ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΡΟΩΝ ΣΥΝΘΕΤΩΝ ΔΕΔΟΜΕΝΩΝ

Βασίλειος Ν. Κοντονής

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιανουάριος 2017

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΡΟΩΝ ΣΥΝΘΕΤΩΝ ΔΕΔΟΜΕΝΩΝ

Βασίλειος Ν. Κοντονής

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιανουάριος 2017

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Νικόλαος Σ. Πελέκης (Επιβλέπων)
- …….
- …….

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

CLUSTERING COMPLEX DATA STREAMS

By

Vasileios N. Kontonis

MSc Dissertation

Submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
January 2017

Στους γονείς μου
Νίκο και Αγγελική

Ευχαριστίες

Για τη διεκπεραίωση της παρούσας Διπλωματικής Εργασίας, θα ήθελα να ευχαριστήσω την οικογένειά μου για την αμέριστη συμπαράσταση και υποστήριξη τους. Επίσης, θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου, Πελέκη Νικόλαο, για τη συνεργασία και την πολύτιμη συμβολή του στην ολοκλήρωση της.

Περίληψη

Η συσταδοποίηση ροών δεδομένων (data stream clustering) αποτελεί μία ερευνητική περιοχή η οποία συνεχώς αναπτύσσεται μέσα από διάφορες πτυχές (Εξόρυξη Μεγάλου Όγκου δεδομένων- Big Data, αλγόριθμοι ομαδοποίησης, αλγόριθμοι ανάλυσης, κλπ). Κοινός τους παρονομαστής είναι το ζητούμενο της επεξεργασίας τεράστιου όγκου δεδομένων σε διάφορους τομείς. Η παρούσα εργασία προσπάθησε να αξιολογήσει τις δυνατότητες που παρέχουν σύγχρονες βιβλιοθήκες (MOA) και εργαλεία (γλώσσα προγραμματισμού R) στην εφαρμογή αλγορίθμων συσταδοποίησης ροών μεγάλου όγκου δεδομένων με αυξημένη πολυπλοκότητα. Με δεδομένα της κυκλοφορίας των ταξί στο Πεκίνο, εξετάστηκε η υλοποίηση συγκεκριμένης αρχιτεκτονικής και εφαρμογής ως σημαντική συνεισφορά της παρούσας εργασίας.

Τα αποτελέσματα εφαρμογής των παραπάνω στο συγκεκριμένο παράδειγμα δεδομένων ανέδειξε την προοπτική που προσφέρει σε ερευνητικό και εμπορικό επίπεδο η αξιοποίηση της βιβλιοθήκης MOA μέσω ενός προγραμματιστικού περιβάλλοντος όπως αυτό της R. Δοκιμάστηκε και αξιολογήθηκε ο αλγόριθμος συσταδοποίησης βάσει πυκνότητας DenStream σε σύγκριση με τον αλγόριθμο CluStream. Ο αλγόριθμος DenStream απέδωσε καλύτερα αποτελέσματα σε σύγκριση με τον CluStream με βάση συγκεκριμένες μετρικές αξιολόγησης. Η αρχικοποίηση των παραμέτρων του εκάστοτε αλγορίθμου προέκυψε έπειτα από πραγματοποίηση δοκιμών. Τέλος, η εργασία εξετάζει αν η προτεινόμενη αρχιτεκτονική μπορεί να αποτελέσει σημείο αναφοράς ή υπο-σύνολο ενός συστήματος στατιστικής ανάλυσης δεδομένων, το οποίο σε πραγματικό χρόνο θα μπορεί να επεξεργάζεται ροές δεδομένων, να τις ταξινομεί σε συμπλέγματα/ομαδοποιήσεις (π.χ. γεωγραφικές περιοχές) και να τις αξιολογεί.

Abstract

Streaming data clustering is a continuous growing research area through various aspects such as big data mining, clustering and analysis algorithms and more, where the need for high volume data processing is required. This thesis's purpose is to evaluate the capabilities provided by modern and widely known libraries (MOA) and tools (R programming language) to apply clustering algorithms on streams of high volume of complex data. Having available a dataset with a big amount of records regarding the trajectories of many taxis in Beijing, the important contribution of this project is the implementation of this particular architecture and application.

The result of the above mentioned application using the specific dataset that is available has revealed the prospect that is provided to the research and the commercial sector by utilizing the MOA library through the R programming environment. DenStream, a density based clustering algorithm, has been applied, evaluated and also compared to the CluStream algorithm, which has also been applied and evaluated as part of this thesis. DenStream algorithm has given better results than the CluStream algorithm in terms of clustering quality, based on specific evaluation metrics used. Furthermore, it turns out that the initialization of the parameters of each algorithm is empirical and it is defined by the analyst based on expertise and the domain knowledge. Eventually, in this thesis it is considered if the suggested architecture could be a point of reference or a subset of a statistical analysis system, which could process data streams, group them into clusters and evaluate the results in real time.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1 **1**

1. Εισαγωγή **1**

Κ Ε Φ Α Λ Α Ι Ο 2 **4**

2. Ανασκόπηση βιβλιογραφίας **4**

2.1 Συσταδοποίηση ροών δεδομένων (data stream clustering) **4**

2.1.1 Συνιστώσες συσταδοποίησης ροών δεδομένων **5**

2.2 Τεχνικές χωρο-χρονικής συσταδοποίησης ροών δεδομένων **8**

2.2.1 Παραδοσιακές τεχνικές συσταδοποίησης δεδομένων **8**

2.2.2 Συσταδοποίηση βάσει πυκνότητας - DBSCAN **12**

2.3 R - MOA **19**

2.3.1 R - Stream **20**

2.3.2 DenStream **22**

2.3.3 CluStream **25**

2.3.4 ClusTree **26**

2.3.5 Μετρικές απόστασης **27**

2.4 Ροές δεδομένων κινούμενων οχημάτων **28**

Κ Ε Φ Α Λ Α Ι Ο 3 **30**

3. Εφαρμογή αλγορίθμων συσταδοποίησης βάσει πυκνότητας **30**

3.1 Αρχιτεκτονική **30**

3.2 Βάση δεδομένων **32**

3.2.1 Ροή δεδομένων **32**

3.3 Εφαρμογή R **34**

3.3.1 Βιβλιοθήκη Stream - MOA **35**

3.3.2 Υλοποίηση εφαρμογής **35**

3.3.2.1 Αρχικοποίηση Παραμέτρων **39**

3.3.2.2 Αρχείο Συναρτήσεων **41**

3.3.2.3 Ροή Εκτέλεσης **44**

3.4 Προβολή δεδομένων και σημείων ομαδοποίησης στο χάρτη **46**

Κ Ε Φ Α Λ Α Ι Ο 4 **48**

4. Εφαρμογή αλγορίθμων συσταδοποίησης βάσει πυκνότητας **48**

4.1 Μετρικές αξιολόγησης αποτελεσμάτων **49**

Κ Ε Φ Α Λ Α Ι Ο 5 **58**

5. Συμπεράσματα **58**

ΒΙΒΛΙΟΓΡΑΦΙΑ **60**

ΚΕΦΑΛΑΙΟ 1

1. Εισαγωγή

Η συσταδοποίηση ροών δεδομένων (data stream clustering) είναι μια διαρκώς επεκτεινόμενη ερευνητική περιοχή καθώς καλείται να αντιμετωπίσει την πρόκληση της επεξεργασίας τεράστιου όγκου δεδομένων σε διάφορους τομείς. Οι τομείς αυτοί μπορεί να είναι δεδομένα αναζήτησης στο διαδίκτυο, δεδομένα βιομηχανικής παραγωγής, συστήματα λήψης και επεξεργασίας δεδομένων από πολλές πηγές σε πραγματικό χρόνο (π.χ. συστήματα κυβερνο-ασφάλειας), τραπεζο-οικονομικά ή άλλα συστήματα που επεξεργάζονται συναλλαγές σε πραγματικό χρόνο (π.χ. χρηματιστηριακές συναλλαγές, παρακολούθηση ασθενών, δεδομένα που λαμβάνονται από δίκτυα οργανισμών, δεδομένα κυκλοφοριακής κίνησης – GPS) (Silva, Jonathan A. and Faria, Elaine R. and Barros, Rodrigo C. and Hruschka, Eduardo R. and Carvalho, 2013).

Το πρόβλημα που καλούνται να αντιμετωπίσουν οι σύγχρονες τεχνικές συσταδοποίησης ροών δεδομένων στους παραπάνω τομείς είναι η αναγνώριση κάποιου προβλήματος που συμβαίνει τη δεδομένη χρονική στιγμή (π.χ. απώλεια δικτύου, πτώση παραγωγής ρεύματος, αυξημένη κίνηση σε ορισμένες περιοχές μίας πρωτεύουσας-πόλης, κ.α), η εξαγωγή και ο ορισμός μοτίβων μέσα στα δεδομένα που θα επιτρέψει την αποτελεσματικότερη λήψη αποφάσεων έχοντας αυτά τα μοτίβα (patterns) ως οδηγό (Mansalis S., Ntoutsi E., Pelekis N., Theodoridis Y., 2016). Πέρα από τις παραπάνω προκλήσεις σε επιχειρησιακό επίπεδο, όποια τεχνική συσταδοποίησης ροών δεδομένων επιλεχθεί ή υλοποιηθεί πρέπει επίσης να επιλύσει προβλήματα σε υπολογιστικό επίπεδο όπως η συνεχής ροή δεδομένων και σε μεγάλο όγκο καθιστούν αδύνατη την αποθήκευση τους στη μνήμη ή τον δίσκο του σταθμού εργασίας. Αυτό έχει ως επακόλουθο ότι ο αλγόριθμος της συσταδοποίησης θα πρέπει να προσπελάσει την εκάστοτε πληροφορία που εισέρχεται μόνο μία φορά ή όσες φορές του επιτρέπει η διαθέσιμη μνήμη ή δίσκος του υπολογιστικού σταθμού. Επίσης η ροή των πληροφοριών δεν ακολουθεί μία συγκεκριμένη λογική σειρά που επιτρέπει την αντίστοιχη σύνθεση και επεξεργασία τους από τον αλγόριθμο.

Οι παραπάνω περιορισμοί λαμβάνονται υπόψη από ένα ερευνητή στον σχεδιασμό ενός αλγορίθμου συσταδοποίησης ροών δεδομένων είτε αυτός ανήκει στην τάξη της συσταδοποίησης βάσει διαχωρισμού (partitioning-based), βάσει πυκνότητας (density-based) βάσει της κατανομής τους σε πλέγμα (grid-based) ή βάσει ενός μοντέλου (model-based). Σε όλες τις περιπτώσεις καλούνται να επιλέξουν τα κατάλληλα λογισμικά και εργαλεία που θα τους επιτρέψουν να βελτιστοποιήσουν τις κύριες συνιστώσες ενός τέτοιου αλγορίθμου. Για παράδειγμα, ο αλγόριθμος θα επεξεργαστεί τα δεδομένα μόνο μία φορά (single-pass) ή θα μπορέσει να τα επεξεργαστεί περισσότερες φορές αξιοποιώντας τη μνήμη του υπολογιστικού πόρου. Θα πρέπει επίσης να επιλέξει τη κατάλληλη τεχνική για τη ρύθμιση του παράθυρου χρόνου (time window) μέσα στο οποίο θα επεξεργάζεται υποσύνολο των δεδομένων που λαμβάνει, καθώς επίσης την κατάλληλη τεχνική για τη συγκέντρωση ενδιάμεσης γνώσης κατά την επεξεργασία (summarization) και την αξιοποίηση της σε μεταγενέστερο στάδιο για

τη δημιουργία των ομάδων δεδομένων (Mansalis S., Ntoutsi E., Pelekis N., Theodoridis Y., 2016).

Αντικείμενο αυτής της διπλωματικής εργασίας είναι η συσταδοποίηση ροών δεδομένων με συσταδοποίηση ροών μεγάλου όγκου δεδομένων (π.χ. εκατομμύρια εγγραφών τα οποία παράγονται τυχαία και ο αριθμός των πραγματικών ομάδων δεδομένων είναι άγνωστος) αξιοποιώντας σύγχρονα εργαλεία και τεχνικές όπως η γλώσσα επίλυσης προβλημάτων εξόρυξης δεδομένων και στατιστικής ανάλυσης R (Forrest J., 2011) και η βιβλιοθήκη MOA (Massive Online Analysis) που επιτρέπει την δημιουργία, βαθμολόγηση και αξιολόγηση μοντέλων με ροές δεδομένων εφαρμόζοντας συγκεκριμένους αλγορίθμους (π.χ. DBSCAN, DenStream, CluStream) (MOA, 2016). Εκτός από τη δημιουργία μοντέλων κατηγοριοποίησης και συσταδοποίησης δεδομένων σε συμπλέγματα υποστηρίζει επίσης τη δημιουργία συστημάτων πρόβλεψης για ροές δεδομένων (Forrest J., 2011).

Λαμβάνοντας υπόψη τα παραπάνω εργαλεία και βιβλιοθήκες αλγορίθμων, σκοπός της εργασίας είναι η επεξεργασία χώρο-χρονικών ροών δεδομένων (spatio-temporal streaming data) με σκοπό την ανακάλυψη σημαντικών, ως προς κάποιο γεγονός, γεωγραφικών σημείων μέσα στις ομάδες των δεδομένων. Μέσα από τη διαδικασία της συσταδοποίησης και εξόρυξης πληροφοριών θα μπορούν να ανακαλυφθούν σημαντικές τοποθεσίες από δεδομένα που περιγράφουν τροχιές κινούμενων αντικειμένων.

Στο παραπάνω πλαίσιο, επικεντρωνόμαστε στην εφαρμογή ενός αλγορίθμου που θα ανήκει στη κατηγορία της συσταδοποίησης βάσει πυκνότητας για να μπορεί να ανταποκριθεί στην πολυπλοκότητα του μεγάλου όγκου δεδομένων τα οποία κατανέμονται σε μεγάλες ή μικρές ομάδες και ο αριθμός των ομάδων αυτών αλλάζει δυναμικά στο χώρο και το χρόνο. Ένας τέτοιος αλγόριθμος είναι ο DBSCAN (Kondaveeti A., 2012). Η υπόθεση που εξετάζεται στην εργασία αυτή είναι:

- Εφαρμόζοντας τον αλγόριθμο DenStream ο οποίος βασίζεται στον αλγόριθμο DBSCAN και κάνοντας χρήση των βιβλιοθηκών της γλώσσας R και συγκεκριμένα της βιβλιοθήκης MOA μπορεί να υλοποιηθεί ένα σύστημα στατιστικής ανάλυσης δεδομένων, το οποίο σε πραγματικό χρόνο θα μπορεί να επεξεργάζεται ροές δεδομένων, να τις ταξινομεί σε συμπλέγματα/ομαδοποιήσεις (π.χ. γεωγραφικές περιοχές) και να τις αξιολογεί (Michael Hahsler, 2015). Στο πλαίσιο αυτό εξετάζεται επίσης και ο αλγόριθμος CluStream και κατά πόσο μπορεί να αντικαταστήσει τον αλγόριθμο DenStream.
- Θα μπορεί το σύστημα αυτό να παρέχεται συμβουλευτικά σε συγκεκριμένους οργανισμούς και στη προσπάθεια τους να αξιολογούν δεδομένα και να λαμβάνουν αποφάσεις σε πραγματικό χρόνο. Σε μια τέτοια μελέτη περίπτωσης εξετάζουμε περίπου 10 εκατομμύρια εγγραφές από δεδομένα κυκλοφορίας των ταξί στο Πεκίνο. Τα δεδομένα είναι 6 ημερών και μπορούν να χρησιμοποιηθούν για να αξιολογήσουμε αν το σύστημα στατιστικής ανάλυσης δεδομένων που θα υλοποιηθεί μπορεί να ανταποκριθεί στις ανάγκες ενός σχετικού οργανισμού (π.χ. να παρακολουθεί την κίνηση για να εντοπίσει μποτιλιάρισμα, να εντοπίσει περιοχές με το μεγαλύτερο ενδιαφέρον και σε ποιες χρονικές στιγμές της ημέρας) (Jing Yuan, 2010).

Το περιεχόμενο της διπλωματικής εργασίας έχει οργανωθεί ως εξής. Το κεφάλαιο 2 παρουσιάζει ανασκόπηση της βιβλιογραφίας σχετικά με την ερευνητική περιοχή της

συσταδοποίησης ροών δεδομένων και των διαφορετικών τεχνικών και αλγορίθμων. Το κεφάλαιο 3 περιγράφει αναλυτικά το αντικείμενο της διπλωματικής εργασίας. Ως πηγή δεδομένων θα οριστεί ένα σύνολο GPS δεδομένων της Microsoft από τη κυκλοφορία 10.500 περίπου ταξί στο Πεκίνο (Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang, 2010). Επίσης αναλύει την αρχιτεκτονική της υλοποίησης της βιβλιοθήκης MOA της γλώσσας R η οποία περιλαμβάνει ρουτίνες ροής δεδομένων, και ενσωματώνει τον αλγόριθμο DenStream και CluStream όπως και άλλες ρουτίνες συσταδοποίησης δεδομένων στο χώρο και το χρόνο. Η R περιλαμβάνει επίσης βιβλιοθήκες γραφικής απεικόνισης των δεδομένων. Το κεφάλαιο 4 καταγράφει τα αποτελέσματα των πειραμάτων για την εξαγωγή χρήσιμων συμπερασμάτων. Το κεφάλαιο 5 κλείνει τη διπλωματική εργασία με τα συνολικά συμπεράσματα για τη συνεισφορά της διπλωματικής εργασίας και τις νέες κατευθύνσεις που ανοίγει στη συγκεκριμένη ερευνητική περιοχή.

Κ Ε Φ Α Λ Α Ι Ο 2

2. Ανασκόπηση βιβλιογραφίας

2.1 Συσταδοποίηση ροών δεδομένων (data stream clustering)

Η συσταδοποίηση (clustering) είναι μία από τις σημαντικότερες τεχνικές της μηχανικής μάθησης, με στόχο την συσταδοποίηση αντικειμένων με παρόμοια χαρακτηριστικά σε σύνολα και έχοντας διαφορές με άλλα τα οποία και τοποθετούνται σε διάφορες ομάδες. Παραδοσιακές τεχνικές συσταδοποίησης χωρίς επίβλεψη εκτελούνται συνήθως σε batch διαδικασίες: αυτό σημαίνει ότι τα δεδομένα φορτώνονται με κάποιο τρόπο στη μνήμη εκτελώντας αρκετά περάσματα μέχρι την εξαγωγή του τελικού αποτελέσματος. Ωστόσο, το νέο πρότυπο που εισάγουν τα Big Data, και έχοντας ως βασικά του συστατικά τον όγκο και την ταχύτητα, έχει δημιουργήσει ένα νέο περιβάλλον όπου τα δεδομένα μπορεί να είναι δυνητικά μη-πεπερασμένα και η είσοδος τους στη προκείμενη διαδικασία συνεχής. Τέτοιες ροές δεδομένων μπορεί να απαιτήσουν υπολογιστικά συστήματα υψηλών ταχυτήτων και μπορεί να περιέχουν διαδικασίες δημιουργίας δεδομένων οι οποίες δεν είναι στατικές (Merino, J.A., 2015).

Σε ένα πλαίσιο απαιτήσεων συσταδοποίησης, θα ήταν αδύνατο να αποθηκευτούν όλα τα δεδομένα στη μνήμη, όπως επίσης άγνωστο αριθμό και μέγεθος των clusters. Τα επίπεδα «θορύβου» (μη σωστής συσταδοποίησης) μπορεί επίσης να είναι υψηλά εξαιτίας είτε της δημιουργίας νέων δεδομένων ή της μετάδοσης τους. Όλοι αυτοί οι παράγοντες καθιστούν τις παραδοσιακές μεθόδους συσταδοποίησης διεθνώς αδύνατο να ανταπεξέλθουν. Κατά συνέπεια, η συσταδοποίηση ροών δεδομένων έχει αναδειχθεί ως ένα πεδίο με έντονη ερευνητική δραστηριότητα, με στόχο την αντιμετώπιση αυτών των προκλήσεων (Merino, J.A., 2015).

Επίσης, ο (Forrest J., 2011) αναφέρει ότι εξαιτίας του μεγέθους και της δυναμικής φύσης των ροών δεδομένων, είναι συχνά δύσκολο να επεξεργαστεί κάποιος ροές δεδομένων σε πραγματικό χρόνο χωρίς να δεχτεί το κόστος της δημιουργίας μίας υποδομής/εργαλειοθήκης η οποία μπορεί να δημιουργήσει δεδομένα με συγκεκριμένες ιδιότητες. Μία τέτοια εργαλειοθήκη έχει υλοποιηθεί στη γλώσσα R.

2.1.1 Συνιστώσες συσταδοποίησης ροών δεδομένων

Μία ροή δεδομένων είναι μία τεράστια ακολουθία δεδομένων. Η κάθε περίπτωση (στη ροή δεδομένων) περιγράφεται από ένα διάνυσμα d-διαστάσεων. Το περιβάλλον ροής δεδομένων στον τομέα των Big Data θέτει ιδιαίτερες προκλήσεις στις οποίες ισχύουν οι ακόλουθοι περιορισμοί (Merino, J.A., 2015):

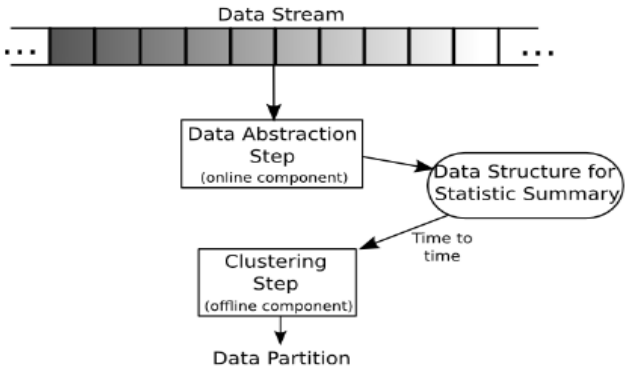
- ο όγκος των δεδομένων μπορεί να είναι υψηλός ή απεριόριστος
- τα δεδομένα φτάνουν συνεχώς
- η σειρά της άφιξη των δεδομένων δεν μπορεί να ελεγχθεί
- οι διαστάσεις των δεδομένων δεν περιορίζονται
- μόνο ένα πέρασμα επιτρέπεται για τα δεδομένα
- τα δεδομένα απορρίπτονται μετά την πρώτη τους επεξεργασία
- τα δεδομένα μπορεί να είναι δυναμικά εκ φύσεως
- οι υπολογισμοί πρέπει να είναι γρήγοροι και γραμμικοί σε κλίμακα

Επίσης στο (Silva, Jonathan A. and Faria, Elaine R. and Barros, Rodrigo C. and Hruschka, Eduardo R. and Carvalho, 2013) ορίζεται ένα πλαίσιο για τον προσδιορισμό των πιο σημαντικών διαστάσεων της διαδικασίας συσταδοποίησης ροών δεδομένων!

- δομή δεδομένων για στατιστική περίληψη
- μοντέλο χρονικού παραθύρου
- μηχανισμός ανίχνευσης ακραίων τιμών
- αριθμό των παραμέτρων που ορίζονται από το χρήστη
- αλγόριθμος συσταδοποίησης εκ των υστέρων (offline)
- σχήμα συσταδοποίησης
- κατηγορία προβλήματος συσταδοποίησης

Η εφαρμογή αλγορίθμων συσταδοποίησης σε ροές δεδομένων διακρίνονται σε αντικειμενοστραφείς προσεγγίσεις ή ομαδοποιήσεις βάσει χαρακτηριστικών. Η Εικόνα 1 παρουσιάζει τις βασικές συνιστώσες της αντικειμενοστραφούς προσέγγισης:

1. ορισμός δεδομένων σε αφαιρετικό επίπεδο (online εκτέλεση) - συνοψίζει τη ροή δεδομένων με τη βοήθεια συγκεκριμένων δομών δεδομένων για την αντιμετώπιση περιορισμών στον αποθηκευτικό χώρο και τη μνήμη του υπολογιστή. Αυτές οι δομές δεδομένων συνοψίζουν τη ροή, προκειμένου να διατηρηθεί η σημασία των αρχικών δεδομένων χωρίς την ανάγκη της αποθήκευση τους.
2. συσταδοποίηση (offline εκτέλεση)



Εικόνα 1: αντικειμενοστραφές πλαίσιο συσταδοποίησης ροών δεδομένων (Silva, Jonathan A. and Faria, Elaine R. and Barros, Rodrigo C. and Hruschka, Eduardo R. and Carvalho, 2013)

3. ορισμός χρονικού παραθύρου

Προκειμένου να ανιχνευτεί μία πιθανή αλλαγή κατά τη δημιουργία μίας κατανομής δεδομένων χωρίς να αυξηθούν οι υπολογιστικές απαιτήσεις (μνήμη, cpu), τα πλέον πρόσφατα δεδομένα στη ροή αναλύονται και κανονικοποιούνται. Αυτό είναι απαραίτητο για να δοθεί μεγαλύτερη σημασία σε νεότερα δεδομένα μέσα στη ροή σε σχέση με παλαιότερα. Αν αυτό δεν συνέβαινε, η αλλαγή στη ροή δεδομένων δεν θα μπορούσε να εντοπιστεί. Η συσταδοποίηση σε ένα περιβάλλον ροής μπορεί επίσης να ποικίλει ανάλογα με τη στιγμή που προσδιορίζεται μέσα στο χρονικό παράθυρο ή τον ορίζοντα επί των οποίων έχουν μετρηθεί. Δηλαδή, η συσταδοποίηση για τα τελευταία 5 δευτερόλεπτα θα μπορούσε να δώσει ένα εντελώς διαφορετικό αποτέλεσμα συσταδοποίησης από εκείνο που μπορεί να υπολογιστεί για τους τελευταίους 5 μήνες.
Από τους επιμέρους τύπους χρονικών παραθύρων, το μοντέλο των Συρόμενων Χρονικών Παραθύρων ταιριάζει στο πλαίσιο της παρούσας διπλωματικής εργασίας. Το μοντέλο αυτό υποθέτει ότι τα πιο πρόσφατα στοιχεία είναι περισσότερο σχετικά σε σχέση με παλαιότερα δεδομένα με τη χρήση FIFO ουράς. Κατά συνέπεια, ένα χρονικό παράθυρο διατηρείται κρατώντας τις πιο πρόσφατες περιπτώσεις που εμπίπτουν στο παράθυρο της ροής δεδομένων απορρίπτοντας παλαιότερα δεδομένα. Για τα στιγμιότυπα δεδομένων που περιέχονται στο παράθυρο, τρεις εργασίες μπορούν να πραγματοποιηθούν: ανίχνευση αλλαγών χρησιμοποιώντας υπο-παράθυρα, λήψη στατιστικών στοιχείων από τα στιγμιότυπα και ενημέρωσης των μοντέλων.
Το μέγεθος του χρονικού παραθύρου είναι μια παράμετρος που ρυθμίζεται από το χρήστη. Ένα μικρό μέγεθος του παραθύρου θα επιλέξει τα τελευταία στιγμιότυπα της ροής και επομένως οι όποιοι αλγόριθμοι θα αντιδράσουν γρήγορα στην άφιξη νέων δεδομένων αντανακλώντας τις αλλαγές στη κατανομή τους. Από την άλλη πλευρά, ένα μεγάλο μέγεθος του χρονικού παραθύρου θα επιλέξει ένα μεγαλύτερο σύνολο στιγμιότυπων που θα αυξάνει την ακρίβεια σε σταθερά χρονικά διαστήματα. Επομένως η εύρεση του ιδανικού μεγέθους του παραθύρου είναι μια πρόκληση. Για την αντιμετώπιση αυτού του ζητήματος, διάφορες τεχνικές (π.χ. Adwin) έχουν υλοποιηθεί για την ανίχνευση αλλαγών στα δεδομένα διανομής και την αυτόματη προσαρμογή ανάλογα με το μέγεθος του παραθύρου. Τα πειράματα με τις τεχνικές αυτές έδειξαν πολύ θετικά αποτελέσματα στη βελτίωση της συνολικής ακρίβειας.

4. ανίχνευση ακραίων τιμών- η διαδικασία ανίχνευσης ακραίων τιμών μπορεί να υλοποιηθεί είτε ως μία στατιστική προσέγγιση ή βάση πυκνότητας.

Ο παρακάτω πίνακας (Πίνακας 1) παρουσιάζει τους πιο συναφείς αλγορίθμους συσταδοποίησης οι οποίοι έχουν ταξινομηθεί από τους Silva, Jonathan A. and Faria, Elaine R. and Barros, Rodrigo C. and Hruschka, Eduardo R. and Carvalho, 2013) βάσει των παραπάνω διαστάσεων.

Αλγόριθμος	**Αλγόριθμος Συσταδοποίησης (Clustering)**	**Δομή Δεδομένων**	**Μοντέλο χρονικού παραθύρου**	**Μέθοδος ανίχνευσης ακραίων τιμών**	**Αριθμός παραμέτρων**
BIRCH	k-means	Διάνυσμα χαρακτηριστικών (feature vector)	Μοντέλο ορόσημων (landmark)	Βάσει μετρικών πυκνότητας	5
CluStream	k-means	Διάνυσμα χαρακτηριστικών	Μοντέλο ορόσημων	Βάσει στατιστικών μετρικών	9
ClusTree	k-means /DBSCAN	Διάνυσμα χαρακτηριστικών	Μοντέλο απόσβεσης (damped)	-	3
D-Stream	DBSCAN	Πίνακας τιμών	Μοντέλο απόσβεσης	Βάσει μετρικών πυκνότητας	5
DenStream	DBSCAN	Διάνυσμα χαρακτηριστικών	Μοντέλο απόσβεσης	Βάσει μετρικών πυκνότητας	4
DGClust	k-means	Πίνακας τιμών	Μοντέλο ορόσημων		5
ODAC	Ιεραρχική συσταδοποίηση	Πίνακας συσχέτισης	Μοντέλο ορόσημων		3
Scalable k-means	k-means	Διάνυσμα χαρακτηριστικών	Μοντέλο ορόσημων		5
Single-pass k-means	k-means	Διάνυσμα χαρακτηριστικών	Μοντέλο ορόσημων		2

Πίνακας 1: αλγόριθμοι συσταδοποίησης ροών δεδομένων (Silva, Jonathan A. and Faria, Elaine R. and Barros, Rodrigo C. and Hruschka, Eduardo R. and Carvalho, 2013)

2.2 Τεχνικές χωρο-χρονικής συσταδοποίησης ροών δεδομένων

2.2.1 Παραδοσιακές τεχνικές συσταδοποίησης δεδομένων

Οι μέθοδοι συσταδοποίησης μπορούν να κατηγοριοποιηθούν με βάση τον τύπο δεδομένων που εισάγονται στον αλγόριθμο, τη μέθοδο που καθορίζει την συσταδοποίηση του συνόλου των δεδομένων, τη θεωρία και τις θεμελιώδεις έννοιες στις οποίες είναι βασισμένες οι τεχνικές ανάλυσης συστάδας (Merino, J.A., 2015).

Η κατηγοριοποίηση με βάση τύπο δεδομένων ασχολείται είτε με τη συσταδοποίηση αριθμητικών δεδομένων, για παράδειγμα η Ευκλείδεια απόσταση,

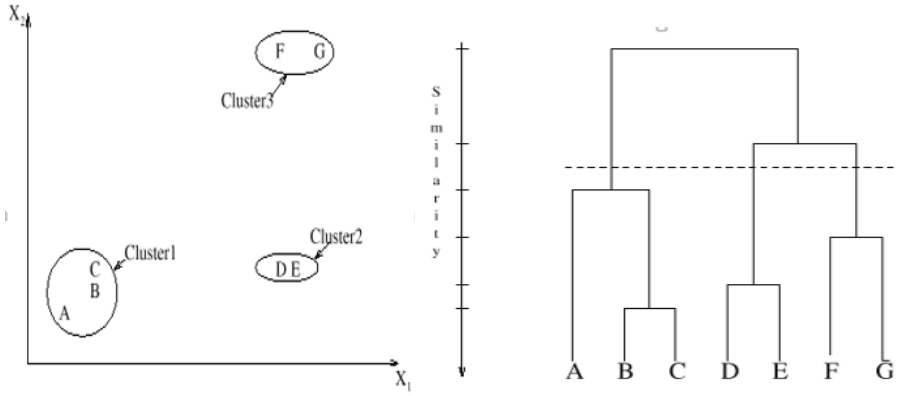
$$\sqrt{(q_{1-}\, p_1)^2 + (q_{2-}\, p_2)^2}$$

Η κατηγορική και η κειμενική συσταδοποίηση. Ενώ η κατηγοριοποίηση με βάση τη μέθοδο συσταδοποίησης περιλαμβάνει τις εξής τεχνικές:

Ιεραρχικοί Αλγόριθμοι (Hierarchical Clustering)

Οι ιεραρχικοί αλγόριθμοι διακρίνονται σε συσσωρευτικούς (Agglomerative) και διαιρετικούς (Divisive). Δημιουργούν συστάδες προς τα κάτω/προς τα πάνω χρησιμοποιώντας δενδρογράμματα και κριτήρια συνάθροισης (single linkage, complete, average linkage και Ward). Τα κριτήρια σύνδεσης διακρίνονται σε:

- Μονή σύνδεση (single linkage): η μικρότερη απόσταση ανάμεσα σε ζεύγη συστάδων.
- Μέση σύνδεση (average linkage): η μέση απόσταση ανάμεσα σε ζεύγη συστάδων.
- Πλήρης σύνδεση (complete linkage): η μέγιστη απόσταση ανάμεσα σε ζεύγη συστάδων.
- Η Απόσταση μεταξύ των κεντρικών σημείων της κάθε συστάδας (centroids).



Εικόνα 2: δενδρόγραμμα συστάδων – εφαρμογή του αλγορίθμου single-link

Διαμεριστικοί Αλγόριθμοι (Partitioning Clustering)

Οι Διαμεριστικοί Αλγόριθμοι δημιουργούν ομάδες σε ένα βήμα επιχειρώντας να ελαχιστοποιηθεί η απόσταση μεταξύ των σημείων που έχουν επισημανθεί ως μέρος μίας συστάδας και του σημείου που έχει οριστεί ως το κέντρο της συστάδας. Καθώς απαιτείται (κατά κανόνα) γνώση του μεγέθους της ομάδας συνήθως διαχειρίζεται στατικά σύνολα. Ο αλγόριθμος K-means είναι η πιο γνωστή τεχνική αυτής της προσέγγισης. Οι k ομάδες είναι συνήθως γνωστές εκ των προτέρων και ένα στοιχείο επιλέγεται ως αντιπρόσωπος για την συστάδα του, δημιουργώντας Gaussian τύπου ομάδες (Merino, J.A., 2015). Μερικοί ιεραρχικοί αλγόριθμοι μπορούν να μετατραπούν σε διαμεριστικούς, παράδειγμα ο Αλγόριθμος Τετραγωνικού Σφάλματος (MST) ο οποίος λειτουργεί ως εξής:

- Το αρχικό σύνολο των συστάδων και των κεντρικών τους σημείων επιλέγεται τυχαία.
- Ανάθεση των αντικειμένων στις πιο κοντινές συστάδες βάσει της απόστασης από τα κεντρικά σημεία.
- Επαναϋπολογισμός του κέντρου της κάθε συστάδας.
- Υπολογισμός του τετραγωνικού σφάλματος:

$$s_i = \sum_{j=1}^{m} \|t_{ji} - c_i\|^2 \qquad S = \sum_{i=1}^{k} S_i$$

- Επανάληψη της παραπάνω διαδικασίας μέχρις ότου η διαφορά μεταξύ δύο διαδοχικών τετραγωνικών σφαλμάτων, να είναι πιο μικρή από ένα όριο.

Ενώ ο k-means επιχειρεί να ελαχιστοποιήσει τη μέση τετραγωνική απόσταση των δεδομένων από τα πλησιέστερα κέντρα των συστάδων και μπορεί επίσης να θεωρηθεί αλγόριθμος τετραγωνικού σφάλματος:

- Το αρχικό σύνολο από συστάδες επιλέγεται τυχαία.
- Ανάθεση των αντικειμένων στις πιο κοντινές συστάδες βάσει της απόστασης από τα κεντρικά τους σημεία.
- Επαναϋπολογισμός του κέντρου της κάθε συστάδας.

- Επαναληπτικά, τα αντικείμενα μετακινούνται ανάμεσα σε σύνολο από συστάδες έως ότου εντοπιστεί το επιθυμητό σύνολο.

Αλγόριθμοι Ασαφούς Ομαδοποίησης

Στη πλειοψηφία των αλγορίθμων συσταδοποίησης δημιουργείται μία διαμέριση ένα-προς-ένα, δηλαδή το κάθε αντικείμενο ανήκει αποκλειστικά και μόνο σε μία ομάδα. Αντίθετα, σε μια ασαφή ομαδοποίηση είναι δυνατόν κάθε αντικείμενο να ανήκει σε πάνω από μία ομάδες. Ο βασικός πυρήνας των αλγορίθμων ασαφούς ομαδοποίησης αποτελείται από τα παρακάτω βήματα:

- Επιλογή μίας αρχικής ασαφούς διαίρεσης των N αντικειμένων σε K συστάδες παίρνοντας τον NxK πίνακα U με τους βαθμούς στους οποίους το κάθε αντικείμενο ανήκει σε κάθε συστάδα. Κάθε στοιχείο u_{ij} στον πίνακα αναπαριστά τον βαθμό στον οποίο το αντικείμενο xi ανήκει στην ομάδα c_j. Συνεπώς $u_{ij} \in [0, 1]$.
- Χρησιμοποιώντας τον πίνακα U υπολογίζεται η τιμή μίας ασαφούς συνάρτησης, για παράδειγμα η συνάρτηση τετραγωνικού σφάλματος με βάρη:

$$E^2(\mathscr{X}, U) = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ij}\|\mathbf{x}_i - \mathbf{c}_k\|^2,$$

Όπου $\mathbf{c}_k = \sum_{i=1}^{N} u_{ij}\mathbf{x}_i$ είναι το κέντρο της κάθε ασαφούς ομάδας. Εκτελείται εκ νέου η κατανομή των αντικειμένων σε συστάδες ώστε να μειωθεί η τιμή της ασαφούς συνάρτησης και υπολογίζεται ξανά ο πίνακας U.

Ο Fuzzy C-Means είναι ο πιο αντιπροσωπευτικός αλγόριθμος ασαφούς ομαδοποίησης.

- Το παραπάνω βήμα επαναλαμβάνεται μέχρι να μην είναι δυνατή η περαιτέρω μεταβολή του πίνακα U.

Αλγόριθμοι βασισμένοι στη πυκνότητα

Βασική αρχή των αλγορίθμων αυτών είναι ότι για κάθε αντικείμενο που εκχωρείται σε μία συστάδα θα πρέπει να υπάρχει ένα σύνολο γειτνίασης (e-neighborhood) με το αντικείμενο αυτό που να περιέχει ένα ελάχιστο αριθμό σημείων $e > 0$. Η μεταβλητή e-neighborhood

ορίζει συνεπώς τη πυκνότητα της συστάδας. Όσα σημεία δεν ανήκουν σε κάποια αντικείμενα αποτελούν θόρυβο. Ο DBSCAN είναι ο πιο αντιπροσωπευτικός αλγόριθμος και παρουσιάζεται στη παρακάτω ενότητα.

Εξελικτικοί αλγόριθμοι

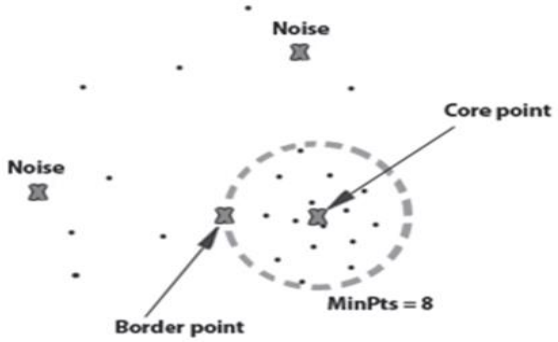
Οι αλγόριθμοι αυτοί αντιγράφουν βασικές αρχές της εξέλιξης στη φύση. Εφαρμόζοντας τελεστές εξέλιξης σε ένα πληθυσμό ομαδοποίησης σκοπός τους είναι η βέλτιστη ομαδοποίηση των αντικειμένων. Οι υποψήφιες λύσεις αντιπροσωπεύονται ως χρωμοσώματα και οι τελεστές εξέλιξης περιλαμβάνουν συναρτήσεις επιλογής, επανασυνδυασμού, και μετάλλαξης. Για κάθε χρωμόσωμα υπολογίζεται η τιμή μίας συνάρτησης βελτιστοποίησης (fitness function) αντιπροσωπεύοντας πόσο πιθανό είναι το συγκεκριμένο χρωμόσωμα να επιβιώσει στην επόμενη γενιά εκτελώντας τα παρακάτω βήματα:

- Τυχαία επιλογή ενός πληθυσμού από λύσεις/συστάδες. Υπολογίζεται η τιμή της συνάρτησης βελτιστοποίησης, ως τιμή αντιστρόφως ανάλογη του τετραγωνικού σφάλματος.
- Εφαρμογή των τελεστών επιλογής, επανασυνδυασμού, και μετάλλαξης για τη δημιουργία της επόμενης γενιάς λύσεων/ομαδοποιήσεων. Υπολογισμός της τιμής της συνάρτησης βελτιστοποίησης.
- Επανάληψη του παραπάνω βήματος μέχρι την ικανοποίηση ενός κριτηρίου τερματισμού

Οι πιο γνωστοί εξελικτικοί αλγόριθμοι περιλαμβάνουν τους Γενετικούς Αλγορίθμους, τις Εξελικτικές Στρατηγικές και τον Εξελικτικό Προγραμματισμό.

2.2.2 Συσταδοποίηση βάσει πυκνότητας - DBSCAN

Η συσταδοποίηση με βάση την πυκνότητα των ομάδων βασίζεται κυρίως στον αλγόριθμο DBSCAN και τις επεκτάσεις του (M. Hahsler, and M. Bolanos, 2016). Ο DBSCAN υπολογίζει τη πυκνότητα γύρω από κάθε σημείο δεδομένων μετρώντας τον αριθμό των σημείων σε μία καθορισμένη από το χρήστη περιοχή γειτνίασης (eps-neighborhood) και επιτρέπει τον ορισμό κατώτατων ορίων από το χρήστη για τον εντοπισμό σημείων στο κέντρο των περιοχών (core), στα σύνορα τους (border) και σημεία που ενέχουν «θόρυβο» (noise).



Εικόνα 2: συσταδοποίηση κάνοντας χρήση του αλγορίθμου DBSCAN (κεντρικά σημεία, ακραία σημεία, σημεία εκτός ορίων – noise) (T. Galba, Z. Balkić, G. Martinović, 2013)

Σε ένα δεύτερο στάδιο, επιπλέον σημεία δεδομένων προστίθενται στο κέντρο των περιοχών αν είναι προσβάσιμα βάσει πυκνότητας (δηλαδή υπάρχει μια αλυσίδα βασικών (core) σημείων όπου το καθένα ανήκει στην περιοχή γειτνίασης (Eps-neighborhood) του επόμενου σημείου. Τέλος, είναι εφικτό στη κάθε συστάδα να διακριθούν ακραίες τιμές (σύνορα τιμών).

Η περιοχή γειτνίασης ενός σημείου p ορίζεται ως το σύνολο των σημείων για τα οποία ισχύει ότι η Ευκλείδεια απόσταση ανάμεσα στα σημεία p, q είναι μικρότερη της παραμέτρου Eps:

- $N_{Eps}(p) = \{ q \in D \mid dist(p, q) \leq Eps \}$

εφόσον **p** = (p_1, p_2) and **q** = (q_1, q_2) η Ευκλείδεια απόσταση ορίζεται ως:

$$\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$$

Ένα σημείο p θεωρείται ότι είναι προσβάσιμο από ένα σημείο q βάσει πυκνότητας που προσδιορίζεται από τις παραμέτρους Eps, MinPts εφόσον:

- $p \in N_{Eps}(q)$ και
- $N_{Eps}(q) \geq MinPts$

```
DBSCAN (SetOfPoints, Eps, MinPts)
        SetOfPoints is UNCLASSIFIED
        ClusterId := nextId(NOISE);
        FOR i FROM 1 TO SetOfPoints.size DO
            Point := SetOfPoints.get(i);
            IF Point.CiId = UNCLASSIFIED THEN
                IF ExpandCluster(SetOfPoints, Point, ClusterId, Eps, MinPts) THEN
                    ClusterId := nextId(ClusterId)
                END IF
            END IF
        END FOR
END;
```

setOfPoints είναι είτε ολόκληρη η βάση δεδομένων ή μία συσταδοποίηση (M. Hahsler, and M. Bolanos, 2016) που προέκυψε από προηγούμενη εκτέλεση. Οι παράμετροι *Eps* και *MinPts* είναι οι καθολικές παράμετροι πυκνότητας οι οποίες υπολογίζονται είτε χειροκίνητα είτε βάσει κανόνων (heuristics). Η μέθοδος *ExpandCluster* είναι η πιο σημαντική διαδικασία του αλγορίθμου.

```
ExpandCluster (SetOfPoints, Point, CiId, Eps, MinPts) : Boolean;
        seeds :=SetOfPoints.regionQuery (Point, Eps );
        IF seeds.size<MinPts THEN // no core point
            SetOfPoint.changeClId (Point, NOISE);
            RETURN False;
        ELSE // all points in seeds are density- reachable from Point
            SetOfpoints.changeCiIds ( seeds, C1 Id) ;
            seeds.delete (Point) ;
            WHILE seeds <> Empty DO
                currentP := seeds.first();
                result := setofPoints.regionQuery(currentP, Eps);

                IF result.size >= MinPts THEN
                    FOR i FROM 1 TO result.size DO
                        resultP := result.get(i);
                        IF resultP.CiId IN (UNCLASSIFIED, NOISE} THEN
```

```
                    IF resultP.ClId = UNCLASSIFIED THEN
                        Seeds.append (resultP);
                    END IF;
                    SetOfPoints.changeClId ( resultP, ClId);
                END IF; // UNCLASSIFIED or NOISE
            END FOR ;
        END IF; // result.size >= MinPts

        Seeds.delete (currentP)
    END WHILE; // seeds <> Empty
    RETURN True;
  END IF
END; // ExpandCluster
```

Η μέθοδος *SetOfPoints.regionQuery (Point, Eps)* επιστρέφει το σύνολο των γειτονικών σημείων (Eps-neighborhood) ενός σημείου Point. Αν κάποια συσταδοποίηση -ClId (ClusterId)- χαρακτηριστεί αρχικά ως "τιμή εκτός συνόρων" (NOISE) αυτό μπορεί να αλλάξει στη πορεία της εκτέλεσης αν τα δεδομένα σημεία είναι πλέον σε κοντινή απόσταση (βάσει πυκνότητας) από άλλα σημεία στη βάση δεδομένων. Κυρίως αυτό αφορά σημεία που είναι στα σύνορα των ομαδοποιήσεων. Επιπλέον, αν δύο ομάδες σημείων είναι σε κοντινή απόσταση υπάρχει το ενδεχόμενο κάποιο σημείο να ανήκει και στις δύο ομάδες. Στη περίπτωση αυτή τέτοια σημεία θα πρέπει να χαρακτηριστούν συνοριακά σημεία διαφορετικά θα έπρεπε οι δύο ομάδες να χαρακτηριστούν ισοδύναμες.

Ειδικότερα, σε προβλήματα συσταδοποίησης γεωγραφικών σημείων βάσει πυκνότητας (density-based spatial clustering) επιλέγεται ο DBSCAN για τους παρακάτω λόγους:

- Είναι κατάλληλος για γεωγραφικά δεδομένα.
- Δεν απαιτεί εκ των προτέρων γνώση για τον αριθμό των ομάδων και αυτό είναι πολύ θετικό για μεγάλα δεδομένα.
- Εντοπίζει ομάδες με διαφορετικό σχήμα.

Μειονεκτήματα του αλγορίθμου DBSCAN

Το κυριότερο μειονέκτημα του αλγορίθμου είναι ότι επηρεάζεται από τις τιμές των παραμέτρων Eps και MinPts, οι οποίες είναι δύσκολο να προσδιοριστούν. Όπως όλοι οι ιεραρχικοί αλγόριθμοι πάσχει από το πρόβλημα της μη ευρωστίας καθώς στην περίπτωση που υπάρχει μία πυκνή σειρά σημείων που συνδέει δύο συστάδες ο DBSCAN μπορεί να τελειώσει συγχωνεύοντας τις δύο συστάδες. Η χρήση δείγματος για να περιοριστεί το μέγεθος της εισόδου στην εφαρμογή των αλγορίθμων που βασίζονται στην πυκνότητα δεν είναι εφικτή. Ο λόγος είναι ότι ακόμα και αν το δείγμα είναι μεγάλο, μπορεί να υπάρχουν μεγάλες διακυμάνσεις στην πυκνότητα των σημείων μέσα σε κάθε συστάδα στο τυχαίο δείγμα.

Παραμένει θετικό ότι ο χρόνος τρεξίματος είναι O(nlogn).

C-DBSCAN

Οι (Lei Gong, Hitomi Sato, Toshiyuki Yamamoto, Tomio Miwa, Takayuki Morikawa, 2015) παρουσιάζουν τη μεθοδολογία και τους περιορισμούς που θέτουν έτσι ώστε χρησιμοποιώντας τον DBSCAN να αναγνωρίσουν στάσεις σε τροχιές και Points Of Interest (POI) σε GPS data.

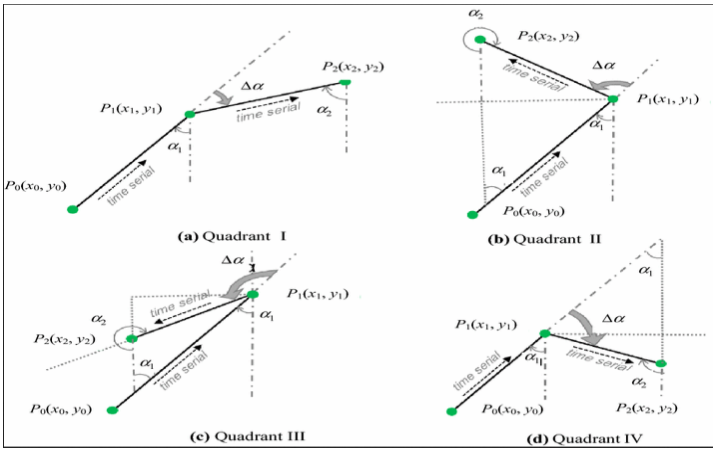
Η συγκεκριμένη εργασία προτείνει μια μέθοδο για τον προσδιορισμό των θέσεων τερματισμού μίας δραστηριότητας χρησιμοποιώντας μόνο συντεταγμένες και χρονικά δεδομένα (timestamps) από τροχιά GPS η οποία είναι συνεχής. Η μέθοδος αυτή αποτελείται από μια βελτιωμένη έκδοση του αλγορίθμου DBSCAN (ως πρώτο βήμα) και Μηχανές Διανυσματικής Υποστήριξης- SVMs (ως δεύτερο βήμα). Έχει δοκιμαστεί με GPS δεδομένα τα οποία έχουν συλλεχθεί από κινητό τηλέφωνο παρέχοντας αποτελέσματα με μεγαλύτερη ακρίβεια από ό, τι άλλες παραλλαγές της DBSCAN στον εντοπισμό θέσεων τερματισμού μίας δραστηριότητας.

Στην περίπτωση των γεωγραφικών σημείων τα οποία προσδιορίζονται από GPS δεδομένα και για τα οποία εφαρμόζεται ο αλγόριθμος DBSCAN, τα σημεία μέσα σε κάθε σύμπλεγμα (cluster) αντιστοιχούν σε σημεία τερματισμού τα οποία συγκεντρώνονται με μεγαλύτερη πυκνότητα. Από την άλλη πλευρά, τα σημεία στο όριο του θορύβου αντιπροσωπεύουν κινούμενα σημεία κατά μήκος του δρόμου ή σιδηροδρομικού δικτύου με χαμηλότερη πυκνότητα.

Τονίζεται ότι ο αλγόριθμος DBSCAN αναπτύχθηκε για να λύσει το πρόβλημα της κατηγοριοποίησης των δεδομένων (σημείων) στο χώρο διαβάθμισης χωρίς εξέταση της χρονικής τους αλληλουχίας. Κατά συνέπεια, σε μια παράκαμψη της τροχιάς, ένα σύμπλεγμα με σημεία τερματισμού μπορεί να περιέχει άλλα κινούμενα σημεία ή σημεία τα οποία έχουν δηλωθεί επίσης σε επόμενα συμπλέγματα και μοιράζονται την ίδια θέση. Επιπλέον, λόγω των ορισμών και εννοιών του αρχικού DBSCAN αλγόριθμου, σημεία που αντιπροσωπεύουν την κίνηση κατά μήκος ενός δρόμου ο οποίος είναι ευθύς σε χαμηλή ταχύτητα μπορούν να ομαδοποιηθούν σε ένα ενιαίο σύμπλεγμα όταν η συχνότητα του GPS μετάδοση σήματος είναι υψηλή και υπό ορισμένες συγκεκριμένες τιμές των παραμέτρων. Ως αποτέλεσμα, η εφαρμογή του αρχικού DBSCAN αλγορίθμου για GPS τροχιές μπορεί να οδηγήσει σε σφάλματα. Στην εργασία αυτή προτείνεται η επέκταση του αλγορίθμου με δύο περιορισμούς προκειμένου να αποφευχθούν αυτά τα δύο πιθανά σφάλματα. τον αλγόριθμο. Ο αλγόρθμος ονομάζεται C-DBSCAN.

Ο πρώτος περιορισμός είναι ότι όλα τα σημεία σε ένα σύμπλεγμα θα πρέπει να είναι χρονικά διαδοχικά. Αυτό σημαίνει ότι η σειρά που θα πρέπει να αυξηθούν είναι ένα-προς-ένα και να μην επιτρέπεται «ξαφνική αύξηση» στο σύμπλεγμα. Εάν βρεθεί μια τέτοια «απότομη αύξηση», τότε σύμπλεγμα θα χωριστεί σε δύο πιθανά συμπλέγματα στο σημείο της απότομης αύξησης και το κάθε σύμπλεγμα θα πρέπει να δοκιμαστεί για το αν πληροί την προϋπόθεση του ελάχιστου αριθμού σημείων σε ένα σύμπλεγμα. Αν όχι, τα σημεία στο δυναμικό σύμπλεγμα θα πρέπει να χαρακτηριστούν ως κινούμενα σημεία. Διαφορετικά, τα σημεία στο δυναμικό σύμπλεγμα θα πρέπει να ελεγχθούν όσον αφορά τον δεύτερο περιορισμό.

Ο δεύτερος περιορισμός είναι ότι το ποσοστό των ανώμαλων σημείων (PCT) σε ένα σύμπλεγμα δεν πρέπει να υπερβαίνει ένα συγκεκριμένο όριο που ονομάζεται (PCT_{AP}). Συγκεκριμένα, θα πρέπει να ισχύει $PCT \leq PCT_{AP}$, όπου $PCT = AP / P$, AP είναι ο αριθμός των ανώμαλων σημείων και P είναι το σύνολο των σημείων στο ίδιο σύμπλεγμα.



Εικόνα 3: Κατεύθυνση ενός σημείου και την αλλαγή της κατεύθυνσης δύο σημείων για το δεύτερο σημείο σε διαφορετικά τεταρτημόρια (Lei Gong, Hitomi Sato, Toshiyuki Yamamoto, Tomio Miwa, Takayuki Morikawa, 2015)

Εάν ένα σύμπλεγμα αντιπροσωπεύει μια θέση τερματισμού, τα διάφορα σημεία συγκεντρώνονται διάσπαρτα γύρω από την τοποθεσία, και τα σημεία θα πρέπει να έχουν μια ομοιόμορφη κατανομή των αλλαγών κατεύθυνσης. Αυτό σημαίνει ότι το συνημίτονο της αλλαγής κατεύθυνσης (ή του συντελεστή αλλαγής κατεύθυνσης, (DCC)) θα πρέπει σχεδόν πάντα να διαφέρουν από 1. Τα σημεία με τιμή DCC κοντά στο 1 πιθανώς αντιπροσωπεύουν την κίνηση του αντικειμένου κατά μήκος μιας ευθείας γραμμής στο δίκτυο. Σε ένα σύμπλεγμα, τα σημεία θα πρέπει να έχουν ομοιόμορφη κατανομή του DCC. Έτσι ανώμαλα σημεία είναι εκείνα τα σημεία χωρίς ομοιόμορφη κατανομή του DCC, συγκεκριμένα, έχουν μία τιμή DCC κοντά στο 1 (DCCAP).

DBSCANSD

Η παρακολούθηση της θαλάσσιας κυκλοφορίας είναι μια σημαντική πτυχή της ασφάλειας και της προστασίας, ιδίως για λειτουργίες κοντά σε λιμάνια. Ενώ υπάρχει μια μεγάλη ποσότητα δεδομένων η ποιότητα των οποίων μεταβάλλεται, οι ιθύνοντες χρειάζονται αξιόπιστες πληροφορίες σχετικά με πιθανές καταστάσεις ή απειλές. Για την αντιμετώπιση αυτής της απαίτησης, οι (Bo Liu, Erico N. de Souza, Stan Matwin, Marcin Sydowzy, 2014) προτείνανε την υλοποίηση μίας επέκτασης του αλγορίθμου DBSCAN, η οποία ονομάζεται DBSCANSD, για την εξόρυξη μορφότυπων από κανονικές τροχιές πλοίων και τη δημιουργία ομάδων δεδομένων. Οι ομαδοποιήσεις των δεδομένων λαμβάνουν υπόψη τα δεδομένα γεωγραφικού εντοπισμού της θέσης των πλοίων καθώς και τους δημόσια διαθέσιμους κανόνες των Διεθνών Ναυτιλιακών Οργανισμών (IMO). Το κύριο αποτέλεσμα της συσταδοποίησης είναι ένα σύνολο από τροχιές πλοίων σε λωρίδες οι οποίες μπορούν να χαρτογραφηθούν και να συγκριθούν με εκείνες που είχαν αρχικά οριστεί σύμφωνα με τις οδηγίες του IMO. Καθώς το μοντέλο λαμβάνει υπόψη του και μη-χωρικά χαρακτηριστικά (ταχύτητα και κατεύθυνση) ο αλγόριθμος επιτρέπει στους ιθύνοντες να εντοπίσουν μη-φυσιολογική συμπεριφορά, για παράδειγμα, σκάφη τα οποία δεν είναι στην υποχρεωτική τροχιά-λωρίδα κίνησης ή σκάφη που ακολουθούν υψηλότερες ή χαμηλότερες ταχύτητες.

Τα δεδομένα τροφοδοτούνται από ένα Σύστημα Αυτόματης Αναγνώρισης (AIS), το οποίο είναι ένα αυτόματο σύστημα παρακολούθησης για την ταυτοποίηση και τον εντοπισμό σκαφών με την ανταλλαγή δεδομένων με άλλα κοντινά πλοία, σταθμούς βάσης AIS και τους δορυφόρους. Πάνω από 400.000 πλοία έχουν εγκαταστήσει το συγκεκριμένο σύστημα για να μπορούν οι Διεθνείς Ναυτιλιακοί Οργανισμοί να ελέγχουν τη κίνηση τους.

Η εξόρυξη μορφότυπων από κανονικές τροχιές πλοίων βασίζεται σε ένα μοντέλο που στη βιβλιογραφία ονομάζεται Stops-and-Moves *(Bo Liu, Erico N. de Souza, Stan Matwin, Marcin Sydowzy, 2014). Η αναζήτηση γίνεται σε ιστορικά AIS δεδομένα και μία από τις παραμέτρους που θα πρέπει να βελτιστοποιηθεί είναι η ταχύτητα που θεωρείται φυσιολογική (Speed over ground – SOG). Για παράδειγμα οι (Bo Liu, Erico N. de Souza, Stan Matwin, Marcin Sydowzy, 2014) ορίζουν το όριο του SOG στους 0.5 κόμβους. Επιπλέον η συγκεκριμένη προσέγγιση του DBSCANSD (*

Πίνακας 2) θεωρεί ότι είναι φυσιολογικό διαφορετικοί τύποι πλοίων να πλεύσουν με διαφορετικές ταχύτητες στην ίδια θαλάσσια περιοχή. Βάσει αυτού η έρευνα τους λαμβάνει υπόψη τα δεδομένα κίνησης και παύσης της κίνησης σε μία συγκεκριμένη περιοχή για να εξορύξουν ολόκληρες τροχιές φυσιολογικής κίνησης καθώς και αυθαίρετα σχήματα από περιοχές παύσης της κίνησης.

```
Procedure DBSCANSD(DatasetM, eps, MinPts, MaxDir, MaxSpd)
        Mark all points in moving dataset DatasetM as unclassified
        clusterList ← empty list
        for each unclassified point P in DatasetM do
                Mark P as classified
                neighborPts ← queryNeighborPoints (Datas-etM, P, eps, MinPts, MaxDir,
                 MaxSpd)
                if neighborPts is not NULL then
                    clusterList.add(neighborPts)
        for each cluster C in clusterList do
                for each cluster C' in clusterList do
                     if C and C' are different clusters then
                        if mergeClusters(C, C') is TRUE then
                            clusterList.remove(C')
          return clusterList

procedure QUERYNEIGHBORPOINTS(data,P,eps,MinPts,MaxDir,MaxSpd)
        cluster empty list
        for each point Q in data do
                if distance(P,Q) < eps then
                        if (P.SOG - Q.SOG) < MaxSpd then
                                if (P.COG - Q.COG) < MaxDir then
                                        cluster.add(Q)
        if cluster.size > MinPts then
             Mark P as core point
        return cluster
return NULL
procedure MERGECLUSTERS(clusterA,clusterB)
        merge FALSE
        for each point Q in clusterB do
            if point Q is core point and clusterA contains Q then
              merge ← TRUE
             for each point Q' in clusterB do
                    clusterA.add(Q')
                     break
            return merge
```

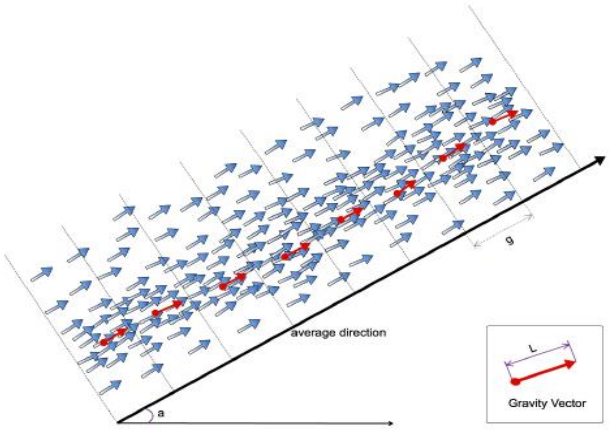
Πίνακας 2: ο αλγόριθμος DBSCANSD (Bo Liu, Erico N. de Souza, Stan Matwin, Marcin Sydowzy, 2014)

Ο παραπάνω αλγόριθμος επεκτείνει τη βασική ιδέα του αλγορίθμου DBSCAN, δηλαδή για κάθε σημείο ενός συμπλέγματος η περιοχή γειτνίασης σε συγκεκριμένη ακτίνα πρέπει να περιλαμβάνει τουλάχιστον έναν ελάχιστο αριθμό σημείων, ενσωματώνοντας δύο άλλους παράγοντες: α) τη μέγιστη διακύμανση της ταχύτητας (MaxSpd) και τη μέγιστη διακύμανση της κατεύθυνσης (MaxDir). Η βασική ιδέα πίσω από αυτό είναι ότι τα γειτονικά σημεία ενός άλλου σημείου πρέπει να είναι όχι μόνο αρκετά κοντά, αλλά και με παρόμοια Κατεύθυνση (Course Over Ground - COG) και Ταχύτητα (Speed Over Ground - SOG). Έτσι, ο ορισμός της παραμέτρου Eps-περιοχή γειτνίασης του DBSCAN τροποποιείται με τις ακόλουθες συναρτήσεις:

$$N_\varepsilon(p) = \{q \in D \,|Dist(p,q) < \varepsilon$$
$$|p.SOG - q.SOG| < MaxSpd$$
$$|p.COG - q.COG| < MaxDir$$

Dist(p,q) είναι η γεωγραφική απόσταση ανάμεσα στα σημεία *p*, και *q* αντί για την Ευκλείδεια απόσταση λαμβάνοντας υπόψη έτσι της καμπυλότητας της Γης. *DatasetM* είναι μια λίστα με όλα τα σημεία της κίνησης πάνω σε μία τροχιά. *eps* και *MinP* είναι η ελάχιστη εφικτή απόσταση και ο ελάχιστα εφικτός αριθμός σημείων μίας δεδομένης τροχιάς. *MaxDir* και *MaxSpd* είναι οι δύο νέες παράμετροι που αναφέραμε παραπάνω.

Μετά την εφαρμογή DBSCANSD στο σύνολο των δεδομένων με τις τροχιές των πλοίων, σημεία επί της τροχιάς τα οποία είναι σε κοντινή απόσταση με παρόμοια κατεύθυνση και ταχύτητα ομαδοποιούνται σε ένα σύμπλεγμα. Κατόπιν το αρχικά μεγάλο σύμπλεγμα διαχωρίζεται σε μικρότερα λαμβάνοντας υπόψη τις διαφορές σε ταχύτητα και κατεύθυνση. Επιπλέον, ο αλγόριθμος μπορεί να χειριστεί την επιτάχυνση ή επιβράδυνση του πλοίου ως αιτία δημιουργίας νέου συμπλέγματος, εφόσον η παράμετρος *MaxSpd* είναι σαφώς καθορισμένη. Η τεχνική αυτή ονομάζεται υπολογισμός Διανυσμάτων Βαρύτητας (Εικόνα 4). Όπως φαίνεται στην εικόνα, μπλε βέλη συμβολίζουν τα σημεία του συμπλέγματος, και τα κόκκινα βέλη είναι τα διανύσματα Βαρύτητας (GV). Το μήκος (L) του διανύσματος δίνεται από τη μέση τιμή της παραμέτρου SOG. Το πλάτος ενός πλέγματος g είναι το προκαθορισμένο μήκος για τη στεγανοποίηση των σημείων.



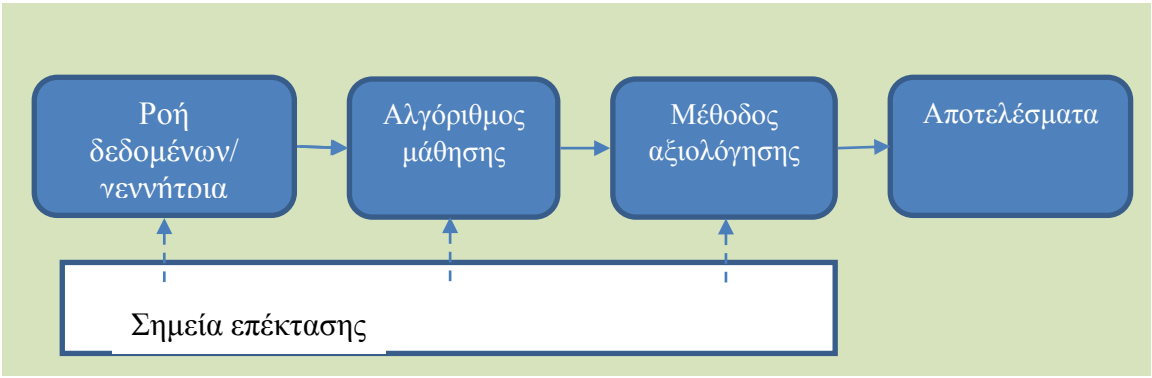
Εικόνα 4: τεχνική υπολογισμού Διανυσμάτων Βαρύτητας (Gravity Vector) (Bo Liu, Erico N. de Souza, Stan Matwin, Marcin Sydowzy, 2014)

Με τον ίδιο τρόπο που τα Διανύσματα Βαρύτητας χρησιμοποιούνται για τον διαχωρισμό των σημείων κίνησης σε συμπλέγματα, χρησιμοποιούνται αντιπροσωπευτικά σημεία παύσης της κίνησης (Sampled Stopping Points – SSP). Τα σημεία παύσης αρχικά ομαδοποιούνται από τον DBSCAN και κατόπιν ο παρακάτω αλγόριθμος τα διαχωρίζει σε επιμέρους ομαδοποιήσεις.

Σε ένα πείραμα στη θαλάσσια περιοχή του Λος Άντζελες και στο χρονικό διάστημα (11-12 του 2012), ο αλγόριθμος εκτελέστηκε σε ένα σύνολο από 327.694 εγγραφές (99.937 κινούμενα σημεία και 227.757 σημεία στάσης). Παρόμοια πειράματα πραγματοποιήθηκαν με μικρότερο όγκο ιστορικών δεδομένων σε άλλες περιοχές (Bo Liu, Erico N. de Souza, Stan Matwin, Marcin Sydowzy, 2014). Ωστόσο δεν αναφέρεται ο τρόπος υλοποίησης των αλγορίθμων καθώς επίσης δεν περιγράφεται κάποια προοπτική εφαρμογής της τεχνικής για την επεξεργασία ροών δεδομένων (streams) σε πραγματικό χρόνο.

2.3 R - MOA

Το MOA είναι μια εξειδικευμένη πλατφόρμα για εξόρυξη ροών δεδομένων. Αναπτύχθηκε από το Πανεπιστήμιο του Waikato, στη Νέα Ζηλανδία και έγινε γνωστή στην ερευνητική κοινότητα καθώς περιλαμβάνει μια συλλογή από αλγορίθμους μάθησης, που εξειδικεύονται στην ταξινόμηση, παλινδρόμηση, συσταδοποίηση, ανίχνευση ακραίων τιμών, την υλοποίηση συστημάτων στατιστικής ανάλυσης (Merino, J.A., 2015). Η πλατφόρμα έχει αναπτυχθεί σε java και διαχειρίζεται απλές ή σύνθετες ροές δεδομένων. Έχει έτοιμους πολλούς αλγορίθμους για συσταδοποίηση (clustering) και ταξινόμηση (classification) σε ροές δεδομένων, καθώς επίσης προσφέρει τη δυνατότητα στον χρήστη να υλοποιήσει τον δικό του αλγόριθμο σε java κάνοντας χρήση ειδικής διεπαφής (interface).
Το MOA είναι μία πλατφόρμα ανοιχτού κώδικα και ακολουθεί μία αρθρωτή σχεδίαση γεγονός που της επιτρέπει να φιλοξενήσει νέους αλγόριθμους και μετρικές (Εικόνα 6). Είναι ίσως το πιο ολοκληρωμένο πακέτο για εφαρμογή συσταδοποίησης (clustering) και ταξινόμησης (classification) αλγορίθμων σε ροές δεδομένων.



Επίσης οι (Philipp
να ενσωματωθούν
δεδομένων μπορούν να δημιουργηθούν χρησιμοποιώντας σύνθετες γεννήτριες, διαβάζοντας ARFF αρχεία, ενώνοντας πολλές ροές, ή φιλτράροντας ροές. Επιτρέπουν την προσομοίωση μίας δυνητικά άπειρης ακολουθίας δεδομένων. Διάφορες γεννήτριες είναι σήμερα διαθέσιμες (π.χ. για ένα πραγματικό σύνολο δεδομένων), σύνολο δεδομένων με περιορισμένες διαστάσεις, σύνολα δεδομένων που συγχωνεύονται.

2.3.1 R - Stream

Η γλώσσα R, είναι μία γλώσσα προγραμματισμού με πολλά πλεονεκτήματα και δυνατότητες στην ανάλυση δεδομένων και στη γραφική απεικόνιση. Παρέχει τη δυνατότητα να συνεργάζεται και με άλλες γλώσσες προγραμματισμού, όπως java, C/C++, Python και άλλες. Η βιβλιοθήκη "stream" της R παρέχει τη δυνατότητα να χρησιμοποιηθεί η βιβλιοθήκη της MOA και οι αλγόριθμοι που υπάρχουν έτοιμοι σε αυτό για τις ροές δεδομένων, με τη διεπαφή της R. Η παραπάνω βιβλιοθήκη είναι ένα ερευνητικό εργαλείο που περιλαμβάνει λειτουργίες μοντελοποίησης και προσομοίωσης ροών δεδομένων καθώς και ένα επεκτάσιμο πλαίσιο για την υλοποίηση, τη διασύνδεση και πειραματισμό με επιμέρους αλγορίθμους για διάφορες εργασίες εξόρυξης ροής δεδομένων. Το κύριο πλεονέκτημα της βιβλιοθήκης είναι ότι συνεργάζεται άψογα με τη μεγάλη υπάρχουσα υποδομή που παρέχεται από την R. Εκτός από δομές δεδομένων και συναρτήσεις που διευκολύνουν τον χειρισμό των δεδομένων, τη γραφική απεικόνιση και τις δυνατότητες σύνταξης νέων προγραμμάτων, η R παρέχει επίσης πολλούς υπάρχοντες αλγορίθμους εξόρυξης δεδομένων. Επιπλέον, επιτρέπει στους χρήστες να διασυνδέουν τον κώδικα τους με κώδικα γραμμένο σε άλλες γλώσσες (π.χ. C/C++, Java, Python).

Αρκετά πακέτα έχουν υλοποιηθεί και αξιοποιούν τις δυνατότητες της βιβλιοθήκης stream, για παράδειγμα, αρκετά πακέτα στατιστικής ανάλυσης μεγάλου όγκου δεδομένων όπως το factas, και το PCA. Το πακέτο birch υλοποιεί τον αλγόριθμο ομαδοποίησης BIRCH ενώ το πακέτο rEMM υλοποιεί ένα τυπικό αλγόριθμο ομαδοποίησης προσθέτοντας τη διάσταση του χρόνου στη δημιουργία των συστάδων. Η βιβλιοθήκη RMOA παρέχει όλες τις απαραίτητες ρουτίνες για την υλοποίηση σύνθετων ροών δεδομένων, και αλγορίθμων συσταδοποίησης. Οι (M. Hahsler, and M. Bolanos, 2016) εφαρμόσανε αυτή τη βιβλιοθήκη στην δημιουργία ομαδοποιήσεων για μεγάλα σύνολα δεδομένων. Παράλληλα, στα πειράματα τους λήφθηκαν υπόψη διαφορετικές μετρικές αξιολόγησης των αποτελεσμάτων (π.χ. silhouette, δείκτης Rand) (M. Hahsler, and M. Bolanos, 2016).

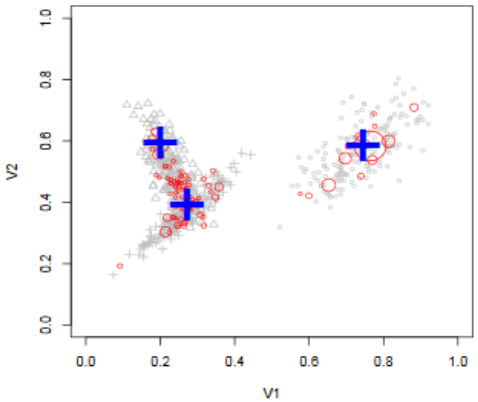
Όπως αναφέρθηκε παραπάνω (βλ. §2.1.1) η υλοποίηση των δυνατοτήτων της πλατφόρμας MOA στην R περιλαμβάνει το παρακάτω πλαίσιο υλοποίησης:

1. Δημιουργία ροής – λήψη δεδομένων από ένα ή περισσότερα CSV αρχεία
2. Εκτέλεση αλγορίθμου συσταδοποίησης (π.χ. DenStream) στα αρχικά σημεία, δημιουργία των πρώτων μικρό-ομάδων (micro-clusters).
3. Δημιουργία των μακρό-ομάδων (macro-clustering).
4. Φόρτωση του επόμενου CSV αρχείου – επανάληψη εκτέλεσης βημάτων 1 έως 3.

Στο παρακάτω παράδειγμα παρουσιάζεται η χρήση της βιβλιοθήκης για τη δημιουργία μίας στατικής ροής τυχαίων (Gaussian) δεδομένων και στη συνέχεια την εφαρμογή του MOA αλγορίθμου DenStream για τη δημιουργία των μικρο-ομάδων. Στη συνέχεια εφαρμόζεται ο αλγόριθμος K-means για την δημιουργία των μακρο-ομάδων.

```
> library(stream)

### define a data stream
> dsd <- DSD_GaussianStatic(k=3, d=2)
> get_points(dsd)

V1 V2
1 0.6051603 0.4822357

### cluster using DenStream
> micro <- DSC_DenStream(initPoints=100, horizon=1000)
> cluster(micro, dsd, 1000)

### recluster using weighted k-means
> macro <- DSC_Kmeans(k=3, weighted=TRUE)
> recluster(macro, micro)

### plot results
> plot(micro, dsd=dsd)
> points(get_centers(macro), col="blue", cex=5, lwd=2, pch="+")
```

Στη συνέχεια τα σημεία των μικρο- και μακρο-ομάδων προβάλλονται σε ένα γράφημα:




```
> evaluate(micro, dsd, n=100)
 recall     precision         f1        ssq       rand numCluster numClasses
 0.1000000  1.0000000  0.1818182  4.5551937  0.7776000 30.0000000  3.0000000
> evaluate(macro, dsd, n=100)
 recall     precision         f1        ssq       rand numCluster numClasses
 0.9876543  0.9912281  0.9894380  7.8401944  0.9874000  3.0000000  3.0000000
```

Εικόνα 5: γράφημα μικρο- και μακρο-ομάδων καθώς και προβολή μετρικών αξιολόγησης του αποτελέσματος της συσταδοποίησης (M. Hahsler, and M. Bolanos, 2016)

2.3.2 DenStream

Ο DenStream είναι ένας αλγόριθμος ομαδοποίησης κατά πυκνότητα. Σε αντίθεση με τους διαμεριστικούς αλγορίθμους που αποφασίζουν την δημιουργία μίας νέας ομάδας βάσει απόστασης, συνεπώς καταλήγουν σε σφαιρικού τύπου ομάδες, αλγόριθμοι όπως ο DenStream είναι η δημιουργία και επέκταση μίας ομάδας με βάση τη πυκνότητα των γειτονικών σημείων. Εφόσον η πυκνότητα ξεπεράσει ένα συγκεκριμένο όριο τότε προχωράει στη σύσταση νέας ομάδας. Για το λόγο αυτό προκύπτουν ομάδες με αυθαίρετο σχήμα. Το παραπάνω επιτρέπει το μη ορισμό εκ των προτέρων του αριθμού των ομάδων που θα δημιουργηθούν (João Henrique Alves Pereira, 2014).

Ο DenStream χρησιμοποιεί την έννοια των μικρο-ομάδων πυρήνων (core-micro-clusters) για να προσδιορίσει ομαδοποιήσεις σημείων που είναι σε κοντινή απόσταση μεταξύ τους και επιτρέπει την εκπόνηση μιας στρατηγικής αφαίρεσης τιμών που απομακρύνονται σημαντικά από το κέντρο της ομάδας χειρίζοντας τις ως ακραίες τιμές. Σύμφωνα με τον (Voigtlaender, 2013) το γεγονός ότι ο DenStream μπορεί να εντοπίζει αυθαίρετα σχήματα του δίνει σημαντικά πλεοκτήματα έναντι άλλων αλγορίθμων όπως ο CluStream που βρίσκει σφαιρικές ομαδοποιήσεις. Ο αλγόριθμος DenStream εκφράζει την ιδέα ότι τα πιο πρόσφατα δεδομένα που εισάγονται αντικατοπτρίζουν με μεγαλύτερη ακρίβεια την τρέχουσα κατανομή των δεδομένων διανομής από ότι τα παλαιότερα τα δεδομένα.

Μία μικρο-ομάδα πυρήνας στη χρονική στιγμή t ορίζεται ως CMC (w, c, r) για ένα αριθμό κοντινών σημείων X_i1…X_in στις χρονικές περιόδους T_i1…T_in ,

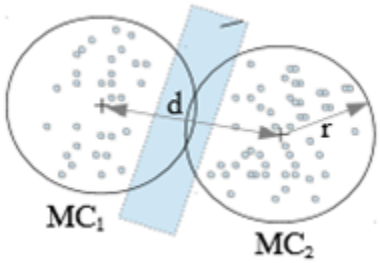
$w = \sum_{j=1}^{n} f\left(t - T_{i_j}\right), w > \mu$ είναι ο συντελεστής βαρύτητας της ομάδας

$c = \frac{\sum_{j=1}^{n} f\left(t - T_{i_j}\right) X_{i_j}}{w}$ είναι το κέντρο της ομάδας

$r = \frac{\sum_{j=1}^{n} f\left(t - T_{i_j}\right) dist\left(X_{i_j}, c\right)}{w}, r < \varepsilon$ είναι η ακτίνα της ομάδας.

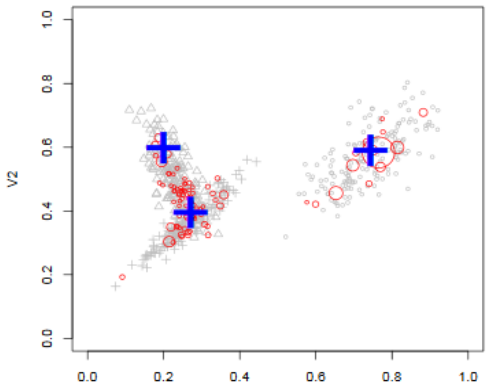
Η διάρκεια ζωής μίας ομάδας προκύπτει από τη σχέση (Voigtlaender, 2013):
$f(t) = 2^{-\lambda \cdot t}, \lambda > 0$, όπου λ είναι παράγοντας επιβράδυνσης στη χρονική στιγμή t.

Η υλοποίηση του αλγορίθμου διακρίνεται σε δύο μέρη: online και offline εκτέλεση (Πίνακας 3). Στο πρώτο μέρος (online) και κατά την αρχικοποίηση της διαδικασίας συσταδοποίησης χρησιμοποιείται ο DBSCAN. Ο αλγόριθμος συνοψίζει τα δεδομένα σε ένα σύνολο μικρο-ομάδων. Οι μικρο-ομάδες είναι αντιπρόσωποι συνόλων με παρόμοια σημεία δεδομένων και δημιουργούνται εκτελώντας το πρώτο πέρασμα πάνω από τα στοιχεία σε πραγματικό χρόνο και όταν εισέρχεται η ροή δεδομένων (online, βλ. §2.1.1). Οι μικρο-ομάδες διαθέτουν ως κύρια χαρακτηριστικά τους α) το κέντρο της συστάδας και β) πρόσθετα στατιστικά στοιχεία όπως ο συντελεστής βαρύτητας (πυκνότητα) και η διασπορά (variance). Κάθε νέο σημείο δεδομένων προστίθεται στη πλησιέστερη μικρο-συστάδα.



Εικόνα 6: μικρο-ομάδες (M. Hahsler, and M. Bolanos, 2016)

Στο online μέρος του αλγορίθμου δημιουργούνται οι μικρο-ομάδες και στο offline μέρος του αλγορίθμου δημιουργούνται οι τελικές ομάδες εφόσον ζητηθεί από τον χρήστη. Καθώς η ροή των δεδομένων μεταβάλλεται και ο αλγόριθμος υποδέχεται νέα σημεία, οι μικρο-ομάδες χωρίζονται σε p-μικρο-ομάδες και σε o-μικρο-ομάδες από μία διαδικασία συγχώνευσης. Η διαδικασία αυτή ελέγχει την ενσωμάτωση των νέων σημείων σε p-μικρο-ομάδες και όσα σημεία περισσεύουν ενσωματώνονται σε o-μικρο-ομάδες. Οι τελικές ομάδες δημιουργούνται χρησιμοποιώντας μία παραλλαγή του DBSCAN επάνω στις p-μικρο-ομάδες.



Εικόνα 7: εκτέλεση DenStream – δημιουργία micro και macro-ομάδων (M. Hahsler, and M. Bolanos, 2016)

Ο εντοπισμός των ομάδων σε μία ροή δεδομένων που ενημερώνεται συνεχώς βασίζεται στην δημιουργία και ενημέρωση μίας συστοιχίας p-μικρο-ομάδων και o-μικρο-ομάδων. Οι o-μικρο-ομάδες αποθηκεύονται σε ένα ενταμιευτή ακραίων τιμών στη μνήμη του υπολογιστή. Αυτό έχει να κάνει με τη παρατήρηση ότι τα περισσότερα από τα νέα σημεία ανήκουν ήδη σε υπάρχουσες ομάδες και επομένως μπορούν να απορροφηθούν από τις p-μικρο-ομάδες. Όταν εισέρχεται ένα νέο σημείο p εκτελείται ο αλγόριθμος που περιγράφεται παρακάτω, όπου $T_p = [\frac{1}{\lambda}\log\left(\frac{\beta\mu}{\beta\mu-1}\right)]$ ορίζει τη χρονική διάρκεια ζωής μίας p-μικρο-ομάδας. Με άλλα λόγια, ελέγχουμε για την ύπαρξη ενός νέου σημείου p κάθε χρόνο T_p. Αυτή η στρατηγική δίνει το

πλεονέκτημα της αποθήκευσης στη μνήμη ενός μέγιστου αριθμού από p-μικρο-ομάδες, της τάξης των $\frac{W}{\beta\mu}$, καθώς ο συνολικός συντελεστής βάρους της ροής δεδομένων ορίζεται από τη σταθερά W. Επιπλέον, στις χρονικές στιγμές T_p συγκρίνεται το βάρος της κάθε ο-μικρο-ομάδας με το χαμηλότερο δυνατό βάρος (συμβολίζεται με ξ) και ορίζεται ως εξής:

$$\xi = \frac{2^{-\lambda(t-t_0+T_p} - 1}{2^{-\lambda T_p - 1}}$$

Αλγόριθμος στο MOA	**Denstream**
Τύπος	Αλγόριθμος συσταδοποίησης βάσει πυκνότητας
Πεδίο Τιμών	Αριθμητικό
Τύπος συσταδοποίησης	Στιγμιότυπα τιμών
Δομή δεδομένων	Διανύσματα με χαρακτηριστικά προς συσταδοποίηση
Μοντέλο χρονικού παραθύρου	Damped /fading χρονικό παράθυρο
Παράμετροι	- ε (τα σημεία γειτνίασης γύρω από υπο-ομάδες) - β (άκρα τιμών) - μ (συντελεστές υπο-ομάδων) - initPoints (αρχικά σημεία) - offline (πολλαπλασιαστής του ε) - λ (καθορίζει τον παράγοντα διάσπασης) - processingSpeed (αριθμός σημείων προς επεξεργασία στη δεδομένη χρονική στιγμή)
Αλγόριθμος συσταδοποίησης	DBSCAN
Σχήμα της κάθε συστάδας	Αυθαίρετο

Περιγραφή αλγορίθμου DenStream	$T_p = [\frac{1}{\lambda}\log\left(\frac{\beta\mu}{\beta\mu - 1}\right)]$
	
	Πάρε το επόμενο σημείο p στη δεδομένη χρονική στιγμή t από τη πηγή (ροή) δεδομένων DS
	Εκτέλεση διαδικασίας: Merging (p)
	Αν (t % Tp) =0 τότε {
	Για κάθε p-micro-cluster cp {
	Αν ο συντελεστής βαρύτητας (wp) του cp < βμ τότε {
	Διαγραφή του cp
	}
	}
	Για κάθε o-micro-cluster co {
	$\xi = \frac{2^{-\lambda(t-t_0+T_p-1}}{2^{-\lambda T_p - 1}}$
	Αν ο συντελεστής βαρύτητας (w0) του c0 < ξ τότε {
	Διαγραφή του c0
	}
	}
	}
	Δημιουργία ομάδων για κάθε νέο αίτημα δημιουργίας ομάδων

Περιγραφή διαδικασίας Merging (p)	Προσπάθεια συγχώνευσης του σημείου p στο κοντινότερο p-micro-cluster cp
	Αν rp (η νέα ακτίνα του cp) < ε τότε {
	Συγχώνευση του p στο cp
	}
	Διαφορετικά {
	Προσπάθεια συγχώνευσης του p στο κοντινότερο o-micro-cluster co
	Αν ro < ε τότε {

	Συγχώνευση του p στο co
	Αν wo > βμ τότε {
	Αφαίρεσε το co από τον ενταμιευτή ακραίων τιμών και δημιούργησε ένα νέο p-micro-cluster με το co
	}
	}
	Διαφορετικά {
	δημιούργησε ένα νέο o-micro-cluster με το p και πρόσθετο στον ενταμιευτή ακραίων τιμών
	}
	}

Πίνακας 3: Σύνοψη αλγορίθμου DenStream

Αντίστοιχα για ροές δεδομένων που εξελίσσονται μέσα στο χρόνο, οι (Feng Cao and Martin Ester and Weining Qian and Aoying Zhou, 2006) προτείνουν επίσης τη χρήση αλγορίθμου συσταδοποίησης βάσει πυκνότητας και προτείνεται επίσης ο DenStream. Με βάση τον συγκεκριμένο αλγόριθμο υλοποιούν μία αντίστοιχη διαδικασία micro και macro clustering όπως αυτή που περιγράφεται παραπάνω. Επιπλέον, θεωρούν ότι σε μία ροή δεδομένων που συνεχώς εξελίσσεται δεν μπορούν όλα τα σημεία –δεδομένα να παραμένουν στη μνήμη εξαιτίας των περιορισμών στην διαθέσιμη μνήμη. Αντί αυτού, βασίζονται στην εισαγωγή της έννοιας του core-micro-cluster (c-micro-cluster) για την αναπαράσταση αντιπροσωπευτικών και όχι όλων των σημείων – δεδομένων.

2.3.3 CluStream

Ο CluStream είναι ένας άλλος διαμεριστικός αλγόριθμος στη βιβλιοθήκη του MOA ο οποίος έχει επίσης 2 συνιστώσες α) δημιουργίας μικρο-ομάδων και β) δημιουργίας μακρο-ομάδων όπως ο DenStream. Ο CluStream ορίζει τις μικρο-ομάδες ως εξής (João Henrique Alves Pereira, 2014):

Για μία πλειάδα σημείων $X_{i1} \dots X_{in}$ και χρονικές περιόδους $T_{i1} \dots T_{in}$, μία μικρο-ομάδα προκύπτει από τη σχέση:

$$(2 \cdot d + 3) \text{ tuple } (\overline{CF2^x}, \overline{CF1^x}, CF2^t, CF1^t, n)$$

$\overline{CF2^x}$	είναι ένα διάνυσμα μεγέθους d το οποίο περιέχει το άθροισμα των τετραγώνων των σημείων για κάθε διάσταση
$\overline{CF1^x}$	είναι ένα διάνυσμα μεγέθους d το οποίο περιέχει το άθροισμα των τιμών –δεδομένων σε κάθε διάσταση
$CF2^t$	είναι το άθροισμα των τετραγώνων των χρονικών ορόσημων
$CF1^t$	είναι το άθροισμα των χρονικών ορόσημων
n	είναι ο αριθμός των σημείων σε μία μικρο-ομάδα

Κατά την άφιξη ενός νέου σημείου, προστίθεται σε μία από τις διαθέσιμες μικρο-ομάδες. Εφόσον ξεπεράσει το μέγιστο όριο της κοντινότερης ομάδας τότε προστίθεται σε νέα ομάδα. Αν πρέπει να γίνει αυτό μειώνεται ο αριθμός των ομάδων κατά μία είτε μέσω διαγραφής ή μέσω συγχώνευση.

Αλγόριθμος στο MOA	**Clustream**
Τύπος	Διαμεριστικός Αλγόριθμος
Πεδίο Τιμών	Αριθμητικό
Τύπος συσταδοποίησης	Στιγμιότυπα τιμών
Δομή δεδομένων	Διανύσματα με χαρακτηριστικά προς συσταδοποίηση
Μοντέλο χρονικού παραθύρου	Damped /fading χρονικό παράθυρο
Παράμετροι	- m: ορίζει το μέγιστο αριθμό μικρο-ομάδων που επιτρέπεται να δημιουργηθούν στον αλγόριθμο - horizon: ορίζει το χρονικό παράθυρο που χρησιμοποιείται στον αλγόριθμο - t: Κατά τη προσθήκη ενός νέου σημείου σε μία μικρο-ομάδα η συγκεκριμένη παράμετρος ορίζεται ως η μέγιστη επιτρεπόμενη απόκλιση της ρίζας μέσου τετραγώνου των τιμών μίας μικρο-ομάδας από το κέντρο της ομάδας - k: ο αριθμός των μακρο-ομάδων που μπορούν να δημιουργηθούν εφαρμόζοντας τον αλγόριθμο k-means με βάρη. Αν δεν οριστεί η τιμή αυτή τότε δεν εκτελείται επανα-συσταδοποίηση.
Σχήμα της κάθε συστάδας	σφαιρικό

Ένα βασικό μειονέκτημα του αλγορίθμου είναι ότι απαιτεί τον ορισμό του μέγιστου αριθμού μικρο-ομάδων. Σε περιπτώσεις που η πηγή δεδομένων εμφανίζει μια ανομοιόμορφη ροή ο λανθασμένος ορισμός της συγκεκριμένης μεταβλητής μπορεί να οδηγήσει σε λανθασμένη ομαδοποίηση.

2.3.4 ClusTree

Ο ClusTree παρέχει στρατηγικές για τον χειρισμό χρονικών περιορισμών κατά την πραγματοποίηση ομαδοποίησης σε οποιαδήποτε χρονική στιγμή καθώς και της διακοπής της διαδικασίας εισαγωγής νέων αντικειμένων στο δέντρο σε οποιαδήποτε χρονική στιγμή προκειμένου να γίνουν ανακατανομές προς βελτίωση της ομαδοποίησης. Ο αλγόριθμος δεν κάνει κάποια υπόθεση για τον αριθμό των ομάδων που μπορούν να δημιουργηθούν καθώς οι

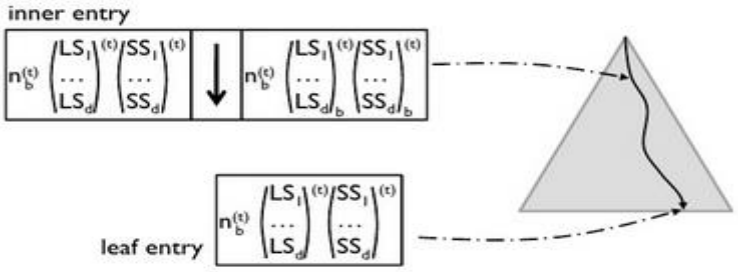
λειτουργίες διάσπασης και ένωσης που διαθέτει προσαρμόζουν το μέγεθος του μοντέλου αυτόματα.

Ο αλγόριθμος εκτελείται με βάση δύο στρατηγικές, της γρήγορης και της αργής εκτέλεσης. Στη γρήγορη εκτέλεση ο αλγόριθμος ενώνει σε ομάδες όμοια αντικείμενα για να δημιουργήσει άμεσα τις νέες ομάδες. Στην αργή εκτέλεση και αφού εισαχθεί ένα νέο αντικείμενο στο δέντρο, πραγματοποιεί επανέλεγχο και τυχόν διορθώσεις αν χρειαστεί βελτιώνοντας τη ποιότητα της ομαδοποίησης.

Οι μικρο-ομάδες ορίζονται ιεραρχικά ως CF= (N, LS, SS). Νέα αντικείμενα (X1…Xd) εισάγονται σε νέο εσωτερικό κόμβο για τον οποίο ισχύει:

$$N = N + 1, \quad LS_i = LS_i + x_i, \quad SS_i = SS_i + (x_i)^2$$

Αντικείμενα τα οποία πρέπει να τοποθετηθούν δυναμικά αποθηκεύονται προσωρινά σε ένα ενταμιευτή για ενδιάμεσες προσθήκες. Ο αλγόριθμος αναζητά τον ενδιάμεσο ή τελικό κόμβο-ομάδα στον οποίο τοποθετεί το κάθε αντικείμενο:



Εικόνα 8: ομαδοποίηση βάσει του αλγορίθμου ClusTree

Αλγόριθμος στο MOA	**ClusTree**
Τύπος	Αλγόριθμος δυναμικής συσταδοποίησης
Πεδίο Τιμών	Αριθμητικό
Τύπος συσταδοποίησης	Στιγμιότυπα τιμών
Δομή δεδομένων	Διανύσματα με χαρακτηριστικά προς συσταδοποίηση
Μοντέλο χρονικού παραθύρου	Damped /fading χρονικό παράθυρο
Παράμετροι	- horizon: ορίζει το χρονικό παράθυρο που χρησιμοποιείται στον αλγόριθμο - maxHeight: ορίζει το μέγιστο ύψος του δέντρου - lambda: αντικαθιστά τον αυτόματα υπολογισμένο παράγοντα επιβράδυνσης (decay) της συσταδοποίησης
Σχήμα της κάθε συστάδας	Δέντρο

2.3.5 Μετρικές απόστασης

Στο πλαίσιο της ομαδοποίησης και των αλγορίθμων που αναφέρθηκαν παραπάνω απαιτείται η εφαρμογή συγκεκριμένων μετρικών απόστασης για τη σύγκριση της ομοιότητας μεταξύ των αντικειμένων καθώς και τη δημιουργία των ομάδων. Η επιλογή της κατάλληλης μετρικής εξαρτάται από το τύπο των δεδομένων που τίθενται σε επεξεργασία (João Henrique Alves Pereira, 2014).

Η τυπική μετρική μίας απόστασης είναι η Ευκλείδεια απόσταση. Η μετρική αυτή ορίζει την απόσταση δύο αντικειμένων X1, και X2 τα οποία αποτελούνται από n ιδιότητες:

$$dist(X_1, X_2) = \sqrt{(X_{1_1} - X_{1_2})^2 + \cdots + (X_{1_n} - X_{2_n})^2}$$

Μία άλλη αντίστοιχη μετρική είναι η Μανχάταν (Manhattan) απόσταση. Ορίζει την απόσταση σε περιοχές ανάμεσα σε 2 σημεία μέσα στην ίδια πόλη:

$$dist(X_1, X_2) = |X_{1_1} - X_{1_2}| + \cdots + |X_{1_n} - X_{2_n}|$$

Επιπλέον ορίζεται η Minkowski απόσταση ως μία γενίκευση των δύο παραπάνω μετρικών:

$$dist(X_1, X_2) = \sqrt[p]{|X_{1_1} - X_{1_2}|^p + \cdots + |X_{1_n} - X_{2_n}|^p}, p > 1$$

Όταν p=1 τότε αντιπροσωπεύει την Μανχάταν (Manhattan) απόσταση. Όταν p=2 τότε αντιπροσωπεύει την Ευκλείδεια απόσταση.

Τέλος ορίζεται η Chebysev απόσταση ως γενίκευση της Minkowski απόστασης για p = ∞.

$$dist(X_1, X_2) = \lim_{p \to \infty} \left(\sum_{j=1}^{n} |X_{1_j} - X_{2_j}|^p \right)^{\frac{1}{p}} = max_j |X_{1_j} - X_{2_j}|$$

2.4 Ροές δεδομένων κινούμενων οχημάτων

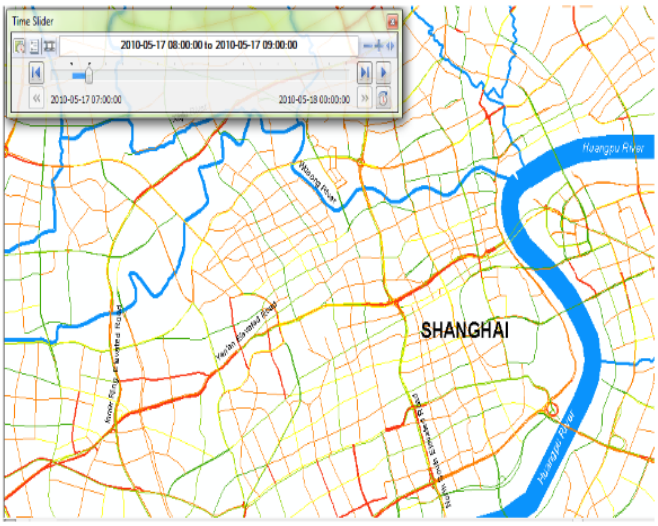
Η κυκλοφοριακή συμφόρηση είναι ένα σοβαρό πρόβλημα προκαλώντας οικονομικές απώλειες και υποβιβασμό της ποιότητας ζωής εξαιτίας των απρόβλεπτων χρόνων ταξιδιού. Τα ακινητοποιημένα οχήματα παράγουν αυξημένο επίπεδο αποτύπωμα άνθρακα ενώ αυξάνεται επίσης ο αριθμός των ατυχημάτων ως άμεσο αποτέλεσμα του άγχους και της κούρασης των οδηγών που έχουν κολλήσει στη κίνηση. Καθώς η επέκταση των δρόμων είναι πολυδάπανη η σύγχρονη επιστήμη μελετά τρόπους αξιοποίησης των ροών δεδομένων που αφορούν τα κινούμενα οχήματα (Floating Car Data – FCD) (Győző, 2015).

Ο DBSCAN και γενικότερα οι αλγόριθμοι χωρο-χρονικής συσταδοποίησης εφαρμόζονται σε ένα πρώτο στάδιο (pre-processing) επεξεργασίας της πληροφορίας σε ένα πεδίο εφαρμογής big data (π.χ. ανάλυση των δεδομένων της κυκλοφορίας σε μία πόλη, της κίνησης των δεδομένων σε ένα δίκτυο κινητής τηλεφωνίας 3G/4G, ελέγχου της θαλάσσιας κυκλοφορίας κλπ). Η εξαγωγή επιπλέον συμπερασμάτων ή η πραγματοποίηση προβλέψεων (π.χ. εντοπισμός σημείων με τα μεγαλύτερο ενδιαφέρον σε μία πόλη, επίπεδα συμφόρησης) μπορεί να πραγματοποιηθεί σε ένα επόμενο στάδιο. Για παράδειγμα, χώρο-χρονικά μορφότυπα σχετικά με δεδομένα συμφόρησης μπορούν να εξαχθούν μέσω γραφικής απεικόνισης και ανάλυσης (χαρτογράφηση πυκνότητας σημείων και της τροχιάς τους, απεικόνιση του επιπέδου κυκλοφοριακής συμφόρησης στο δίκτυο οδοστρώματος). Άλλες τεχνικές που μπορούν να εφαρμοστούν, περιλαμβάνουν (Győző, 2015):

1. Στατιστικές μετρικές ροών δεδομένων με κατεύθυνση σε πλέγμα (grid-based directional flow statistics).
2. Ανίχνευση συμφόρησης σε μία διαδρομή γνωρίζοντας την κατεύθυνση (Directional congestion detection).
3. Μετρικές που βασίζονται στην αποθήκευση και εκτέλεση ερωτημάτων σε βάση δεδομένων (SQL-based implementation).

Η παραπάνω ανάλυση μπορεί να συμβάλλει στην εξαγωγή συμπερασμάτων όπως (C. Kartika, 2015):

1. Ποια είναι η περίοδος αιχμής, δηλαδή τα χρονικά σημεία της ημέρας με τη μεγαλύτερη κίνηση στη πόλη;
2. Ποια είναι η χρονική διάρκεια της συμφόρησης;
3. Σε ποιες περιοχές (δρόμους) ισχύουν τα παραπάνω με τη μεγαλύτερη συχνότητα;



Εικόνα 9: συσταδοποίηση σημείων (οχημάτων) στους δρόμους της Σαγκάης – δυναμική ενημέρωση – διαστήματα αιχμής (C. Kartika, 2015)

Η μέση ταχύτητα όλων των σημείων – οχημάτων σε κάθε χρονική περίοδο μπορεί να υπολογιστεί από τα δεδομένα συγκρίνοντας τα κυκλοφοριακά δεδομένα κάθε χρονικής περιόδου με την επόμενη περίοδο (π.χ. η απόσταση που διένυσε το ίδιο όχημα σε ένα

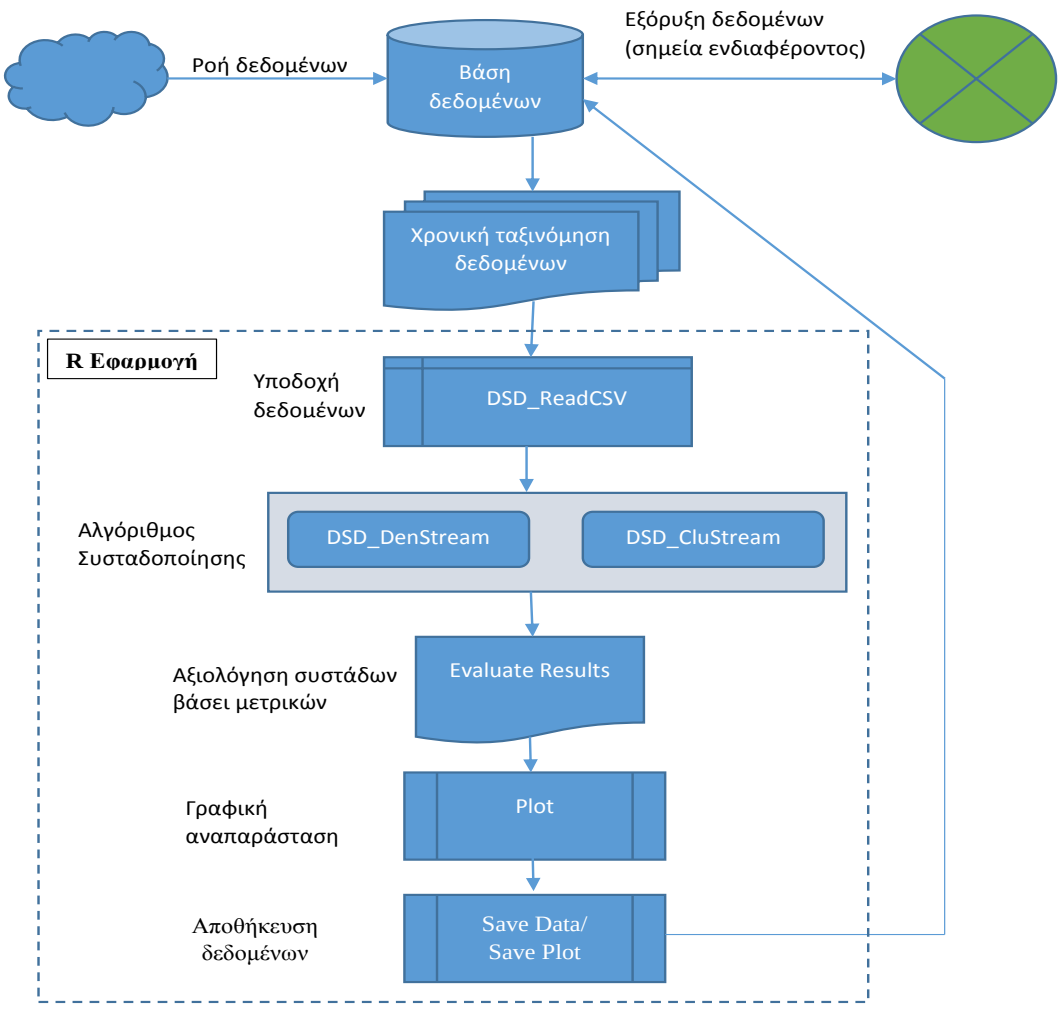
διάστημα 5 λεπτών ορίζει τη ταχύτητα) δημιουργώντας έτσι τη συνολική κυκλοφοριακή τάση σε μία πόλη. Στις περιόδους αιχμής η μέση ταχύτητα είναι χαμηλότερη από τις άλλες περιόδους πράγμα που σημαίνει ότι είναι πιο πιθανό να εντοπιστεί συμφόρηση σε αυτές τις περιόδους.

Κ Ε Φ Α Λ Α Ι Ο 3

3. Εφαρμογή αλγορίθμων συσταδοποίησης βάσει πυκνότητας

3.1 Αρχιτεκτονική

Αντικείμενο αυτής της διπλωματικής εργασίας είναι η υλοποίηση μίας εφαρμογής η οποία επιτρέπει τη συσταδοποίηση ροών δεδομένων τα οποία είναι διαθέσιμα σε πραγματικό χρόνο ή όχι και με αυξημένη πολυπλοκότητα. Για την υλοποίηση της αρχιτεκτονικής της προτεινόμενης εφαρμογής θα αξιοποιηθούν οι βιβλιοθήκες της γλώσσας επίλυσης προβλημάτων εξόρυξης δεδομένων και στατιστικής ανάλυσης R (Forrest J., 2011).



Εικόνα 10: Αρχιτεκτονική Προτεινόμενης Λύσης

Η Εικόνα 12 παρουσιάζει την αρχιτεκτονική της προτεινόμενης λύσης. Σε αυτή διακρίνονται τρία κύρια συστατικά μέρη:

1. Βάση δεδομένων: η ροή δεδομένων που λαμβάνονται από διάφορα μέρη (π.χ. GPS δεδομένα) και η αποθήκευση τους στη βάση δεδομένων. Αυτό επιτρέπει την χρονική ταξινόμηση των δεδομένων. Για παράδειγμα, η αρχιτεκτονική αυτή επιτρέπει τον επιμερισμό των δεδομένων σε διαστήματα (π.χ. δεδομένα που είναι διαθέσιμα τα τελευταία 5 λεπτά της ώρας, την τελευταία μισή ή μία ώρα, κλπ). Σημειώνεται ότι η βάση δεδομένων δεν αποτελεί συστατικό της R εφαρμογής. Η ύπαρξη της βοηθάει την εκτέλεση του R προγράμματος καθώς δεν έχουμε τους απαραίτητους υπολογιστικούς πόρους. Η R εφαρμογή δεν απαιτεί για την εκτέλεσή της την βάση δεδομένων.
2. Εφαρμογή R: η εφαρμογή ενεργοποιεί τις απαραίτητες βιβλιοθήκες για την υποδοχή των δεδομένων ανά χρονικά διαστήματα και την εκτέλεση του επιλεγμένου αλγορίθμου συσταδοποίησης βάσει των χωρικών δεδομένων. Ο εκάστοτε αλγόριθμος παραμετροποιείται για την αντιμετώπιση της πολυπλοκότητας των συγκεκριμένων δεδομένων. Τα αποτελέσματα (γεωγραφικά σημεία) απεικονίζονται γραφικά πάνω

στο χάρτη ως σημεία ενδιαφέροντος (π.χ. μιας πόλης). Τα αποτελέσματα εξάγονται σε αρχεία και επιστρέφουν στη βάση δεδομένων για περαιτέρω εκμετάλλευση.

3. Εξόρυξη δεδομένων: τα αποτελέσματα της συσταδοποίησης είναι διαθέσιμα στη βάση δεδομένων προκειμένου εφαρμογές τρίτων να μπορούν να εξορύξουν δεδομένα ως προς τα σημεία ενδιαφέροντος που παρουσιάζουν τη μεγαλύτερη επισκεψιμότητα ή περιοχές που παρουσιάζουν χαμηλές ταχύτητες των κινούμενων οχημάτων σε δεδομένα σημεία του χρόνου.

Τα παραπάνω συστατικά μέρη παρουσιάζονται αναλυτικά παρακάτω.

3.2 Βάση δεδομένων

3.2.1 Ροή δεδομένων

Για την υλοποίηση της παραπάνω αρχιτεκτονικής εφαρμόσαμε ως μελέτη περίπτωσης ένα σύνολο GPS δεδομένων της Microsoft από τη κυκλοφορία 10.357 ταξί στο Πεκίνο (Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang, 2010). Αφορά δείγμα με δεδομένα κυκλοφορίας στη διάρκεια μίας εβδομάδας. Το αρχικό σύνολο των δεδομένων ήταν περίπου 15 εκατομμύρια σε μία συνολική απόσταση τροχιάς που φτάνει τα 9 χιλιόμετρα. Από τα 15 εκατομμύρια που φορτώθηκαν στη βάση δεδομένων αξιοποιήθηκαν οι 10.022.470 εγγραφές καθώς οι υπόλοιπες είτε περιείχαν σφάλματα ή ήταν εκτός των συνόρων του Πεκίνου.

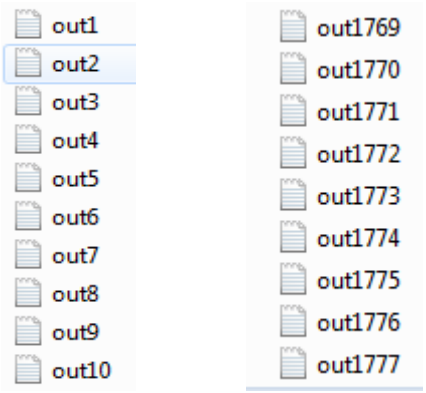
Τα δεδομένα με τις συντεταγμένες των ταξί του Πεκίνου ήταν διαθέσιμα σε txt μορφή που περιείχαν εγγραφές με τα ακόλουθα στοιχεία:

1. Μοναδικό αναγνωριστικό ταξί.
2. Ημερομηνία και ώρα με ακρίβεια δευτερολέπτου.
3. Συντεταγμένες.

Εξαιτίας των περιορισμένων υπολογιστικών πόρων που ήταν διαθέσιμοι για την εκτέλεση της εφαρμογής θεώρηθηκε σκόπιμη η σταδιακή ροή των δεδομένων προς επεξεργασία (ομαδοποίηση) είτε από ένα μεγάλο αρχείο ή από μικρότερα αρχεία τα οποία θα είναι χρονικά ταξινομημένα. Προτιμήθηκε η δεύτερη μέθοδος καθώς με τη διαίρεση σε μικρότερα αρχεία (π.χ. σε διάστημα 5λεπτων) θα μπορεί να ελεγχθεί η εκτέλεση του προγράμματος και η βέλτιση χρήση των υπολογιστικών πόρων.

Συνεπώς, αρχικά τα δεδομένα αυτά εισαχθήκαν στη βάση δεδομένων της αρχιτεκτονικής (MySQL) (Εικόνα 12). Στη συνέχεια εκτελώντας συγκεκριμένα ερωτήματα στη βάση υλοποιήθηκε η χρονική ταξινόμηση των δεδομένων και η εξαγωγή τους σε αρχεία κειμένου (txt) ανά χρονικό διάστημα πέντε λεπτών με τις ακόλουθες εγγραφές:

1. Μοναδικό αναγνωριστικό ταξί.
2. Συντεταγμένες.




```
7659    2008-02-02 13:30:45     116.35203       39.84086
2608    2008-02-02 13:30:47     115.96655       40.45673
4447    2008-02-02 13:30:51     116.32862       39.97532
719     2008-02-02 13:30:52     116.34647       40.00016
5132    2008-02-02 13:30:52     116.38024       39.85011
4447    2008-02-02 13:30:56     116.32832       39.97535
2417    2008-02-02 13:30:57     116.40153       39.99036
8769    2008-02-02 13:30:57     116.40037       39.90720
3443    2008-02-02 13:30:58     116.32132       39.97999
```

Εικόνα 11: λίστα χρονικά ταξινομημένων αρχείων – περιεχόμενο κάθε αρχείου

Η βάση δεδομένων δεν είχε άλλο σημαντικό ρόλο στην εφαρμογή των αλγορίθμων ομαδοποίησης.

Κάθε txt αρχείο περιέχει όλα τα ταξί του Πεκίνου για εκείνο το χρονικό διάστημα. Το αρχικό σύνολο δεδομένων αποτελείτο από txt αρχεία ανά ταξί , δηλαδή σημεία για ένα ταξί σε κάθε αρχείο, που εμπεριείχαν τις συντεταγμένες και το χρόνο. Παρατηρήθηκε ότι η αρχική δειγματοληψία ήταν ανά πέντε λεπτά, δηλαδή οι συντεταγμένες του κάθε ταξί ανανεώνονταν ανά πέντε λεπτά. Η βάση δεδομένων χρησιμοποιήθηκε για να συγκεντρωθεί μαζί η πληροφορία για όλα τα ταξί και έπειτα τα επιμέρους txt αρχεία να περιέχουν την πληροφορία για όλα τα ταξί ανά πεντάλεπτο. Τα παραγόμενα txt αρχεία αποτέλεσαν την κύρια πηγή ροής δεδομένων στην εφαρμογή της R που περιγράφεται παρακάτω.

3.3 Εφαρμογή R

Για την υλοποίηση της εφαρμογής χρησιμοποιήθηκε η γλώσσα προγραμματισμού R για την ομαδοποίηση των δεδομένων και την παραγωγή των αποτελεσμάτων. Η έκδοση της R που χρησιμοποιήσαμε ήταν η 3.3.1.

Η R είναι μια γλώσσα προγραμματισμού που χρησιμοποιείται για στατιστικούς υπολογισμούς καθώς και για την δημιουργία γραφημάτων. Συνεπώς, είναι ιδανική για το αντικείμενο της διπλωματικής εργασίας. Στην υλοποίησή χρησιμοποιήσαμε τις ακόλουθες βιβλιοθήκες (πακέτα) της R:

1. **stream**: Περιέχει λειτουργικότητα για εξόρυξη δεδομένων από ροές (data stream mining).
2. **mvtnorm**: Έχει πολυμεταβλητές, κανονικές και t κατανομές.
3. **streamMOA**: Έχει αλγόριθμους ομαδοποίησης που χρησιμοποιούνται για μαζική ανάλυση σε πραγματικό χρόνο (MOA, Massive Online Analytics) ροών.
4. **ggmap**: Περιέχει λειτουργικότητα για παρουσίαση γεωγραφικών δεδομένων.
5. **gtools**: Έχει διάφορα προγραμματιστικά εργαλεία της R.
6. **cluster**: Χρησιμοποιείται για την εύρεση ομάδων σε δεδομένα.

Η εφαρμογή υλοποιήθηκε σύμφωνα με το πλαίσιο της προτεινόμενης αρχιτεκτονικής (Εικόνα 12) με τις ακόλουθες λειτουργίες:

1. Διάβασμα αρχείων – Υποδοχή ροών δεδομένων.
2. Δημιουργία συστάδων (clustering) – παραμετροποίηση επιλεγμένου αλγορίθμου.
3. Προσωρινή αποθήκευση αποτελεσμάτων στη μνήμη.
4. Προβολή μετρικών αξιολόγησης αποτελεσμάτων.
5. Προβολή σημείων στο χάρτη – σύγκριση με το κεντρικό σημείο της κάθε ομάδας.
6. Αποθήκευση χαρτών – μετρικών.

3.3.1 Βιβλιοθήκη Stream - MOA

Η βασική λειτουργία που εκτελεί η εφαρμογή R που έχουμε αναπτύξει είναι η ομαδοποίηση δεδομένων. Για αυτή την λειτουργία έχουμε βασιστεί στη βιβλιοθήκη της R **streamMOA**. Χρησιμοποιήσαμε την μέθοδο **DSC_DenStream** για τη παραμετροποίηση και εκτέλεση του DenStream αλγορίθμου (βλ. §2.3.2). Αντίστοιχα χρησιμοποιήσαμε τη μέθοδο **DSC_CluStream** για τη παραμετροποίηση και εκτέλεση του CluStream αλγορίθμου (βλ. §2.3.3).

Η δημιουργία των ομάδων είτε με τον DenStream ή με τον CluStream δεν θα μπορούσε να επιτευχθεί χωρίς την χρήση των ακόλουθων συναρτήσεων από το R πακέτο **stream**:

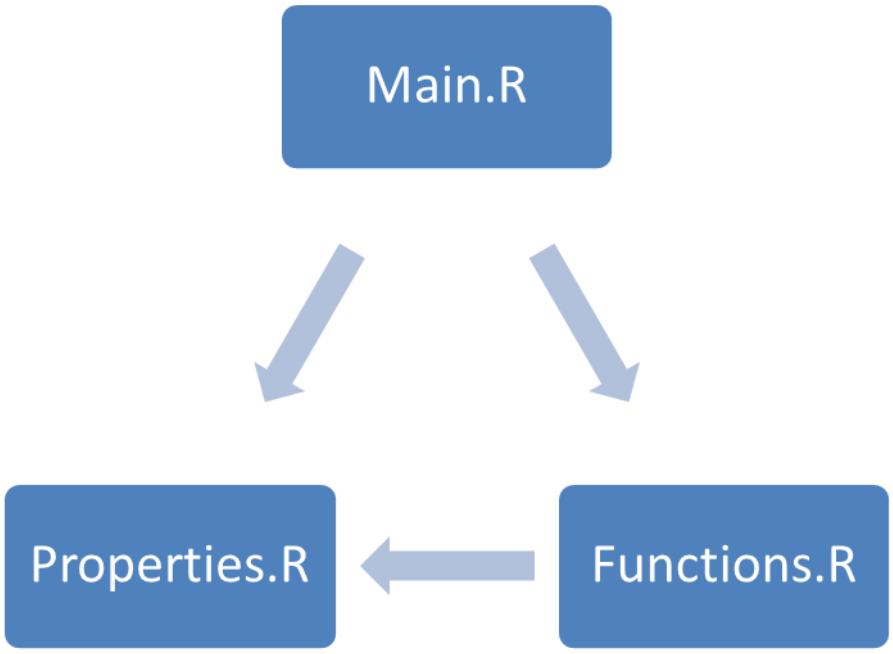
1. **get_points**: επιστρέφει τα σημεία από ένα DSD (Data Stream Data) αντικείμενο.
2. **get_assignment**: επιστρέφει την ανάθεση / τοποθέτηση των δεδομένων / σημείων σε ομάδες σε ένα DSC (Data Stream Cluster) αντικείμενο.
3. **DSD_ReadCSV**: επιστρέφει ένα DSD αντικείμενο από ένα CSV αρχείο.
4. **reset_stream**: επαναφέρει τον μετρητή σε ένα DSD αντικείμενο στην αρχή ή σε οποιδήποτε άλλη θέση στη ροή δεδομένων.
5. **update**: ενημερώνει ένα μοντέλο ομαδοποίησης ομαδοποιώντας έναν αριθμό δεδομένων / σημείων εισόδου από μια ροή δεδομένων σε ένα ομαδοποιημένο αντικείμενο.
6. **evaluate**: επιστρέφει μετρικές αξιολόγησης για μίκρο και μάκρο ομαδοποιήσεις από ένα DSD αντικείμενο δεδομένου του αρχικού DSD αντικειμένου.
7. **recluster**: χρησιμοποιεί έναν αλγόριθμο μάκρο-ομαδοποίησης για να ομαδοποιήσει εκ’ νέου τις μίκρο-ομαδοποιήσεις σε τελικέςομαδοποιήσεις.

3.3.2 Υλοποίηση εφαρμογής

Για την υλοποίηση και εκτέλεση της εφαρμογής στη γλώσσα R χρησιμοποιήθηκε το RStudio. Ποιο συγκεκριμένα δημιουργήθηκε ένας χώρος (Project) για την οργάνωση των αρχείων του πηγαίου κώδιμα με όνομα **BeijingTaxisAnalysis** που περιέχει τα ακόλουθα R αρχεία:

- **Main.R**: Είναι ο κορμός του κώδικα που υλοποιήσαμε καθώς είναι το αρχείου που εκτελούμε. Αυτό καλεί τα άλλα δύο αρχεία (Properties.R και Functions.R).
- **Properties.R**: Περιέχει όλες τις παραμέτρους της εφαρμογής συμπεριλαμβάνοντας τις παραμέτρους των αλγορίθμων DenStream και CluStream. Πριν την εκτέλεση της εφαρμογής ο χρήστης μπορεί να τροποποιεί τις παραμέτρους που περιέχονται σε αυτό το αρχείο. Αυτό το αρχείο καλείται από τα άλλα δύο αρχεία (Main.R και Properties.R).
- **Functions.R**: Σε αυτό το αρχείο υπάρχουν όλες οι συναρτήσεις της εφαρμογής. Αυτό το αρχείο καλείται από το Main.R.

Στην Εικόνα 14 παρουσιάζεται η σειρά εκτέλεσης των αρχείων R.



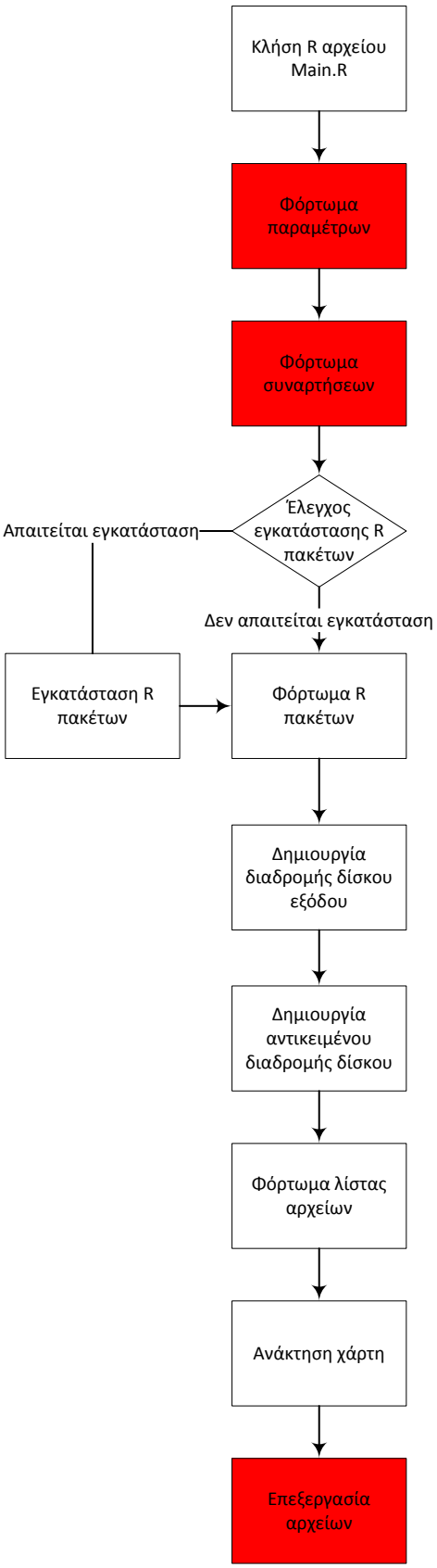
Εικόνα 14: Κλήσεις Αρχείων R

Πιο αναλυτικά, στην Εικόνα 15 παρουσιάζεται η ροή εκτέλεσης ξεκινώντας από το αρχείο πηγαίου κώδικα **Main.R**. ΜΕ κόκκινο χρώμα έχουν επισημανθεί οι κύριες διαδικασίες οι οποίες αποτελούν τον πυλώνα της υλοποίησης. Αυτές παρουσιάζονται αναλυτικά στις επόμενες παραγράφους.

Για καθένα από τα αρχεία εισόδου δημιουργούνται τα ακόλουθα αποτελέσματα, τα οποία αποθηκεύονται στον δίσκο (και κατόπιν στη βάση δεδομένων) εφόσον χρήστης το έχει επιλέξει:

1. **inX[1]macro.txt**: η μάκρο-ομάδα στην οποία ανήκει το κάθε σημείο.
2. **inX[1]micro.txt**: η μίκρο-ομάδα στην οποία ανήκει το κάθε σημείο.
3. **plotX[1]MacroClusters.jpg**: διαγραμματική απεικόνιση των μάκρο-ομαδοποιήσεων.
4. **plotX[1]macroSilhouette.jpg**: διαγραμματική απεικόνιση του Silhouette για τις μάκρο-ομαδοποιήσεις.
5. **plotX[1]Map.jpg**: απεικόνιση του χάρτη με τα σημεία.
6. **plotX[1]MicroClusterPurity.jpg**: διαγραμματική απεικόνιση της μέσης καθαρότητας (purity) για τις μίκρο-ομαδοποιήσεις.
7. **plotX[1]MicroClusters.jpg**: διαγραμματική απεικόνιση των μίκρο-ομαδοποιήσεων.
8. **plotX[1]MicroClustersAssignementArea.jpg**: διαγραμματική απεικόνιση της περιοχής των αναθέσεων για τις μικρό-ομαδοποιήσεις.
9. **plotX[1]microSilhouette.jpg**: διαγραμματική απεικόνιση του Silhouette για τις μίκρο-ομαδοποιήσεις.

[1] Όπου X ο αύξων αριθμός του αρχείου που επεξεργαζόμαστε



Εικόνα 12: Ροή Εκτέλεσης Κώδικα

3.3.2.1 Αρχικοποίηση Παραμέτρων

Οι παράμετροι αρχικοποιούνται στο R αρχείο **Properties.R**. Στον Πίνακας 4 παρουσιάζονται αναλυτικά οι παράμετροι που έχουμε θεσπίσει

Όνομα	Τύπος	Περιγραφή
INSTALL_PACKAGES	Δυαδική Τιμή	Καθορίζει εάν θα πρέπει να εγκατασταθούν ή όχι τα R πακέτα. Πρέπει να έχει τιμή true μόνο κατά την πρώτη φορά εκτέλεσης του R κώδικα.
SAVE_ASSIGNMENTS	Δυαδική Τιμή	Καθορίζει εάν θα πρέπει να αποθηκευτούν ή όχι οι αναθέσεις / τοποθετήσεις των δεδομένων / σημείων στον δίσκο. Βάζουμε τιμή false για λόγους ταχύτητας όταν κάνουμε αποσφαλμάτωση (debug) του κώδικα.
SAVE_PLOTS	Δυαδική Τιμή	Καθορίζει εάν θα αποθηκευτούν τα διαγράμματα ή όχι στον δίσκο. Βάζουμε τιμή false για λόγους ταχύτητας όταν κάνουμε αποσφαλμάτωση (debug) του κώδικα.
INPUT_PATH	Αλφαριθμητικό	Η διαδρομή στο δίσκο την οποία θα χρησιμοποιήσουμε για ανάγνωση αρχείων.
OUTPUT_PATH	Αλφαριθμητικό	Η διαδρομή στο δίσκο την οποία θα χρησιμοποιήσουμε για εγγραφή αρχείων.
INPUT_FILE_PATTERN	Αλφαριθμητικό	Είναι το πρότυπο που πρέπει να έχουν τα αρχείο εισόδου.
OUTPUT_FILE_PREFIX	Αλφαριθμητικό	Είναι το πρόθεμα που πρέπει να έχουν τα αρχεία εξόδου, αυτά που αποθηκεύονται στον δίσκο. Είναι το γενικό πρόθεμα.
OUTPUT_FILE_TYPE	Αλφαριθμητικό	Περιγράφει τον τύπο που είναι τα αρχεία εξόδου. Είναι ο γενικός τύπος
PLOT_FILE_PREFIX	Αλφαριθμητικό	Είναι το πρόθεμα που πρέπει να έχουν τα αρχεία που έχουν αποκλειστικά και μόνο διαγράμματα, αυτά που αποθηκεύονται στον δίσκο.
PLOT_FILE_TYPE	Αλφαριθμητικό	Περιγράφει τον τύπο που είναι τα αρχεία που έχουν αποκλειστικά και μόνο διαγράμματα.
ASSIGNMENT_METHOD	Αλφαριθμητικό	Υπαγορεύει τον τρόπο με τον οποίο θα γίνουν οι αναθέσεις των σημείων. Οι τιμές που μπορεί να πάρει είναι default και custom.

DIST_METHOD	Αλφαριθμητικό	Υπαγορεύει τον αλγόριθμο με τον οποίο θα γίνει η μέτρηση των αποστάσεων των σημείων για να ομαδοποιηθούν. Οι τιμές που μπορεί να πάρει είναι Euclidean και Manhattan.
CLUSTER_ALGORITHM	Αλφαριθμητικό	Υπαγορεύει τον αλγόριθμο ομαδοποίησης που θα χρησιμοποιηθεί. Οι τιμές που μπορεί να πάρει είναι DenStream και CluStream.
AREA	Αλφαριθμητικό	Η περιοχή για την οποία θέλουμε να φορτώσουμε χάρτη.
DEN_STREAM_EPSILON	Δεκαδικός	Έχει την τιμή της παραμέτρου epsilon της συνάρτησης DSC_DenStream. Αρχική τιμή 0.05
DEN_STREAM_BETA	Δεκαδικός	Έχει την τιμή της παραμέτρου β της συνάρτησης DSC_DenStream. Αρχική τιμή 0.2
DEN_STREAM_LAMPDA	Δεκαδικός	Έχει την τιμή της παραμέτρου λ της συνάρτησης DSC_DenStream. Αρχική τιμή 0.001.
CLUSTREAM_M	Ακέραιος	Έχει την τιμή της παραμέτρου m της συνάρτησης DSC_CluStream. Αρχική τιμή 150.
CLUSTREAM_K	Ακέραιος	Έχει την τιμή της παραμέτρου k της συνάρτησης DSC_CluStream. Αρχική τιμή 5.
CLUSTREAM_HORIZON	Ακέραιος	Έχει την τιμή της παραμέτρου horizon της συνάρτησης DSC_CluStream. Αρχική τιμή 1000.
CLUSTREAM_T	Ακέραιος	Έχει την τιμή της παραμέτρου t της συνάρτησης DSC_CluStream. Αρχική τιμή 2.
MAP_ALPHA	Δεκαδικός	Καθορίζει την διαφάνεια των σημείων / δεδομένων που απεικονίζονται σε ένα χάρτη.
MAP_ZOOM	Ακέραιος	Καθορίζει το ζουμ που θα έχει ο χάρτης.

Πίνακας 4, Παράμετροι

3.3.2.2 Αρχείο Συναρτήσεων

Οι συναρτήσεις υπάρχουν στο R αρχείο **Functions.R**. Στον Πίνακας 5 παρουσιάζουμε αναλυτικά τις παραμέτρους που έχουμε θεσπίσει.

Συνάρτηση	Περιγραφή	Παράμετροι	Έξοδος
outputPath	Δημιουργία φακέλου αρχείων εξόδου	Καμία	Το αρχείο εξόδου
savePlot	Αποθήκευση διαγράμματος στον δίσκο	1. **plotToSave**: το διάγραμμα που πρέπει να αποθηκευτεί 2. **localIndex**: αύξων αριθμός αρχείου εισόδου 3. **name**: όνομα διαγράμματος στον δίσκο	Καμία
retrieveMap	Ανάκτηση / φόρτωμα χάρτη από το διαδίκτυο	1. **area**: περιοχή για την οποία θα ανακτηθεί / φορτωθεί ο χάρτης	Το αντικείμενο του χάρτη
getGeomPoints	Δημιουργία αντικειμένου σημείων που θα απεικονιστεί στο χάρτη	1. **localInputStream**: η ροή δεδομένων / σημείων που πρέπει να απεικονιστεί στον χάρτη 2. **localIndex**: αύξων αριθμός αρχείου εισόδου 3. **isStream**: εάν είναι ροή δεδομένων ή όχι 4. **localColor**: το χρώμα του σημείου στο χάρτη 5. **localShape**: το σχήμα του σημείου στο χάρτη	Το αντικείμενο των σημείων που θα απεικονιστεί στο χάρτη
printMapSingleGeomPoints	Εκτύπωση και αποθήκευση χάρτη (εάν έχει επιλεχθεί) από ένα αντικείμενο σημείων	1. **localInputStream**: η ροή δεδομένων / σημείων που πρέπει να απεικονιστεί στον χάρτη 2. **localIndex**: αύξων αριθμός αρχείου εισόδου 3. **isStream**: εάν είναι ροή δεδομένων ή όχι 4. **localColor**: το χρώμα του σημείου στο χάρτη 5. **localShape**: το σχήμα του σημείου στο χάρτη	Καμία
printMapMultipleGeomPoints	Εκτύπωση και αποθήκευση χάρτη (εάν έχει	1. **localIndex**: αύξων αριθμός αρχείου εισόδου 2. **points**: αντικείμενο σημείων ταξί	Καμία

	επιλεχθεί) από πολλαπλά αντικείμενα σημείων	3. **microCenters**: αντικείμενο σημείων με τα κέντρα των μίκρο-ομαδοποιήσεων 4. **macroCenters**: αντικείμενο σημείων με τα κέντρα των μάκρο-ομαδοποιήσεων	
findAssignments	Εύρεση αναθέσεων σημείων / δεδομένων	1. **localInputStream**: η ροή δεδομένων / σημείων 2. **localAlgStream**: το αντικείμενο που περιέχει τον αλγόριθμο (DenStream/CluStream) 3. **localIndex**: αύξων αριθμός αρχείου εισόδου 4. **assignmentType**: ο τύπος των αναθέσεων που θέλουμε να βρούμε (micro ή macro)	Καμία
saveAssignments	Αποθήκευση των αναθέσεων σημείων / δεδομένων στον δίσκο	1. **assignments**: οι αναθέσεις προς αποθήκευση 2. **assignmentType**: ο τύπος των αναθέσεων 3. **localIndex**: αύξων αριθμός αρχείου εισόδου 4. **localPointsNumber**: το πλήθος των σημείων / δεδομένων που απεικονίζονται στις αναθέσεις	Καμία
getInputStream	Διάβασμα σημείων από τον δίσκο	1. **localFile**: η πλήρης διαδρομή του αρχείου από τον δίσκου (το αρχείο είναι σε CSV μορφή)	Η ροή δεδομένων που περιέχει τα σημεία που διαβάστηκαν
getDenStream	Επιστρέφεται το αντικείμενο του DenStream	1. **localInputStream**: η ροή δεδομένων / σημείων	Αντικείμενο του DenStream
getCluStream	Επιστρέφεται το αντικείμενο του CluStream	2. **localInputStream**: η ροή δεδομένων / σημείων	Αντικείμενο του CluStream
getCentersAndPrintMap	Εύρεση των κέντρων των ομαδοποιήσεων (μίκρο και μάκρο) και εκτύπωση τους χάρτη μαζί με τα σημεία (εάν έχει επιλεχθεί)	1. **localInputStream**: η ροή δεδομένων / σημείων 2. **pointsGeomPoints**: αντικείμενο σημείων ταξί 3. **localIndex**: αύξων αριθμός αρχείου εισόδου	Καμία
countFileLines	Καταμέτρηση του πλήθους των σημείων	1. **inputFile**: το αρχείο που περιέχει τα σημεία	Πλήθος σημείων

	υπάρχουν σε ένα αρχείο		
displayPlots	Εκτύπωση και αποθήκευση διαγραμμάτων (εάν έχει επιλεχθεί)	1. **localAlgStream**: το αντικείμενο που περιέχει τον αλγόριθμο (DenStream/CluStream) 2. **localInputStream**: η ροή δεδομένων / σημείων 3. **localIndex**: αύξων αριθμός αρχείου εισόδου 4. **localPointsNumber**: πλήθος σημείων	Καμία
displayEvaluations	Εκτύπωση και αποθήκευση αξιολογήσεων (εάν έχει επιλεχθεί)	1. **localAlgStream**: το αντικείμενο που περιέχει τον αλγόριθμο (DenStream/CluStream) 2. **localInputStream**: η ροή δεδομένων / σημείων 3. **localIndex**: αύξων αριθμός αρχείου εισόδου 4. **localPointsNumber**: πλήθος σημείων	Καμία
get_assignment.custom	Υλοποίηση προσαρμοσμένης μεθόδου get_assignments ορίζοντας ως παράμετρο μία εναλλακτική μετρική απόστασης	1. **localAlgStream**: το αντικείμενο που περιέχει τον αλγόριθμο (DenStream/CluStream) 2. **localInputStream**: η ροή δεδομένων / σημείων 3. **assignmentType**: αντιστοίχιη δεδομένων είτε σε μικρο-, ή σε μακρ-ομάδες. **4. distanceMethod**: επιλογή της μεθόδου υπολογισμού απόστασης του κάθε σημείου από το κέντρο της μικρο- ή μακρο-ομάδας βάσει Ευκλείδειας ή Μανχάταν μετρικής	

Πίνακας 5: Συναρτήσεις

3.3.2.3 Ροή Εκτέλεσης

Η υποδοχή της ροής των δεδομένων (txt αρχεία) και η δημιουργία των ομαδοποιήσεων είναι μια διαδικασία που γίνεται σειριακά. Κάθε αρχείο επεξεργάζεται ανεξάρτητα από το προηγούμενο. Η μόνη συσχέτιση που υπάρχει μεταξύ των αρχείων είναι στην διαδικασία της ομαδοποίησης εκ' νέου (recluster). Σε αυτή την διαδικασία λαμβάνεται υπόψη το αντικείμενου του αλγορίθμου (DenStream ή CluStream) που είχε προκύψει από την εκτέλεση του προηγούμενου αρχείου. Οι ακόλουθες ενέργειες γίνονται σειριακά κατά την επεξεργασία των αρχείων:

1. Αποθήκευση ώρας έναρξης επεξεργασίας του αρχείου

```
startTime <- Sys.time()
```

2. Αποθήκευση της διεύθυνσης του αρχείου σε μια τοπική μεταβλητή

```
localFile <- fileList[i]
```

3. Υπολογισμός και αποθήκευση του πλήθους των σημείων που υπάρχουν στο αρχείο

```
pointsNumber <- countFileLines(localFile)
```

4. Φόρτωμα των σημείων του αρχείου σε μια ροή δεδομένων (ανάγνωση του αρχείου)

```
inputStream <- getInputStream(localFile)
```

5. Εμφάνιση των δεδομένων / σημείων στον χάρτη. Εφόσον ο χρήστης το έχει επιλέξει, στις παραμέτρους που έχουν φορτωθεί, ο χάρτης αποθηκεύεται στον δίσκο

```
printMap(inputStream, localMap, i)
```

6. Δημιουργία ενός αντικειμένου του DenStream ή CluStream για τα δεδομένα που περιέχονται σε αυτό το αρχείο

```
outputStream <- getDenStream(inputStream) ή
outputStream <- getCluStream(inputStream)
```

7. Έλεγχος εάν πρέπει να γίνει ομαδοποίηση εκ' νέου. Εάν το αρχείο που επεξεργαζόμαστε δεν μπορεί να γίνει ομαδοποίηση εκ' νέου

```
if (i > 1) {
  recluster (tempStream, outputStream)
}
```

8. Εύρεση αναθέσεων για μίκρο και μάκρο-ομαδοποιήσεις. Εφόσον ο χρήστης το έχει επιλέξει, στις παραμέτρους που έχουν φορτωθεί, οι αναθέσεις αποθηκεύονται στον δίσκο. Η συνάρτηση αυτή αντικαθιστά τη ρουτίνα get_assignments που παρέχει η R (βιβλιοθήκη: stream) με μία νέα συνάρτηση get_assignments.custom η οποία υλοποιήθηκε στα πλαίσια της διπλωματικής εργασίας. Η νέα ρουτίνα get_assignments.custom λαμβάνει υπόψη τη τιμή της παραμέτρου DIST_METHOD και προσαρμόζει την μετρική υπολογισμού της απόστασης των υποψήφιων σημείων με το κέντρο της κάθε μικρο- ή μακρο-ομάδας με σκοπό να καθορίσει αν αυτό το σημείο θα ενταχθεί στην ομάδα ή όχι. Η μετρική μπορεί να είναι η Ευκλείδεια ή η Μανχάταν αλλά δίνει επίσης τη δυνατότητα ορισμού μίας οποιασδήποτε άλλης μετρικής από τον χρήστη.

```
findAssignments(inputStream, outputStream, i, "micro")
findAssignments(inputStream, outputtream, i, "macro")
```

9. Εμφάνιση αξιολογήσεων και εφόσον ο χρήστης το έχει επιλέξει, στις παραμέτρους που έχουν φορτωθεί, αποθηκεύονται στον δίσκο

```
displayEvaluations(outputStream, inputStream, i, pointsNumber)
```

10. Εμφάνιση διαγραμμάτων και εφόσον ο χρήστης το έχει επιλέξει, στις παραμέτρους που έχουν φορτωθεί, αποθηκεύονται στον δίσκο

```
displayPlots(outputStream, inputStream, i, pointsNumber)
```

11. Υπολογισμός χρόνου επεξεργασίας του αρχείου

```
endTime <- Sys.time()
timeTaken <- endTime - startTime
print(paste0("Time taken for execution (sec) for run ", i,": ", timeTaken))
```

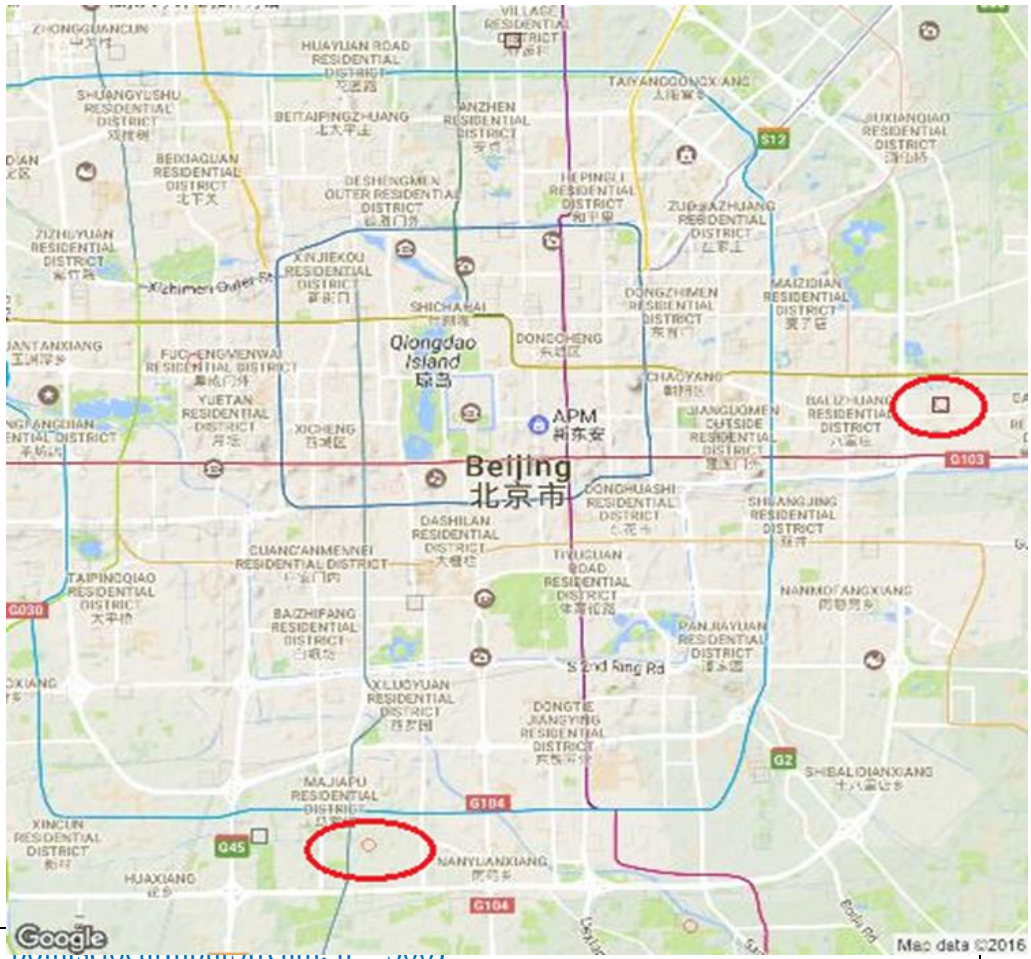
Στο σημείο αυτό πρέπει να αναφέρουμε ότι η επεξεργασία των δεδομένων λαμβάνει χώρα σε μια επαναλαμβανόμενη δομή (for loop).

Όπως αναφέρθηκε παραπάνω, εξαιτίας των περιορισμένων υπολογιστικών πόρων που ήταν διαθέσιμοι για την εκτέλεση της εφαρμογής θεώρηθηκε σκόπιμη η σταδιακή ροή των δεδομένων προς επεξεργασία (ομαδοποίηση) είτε από ένα μεγάλο αρχείο ή από μικρότερα αρχεία τα οποία θα είναι χρονικά ταξινομημένα. Προτιμήθηκε η δεύτερη μέθοδος καθώς με τη διαίρεση σε μικρότερα αρχεία (π.χ. σε διάστημα πέντε λεπτών) θα μπορεί να ελεγχθεί η εκτέλεση του προγράμματος και η βέλτιση χρήση των υπολογιστικών πόρων. Επιπλέον, η αρχική δειγματοληψία των δεδομένων ήταν ανά πεντάλεπτο. Συνεπώς χρησιμοποιείται η εντολή for-loop για την συνεχή ροή δεδομένων από όλα τα αρχεία που βρίσκονται σε ένα συγκεκριμένο φάκελο.
Η εφαρμογή του αλγορίθμου και στη μία ή την άλλη περίπτωση δεν μεταβάλλεται. Στο βήμα 7 παραπάνω γίνεται το recluster, δηλαδή η εκτέλεση της ομαδοποίησης με τον ίδιο αλγόριθμο (DenStream) λαμβάνοντας υπόψη α) τα δεδομένα της μέχρι τώρα ομαδοποίησης, και β) τα νέο-εισερχόμενα δεδομένα από το νέο αρχείο.

3.4 Προβολή δεδομένων και σημείων ομαδοποίησης στο χάρτη

Για την απεικόνιση των δεδομένων στον χάρτη της πόλης του Πεκίνου αξιοποιούνται οι ρουτίνες των βιβλιοθηκών ggmap και gtools της R (Εικόνα 16). Η συνολική απεικόνιση των δεδομένων σε μία ροή δεν ήταν εφικτή (π.χ. 8000 ταξί) καθώς οι συγκεκριμένες βιβλιοθήκες διαθέτουν περιορισμένες δυνατότητες μεγέθυνσης του χάρτη όπως οι χάρτες της Google. Συνεπώς ορίστηκε η δυνατότητα προβολής των πρώτων 500 σημείων που ήταν διαθέσιμα στη κάθε ροή δεδομένων. Στο χάρτη απεικονίζονται:

1. Ταξί: σε μαύρα τετράγωνα.
2. Κέντρα μίκρο-ομαδοποιήσεων: σε κόκκινους κύκλους.




```
mapPoints <- get_points(localInputStream, n = 500)

colnames(mapPoints) <- c("lat", "lon")

mapPoints <- data.frame(mapPoints)

names(mapPoints) <- c("lat", "lon")
```

```
qplot(lat, lon, data = mapPoints)
```

Ο χάρτης αρχικοποιείται σε ένα αντικείμενο qmap με το περιεχόμενο που ανακτάται από τις βιβλιοθήκες της Google (Google Maps Api) βάσει των συντεταγμένων του Πεκίνου. Στο σταθερό αυτό γράφημα του χαρτή προστίθεται ένα διάνυσμα γεωμετρικών σημείων τα οποία αντανακλούν τα σημεία των ταξί στο χάρτη. Επιπλέον προστίθεται άλλο ένα διάνυσμα γεωμετρικών σημείων τα οποία συμβολίζουν τα κέντρα των παραγόμενων ομαδοποιήσεων (cluster centers).

```
# show it on the map

mapObject <- qmap('Beijing', zoom=MAP_ZOOM, maptype = "terrain") +

geom_point(aes(x = lat, y = lon), data = mapPoints, color="black", fill = 'black', alpha = MAP_ALPHA, size = 3)

mapObject <- mapObject +

geom_point(aes(x = lat, y = lon), data = centerPoints, color="red", fill = 'red', alpha = MAP_ALPHA, size = 5)
```

Ο παραγόμενος χάρτης απεικονίζεται στην οθόνη και αποθηκεύεται επίσης στον δίσκο (αρχείο jpg). Το σύνολο των παραγόμενων εικόνων ενώνονται σε μία ενιαία εικόνα τύπου animated gif για την αντανάκλαση των κινούμενων σημείων στο χάρτη.

```
plotToSave <- plot(mapObject)

savePlot(plotToSave, localIndex, "Map")
```

Κ Ε Φ Α Λ Α Ι Ο 4

4. Εφαρμογή αλγορίθμων συσταδοποίησης βάσει πυκνότητας

Η εφαρμογή των αλγορίθμων DenStream και CluStream στο κεφάλαιο 3 μας επιτρέπει να πραγματοποιήσουμε συγκεκριμένα πειράματα ομαδοποίησης ροών δεδομένων με χωρικά και χρονικά χαρακτηριστικά. Όπως αναφέρθηκε στην εισαγωγή του ίδιου κεφαλαίου, θα εξετάσουμε το σύνολο GPS δεδομένων της Microsoft από τη κυκλοφορία 10.500 περίπου ταξί στο Πεκίνο (Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang, 2010). Ως κύριο κορμό των πειραμάτων εξετάζουμε τις παρακάτω περιπτώσεις:

1. Εκτέλεση του αλγορίθμου DenStream με κύρια μετρική υπολογισμού της απόστασης την Ευκλείδεια απόσταση (βλ. §2.3.5).
2. Εκτέλεση του αλγορίθμου DenStream με κύρια μετρική υπολογισμού της απόστασης την απόσταση Μανχάταν (βλ. §2.3.5).
3. Εκτέλεση του αλγορίθμου CluStream με κύρια μετρική υπολογισμού της απόστασης την Ευκλείδεια απόσταση (βλ. §2.3.5).

Επιπρόσθετα, για να μπορέσουμε να αναλύσουμε τα αποτελέσματα στο κάθε στάδιο των δοκιμών (1 έως 4) αλλά και να τα συγκρίνουμε μεταξύ τους, θα πρέπει να αρχικά να ορίσουμε ποιες είναι αυτές οι μετρικές αξιολόγησης των αποτελεσμάτων. Οι μετρικές αυτές περιγράφονται στην παρακάτω ενότητα.

4.1 Μετρικές αξιολόγησης αποτελεσμάτων

Ο έλεγχος της ποιότητας των ομαδοποιήσεων σε κάθε στιγμιότυπο της εκτέλεσης (υποδοχή αντίστοιχου txt αρχείου) θα βασιστεί στις παρακάτω παραμέτρους:

- «καθαρότητα» ομαδοποίησης (micro-cluster purity)
- ακρίβεια της ομαδοποίησης
- άθροισμα τετραγώνων μέσα στην ομάδα (within cluster sum of squares - SSQ)

Η μετρική Purity ορίζει την «καθαρότητα» ή αλλιώς η καθαρότητας τη ποιότητα της ομαδοποίησης. Για τον υπολογισμό της καθαρότητας κάθε ομαδοποίηση συσχετίζεται με τη κλάση (αντιπροσωπευτική γεωγραφική θέση) που εμφανίζεται πιο συχνά μέσα στην ομάδα. Η μετρική αυτή υπολογίζεται διαιρώντας τον αριθμών των σωστά ταξινομημένων γεωγραφικών θέσεων με τον αριθμό των ομαδοποιήσεων:

$$purity = \frac{\sum_{i=1}^{K} \frac{|C_i^d|}{|C_i|}}{K} * 100\%$$ (F. Chao;M. Ester, 2006)

- Η μεταβλητή K ορίζει τον αριθμό των ομαδοποιήσεων
- $|C_i^d|$ ορίζει τον αριθμό των σημείων γύρω από την επικρατέστερη τιμή (κλάση) σε μία ομάδα i.
- $|C_i|$ ορίζει των αριθμό των σημείων σε σε μία ομάδα i.

Ουσιαστικά η μεταβλητή purity μετράει την καθαρότητα ή αλλιώς ποιότητα των ομαδοποιήσεων σε σχέση με τις πραγματικές τιμές (κλάσεις) οι οποίες είναι γνωστές εκ των προτέρων για τα διαθέσιμα σύνολα δεδομένων. Καθώς η βαρύτητα των σημείων εξασθενεί σταδιακά, η καθαρότητα υπολογίζεται με βάση ένα αριθμό σημείων τα οποία εισέρχονται σε ένα προκαθορισμένο διάστημα χρόνου.

Η μετρική Rand υπολογίζει τον αριθμό των αποφάσεων που είναι έγκυρες ή αλλιώς την ακρίβεια της ομαδοποίησης:

$$RI = \frac{TP+TN}{TP+FP+TN+FN}$$ (C. D. Manning, P. Raghavan and H. Schütze, 2008)

- Μία θετικά –ορθή απόφαση (TP) αντιστοιχεί δύο όμοια σημεία στην ίδια ομάδα.
- Μία αρνητικά –ορθή απόφαση (TN) αντιστοιχεί δύο μη όμοια σημεία σε διαφορετικές ομάδες.
- Μία θετικά –εσφαλμένη απόφαση (FP) αντιστοιχεί δύο μη όμοια σημεία στην ίδια ομάδα.
- Μία αρνητικά – εσφαλμένη απόφαση (FN) αντιστοιχεί δύο όμοια σημεία σε διαφορετικές ομάδες.

cRand είναι προσαρμοσμένη έκδοση του δείκτη Rand. Καθώς ο δείκτης Rand μπορεί να πάρει τιμές 0 έως 1, ο δείκτης cRand μπορεί να πάρει και αρνητικές τιμές εφόσον ο δείκτης

που έχει υπολογιστεί έχει τιμή μικρότερη από την αναμενόμενη τιμή (C. D. Manning, P. Raghavan and H. Schütze, 2008).

Ο δείκτης *Within cluster sum of squares by cluster* (SSQ) ορίζει το ποσοστό διακύμανσης που υπάρχει ανάμεσα στα σημεία. Όσο πιο μεγάλη αυτή η τιμή τόσο μεγαλύτερη η μείωση της διακύμανσης. Το κάθε σημείο καταχωρείται στη κοντινότερη ομάδα βάση απόστασης μεταξύ σημείου και κέντρου της ομάδας. Το κέντρο της κάθε ομάδας ορίζεται κατόπιν ως η μέση τιμή (mean) μεταξύ των σημείων που έχουν εκχωρηθεί σε μία ομάδα. Συνεπώς το άθροισμα τετραγώνων ανά ομάδα ορίζεται ως:

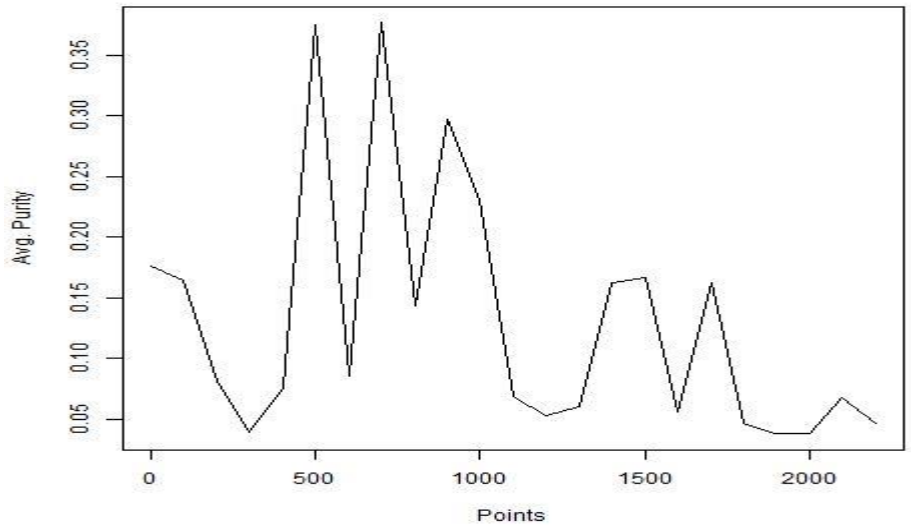
$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \overline{x_{kj}})^2$ (C. D. Manning, P. Raghavan and H. Schütze, 2008)

Όπου S_k είναι το σύνολο των σημείων στη K-οστή ομάδα και $\overline{x_{kj}}$ είναι η μεταβλητή στη σειρά j που αντιπροσωπεύει το κέντρο της ομάδας στην ομάδα K.

Το εύρος των μικρο και μακρο-ομάδων που παράγονται για την ομαδοποίηση των σημείων επηρεάζει άμεσα τις τιμές των παραπάνω μετρικών. Όσες περισσότερες μικρο και μακρο-ομάδες δημιουργούνται, τόσο καλύτερα τα αποτελέσματα των παραπάνω μετρικών. Για παράδειγμα, η τιμή της μετρικής «καθαρότητα» ομαδοποίησης (purity) είναι μεταξύ [0, 1]. Όσο πιο κοντά είναι στη τιμή 1 τόσο πιο ποιοτικό είναι το αποτέλεσμα της ομαδοποίησης (δημιουργία ομάδων σε κοντινή απόσταση, αντιστοίχιση όλων των σημείων) όπως στο παράδειγμα που περιγράφεται στον παρακάτω πίνακα.

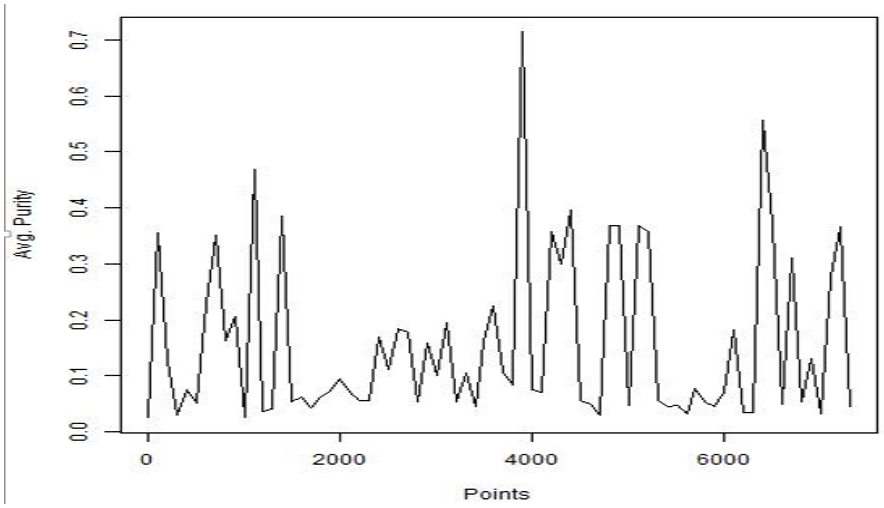
purity	Rand	cRand	SSQ
0.6675	*0.*4362	*0.*1486	*0.*1059
0.6617	*0.*3832	*0.*7884	*0.*9573
0.6728	*0.*4293	*0.*1415	*0.*1013
0.6830	*0.*4330	*0.*9903	*0.*8194

Επιπλέον, ο αριθμός των ομάδων και η αντιστοιχία των σημείων σε αυτές είναι συνάρτηση των παραμέτρων του εκάστοτε αλγορίθμου και του όγκου των δεδομένων που λαμβάνονται προς επεξεργασία. Όπως αναφέρθηκε παραπάνω η μεταβλητή purity μετράει την καθαρότητα των ομαδοποιήσεων σε σχέση με τις αντιπροσωπευτικές τιμές (κλάσεις), ουσιαστικά τις τιμές που είναι κοντά στο κέντρο της ομάδας. Καθώς η βαρύτητα των σημείων εξασθενεί σταδιακά, η καθαρότητα υπολογίζεται με βάση ένα αριθμό σημείων τα οποία εισέρχονται σε ένα προκαθορισμένο διάστημα χρόνου. Στο παρακάτω διάγραμμα επιβεβαιώνεται ο παραπάνω κανόνας: η μέση καθαρότητα φτάνει μία μέγιστη τιμή άνω του 0.35 για τα πρώτα 1000 σημεία και εξασθενεί καθώς περισσότερα σημεία εισάγονται στην ομάδα.



Εικόνα 13: καμπύλη μέσης καθαρότητας κατά την ομαδοποίηση μίας ροής δεδομένων

Καθώς αυξάνεται ο αριθμός των σημείων που εκτελούνται μέσω του αλγορίθμου (clustering, re-clustering) ο υπολογισμός της παραπάνω μετρικής λαμβάνει υπόψη της περισσότερα σημεία ανά ομάδα. Στο παρακάτω διάγραμμα η τιμή της μετρικής είναι διπλάσια (ως μέγιστη τιμή) σε σχέση με το προηγούμενο διάγραμμα (0.70) και σε ένα παράθυρο επεξεργασίας μέχρι τότε έως 4000 σημεία.



Εικόνα 14: καμπύλη μέσης καθαρότητας κατά την ομαδοποίηση μίας ροής δεδομένων

1. Εκτέλεση του αλγορίθμου DenStream – Ευκλείδεια απόσταση

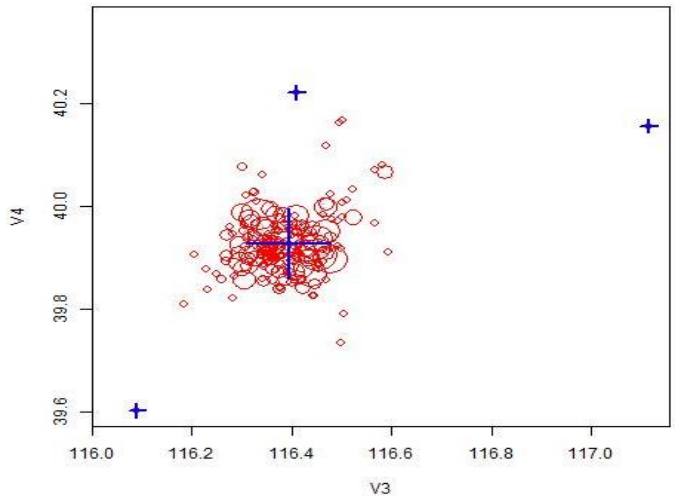
Ο παρακάτω πίνακας παρουσιάζει ένα στιγμιότυπο της εφαρμογής του αλγορίθμου DenStream στην ροή και επεξεργασία 4 αρχείων από τα συνολικά 1766 αρχεία στα οποία είχαν επιμεριστεί χρονικά τα δεδομένα σχετικά με τη κίνηση των ταξί στο Πεκίνο (Κίνα). Στον πίνακα παρατηρούμε τη διακύμανση του αριθμού των μικρο και μακρο-ομαδοποιήσεων που δημιουργούνται ως συνάρτηση του όγκου των δεδομένων. Επιπλέον, λαμβάνονται υπόψη οι παράμετροι eps (0.05), β (0.02) και λ (0.001) του αλγορίθμου καθώς και το γεγονός της επιλογής της Ευκλείδειας μετρικής υπολογισμού της απόστασης. Επίσης, ο χρόνος εκτέλεσης αυξάνεται σημαντικά, λόγω του recluster, καθώς αυξάνεται ο αριθμός των σημείων στη ροή δεδομένων, ωστόσο η τιμή της καθαρότητας καθώς και του αθροίσματος των τετραγώνων είναι ικανοποιητικές. Συγκεκριμένα, η τιμής της καθαρότητας παραμένει σταθερά κοντά στο ανώτερο όριο (1), όπως επίσης και η τιμή του αθροίσματος των τετραγώνων είναι υψηλή.

Α/Α	Πλήθος σημείων	Πλήθος μικρό-ομαδοποιήσεων	Πλήθος μάκρο-ομαδοποιήσεων	purity	Rand	cRand	SSQ	Χρόνος (δευτ.)
1	2385	272	4	0.60	0.399	0.001	25.95	2.09
2	7517	466	6	0.52	0.386	0.0003	92.09	9.00
3	7997	539	10	0.82	0.384	1.819	96.67	11.35
4	7983	513	10	0.77	0.394	0.239	77.95	10.49

Πίνακας 4: Αποτελέσματα για τον 1ο συνδυασμό παραμέτρων

Εικόνα 15: απεικόνιση σημείων ταξί στο χάρτη του Πεκίνου (χάρτες google)

Στον παραπάνω χάρτη απεικονίζονται, αντιπροσωπευτικά, με κόκκινους κύκλους οι θέσεις 500 ταξί στο χάρτη του Πεκίνου τη δεδομένη χρονική στιγμή. Δεν επιλέγουμε να απεικονίσουμε το σύνολο των (π.χ. 2375 ή 7517, κ.ο.κ) θέσεων καθώς υπάρχει μεγάλη συσσώρευση των σημείων-ταξί στο κέντρο της πόλης και η απεικόνιση τους δεν είναι διακριτή.



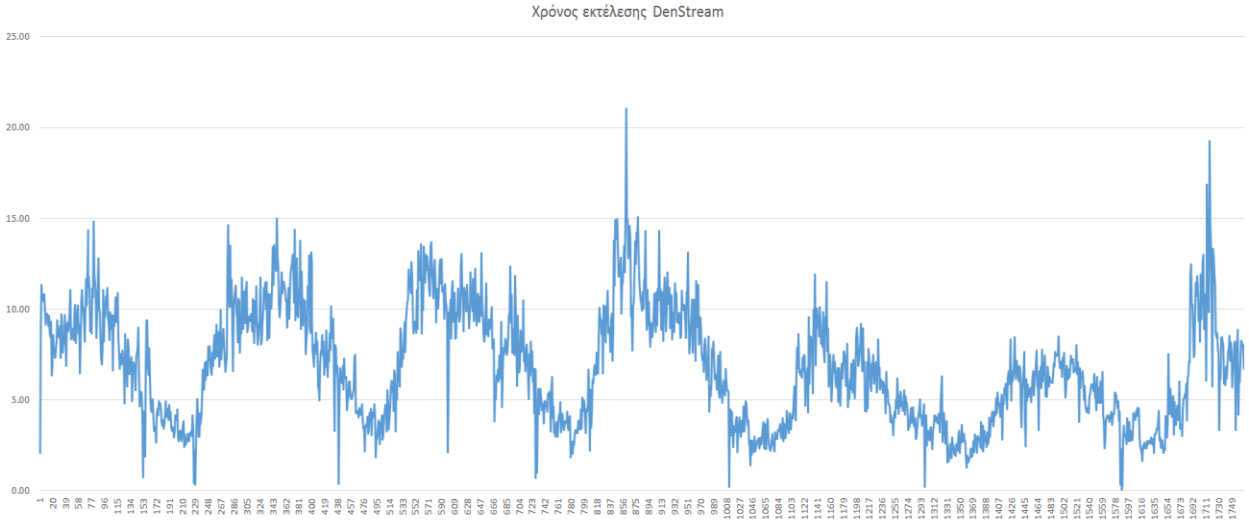
Εικόνα 16: δημιουργία μικρο και μακρο-ομάδων – εκτέλεση DenStream με Ευκλείδεια απόσταση

Η μεγάλη συσσώρευση των ταξί σε πολύ κοντινή απόσταση αποδεικνύεται και από τη συσσώρευση των μικρο-ομάδων σε πολύ κοντινή απόσταση σύμφωνα με το παραπάνω διάγραμμα. Οι μπλε σταυροί ορίζουν τις μακρο-ομάδες και οι κόκκινοι κύκλοι τις μικρο ομάδες. Η ακτίνα – μέγεθος του κάθε κύκλου δείχνει τη διαφοροποίηση στην αντιστοίχιση των σημείων στις ομάδες βάσει απόστασης ή αλλιώς βάσει των παραμέτρων του αλγορίθμου (π.χ. eps).

Στην υποδοχή του δεύτερου txt αρχείου με τα επόμενα σημεία (7517 θέσεις ταξί) αποτυπώνονται οι μετακινήσεις των σημείων πάνω στο χάρτη.

Εικόνα 17: απεικόνιση σημείων ταξί στο χάρτη του Πεκίνου (χάρτες google)

Το διάγραμμα του χρόνου που απαιτήθηκε για να τρέξει η εφαρμογή για όλα τα αρχεία παρουσιάζεται ακολούθως.



Εικόνα 1818: διάγραμμα χρόνων εκτέλεσης του αλγορίθμου DenStream

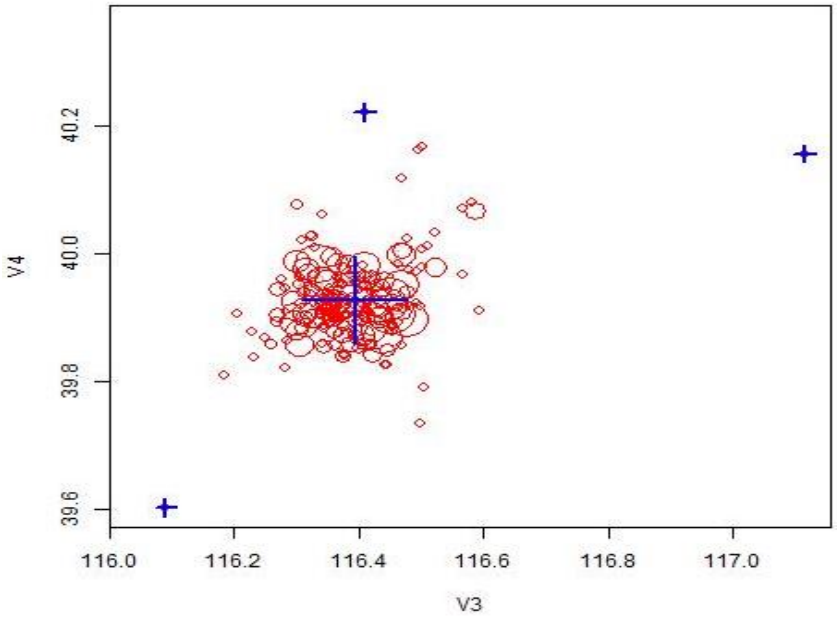
2. Εκτέλεση του αλγορίθμου DenStream – Μανχάταν απόσταση

Η εκτέλεση του αλγορίθμου DenStream επαναλαμβάνεται ορίζοντας ως μετρική απόστασης για την αντιστοίχιση των σημείων –θέσεων ταξί στις ομάδες τη Μανχάταν. Παρατηρούμε ότι το πλήθος των μικρο- και μακρο-ομάδων παραμένει το ίδιο σε σύγκριση με την προηγούμενη εκτέλεση του DenStream (Ευκλείδεια απόσταση). Αντίστοιχα δεν μεταβάλλονται οι ποιοτικές τιμές του αλγορίθμου όπως και ο χρόνος εκτέλεσης.

Α/Α	Πλήθος σημείων	Πλήθος μικρό-ομαδοποιήσεων	Πλήθος μάκρο-ομαδοποιήσεων	purity	Rand	cRand	SSQ	Χρόνος (δευτ.)
1	2385	272	4	0.61	0.40	0.002	25.94	2.09
2	7517	466	6	0.52	0.39	0.0002	92.09	9.00
3	7997	539	10	0.82	0.38	0.182	96.67	11.35
4	7983	513	10	0.77	0.39	0.239	77.95	10.49

Πίνακας 5: Αποτελέσματα για τον 2ο συνδυασμό παραμέτρων

Από το παρακάτω διάγραμμα αποδεικνύεται επίσης ότι η αντιστοίχιση των σημείων στις ομάδες λαμβάνει υπόψη τη τιμή της παραμέτρου γειτνίασης (eps) για να καθορίσει το πλήθος των σημείων που πρέπει να ανήκουν σε μία ομάδα ανεξάρτητα από τη μετρική υπολογισμού της γεωγραφικής απόστασης (είτε Ευκλίδεια ή Μανχάταν).



Εικόνα 20: δημιουργία μικρο- και μακρο-ομάδων – εκτέλεση DenStream με Μανχάταν απόσταση

3. Εκτέλεση του αλγορίθμου CluStream – Ευκλείδεια απόσταση

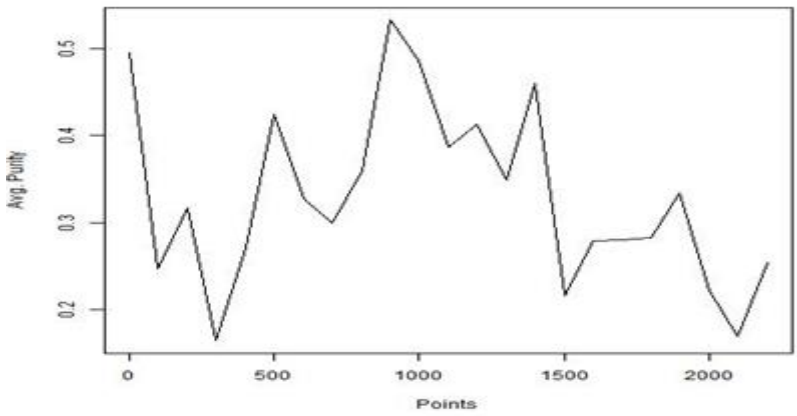
Ο παρακάτω πίνακας παρουσιάζει ένα στιγμιότυπο της εφαρμογής του αλγορίθμου CluStream στην ροή και επεξεργασία 4 επίσης αρχείων από τα συνολικά 1766 αρχεία. Σε αντίθεση με τον DenStream, στον αλγόριθμο αυτό δεν παρατηρούμε διακύμανση του αριθμού των μικρο- και μακρο-ομαδοποιήσεων ως συνάρτηση του όγκου των δεδομένων. Αυτό πηγάζει από το γεγονός ότι ορίζεται η παράμετρος k που αφορά το μέγιστο αριθμό μακρο-ομάδων (k=10), και η παράμετρος m που ορίζει τον αριθμό των μικρο-ομάδων (m=150). Η παράμετρος horizon ορίζει το χρονικό παράθυρο (2000) και η t αντιπροσωπεύει τη μέγιστη επιτρεπόμενη απόκλιση της ρίζας μέσου τετραγώνου των τιμών μίας μικρο-ομάδας από το κέντρο της ομάδας (=2).

Στο στάδιο αυτό των δοκιμών επιλέγεται η Ευκλείδεια μετρική υπολογισμού της απόστασης. Παρατηρούμε ότι αφενός ο χρόνος εκτέλεσης αυξάνεται καθώς μεγαλώνει ο αριθμός των σημείων στη ροή δεδομένων, ωστόσο ο απαιτούμενος χρόνος είναι κατά πολύ μικρότερος του DenStream. Παρόλα αυτά η τιμή της καθαρότητας καθώς και του αθροίσματος των τετραγώνων δεν είναι ικανοποιητικές σε σύγκριση με τον DenStream.

A/A	Πλήθος σημείων	Πλήθος μικρό-ομαδοποιήσεων	Πλήθος μάκρο-ομαδοποιήσεων	purity	Rand	cRand	SSQ	Χρόνος (δευτ.)
1	2385	150	10	0.21	0.85	0.015	6.68	3.10
2	7517	150	10	0.06	0.84	0.002	32.27	3.46
3	7997	150	10	0.08	0.82	0.002	45.18	3.50
4	7983	150	10	0.09	0.82	0.002	34.87	3.19

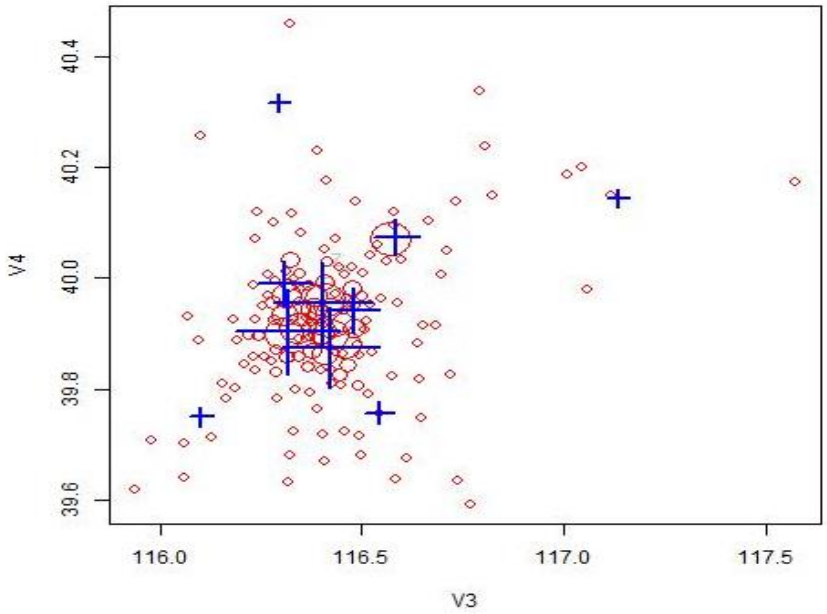
Πίνακας 8: Αποτελέσματα για τον 3ο συνδυασμό παραμέτρων

Συγκεκριμένα, η τιμή της καθαρότητας παραμένει σταθερά κάτω από 0.5, όπως επίσης και η τιμή του αθροίσματος των τετραγώνων είναι χαμηλή.



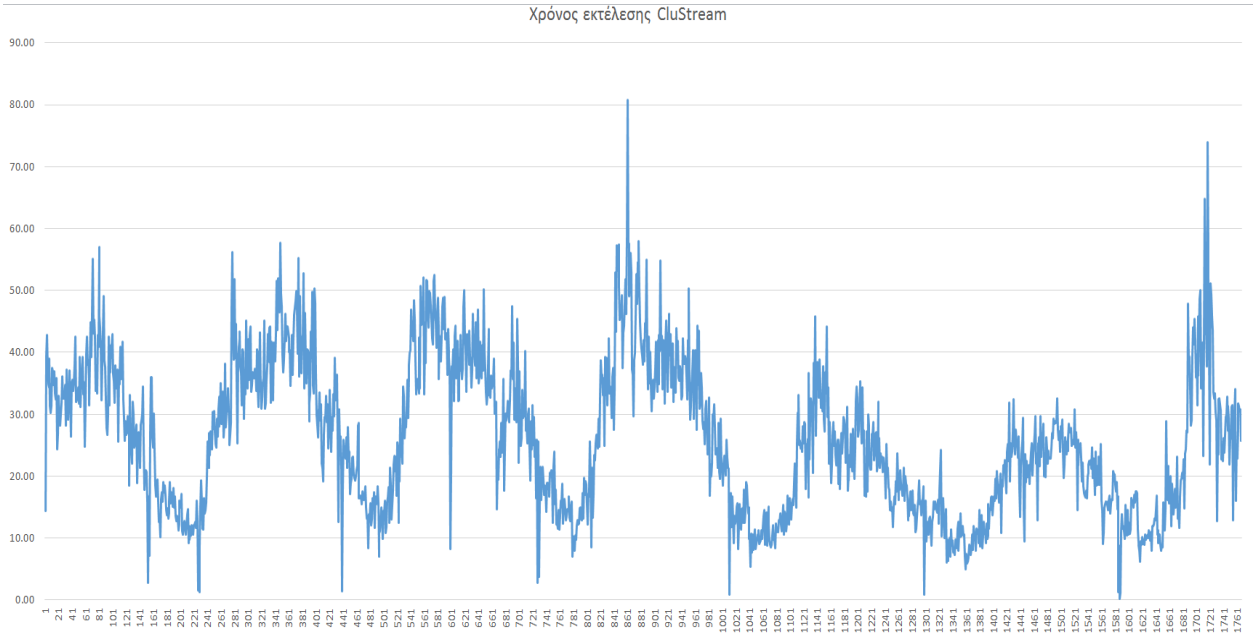
Εικόνα 19: εκτέλεση CluStream – καμπύλη μετρικής καθαρότητας (purity)

Πράγματι, στο παρακάτω διάγραμμα διαφαίνεται η μεγάλη σποραδικότητα των μικρο-ομάδων ως αποτέλεσμα του CluStream.



Εικόνα 20: δημιουργία μικρο- και μακρο-ομάδων – εκτέλεση CluStream με Μανχάταν απόσταση

Το διάγραμμα του χρόνου που απαιτήθηκε για να τρέξει η εφαρμογή για όλα τα αρχεία παρουσιάζεται ακολούθως.



Εικόνα 2121: διάγραμμα χρόνων εκτέλεσης του αλγορίθμου CluStream

ΚΕΦΑΛΑΙΟ 5

5. Συμπεράσματα

Η παρούσα εργασία αξιολόγησε τις δυνατότητες που παρέχουν σύγχρονες βιβλιοθήκες (π.χ. MOA - Massive Online Analysis) και εργαλεία (R) στη συσταδοποίηση ροών δεδομένων με αυξημένη πολυπλοκότητα. Σε ένα τέτοιο σενάριο όπως αυτό των εκατομμυρίων εγγραφών με τα δεδομένα των κινήσεων των ταξί στο Πεκίνο έπρεπε να ληφθούν υπόψη οι περιορισμοί: α) του μη προβλέψιμου όγκου σημείων – ταξί που μετακινούνται σε μία συγκεκριμένη περιοχή ανά δεδομένη χρονική στιγμή και β) και του γεγονότος ότι ο αριθμός των πραγματικών ομάδων δεδομένων είναι άγνωστος. Στο πλαίσιο επίλυσης τέτοιων προβλημάτων με εφαρμογή αλγορίθμων συσταδοποίησης βάσει πυκνότητας η παρούσα εργασία συνέβαλλε στην εξαγωγή των παρακάτω συμπερασμάτων:

1. Η βιβλιοθήκη της MOA (Massive Online Analysis) και η διαθεσιμότητα της μέσω του προγραμματιστικού περιβάλλοντος της R διευκολύνουν σημαντικά την εφαρμογή αλγορίθμων, επεξεργασία ροών δεδομένων και τη γραφική απεικόνιση των αποτελεσμάτων. Η εφαρμογή που υλοποιήθηκε βασισμένη στις βιβλιοθήκες της R και αυτή της MOA διευκολύνθηκε σημαντικά στην εφαρμογή του αλγορίθμου DenStream και την άμεση σύγκριση του με τον αλγόριθμο CluStream χωρίς να απαιτείται να αναπτυχθούν προγραμματιστικές γνώσεις σε βάθος.
2. Ο αλγόριθμος DenStream απέδωσε καλύτερα αποτελέσματα σε σύγκριση με τον CluStream και με βάση συγκεκριμένες μετρικές αξιολόγησης που ορίζονται στη βιβλιογραφία (βλ. §4.1). Τα αποτελέσματα αξιολόγησης του CluStream τονίζουν ότι τα αποτελέσματα της πρώτης (online) φάσης ομαδοποίησης θα πρέπει να μπορούν να βελτιωθούν όχι απλά κάνοντας μία δεύτερη (offline) ομαδοποίηση με τα τωρινά αποτελέσματα ομαδοποίησης που υπάρχουν στη μνήμη, αλλά και με ένα επιπλέον αλγόριθμο όπως ο k-means. Στη περίπτωση αυτή θα πρέπει να αναλυθεί η επιλογή της παραμέτρου k, για παράδειγμα k=3.
3. Η αρχικοποίηση των παραμέτρων του εκάστοτε αλγορίθμου είναι εμπειρική και επηρεάζει άμεσα τη δημιουργία συγκεκριμένου αριθμού μικρο και μακρο-ομάδων. Κάποιος αναλυτής μπορεί να καταλήξει σε συγκεκριμένες τιμές για τις παραμέτρους αυτές αφού μελετήσει το συγκεκριμένο πεδίο του προβλήματος (π.χ. κυκλοφοριακό πρόβλημα μίας πόλης). Στην περίπτωση των δοκιμών που εφαρμόστηκαν παραπάνω η βέλτιστες τιμές των παραμέτρων προέκυψαν κατόπιν δοκιμών. Ωστόσο, η αντικατάσταση της Ευκλείδειας απόστασης, που αποτελεί την εκ των προτέρων μετρική του αλγορίθμου για τον υπολογισμό της απόστασης μεταξύ του κέντρου μίας ομάδας και της θέσης ενός υποψήφιου σημείου, με μία άλλη μετρική όπως η Μανχάταν δεν επηρεάζει το αποτέλεσμα της εκτέλεσης του αλγορίθμου.
4. Πράγματι εφαρμόζοντας τον αλγόριθμο DenStream ο οποίος βασίζεται στον αλγόριθμο DBSCAN μπορεί να υλοποιηθεί ένα σύστημα πρόβλεψης το οποίο σε πραγματικό χρόνο θα μπορεί να επεξεργάζεται ροές δεδομένων, να τις ταξινομεί σε συμπλέγματα/ομαδοποιήσεις (π.χ. γεωγραφικές περιοχές) και να τις αξιολογεί. Ωστόσο για να μπορεί να πραγματοποιείται αυτό σε πραγματικό χρόνο απαιτούνται σημαντικοί υπολογιστικοί πόροι. Τέτοιοι πόροι δεν ήταν διαθέσιμοι στη διάρκεια εκτέλεσης των πειραμάτων.

Επίσης, σημαντική συνεισφορά της εργασίας αποτελεί τόσο η υλοποίηση της συγκεκριμένης εφαρμογής αλγορίθμων συσταδοποίησης βάσει πυκνότητας στη συγκεκριμένη μελέτη περίπτωσης όσο και η σύγκριση των αποτελεσμάτων μεταξύ των δύο αλγορίθμων. Για το τελευταίο θα πρέπει να αναφερθεί ότι δεν έχει αναφερθεί στη βιβλιογραφία αντίστοιχη σύγκριση με αυτή που πραγματοποιήθηκε στη συγκεκριμένη εργασία. Επιπλέον, η εφαρμογή αντικαθιστά την get_assignment μέθοδο της βιβλιοθήκης stream προκειμένου να ανιχνεύσει τη σύνδεση της μετρικής απόστασης με τα ποιοτικά αποτελέσματα στην εκτέλεση των αλγορίθμων. Παρόλο που δεν αποδείχθηκε μία τέτοια διασύνδεση, αποτελεί χρήσιμη μελλοντική αναφορά για αντίστοιχη ερευνητική προσπάθεια.

Σε μελλοντική επέκταση της εργασίας θα πρέπει να βελτιωθεί η απεικόνιση των σημείων στο χάρτη καθώς και των μικρο και μακρο-ομάδων. Στο σύντομο χρόνο που διαθέταμε για την εκμάθηση της γλώσσας R και την υλοποίηση της παραπάνω εφαρμογής μπορέσαμε μόνο να αξιοποιήσουμε βασικές βιβλιοθήκες γραφικής απεικόνισης της R. Αδιαμφισβήτητα, η γλώσσα παρέχει πολύ περισσότερες βιβλιοθήκες και ρουτίνες για την δυναμική απεικόνιση σημείων σε χάρτες και την προβολή τους στο επίπεδο που διευκολύνει ένα χρήστη. Η αξιολόγηση και υλοποίηση τους δεν ήταν εφικτή στη παρούσα φάση.

Τέλος, η αρχιτεκτονική της εφαρμογής (βλ. §3.1) επιτρέπει την αποθήκευση των περιεχομένων των μικρο και μακρο-ομάδων στον τοπικό δίσκο του υπολογιστικού πόρου στον οποίο εκτελείται. Αυτό δίνει τη δυνατότητα της μετέπειτα επεξεργασίας τους από τρίτους οργανισμούς που καλούνται να αξιολογούν κυκλοφοριακά δεδομένα και να λαμβάνουν αποφάσεις ή να ενημερώνουν σε πραγματικό χρόνο. Εφόσον τα δεδομένα αποθηκεύονται σε μία βάση δεδομένων συγκεκριμένες τεχνικές ανάλυσης θα μπορούν να υλοποιηθούν για να απαντήσουν σε ερωτήσεις όπως:

- Σε ποιες περιοχές υπάρχει μποτιλιάρισμα: ένας κανόνας στη περίπτωση αυτή θα εξέταζε τη διακύμανση του μεγέθους της κάθε ομάδας μέσα στο χρόνο. Για παράδειγμα, αν μία μικρο-ομάδα περιείχε σταθερά μεγάλο αριθμό σημείων τότε η ομάδα αυτή είναι μποτιλιαρισμένη, εφόσον αυτό εξετάζεται σε ένα συγκεκριμένο πλαίσιο, π.χ. τα οχήματα έχουν αλλάξει ελαφρώς θέση, δεν είναι αργία ή βραδινές ώρες.
- Να υπολογίσει τη μέση ταχύτητα σε μία περιοχή: στη περίπτωση αυτή εξετάζει τον χρόνο μέσα στον οποίο ένα σημείο (π.χ. ταξί) μεταφέρθηκε από μία μικρο-ομάδα σε μία άλλη μικρο-ομάδα και ποια είναι η απόσταση που κάλυψε. Στη περίπτωση αυτή χρειάζεται η μετρική Μανχάταν που λαμβάνει υπόψη τους δρόμους της πόλης και όχι η Ευκλείδεια απόσταση.
- Να εντοπίσει περιοχές με το μεγαλύτερο ενδιαφέρον και σε ποιες χρονικές στιγμές της ημέρας. Μικρο-ομάδες που συρρικνώνονται σταθερά σε συγκεκριμένες περιοχές καταδεικνύουν τις περιοχές με το μεγαλύτερο ενδιαφέρον σε συγκεκριμένες ώρες της ημέρας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Mansalis S., Ntoutsi E., Pelekis N., Theodoridis Y., 2016. An Evaluation of Data Stream Clustering Algorithms, s.l.: Under Review.

Ξένη

Bo Liu, Erico N. de Souza, Stan Matwin, Marcin Sydowzy, 2014. Knowledge-based Clustering of Ship Trajectories Using Density-based Approach. s.l., 2014 IEEE International Conference on Big Data.

C. Kartika, 2015. Visual Exploration of Spatial-Temporal Traffic Congestion Patterns Using Floating Car Data, s.l.: Technische Universität München.

Cambridge University Press, 2008. Evaluation of clustering. [Ηλεκτρονικό]
Available at: http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html
[Πρόσβαση 06 12 2016].

Feng Cao and Martin Ester and Weining Qian and Aoying Zhou, 2006. Density-based clustering over an evolving data stream with noise. s.l., In 2006 SIAM Conference on Data Mining.

Forrest J., 2011. stream: A Framework for Data Stream Modeling in R, https://lyle.smu.edu/IDA/TRACDS/docs/jhforrest_thesis.pdf: Southern Methodist University.

Győző, 2015. Scalable Detection of Traffic Congestion from Massive Floating Car Data Streams. ACM- Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics, pp. 114-121.

J. A. SILVA, E. R. FARIA, R. C. BARROS,E. R. HRUSCHKA, A. C. P. L. F. DE CARVALHO, 2013. Data Stream Clustering: A Survey. Journal of the ACM, V(N), pp. 1-37.

Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang, 2010. T-drive: driving directions based on taxi trajectories. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10, New York, NY, USA,2010, ACM..

Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun, 2011. Driving with knowledge from the physical world.. In The 17th ACM SIGKDD international conference on Knowledge Discovery and Data mining, KDD'11, New York, NY, USA, 2011., ACM..

João Henrique Alves Pereira, 2014. TweeProfiles2: real-time detection of spatio-temporal patterns in Twitter, Portugal: University of Porto.

Kondaveeti A., 2012. Spatio-Temporal Data Mining to Detect Changes and Clusters in Trajectories, s.l.: ARIZONA STATE UNIVERSITY, PhD Disseration.

Lei Gong, Hitomi Sato, Toshiyuki Yamamoto, Tomio Miwa, Takayuki Morikawa, 2015. Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. Journal of Modern Transportation, 23(3), p. 202–213.

M. Hahsler, M. Bolanos, J. Forrest, n.d. Introduction to Stream: An Extensible Framework for Data Stream Clustering Research with R. [Ηλεκτρονικό]

Available at: https://cran.r-project.org/web/packages/stream/

M. Hahsler, and M. Bolanos, 2016. Clustering Data Streams Based on Shared Density Between Micro-Clusters. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.

Merino, J.A., 2015. Master Thesis: Streaming Data Clustering in MOA using the Leader Algorithm, FACULTAT D'INFORMATICA DE BARCELONA: UNIVERSITAT POLITECNICA DE CATALUNYA.

Michael Hahsler, Matthew Bolanos, John Forrest, 2015. Package 'streamMOA', https://cran.r-project.org/web/packages/streamMOA/streamMOA.pdf: CRAN.

Philipp Kranen, Hardy Kremer, Timm Jansen, Thomas Seidl, Albert Bifet, Geoff Holmes, Bernhard Pfahringer, 2010. Benchmarking Stream Clustering Algorithms within the MOA Framework. Washington, DC, USA, In 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010).

RDM, 2016. Building an R Hadoop System. [Ηλεκτρονικό]
Available at: http://www.rdatamining.com/big-data/rhadoop
[Πρόσβαση 10 09 2016].

Silva, Jonathan A. and Faria, Elaine R. and Barros, Rodrigo C. and Hruschka, Eduardo R. and Carvalho, 2013. Data Stream Clustering: A Survey. ACM Comput. Surv., 46(October 2013), pp. 13:1--13:31.

T. Galba, Z. Balkić, G. Martinović, 2013. Public Transportation BigData Clustering. Preliminary Communication, 4(1).

Voigtlaender, P., 2013. DenStream: Density-based Stream Clustering Algorithm. University of Aachen, http://dme.rwth-aachen.de/en/system/files/file_upload/course/12/elementary-data-mining-techniques/denstreamfinal.pdf.

MOA, 2016. MOA (Massive Online Analysis). [Ηλεκτρονικό]
Available at: http://moa.cms.waikato.ac.nz/
[Πρόσβαση 21 10 2016].