοποιημένα δεδομένα. Σαν τρίτη παράμετρο δέχεται το επίπεδο της ανωνυμοποίησης που θα πραγματοποιήσει η εφαρμογή. Σαν τέταρτη παράμετρο ορίζεται η μέθοδος που θα χρησιμοποιήσει και στην προκειμένη περίπτωση είναι η μέθοδος 3 cut στην οποία προσθέτουμε επιπλέον την παράμετρο segmentationLcs ή nosegmentationLcs για την εφαρμογή των αλγορίθμων KAM_REC και KAM_CUT αντίστοιχα. Μετά την ολοκλήρωση της ανωνυμοποίησης τα δεδομένα που εξάγονται έχουν την ίδια μορφή με αυτή που εισήχθησαν. Έτσι με μια απλή τροποποίηση τα επαναφέρουμε ξανά σε στήλες ως εξής, η στήλη με το ID της τροχιάς, η στήλη με την ημερομηνία και τέλος οι στήλες longitude και latitude.

Με την ολοκλήρωση του δεύτερου βήματος και την ανωνυμοποίηση των τροχιών για όλες τις μεθόδους και όλες τις περιπτώσεις έχει αποκτηθεί επιπλέον η πληροφορία για το κριτήριο αξιολόγησης *Runtime Performance.* Η απόκτηση αυτής της πληροφορίας έγινε είτε με την χρήση της εντολής *time* για τις περιπτώσεις των μεθόδων NWA, W4M και Generalization που πραγματοποιήθηκαν σε γραμμή εντολών, είτε με την χρήση της συνάρτησης `System.currentTimeMillis()` για την περίπτωση της AWO. Η εντολή *time* χρησιμοποιεί σαν παράμετρο ολόκληρη την εντολή ανωνυμοποίησης, την εκτελεί κανονικά και με την ολοκλήρωση της διαδικασίας επιστρέφει επιπλέον αποτελέσματα σχετικά με τον χρόνο εκτέλεσης στους πόρους του συστήματος. Τα αποτελέσματα αυτά είναι ο πραγματικός χρόνος εκτέλεσης “real”, ο χρόνος που ο χρήστης χρησιμοποίει τους πόρους του συστήματος “user” και ο χρόνος που χρησιμοποιείται η διαδικασία απευθείας από το σύστημα “sys”. Στη προκειμένη περίπτωση για την καταγραφή του *Runtime Performance*, ο χρόνος που χρησιμοποιήθηκε είναι ο πραγματικός. Τέλος η χρήση της Java συνάρτησης `System.currentTimeMillis()`επιστρέφει την χρονική στιγμή στην οποία εκτελείται, οπότε με την καταγραφή της χρονικής στιγμής πριν και μετά την έναρξη της διαδικασίας μπορούμε να πάρουμε την συνολική διάρκεια εκτέλεσης σε millisecond.

Λόγω της πολυπλοκότητας των παραμέτρων και του χρόνου που απαιτείται για να εφαρμόσει κάποιος όλες αυτές τις μεθόδους ανωνυμοποίησης, υλοποιήθηκε ένα shellscript “*anonymization.sh*” το οποίο αυτοματοποιεί όλο το βήμα της ανωνυμοποίησης για τις μεθόδους NWA W4M AWO και Generalization. Η τροποποίηση της εκτέλεσης του script για το πώς θα τρέξουν οι μέθοδοι ανωνυμοποίησης γίνεται μέσω πέντε αρχείων τα οποία περιέχουν όλες τις εναλλακτικές παραμέτρους που θα χρησιμοποιηθούν από τις αντίστοιχες μεθόδους. Τα αρχεία αυτά είναι τα εξής:

1. *k*: Περιέχει την λίστα όλα τα επίπεδα ανωνυμοποίησης που θα χρησιμοποιηθούν
2. *r*: Περιέχει την λίστα με τις ακτίνες που θα χρησιμοποιηθούν από τις μεθόδους NWA και W4M
3. *awo_method*: Περιέχει την λίστα με τους αλγορίθμους της μεθόδου AWO που θα χρησιμοποιήσει το script
4. *gen_method*: Περιέχει την λίστα τους αλγορίθμους της μεθόδου Generalization που θα χρησιμοποιήσει το script

5. *params.cfg*: Περιέχει όλες τις απαραίτητες παραμέτρους για την απλοποίηση των τροχιών της μεθόδου Generalization

Για την εκτέλεση του shellscript απαιτείται από το σύστημα να είναι εγκατεστημένη η *Python* και το *JDK v1.8.0_66* ή νεότερο. Επίσης το shellscript δέχεται τρεις παραμέτρους. Οι παράμετροι αυτοί είναι τα ονόματα των προεπεξεργασμένων συνόλων δεδομένων τα οποία λόγω των ιδιαιτεροτήτων των μεθόδων χρειάζεται να έχουν κάποιες διαφορές στο πεδίο της ημερομηνίας. Το πρώτο αρχείο χρησιμοποιείται από τις μεθόδους W4M και AWO, το δεύτερο αρχείο χρησιμοποιείται από την μέθοδο NWA και το τρίτο αρχείο χρησιμοποιείται από την μέθοδο Generalization. Τα αποτελέσματα από την ολοκλήρωση των ανωνυμοποιήσεων των μεθόδων του shellscript αποθηκεύονται στο κατάλογο ./Results/ ενώ όλα τα logs αποθηκεύονται στον κατάλογο ./logs/ . Τέλος για τον λόγο του ότι στα logs αποθηκεύεται και η συνολική χρονική διάρκεια της κάθε ανωνυμοποίησης μπορούμε εύκολα να εξάγουμε τις τιμές αυτές με την χρήση της bash εντολής “grep” στα αρχεία αυτά.

4.3. Βήμα 3ο: Αξιολόγηση αποτελεσμάτων

Με την εκτέλεση όλων των πιθανών περιπτώσεων ανωνυμοποίησης από τις μεθόδους ξεκινά το τρίτο και τελευταίο βήμα της διαδικασίας αυτής για την αξιολόγηση των αποτελεσμάτων. Ξεκινάμε εισάγοντας την κάθε περίπτωση ανωνυμοποίησης σε έναν ξεχωριστό πινάκα στην βάση δεδομένων δίνοντας ένα χαρακτηριστικό όνομα στον πίνακα ώστε να είναι ευδιάκριτος ο τρόπος με τον οποίο ανωνυμοποιήθηκαν τα αποτελέσματα. Με την ολοκλήρωση της εισαγωγής των δεδομένων στην βάση, μπορούν πλέον να εφαρμοστούν τα ερωτήματα εκείνα τα οποία θα μας δώσουν τα αποτελέσματα για τα κριτήρια αξιολόγησης *Distortion, F-measure* και *Removed Points*.

Για την εξόρυξη των αποτελεσμάτων του κριτηρίου *Distortion* θα πρέπει πρώτα να δημιουργηθεί ένας πινάκας που θα περιέχει τυχαία χωροχρονικά παραδείγματα πάνω στα οποία θα εξεταστεί εάν θα ισχύουν τα ερωτήματα τύπου *Possibly_Sometime_Inside (P_S_I)* [2] και *Definitely_Always_Inside (D_A_I)* [2] . Η δημιουργία αυτών των παραδειγμάτων πραγματοποιείται με την εκτέλεση του ερωτήματος 6 το οποίο επιλέγει χίλια τυχαία χωροχρονικά σημεία εντός των χωρικών ορίων του συνόλου δεδομένων και σε αυτά δίνει μια τυχαία ακτίνα και διάρκεια χρόνου. Με τον τρόπο αυτό δημιουργούνται τυχαίες κυλινδρικές χωροχρονικές περιοχές. Με την δημιουργία αυτών των παραδειγμάτων μπορούν να εκτελεστούν τα ερωτήματα που θα επιστρέψουν την αναλογία σφάλματος για τις περιπτώσεις του PSI και DAI. Για την αναλογία σφάλματος του DAI χρησιμοποιείται το ερώτημα 7 ενώ για την αναλόγια σφάλματος του PSI χρησιμοποιείται το ερώτημα 8. Τα ερωτήματα αυτά ουσιαστικά επιστρέφουν την απόλυτη τιμή της διαφοράς των επιτυχημένων απαντήσεων του πίνακα των αρχικών δεδομένων με τον πίνακα των ανωνυμοποιημένων δεδομένων προς το πλήθος των επιτυχημένων απαντήσεων του αρχικού πίνακα δεδομένων.

```
--Example Generator
Create table examples1 as (
select row_number() over(),substring(to_char(c.drmax, 'MMDDHH24MISS') from 2 for
9)::integer drmax,
substring(to_char(c.drmin, 'MMDDHH24MISS') from 2 for 9)::integer
drmin,c.lnr,c.ltr,c.radius
from(
  select --Create of random max timestamp
  b.drmin + '1 days'::INTERVAL * ROUND(RANDOM() * 10) + '1 days'::INTERVAL * 5 as
drmax,b.drmin,
  b.lnr, b.ltr,radius
  from (
    SELECT --Random times
    (a.dmax - a.dmin) * random() + a.dmin as drmin, a.dmax,
    (a.lnmax - a.lnmin) * random() + a.lnmin as lnr,
    (a.ltmax - a.ltmin) * random() + a.ltmin as ltr,
    (0.5 - 0.05) * random() + 0.05 as radius--Diameter
    FROM   generate_series(1, 1000), --1000 examples will be created
        (       select --Max Min values from the base table
                to_timestamp('020' || max(datetime), 'YYMMDDHH24MISS') dmax,
                to_timestamp('020' || min(datetime), 'YYMMDDHH24MISS') dmin,
                max(lon) lnmax, min(lon) lnmin, max(lat) ltmax, min(lat) ltmin
                from base
        ) a
    ) b
)c
)
```

Ερώτημα 6: Δημιουργία τυχαίων χωροχρονικών περιοχών

```
--Definitele Always Inside error ratio
select abs(a.dai-b.dai_anon)::real/a.dai::real from
(       select count(*) DAI from
        (       select b.tr_id , e.row_number, count(*) tr_points_in from base
b,examples1 e
                where b.datetime between e.drmin and e.drmax
                and sqrt(pow((b.lon - e.lnr),2) + pow((b.lat - e.ltr),2)) <= e.radius
                group by b.tr_id , e.row_number
        )a,
        (       select b.tr_id, count(*) tr_points from base b group by b.tr_id
        )b
        where a.tr_id=b.tr_id
        and a.tr_points_in=b.tr_points
) a,
(       select count(*) DAI_anon from
        (       select b.tr_id , e.row_number, count(*) tr_points_in
                from examples1 e ,extr_w4m_edr_4_005 b --Anon table
                where b.datetime between e.drmin and e.drmax
                and sqrt(pow((b.lon - e.lnr),2) + pow((b.lat - e.ltr),2)) <= e.radius
                group by b.tr_id , e.row_number
        )a,
        (       select b.tr_id, count(*) tr_points from extr_w4m_edr_4_005 b--Anon table
                group by b.tr_id
        )b
        where a.tr_id=b.tr_id
        and a.tr_points_in=b.tr_points
) b
```

Ερώτημα 7: Δείκτης σφάλματος DAI

```
--Possibly_Sometime_Inside error ratio
select abs(a.PSI-b.PSI_anon)::real/a.PSI::real from
(
	select count(*) PSI from
	(
		select b.tr_id , e.row_number, count(*) tr_points_in from base
b,examples1 e
		where b.datetime between e.drmin and e.drmax
		and sqrt(pow((b.lon - e.lnr),2) + pow((b.lat - e.ltr),2)) <= e.radius
		group by b.tr_id , e.row_number
	)a,
	(
		select b.tr_id, count(*) tr_points from base b group by b.tr_id
	)b
	where a.tr_id=b.tr_id
	and a.tr_points_in between 1 and b.tr_points
) a,
(
	select count(*) PSI_anon from
	(
		select b.tr_id , e.row_number, count(*) tr_points_in from examples1 e
,extr_w4m_edr_4_005 b --Anon table
		where b.datetime between e.drmin and e.drmax
		and sqrt(pow((b.lon - e.lnr),2) + pow((b.lat - e.ltr),2)) <= e.radius
		group by b.tr_id , e.row_number
	)a,
	(
		select b.tr_id, count(*) tr_points from extr_w4m_edr_4_005 b --Anon table
		group by b.tr_id
	)b
	where a.tr_id=b.tr_id
	and a.tr_points_in between 1 and b.tr_points
) b
```

Ερώτημα 8: Δείκτης σφάλματος PSI

Για την εξαγωγή των αποτελεσμάτων του κριτηρίου *Removed Points* εφαρμόζουμε ένα απλό ερώτημα που μετράει το πλήθος των σημείων που περιέχει ο αρχικός πίνακας δεδομένων και ο πίνακας με τα ανωνυμοποιημένα δεδομένα και εμφανίζει την διαφορά τους. Για αυτή την περίπτωση το κριτήριο που χρησιμοποιήθηκε είναι το ερώτημα 9.

```
--Removed Points
select a.cn-b.cn "Removed Points" from
(select count(*) cn from base) a,
(select count(*) cn from extr_w4m_edr_4_005) b --Anon table
```

Ερώτημα 9: Πλήθος διαγεγραμμένων σημείων

Για το κριτήριο *F-measure* είναι το ερώτημα 10 αυτό που χρησιμοποιείται ώστε να εξάγουμε τα αποτελέσματα του. Το ερώτημα αυτό εφαρμόζει πλήρως όλη την διαδικασία όπως αυτή έχει περιγραφεί στο κεφάλαιο 3. Προεπεξεργάζεται τις τροχιές μετατρέποντας τες σε μια αλληλουχία σημείων σε ένα δισδιάστατο πίνακα. Εν συνεχεία αφαιρεί όλα τα σπανίως εμφανιζόμενα σημεία των τροχιών και αφού πλέον έχει υπολογίσει την ακρίβεια και την ανάκληση από τα αρχικά και ανωνυμοποιημένα δεδομένα τα χρησιμοποιεί ώστε να πάρουμε σαν αποτέλεσμα τον αρμονικό μέσο όρο *F-measure*.

```
--F-Measure
with array_table_base as (
        select a.tr_id, a.datetime,
        trunc(20*(a.lon-b.minlon)/(b.maxlon-b.minlon)) grid_lon,
        trunc(20*(a.lat-b.minlat)/(b.maxlat-b.minlat)) grid_lat from
        (
                select * from base ------------------------
        ) a,
        (select min(lon) minlon, max(lon) maxlon, min(lat) minlat, max(lat) maxlat  from
base) b
), array_table_anon as (
        select a.tr_id, a.datetime,
        trunc(20*(a.lon-b.minlon)/(b.maxlon-b.minlon)) grid_lon,
        trunc(20*(a.lat-b.minlat)/(b.maxlat-b.minlat)) grid_lat from
        (
                select * from extr_w4m_edr_4_005 --Anon table
        ) a,
        (select min(lon) minlon, max(lon) maxlon, min(lat) minlat, max(lat) maxlat  from
base) b
)
select 2*a6.rec*b6.prec/(a6.rec+b6.prec) fmeasure from
(
        select a5.sst/b5.s rec from --Recall
        (
                select count(a4.*)::real sst from -- S/\S'
                (
                        select b3.* from --All frequent original patterns
                        (
                                select a2.* from --frequent original patterns >0.01%
                                (
                                        select a1.grid_lon, a1.grid_lat, count(*) cnt from -
-grouped patterns
                                        array table base  a1
                                        group by a1.grid_lon, a1.grid_lat
                                ) a2,
                                (select count(*) cnt from base) b2
                                where a2.cnt::real/b2.cnt::real>=0.0001
                        ) a3 inner join array_table_base b3
                        on a3.grid_lon=b3.grid_lon and a3.grid_lat=b3.grid_lat
                )a4 inner join
                (
                        select a2.* from
                        (
                                select a1.grid_lon, a1.grid_lat, count(*) cnt from
                                array table anon a1
                                group by a1.grid_lon, a1.grid_lat
                        ) a2,
                        (select count(*) cnt from base) b2
                        where a2.cnt::real/b2.cnt::real>=0.0001
                ) b4
                on a4.grid_lon=b4.grid_lon and a4.grid_lat=b4.grid_lat
        ) a5,
        (
                select count(b3.*) s from --All frequent original patterns
                (
                        select a2.* from --frequent original patterns >1%
                        (
                                select a1.grid_lon, a1.grid_lat, count(*) cnt from --grouped
patterns
                                array table base a1
                                group by a1.grid_lon, a1.grid_lat
                        ) a2,
                        (select count(*) cnt from base) b2
                        where a2.cnt::real/b2.cnt::real>=0.0001
                ) a3 inner join
                array table base b3
                on a3.grid_lon=b3.grid_lon and a3.grid_lat=b3.grid_lat
        ) b5
) a6,
```

```
(
        select a5.sts/b5.st prec from --Precition
        (
                select count(a4.*)::real sts from -- S'/\ S
                (
                        select b3.* from --All frequent anonymized patterns
                        (
                                select a2.* from --frequent anonymized patterns >0.01%
                                (
                                        select a1.grid_lon, a1.grid_lat, count(*) cnt from --
grouped patterns
                                        array table anon a1
                                        group by a1.grid_lon, a1.grid_lat
                                ) a2,
                                (select count(*) cnt from base) b2
                                where a2.cnt::real/b2.cnt::real>=0.0001
                        ) a3 inner join
                        array table anon b3
                        on a3.grid_lon=b3.grid_lon and a3.grid_lat=b3.grid_lat
                ) a4 inner join
                (
                        select a2.* from
                        (
                                select a1.grid_lon, a1.grid_lat, count(*) cnt from
                                array_table_base a1
                                group by a1.grid_lon, a1.grid_lat
                        ) a2,
                        (select count(*) cnt from base) b2
                        where a2.cnt::real/b2.cnt::real>=0.0001
                )b4
                on a4.grid_lon=b4.grid_lon and a4.grid_lat=b4.grid_lat
        )a5,
        (
                select count(b3.*)::real st from -- All frequent anonymized patterns
                (
                        select a2.* from -- frequent anonymized patterns >0.01%
                        (
                                select a1.grid_lon, a1.grid_lat, count(*) cnt from --
grouped patterns
                                array_table_anon a1
                                group by a1.grid_lon, a1.grid_lat
                        ) a2,
                        (select count(*) cnt from base) b2
                        where a2.cnt::real/b2.cnt::real>=0.0001
                ) a3 inner join
                array_table_anon b3
                on a3.grid_lon=b3.grid_lon and a3.grid_lat=b3.grid_lat
        )b5
) b6
```

Ερώτημα 10: Υπολογισμός F-Measure

5. Πειραματική Μελέτη

5.1. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν

Σε αυτό το μέρος της εργασίας θα πραγματοποιηθεί η πειραματική μελέτη και θα παρουσιαστούν αποτελέσματα για 3 διαφορετικά σύνολα δεδομένων τροχιών κινούμενων αντικειμένων με διαφορετικές ιδιαιτερότητες που επιλέχθηκαν από τον ιστότοπο http://chorochronos.datastories.org/.

Το πρώτο σύνολο δεομένων *Hermes Trucks* προέρχονται από τροχιές φορτηγών εξοπλισμένα με GPS και αναφέρεται στην περίοδο από τον Αύγουστο έως και το Σεπτέμβριο του 2002. Κατά την προεπεξεργασία στο συνόλου δεδομένων, διασπάστηκαν όλες εκείνες οι τροχιές που είχαν χρονικά κενά με πάνω από 12 ώρες αδράνειας, διεγράφησαν όλες εκείνες οι τροχιές με διάρκεια κάτω από 3 ώρες και τέλος διεγράφησαν οι τροχιές εκείνες που είχαν μήκος κάτω από 5000 μέτρα. Στην τελική μορφή του, το σύνολο δεδομένων έχει 163 τροχιές και συνολικά 110804 σημεία.

Το δεύτερο σύνολο δεομένων *Bus* προέρχεται από τροχιές λεωφορείων εξοπλισμένα με GPS και αναφέρεται στην περίοδο από τον Οκτώβριο του 2000 έως και τον Οκτώβριο 2001. Στην προεπεξεργασία αυτού του συνόλου δεδομένων διασπάστηκαν οι τροχιές με χρονικά κενά πάνω από 12 ώρες. Επίσης, θεωρήθηκαν θορυβώδεις εκείνες οι τροχιές με διάρκεια κάτω από 3 ώρες και αυτές με συνολική τροχιά κάτω από 5000 μέτρα. Στην τελική μορφή του το σύνολο δεδομένων έχει 87 τροχιές και συνολικά 61859 σημεία.

Το τρίτο σύνολο δεομένων *Milan* προέρχεται από τροχιές οχημάτων στη πόλη του Μιλάνου και αναφέρεται στον τέταρτο μήνα του 2007. Στην προεπεξεργασία αυτού του συνόλου δεδομένων διασπάστηκαν οι τροχιές με χρονικά κενά πάνω από 6 ώρες. Επίσης θεωρήθηκαν θορυβώδεις εκείνες οι τροχιές με διάρκεια κάτω από 30 λεπτά και αυτές με συνολική τροχιά κάτω από 2000 μέτρα. Στην τελική μορφή του το σύνολο δεδομένων έχει 43571 τροχιές και συνολικά 1478399 σημεία.

Πίνακας 3: Χαρακτηριστικά των συνόλων δεδομένων

D	radius(D)	|D|	Points	Max length	Avg. step
Trucks	17501	163	110804	4676	120
Bus	44646	87	61859	3062	91
Milan	6929	43571	1478399	838	394

Στον παραπάνω πίνακα παρουσιάζονται τα κύρια χαρακτηριστικά των τριών συνόλων δεδομένων *D* που χρησιμοποιήθηκαν. Όπου *radius(D)* είναι το μισό της διαγωνίου του πλαισίου οριοθέτησης των σημείων του D, *|D|* είναι το πλήθος των τροχιών του D, *Points* είναι το πλήθος των σημείων του D, *Max length* είναι το μεγαλύτερο πλήθος των σημείων σε μια τροχιά και τέλος το *Average step* είναι το μέσο χωρικό βήμα σε δυο διαδοχικά σημεία για όλες τις τροχιές.

5.2. Παράμετροι πειράματος

Οι μέθοδοι NWA και W4M έχουν υλοποιηθεί σε γλώσσα C, επίσης οι μέθοδοι AWO και Generalization είναι υλοποιημένες σε Java. Όλες οι ανωνυμοποιήσεις των μεθόδων έχουν πραγματοποιηθεί σε ένα Virtual Machine με λειτουργικό σύστημα Linux Ubundu 14.04 LTS. Τα τεχνικά χαρακτηριστικά του VM είναι τα εξής:

Μνήμη RAM: 2 GB

Επεξεργαστής: 2xIntel Core i7 CPU 920 @ 2.67GHz

Χωρητικότητα δίσκου: 20 GB

Για τις μεθόδους NWA και W4M η παράμετρος Max_Trush, δηλαδή ο μέγιστος αριθμός τροχιών που μπορεί να διαγράφουν, ορίστηκε στο 10% του συνόλου του συνόλου δεδομένων. Για την παράμετρο δ των μεθόδων αυτών έγιναν οι δοκιμές για δ=0,02 και για δ=0,1 ως ευκλείδεια απόσταση μεταξύ των σημείων. Επίσης το Max radius ορίστηκε στο 1 ως ευκλείδεια απόσταση μεταξύ των σημείων. Για την απλοποίηση των τροχιών πριν από την ανωνυμοποίηση της μεθόδου Generalization ορίστηκαν οι παρακάτω παράμετροι:

- MinAngle, είναι η ελάχιστη γωνία αλλαγής κατεύθυνσης της τροχιάς για να θεωρηθεί σημαντική αλλαγή πορείας και ορίστηκε στις 20 μοίρες.
- MinStopDuration, είναι ο ελάχιστος χρόνος που απαιτείται να περάσει ένα αντικείμενο σε ένα σημείο για να θεωρηθεί ότι πραγματοποιείται στάση και ορίστηκε στα 300000 ms.
- MinDistance, είναι η μέγιστη απόσταση μεταξύ δυο διαδοχικών σημείων για να θεωρηθεί ότι πρόκειται για την ίδια θέση και ορίστηκε στα 100 μέτρα.
- MaxDistance, είναι η μέγιστη επιτρεπτή απόσταση μεταξύ δυο διαδοχικών χαρακτηριστικών σημείων και ορίστηκε στα 10000 μέτρα.
- MaxRadius, είναι η επιθυμητή ακτίνα της κάθε συστάδας που εμπεριέχει όλα της τα χαρακτηριστικά σημεία και ορίστηκε στα 500 μέτρα. Επίσης για το σύνολο δεδομένων Bus ορίστηκε στα 2000 μέτρα λόγω της αραιότητας των τροχιών.

Σημειώνεται επίσης ότι για τις μεθόδους NWA και Generalization χρειάστηκε να αλλοιωθούν τα χρονικά δεδομένα των τροχιών σε επίπεδο ημέρας. Στην NWA έγινε η αλλοίωση για να μπορέσουν να δημιουργηθούν όσο το δυνατό λιγότερες ισοδύναμες κλάσεις για το στάδιο της συσταδοποίησης όπως αναφέρεται στην παράγραφο 2.6. Και για την Generalization η αλλοίωση ήταν προαπαιτούμενη από τη εφαρμογή ανωνυμοποίησης.

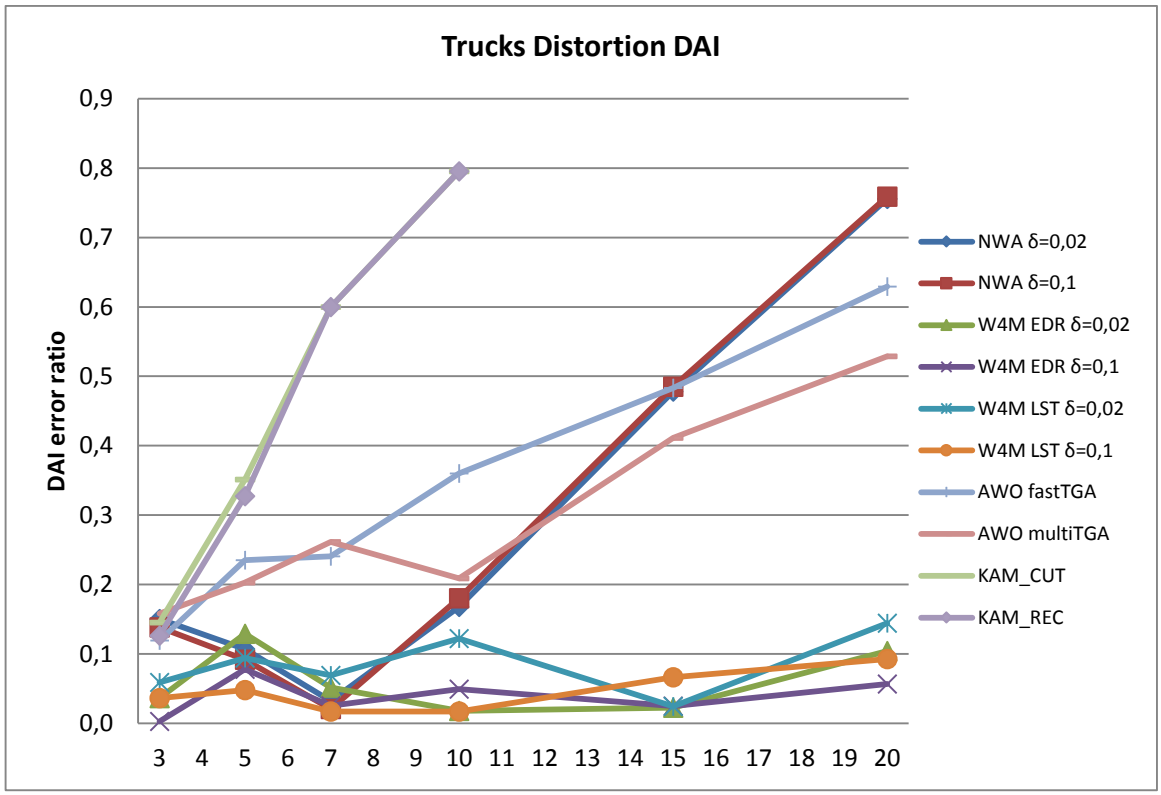
5.3. Παρουσίαση αποτελεσμάτων ανά κριτήριο

Τα αποτελέσματα προέκυψαν από την εφαρμογή των κριτήριων, που αναφέρονται στο κεφάλαιο 3, πάνω στα ανωνυμοποιημένα σύνολα δεδομένων. Για την πραγματοποίηση ανωνυμοποιημένων συνόλων δεδομένων που να καλύπτουν ένα μεγάλο εύρος περιπτώσεων πραγματοποιήθηκαν ανωνυμοποιήσεις με k από 3 έως και 20 ενώ για το σύνολο δεδομένων Milan έως και 100. Για τις μεθόδους NWA και W4M εφαρμόστηκαν ανωνυμοποιήσεις με δ=0,02% και δ=0,1% της συνολικής ακτίνας του συνόλου δεδομένων. Επιπλέον για την μέθοδο W4M πραγματοποιήθηκαν ανωνυμοποιήσεις και για τις δυο μεθόδους μέτρησης απόστασης τροχιών EDR και LST. Για την μέθοδο AWO εφαρμόστηκαν και στους δυο αλγορίθμους που διαθέτει η μέθοδος fastTGA και multiTGA. Και τέλος για την μέθοδο Generalization εφαρμόστηκαν για τους αλγορίθμους ανωνυμοποίησης KAM_CUT και KAM_REC.

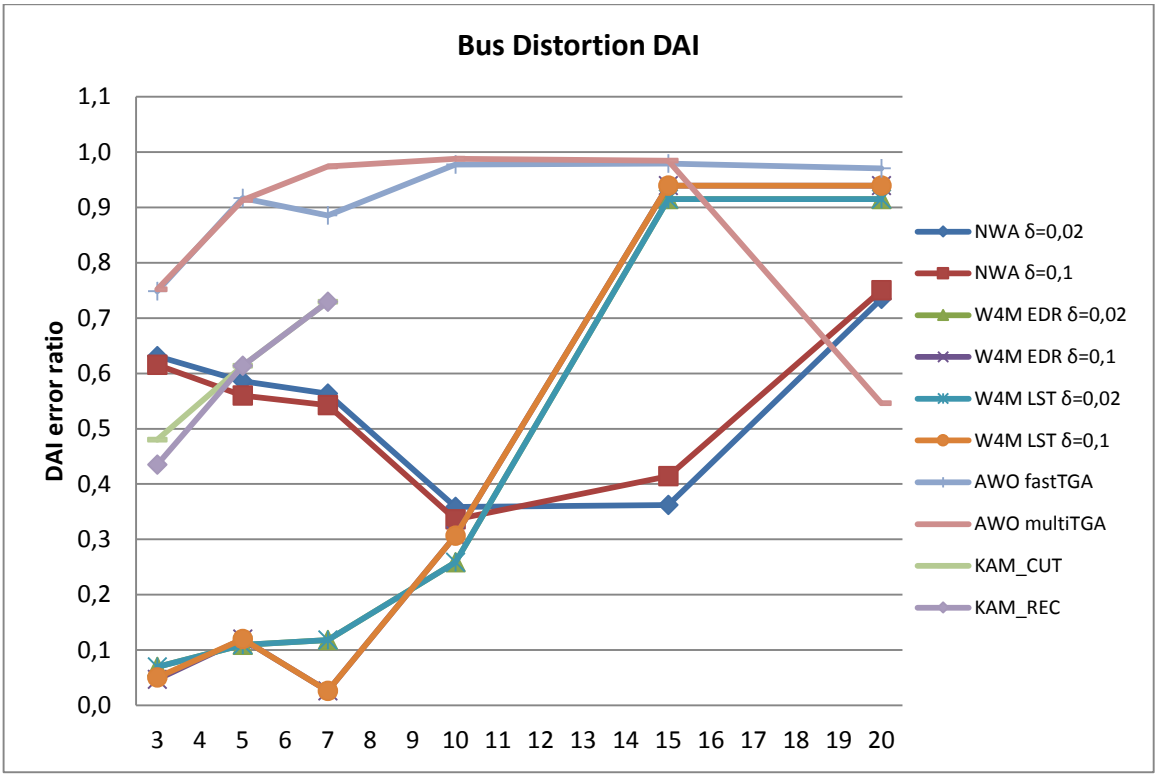
5.3.1. Distortion

Σε αυτό το κριτήριο σύγκρισης μεθόδων χρησιμοποιούνται τα ερωτήματα Definitely Always Inside (DAI) και η Possibly Sometime Inside (PSI) που περιγράφονται πιο αναλυτικά το κεφάλαιο 4. Βάση των οποίων έχουμε τα ερωτήματα Q1 και Q2, με το Q1 να ορίζετε ως «επέστρεψε μου το πλήθος των τροχιών για τα οποία ισχύει το DAI σε κάποιο ή κάποια από τα χωροχρονικά δείγματα που δημιουργήσαμε» και αντίστοιχα για το Q2 για το «επέστρεψε μου το πλήθος των τροχιών για τα οποία ισχύει το PSI σε κάποιο ή κάποια από τα χωροχρονικά δείγματα που δημιουργήσαμε». Τα χωροχρονικά δείγματα είναι ουσιαστικά τυχαίες κυλινδρικές χωροχρονικές περιοχές, με διακύμανση ακτίνας και χρονικής διάρκειας να ποικίλει αναλόγως το σύνολο δεδομένων στο οποίο χρησιμοποιήθηκαν και είναι χίλια στον αριθμό για κάθε σύνολο δεδομένων. Για παράδειγμα τα χωροχρονικά δείγματα για το σύνολο δεδομένων Bus είχαν μεγαλύτερη χρονική διάρκεια λόγω του ότι καλύπτει ολόκληρο χρόνο και έχει συνολικά τα λιγότερα δείγματα τροχιών. Για να γίνει λοιπόν η μέτρηση ανάμεσα στα αρχικά (D) και ανωνυμοποιημένα (D') αποτελέσματα χρησιμοποιήθηκαν τα ερωτήματα Q1 και Q2 ως εξής $|Q_1(D)-Q_1(D')|/Q_1(D)$ και αντίστοιχα για το Q_2. Τα αποτελέσματα των μετρήσεων αυτών ακολουθούν σε μορφή διαγράμματος.

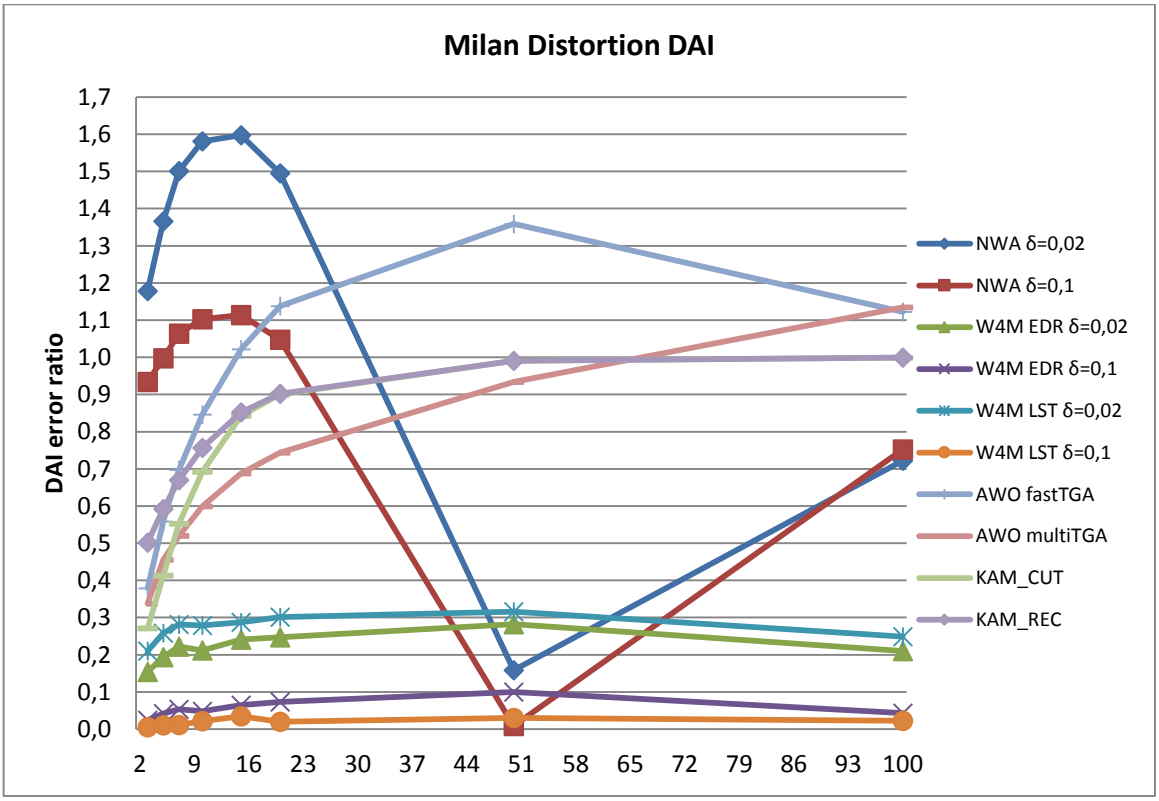
5.3.1.1 Definitely Always Inside



Γράφημα 1: DAI error ratio για το σύνολο δεδομένων Trucks



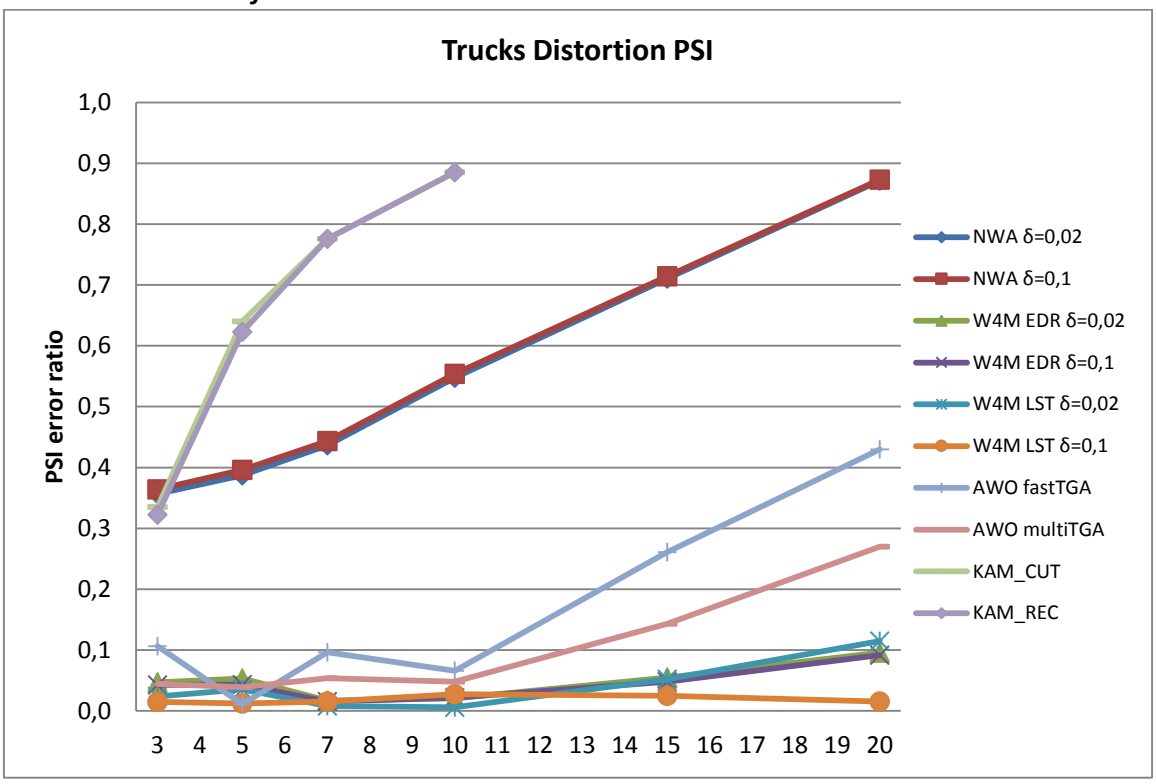
Γράφημα 2: DAI error ratio για το σύνολο δεδομένων Bus



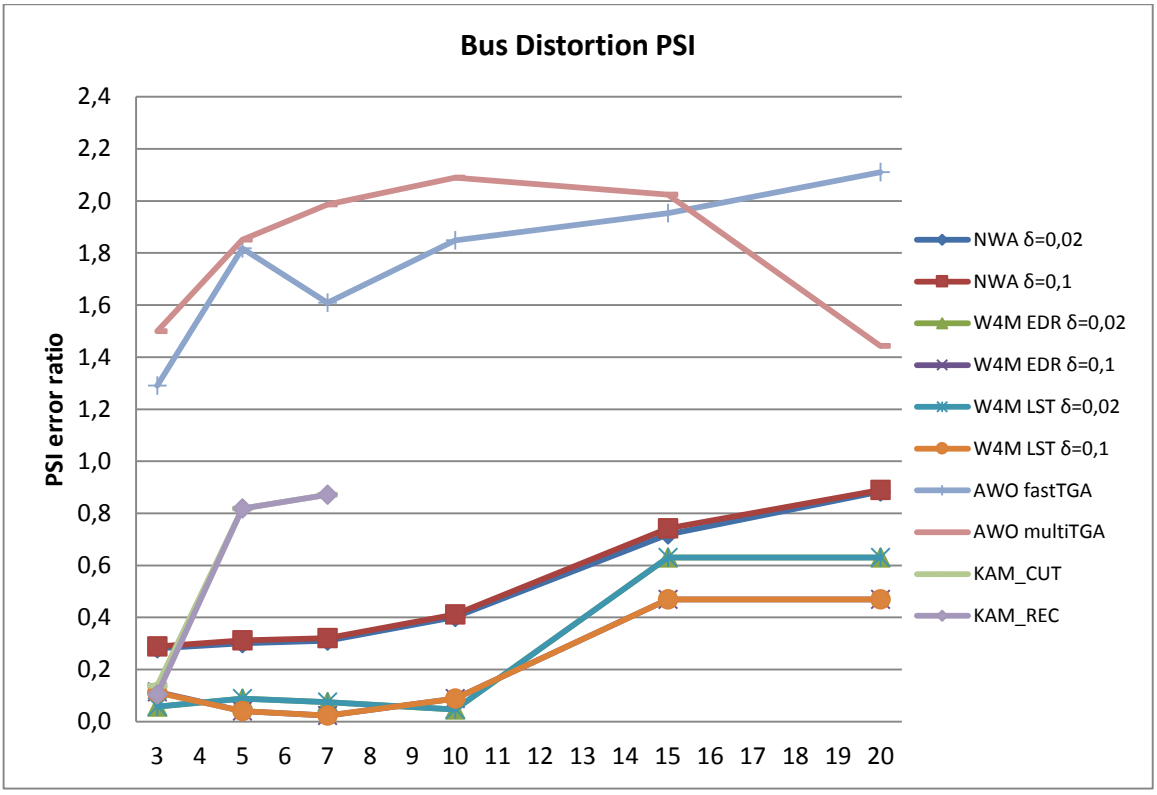
Γράφημα 3: DAI error ratio για το σύνολο δεδομένων Milan

Παρατηρούμε ότι για όλες τις περιπτώσεις η W4M παρουσιάζει τα καλύτερα αποτελέσματα ακόμα και στα υψηλά επίπεδα ανωνυμοποίησης εκτός από την περίπτωση του συνόλου δεδομένων Bus το οποίο έχει τα λιγότερα δείγματα τροχιών. Για την μέθοδο NWA βλέπουμε να παρουσιάζει υψηλό βαθμό error ratio στα μεγάλα επίπεδα ανωνυμοποίησης εκτός από την περίπτωση του συνόλου δεδομένων Milan όπου παρόλο το υψηλό error ratio δείχνει να παρουσιάζει μια απότομη βελτίωση. Αυτό οφείλεται ουσιαστικά στην σταδιακή μείωση του Q(D') για *k* από 20 έως 100 που ουσιαστικά γίνεται αλλαγή προσήμου στο Q(D)-Q(D'). Για τους δυο αλγορίθμους της μεθόδου AWO παρουσιάζουν υψηλό error ratio παρόλο που στα μικρά επίπεδα ανωνυμοποίησης κυμαίνεται μέτρια σχετικά επίπεδα error ratio. Για τους αλγορίθμους KAM_REC και KAM_CUT της μεθόδου Generalization παρατηρούμε μια συνεχή αύξηση του error ratio όσο ανεβαίνει το επίπεδο της ανωνυμοποίησης. Διαπιστώνουμε επίσης από τα γραφήματα ότι η μέθοδος Generalization αδυνατεί να πραγματοποιήσει ανωνυμοποιήσεις για υψηλά *k* σε σύνολα δεδομένων που δεν έχουν μεγάλο πλήθος δειγμάτων όπως αυτό του Trucks και του Bus παρά το υψηλό MaxRadius που τους ορίστηκε.

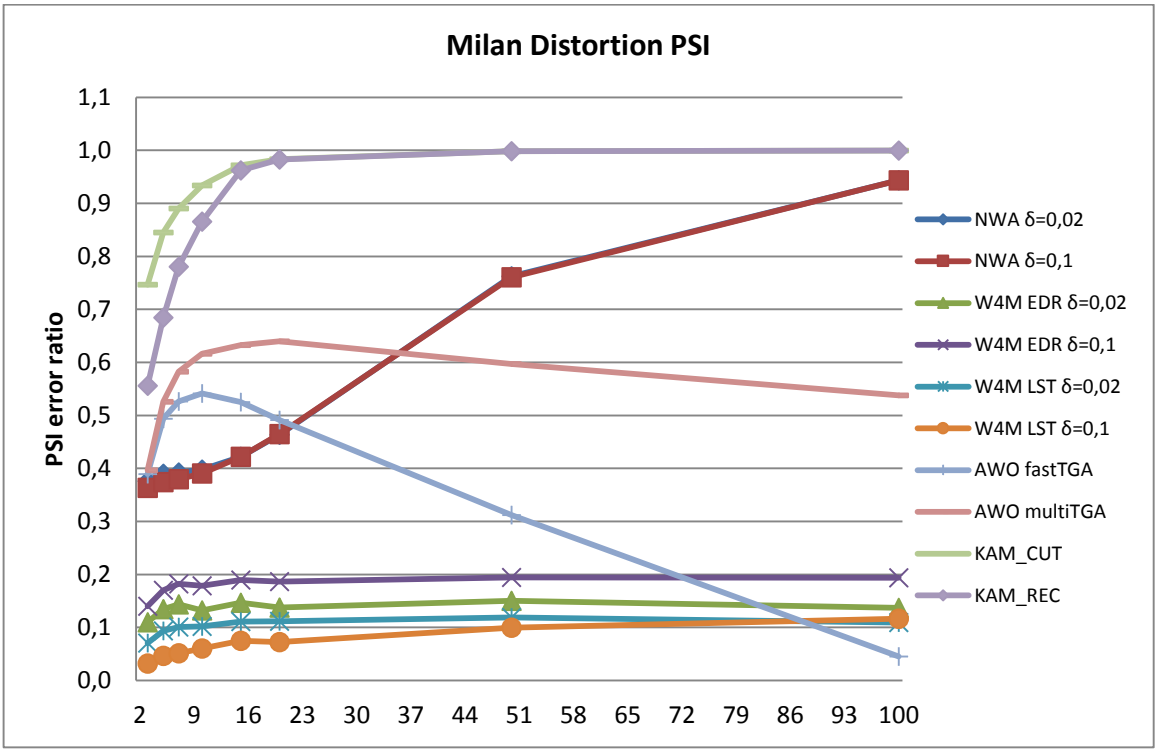
5.3.1.2 Possibly Sometime Inside



Γράφημα 4: PSI error ratio για το σύνολο δεδομένων Trucks



Γράφημα 5: PSI error ratio για το σύνολο δεδομένων Bus

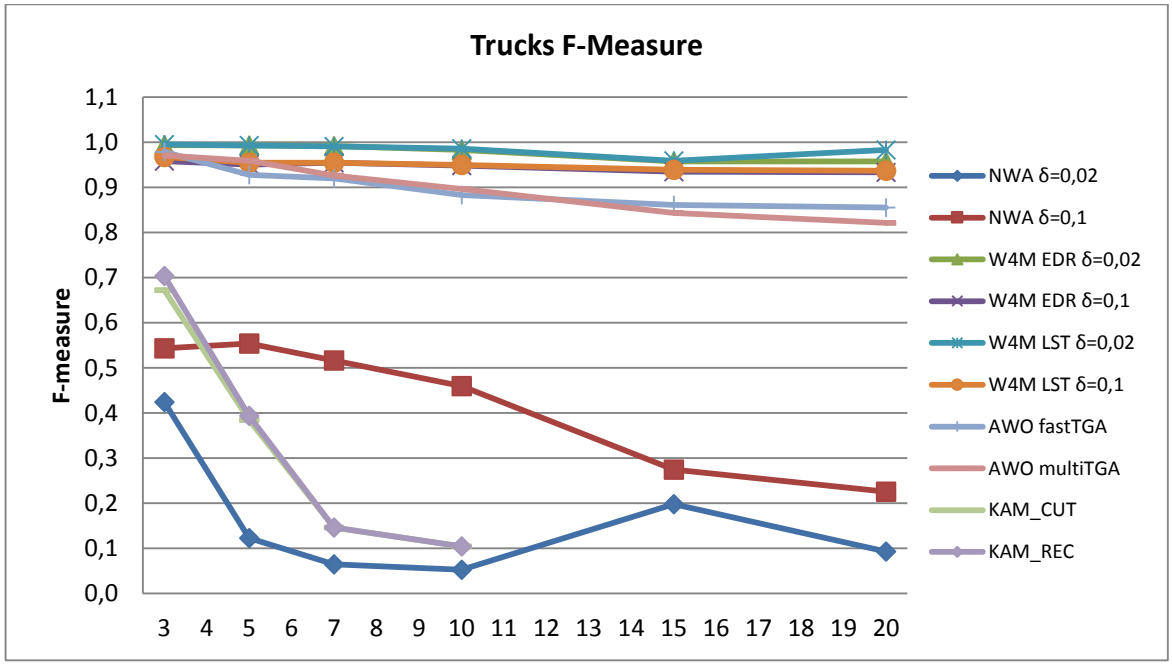


Γράφημα 6: PSI error ratio για το σύνολο δεδομένων Milan

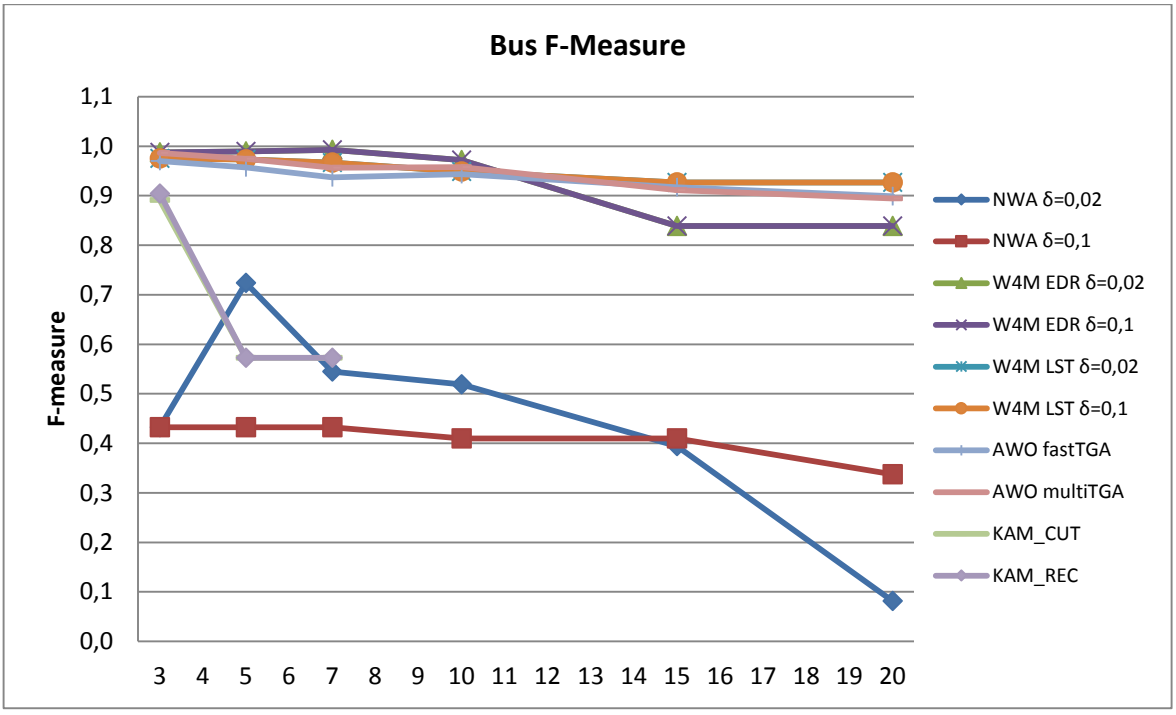
Παρατηρούμε για το PSI ότι για όλα τα σύνολα δεδομένων η μέθοδος W4M παρουσιάζει τα καλύτερα αποτελέσματα ακόμα και στα υψηλά επίπεδα ανωνυμοποίησης εκτός από την περίπτωση του συνόλου δεδομένων Bus όπως και στο DAI. Για την μέθοδο NWA βλέπουμε να έχει μέτριο error ratio και να αυξάνεται σταδιακά όσο το επίπεδο ανωνυμοποίησης ανεβαίνει. Το ίδιο συμβαίνει και για τους αλγορίθμους του Generalization που παρουσιάζουν μεγαλύτερο ρυθμό αύξησης του error ratio με τον αλγόριθμο KAM_REC να είναι ελαφρώς καλύτερος. Η μέθοδος AWO παρουσιάζει υψηλό error ratio στο σύνολο δεδομένων Bus το οποίο έχει τα λιγότερα δείγματα τροχιών, ενώ στα μεγάλα επίπεδα ανωνυμοποίησης για *k* παρουσιάζει σταδιακή βελτίωση.

5.3.2. F-measure

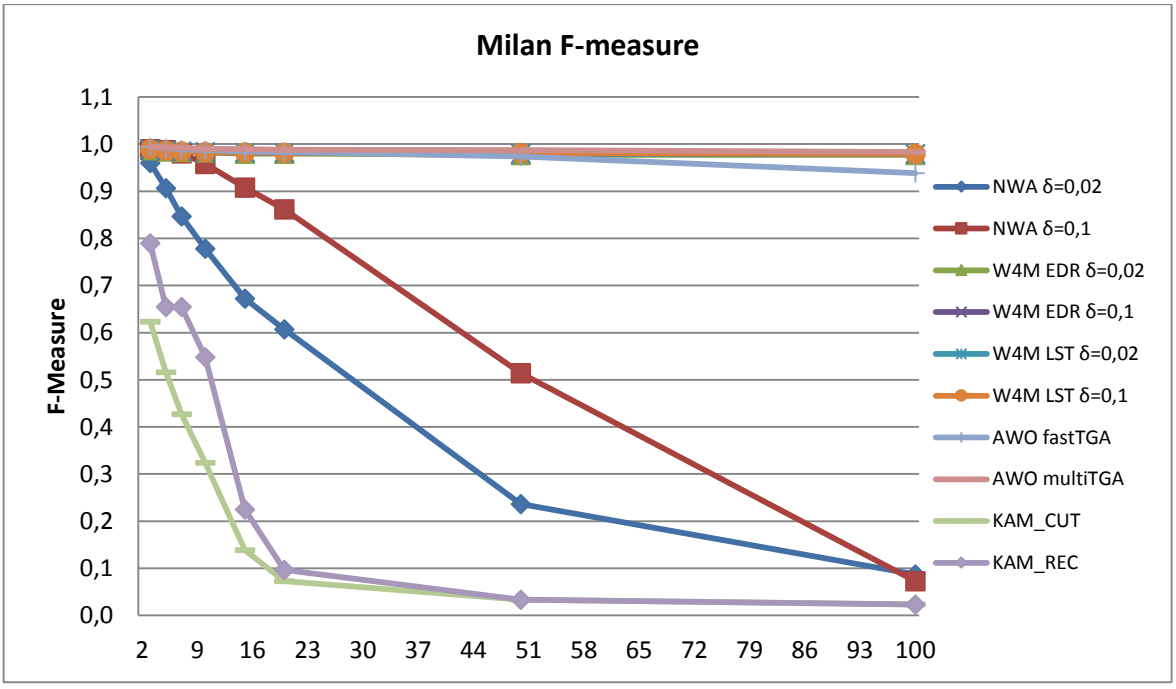
Για την εφαρμογή του κριτηρίου F-measure [11] πάνω στις αυθεντικές και ανωνυμοποιημένες τροχιές χρειάστηκε να γίνει η ακόλουθη προετοιμασία. Αρχικά χωρίστηκε η γεωγραφική περιοχή του συνόλου δεδομένων σε πίνακα διαστάσεων 25x25. Εν συνεχεία όλες οι τροχιές από τα ανωνυμοποιημένα και αυθεντικά σύνολα δεδομένων μεταφράζονται σε μια αλληλουχία συντεταγμένων πάνω στον πίνακα. Τέλος θεωρούνται θόρυβος και αφαιρούνται εκείνα τα σημεία των οποίων το πλήθος τους σε περιοχή του πίνακα είναι μικρότερο από το 1% επί του πλήθους των σημείων του συνόλου δεδομένων. Έτσι έχοντας προετοιμάσει τα δεδομένα μπορούμε να εφαρμόσουμε τον αρμονικό μέσο όρο της ακρίβειας (α) ως $|S'\cap S|/|S'|$ και της ανάκλησης (β) ως $|S'\cap S|/|S|$ όπου S τα αυθεντικά και S’ τα ανωνυμοποιημένα δεδομένα του συνόλου δεδομένων όπως αυτό αναφέρεται στην περιγραφή του μέτρου στην παράγραφο 3 που ορίζεται ως $2\alpha\beta/(\alpha+\beta)$.



Γράφημα 7: F-Measure για το σύνολο δεδομένων Trucks



Γράφημα 8: F-Measure για το σύνολο δεδομένων Bus

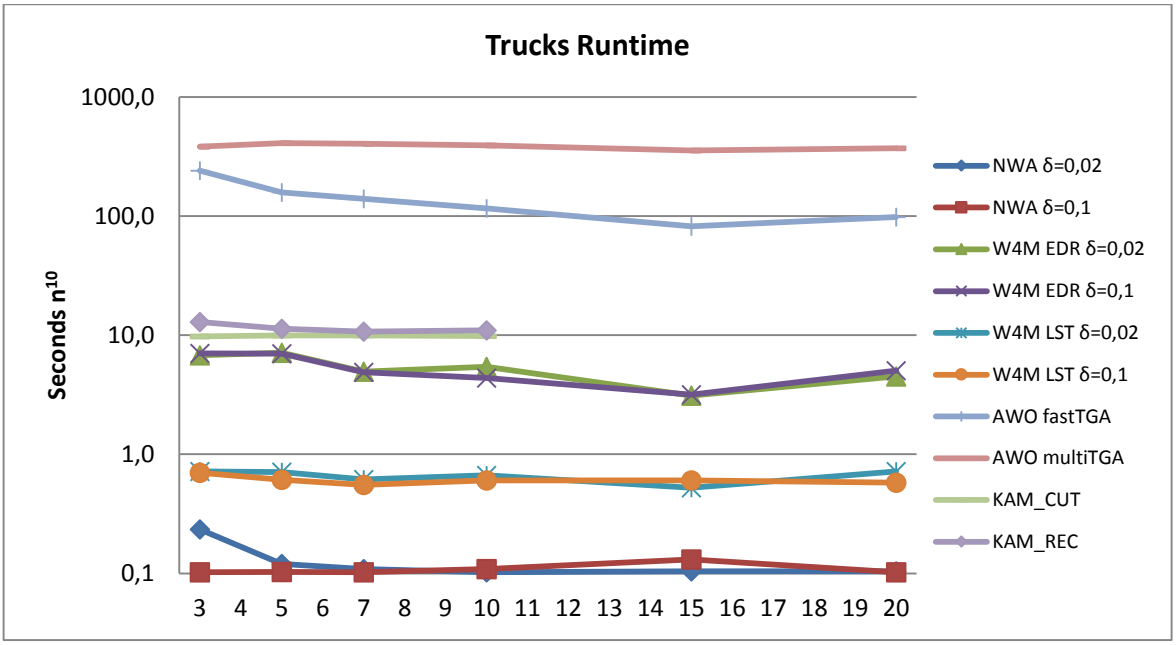


Γράφημα 9: F-Measure για το σύνολο δεδομένων Milan

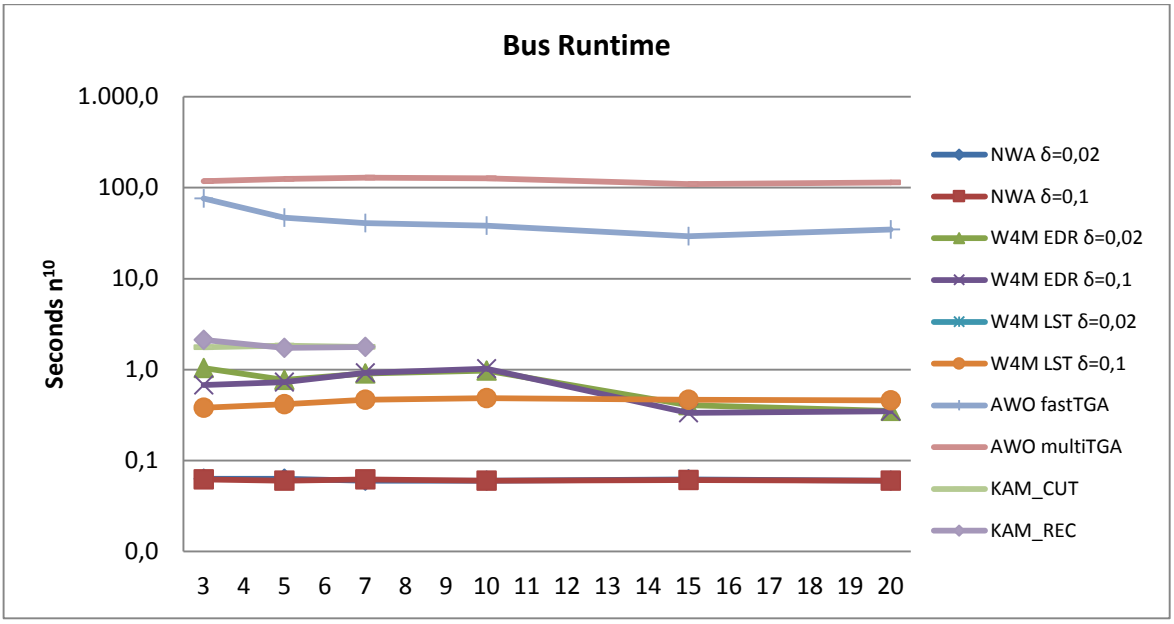
Παρατηρούμε ότι στον αρμονικό μέσο όρο του κριτηρίου F-measure τα καλύτερα αποτελέσματα μας τα δίνει η μέθοδος W4M με όλες τις παραλλαγές της. Αμέσως μετά και πολύ κοντά στην W4M είναι οι αλγόριθμοι της μεθόδου AWO που παρουσιάζουν μια ελαφριά φθορά μόνο στα υψηλά επίπεδα της ανωνυμοποίησης. Μετά ακολουθούν οι παραλλαγές του NWA και οι δυο αλγόριθμοι του Generalization οι οποίοι παρουσιάζουν μεγάλη φθορά όσο ανεβαίνει η ανωνυμία.

5.3.3. Runtime

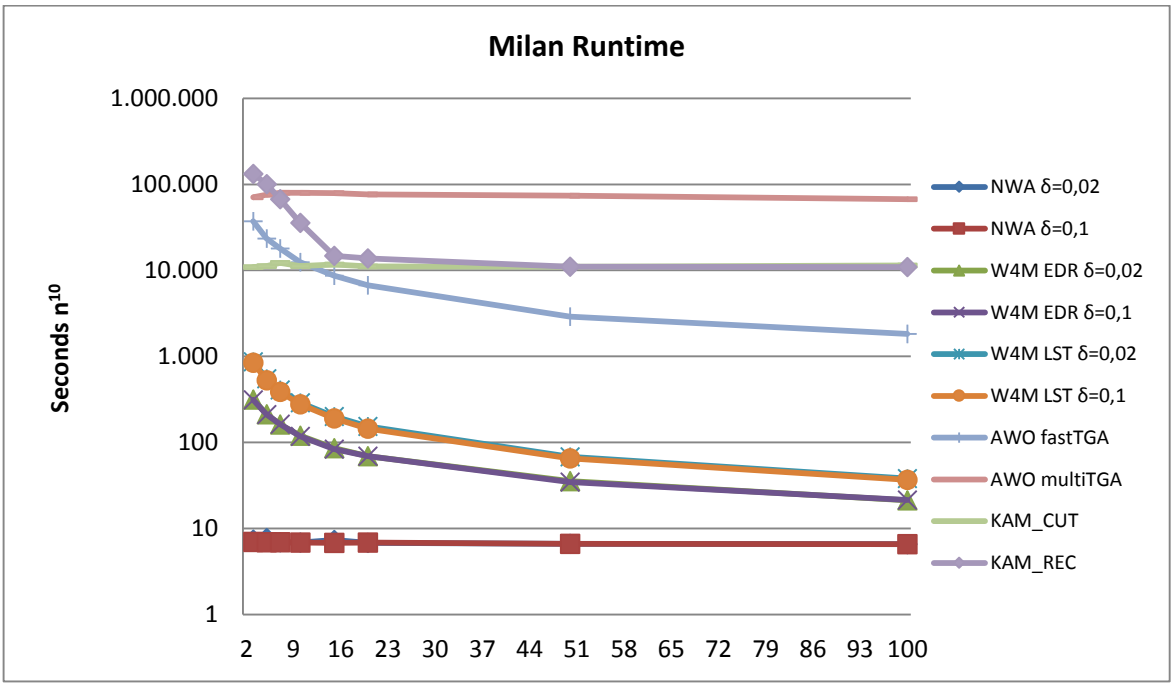
Σε αυτό το κριτήριο μετράμε ουσιαστικά τον χρόνο εκτέλεσης της ανωνυμοποίησης για κάθε μέθοδο και σε κάθε επίπεδο ανωνυμοποίησης *k* χρησιμοποιώντας τους ίδιους πόρους συστήματος που αναφέρθηκαν στην παράγραφο 6. Σημειώνεται επίσης ότι για την μέθοδο Generalization ο χρόνος απλοποίησης των τροχιών δεν έχει προστεθεί στον συνολικό χρόνο ανωνυμοποίησης. Στην συνέχεα παρουσιάζονται τα αποτελέσματα για το Runtime κριτήριο όπου ο άξονας του χρόνου εμφανίζεται σε λογαριθμική κλίμακα του 10 λόγω της μεγάλης διαφοράς που υπάρχει από μέθοδο σε μέθοδο.



Γράφημα 10: Runtime για το σύνολο δεδομένων Trucks



Γράφημα 11: Runtime για το σύνολο δεδομένων Bus



Γράφημα 12: Runtime για το σύνολο δεδομένων Milan

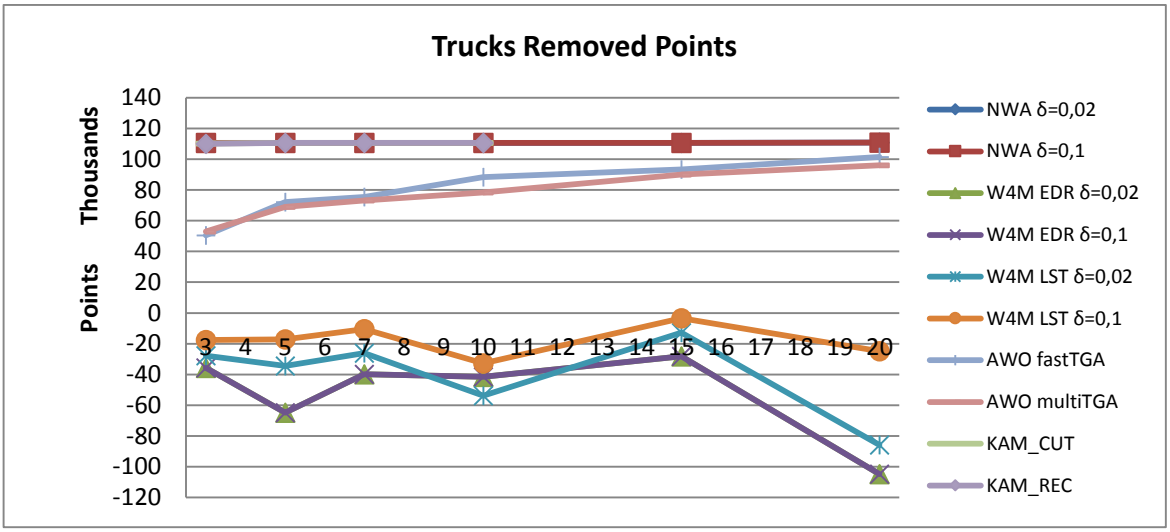
Παρατηρούμε ότι τον υψηλότερο χρόνο εκτέλεσης της ανωνυμοποίησης εμφανίζει η μέθοδος NWA και στα τρία σύνολα δεδομένων με διαφορά από τις υπόλοιπες τρεις μεθόδους. Εν συνεχεία ακολουθούν οι παραλλαγές του W4M παρουσιάζοντας σταδιακή μείωση του χρόνου ανωνυμοποίησης όσο αυξάνεται το *k* για το μεγάλο σύνολο δεδομένων του Μιλάνου. Μετά ακολουθούν οι αλγόριθμοι των μεθόδων AWO και Generalization με πολύ υψηλό απαιτούμενο χρόνο ανωνυμοποίησης. Αξίζει να σημειωθεί όμως η βελτίωση του απαιτούμενου χρόνου του αλγορίθμου fastTGA ως προς τον multiTGA για τον AWO και του

αλγορίθμου KAM_REC ως προς τον KAM_CAT για τον Generalization όσο ανεβαίνει το επίπεδο ανωνυμοποίησης.

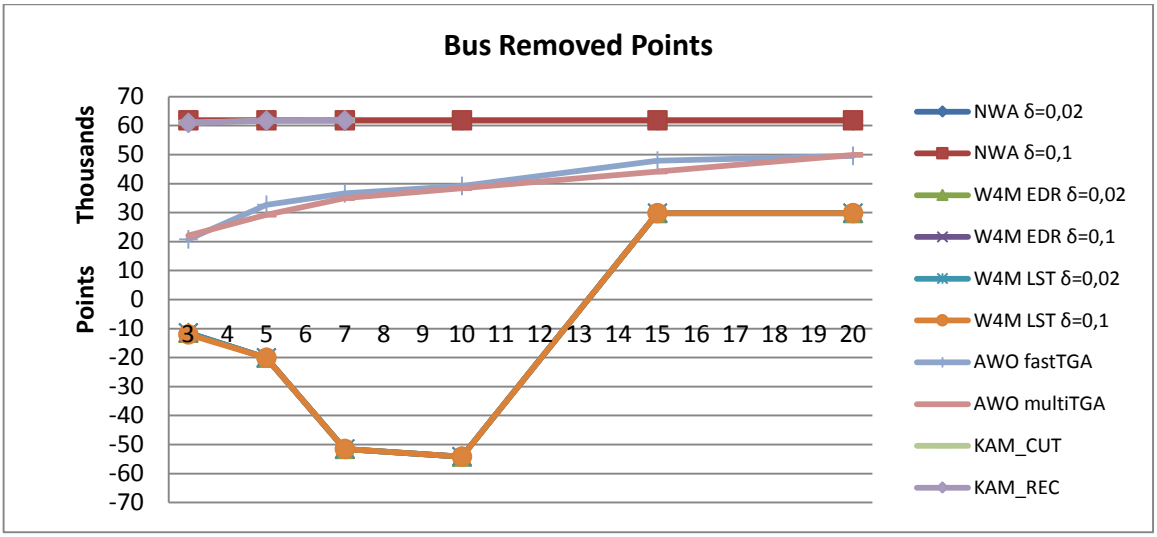
5.3.4. Removed Points

Στο κριτήριο αυτό τα αποτελέσματα προκύπτουν από την διαφορά που υπάρχει ανάμεσα στο σύνολο των σημείων των ανωνυμοποιημένων και αυθεντικών συνόλων δεδομένων ανά μέθοδο και ανά επίπεδο ανωνυμοποίησης *k*. Σημειώνεται ότι το αρνητικό πρόσημο στα Removed Points της W4M είναι λόγω της δημιουργίας σημείων όπως αυτό περιγράφεται στην παράγραφο 2.7.

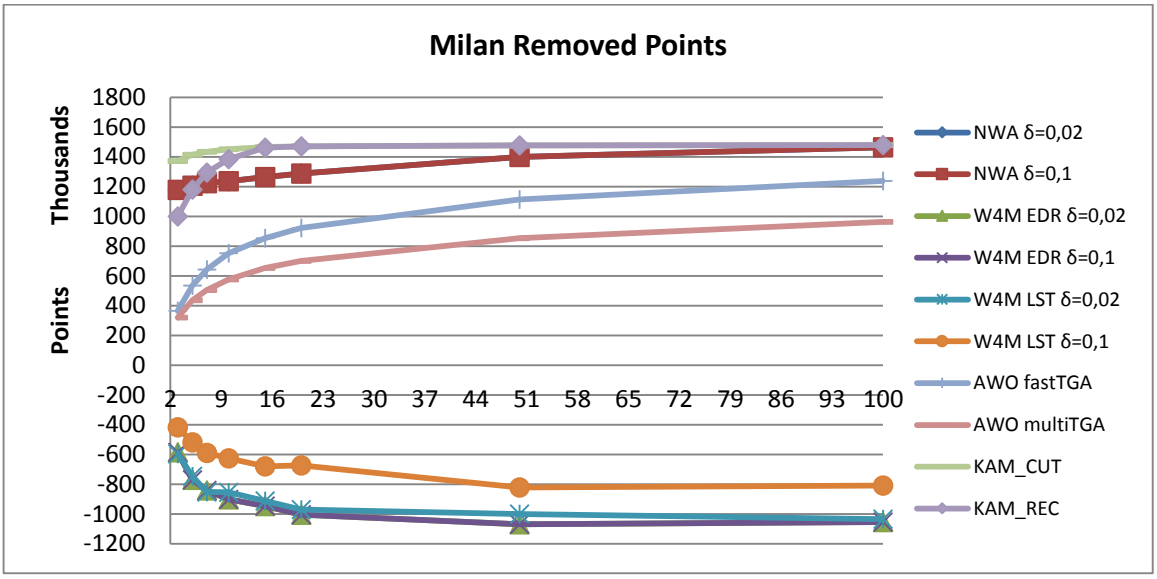
Παρατηρούμε από τα γραφήματα που ακολουθούν ότι στην μέθοδο W4M δημιουργούνται περισσότερα σημεία από όσα διαγράφονται εκτός από την περίπτωση του συνόλου δεδομένων Bus με τις λιγότερες τροχιών όπου στα υψηλά επίπεδα ανωνυμοποίησης ισχύει το αντίστροφο. Οι αλγόριθμοι του Generalization μαζί με τις παραλλαγές του NWA παρουσιάζουν τον υψηλότερο αριθμό διαγεγραμμένων σημείων για όλα τα επίπεδα ανωνυμίας σε όλα τα σύνολα δεδομένων παρουσιάζοντας τις πιο απλοποιημένες ανωνυμοποιημένες τροχιές. Και τέλος οι αλγόριθμοι της μεθόδου AWO παρουσιάζουν υψηλό αριθμό διαγραφών, με τον αριθμό αυτό να αυξάνεται όσο ανεβαίνει το επίπεδο ανωνυμοποίησης.



Γράφημα 13: Deleted Points για το σύνολο δεδομένων Trucks



Γράφημα 14: Deleted Points για το σύνολο δεδομένων Bus



Γράφημα 15: Deleted Points για το σύνολο δεδομένων Milan

6. Συμπεράσματα

Με την ανάγκη που υπάρχει σήμερα για την προστασία της ιδιωτικής πληροφορίας από κακόβουλη χρήση και τη συνεχή ανάπτυξη και εξέλιξη νέων μεθόδων ανωνυμοποίησης σε δεδομένα κίνησης που υπάρχει τα τελευταία χρόνια, ήταν απαραίτητη η πραγματοποίηση μιας τέτοιας εργασίας για την σύγκριση των μεθόδων αυτών πάνω σε κοινά κριτήρια. Με την ολοκλήρωση της εργασίας προέκυψαν κάποια συμπεράσματα των οποίων πρέπει να τονιστεί η σημασία τους. Η ανάλυση αυτών των συμπερασμάτων παρουσιάζεται στην συνέχεα αναλύοντας ξεχωριστά για κάθε μέθοδο τι μας προσέφερε από την σύγκριση της με τις υπόλοιπες μεθόδους.

Για την μέθοδο ανωνυμοποίησης W4M και για τους δυο τρόπους μέτρησης απόστασης τροχιές EDR και LST, παρατηρώντας και μόνο τα γραφήματα βλέπουμε ότι είναι ανώτερη σε σχέση με τις υπόλοιπες παρουσιάζοντας χαμηλή παραμόρφωση στα δεδομένα και μικρό χρόνο ολοκλήρωσης της ανωνυμοποίησης. Θα μπορούσαμε λοιπόν να πούμε ότι είναι μια μέθοδος γενικής χρήσης, παρόλα αυτά όμως υπάρχουν κάποια χαρακτηριστικά στις άλλες μεθόδους που θα μας κάνουν να τις προτιμήσουμε για κάποιες ειδικές περιπτώσεις.

Τους πιο χαμηλούς χρόνους εκτέλεσης της ανωνυμοποίησης τους παρουσιάζει η μέθοδος NWA. Ακόμα και για το σύνολο δεδομένων του Μιλάνου ο χρόνος εκτέλεσης της ανωνυμοποίησης διήρκησε μόλις μερικά δευτερόλεπτα σε όλα τα επίπεδα ανωνυμοποίησης. Δεδομένου του χαμηλού χρόνου ανωνυμοποίησης μαζί με τον υψηλό βαθμό απλοποίησης των ανωνυμοποιημένων τροχιών, καθιστά αυτή την μέθοδο ιδανική για εφαρμογές που απαιτείται ανωνυμοποίηση δεδομένων πραγματικού χρόνου. Μια τέτοια εφαρμογή θα μπορούσε να επιστρέφει την τάση των οδηγών την τελευταία μια ώρα στο κέντρο της Αθήνας. Σε αυτή την περίπτωση η ευκλείδεια μέτρηση απόστασης της NWA δεν αποτελεί πρόβλημα αφού όλα τα δεδομένα θα έχουν κοινή χρονική αρχή και τέλος.

Οι μέθοδοι AWO και Generalization είναι οι μέθοδοι που απαιτούν τους υψηλότερους χρόνους ανωνυμοποίησης. Παρόλα αυτά θα μπορούσαμε να πούμε για την περίπτωση της μεθόδου Generalization ότι αν δεν μας ενδιαφέρει ο χρόνος ανωνυμοποίησης μπορεί να μας παρουσιάζει τις πιο απλοποιημένες τροχιές σε σύνολα δεδομένων πολύ μεγάλου όγκου. Και παρότι η NWA παρουσιάζει μικρότερη παραμόρφωση στα δεδομένα, είναι η ευκλείδεια μέτρηση απόστασης των τροχιών που θέλουμε να αποφύγουμε.

Τέλος για την μέθοδο AWO παρά τον πολύ μεγάλο χρόνο ανωνυμοποίησης που απαιτεί παρουσιάζει πολύ χαμηλή παραμόρφωση στα ανωνυμοποιημένα δεδομένα που συγκρίνεται με αυτή της μεθόδου W4M. Επίσης λόγω της σταδιακής μείωσης των απαιτούμενων σημείων των ανωνυμοποιημένων τροχιών όσο το επίπεδο ανωνυμοποίησης *k* ανεβαίνει, θα μπορούσε να χρησιμοποιηθεί αντί της μεθόδου Generalization σε μικρότερα σύνολα δεδομένων. Άλλος ένα λόγος που μας ωθεί στην αντικατάσταση της μεθόδου Generalization είναι η αδυναμία της να πραγματοποιήσει υψηλού επίπεδου ανωνυμοποίησης *k* σε μικρά σύνολα δεδομένων όπως είναι τα Trucks και Bus.

7. Βιβλιογραφικές αναφορές

[1] Abul, O., Bonchi, F., and Nanni, M. Never walk alone: Uncertainty for anonymity in moving objects databases. In Proceedings of ICDE, pages 376-385, 2008.

[2] Abul, O., Bonchi, F., and Nanni, M. Anonymization of moving objects databases by clustering and perturbation. Information Systems, 35(8):884-910, 2010.

[3] Nergiz, M. E., Atzori, M. and Saygin, Y. Towards trajectory anonymization: A generalization-based approach. In ACM GIS Workshop on Security and Privacy in GIS and LBS, pages 1-10, 2008.

[4] Monreale, A., Andrienko, G., Andrienko, N., Giannotti, F., Pedreschi, D., Rinzivillo, S. Wrobel, S. (2010). Movement Data Anonymity through Generalization. Transactions on Data Privacy 3(2), 91-12

[5] Sweeney, Latanya. "*k*-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.

[6] Machanavajjhala, Ashwin, et al. "*l*-diversity: Privacy beyond *k*-anonymity." ACM Transactions on Knowledge Discovery from Data (TKDD) 1.1 (2007): 3.

[7] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "*t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity." Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.

[8] Hoh, B. and Gruteser, M.. Protecting location privacy through path confusion. In SECURECOMM, pages 194-205, 2005.

[9] Terrovitis M, Mamoulis N (2008) Privacy preservation in the publication of trajectories. In MDM, pp.65-72

[10] Mahdavifar, Samaneh, et al. "A clustering-based approach for personalized privacy preserving publication of moving object trajectory data." Network and System Security. Springer Berlin Heidelberg, 2012. 149-165.

[11] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. Modern Information Retrieval. ACM Press / Addison-Wesley, 1999.

[12] S. Stieniger, M. Neun, A. Edwardes, “Foundations of location based services,” Project CartouCHe - Lecture Notes on LBS, version 1.0

[13] G. Trajcevski, O. Wolfson, K. Hinrichs, S. Chamberlain, Managing uncertainty in moving objects databases, ACM Transactions on Database Systems 29 (3) (2004) 463–507

[14] Natalia Andrienko and Gennady Andrienko, Spatial Generalization and Aggregation of Massive Movement Data, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 16, NO. X, XXX/XXX 2010