ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

Αξιολόγηση αλγορίθμων συσταδοποίησης ακολουθιακών χωροχρονικών δεδομένων με χρήση διαφορετικών συναρτήσεων απόστασης

Ιωάννης Καρανίκας

Ιούλιος, 2016

Επιβλέπων:

Επικ. Καθηγητής Πελέκης Νικόλαος

UNIVERSITY OF PIRAEUS
Department of Statistics and Insurance Science

Evaluation of clustering algorithms of sequential spatio-temporal data with various distance functions

Ioannis Karanikas

July, 2016

Supervisor:

Assist. Professor Pelekis Nikolaos

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

Αξιολόγηση αλγορίθμων συσταδοποίησης ακολουθιακών χωροχρονικών δεδομένων με χρήση διαφορετικών συναρτήσεων απόστασης

Ιωάννης Καρανίκας

Ιούλιος, 2016

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ

Νικόλαος Πελέκης

Επίκουρος Καθηγητής , Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς (Επιβλέπων Καθηγητής)

Ιωάννης Θεοδωρίδης

Καθηγητής, Τμήμα Πληροφορικής, Πανεπιστήμιο Πειραιώς

Ελευθέριος Κοφίδης

Επίκουρος Καθηγητής, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς

Στην οικογένειά μου για τη συνεχή στήριξη και την υπομονή τους

Ευχαριστίες

Κατά τη διάρκεια των μεταπτυχιακών μου σπουδών στο Πανεπιστήμιο Πειραιώς, αρκετοί άνθρωποι συνέβαλαν, ο καθένας με τον τρόπο του, στο να αποκτήσω το μορφωτικό επίπεδο που έχω σήμερα, και θα ήθελα σε αυτό το σημείο να τους εκφράσω τις ευχαριστίες μου, με την πεποίθηση ότι φάνηκα αντάξιος των προσδοκιών τους.

Η εκπόνηση αυτής της μεταπτυχιακής εργασίας θα ήταν αδύνατη χωρίς τη συμβολή του Επιβλέποντος κ. Νικόλαου Πελέκη, Επικ. Καθηγητή του Πανεπιστημίου Πειραιώς. Η καθοδήγησή του υπήρξε αρωγός στο να εμβαθύνω στο θέμα της εργασίας μου και να παρουσιάσω αρτιότερα τα εξαγόμενα ερευνητικά αποτελέσματα. Συγχρόνως, η διαρκής υποστήριξη και η ενθάρρυνση που μου παρείχε αποτέλεσαν κινητήριο μοχλό για την επιτυχή ολοκλήρωση της μεταπτυχιακής αυτής εργασίας. Για αυτόν το λόγο, όπως και για τη δυνατότητα που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον ερευνητικό θέμα, τον ευχαριστώ θερμά.

Τις ευχαριστίες μου επίσης θα ήθελα να εκφράσω στα άλλα δύο μέλη της Τριμελούς Συμβουλευτικής Επιτροπής, τον κο Ιωάννη Θεοδωρίδη, Καθηγητή του Πανεπιστημίου Πειραιώς, και τον κο Ελευθέριο Κοφίδη, Επικ. Καθηγητή του Πανεπιστημίου Πειραιώς, για το χρόνο που αφιέρωσαν στην ανάγνωση της εργασίας και για τις χρήσιμες υποδείξεις τους.

Εκφράζω επίσης τις ευχαριστίες μου στους κο Κυριάκο Βελισσαρίου και κο Ευστράτιο Μάνσαλη για την πολύτιμη βοήθειά τους στην υλοποίηση των πειραματικών εκτελέσεων της παρούσας μεταπτυχιακής εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω θερμά την αδελφή μου Κλεονίκη, και τους γονείς μου Γεώργιο και Ελένη, για την αμέριστη ηθική συμπαράσταση, υλική υποστήριξη και καθημερινή ενθάρρυνσή τους, σε όλη τη διάρκεια των προπτυχιακών και μεταπτυχιακών μου σπουδών, και ως ένα ελάχιστο δείγμα ευγνωμοσύνης η παρούσα μεταπτυχιακή εργασία αφιερώνεται σε αυτούς.

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως βασικό στόχο τη μελέτη και την αξιολόγηση της απόδοσης αλγορίθμων συσταδοποίησης ακολουθιακών χωροχρονικών δεδομένων, χρησιμοποιώντας κάθε φορά διαφορετικά μέτρα απόστασης/ομοιότητας με σκοπό την ανάδειξη των πλεονεκτημάτων και των μειονεκτημάτων τους. Συγκεκριμένα, εφαρμόζονται μετρικές συναρτήσεις απόστασης (Ευκλείδεια, Manhattan, Chebyshev, Ευκλείδια STARTEND), και μη μετρικές συναρτήσεις απόστασης είτε βασισμένες στη δυναμική χρονική στρέβλωση (Dynamic Time Warping), είτε βασισμένες στην "επεξεργασία" της απόστασης (Edit Distance on Real sequence) ή βασισμένες στη μεγαλύτερη κοινή υποαλληλουχία (Longest Common Subsequence). Ποικίλοι μετασχηματισμοί τροχιών (επαναδειγματολειψία, προσθήκη θορύβου και μετατόπιση σημείου) ελεγχόμενοι από δυο παραμέτρους, τον ρυθμό και την απόσταση, εφαρμόζονται σε πραγματικά και συνθετικά σύνολα δεδομένων τροχιών. Για κάθε μετασχηματισμό, αξιολογείται η ομαδοποίηση του αρχικού συνόλου δεδομένων και των μετασχηματισμένων συνόλων δεδομένων ανάλογα με την τιμή της παραμέτρου που "τρέχει". Τα εξαγόμενα αποτελέσματα της εκτενούς πειραματικής μελέτης χρησιμοποιούνται για την αξιολογήση της εγκυρότητας των ομαδοποιήσεων που επιτυγχάνονται από τον αλγόριθμο optics και την ιεραρχική ομαδοποίηση με τη μέθοδο Ward, αντίστοιχα.

Abstract

This thesis has as main objective to study and evaluate the performance of clustering algorithms considering sequential spatiotemporal data, each time using a different distance/similarity measure in order to highlight its advantages and disadvantages. Specifically, we apply metric distance functions (Euclidean, Manhattan, Chebyshev EuclideanSTARTEND), as well as non-metric distance functions, based either on dynamic time warping (DTW), or on editing distance on real sequence (EDR) or on longest common subsequence (LCSS). Various trajectories transformations (re-sampling, adding noise and point shift) controlled by two parameters, the rate and distance, are applied to real and synthetic trajectory datasets. For each transformation, the clustering of the original data set and the transformed data sets is evaluated depending on the value of the parameter which is not fixed. The results derived from the extensive experimental study are used to assess the validity of clusters obtained by the Optics clustering algorithm and hierarchical clustering via the Ward method, respectively.

Περιεχόμενα

Περιεχόμενα Εικόνων

Περιεχόμενα Πινάκων

1. Εισαγωγή

Η παρούσα διπλωματική εργασία με τίτλο "Αξιολόγηση αλγορίθμων συσταδοποίησης ακολουθιακών χωροχρονικών δεδομένων με χρήση διαφορετικών συναρτήσεων απόστασης" έχει ως βασικό στόχο τη μελέτη και την αξιολόγηση της απόδοσης αλγορίθμων συσταδοποίησης ακολουθιακών χωροχρονικών δεδομένων, χρησιμοποιώντας κάθε φορά διαφορετικά μέτρα απόστασης/ομοιότητας, όπως για παράδειγμα την Ευκλείδεια απόσταση, μέτρα βασισμένα στη δυναμική χρονική στρέβλωση, καθώς επίσης και μέτρα βασισμένα στη μεγαλύτερη κοινή υποαλληλουχία. Συγκεκριμένα, θα μελετήσουμε την απόδοση αλγορίθμων συσταδοποίησης εφαρμόζοντας ποικίλα μέτρα ομοιότητας, με σκοπό να αναδείξουμε τα πλεονεκτήματα και τα μειονεκτήματά τους. Επίσης, θα παρουσιάσουμε συγκριτικά αποτελέσματα μεταξύ των διαφόρων μέτρων ομοιότητας βασισμένοι είτε σε συνθετικά είτε σε πραγματικά χωροχρονικά δεδομένα. Τέλος, θα μελετήσουμε τη συμπεριφορά των προαναφερθέντων μέτρων στα διάφορα προβλήματα που προκύπτουν κατά τη σύγκριση των τροχιών.

Τα χωροχρονικά σύνολα δεδομένων αυξάνονται με ταχύτατους ρυθμούς στην εποχή μας, λόγω των τεχνολογικών και κοινωνικών συνθηκών, καθώς επίσης χρησιμοποιούνται ευρέως και για εμπορικούς σκοπούς. Παρατηρείται καθημερινή συλλογή δεδομένων συναλλαγής μέσω των συστημάτων βάσεων δεδομένων, των ελεγκτών κυκλοφορίας δικτύου, διακομιστές web, αισθητήρες, δίκτυα κ.α. Η σημαντική πρόοδος στη τεχνολογία αισθητήρων, στα GPS και στην ασύρματη επικοινωνία, δημιουργούν μεγάλες ποσότητες δεδομένων που περιγράφουν την ιστορική κίνηση κινούμενων αντικειμένων, που είναι γνωστή ως τροχιά. Τα χωροχρονικά αυτά δεδομένα τροχιάς μπορούν να αναλυθούν μέσω αλγορίθμων εξόρυξης δεδομένων και να υπάρξει σημαντική συνεισφορά σε τομείς εφαρμογών, όπως τα συστήματα περιβαλλοντικών πληροφοριών, μετεωρολογίας, ασύρματης τεχνολογίας, παρακολούθησης βίντεο, ή βίντεο καταγραφής κίνησης.

Όπως προαναφέραμε, οι όλο και μεγαλύτερες ποσότητες δεδομένων που συλλέγονται και αποθηκεύονται σε βάσεις δεδομένων, αυξάνουν και την ανάγκη για αποδοτικές και αποτελεσματικές μεθόδους ανάλυσης, με σκοπό τη χρήση της έμμεσης πληροφορίας που περιέχουν τα δεδομένα και τη βέλτιστη λήψη αποφάσεων. Η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων μέσα από τα στάδια της συλλογής, προεπεξεργασίας, μετασχηματισμού, εξόρυξης δεδομένων, επικύρωσης και ερμηνείας/αξιολόγησης, του αποτελέσματος, στοχεύει στην αυτόματη ή ημιαυτόματη ανάλυση αυτών των μεγάλων ποσοτήτων δεδομένων για την εξαγωγή κάποιου ενδιαφέροντος προτύπου που ήταν άγνωστο μέχρι εκείνη τη στιγμή, όπως για παράδειγμα ομάδες από εγγραφές δεδομένων (συσταδοποίηση), γενίκευση γνωστών δομών για την εφαρμογή τους πάνω σε νέα δεδομένα (κατηγοριοποίηση), τον προσδιορισμό συσχετίσεων ή ακόμη και τον προσδιορισμό ασυνήθιστων εγγραφών δεδομένων.

Μία από τις πιο διαδεδομένες και ευρέως εφαρμόσιμες τεχνικές εξόρυξης δεδομένων είναι η ανάλυση κατά συστάδες. Χρησιμοποιείται ως αυτόνομο εργαλείο για την κατανόηση της κατανομής ενός συνόλου δεδομένων, είτε ως ένα στάδιο επεξεργασίας για άλλους αλγορίθμους που δραστηριοποιούνται στην ανίχνευση συστάδων. Όλος αυτός ο τεράστιος όγκος πληροφορίας είναι και το εφαλτήριο για την εξόρυξη γνώσης με απώτερο στόχο τη δημιουργία προτύπων. Ως εκ τούτου, η ανάπτυξη βελτιωμένων αλγορίθμων ομαδοποίησης έχει λάβει πολλή προσοχή στη διεθνή βιβλιογραφία τα τελευταία χρόνια.

Για τους λόγους που προαναφέραμε, η μέτρηση της ομοιότητας μεταξύ τροχιών είναι αναμφισβήτητα ένα από τα πιο σημαντικά ζητήματα για τη διαχείριση δεδομένων τροχιάς, και αποτελεί τη βάση για πολλές προηγμένες αναλύσεις, όπως είναι η αναζήτηση ομοιότητας, η ομαδοποίηση και η ταξινόμηση. Βιβλιογραφικά έχει προταθεί ένας μεγάλος αριθμός μέτρων ομοιότητας τροχιών και ποικίλα μέτρα απόστασης. Σε γενικές γραμμές, δίνεται έμφαση στην εξέταση της επεκτασιμότητας των μέτρων ομοιότητας, και δεν έχουν διεξαχθεί εκτενή πειράματα για την αξιολόγηση της αποτελεσματικότητάς τους. Στην παρούσα διπλωματική εργασία θα μελετήσουμε και θα αξιολογήσουμε αλγορίθμους συσταδοποίησης θεωρώντας ποικίλα μέτρα ομοιότητας με σκοπό να μελετήσουμε το εάν και το κατά πόσο επηρεάζει κάθε μέτρο ομοιότητας ξεχωριστά την απόδοση των αλγορίθμων.

Η παρουσίαση των επιμέρους θεμάτων και αποτελεσμάτων της εργασίας αυτής οργανώνεται ως εξής:

Στο Κεφάλαιο 2 παρουσιάζονται διάφορες μετρικές και μη μετρικές συναρτήσεις απόστασης οι οποίες χρησιμοποιούνται ευρέως για τον υπολογισμό της ομοιότητας μεταξύ τροχιών κινούμενων αντικειμένων.

Στο Κεφάλαιο 3 παρουσιάζονται οι βασικές έννοιες και ιδιότητες που αφορούν στη διαδικασία της συσταδοποίησης, καθώς και οι βασικοί ορισμοί για τους αλγορίθμους ομαδοποίησης που χρησιμοποιούνται στην παρούσα μελέτη.

Στο Κεφάλαιο 4, παρουσιάζονται οι θεμελιώδεις έννοιες για την αξιολόγηση της εγκυρότητας της συσταδοποίησης. Περιγράφονται αναλυτικά τα κριτήρια (εσωτερικά, εξωτερικά και σχετικά) τα οποία χρησιμοποιούνται για να αξιολογήσουν την ισχύ των ομαδοποιήσεων οι οποίες προκύπτουν από την εφαρμογή αλγορίθμων σε ένα σύνολο δεδομένων.

Στο Κεφάλαιο 5, μελετώνται τα στατιστικά τεστ τα οποία θα μπορούσαν να χρησιμοποιηθούν ή χρησιμοποιούνται ήδη στη διεθνή βιβλιογραφία για τη σύγκριση δυο ή περισσοτέρων αλγορίθμων σε πολλαπλά σύνολα δεδομένων.

Στο Κεφάλαιο 6, περιγράφονται αναλυτικά τα σύνολα δεδομένων τα οποία χρησιμοποιούνται στην παρούσα μελέτη, οι μετασχηματισμοί τροχιών που

εφαρμόζονται στα σύνολα αυτά, καθώς και οι προκύπτουσες πειραματικές παρατηρήσεις-εξαγόμενα αποτελέσματα σε σχέση με την εγκυρότητα των ομαδοποιήσεων που επιτυγχάνονται από τον αλγόριθμο optics και την ιεραρχική ομαδοποίηση με την μέθοδο Ward, αντίστοιχα.

Στο Κεφάλαιο 7, παρουσιάζονται συγκεντρωτικά τα εξαγόμενα συγκριτικά αποτελέσματα της εκτενούς πειραματικής μελέτης για τον αλγόριθμο optics και την ιεραρχική ομαδοποίηση, για κάθε έναν από τους μετασχηματισμούς που εφαρμόστηκαν, για την εσωτερική και εξωτερική αξιολόγηση, αντίστοιχα.

2. Συναρτήσεις Ομοιότητας

2.1 Εισαγωγή

Στην παρούσα μελέτη βασικός σκοπός μας είναι να μελετήσουμε και να αξιολογήσουμε την απόδοση αλγορίθμων συσταδοποίησης ακολουθιακών χωροχρονικών δεδομένων με χρήση διαφορετικών μέτρων υπολογισμού απόστασης.

Για τον καθορισμό της ομοιότητας μεταξύ των σημείων των συστάδων (τα οποία είναι τροχιές κινούμενων αντικειμένων) απαιτείται ο υπολογισμός της απόστασης μεταξύ τους. Στη συνέχεια παρουσιάζονται διάφορες συναρτήσεις υπολογισμού απόστασης οι οποίες χρησιμοποιούνται ευρέως σε τέτοιου είδους μελέτες, οι οποίες και ταξινομούνται σε δύο βασικές κατηγορίες, τις μετρικές και μη μετρικές συναρτήσεις απόστασης.

2.2 Μετρικές Συναρτήσεις Απόστασης

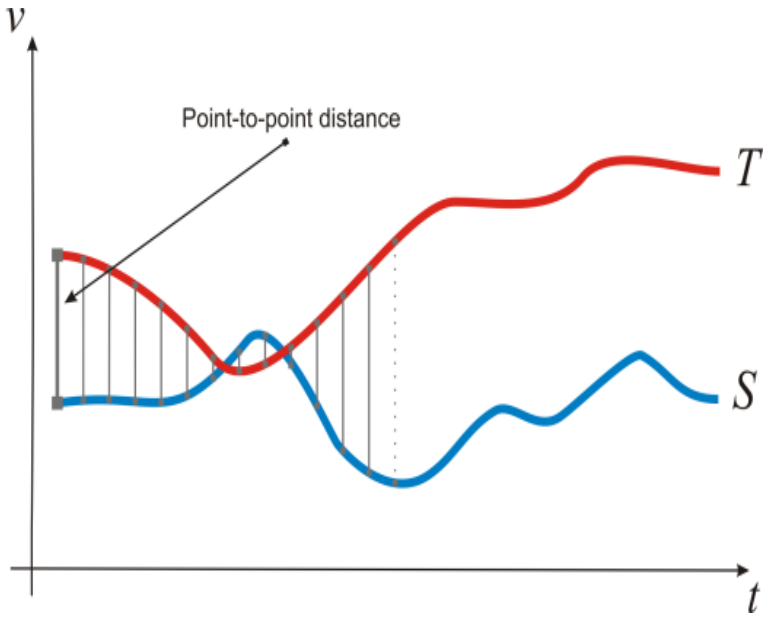
2.2.1 Ευκλείδεια απόσταση

Ο υπολογισμός του αθροίσματος των ευκλείδειων αποστάσεων μεταξύ των αντίστοιχων θέσεων στις δύο τροχιές αποτελεί την πιο απλή προσέγγιση για να προσδιοριστεί η ομοιότητα δύο τροχιών. Δοθέντων δύο τροχιών T_1 και T_2 η Ευκλείδεια απόσταση ορίζεται ως

$$d(T_1, T_2) = \frac{\sum d(p_{1,i}, p_{2,i})}{n},$$

όπου $p_{1,i}$ το σημείο i στην τροχιά T_1, $p_{2,i}$ το σημείο i στην τροχιά T_2, $d(p_{1,i}, p_{2,i})$ είναι η χωρική απόσταση μεταξύ τους, και n το πλήθος των σημείων.

Βασικό πλεονέκτημα της Ευκλείδειας απόστασης αποτελεί ο εύκολος υπολογισμός της, παρ'όλα αυτά όμως δεν αποτελεί το καταλληλότερο μέτρο απόστασης τροχιών για όλες τις περιπτώσεις, καθώς δε λαμβάνει υπόψη της βασικά χαρακτηριστικά των τροχιών, όπως είναι η διεύθυνση, το μήκος τροχιάς κ.α. Ένα επιπλέον μειονέκτημα της Ευκλείδειας απόστασης είναι ότι με τον υπολογισμό των τετραγωνικών αποκλίσεων, οι ακραίες τιμές έχουν μεγάλο αντίκτυπο στη συνολική απόσταση. Μια αποτελεσματική εναλλακτική λύση η οποία αντιμετωπίζει το πρόβλημα των ακραίων τιμών της Ευκλείδειας απόστασης, είναι η χρήση της απόστασης Μανχάταν ή της απόστασης Chebyshev.



Εικόνα 1: Γραφική αναπαράσταση του υπολογισμού της Ευκλείδειας απόστασης

2.2.2 Συναρτήσεις ομοιότητας βασισμένες στην Ευκλείδεια απόσταση

Θεωρώντας δύο τροχιές $p_1, p_2, p_3, \ldots, p_n$ και $q_1, q_2, q_3, \ldots, q_n$, η Ευκλείδεια απόσταση μπορεί να χρησιμοποιηθεί στις συναρτήσεις ομοιότητας Starts only, Ends only και Starts ends (Clarke 1976, Richalet et al. 1978, Jonkery et al. 1980, Sanderson & Wong 1980, Takens 1980, Kallitzaki 2014). Στη συνάρτηση ομοιότητας Starts only, η απόσταση των τροχιών ορίζεται ως η Ευκλείδεια απόσταση των σημείων p_1 και q_1 έναρξης των τροχιών. Στη συνάρτηση ομοιότητας Ends only, η απόσταση των τροχιών ορίζεται ως η Ευκλείδεια απόσταση των σημείων p_n και q_n τερματισμού των τροχιών. Τέλος, στη συνάρτηση ομοιότητας Starts ends, η απόσταση των τροχιών ορίζεται ως η μέση τιμή της Ευκλείδειας απόστασης των σημείων p_1 και q_1 έναρξης των τροχιών, και των σημείων p_n και q_n τερματισμού των τροχιών. Στις προαναφερθείσες συναρτήσεις ομοιότητας η απόσταση των τροχιών υπολογίζεται κάθε φορά χωρίς να λαμβάνεται υπόψη ο χρόνος.

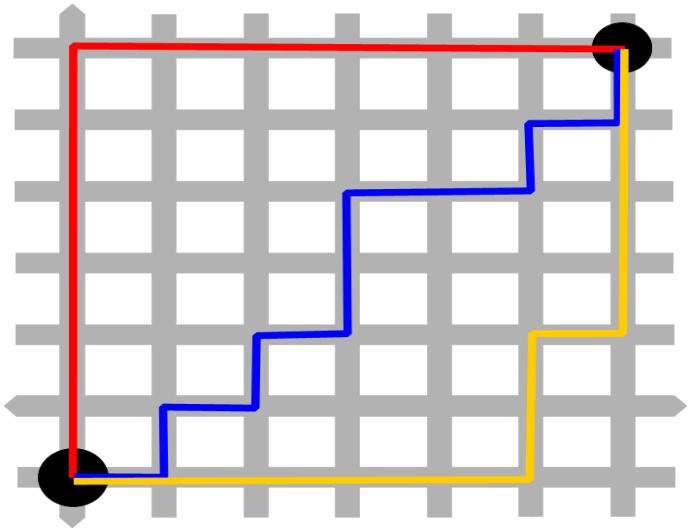
2.2.3 Manhattan απόσταση

Σε ένα n-διάστατο διανυσματικό χώρο, δοθέντων δύο διανυσμάτων R και S, η μαθηματική έκφραση της απόστασης Μανχάταν ορίζεται ως

$$L_1(R,S) = \sum_{i=1}^{n} |r_{i,x} - s_{i,x}| + |r_{i,y} - s_{i,y}|,$$

όπου $r_{i,x}$ η χ-συντεταγμένη του i-οστού στοιχείου του διανύσματος R και $r_{i,y}$ η y-συντεταγμένη του i-οστού στοιχείου του διανύσματος R, $s_{i,x}$ η χ-συντεταγμένη του i-οστού στοιχείου του διανύσματος S και $s_{i,y}$ η y-συντεταγμένη του i-οστού στοιχείου του διανύσματος S, και n το συνολικό πλήθος των σημείων της τροχιάς.

Η απόσταση Μανχάταν είναι περισσότερο αποτελεσματική στη διαχείριση των ακραίων τιμών και του θορύβου συγκριτικά με την Ευκλείδεια απόσταση, επειδή χρησιμοποιεί την απόλυτη τιμή (Kallitzaki 2014).



Εικόνα 2: Γραφική αναπαράσταση του υπολογισμού της απόστασης Μανχάταν

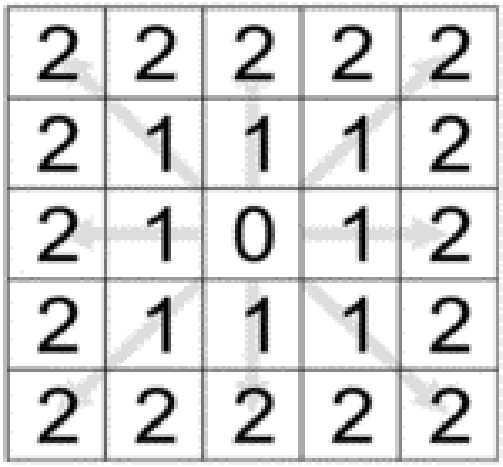
2.2.4 Chebyshev απόσταση

Η απόσταση Chebyshev σε αντίθεση με τις υπόλοιπες αποστάσεις που προαναφέραμε, δεν χρησιμοποιεί όλες τις αποκλίσεις αλλά μόνο τη μεγαλύτερη εξ' αυτών.

Η απόσταση Chebyshev για τις τροχιές R και S ορίζεται ως

$$L(R,S)= \max_{i=1}^{n}((r_{i,x} - s_{i,x}),(r_{i,y} - s_{i,y})),$$

όπου $r_{i,x}$ η χ-συντεταγμένη του i-οστού στοιχείου του διανύσματος R και $r_{i,y}$ η y-συντεταγμένη του i-οστού στοιχείου του διανύσματος R, $s_{i,x}$ η χ-συντεταγμένη του i-οστού στοιχείου του διανύσματος S και $s_{i,y}$ η y-συντεταγμένη του i-οστού στοιχείου του διανύσματος S, και n το συνολικό πλήθος των σημείων της τροχιάς.



Εικόνα 3: Γραφική αναπαράσταση του υπολογισμού της απόστασης Chebyshev

2.2.5 Πλεονεκτήματα και μειονεκτήματα των L_p-μετρικών

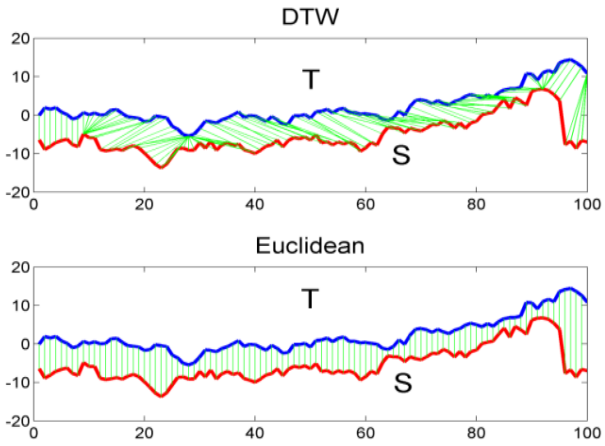
Συνοψίζοντας, τα πλεονεκτήματα των τριών προαναφερθέντων L_p-μετρικών (Ευκλείδεια, Manhattan και Chebyshev) είναι ότι η πολυπλοκότητα του υπολογισμού τους είναι γραμμική, ότι δεν απαιτούν ιδιαίτερες παραμέτρους και ότι είναι ιδιαίτερα απλές στην εφαρμογή και την ευρετηρίασή τους (Agrawal et al. 1993). Τα βασικά μειονεκτήματα αυτών των μετρικών είναι η αναγκαιότητα να έχουν το ίδιο μήκος οι τροχιές, η ευαισθησία τους στο θόρυβο και στις χρονικές αποκλίσεις, και η ανικανότητα χειρισμού της τοπικής χρονικής μετατόπισης.

2.3 Μέτρα βασισμένα στη δυναμική χρονική στρέβλωση

2.3.1 Dynamic Time Warping (DTW)

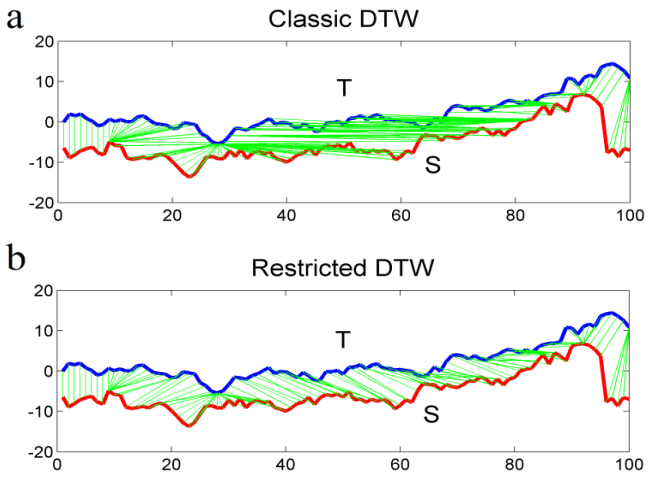
Η συνάρτηση απόστασης της δυναμικής χρονικής στρέβλωσης (Dynamic Time Warping-DTW) η οποία αρχικά εφαρμόστηκε για την αναγνώριση ομιλίας (Berndt and Clifford 1996), μπορεί επίσης να χρησιμοποιηθεί για τον υπολογισμό της ομοιότητας μεταξύ δύο τροχιών που διαφέρουν σε μήκος ή ταχύτητα. Το βασικό χαρακτηριστικό της DTW είναι ότι επιτρέπει μια αλληλουχία να "τεντώνεται" ή να "συρρικνώνεται" προκειμένου να ταιριάζει καλύτερα με μία άλλη αλληλουχία (Pelekis & Theodoridis 2014). Αν δύο τροχιές είναι παρόμοιες, το γεγονός ότι δεν είναι ευθυγραμμισμένες στο χρόνο, έχει ως αποτέλεσμα η Ευκλείδεια απόσταση να ταιριάζει ένα-ένα τα σημεία των δύο σειρών και ψευδώς να καταλήγει στο συμπέρασμα ότι υπάρχει μεγάλη απόσταση μεταξύ τους. Αυτό διορθώνεται από τη συνάρτηση απόστασης DTW με γραμμική αντιστοίχιση. Η DTW είναι μία βελτίωση των μετρικών συναρτήσεων, αλλά και αυτή είναι ευαίσθητη στο θόρυβο καθώς όλα τα σημεία των τροχιών πρέπει να αντιστοιχηθούν, ακόμα και οι ακραίες τιμές. Η

DTW σε αντίθεση με τις προαναφερθείσες μετρικές, ταιριάζει κάθε σημείο της πρώτης τροχιάς με το κοντινότερο της δεύτερης, και τελικά επιλέγει τη συντομότερη απόσταση.



Εικόνα 4: Σύγκριση λειτουργίας DTW και Ευκλείδιας απόστασης

Η υπολογιστική πολυπλοκότητα της DTW είναι τετραγωνική $O(n^2)$. Στην αρχική του εκδοχή ο αλγόριθμος υπολογισμού της απόστασης DTW δεν περιέχει παραμέτρους (Wang et al. 2013). Ωστόσο, η επιβολή ενός προσωρινού περιορισμού στο μέγεθος του παραθύρου στρέβλωσης, βελτιώνει την αποτελεσματικότητα του αλγορίθμου στους υπολογισμούς και την ακρίβειά του, όσον αφορά τη μέτρηση της ομοιότητας των τροχιών, επειδή στην εκτεταμένη στρέβλωση υπάρχει μεγάλη πιθανότητα εσφαλμένου ταιριάσματος μεταξύ δύο τροχιών καθώς και εμφάνιση ομοιότητας που δεν υπάρχει στην πραγματικότητα. Η στρέβλωση με περιορισμό χρησιμοποιείται για την ανάπτυξη της μικρότερης ορισμένης απόστασης και για την ευρετηρίαση των τροχιών (Vlachos et al. 2002a).



Εικόνα 5: Σύγκριση λειτουργίας της "classic" και "restricted" DTW

Έστω οι τροχιές δύο κινούμενων αντικειμένων R και S, τότε η συνάρτηση DTW ορίζεται ως ακολούθως:

$$DTW(R,S) = L_p(r_n, s_m) + min\begin{Bmatrix} DTW(R, Head(S)), \\ DTW(Head(R), S), \\ DTW(Head(R), Head(S) \end{Bmatrix}$$

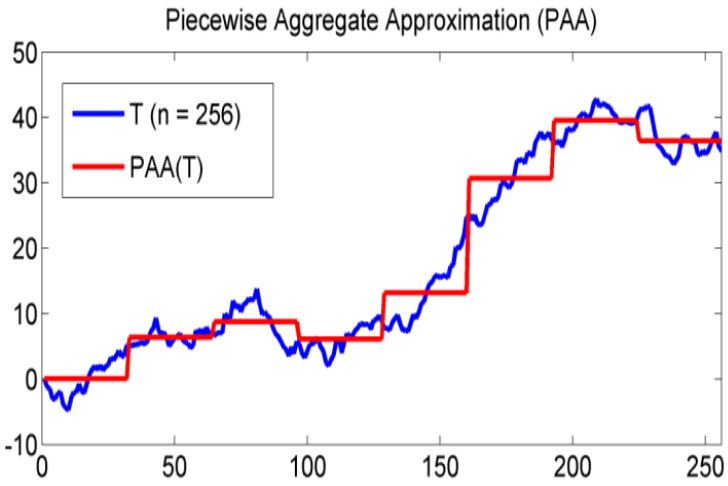
όπου Head (R) = (($r_{1,x}$, $r_{1,y}$) ... ($r_{n-1,\ x}$, $r_{n-1,\ y}$)) και η L_p μπορεί να είναι οποιαδήποτε L μετρική.

2.3.2 Piecewise Dynamic Time Warping (PDTW)

Για να επιταχυνθεί η συνάρτηση απόστασης DTW και να μειωθεί το υπολογιστικό της κόστος, έχουν εισαχθεί αρκετές μέθοδοι κλαδέματος, όπως είναι η μέθοδος FastMap και η μέθοδος κατωτάτου ορίου (Sakurai et al. 2005,Yi et al. 1998). Η συνάρτηση απόστασης PDTW (Keogh & Pazzani, 2000), επιταχύνει την DTW με μια σταθερά c η οποία εξαρτάται από τα δεδομένα. Δοθέντων δύο τροχιών T_1 και T_2, η PDTW εκτελεί τα ακόλουθα δύο βήματα:

1. Piecewise Aggregate Approximation (PAA) δηλ. διαχωρίζει τη τροχιά σε c τμήματα, όπου το i-οστό τμήμα είναι το παρακάτω:

$$[p_{c\times(i-1)+1}, p_{c\times(i-1)+2}, \dots, p_{c\times i}].$$



Εικόνα 6: Η διαδικασία PAA

Για κάθε τμήμα *i*, η PAA υπολογίζει το $\overline{p_i}$ ως αντιπροσωπευτικό σημείο, και με αυτό τον τρόπο μετασχηματίζει την τροχιά T σε $\overline{T} = [\overline{p_1}, \overline{p_2}, \overline{p_3}, \dots, \overline{p_c}]$.

2. Η PDTW τρέχει την DTW απόσταση για να βρει όμοια μοτίβα τροχιών μεταξύ των μετασχηματισμένων τροχιών $\overline{T_1}$ *και* $\overline{T_2}$.

2.4 Μεγαλύτερη κοινή υποακολουθία

2.4.1 Longest Common Subsequence (LCSS)

Η συνάρτηση απόστασης της "μεγαλύτερης κοινής υποακολουθίας" (Longest Common Subsequence-LCSS) είναι ένα άλλο μέτρο ομοιότητας που προσπαθεί να ταιριάξει δύο χρονοσειρές, επιτρέποντάς τους το «τέντωμα», χωρίς όμως να αλλάζει η αλληλουχία των στοιχείων, ενώ επιτρέπει κάποια στοιχεία από τις σειρές να μείνουν χωρίς ταίρι (Bollobas et al. 2001). Το κύριο πλεονέκτημά της σε σχέση με τις τεχνικές που παρουσιάστηκαν παραπάνω, είναι ότι η LCSS χειρίζεται το θόρυβο πιο αποτελεσματικά, διότι μπορεί να αγνοήσει θορυβώδη σημεία (με το να μην έχουν ταίρι απαραίτητα) σε αντίθεση με άλλες τεχνικές των οποίων όλα τα σημεία πρέπει να έχουν ταίρι (ακόμα και τα θορυβώδη).

Η LCSS χρησιμοποιεί δύο παραμέτρους, την παράμετρο δ η οποία δείχνει τη χρονική κλίμακα όπου η μέθοδος προσπαθεί να ταιριάξει τα στοιχεία μεταξύ τους, και την παράμετρο ε η οποία είναι ένα όριο απόστασης το οποίο δείχνει το κατά πόσο δύο σημεία ταιριάζουν ή όχι (Vlachos et al. 2002b).

Η LCSS μεταξύ δύο τροχιών κινούμενων αντικειμένων, *R* και *S*, δίνεται από την ακόλουθη εξίσωση:

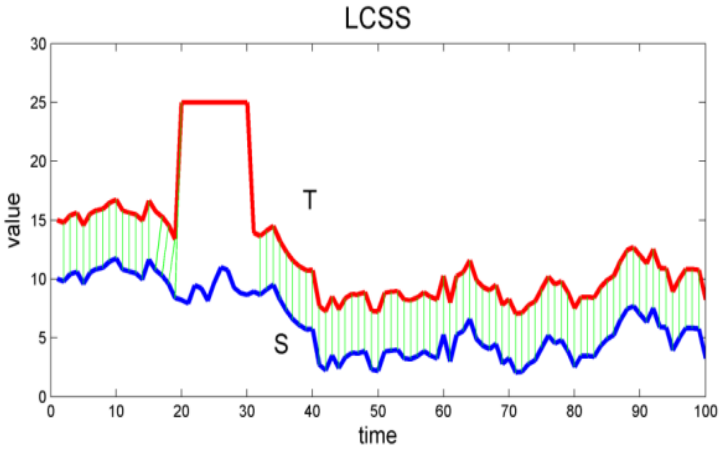
$$LCSS(R,S) = \begin{cases} 0, & m = 0 \; or \; n = 0 \\ LCSS\big(Head(R), Head(S)\big) + 1, & if \; |r_{n,x} - s_{m,x}| < \varepsilon \\ & and \; |r_{n,y} - s_{m,y}| < \varepsilon \\ & and \; |n - m| \le \delta \\ \max\{LCSS(Head(R), S), LCSS\big(R, Head(S)\big)\}, & otherwise \end{cases}$$

όπου Head (R) = (($r_{1,x}$, $r_{1,y}$) ... ($r_{n-1,x}$, $r_{n-1,y}$)), δ είναι ένας ακέραιος αριθμός και ε είναι ένας πραγματικός αριθμός.

Έχοντας υπολογίσει την ομοιότητα μεταξύ δύο τροχιών (όπως δείξαμε παραπάνω), η απόσταση μεταξύ τους υπολογίζεται με την εξίσωση

$$D_{\delta,\varepsilon}(R,S) = 1 - \frac{LCSS_{\delta,\varepsilon}(R,S)}{\min(n,m)}$$

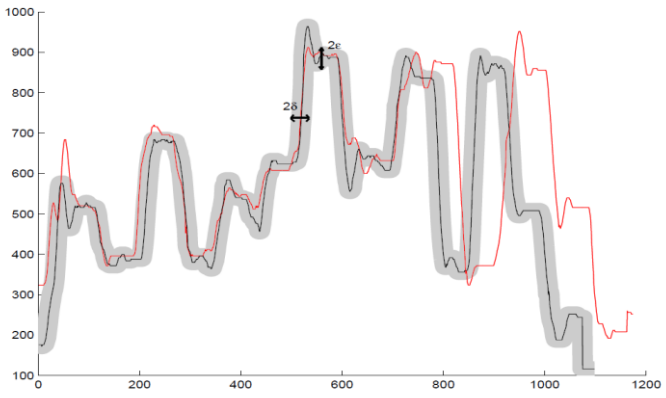
η οποία κανονικοποιεί την τιμή ομοιότητας σε σχέση με το μήκος των συγκρινόμενων τροχιών:



.

Εικόνα 7: Η συνάρτηση απόστασης LCSS

Στην Εικόνα 8 που ακολουθεί, παρατηρούμε διαισθητικά τη βασική ιδέα των παραμέτρων δ και ε της συνάρτησης απόστασης LCSS. Συγκεκριμένα, τα σημεία των δύο τροχιών που βρίσκονται μέσα στην γκρίζα περιοχή, η οποία καθορίζεται από τις παραμέτρους δ και ε, μπορούν να συνδυαστούν.



Εικόνα 8: Οι παράμετροι δ και ε της LCSS

2.5 Μέτρα βασισμένα στην "επεξεργασία" της απόστασης

2.5.1 Edit Distance on Real sequence (EDR)

Η συνάρτηση απόστασης EDR (Chen et al. 2005) βασίζεται στη γνωστή λειτουργία επεξεργασίας της απόστασης που έχει χρησιμοποιηθεί με επιτυχία σε ποικίλες εφαρμογές, όπως για παράδειγμα στη βιοπληροφορική και στην αναγνώριση φωνής, όπου ένα σημαντικό ζητούμενο είναι να ποσοτικοποιηθεί η ομοιότητα μεταξύ δύο συμβολοσειρών. Δοθέντων δύο συμβολοσειρών, η συνάρτηση απόστασης EDR υπολογίζει τον ελάχιστο αριθμό των εισαγωγών, των διαγραφών και αντικαταστάσεων προκειμένου οι δύο συμβολοσειρές να γίνουν πανομοιότυπες. Αντίστοιχα με τη συνάρτηση LCSS, η EDR χρησιμοποιεί επίσης ένα όριο ε για να ταιριάξει τα δύο σημεία, όμως εδώ η απόσταση μπορεί να πάρει τις τιμές 0 ή 1, ανάλογα με το αν ταιριάζουν τα σημεία ή όχι. Η εφαρμογή αυτού του κριτηρίου, δηλαδή το αν ταιριάζουν τα δύο σημεία r_i και s_j των τροχιών κινούμενων αντικειμένων περιγράφεται ως ακολούθως:

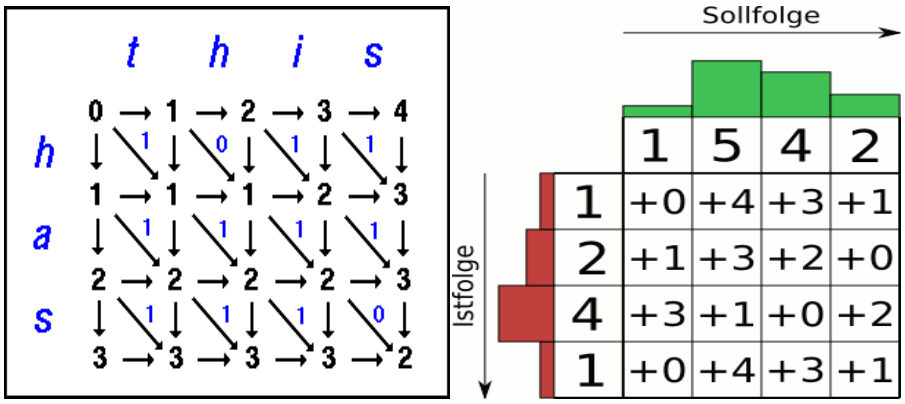
$$match(r_i, s_j) = \begin{cases} 1, & if\ |r_{i,x} - s_{j,x}| \le \varepsilon\ and\ |r_{i,y} - s_{j,y}| \le \varepsilon \\ 0, & else \end{cases}.$$

Λόγω του προαναφερθέντος "ταιριάσματος" των σημείων, οι ακραίες τιμές έχουν πολύ μικρότερη επίδραση στη συνολική απόσταση σε σύγκριση με την Ευκλείδεια απόσταση ή τη συνάρτηση απόστασης DTW. Από την άλλη πλευρά, μία κρίσιμη διαφορά με τη συνάρτηση απόστασης LCSS, είναι ότι η EDR προσθέτει μία ποινή για τα διάκενα μεταξύ δύο σημείων που έχει ταιριάξει, η οποία είναι ανάλογη προς το μήκος διάκενου. Αυτό οδηγεί σε υψηλότερη ακρίβεια όσον αφορά τον υπολογισμό της απόστασης, σε αντίθεση με τη συνάρτηση LCSS. Λαμβάνοντας υπόψη τα ανωτέρω, η συνάρτηση απόστασης EDR για δυο τροχιές κινούμενων αντικειμένων, έστω *R* και *S*, με μήκη *n* και *m*, αντίστοιχα, ορίζεται ως:

$$EDR(R,S) = \begin{cases} n, & if\ m = 0 \\ m, & if\ n = 0 \\ \min\begin{cases} EDR\big(Rest(R), Rest(S)\big) + subcost, \\ EDR(Rest(R), S) + 1, \\ ERP\big(R, Rest(S)\big) + 1 \end{cases}, & otherwise \end{cases}$$

όπου $Rest(R) = ((r_{2,x}, r_{2,y}) \dots (r_{n,x}, r_{n,y}))$, και το subcost ορίζεται ως:

$$subcost = \begin{cases} 0, & if\ match(r_1, s_1) = 1 \\ 1, & otherwise \end{cases}.$$



Εικόνα 9: Η λειτουργία της EDR

Οι συναρτήσεις που παρουσιάστηκαν παραπάνω (βλ. Ενότητες 2.1, 2.2, 2.3, 2.4) μπορούν να ταξινομηθούν σε δύο ευρείες κατηγορίες: μετρικές συναρτήσεις, οι οποίες δεν μπορούν να χειριστούν τη χρονική μετατόπιση των τροχιών, και μη μετρικές συναρτήσεις οι οποίες υποστηρίζουν τη χρονική μετατόπιση.

2.5.2 Edit Distance with Real Penalty (ERP)

Η συνάρτηση ERP (Chen and Ng 2004) αποτελεί έναν συνδυασμό μετρικών συναρτήσεων απόστασης που δεν μπορούν να υποστηρίξουν τη χρονική μετατόπιση, και μη μετρικών συναρτήσεων απόστασης που υποστηρίζουν τη χρονική μετατόπιση. Η συνάρτηση ERP συνδυάζει τα πλεονεκτήματα των μετρικών και των μη μετρικών συναρτήσεων απόστασης καθώς υποστηρίζει την τοπική χρονική μετατόπιση και είναι και μετρική.

Η συνάρτηση ERP χρησιμοποιεί πραγματικές ποινές μεταξύ δύο μη κενών στοιχείων και μία σταθερή τιμή για τον υπολογισμό της απόστασης των κενών. Ο αλγόριθμος που χρησιμοποιεί είναι ο ακόλουθος:

$$ERP(R,S) = \begin{cases} \sum_{i=1}^{n} dist(s_i, g), & if\ m = 0 \\ \sum_{i=1}^{n} dist(r_i, g), & if\ \ n = 0 \\ \min \begin{cases} ERP\big(Rest(R), Rest(S)\big) + dist(t_1, s_1), \\ ERP(Rest(R), S) + dist(r_1, g), \\ ERP\big(R, Rest(S)\big) + dist(s_1, g) \end{cases}, & otherwise \end{cases}$$

όπου: $dist_{erp}(s_i, r_i) = \begin{cases} \sum_{i=1}^{n} |r_i - s_i|, & \text{εάν } r_i, s_i \text{ δεν είναι κενά} \\ \sum_{i=1}^{n} |r_i - g|, & \text{εάν } s_i \text{ είναι κενό} \\ \sum_{i=1}^{n} |s_i - g|, & \text{εάν } r_i \text{ είναι κενό} \end{cases}$

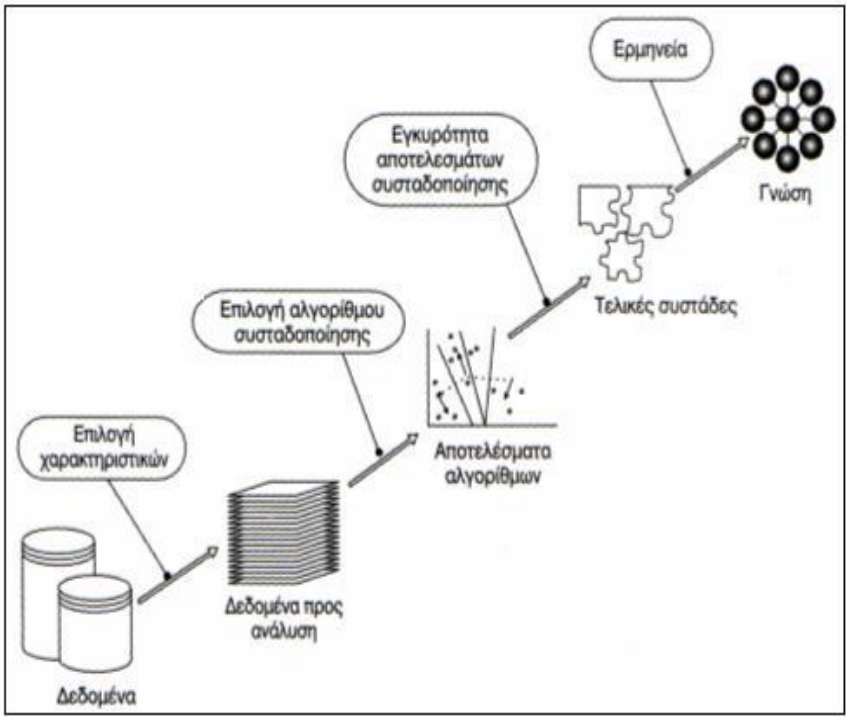
όπου Rest(R) =(($r_{2,x}$, $r_{2,y}$)… ($r_{n,x}$, $r_{n,y}$)), και g είναι μία σταθερή τιμή.

3. Συσταδοποίηση

3.1 Εισαγωγή

Η ομαδοποίηση (συσταδοποίηση) δεδομένων είναι ο συνηθέστερος τρόπος εξόρυξης γνώσης από δεδομένα, καθώς αποτελεί τρόπο μάθησης χωρίς επίβλεψη, που σημαίνει ότι δεν χρειάζεται ιδιαίτερη γνώση για τη φύση των δεδομένων προκειμένου να διαχωριστούν αυτά σε ομάδες. Ένας ορισμός της ομαδοποίησης (clustering) είναι ο ακόλουθος. Ομαδοποίηση είναι η διαδικασία κατά την οποια ένας πληθυσμός αντικειμένων με διαφορετικά χαρακτηριστικά ομαδοποιείται σε ένα σύνολο από διαφορετικές ομάδες. Σε αντίθεση με την κατηγοριοποίηση, η ομαδοποίηση δεν χρησιμοποιεί προκαθορισμένες κατηγορίες, αλλά τα δεδομένα ομαδοποιούνται με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους. Τέλος, βασικός σκοπός της ομαδοποίησης είναι η μεγιστοποίηση της ομοιότητας μέσα στις ομάδες και η ελαχιστοποίηση της ομοιότητας ανάμεσα στις ομάδες.

Η διαδικασία της ομαδοποίησης μπορεί να χωριστεί στα ακόλουθα στάδια (Kallitzaki 2014). Να σημειωθεί ότι πριν την εφαρμογή της διαδικασίας είναι απαραίτητη η προεπεξεργασία των δεδομένων συσταδοποίησης. Πρώτον, επιλέγονται τα πιο χαρακτηριστικά γνωρίσματα στα οποία μπορεί να εφαρμοστεί η συσταδοποίηση με στόχο την επίτευξη της βέλτιστης ομοιογένειας σε κάθε συστάδα. Δεύτερον, επιλέγεται ο καλύτερος αλγόριθμος συσταδοποίησης για το σύνολο δεδομένων με στόχο την επίτευξη ενός βέλτιστου σχήματος συσταδοποίησης με την καλύτερη δυνατή προσαρμογή στα δεδομένα μας. Τρίτον, αξιολογούνται τα εξαγόμενα αποτελέσματα του αλγορίθμου συσταδοποίησης με τη χρήση κατάλληλων κριτηρίων επικύρωσης. Τέλος, ερμηνεύονται και παρουσιάζονται τα εξαγόμενα αποτελέσματα τα οποία προέκυψαν από την ως άνω διαδικασία συσταδοποίησης.



Εικόνα 10: Τα βήματα της συσταδοποίησης

Κατά την υλοποίηση των αλγορίθμων συσταδοποίησης, μπορεί να προκύψουν ορισμένα σημαντικά ζητήματα προς επίλυση (Kallitzaki 2014) ορισμένα από τα οποία είναι:

- Η ύπαρξη ακραίων σημείων τα οποία συνατάμε σε κάθε πραγματικό σύνολο δεδομένων. Τα ακραία σημεία αποτελούν τιμές που δεν ανταποκρίνονται σε υπαρκτά σημεία. Το φαινόμενο αυτό μπορεί να προκαλείται από παρεμβολές σήματος.

- Ο άγνωστος επιθυμητός αριθμός συστάδων. Το πλήθος των συστάδων που θα ήθελε ο πειραματιστής να σχηματιστεί δεν είναι εκ των προτέρων γνωστό, και γι'αυτό το λόγο ο αλγόριθμος θα πρέπει να έχει τη δυνατότητα να εντοπίζει αυτό το σύνολο των συστάδων.

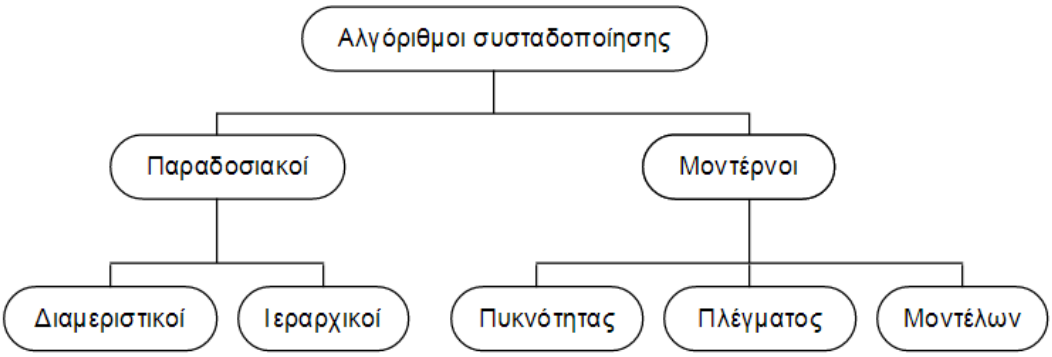
- Το να έχουμε δυναμικά μεταβαλλόμενα δεδομένα ή με πιο απλά λόγια οι συστάδες να αλλάζουν στην πορεία του χρόνου.

- Να μπορεί να γίνει κλιμάκωση στον αριθμό των σημείων και των διαστάσεων των αντικειμένων.

- Τα εξαγόμενα αποτελέσματα κάθε αλγορίθμου να μπορούν να αξιολογηθούν και να ερμηνευτούν ούτως ώστε να καταλήγουμε σε ασφαλή εξαγωγή συμπερασμάτων.

3.2 Κατηγορίες μεθόδων συσταδοποίησης

Υπάρχει ένας μεγάλος αριθμός αλγορίθμων που ασχολούνται με τη συσταδοποίηση, οι οποίοι μπορούν να ομαδοποιηθούν σε πέντε ομάδες όπως αυτές απεικονίζονται στο παρακάτω διάγραμμα. Στις σύγχρονες εφαρμογές, όπου το πλήθος των δεδομένων είναι πολύ μεγάλο και οι αντίστοιχες υπολογιστικές ανάγκες υπερβολικές, το πρόβλημα της σωστής επιλογής της κατάλληλης μεθόδου γίνονται πολύ πιο έντονο.



Εικόνα 11: Αλγόριθμοι συσταδοποίησης

Η παρούσα εργασία μελετά από τους παραδοσιακούς αλγορίθμους συσταδοποίησης τους ιεραρχικούς και από τους μοντέρνους τους αλγορίθμους των οποίων η ομαδοποίηση βασίζεται στην πυκνότητα. Πιο συγκεκριμένα, οι εν λόγω αλγόριθμοι οι οποίοι εφαρμόστηκαν και στην πειραματική μας μελέτη είναι ο αλγόριθμος της ιεραρχικής ομαδοποίησης με χρήση της μεθόδου Ward, και ο βασισμένος στην πυκνότητα αλγόριθμος Optics, αντίστοιχα.

3.2.1 Ιεραχικοί αλγόριθμοι συσταδοποίησης

Στις ιεραρχικές μεθόδους οι ομάδες σχηματίζονται σταδιακά είτε με συνένωση μικρότερων ομάδων σχηματίζοντας συνεχώς μεγαλύτερες ομάδες μέχρι να φτάσουμε να έχουμε όλα τα δεδομένα σε μια ομάδα (συσσωρευτικές μέθοδοι), είτε με διαίρεση

ομάδων σε μικρότερες μέχρι να φτάσουμε σε μια κατάσταση όπου κάθε παρατήρηση να είναι από μόνη της μια ομάδα (διαιρετικές μέθοδοι).

Οι ιεραρχικές μέθοδοι καλό είναι να αποφεύγεται να χρησιμοποιούνται για μεγάλο πλήθος δεδομένων αφού απαιτούν πολύ χρόνο, μνήμη και υπολογιστική ισχύ. Επίσης, υπάρχει η τάση να δημιουργούνται ομάδες με ανομοιογενές μέγεθος.

3.2.2 Αλγόριθμοι ομαδοποίησης βασισμένοι στην πυκνότητα

Οι αλγόριθμοι συσταδοποίησης οι βασισμένοι στην πυκνότητα χαρακτηρίζονται από δύο βασικά γνωρίσματα, την ανθεκτικότητά τους σε προβλήματα ύπαρξης θορύβου και ακραίων τιμών και την ικανότητα αναγνώρισης διαφορετικών συχνοτήτων στο σύνολο των σημείων. Οι προαναφερθείσες ιδιότητες είναι ιδιαίτερα σημαντικές για ένα πραγματικό σύνολο δεδομένων το οποίο είναι πολύ πιθανόν να έχει ακραίες τιμές και να μην έχει ταξινομημένη πυκνότητα στο εσωτερικό της δομής της συστάδας. Σε μία τέτοια περίπτωση, δεν αρκεί η χρήση μιας γενικής παραμέτρου πυκνότητας αλλά απαιτείται ο υπολογισμός ενός μεγάλου συνόλου τοπικών πυκνοτήτων ούτως ώστε να αναγνωριστούν οι συστάδες σε διαφορετικές περιοχές.

3.3 Ιεραρχική ομαδοποίηση με τη μέθοδο Ward

Στους αλγορίθμους ιεραρχικής ομαδοποίησης, για τον καθορισμό των αποστάσεων μεταξύ των ομάδων, έχουν προταθεί στη μέχρι σήμερα διεθνή βιβλιογραφία διάφορες τεχνικές, με μία από αυτές να είναι η μέθοδος Ward.

Ο αλγόριθμος της ιεραρχικής ομαδοποίησης με τη μέθοδο Ward διαφέρει από τις υπόλοιπες μεθόδους και είναι σχεδιασμένος με στόχο την ελαχιστοποίηση της διακύμανσης μέσα στις ομαδες. Η μέθοδος δημιουργεί ομάδες με παρόμοιο αριθμό παρατηρήσεων, χαρακτηρίζεται από πολύ καλές ιδιότητες, και για αυτό το λόγο χρησιμοποιείται σε διάφορες πρακτικές εφαρμογές.

Συνοπτικά ο αλγόριθμος της ιεραρχικής ομαδοποίησης ακολουθεί τα παρακάτω βήματα. Ξεκινάμε με *n* ομάδες του ενός ατόμου η καθεμιά. Εντοπίζει τις δύο πλησιέστερες ομάδες, έστω *Q* και *R* και έπειτα συγχωνεύει τις εν λόγω ομάδες σε μία, οπότε ο αριθμός των ομάδων μειώνεται κατά ένα. Με τον ανανεωμένο πλέον πίνακα ομοιότητας επαναλαμβάνει τα προηγούμενα βήματα συνολικά *n-1* φορές έως ότου όλα τα άτομα αποτελούν μία μόνο ομάδα.

Με τη μέθοδο Ward η απόσταση μεταξύ δύο ομάδων ορίζεται συναρτήσει της Ευκλείδειας απόστασης μεταξύ των κέντρων βαρών τους μέσω της ακόλουθης μαθηματικής έκφρασης:

$$d^2(R,Q) = \frac{2|R||Q|}{|R|+|Q|} d^2\big(\bar{x}(R), \bar{x}(Q)\big),$$

και ο μαθηματικός τύπος ανανέωσης ορίζεται ως ακολούθως:

$$d^2(R,Q) = \frac{|A|+|Q|}{|R|+|Q|}d^2(A,Q) + \frac{|B|+|Q|}{|R|+|Q|}d^2(B,Q) - \frac{|Q|}{|R|+|Q|}d^2(A,B).$$

3.4 Ο αλγόριθμος Optics

Ένας από τους σημαντικότερους και ευρέως διαδεδομένους αλγορίθμους συσταδοποίησης είναι ο αλγόριθμος OPTICS (Ankerst et al. 1999) και αποτελεί επέκταση του αλγορίθμου DBSCAN. Ο αλγόριθμος DBSCAN είναι ένας αλγόριθμος βασισμένος στην πυκνότητα η οποία ορίζεται ως ο ελάχιστος προκαθορισμένος αριθμός σημείων (MinPts) μέσα σε μια προκαθορισμένη ακτίνα (Eps). Ο DBSCAN διαχωρίζει τα σημεία σε βασικά ή σημεία πυρήνα (ένα σημείο για το οποίο υπάρχουν περισσότερα από ένα προκαθορισμένο αριθμό (MinPts) σημεία σε ακτίνα Eps και είναι τα σημεία που είναι στο εσωτερικό μιας συστάδας), σε οριακά (ένα σημείο για το οποίο υπάρχουν λιγότερα από ένα προκαθορισμένο αριθμό (MinPts) σημεία σε ακτίνα Eps, αλλά είναι στη γειτονιά ενός βασικού σημείου) και σε σημεία θορύβου (ένα σημείο που δεν είναι ούτε βασικό ούτε οριακό). Αφού χαρακτηρίσει κάθε σημείο ως βασικό (core), οριακό (border) ή θορύβου (noise), διαγράφει τα σημεία θορύβου. Έπειτα, τοποθετεί μια ακμή μεταξύ όλων των βασικών σημείων που είναι σε απόσταση έως Eps μεταξύ τους, και κάνει κάθε ομάδα των συνδεδεμένων βασικών σημείων μια διαφορετική συστάδα. Τέλος, αναθέτει κάθε οριακό σημείο σε μία από τις συστάδες των συσχετιζόμενών του βασικών σημείων.

Ως κριτήριο ομαδοποίησης ο αλγόριθμος OPTICS χρησιμοποιεί την πυκνότητα χωρικών δεδομένων. Χαρακτηριστικό γνώρισμα του αλγορίθμου συσταδοποίησης Optics είναι ότι δοθείσης μιας ακτίνας *ε*, για κάθε στοιχείο μιας συστάδας, η γειτονιά του πρέπει να περιέχει τουλάχιστον ένα ελάχιστο πλήθος αντικειμένων. Οι τιμές εισόδου, οι οποίες καθορίζουν την λειτουργία του αλγορίθμου, είναι τα ελάχιστα στοιχεία από τα οποία μπορεί να αποτελείται μια ομάδα και ένα κατώφλι απόστασης *ε* το οποίο αντιπροσωπεύει την ακτίνα της γειτονιάς στην οποία βρίσκονται τα στοιχεία της εκάστοτε ομάδας. Το γραφικό αποτέλεσμα του αλγορίθμου Optics είναι το διάγραμμα γειτνίασης (reachability plot), το οποίο αποτελεί μια ανεξάρτητη από τα δεδομενα οπτικοποίηση της δομής της ομαδοποίησης. Η οπτικοποίηση των δεδομένων είναι ένα επιπλέον εργαλείο για την καλύτερη δυνατή κατανόηση του συνόλου δεδομένων το οποίο προσδίδει πληροφορία για το πού εκχωρείται κάθε στοιχείο, δηλαδή στην αντίστοιχη συστάδα ή στο θόρυβο, αντίστοιχα.

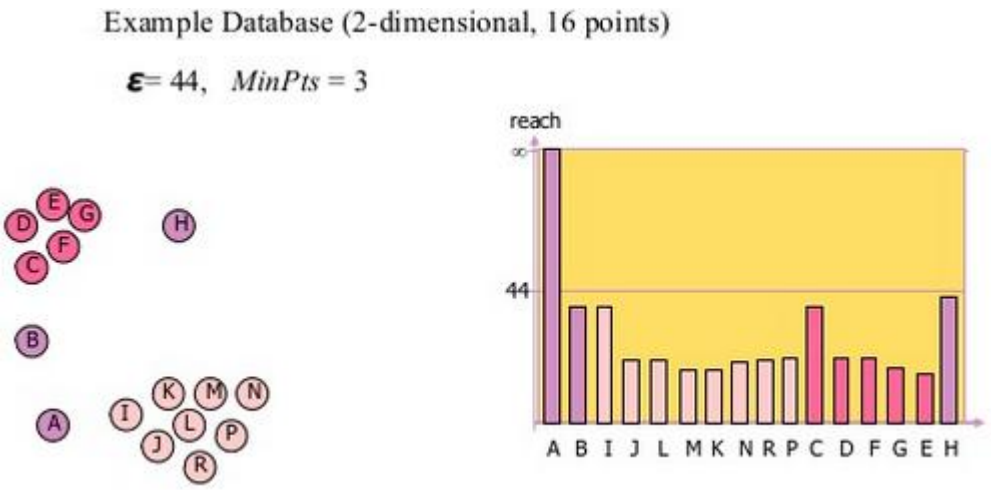
Η απόσταση εγγύτητας μεταξύ δύο σημείων είναι η απόστασή τους. Έστω δύο σημεία *p* και *o*, όταν το *p* είναι πολύ κοντά στο *o*, η απόσταση κανονικοποιείται σε μία κατάλληλη τιμή. Τα βασικά βήματα του αλγορίθμου Optics είναι τα ακόλουθα. Έστω *D* το σύνολο των τροχιών κινούμενων αντικειμένων.Πρώτον, επιλέγεται ένα

αντικείμενο p_0 με τυχαίο τρόπο. Στη συνέχεια, για i επαναλήψεις, το επόμενο αντικείμενο p_i που επιλέγεται από το σύνολο D, είναι αυτό που έχει τη μικρότερη απόσταση γειτνίασης σε σχέση με όλα τα υπόλοιπα αντικείμενα του πυρήνα που έχουν ήδη ελεγχθεί. Τέλος, η διαδικασία επαναλαμβάνεται έως ότου ελεγχθούν όλα τα αντικείμενα του συνόλου D (Kallitzaki 2014).

Το σύνολο της ως άνω διαδικασίας υλοποίησης του αλγορίθμου Optics μπορεί να αναπαρασταθεί σε ένα διάγραμμα γειτνίασης όπου στον κατακόρυφο άξονα κάθε i απόσταση γειτνίασης αντιστοιχίζεται στο αντικείμενο p_i που αναπαριστάται διαγραμματικά, και στον οριζόντιο άξονα παρουσιάζονται τα αντικείμενα ταξινομημένα με τη σειρά που έγινε η επιλογή των στοιχείων $0,\ldots,|D|$-1. Η συχνότητα $(p_o,\ldots,p_{|D|-1})$ καλείται και συσταδική αναταξινόμηση των αντικειμένων στο D.

3.5 Ο αλγόριθμος T-Optics

Ο αλγόριθμος OPTICS παράγει την επαυξημένη ταξινόμηση της συστάδας λαμβάνοντας υπόψη την ταξινόμηση των σημείων, τις τιμές γειτνίασης και τις τιμές του πυρήνα, όπως αυτές έχουν περιγραφεί ανωτέρω. Επεκτάσεις του OPTICS στις τροχιές αποτελούν οι αλγόριθμοι T-OPTICS και TF-OPTICS (Nanni and Pedreschi 2006). Όσον αφορά τους χρόνους εκτέλεσης του αλγορίθμου T-OPTICS ισχύει ότι (α) το μεγαλύτερο μέρος του υπολογιστικού κόστους του αλγορίθμου προέρχεται συνήθως από τον υπολογισμό των αποστάσεων και (β) ο αριθμός των αποστάσεων που απαιτούνται για την κατασκευή του δείκτη του M-δέντρου αυξάνεται εμφανώς τετραγωνικά ακόμα και για μικρές σταθερές. Το M-δέντρο (Ciaccia, Patella & Zezula 1997) αποτελεί μία αποτελεσματική μέθοδο για την αναζήτηση ομοιότητας σε μετρικούς χώρους και ο δείκτης αντίστοιχα έχει ως στόχο την υποστήριξη αποδοτικών ερωτημάτων προς τη βάση δεδομένων, τα οποία και αποτελούν τον πυρήνα λειτουργίας του αλγορίθμου OPTICS. Ωστόσο, για να μικρύνει το υπολογιστικό κόστος μπορούμε να εργαστούμε σε υποδιαστήματα τροχιών, υλοποιώντας κάθε βήμα του αλγορίθμου TF-OPTICS, που συγκριτικά η διαδικασία αυτή είναι υπολογιστικά φθηνότερη από ότι μια εκτέλεση του T-OPTICS.



Εικόνα 12: Διάγραμμα γειτνίασης

4. Αξιολόγηση εγκυρότητας ομαδοποίησης

Η αξιολόγηση των αποτελεσμάτων της ομαδοποίησης μερικές φορές αναφέρεται και ως επικύρωση συμπλέγματος. Υπήρξαν πολλές προτάσεις για τα μέτρα ομοιότητας μεταξύ δύο ομαδοποιήσεων. Ένα τέτοιο μέτρο μπορεί να χρησιμοποιηθεί για να συγκρίνει πόσο καλά μπορούν να λειτουργήσουν διαφορετικοί αλγόριθμοι ομαδοποίησης σε ένα σύνολο δεδομένων. Τα μέτρα αυτά συνήθως συνδέονται με το είδος του κριτηρίου που εξετάζεται κατά την αξιολόγηση της ποιότητας της μεθόδου ομαδοποίησης.

4.1 Εισαγωγή

Η ομαδοποίηση είναι ένα από τα πιο χρήσιμα εργαλεία για την διαδικασία εξόρυξης δεδομένων με σκοπό την ανακάλυψη και τον εντοπισμό ομάδων και προτύπων. Έτσι, η κύρια ανησυχία στη διαδικασία της ομαδοποίησης είναι να αποκαλυφθούν οι ομάδες, οι οποίες θα μας επιτρέψουν να ανακαλύψουμε ομοιότητες και διαφορές, καθώς και να εξάγουμε χρήσιμα συμπεράσματα.

Στη βιβλιογραφία υπάρχει πλήθος αλγορίθμων ομαδοποίησης που έχουν προταθεί με σκοπό να εξυπηρετούν διαφορετικές εφαρμογές και διαφορετικά μεγέθη των συνόλων δεδομένων. Η εφαρμογή ενός αλγορίθμου συσταδοποίησης σε ένα σύνολο στοιχείων αποσκοπεί στην ανακάλυψη των φυσικών διαμερίσεών του. Ωστόσο, η διαδικασία ομαδοποίησης γίνεται αντιληπτή ως μια διαδικασία χωρίς επίβλεψη, δεδομένου ότι δεν υπάρχουν προκαθορισμένες κλάσεις και κανένα υπόδειγμα που θα δείξει ποιες είναι οι επιθυμητές σχέσεις που πρέπει να ισχύουν μεταξύ των δεδομένων.

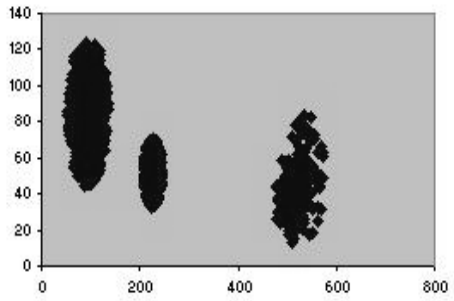
Οι διάφοροι αλγόριθμοι ομαδοποίησης βασίζονται σε ορισμένες παραδοχές προκειμένου να προσδιορίσουν την κατάτμηση ενός συνόλου δεδομένων. Ως συνέπεια αυτού, μπορούν να συμπεριφέρονται με διαφορετικό τρόπο εξαρτώμενοι από

i) τα χαρακτηριστικά του συνόλου δεδομένων (τη γεωμετρία και την κατανομή της πυκνότητας των συστάδων)
ii) την είσοδο των τιμών των παραμέτρων.

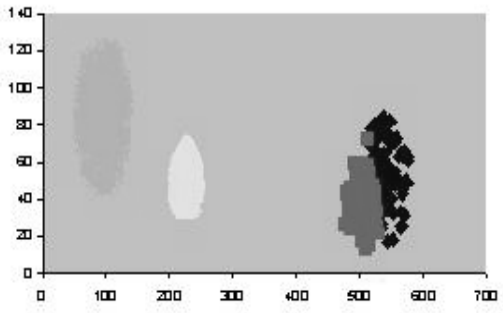
Είναι προφανές ότι ένα πρόβλημα που αντιμετωπίζουμε στην ομαδοποίηση είναι να αποφασιστεί ο βέλτιστος αριθμός των συστάδων που ταιριάζει σε ένα σύνολο δεδομένων.

Στις περισσότερες πειραματικές αξιολογήσεις των αλγορίθμων ομαδοποίησης χρησιμοποιούνται σύνολα δεδομένων δύο διαστάσεων, έτσι ώστε ο πειραματιστής να είναι σε θέση να ελέγξει οπτικά την εγκυρότητα των αποτελεσμάτων (δηλαδή, το πόσο καλά ο αλγόριθμος ομαδοποίησης ανακάλυψε τις συστάδες του συνόλου δεδομένων). Είναι σαφές ότι η απεικόνιση του συνόλου δεδομένων είναι μια κρίσιμη διαδικασία για την επαλήθευση των αποτελεσμάτων ομαδοποίησης. Σε περίπτωση που έχουμε σύνολα δεδομένων με περισσότερες διαστάσεις από τρεις είναι δύσκολη η απεικόνιση των αποτελεσμάτων της ομαδοποίησης. Επίσης, η αντίληψη των συστάδων με τη χρήση διαθέσιμων εργαλείων οπτικοποίησης είναι ένα δύσκολο έργο για τους ανθρώπους, που δεν είναι εξοικειωμένοι με τις υψηλότερες διαστάσεις των χώρων.

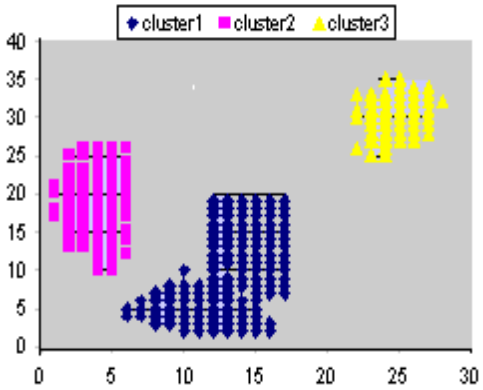
Παράδειγμα συσταδοποίησης



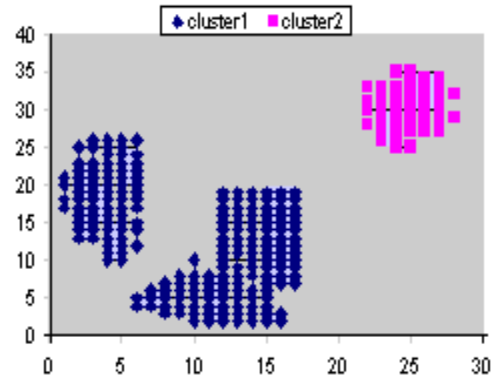
Εικόνα 13 α



Εικόνα 13 β



Εικόνα 14 α



Εικόνα 14 β

Στην Εικόνα 13α παρατηρούμε σαφώς ότι το σύνολο δεδομένων στο παράδειγμά μας ομαδοποιείται σε τρεις συστάδες. Ωστόσο, αν εφαρμόσουμε τον αλγόριθμο K-Means με καθορισμένες τιμές των παραμέτρων εισόδου (στην περίπτωση του K-Means

είναι ο αριθμός των συστάδων), ώστε να διαμερίσει τα δεδομένα μέσα σε τέσσερις συστάδες, το αποτέλεσμα της διαδικασίας της ομαδοποίησης παρουσιάζεται στην Εικόνα 13β. Στο παράδειγμά μας (Halkidi et al. 2002a), ο αλγόριθμος ομαδοποίησης K-Means βρήκε την καλύτερη διαμέριση σε τέσσερις συστάδες. Ωστόσο, αυτή δεν είναι η βέλτιστη διαμέριση για το υπό εξέταση σύνολο δεδομένων. Εδώ ορίζουμε τον όρο "βέλτιστη" συσταδοποίηση ως το αποτέλεσμα της λειτουργίας ενός αλγορίθμου ομαδοποίησης που ταιριάζει καλύτερα με τις φυσικές διαμερίσεις του συνόλου δεδομένων.

Παρομοίως, η Εικόνα 14 παρουσιάζει τη συμπεριφορά του αλγορίθμου DBSCAN εξετάζοντας κάθε φορά διαφορετικές τιμές για τις παραμέτρους εισόδου. Ο DBSCAN επιτυγχάνει να διαμερίσει βέλτιστα τα δεδομένα σε τρεις συστάδες (βλέπε Εικόνα 14α) μόνο κάτω από τη θεώρηση των κατάλληλων τιμών των παραμέτρων (Eps = 2, Nps = 4). Χρησιμοποιώντας διαφορετικές τιμές παραμέτρων εισόδου, αποτυγχάνει να βρει τη βέλτιστη διαμέριση του συνόλου δεδομένων (βλέπε Εικόνα 14β). Κατά συνέπεια, εάν στις παραμέτρους του αλγόριθμου ομαδοποίησης δοθούν ακατάλληλες τιμές, η μέθοδος ομαδοποίησης οδηγεί σε μια διαμέριση που δεν είναι η βέλτιστη για τα συγκεκριμένα δεδομένα, κάτι που θα οδηγούσε και σε λανθασμένες αποφάσεις. Το πρόβλημα που προκύπτει σχετικά με τον καθορισμό του βέλτιστου αριθμού των συστάδων ενός συνόλου δεδομένων, καθώς και η αξιολόγηση των αποτελεσμάτων ομαδοποίησης έχει αποτελέσει αντικείμενο πολλών ερευνητικών προσπαθειών. Στη συνέχεια, θα συζητήσουμε τις θεμελιώδεις έννοιες της αξιολόγησης της ομαδοποίησης.

4.2 Θεμελιώδεις έννοιες αξιολόγησης ομαδοποίησης

Η διαδικασία της αξιολόγησης των αποτελεσμάτων ενός αλγορίθμου ομαδοποίησης είναι γνωστή με τον όρο "ισχύς συστάδας". Σε γενικές γραμμές, υπάρχουν τρεις διαφορετικές προσεγγίσεις για τη διερεύνηση της ισχύος της συστάδας .

Η πρώτη βασίζεται σε εξωτερικά κριτήρια. Αυτό σημαίνει ότι θα αξιολογήσουμε τα αποτελέσματα ενός αλγορίθμου ομαδοποίησης με βάση μια προκαθορισμένη δομή ομαδοποίησης, η οποία επιβάλλεται σε ένα σύνολο δεδομένων και αντικατοπτρίζει αυτό το οποίο θέλουμε να γίνει μέσω της ομαδοποίησης του συγκεκριμένου συνόλου δεδομένων. Η δεύτερη προσέγγιση βασίζεται σε εσωτερικά κριτήρια. Στην περίπτωση αυτή, τα αποτελέσματα ομαδοποίησης αξιολογούνται με βάση τις ποσότητες που περιλαμβάνουν τα ίδια τα διανύσματα του συνόλου δεδομένων. Η τρίτη προσέγγιση της εγκυρότητας της ομαδοποίησης βασίζεται σε συγκριτικά κριτήρια. Εδώ, η βασική ιδέα είναι η αξιολόγηση μιας δομής ομαδοποίησης συγκρίνοντάς τη με άλλες δομές ομαδοποίησης που προκύπτουν από τον ίδιο αλγόριθμο, αλλά με διαφορετικές τιμές παραμέτρων εισόδου.

Οι δύο πρώτες προσεγγίσεις που βασίζονται σε στατιστικές δοκιμές έχουν ως σημαντικό μειονέκτημά τους το υψηλό υπολογιστικό κόστος. Επιπλέον, οι δείκτες που σχετίζονται με αυτές τις προσεγγίσεις στοχεύουν στη μέτρηση του βαθμού στον οποίο ένα σύνολο δεδομένων επιβεβαιώνει την εκ των προτέρων καθορισμένη δομή. Από την άλλη πλευρά, η τρίτη προσέγγιση στοχεύει στο να βρει την καλύτερη συσταδοποίηση που ένας αλγόριθμος ομαδοποίησης μπορεί να κάνει υπό ορισμένες παραδοχές και παραμέτρους.

Παρουσιάζουμε τα θεμελιώδη κριτήρια και τους αντιπροσωπευτικούς δείκτες για τις δύο πρώτες προσεγγίσεις στις οποίες προαναφερθήκαμε (η πρώτη βασίζεται σε εξωτερικά κριτήρια και η δεύτερη σε εσωτερικά κριτήρια). Συγκεκριμένα, παρουσιάζονται κατάλληλες μέθοδοι για την ποσοτική αξιολόγηση των αποτελεσμάτων ομαδοποίησης, γνωστές και ως μέθοδοι εγκυρότητας συστάδας. Ωστόσο, αυτές οι μέθοδοι δίνουν μόνο μια ένδειξη της ποιότητας του προκύπτοντος διαχωρισμού, και έτσι μπορούν να θεωρηθούν μόνο ως ένα εργαλείο στη διάθεση των εμπειρογνωμόνων προκειμένου να αξιολογηθούν τα αποτελέσματα της ομαδοποίησης. Οι διάφορες προσεγγίσεις εγκυρότητας της συσταδοποίησης, με εξωτερικά και εσωτερικά κριτήρια, βασίζονται κυρίως στο στατιστικό έλεγχο υποθέσεων. Παρακάτω παρουσιάζεται μια εισαγωγή στις βασικές έννοιες του ελέγχου υποθέσεων για την εγκυρότητα της συστάδας.

4.2.1 Έλεγχος υποθέσεων στην αξιολόγηση συστάδων

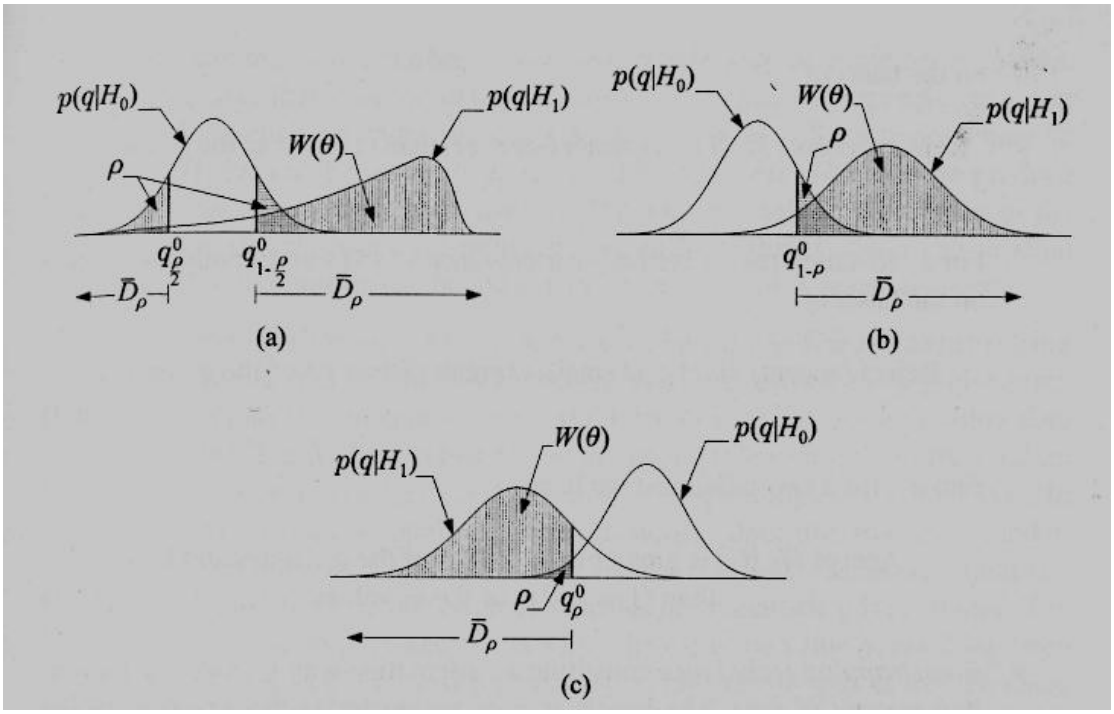
Στη διαδικασία εξέτασης της εγκυρότητας των συστάδων, η ανάλυση είναι βασισμένη στον έλεγχο της τυχαίας δομής ενός συνόλου δεδομένων X, με τη μηδενική υπόθεση να αντιστοιχεί στην H_0: τα σημεία του συνόλου δεδομένων είναι τυχαία δομημένα έναντι της εναλλακτικής H_1: τα σημεία του συνόλου δεδομένων δεν είναι τυχαία δομημένα. Για να εξετάσουμε αυτή την υπόθεση χρησιμοποιούμε στατιστικά τεστ, τα οποία όμως οδηγούν σε μεγάλη υπολογιστική πολυπλοκότητα. Έτσι για να αντιμετωπίσουμε το πρόβλημα της πολυπλοκότητας των υπολογισμών χρησιμοποιούμε κυρίως τεχνικές Monte Carlo.

4.2.2 Τεχνικές Monte Carlo στην αξιολόγηση συστάδων

Ο στόχος των τεχνικών Monte Carlo είναι ο υπολογισμός της συνάρτησης πυκνότητας πιθανότητας των δεικτών ισχύος. Στηρίζονται στην προσομοίωση της διαδικασίας εκτίμησης της συνάρτησης πυκνότητας πιθανότητας του δείκτη ισχύος, χρησιμοποιώντας επαρκή αριθμό "κατασκευασμένων" δεδομένων μέσω υπολογιστή.

Αρχικά, ένα μεγάλος αριθμός των κατασκευασμένων συνόλων δεδομένων παράγεται από μια κανονική κατανομή. Για κάθε ένα από αυτά τα κατασκευασμένα σύνολα δεδομένων, έστω X_i, υπολογίζεται η τιμή του δείκτη η οποία συμβολίζεται με q_i. Στη

συνέχεια, με βάση τις αντίστοιχες τιμές των q_i, για κάθε ένα από τα σύνολα δεδομένων X_i, δημιουργείται ένα διάγραμμα διασποράς. Αυτό το διάγραμμα διασποράς (Halkidi et al. 2002a) είναι μια προσέγγιση της συνάρτησης πυκνότητας πιθανότητας του δείκτη.



Εικόνα 15: Διάστημα εμπιστοσύνης για (α) δίπλευρο δείκτη (β) μονόπλευρο δείκτη (δεξιά ουρά), (γ) μονόπλευρο δείκτη (αριστερή ουρά), όπου q^0_ρ είναι το ποσοστό του *q* στο πλαίσιο της υπόθεσης H_0

Στην Εικόνα 15, παρουσιάζονται οι τρεις πιθανές περιπτώσεις συναρτήσεων πυκνότητας πιθανότητας ενός δείκτη *q*. Υπάρχουν τρεις διαφορετικές δυνατές μορφές, ανάλογα με το διάστημα εμπιστοσύνης $\overline{D_\rho}$ που αντιστοιχεί σε επίπεδο σημαντικότητας ρ.

Διαπιστώνουμε ότι η συνάρτηση πυκνότητας πιθανότητας του στατιστικού δείκτη *q*, υπό την H_0, έχει ένα μόνο μέγιστο, και η περιοχή $\overline{D_\rho}$ είναι είτε μια μισή γραμμή ή μία ένωση δύο μισών γραμμών. Υποθέτοντας ότι το διάγραμμα διασποράς έχει δημιουργηθεί χρησιμοποιώντας τιμές *r* του δείκτη *q*, που ονομάζεται q_i, προκειμένου να αποδεχθούμε ή να απορρίψουμε τη μηδενική υπόθεση H_0 εξετάζουμε τις ακόλουθες προϋποθέσεις:

- Εάν η μορφή της συνάρτησης πυκνότητας πιθανότητας έχει δεξιά ουρά και η τιμή του *q* για το σετ δεδομένων μας είναι μεγαλύτερη από *(1-ρ)*r* των τιμών q_i, τότε απορρίπτουμε την H_0, αλλιώς δεχόμαστε την H_0.
- Εάν η μορφή της συνάρτησης πυκνότητας πιθανότητας έχει αριστερή ουρά και η τιμή του *q* για το σετ δεδομένων μας είναι μικρότερη από *ρ*r* των τιμών q_i, τότε απορρίπτουμε την H_0, αλλιώς δεχόμαστε την H_0.

- Εάν η μορφή της συνάρτησης πυκνότητας πιθανότητας έχει αριστερή και δεξιά ουρά, και η τιμή του q για το σετ δεδομένων μας είναι μεγαλύτερη από $(\rho/2)*r$ των τιμών q_i, και μικρότερη από $(1-\rho/2)*r$ των τιμών q_i, τότε δεχόμαστε την H_0.

4.3 Εσωτερική αξιολόγηση εγκυρότητας συστάδας

Χρησιμοποιώντας την προσέγγιση εσωτερικής αξιολόγησης της εγκυρότητας συστάδας (Halkidi et al. 2002a & 2002b), ο στόχος είναι να αξιολογηθεί το αποτέλεσμα ομαδοποίησης ενός αλγορίθμου χρησιμοποιώντας μόνο ποσοτικά χαρακτηριστικά που κληρονόμησε από το σύνολο δεδομένων. Υπάρχουν δύο περιπτώσεις στις οποίες εφαρμόζονται τα εσωτερικά κριτήρια εγκυρότητας συστάδας, ανάλογα με τη δομή ομαδοποίησης, 1. ιεραρχία των "σχημάτων" συσταδοποίησης (hierarchy of clustering schemes) και 2. μοναδικό σχήμα συσταδοποίησης (single clustering scheme).

4.3.1 Αξιολόγηση της ιεραρχίας των σχημάτων συσταδοποίησης

Ένας πίνακας (με την ονομασία "cophenetic matrix") ο οποίος συμβολίζεται με P_c, μπορεί να αντιπροσωπεύει ένα ιεραρχικό διάγραμμα που παράγεται από έναν ιεραρχικό αλγόριθμο. Το στοιχείο $P_{(i,\ j)}$ του εν λόγω πίνακα αντιπροσωπεύει το επίπεδο γειτνίασης στο οποίο δύο διανύσματα x_i και x_j, βρίσκονται στην ίδια συστάδα για πρώτη φορά. Μπορούμε να ορίσουμε ένα στατιστικό δείκτη για τη μέτρηση του βαθμού της ομοιότητας μεταξύ του πίνακα P_c και του P πίνακα γειτνίασης.

Ο δείκτης αυτός ονομάζεται "Cophenetic Correlation Coefficient" και ορίζεται ως εξής:

$$CPCC = \frac{(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}\, c_{ij} - \mu_P \mu_C}{\sqrt{\left[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 - \mu_P^2\right]\left[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2 - \mu_C^2\right]}},$$

όπου $M = N*(N-1)/2$ και N είναι ο αριθμός των σημείων σε ένα σύνολο δεδομένων. Επίσης, μ_p και μ_C είναι οι μέσοι των πινάκων P και P_c, αντίστοιχα, και ορίζονται από τις παρακάτω εξισώσεις:

$$\mu_P = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P(i,j)\ ,\mu_C = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P_C(i,j).$$

Επιπλέον, d_{ij} και c_{ij} είναι τα *(i, j)* στοιχεία των πινάκων *P* και P_c, αντίστοιχα. Ο CPCC δείκτης λαμβάνει τιμές μεταξύ των -1 και 1. Η τιμή του δείκτη κοντά στο 1 είναι ένδειξη σημαντικής ομοιότητας μεταξύ των δύο πινάκων. Η διαδικασία των τεχνικών Monte Carlo (οι οποίες περιγράφηκαν παραπάνω) χρησιμοποιείται επίσης και σε αυτή την περίπτωση της αξιολόγησης.

4.3.2 Αξιολόγηση μοναδικού σχήματος συσταδοποίησης

Ο στόχος εδώ, είναι να βρούμε τον βαθμό του ταιριάσματος μεταξύ ενός δεδομένου συστήματος ομαδοποίησης, έστω *C*, που αποτελείται από n_c συστάδες και του πίνακα γειτνίασης *P*. Ο κατάλληλος δείκτης για αυτή την προσέγγιση είναι ο δείκτης *Γ* του Hubert ή ο κανονικοποιημένος δείκτης *Γ*. Ένας επιπρόσθετος πίνακας που χρησιμοποιείται στον υπολογισμό του δείκτη είναι ο

$$Y(i,j) = \begin{cases} 1, & \text{αν } x_i \text{ και } x_j \text{ ανήκουν σε διαφορετικές συστάδες} \\ 0, & \text{αλλιώς} \end{cases}$$

όπου *i,j =1,…,N.*

Η εφαρμογή των τεχνικών Monte Carlo αποτελεί και σε αυτή την περίπτωση τον τρόπο για να ελεγχθεί η υπόθεση της τυχαιότητας σε ένα συγκεκριμένο σύνολο δεδομένων.

4.3.3 Κριτήρια εσωτερικής αξιολόγησης

Τα κριτήρια εσωτερικής αξιολόγησης συνήθως αναθέτουν την καλύτερη βαθμολογία στον αλγόριθμο που παράγει συστάδες με μεγάλη ομοιότητα μέσα σε μία συστάδα και χαμηλή ομοιότητα μεταξύ των συστάδων. Ένα μειονέκτημα της χρήσης εσωτερικών κριτηρίων για την αξιολόγηση της συσταδοποίησης είναι ότι οι υψηλές βαθμολογίες σε ένα εσωτερικό μέτρο δεν οδηγούν απαραίτητα σε αποτελεσματικές εφαρμογές ανάκτησης πληροφοριών . Επιπλέον, η αξιολόγηση αυτή μεροληπτεί υπέρ των αλγορίθμων που χρησιμοποιούν το ίδιο μοντέλο συσταδοποίησης. Για παράδειγμα, ο k-Means αλγόριθμος ομαδοποίησης βελτιστοποιεί φυσικά αποστάσεις αντικειμένων και ένα εσωτερικό κριτήριο με βάση την απόσταση θα υπερεκτιμά πιθανώς την προκύπτουσα ομαδοποίηση.

Τα εσωτερικά μέτρα αξιολόγησης είναι τα πλέον κατάλληλα μέτρα για να παρέχουν μία εικόνα για το πότε ένας αλγόριθμος αποδίδει καλύτερα από έναν άλλο, αλλά αυτό δεν σημαίνει ότι ένας αλγόριθμος παράγει και πιο έγκυρα αποτελέσματα από έναν άλλο. Η ισχύς, όπως αυτή μετράται από έναν δείκτη, εξαρτάται από τον ισχυρισμό ότι αυτού του είδους η δομή υπάρχει στο σύνολο δεδομένων. Ένας αλγόριθμος που έχει σχεδιαστεί για συγκεκριμένα μοντέλα δεν έχει καμία πιθανότητα εάν το σύνολο δεδομένων περιέχει ένα εντελώς διαφορετικό σύνολο μοντέλων ή τα μέτρα αξιολόγησης περιέχουν ένα διαφορετικό κριτήριο. Για παράδειγμα, η k-means συσταδοποίηση μπορεί να βρει μόνο κυρτά “συμπλέγματα” και πολλοί δείκτες

αξιολόγησης υποθέτουν κυρτές συστάδες. Σε ένα σύνολο δεδομένων με μη-κυρτές συστάδες, ούτε η χρήση του k-means αλλά ούτε και η χρήση ενός κριτήριου αξιολόγησης που υποθέτει κυρτότητα είναι σωστή.

Παρακάτω παρουσιάζονται εσωτερικά κριτήρια τα οποία μπορούν να χρησιμοποιηθούν για την αξιολόγηση της ποιότητας των αλγορίθμων ομαδοποίησης.

Davies–Bouldin index

Ο δείκτης Davies-Bouldin υπολογίζεται με τον ακόλουθο τύπο

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right),$$

όπου n είναι ο αριθμός των συστάδων, c_x είναι το κέντρο βάρους της συστάδας x, σ_x είναι η μέση απόσταση όλων των στοιχείων της συστάδας x από το κέντρο βάρους c_x, και $d(c_i, c_j)$ είναι η απόσταση μεταξύ των κέντρων c_i και c_j. Οι αλγόριθμοι που παράγουν συστάδες με χαμηλή απόσταση μεταξύ των στοιχείων των συστάδων, και μεγάλη μεταξύ των συστάδων, θα έχουν χαμηλό δείκτη Davies–Bouldin. Ο αλγόριθμος ομαδοποίησης με την καλύτερη απόδοση σύμφωνα με αυτό το κριτήριο είναι αυτός για τον οποίο η τιμή του δείκτη Davies–Bouldin είναι η μικρότερη.

Dunn index

Ο δείκτης Dunn έχει ως στόχο τον εντοπισμό πυκνών και καλά διαχωρισμένων συστάδων. Ορίζεται ως ο λόγος μεταξύ της ελάχιστης απόστασης μεταξύ των συστάδων και της μέγιστης απόστασης ανάμεσα στα στοιχεία των συστάδων.

Για κάθε διαμερισμό συστάδας, ο δείκτης Dunn μπορεί να υπολογιστεί με τον ακόλουθο τύπο:

$$D = \frac{\min_{1<i<j<n} d(i,j)}{\max_{1\leq k\leq n} d'(k)},$$

όπου το $d(i, j)$ αντιπροσωπεύει την απόσταση μεταξύ των συστάδων i και j, και $d'(k)$ η απόσταση ανάμεσα στα στοιχεία της συστάδας k. Η απόσταση $d(i, j)$ μεταξύ δύο συστάδων μπορεί να υπολογιστεί με βάση οποιοδήποτε μέτρο απόστασης, όπως η απόσταση μεταξύ των κεντροειδων των συστάδων. Ομοίως, η απόσταση $d'(k)$ μπορεί να μετρηθεί με ποικίλους τρόπους, όπως η μέγιστη απόσταση μεταξύ οποιουδήποτε

ζεύγους των στοιχείων από τη συστάδα *k*. Γενικά, ένα εσωτερικό κριτήριο επιδιώκει συστάδες με μεγάλη ομοιότητα μέσα στη συστάδα και χαμηλή ομοιότητα μεταξύ των συστάδων, και οι αλγόριθμοι οι οποίοι παράγουν συστάδες με υψηλό δείκτη Dunn είναι οι καλύτεροι.

Silhouette coefficient

Ο συντελεστής Silhouette αντιπαραβάλλει τη μέση απόσταση των στοιχείων από τη μία ομάδα με τη μέση απόσταση των στοιχείων από άλλες ομάδες. Τα στοιχεία με υψηλή τιμή για το συντελεστή Silhouette θεωρούνται ότι έχουν διαχωριστεί σωστά ενώ τα στοιχεία με χαμηλή τιμή μπορεί να είναι ακραίες τιμές. Ο δείκτης αυτός λειτουργεί καλά με τον αλγόριθμο συσταδοποίησης k-means, και χρησιμοποιείται επίσης για να προσδιοριστεί ο βέλτιστος αριθμός των συστάδων.

4.4 Εξωτερική αξιολόγηση εγκυρότητας συστάδας

Χρησιμοποιώντας τα κριτήρια εξωτερικής αξιολόγησης (Halkidi et al. 2002a & 2002b), τα αποτελέσματα ομαδοποίησης αξιολογούνται με βάση τα δεδομένα που δεν χρησιμοποιούνται για την ομαδοποίηση, όπου είναι γνωστές οι ετικέτες κλάσης και τα εξωτερικά σημεία αναφοράς. Τα εν λόγω κριτήρια αποτελούνται από ένα σύνολο "προ-ταξινομημένων" στοιχείων, και αυτές οι ομάδες στοιχειων συχνά δημιουργούνται από πειραματιστές. Έτσι, οι ομάδες αναφοράς μπορεί να θεωρηθούν ως ένα πρότυπο για την αξιολόγηση. Αυτές οι μέθοδοι αξιολόγησης μετρούν το πόσο κοντά είναι η ομαδοποίηση στις προκαθορισμένες ομάδες αναφοράς. Ωστόσο, πρόσφατα έχει συζητηθεί το κατά πόσο είναι επαρκείς οι μέθοδοι αυτές για πραγματικά δεδομένα, ή για τα "κατασκευασμένα» σύνολα δεδομένων. Παρακάτω παρουσιάζονται εξωτερικά κριτήρια τα οποία μπορούν να χρησιμοποιηθούν για την αξιολόγηση της ποιότητας των αλγορίθμων ομαδοποίησης.

Όσον αφορά τα εξωτερικά κριτήρια αξιολόγησης μπορούμε να εργαστούμε με δύο διαφορετικούς τρόπους. Πρώτον, μπορούμε να αξιολογήσουμε την προκύπτουσα δομή ομαδοποίησης *C*, συγκρίνοντάς την με μία ανεξάρτητη διαμέριση των στοιχείων *P*, η οποία γίνεται σύμφωνα με τη διαίσθηση του πειραματιστή για το συγκεκριμένο σύνολο δεδομένων. Δεύτερον, μπορούμε να συγκρίνουμε τον πίνακα γειτνίασης με τη διαμέριση των στοιχείων *P*.

4.4.1 Σύγκριση της δομής ομαδοποίησης *C* με τη διαμέριση των στοιχείων *P*

Υποθέτουμε ότι $C = \{C_1,...,C_m\}$ είναι μία ομαδοποίηση ενός συνόλου στοιχείων *X*, και $P = \{P_1,...,P_s\}$ είναι η διαμέριση η οποία έχουμε εφαρμόσει. Αναφερόμαστε σε ένα ζεύγος σημείων (x_v, x_u) από το σύνολο των στοιχείων χρησιμοποιώντας τους ακόλουθους όρους:

- SS: αν και οι δύο μονάδες ανήκουν στην ίδια συστάδα της ομαδοποίησης *C* και της διαμερισης *P*.
- SD: εάν οι δύο μονάδες ανήκουν στην ίδια συστάδα της ομαδοποίησης *C* και σε διαφορετικές ομάδες της διαμέρισης *P*.
- DS: αν τα σημεία ανήκουν σε διαφορετικές ομάδες της ομαδοποίησης *C* και στην ίδια ομάδα της διαμέρισης *P*.
- DD: αν και οι δύο μονάδες ανήκουν σε διαφορετικές ομάδες της ομαδοποίησης *C* και σε διαφορετικές ομάδες της διαμέρισης *P*.

Υποθέτοντας τώρα ότι *a, b, c* και *d* είναι ο αριθμός των SS, SD, DS και DD ζευγών, αντίστοιχα, τότε $a + b + c + d = M$, όπου *M* είναι ο μέγιστος αριθμός όλων των ζευγών στο σύνολο δεδομένων (δηλαδή, $M = N(N-1)/2$, όπου *N* είναι ο συνολικός αριθμός των σημείων του συνόλου δεδομένων).

Τώρα μπορούμε να ορίσουμε τους ακόλουθους δείκτες για τη μέτρηση του βαθμού ομοιότητας μεταξύ της ομαδοποίησης *C* και της διαμέρισης *P*.

Ο δείκτης Rand

Ο δείκτης Rand υπολογίζει το πόσο όμοιες είναι οι συστάδες με τις συστάδες αναφοράς. Κάποιος μπορεί επίσης να θεωρήσει το δείκτη Rand ως μέτρο του ποσοστού των σωστών αποφάσεων που λαμβάνονται από τον αλγόριθμο.

Ο δείκτης Rand ορίζεται ως:

$$R = (a + d) / M.$$

Ο δείκτης Jaccard

Ο δείκτης Jaccard χρησιμοποιείται για τον ποσοτικό προσδιορισμό της ομοιότητας μεταξύ δύο συνόλων δεδομένων. Ο δείκτης Jaccard παίρνει τιμή μεταξύ του 0 και 1. Ο δείκτης με τιμή 1 σημαίνει ότι τα δύο σετ δεδομένων είναι ταυτόσημα, και ο δείκτης με τιμή 0 δείχνει ότι τα σύνολα δεδομένων δεν έχουν καθόλου κοινά στοιχεία.

Ο δείκτης Jaccard ορίζεται ως:

$$J = a / (a + b + c).$$

Ουσιαστικά αποτελεί τον αριθμό των μοναδικών στοιχείων που είναι κοινά και στα δύο σύνολα, διαιρούμενο με το συνολικό αριθμό των μοναδικών στοιχείων στα δύο σύνολα.

Σημείωση: Οι δύο παραπάνω δείκτες (Rand & Jaccard) κυμαίνονται μεταξύ 0 και 1, και μεγιστοποιούνται όταν το $m = s$.

Ο δείκτης Fowlkes–Mallows

Ο δείκτης Fowlkes-Mallows υπολογίζει την ομοιότητα μεταξύ των ομάδων που επιστρέφονται από τον αλγόριθμο ομαδοποίησης, και των συστάδων αναφοράς. Όσο υψηλότερη είναι η τιμή του δείκτη Fowlkes-Mallows, τόσο πιο όμοιες είναι οι συστάδες και οι συστάδες αναφοράς.

Ο δείκτης FM εναλλακτικά ορίζεται ως ο γεωμετρικός μέσος της ακρίβειας (precision) και της ανάκλησης (recall), P και R, αντίστοιχα, ενώ το F-μέτρο (F-measure) είναι ο αρμονικός μέσος τους. Τα μέτρα της ακρίβειας και της ανάκλησης είναι επίσης γνωστά ως δείκτες Wallace's indices B^I and B^{II}.

Ο δείκτης Fowlkes–Mallows ορίζεται ως

$$FM = \frac{a}{\sqrt{m_1 \times m_2}} = \sqrt{\frac{a}{a+b} \times \frac{a}{a+c}}, \text{ όπου } m_1 = \frac{a}{a+b}, m_2 = \frac{a}{a+c}.$$

Σημείωση: Για τους τρεις προαναφερθέντες δείκτες έχει αποδειχθεί ότι οι υψηλές τιμές των δεικτών δείχνουν μεγάλη ομοιότητα μεταξύ της ομαδοποίησης C και της διαμέρισης P. Με άλλα λόγια οι υψηλές τιμές των δεικτών υποδεικνύουν μεγάλη ομοιότητα μεταξύ των ομάδων που επιστρέφονται από τον αλγόριθμο ομαδοποίησης και τις συστάδες αναφοράς. Όσο υψηλότερες είναι οι τιμές αυτών των δεικτών, τόσο πιο πανομοιότυπες είναι μεταξύ τους οι ομάδες που επιστρέφονται από τον αλγόριθμο ομαδοποίησης και τις συστάδες αναφοράς.

Άλλοι αντίστοιχοι εξωτερικοί δείκτες παρουσιάζονται ως ακολούθως.

Hubert’s Γ Statistic

Ο δείκτης Γ του Hubert ορίζεται ως

$$\Gamma = (1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i,j) \times Y(i,j).$$

Υψηλές τιμές του δείκτη υποδηλώνουν μια ισχυρή ομοιότητα μεταξύ των *X* και *Y*.

Normalized Γ Statistic

Ο κανονικοποιημένος δείκτης *Γ* ορίζεται ως

$$\hat{\Gamma} = \frac{\left[(1/M) \times \sum_{i=1}^{N} \sum_{j=i+1}^{N-1} (X(i,j) - \mu_X) \times \left(Y\left((i,j) - \mu_Y\right)\right)\right]}{\sigma_x \times \sigma_y},$$

όπου το *X(i, j)* και *Y(i, j)* είναι το *(i, j)* στοιχείο των πινάκων *X* και *Y*, αντίστοιχα, τους οποίους και συγκρίνουμε. Επίσης, μ_x, μ_y και σ_x, σ_y είναι οι αντίστοιχες μέσες τιμές και διακυμάνσεις των X, Y πινάκων. Αυτός ο δείκτης παίρνει τιμές μεταξύ -1 και 1.

Οι προαναφερθέντες δείκτες έχουν συναρτήσεις πυκνότητας πιθανότητας με δεξιά ουρά, υπό την τυχαία υπόθεση. Για να χρησιμοποιήσουμε αυτούς τους δείκτες σε στατιστικές δοκιμές πρέπει να γνωρίζουμε τις αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας κάτω από τη μηδενική υπόθεση H_0, που είναι η υπόθεση της τυχαίας δομής των δεδομένων μας. Ωστόσο, ο υπολογισμός της συνάρτησης πυκνότητας πιθανότητας των δεικτών αυτών είναι υπολογιστικά δαπανηρός. Μία λύση στο πρόβλημα αυτό είναι να χρησιμοποιήσουμε τεχνικές Monte Carlo. Η διαδικασία έχει ως εξής:

Για $i = 1$ έως r

- Δημιουργούμε ένα σύνολο δεδομένων X_i με N διανύσματα (σημεία) στην περιοχή του X (δηλαδή, έχει την ίδια διάσταση με εκείνα του σετ δεδομένων X).
- Αντιστοιχούμε κάθε διάνυσμα $Y_{j,I}$ του X_i με την ομάδα που ανήκει, $X_j \in X$, σύμφωνα με τη διαμέριση P.
- Εκτελούμε τον ίδιο αλγόριθμο που χρησιμοποιείται για την παραγωγή της ομαδοποίησης C για κάθε X_i, και αφήνουμε την προκύπτουσα ομαδοποίηση C_i.
- Υπολογίζουμε το $q(C_i)$ του δείκτη q, για το C_i και το P.
- Κατασκευάζουμε διάγραμμα διασποράς των τιμών r του δείκτη εγκυρότητας $q(Ci)$ (που υπολογίζονται μέσα στον βρόχο).
- Σχεδιάζουμε την προσέγγιση της συνάρτησης πυκνότητας πιθανότητας των καθορισμένων στατιστικών δεικτών, οι οποίοι είναι έστω q και συγκρίνουμε με τα $q(C_i)$, τα οποία τα ονομάζουμε έστω q_i. Οι δείκτες *R, J, FM* και *Γ* χρησιμοποιούνται ως δείκτες q, όπως αναφέρεται στην παραπάνω διαδικασία.

4.4.2 Σύγκριση του πίνακα γειτνίασης *C* με τη διαμέριση των στοιχείων *P*

Έστω P είναι ο πίνακας γειτνίασης ενός συνόλου στοιχείων X, και P είναι η διαμέρισή του. Η διαμέριση μπορεί να θεωρηθεί ως μια χαρτογράφηση

$$g: X \to \{1 \dots n_c\},$$

όπου n_c είναι ο αριθμός των συστάδων.

Υποθέτουμε ότι ο πίνακας Y ορίζεται ως

$$Y(i,j) = \begin{cases} 1\,, & \alpha\nu\ g\left(X_i\right) \neq g\left(X_j\right)\ \ i,j = 1 \dots N \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}.$$

Ο στατιστικός δείκτης Γ (η ο κανονικοποιημένος δείκτης Γ) μπορεί να υπολογιστεί χρησιμοποιώντας τον πίνακα γειτνίασης P και τον πίνακα Y. Με βάση την τιμή του δείκτη, μπορεί να έχουμε μία ένδειξη για την ομοιότητα των δύο πινάκων.

Για να συνεχίσουμε με τη διαδικασία αξιολόγησης, χρησιμοποιούμε τις τεχνικές Monte Carlo, όπως αυτές περιγράφηκαν ανωτέρω. Στο βήμα της "κατασκευής" της διαδικασίας, η αντίστοιχη χαρτογράφηση g_i δημιουργείται για κάθε σύνολο δεδομένων X_i που έχει κατασκευαστεί. Έτσι, στο βήμα του υπολογισμού του πίνακα Yi, υπολογίζεται για κάθε $X_{i,}$ προκειμένου να βρεί τον αντίστοιχο στατιστικό δείκτη Γ_i.

4.5 Σχετικά κριτήρια (relative criteria)

Η βάση πάνω στην οποία στηρίζονται οι προαναφερθείσες μέθοδοι επικύρωσης (εσωτερική και εξωτερική αξιολόγηση) είναι η στατιστική δοκιμή. Το βασικότερο μειονέκτημα των παραπάνω μεθόδων είναι η υπολογιστική πολυπλοκότητα. Μία εναλλακτική προσέγγιση έτσι ώστε να αποφευχθεί αυτό το μειονέκτημα, είναι η χρήση σχετικών κριτηρίων που δεν περιλαμβάνουν στατιστικές δοκιμές (Halkidi et al. 2002b). Η βασική ιδέα αυτής της προσέγγισης είναι να επιλέξουμε την καλύτερη ομαδοποίηση σύμφωνα με ένα προκαθορισμένο κριτήριο. Πιο συγκεκριμένα, το πρόβλημα μπορεί να διατυπωθεί ως ακολούθως όπως αυτό περιγράφεται στην εργασία Halkidi et al. (2002b).

«Έστω P_{alg} είναι το σύνολο των παραμέτρων που σχετίζονται με ένα συγκεκριμένοο αλγόριθμο ομαδοποίησης (π.χ. ο αριθμός των συστάδων n_c). Μεταξύ των σχημάτων ομαδοποίησης $C_{i,}$ με $i = 1,\dots,n_c$, που ορίζονται από ένα συγκεκριμένο αλγόριθμο για διαφορετικές τιμές των παραμέτρων P_{alg}, επιλέξετε αυτό που ταιριάζει καλύτερα στο σύνολο των δεδομένων.»

Στη συνέχεια έχουμε τις ακόλουθες περιπτώσεις του ανωτέρου προβλήματος.

- Το P_{alg} δεν περιλαμβάνει τον αριθμό των συστάδων n_c ως παράμετρο

Σε αυτή την περίπτωση, η επιλογή των βέλτιστων τιμών των παραμέτρων περιγράφονται ως ακολούθως. Τρέχουμε τον αλγόριθμο για ένα μεγάλο εύρος τιμών των παραμέτρων του, και επιλέγουμε το μεγαλύτερο εύρος για το οποίο ο αριθμός n_c παραμένει σταθερός (συνήθως $n_c << N$ (αριθμός των πλειάδων)). Στη συνέχεια, επιλέγουμε ως κατάλληλες τιμές των παραμέτρων του P_{alg} τις τιμές που αντιστοιχούν στο μέσο του εύρους αυτής της περιοχής. Επιπρόσθετα, η διαδικασία αυτή προσδιορίζει τον αριθμό των συστάδων που κρύβονται κάτω από το σύνολο δεδομένων μας.

- Το P_{alg} περιλαμβάνει τον αριθμό των συστάδων n_c ως παράμετρο

Η διαδικασία ταυτοποίησης της καλύτερης ομαδοποίησης βασίζεται σε έναν δείκτη εγκυρότητας. Για την επιλογή του καταλληλότερου δείκτη αποδόσεων q ακολουθούμε τα επόμενα βήματα:

1. Ο αλγόριθμος ομαδοποίησης λειτουργεί για όλες τις τιμές των n_c μεταξύ ενός ελάχιστου n_{cmin} και μέγιστου n_{cmax}. Οι ελάχιστες και μέγιστες τιμές έχουν ορισθεί εκ των προτέρων από τον χρήστη.
2. Για κάθε μία από τις τιμές του n_c, ο αλγόριθμος εκτελείται r φορές, χρησιμοποιώντας διαφορετικό σύνολο τιμών για τις άλλες παραμέτρους του αλγορίθμου (π.χ. διαφορετικές αρχικές συνθήκες).
3. Οι βέλτιστες τιμές του δείκτη q που λαμβάνονται από κάθε n_c απεικονίζονται σε μορφή διαγράμματος σαν συνάρτηση του n_c.

Με βάση αυτό το διάγραμμα μπορεί να προσδιοριστεί η καλύτερη συσταδοποίηση. Πρέπει να τονίσουμε ότι υπάρχουν δύο προσεγγίσεις για τον καθορισμό της καλύτερης ομαδοποίησης ανάλογα με τη συμπεριφορά του q σε σχέση με το n_c. Συγκεκριμένα, αν ο δείκτης εγκυρότητας δεν παρουσιάζει αυξητική ή πτωτική τάση καθώς το n_c αυξάνει, επιδιώκουμε το μέγιστο (ελάχιστο) του διαγράμματος. Από την άλλη πλευρά, για τους δείκτες που αυξάνονται (ή μειώνονται) καθώς ο αριθμός των συστάδων αυξάνει, ψάχνουμε για τις τιμές εκείνες του n_c στις οποίες παρουσιάζεται σημαντική τοπική μεταβολή της τιμής του δείκτη. Αυτή η αλλαγή εμφανίζεται στο διάγραμμα με τη μορφή “γονάτου” (knee) και είναι μία ένδειξη του αριθμού των συστάδων για το σύνολο δεδομένων μας. Επιπλέον, η απουσία του “knee” μπορεί να αποτελεί μία ένδειξη ότι το σύνολο των δεδομένων δεν διαθέτει δομή ομαδοποίησης.

5. Στατιστικά τεστ για την σύγκριση της απόδοσης διαφορετικών αλγορίθμων

5.1 Εισαγωγή

Ενώ οι μέθοδοι για την σύγκριση δυο αλγορίθμων εκμάθησης σε ένα μοναδικό σύνολο δεδομένων έχουν μελετηθεί εξονυχιστικά εδώ και πολύ καιρό , το ζήτημα της χρήσης των στατιστικών τεστ για την σύγκριση περισσοτέρων αλγορίθμων σε πολλαπλά σύνολα δεδομένων , το οποίο είναι ιδιαίτερα σημαντικό σε τυπικές μελέτες μηχανικής μάθησης, δεν έχει μελετηθεί όσο θα έπρεπε. Στο κεφάλαιο αυτό μελετώνται τα στατιστικά τεστ τα οποία θα μπορούσαν να χρησιμοποιηθούν ή χρησιμοποιούνται ήδη για την σύγκριση δυο ή περισσοτέρων αλγορίθμων σε πολλαπλά σύνολα δεδομένων. Από εδώ και στο εξής συμβολίζουμε με *k* τους αλγορίθμους εκμάθησης οι οποίοι εξετάζονται πάνω σε N σύνολα δεδομένων.

5.2 Στατιστικά τεστ για την σύγκριση αλγορίθμων

Πολλοί ερευνητές υιοθετούν διαφορετικές στατιστικές τεχνικές ή τεχνικές βασισμένες στην κοινή λογική με σκοπό να αποφασίσουν εάν οι διαφορές μεταξύ των αλγορίθμων είναι πραγματικές ή τυχαίες. Σε αυτήν την ενότητα, εξετάζουμε αρκετά γνωστά και λιγότερο γνωστά στατιστικά τεστ, και μελετούμε την καταλληλότητα τους σχετικά με το τι μετρούν στην πραγματικότητα, καθώς επίσης και τις υποθέσεις που υιοθετούν για τα δεδομένα (Demsar 2006). Υπάρχει μια καθοριστική διαφορά μεταξύ των τεστ τα οποία χρησιμοποιούνται για να αξιολογήσουν την διαφορά μεταξύ δυο αλγορίθμων σε ένα σύνολο δεδομένων, και των τεστ που χρησιμοποιούνται για να αξιολογήσουν τις διαφορές σε πολλαπλά σύνολα δεδομένων. Όταν αξιολογούμε ένα μοναδικό σύνολο δεδομένων, συνήθως υπολογίζουμε την μέση απόδοση και τη διακύμανση κάνοντας χρήση εφαρμόζοντας επαναληπτική εκπαίδευση και εξέταση σε τυχαία δείγματα των παραδειγμάτων μας . Δεδομένου ότι αυτά τα δείγματα συνήθως σχετίζονται ιδιαίτερη προσοχή χρειάζεται στο σχεδιασμό των στατιστικών διαδικασιών και τεστ τα οποία αποφεύγουν προβλήματα με εσφαλμένες εκτιμήσεις διακύμανσης. Επιπλέον, το πρόβλημα εύρεσης σωστών στατιστικών τεστ για την σύγκριση αλγορίθμων σε ένα μοναδικό σύνολο δεδομένων ουσιαστικά δεν σχετίζεται με την σύγκριση σε πολλαπλά σύνολα δεδομένων από την άποψη ότι πρώτα θα έπρεπε να λύσουμε το πρόβλημα στο

μοναδικό σύνολο δεδομένων με στόχο να αντιμετωπίσουμε το πρόβλημα στα πολλαπλά σύνολα δεδομένων. Δεδομένου ότι η υλοποίηση των αλγορίθμων σε πολλαπλά σύνολα δεδομένων δίνει ως αποτέλεσμα ένα δείγμα ανεξάρτητων μετρήσεων, τέτοιου τύπου συγκρίσεις είναι απλούστερες από τις συγκρίσεις σε ένα σύνολο δεδομένων. Από εδώ και στο εξής όταν αναφερόμαστε στο "μέγεθος του δείγματος" εννοούμε τον αριθμό των χρησιμοποιούμενων συνόλων δεδομένων.

5.2.1 Σύγκριση δυο αλγορίθμων

Όσο αναφορά τα τεστ και την σύγκριση δυο αλγορίθμων σε πολλαπλά σύνολα δεδομένων πρέπει να τονίσουμε δυο σημεία. Πρώτον, το ευρέως χρησιμοποιούμενο t-test είναι συνήθως ακατάλληλο και στατιστικά μη ασφαλές. Δεύτερον, τα στατιστικά τεστ μετρούν διαφορές μεταξύ αλγορίθμων για διαφορετικούς σκοπούς, και έτσι η επιλογή του τεστ θα πρέπει να βασίζεται όχι μόνο στην στατιστική καταλληλότητα αλλά και στο τι έχουμε πρόθεση να μετρήσουμε.

5.2.1.1 Μέση απόδοση των συνόλων δεδομένων

Σε πολλές μελέτες της μηχανικής μάθησης οι ερευνητές υπολογίζουν την μέση ακρίβεια των αλγορίθμων σε όλα τα εξεταζόμενα σύνολα δεδομένων. Εάν όμως τα αποτελέσματα των διαφορετικών συνόλων δεδομένων δεν είναι συγκρίσιμα, το να υπολογίσουμε την μέση ακρίβεια δεν έχει νόημα. Ένα άλλο μειονέκτημα είναι ότι ο μέσος ορος είναι ευαίσθητος στις ακραίες τιμές. Επίσης ο υπολογισμός της μέσης ακρίβειας δεν μας βοηθά να διακρίνουμε ποιο από τα σύνολα δεδομένων είχε την καλύτερη ή χειρότερη απόδοση. Για παράδειγμα μπορεί να προκύψει ικανοποιητική μέση ακρίβεια στην περίπτωση που ένα από τα σύνολα δεδομένων αποδίδει τέλεια και τα υπόλοιπα έχουν μέτρια η κακή απόδοση ή ακόμη και στην περίπτωση που μια κακή απόδοση ενός συνόλου δεδομένων αντισταθμίζεται από την πολύ καλή απόδοση των υπολοίπων συνόλων δεδομένων. Για τους προαναφερθέντες λόγους , τα τελευταία χρόνια, η επιστημονική κοινότητα δεν χρησιμοποιεί πλέον την μέση απόδοση των αλγορίθμων ως τεστ και συνεπώς δεν την χρησιμοποιεί και για σκοπούς στατιστικής συμπερασματολογιας με το z ή t-test.

5.2.1.2 Paired T-Test

Ένας τρόπος για να ελεγχθεί αν η διαφορά μεταξύ των αποτελεσμάτων δύο αλγορίθμων σε διάφορα σύνολα δεδομένων είναι μη-τυχαία είναι να υπολογίσουμε ένα paired t-test, το οποίο ελέγχει αν η μέση διαφορά στην απόδοση τους στα σύνολα δεδομένων είναι σημαντικά διαφορετική από το μηδέν.

Αν c_i^1 και c_i^2 τα σκορ των αποδόσεων των δυο αλγορίθμων για το i-στο από N σύνολο δεδομένων d_i η διαφορά c_i^2- c_i^1. Το στατιστικό t υπολογίζεται ως $\frac{\bar{d}}{\sigma_{\bar{d}}}$ και ακολουθει την κατανομή Student με N -1 βαθμούς ελευθερίας. Το t-test παρουσιάζει τρεις αδυναμίες. Η πρώτη είναι ότι χρησιμοποιώντας το paired t-test για να συγκρίνουμε ένα ζευγάρι αλγορίθμων έχει την ίδια λογική σαν να υπολογίζαμε την

μέση απόδοση στα σύνολα δεδομένων. Η μέση διαφορά $\bar{d}$ ισούται με την διαφορά μεταξύ των μέσων σκορ δυο αλγορίθμων $\bar{d} = \overline{c^2} - \overline{c^1}$. Η μόνη διαφοροποίηση μεταξύ της μορφής του paired t-test και της σύγκρισης δυο μέσων σκορ (όπως αυτά που προαναφέρθηκαν) χρησιμοποιώντας απευθείας το t-test για μη συσχετιζόμενα δείγματα είναι στον παρανομαστή: το paired t-test μειώνει το τυπικό σφάλμα $\sigma_{\bar{d}}$ διαμέσου της διακύμανσης μεταξύ των συνόλων δεδομένων ή με άλλα λόγια διάμεσο της συνδιακυμανσης μεταξύ των αλγορίθμων. Το δεύτερο πρόβλημα με το t-test είναι ότι το paired t-test απαιτεί οι διαφορές μεταξύ των δύο συγκρινόμενων τυχαίων μεταβλητών να ακολουθουν την τυχαία κατανομή, εκτός αν το μέγεθος του δείγματος είναι αρκετά μεγάλο (περίπου 30 σύνολα δεδομένων). Το Kolmogorov-Smirnov και παρόμοια τεστ για τον έλεγχο της κανονικότητας των κατανομών έχουν μικρή ισχύ σε μικρά δείγματα, δηλαδή, είναι απίθανο να ανιχνεύσουν ανωμαλίες και προειδοποιήσουν για κακή απόδοση του t-test. Το τρίτο πρόβλημα είναι ότι το t-test , όπως ακριβώς και οι μέσες αποδόσεις στα σύνολα δεδομένων, επηρεάζεται από τις ακραίες τιμές οι οποίες παραποιούν τα στατιστικά στοιχειά του τεστ και τη μειώνουν την ισχύ του τεστ με την αύξηση του εκτιμώμενου τυπικού σφάλματος.

5.2.1.3 Wilcoxon Singed-Ranks Test

Το Wilcoxon Singed-Ranks Test (Wilcoxon, 1945) είναι ένα μη-παραμετρικο εναλλακτικό τεστ για το paired t-test, το οποίο κατατάσσει τις διαφορές των επιδόσεων των δύο αλγορίθμων για κάθε σύνολο δεδομένων, αγνοώντας τα πρόσημα και συγκρίνει τις κατατάξεις για την θετικών και αρνητικών διαφορών. Έστω d_i είναι η διαφορά μεταξύ των σκορ των αποδόσεων των δύο αλγορίθμων για i-οστό από N σύνολα δεδομένων. Οι διαφορές κατατάσσονται σύμφωνα με τις απόλυτες τιμές τους. Έστω R^+ είναι το άθροισμα των κατατάξεων για τα σύνολα δεδομένων στα οποία ο δεύτερος αλγόριθμος ξεπέρασε τον πρώτο, και R^- το άθροισμα των κατατάξεων για το αντίθετο. Οι κατατάξεις που έχουν $d_i=0$ μοιράζονται ισόποσα μεταξύ των αθροισμάτων, εάν είναι μονός ο αριθμός τους, μια αγνοείται:

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

Έστω T είναι το μικρότερο από τα αθροισματα,$T=\min(R^+,R^-)$. Οι κριτικές τιμές για το T για τιμές του N μέχρι 25 βρίσκονται σε πίνακες σε πολλά βιβλία. Για μεγαλύτερο αριθμό συνόλων δεδομένων , η στατιστική συνάρτηση z υπολογίζεται ως εξής:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}$$

και ακολουθει περίπου την κανονική κατανομή. Με επίπεδο σημαντικότητας a = 0.05, η μηδενική-υπόθεση μπορεί να απορριφθεί εάν το z είναι μικρότερο από -1.96.

5.2.1.4 Counts of Wins, Losses and Ties: Sign Test

Ένας τρόπος για να συγκριθούν οι συνολικές επιδόσεις των αλγορίθμων είναι να μετρήσουμε τον αριθμό των συνόλων δεδομένων στα οποία κάθε αλγόριθμος είναι ο καλύτερος. Όταν συγκρίνονται πολλαπλοί αλγόριθμοι με κατά ζεύγη συγκρίσεις μερικές φορές τα αποτελέσματα οργανώνονται σε έναν πίνακα.

Επίσης αυτές τις μετρήσεις χρησιμοποιούνται στην επαγωγικής στατιστική, με τη μορφή διωνυμικής δοκιμής που είναι γνωστό αλλιώς ως τεστ πρόσημων ή αλλιώς Sing Test (Sheskin, 2000, Salzberg, 1997). Έστω δυο αλγόριθμοι που συγκρίνονται ότι είναι ισοδύναμοι κάτω από την μηδενική υπόθεση τότε ο καθένας από αυτούς θα πρέπει να είναι ο καλύτερος σε N/2 από τα N σύνολα δεδομένων. Ο αριθμός των νικών κατανέμεται σύμφωνα με τη δυονυμικη κατανομή. Για έναν μεγαλύτερο αριθμό συνόλων δεδομένων, ο αριθμός των νικών που είναι υπό τη μηδενική υπόθεση διανέμονται σύμφωνα με την κατανομή N (N / 2, $\sqrt{N}$/ 2), η οποία επιτρέπει τη χρήση του z-test: εάν ο αριθμός των νικών είναι τουλάχιστον N / 2 + 1.96$\sqrt{N}$ / 2, ο αλγόριθμος είναι σημαντικά καλύτερος με p<0,05.

5.2.2 Σύγκριση πολλαπλών αλγορίθμων

Κανένα από τα προαναφερθέντα στατιστικά τεστ δεν σχεδιάστηκε με το σκεπτικό να χρησιμοποιηθεί για τους μέσους πολλαπλών τυχαίων μεταβλητών. Πολλοί ερευνητές σε μελέτες μηχανικής μάθησης, ωστόσο, τα χρησιμοποιούν για το σκοπό αυτό, παρ' όλο που όταν πραγματοποιούνται πολλαπλά τεστ, ένα συγκεκριμένο ποσοστό των μηδενικών υποθέσεων απορρίπτεται λόγω τυχαίας πιθανότητας.

Το ζήτημα των δοκιμών πολλαπλών υποθέσεων είναι ένα ευρέως γνωστό στατιστικό πρόβλημα. Ο συνήθης στόχος είναι να ελέγξουμε το σφάλμα, γνωστό ως family-wise-error, δηλαδή την πιθανότητα να κάνουμε τουλάχιστον ένα σφάλμα Τύπου I σε οποιαδήποτε από τις συγκρίσεις. Μία γενική λύση για το πρόβλημα αυτό είναι να υλοποιήσουμε τη διόρθωση Bonferroni, παρόλο που είναι συνήθως αρκετά συντηρητική και αδύναμη ως προς το ότι υποθέτει την ανεξαρτησία των υποθέσεων.

Η στατιστική προσφέρει πολύ πιο ισχυρές εξειδικευμένες διαδικασίες για τον έλεγχο της σημαντικότητας των διαφορών μεταξύ πολλαπλών μέσων. Οι δύο πιο ενδιαφέρουσες είναι η γνωστή μέθοδος της ανάλυσης διακύμανσης (ANOVA), και η αντίστοιχη μη παραμετρική ομόλογή της, το τεστ του Friedman. Το τεστ Friedman, και ιδιαίτερα το αντίστοιχό του post-hoc τεστ, το Nemenyi post-hoc τεστ, είναι λιγότερο γνωστά και περιγράφονται ελάχιστα στη διεθνή βιβλιογραφία, και περιγράφονται αναλυτικά ως ακολούθως.

5.2.2.1 Ανάλυση διακύμανσης

Η κλασική στατιστική μέθοδος που χρησιμοποιείται για τον έλεγχο των διαφορών μεταξύ περισσοτέρων από δύο σχετιζόμενων μέσων δειγμάτων, είναι η ανάλυση διακύμανσης επαναλαμβανόμενων μετρήσεων (*repeated-measures ANOVA).* Οι αποδόσεις των αλγορίθμων μετρούμενες στα ίδια σύνολα δεδομένων, και κατά προτίμηση χρησιμοποιώντας τον ίδιο διαχωρισμό στα σύνολα εκπαίδευσης και εξέτασης, αποτελούν τα "σχετιζόμενα δείγματα" σε αυτήν την περίπτωση. Η μηδενική υπόθεση η οποία εξετάζεται είναι ότι όλοι οι αλγόριθμοι αποδίδουν το ίδιο και οι παρατηρούμενες διαφορές είναι απλώς τυχαίες.

Η ανάλυση διακύμανσης διαχωρίζει τη συνολική μεταβλητότητα, στη μεταβλητότητα μεταξύ των αλγορίθμων, τη μεταβλητότητα μεταξύ των συνόλων δεδομένων και τη μεταβλητότητα που οφείλεται στα σφάλματα, δηλ., στα υπόλοιπα (residuals). Εάν η μεταβλητότητα μεταξύ των αλγορίθμων είναι σημαντικά μεγαλύτερη από τη μεταβλητότητα του σφάλματος, τότε μπορούμε να απορρίψουμε τη μηδενική υπόθεση, και καταλήγουμε στο συμπέρασμα ότι υπάρχουν ορισμένες διαφορές μεταξύ των αλγορίθμων. Σε αυτή την περίπτωση, μπορούμε να προχωρήσουμε με ένα εκ των υστέρων τεστ (post-hoc test) για να βρούμε ποιοι αλγόριθμοι διαφέρουν στην πραγματικότητα.

Μέσα από τα πολλά τεστ που χρησιμοποιούνται για την ανάλυση διακύμανσης, δύο από τα καταλληλότερα για τη σύγκριση πολλαπλών αλγορίθμων είναι το τεστ του Tukey, για τη σύγκριση όλων των αλγορίθμων μεταξύ τους ο ένας με το άλλον, και το τεστ Dunnett όταν θέλουμε να συγκρίνουμε αλγορίθμους με τις αντίστοιχες βελτιωμένες εκδοχές τους ή επεκτάσεις τους, είτε συγκρίνοντας νέους προτεινόμενους αλγόριθμους με διάφορους ήδη υπάρχοντες. Αμφότερα τα δύο τεστ υπολογίζουν το τυπικό σφάλμα της διαφοράς μεταξύ δύο αλγορίθμων, διαιρώντας τη διακύμανση των υπολοίπων από τον αριθμό των συνόλων δεδομένων. Για να γίνουν κατά ζεύγη συγκρίσεις μεταξύ των αλγορίθμων, οι αντίστοιχες διαφορές στις αποδόσεις διαιρούνται με το τυπικό σφάλμα και συγκρίνονται με την κρίσιμη τιμή. Διαπιστώνουμε έτσι, ότι τα δύο αυτά τεστ, είναι παρόμοιας λογικής με το t-test, εκτός του ότι οι κρίσιμες τιμές για τα τεστ των Tukey και Dunnett είναι υψηλότερες, ούτως ώστε εξασφαλιστεί ότι υπάρχει το πολύ 5% πιθανότητα ότι μία από τις κατά ζεύγη διαφορές βρέθηκε να είναι εσφαλμένα σημαντική.

Δυστυχώς, η ανάλυση διακύμανσης βασίζεται σε παραδοχές οι οποίες πιθανότατα παραβιάζονται όταν αναλύουμε την απόδοση αλγορίθμων μηχανικής μάθησης. Πρώτον, η ανάλυση διακύμανσης υποθέτει ότι τα δείγματα προέρχονται από την κανονική κατανομή. Η δεύτερη και πιο σημαντική υπόθεση για την ανάλυση διακύμανσης επαναλαμβανόμενων μετρήσεων είναι η σφαιρικότητα (μια ιδιότητα που είναι παρόμοια με την ομοιογένεια της διακύμανσης στη συνήθη ανάλυση διακύμανσης, η οποία απαιτεί οι τυχαίες μεταβλητές να έχουν την ίδια διακύμανση).

Λόγω της φύσης των αλγορίθμων μηχανικής μάθησης καθώς και των συνόλων δεδομένων, οι παραδοχές αυτές δεν μπορούν να θεωρηθούν δεδομένες. Μάλιστα, οι παραβιάσεις αυτών των παραδοχές θα έχουν ακόμη μεγαλύτερη επίδραση στα post-hoc τεστ. Ως εκ τούτου, η ανάλυση διακύμανσης δεν φαίνεται να είναι ένα κατάλληλο τεστ για γενική χρήση σε τυπικές μελέτες μηχανικής μάθησης.

5.2.2.2 Το τεστ του Friedman

Το τεστ του Friedman αποτελεί ένα μη παραμετρικό στατιστικό τεστ ισοδύναμο με τη μέθοδο της ανάλυσης διακύμανσης επαναλαμβανόμενων μετρήσεων. Συγκεκριμένα, κατατάσσει τους αλγορίθμους για κάθε σύνολο δεδομένων ξεχωριστά, με τον αλγόριθμο που εμφανίζει την καλύτερη απόδοση να λαμβάνει τιμή κατάταξης 1, τον αλγόριθμο με τη δεύτερη καλύτερη κατάταξη να λαμβάνει τη τιμή 2 κ.λπ. Έστω r_{ij} του j-οστού από τους k αλγορίθμους στο i-οστό από τα N σύνολα δεδομένων. Το τεστ του Friedman συγκρίνει τις μέσες τιμές κατάταξης των αλγορίθμων, $R_j = \frac{1}{N}\sum_i r_i^j$. Υπό την μηδενική υπόθεση όλοι οι αλγόριθμοι είναι ισοδύναμοι και έτσι οι τιμές κατάταξής τους Rj πρέπει να ίσες.

Το στατιστικό τεστ του Friedman

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

ακολουθεί την χ_F^2 κατανομη με k-1 βαθμούς ελευθερίας όταν το N και το k είναι αρκετά μεγάλα (κανόνας κατωφλιού N>10 k>5). Για ένα μικρότερο αριθμό αλγορίθμων και συνόλων δεδομένων, έχουν υπολογιστεί ακριβώς οι κρίσιμες τιμές τους (Demsar 2006).

Ο Iman και Davenport (1980) έδειξαν ότι το στατιστικό τεστ του Friedman είναι ανεπιθύμητα συντηρητικό και πρότειναν ένα καλύτερο στατιστικό τεστ που ακολουθει την F κατανομή με k-1 και (k-1)(N-1) βαθμούς ελευθερίας.

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

Στις συγκρίσεις αλγορίθμων ένα μη-παραμετρικο τεστ (Friedman) έχει λιγότερη ισχύ από ένα παραμετρικό (ANOVA) όταν οι υποθέσεις του παραμετρικού ισχυουν, δεν ισχύει εάν δεν τηρούνται οι υποθέσεις.

Στην περίπτωση που απορριφθεί η μηδενική υπόθεση προχωράμε σε ένα post-hoc τεστ. Το τεστ Nemenyi (1963) είναι παρόμοιο με το τεστ Tukey, για την ανάλυση διακύμανσης και χρησιμοποιείται όταν όλοι οι αλγόριθμοι συγκρίνονται μεταξύ τους ο ένας με τον άλλο. Η απόδοση δυο αλγορίθμων διαφέρει σημαντικά όταν οι αντίστοιχες τιμές κατάταξης διαφέρουν τουλάχιστον από την κρίσιμη διαφορά

$$CD = q_a \sqrt{\frac{k(k+1)}{6N}}$$

όπου κρίσιμες τιμές QA βασίζονται στη στατιστική studentized φάσμα διαιρείται με P2.

Στην περίπτωση όπου ένας αλγόριθμος συγκριθεί με βελτιωμένη εκδοχή του ή ένας καινούργιος προτεινόμενος αλγόριθμος συγκριθεί με άλλους ήδη χρησιμοποιημένους, εναλλακτικά αντί του Nemenyi τεστ μπορούμε να εφαρμόσουμε το κριτήριο Bonferroni για να ελέγξουμε το σφάλμα famile-wise error στα τεστ πολλαπλών υποθέσεων.

Το στατιστικό τεστ για να συγκρίνουμε το i-οστό και τον j-οστό αλγόριθμο χρησιμοποιώντας είτε το Nemenyi τεστ είτε το Bonferroni είναι το

$$z = (R_i - R_j / \sqrt{\frac{k(k+1)}{6N}})$$

Η τιμή του z χρησιμοποιείται για να βρει την αντίστοιχη πιθανότητα από τον πίνακα της κατανομής και έπειτα συγκρίνεται με ένα κατάλληλο α. Τα δυο αυτά τεστ διαφέρουν ως προς τον τρόπο με τον οποίο προσαρμόζουν το α σε σχέση με τον αριθμό των συγκρίσεων που πραγματοποιούνται.

6. Πειραματική μελέτη

Παρακάτω περιγράφονται αναλυτικά τα σύνολα δεδομένων τα οποία χρησιμοποιούνται στην παρούσα μελέτη, οι μετασχηματισμοί τροχιών που εφαρμόζονται στα σύνολα αυτά, καθώς και οι προκύπτουσες πειραματικές παρατηρήσεις-εξαγόμενα αποτελέσματα σε σχέση με την εγκυρότητα των ομαδοποιήσεων που επιτυγχάνονται από τον αλγόριθμο optics και την ιεραρχική ομαδοποίηση με την μέθοδο Ward, αντίστοιχα.

6.1 Περιγραφή συνόλων δεδομένων

Στην παρούσα διπλωματική εργασία, θα χρησιμοποιήσουμε κατά τη διαδικασία του πειραματισμού ένα πραγματικό και ένα συνθετικό σύνολο δεδομένων.

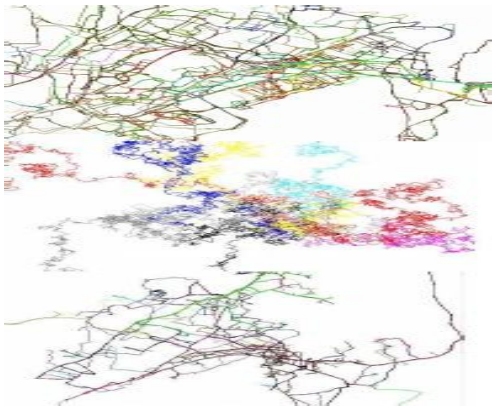
Συγκεκριμένα στα πειράματα που ακολουθούν έχουμε χρησιμοποιήσει ένα μέρος από ένα πραγματικό σύνολο δεδομένων που αφορά σε στόλο από φορτηγά καλούμενο εφεξής ως “Trucks”. Το σύνολο δεδομένων Trucks αποτελείται από 276 τροχιές, οι οποίες αντιστοιχούν σε 50 φορτηγά παράδοσης σκυροδέματος σε διάφορες περιοχές γύρω από την μητροπολιτική περιοχή της Αθήνας, σε χρονικό διάστημα τριαντατριών ημερών. Στη συγκεκριμένη εργασία θα χρησιμοποιήσουμε 53 τροχιές από 50 φορτηγά.

Η δομή της κάθε εγγραφής για το σύνολο Trucks έχει ως εξής:

{obj_id, traj_id, ημερομηνία (ηη/μμ/εεεε), ώρα (ωω:λλ:δδ), lat, lon, x, y},

όπου οι (lat, lon) είναι σε σύστημα αναφοράς WGS84, και οι (x, y) είναι σε σύστημα αναφοράς GGRS87.

Εικόνα 16: Οπτικοποίηση του συνόλου Trucks



Κατά τη διαδικασία του πειραματισμού εκτός από το πραγματικό σύνολο δεδομένων χρησιμοποιήθηκε και ένα συνθετικό με σκοπό τα πιο σφαιρικά συμπεράσματα

Το συνθετικό σύνολο δεδομένων 1d4p100lonlat το οποίο έχει μετονομαστεί σε "labeled" αποτελείται από 100 τροχιές οι οποίες αντιστοιχούν σε 100 κινούμενα αντικείμενα σε διάστημα μιας ημέρας, και οι ομάδες οι οποίες ανήκουν είναι γνωστές εκ των προτέρων.

Η δομή της κάθε εγγραφής για το σύνολο "labeled" έχει ως εξής:

{obj_id, traj_id, ημερομηνία (ηη / μμ / εεεε), ώρα (ωω: λλ: δδ), lat, lon},

όπου (lat, lon) είναι σε σύστημα αναφοράς WGS84.

6.2 Μετασχηματισμοί τροχιών

Στην ενότητα αυτή περιγράφουμε τα είδη των μετασχηματισμών τροχιών που εφαρμόστηκαν στην πειραματική μας μελέτη, καθώς και τα βήματα υλοποίησής τους.

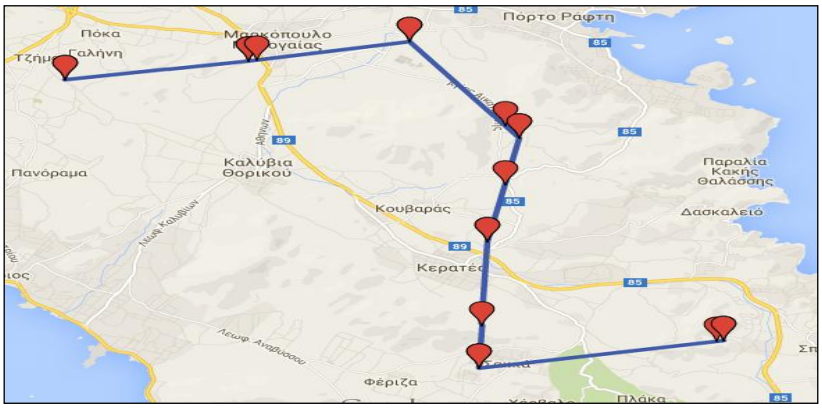
Αρχικά έχουμε την πρωτότυπη τροχιά. Στη συνέχεια κάνουμε τους διάφορους τύπους των μετασχηματισμών στην πρωτότυπη τροχιά με ελεγχόμενο τρόπο (με τη χρήση παραμέτρων), και έτσι καταλήγουμε σε σύνολα δεδομένων με μετασχηματισμένες τροχιές. Για κάθε μετασχηματισμό, θα αξιολογήσουμε την ομαδοποίηση του αρχικού συνόλου δεδομένων και των μετασχηματισμένων συνόλων δεδομένων ανάλογα με την τιμή της παραμέτρου που "τρέχει". Το λογικό είναι η ομαδοποίηση των συνόλων δεδομένων με τον χαμηλότερο βαθμό μετασχηματισμού να μην επηρεάζει την εγκυρότητα της ομαδοποίησης ιδιαίτερα.

Οι μετασχηματισμοί οι οποίοι θα χρησιμοποιήσουμε είναι η επαναδειγματολειψία, η προσθήκη θορύβου και η μετατόπιση (Wang et al. 2013) Αυτοί οι μετασχηματισμοί ελέγχονται από δυο παραμέτρους, τον ρυθμό και την απόσταση. Η παράμετρος ρυθμός χρησιμοποιείται για να προσδιορίσει το ποσοστό της τροχιάς το οποίο θα μετασχηματιστεί. Για παράδειγμα, εάν είχαμε μια τροχιά η οποια αποτελείται από 7 σημεία και θέλαμε να μετασχηματιστεί το 60% της τροχιάς, η τιμή της παραμέτρου ρυθμός θα ήταν rate=0.6 και έτσι θα μετασχηματίζονταν τα 4 από τα 7 σημεία. Επίσης, αν για παράδειγμα, ρυθμίσουμε την απόσταση distance=5000 σημαίνει ότι τα σημεία της τροχιάς που θα μετασχηματιστούν, θα έχουν μετατοπιστεί περίπου 5 χιλιόμετρα . Η απόσταση παράμετρος είναι ένα κατώφλι για το πόσο μακριά ένα σημείο της τροχιάς θα μπορούσε να μετατοπιστεί σε σχέση με το αρχικό σημείο.

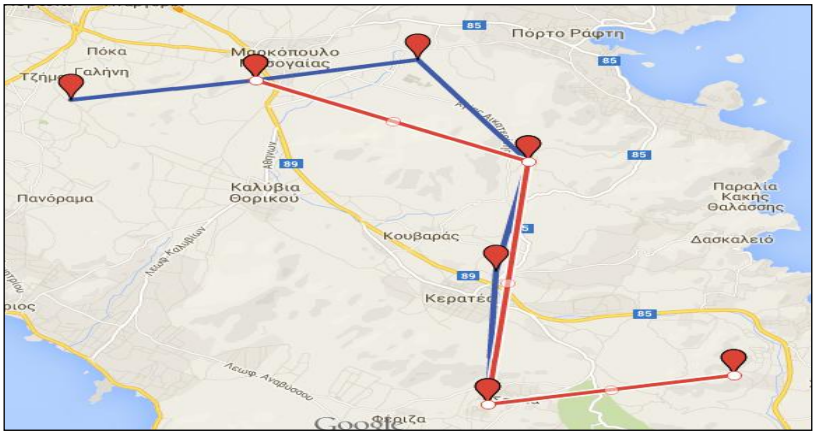
Επαναδειγματοληψία

Υπάρχουν δυο τρόποι για να γίνει επαναδειγματολειψία σε μια τροχιά. Ο ένας είναι να αυξήσουμε τα σημεία δειγματοληψίας και ο δεύτερος είναι να τα μειώσουμε. Στην αύξηση του ρυθμού δειγματοληψίας θα προσθέσουμε το αντίστοιχο ποσοστό σημείων που θέλουμε να προστεθούν στην τροχιά, ενώ αντίθετα στη μείωση του ρυθμού δειγματοληψίας θα αφαιρεθούν σημεία από την τροχιά.

Εικόνα 17: Αύξηση του ρυθμού δειγματοληψίας



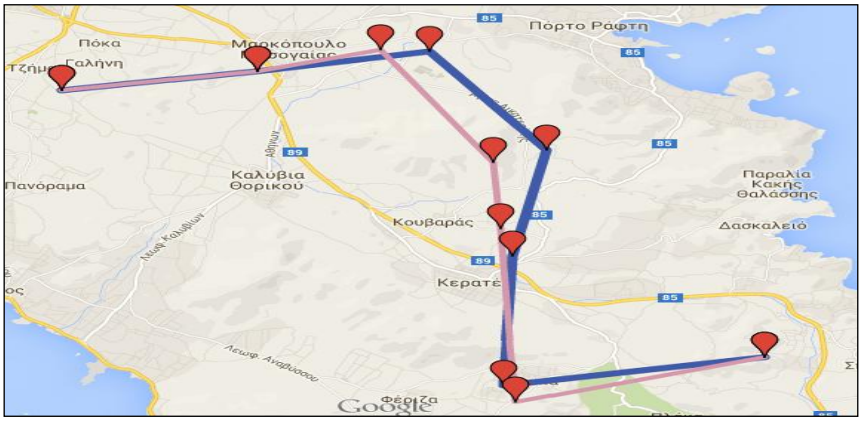
Εικόνα 18: Μείωση του ρυθμού δειγματοληψίας



Μετατόπιση σημείων

Σε αντίθεση με τον μετασχηματισμό της επαναδειγματοληψίας, στον μετασχηματισμό της μετατόπισης σημείων δεν υπάρχει πρόσθεση η αφαίρεση σημείων παρά μόνον η μετατόπισή τους ανάλογα με τις τιμές των παραμέτρων distance και rate. Υπάρχουν δυο τρόποι για να γίνει αυτός ο μετασχηματισμός. Ο πρώτος τρόπος είναι η τυχαία μετατόπιση η οποία αλλάζει την θέση των σημείων τυχαία χωρίς να λαμβάνει υπόψη την μετατόπιση ή όχι των άλλων σημείων, ενώ αντίθετα στην συγχρονισμένη μετατόπιση έχουμε την ίδια συμπεριφορά προς όλα τα επιλεγμένα σημεία της τροχιάς που θα μετατοπιστούν.

Εικόνα 19: Τυχαία μετατόπιση



Εικόνα 20: Συγχρονισμένη μετατόπιση

Προσθήκη θορύβου

Ο τελευταίος μετασχηματισμός είναι ο μετασχηματισμός της προσθήκης θορύβου ο οποίος προσθέτει στην αρχική τροχιά σημεία θορύβου. Το χάσμα ανάμεσα στην αρχική τροχιά και στα σημεία θορύβου ελέγχεται από την παράμετρο distance, ενώ το ποσοστό των σημείων θορύβου που θα προστεθεί από την παράμετρο rate.

Στο Hermes (Pelekis et al. 2015) υπάρχει η δυνατότητα να εκτελεστούν οι παραπάνω μετασχηματισμοί σε ένα σύνολο το οποίο έχει αποθηκευτεί στην μηχανή. Έτσι παρέχεται η δυνατότητα στον χρήστη για την παραγωγή πολλών συνόλων δεδομένων με βάση το αρχικό. Ο χρήστης έχει τη δυνατότητα να εξαγάγει το αποτελέσματα των μετασχηματισμών σε csv αρχείο ή απλά να το αποθηκεύσει μέσα στην βάση δεδομένων.

Εικόνα 21: Προσθήκη θορύβου

Στον Πίνακα 1 που ακολουθεί συνοψίζονται οι μετασχηματισμοί οι οποίοι θα χρησιμοποιήσουμε.

Πίνακας 1: Τύποι μετασχηματισμών και ελεγχόμενες παράμετροι κατά αντιστοιχία

Τύπος Μετασχηματισμού	**Λειτουργία**	**Ελεγχόμενες Παράμετροι**
Επαναδειγματοληψία	Αύξηση ρυθμού δειγματοληψίας(προσθήκη σημείων)	Ρυθμός (rate)
	Μείωση ρυθμού δειγματοληψίας(αφαίρεση σημείων)	Ρυθμός (rate)
Μετατόπιση Σημείου	Τυχαία μετατόπιση	Ρυθμός (rate), Απόσταση(distance)
	Συγχρονισμένη μετατόπιση	Ρυθμός(rate), Απόσταση(distance)
Προσθήκη Θορύβου	Προσθήκη θορύβου	Ρυθμός(rate), Απόσταση(distance)

6.3 Αξιολόγηση εγκυρότητας ομαδοποίησης του αλγορίθμου Optics

Στην ενότητα αυτή θα μελετήσουμε τη συμπεριφορά της ομαδοποίησης του αλγορίθμου Optics, κάνοντας χρήση εσωτερικών και εξωτερικών δεικτών αξιολόγησης της εγκυρότητας των συστάδων. Θεωρούμε κάθε φορά διαφορετικές συναρτήσεις απόστασης και διάφορα είδη μετασχηματισμών.

Για την υλοποίηση του μετασχηματισμού για την προσθήκη θορύβου, αρχικά σταθεροποιούμε την παράμετρο distance (=5.5km) και αλλάζουμε την τιμή της παραμέτρου rate από το 0.2 έως το 0.8, με βήμα=0.2. Έπειτα, σταθεροποιούμε την παράμετρο rate (=0.3) και αλλάζουμε την τιμή της παραμέτρου distance από 3 km έως 9 km, με βήμα=2 km.

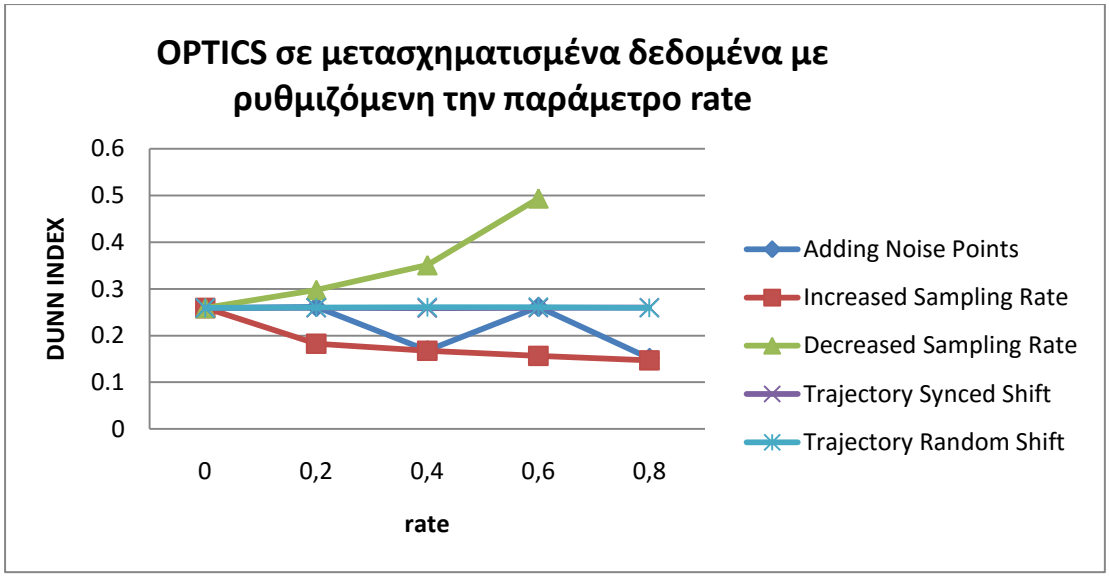
Για την υλοποίηση του μετασχηματισμού της μετατόπισης αρχικά σταθεροποιούμε την παράμετρο distance (=3km) και αλλάζουμε την τιμή της παραμέτρου rate από το 0.2 έως το 0.8, με βήμα=0.2. Έπειτα, σταθεροποιούμε την παράμετρο rate (=0.3) και αλλάζουμε την τιμή της παραμέτρου distance από το 0.5 km έως 5 km, με βήμα=1.5 km.

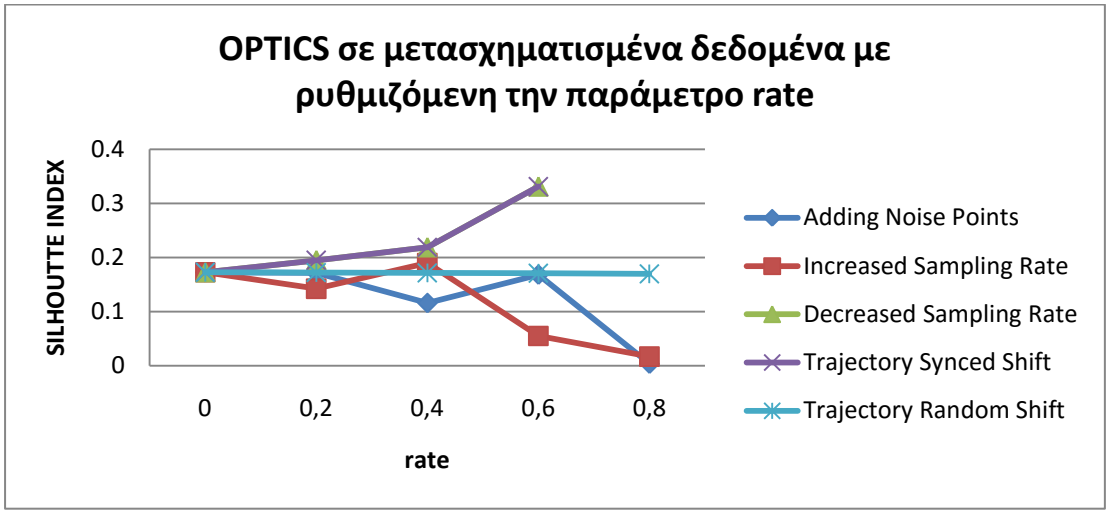
Για την υλοποίηση του μετασχηματισμού της επαναδειγματοληψίας, ρυθμίζουμε μόνο την τιμή της παραμέτρου rate από το 0.2 έως το 0.8, με βήμα=0.2.

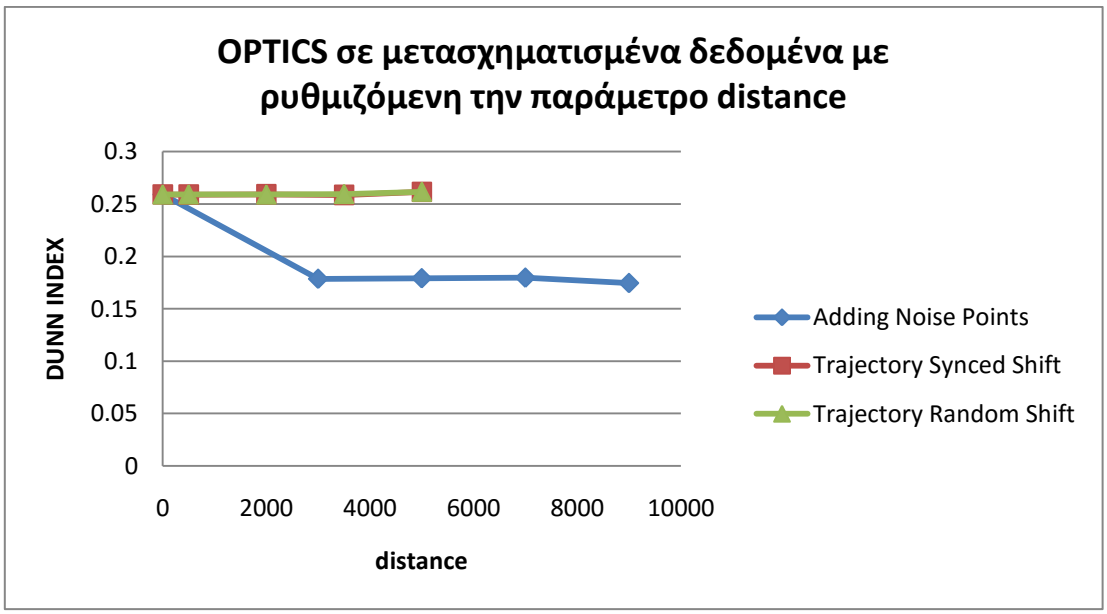
6.3.1 Εσωτερική αξιολόγηση ομαδοποίησης Optics

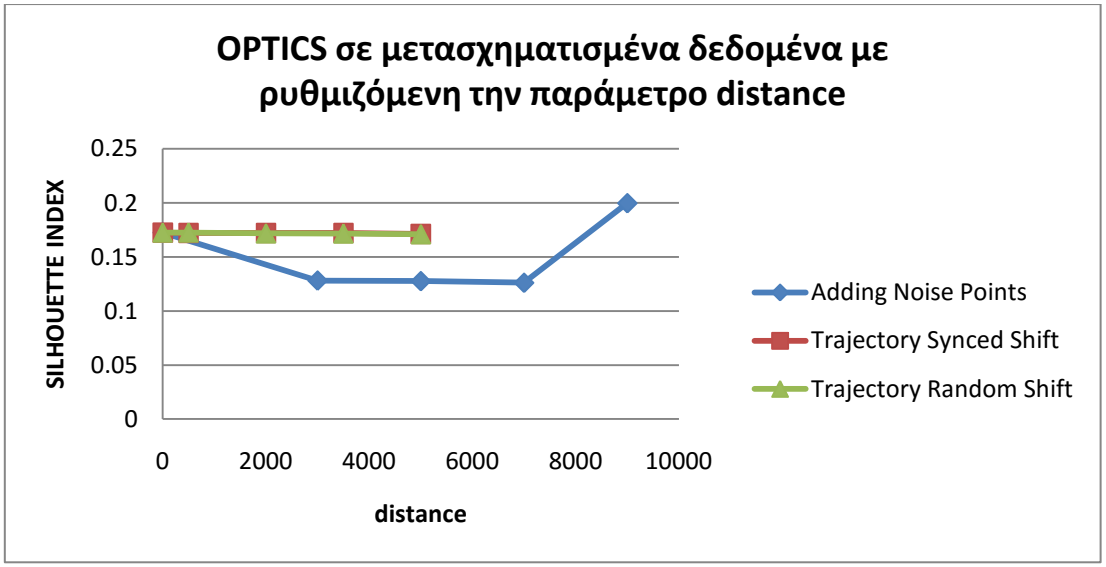
- Ευκλείδεια απόσταση

Με χρήση της ευκλείδειας απόστασης ο αλγόριθμος Optics δείχνει να είναι ευαίσθητος στην προσθήκη θορύβου, και στις δύο περιπτώσεις (είτε θεωρήσουμε σταθερή την απόσταση και ρυθμιζόμενη την παράμετρο rate είτε το αντίστροφο). Επιπλέον δείχνει να είναι ευαίσθητος στην αύξηση και στη μείωση του ρυθμού δειγματοληψίας με ρυθμιζόμενη την παράμετρο rate. Στην συγχρονισμένη, όπως και στην τυχαία μετατόπιση και στις δύο περιπτώσεις (είτε θεωρήσουμε σταθερή την απόσταση και ρυθμιζόμενη την παράμετρο rate είτε το αντίστροφο) φαίνεται ο αλγόριθμος Optics να είναι ανθεκτικός.


OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
DUNN INDEX
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0,2
0,4
0,6
0,8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

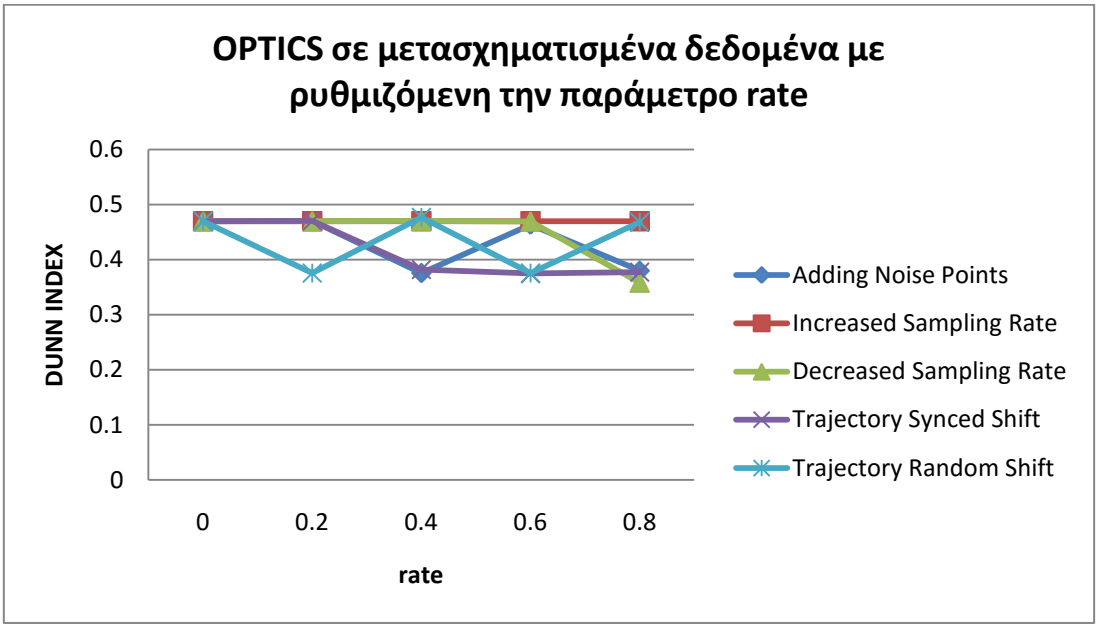

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUTTE INDEX
0.4
0.3
0.2
0.1
0
0
0,2
0,4
0,6
0,8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

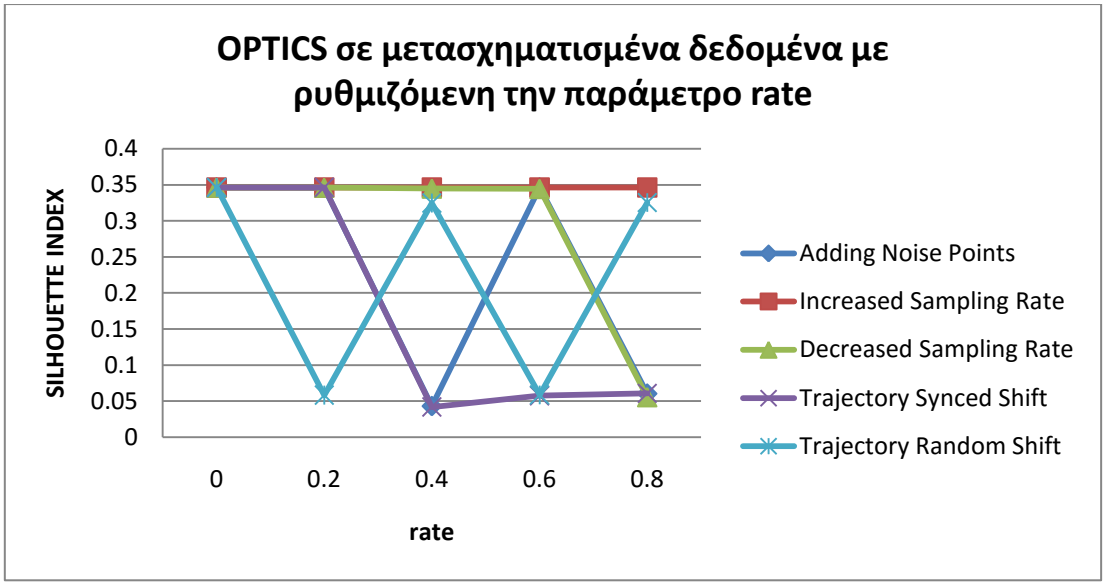

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
DUNN INDEX
0.3
0.25
0.2
0.15
0.1
0.05
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

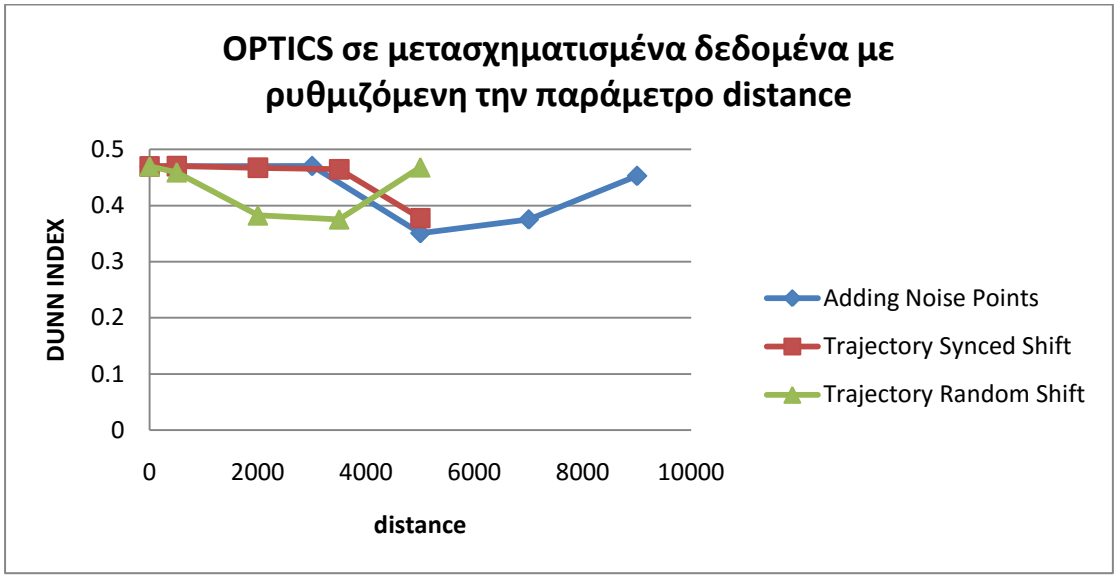


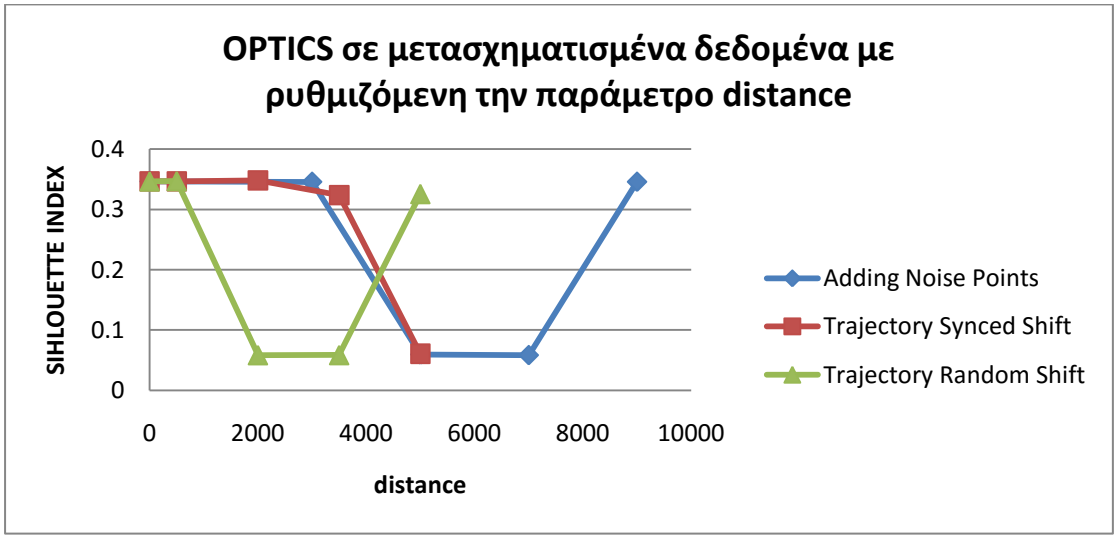
- Ευκλείδεια STARTEND

Με χρήση της απόστασης Euclideanstartend, ο αλγόριθμος Optics δείχνει να είναι ευαίσθητος στην προσθήκη θορύβου και στους δυο μετασχηματισμούς της μετατόπισης. Ακόμη φαίνεται να είναι ανθεκτικός στον μετασχηματισμό της αύξησης του ρυθμού δειγματοληψίας. Τέλος, στη μείωση του ρυθμού δειγματοληψίας παρατηρούμε ότι έχει αποδεκτή συμπεριφορά στις αρχικές τιμές της ρυθμιζόμενης παραμέτρου και είναι πολύ ευαίσθητος στην τελευταία.



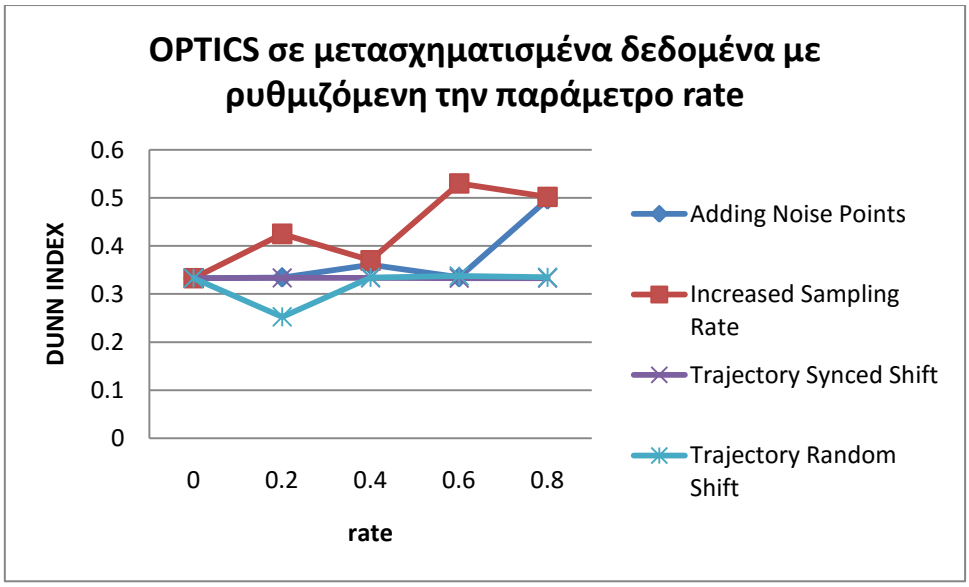

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUETTE INDEX
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

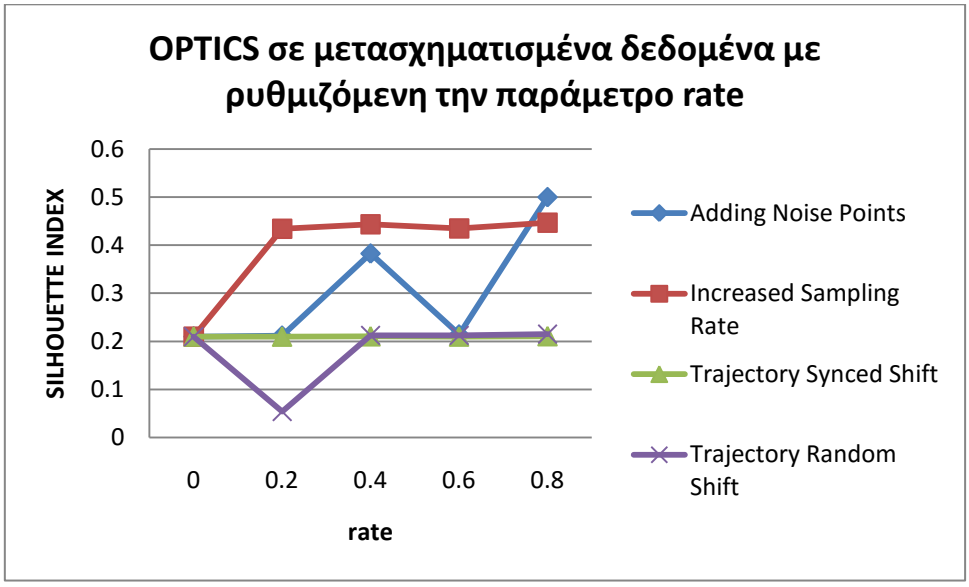

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
DUNN INDEX
0.5
0.4
0.3
0.2
0.1
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

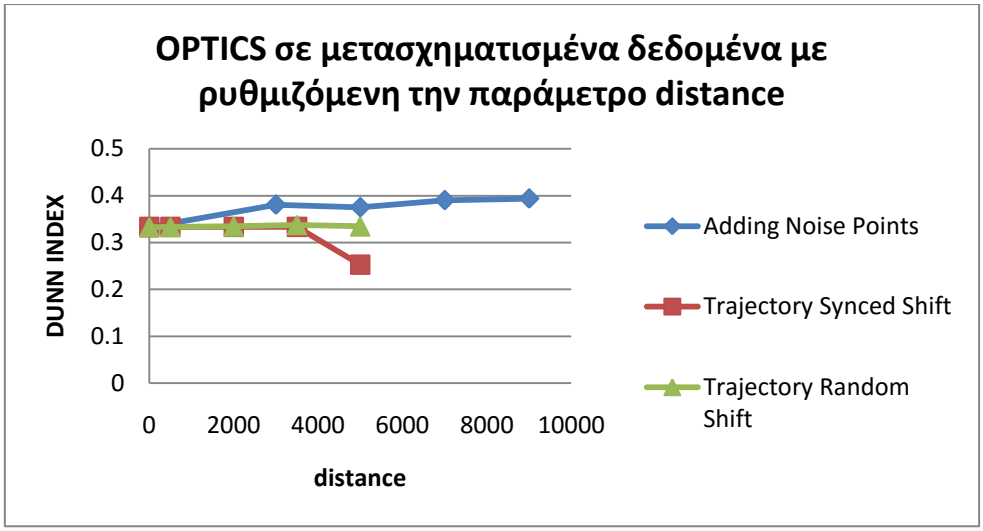

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
SIHLOUETTE INDEX
0.4
0.3
0.2
0.1
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

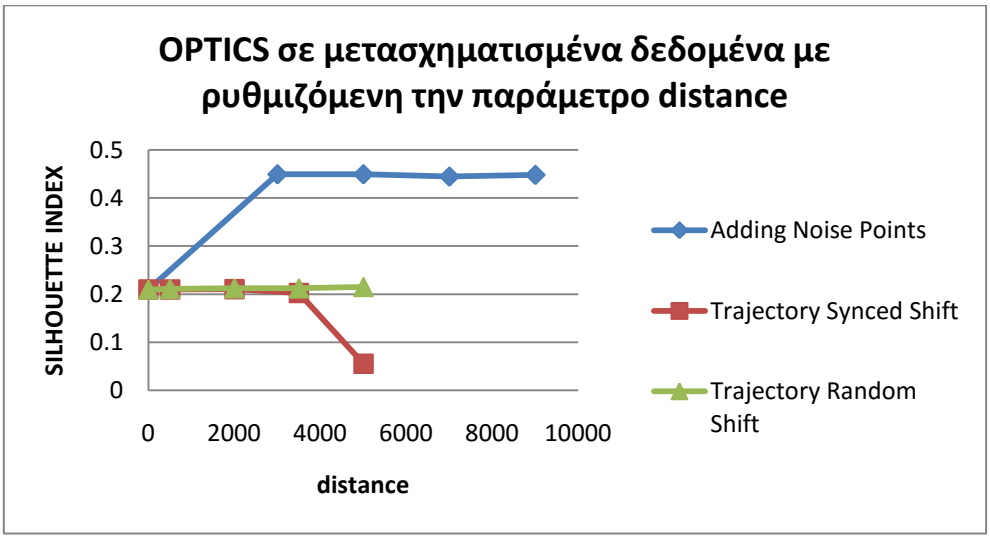
- Manhattan απόσταση

Με χρήση της manhattan απόστασης ο αλγόριθμος Optics δείχνει να επηρεάζεται από την προσθήκη θορύβου και την αύξηση του ρυθμού δειγματοληψίας. Αντίθετα, δείχνει να είναι ανθεκτικός στον μετασχηματισμό της συγχρονισμένης μετατόπισης και να έχει μια αποδεκτή συμπεριφορά στην περίπτωση της τυχαίας μετατόπισης. Για την περίπτωση του μετασχηματισμού της μείωσης του ρυθμού δειγματοληψίας δεν μπορούν να βγουν συμπεράσματα επειδή η ομαδοποίηση γίνεται μόνο σε μια ομάδα.



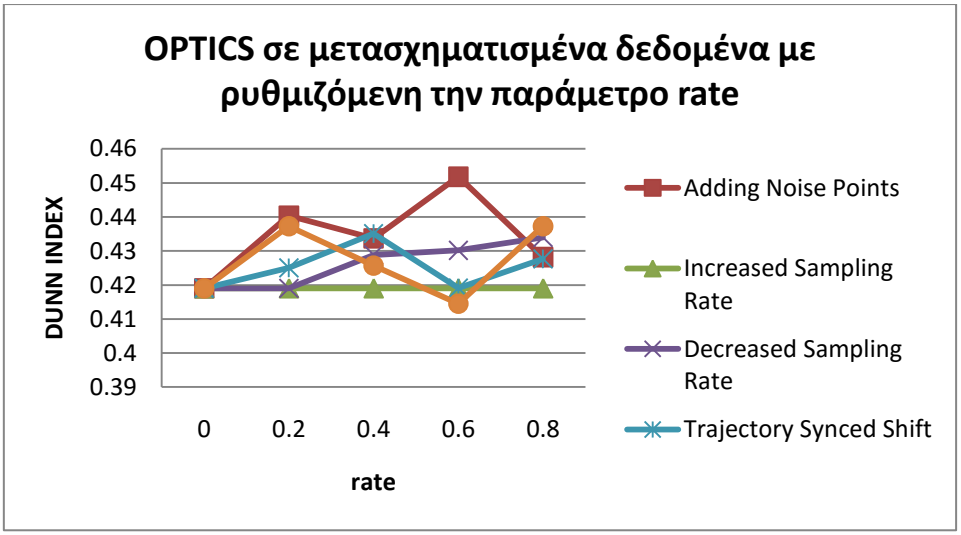


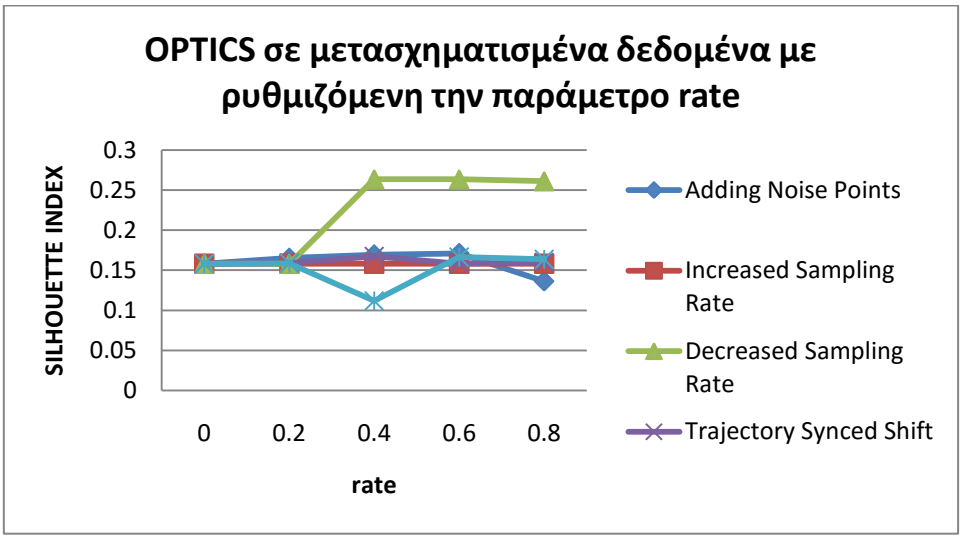


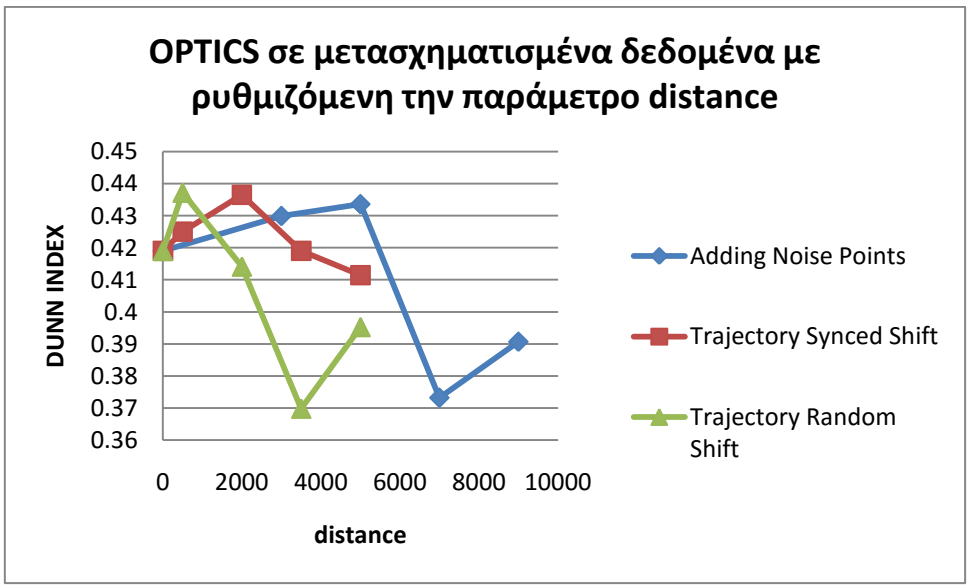


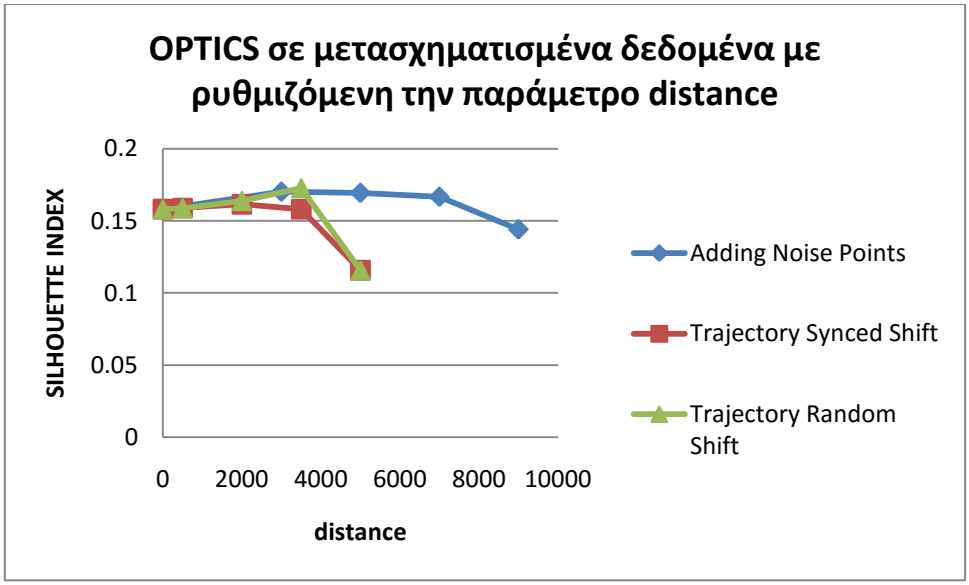
- Chebyshev απόσταση

Με χρήση της απόστασης Chebyshev, ο αλγόριθμος Optics δείχνει να είναι ευαίσθητος στην προσθήκη θορύβου. Επιπλέον, εφαρμόζοντας τον μετασχηματισμό αύξησης ρυθμού δειγματοληψίας ο αλγόριθμος δείχνει να είναι ανθεκτικός. Αντίθετα, στη μείωση του ρυθμού δειγματοληψίας δείχνει να είναι ευαίσθητος. Τέλος, ευαίσθητος δείχνει να είναι και στους μετασχηματισμούς της μετατόπισης.


OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
DUNN INDEX
0.46
0.45
0.44
0.43
0.42
0.41
0.4
0.39
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift

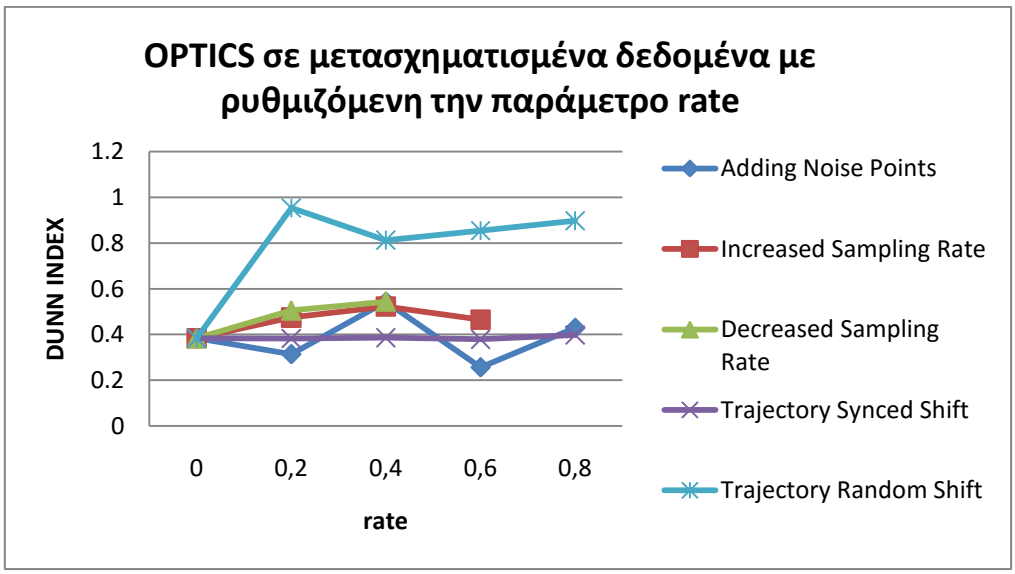

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUETTE INDEX
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift

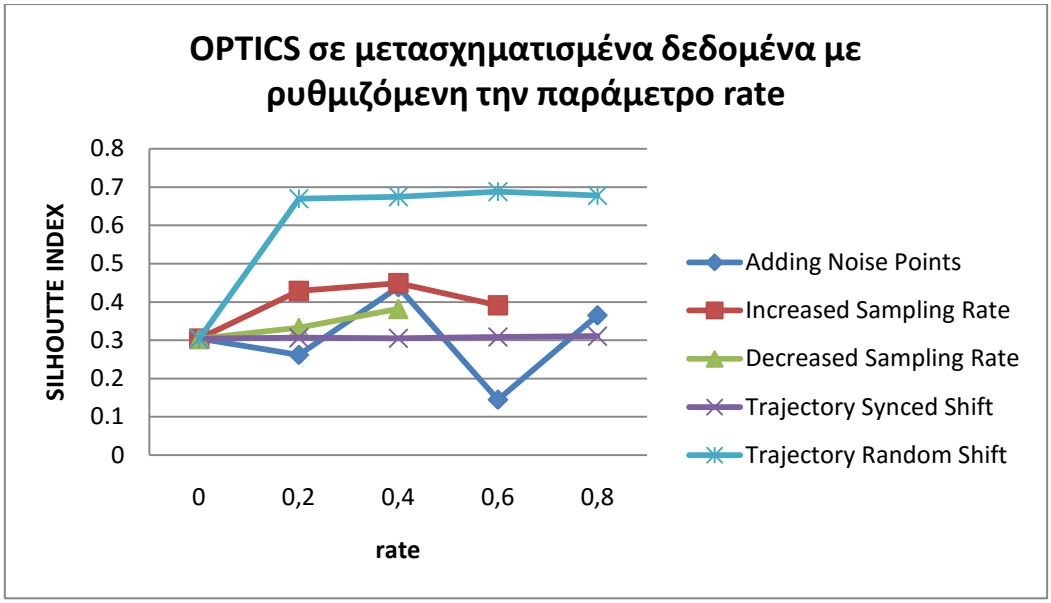

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
DUNN INDEX
0.45
0.44
0.43
0.42
0.41
0.4
0.39
0.38
0.37
0.36
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

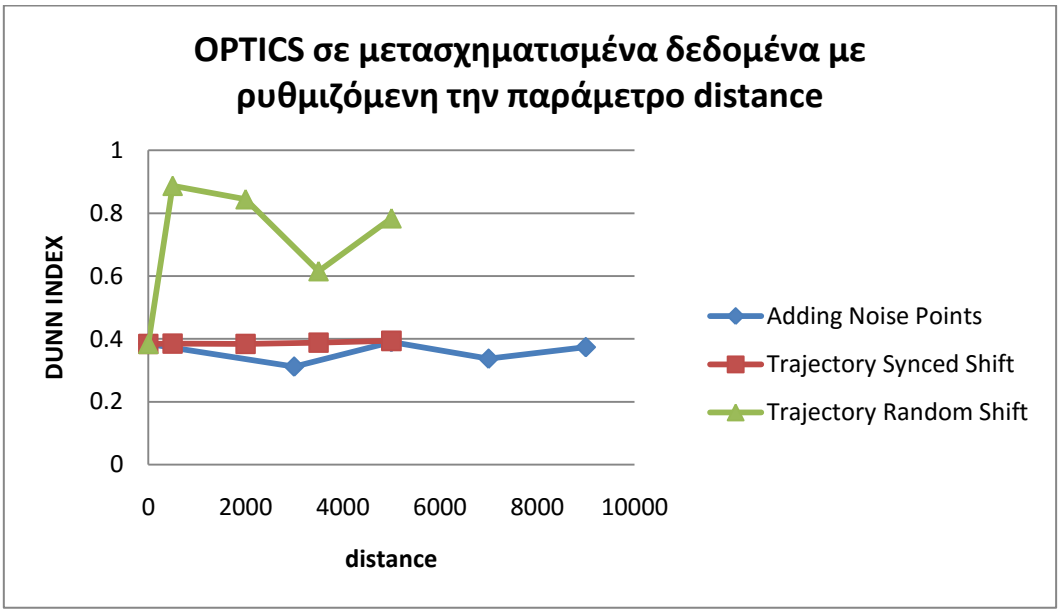


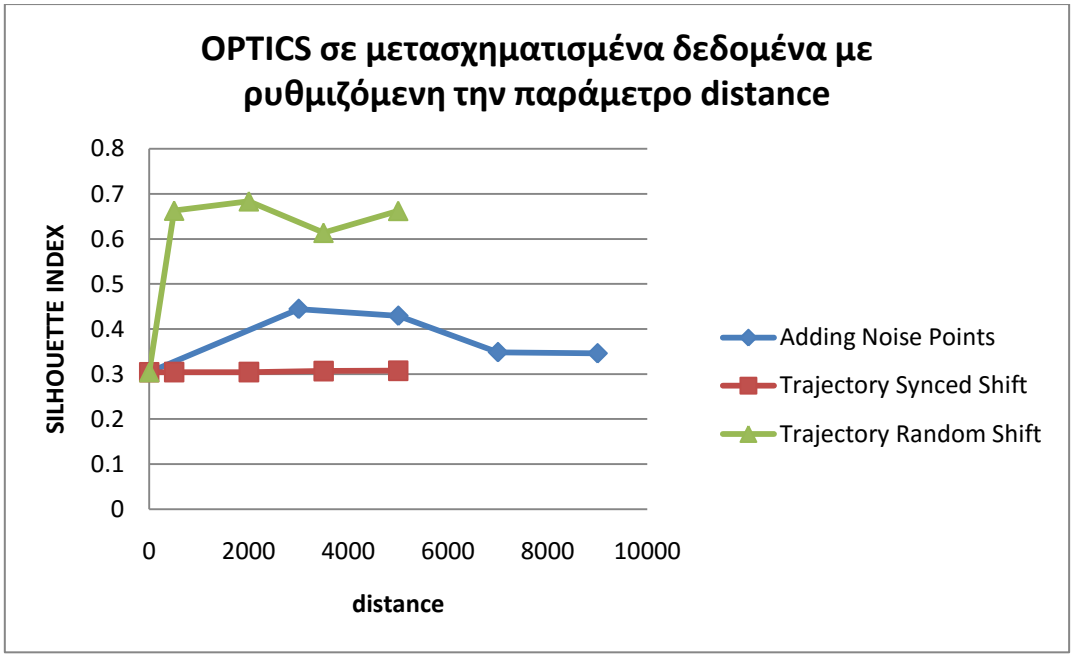
- Dynamic Time Warping

Ο αλγόριθμος Optics με την χρήση της απόστασης DTW φαίνεται να είναι ευαίσθητος στον μετασχηματισμό της προσθήκης θορύβου. Αποδεκτή συμπεριφορά φαίνεται να παρουσιάζει στον μετασχηματισμό της αύξησης του ρυθμού δειγματοληψίας, ενώ δείχνει να είναι ευαίσθητος στη μείωση του ρυθμού δειγματοληψίας. Τέλος, είναι ευαίσθητος στην τυχαία μετατόπιση και παρουσιάζεται ανθεκτικός στην συγχρονισμένη μετατόπιση.



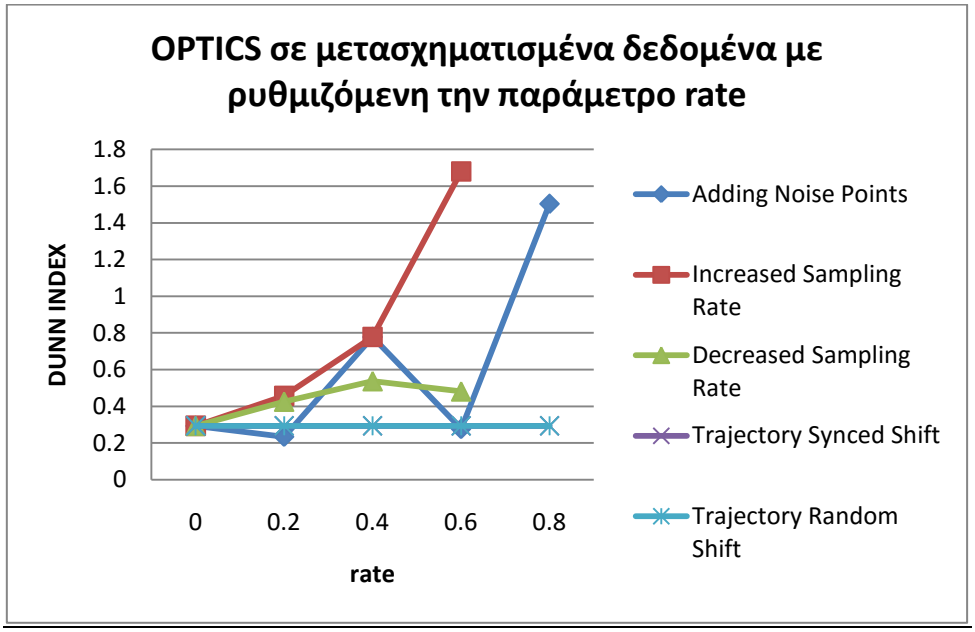

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUTTE INDEX
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0,2
0,4
0,6
0,8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

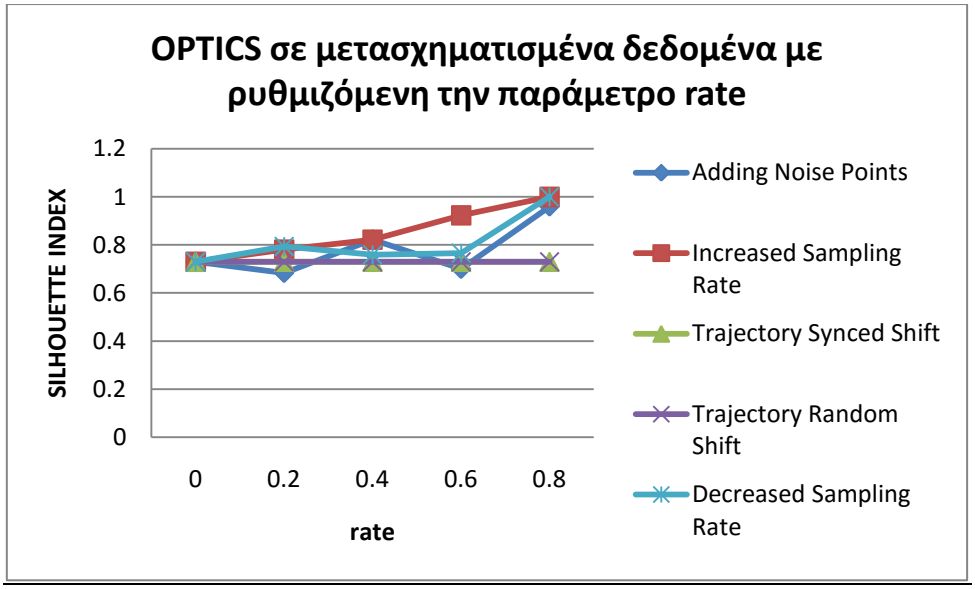

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
DUNN INDEX
1
0.8
0.6
0.4
0.2
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

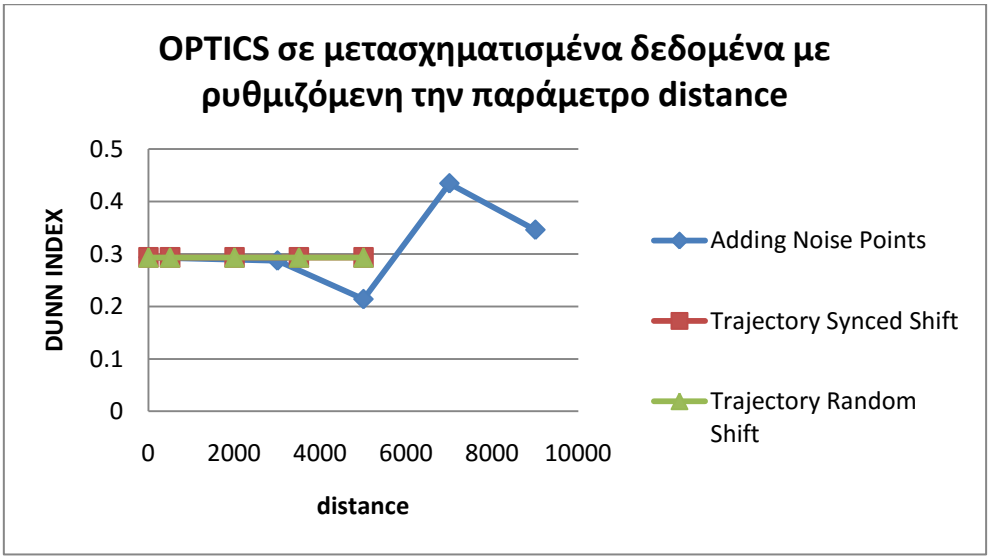


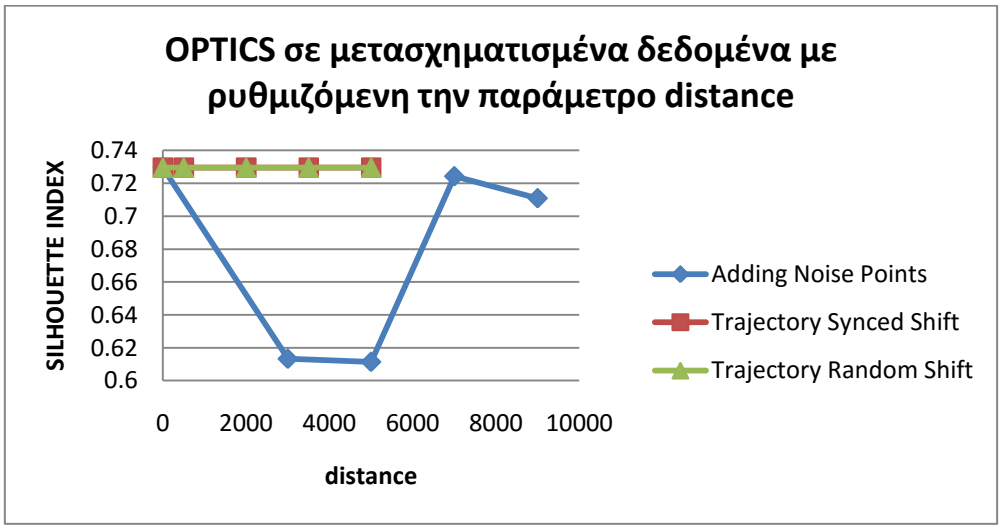
- Edit Distance on Real sequence

Ο αλγόριθμος Optics με την χρήση της απόστασης EDR φαίνεται να είναι ευαίσθητος στον μετασχηματισμό της προσθήκης θορύβου και στον μετασχηματισμο της αυξησης του ρυθμού δειγματολειψίας. Αποδεκτή συμπεριφορά φαίνεται να παρουσιάζει στον μετασχηματισμό της μείωσης του ρυθμού δειγματοληψίας, ενώ δείχνει να είναι ανθεκτικός στην τυχαία και στη συγχρονισμένη μετατόπιση.



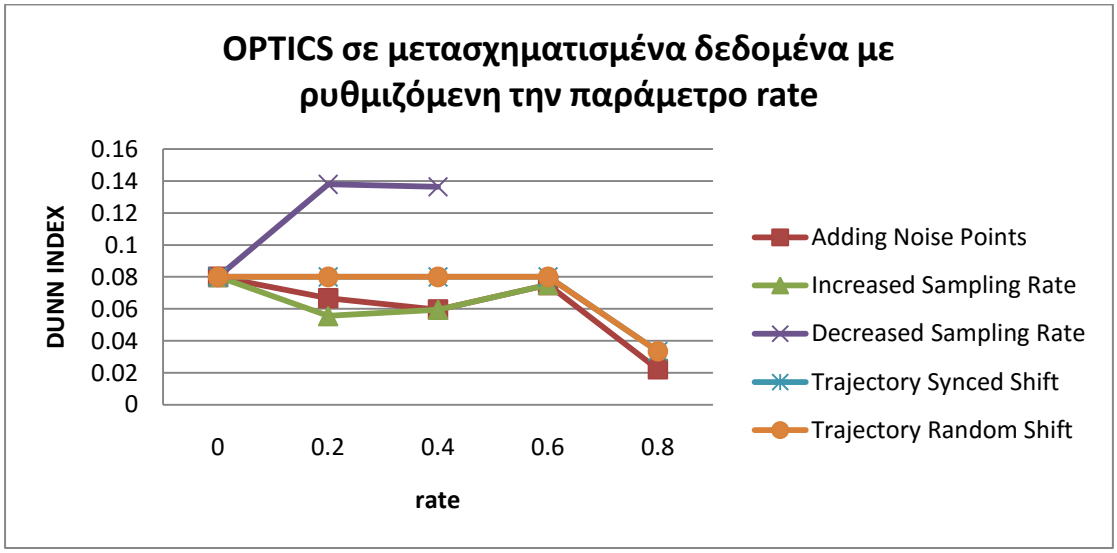

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUETTE INDEX
1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift
Decreased Sampling Rate

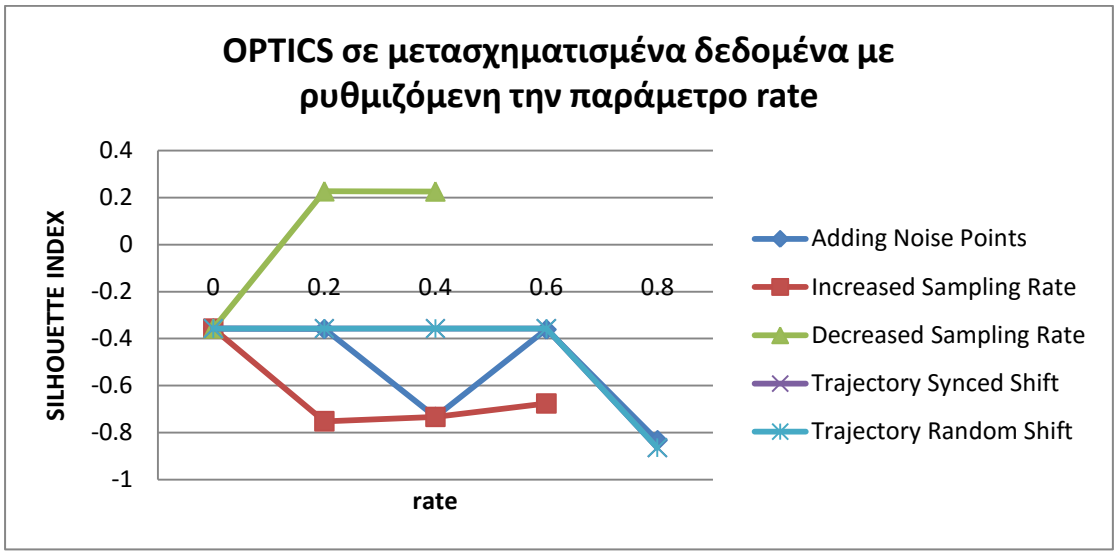

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
DUNN INDEX
0.5
0.4
0.3
0.2
0.1
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

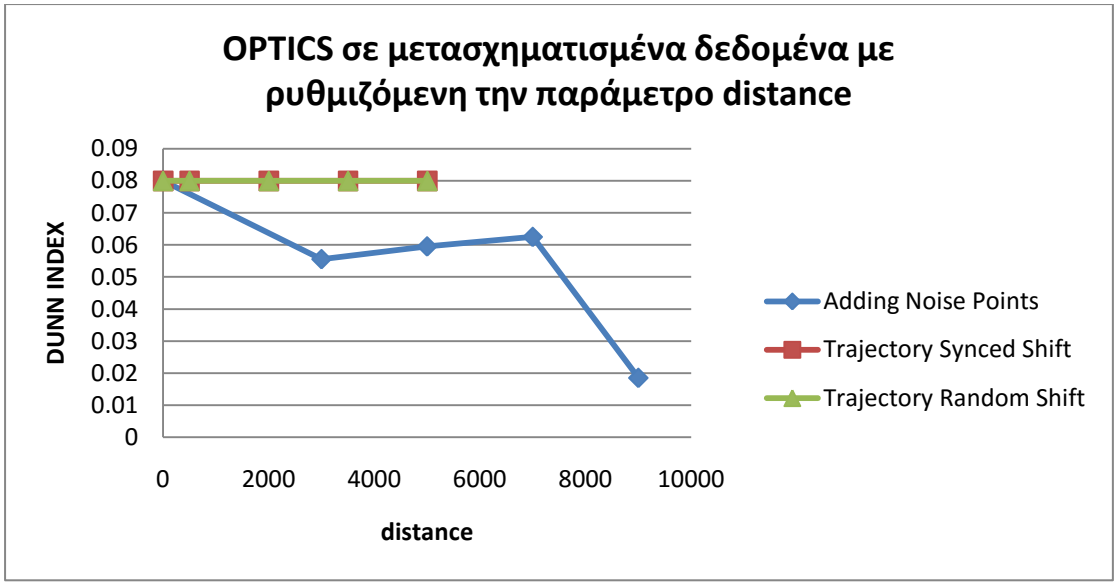

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
SILHOUETTE INDEX
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

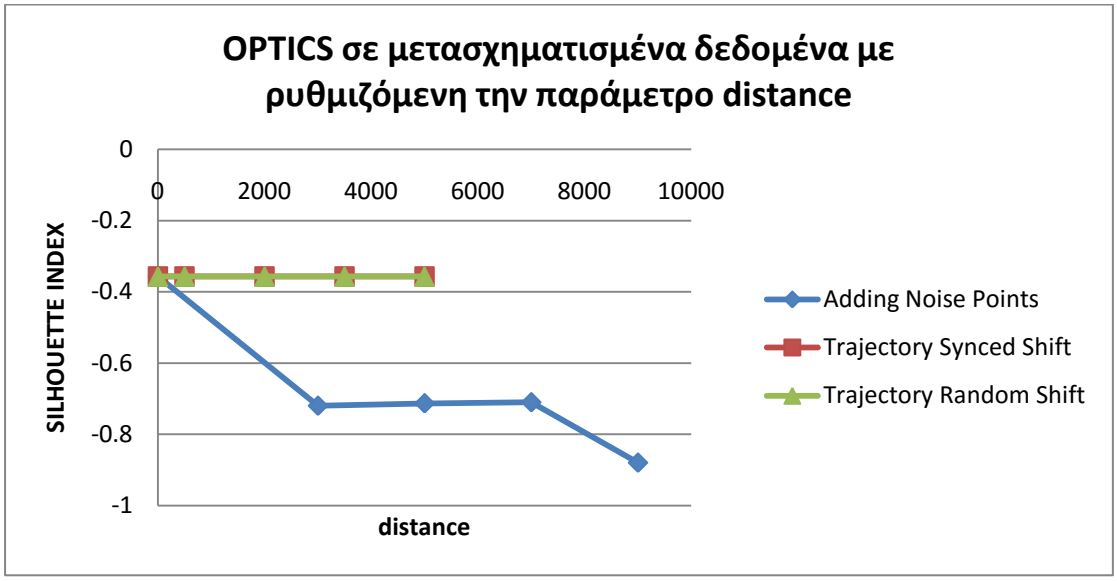
- Longest Common Subsequence

Ο αλγόριθμος Optics με τη χρήση της απόστασης LCSS φαίνεται να είναι ευαίσθητος στο μετασχηματισμό της προσθήκης θορύβου. Στο μετασχηματισμό της αύξησης του ρυθμού δειγματοληψίας φαίνεται να έχει αποδεκτή συμπεριφορά, ενώ αντίθετα στη μείωση του ρυθμού δειγματοληψίας παρουσιάζεται ευαίσθητος. Αποδεκτή συμπεριφορά φαίνεται να παρουσιάζει στους μετασχηματισμούς της τυχαίας και της συγχρονισμένης μετατόπισης.





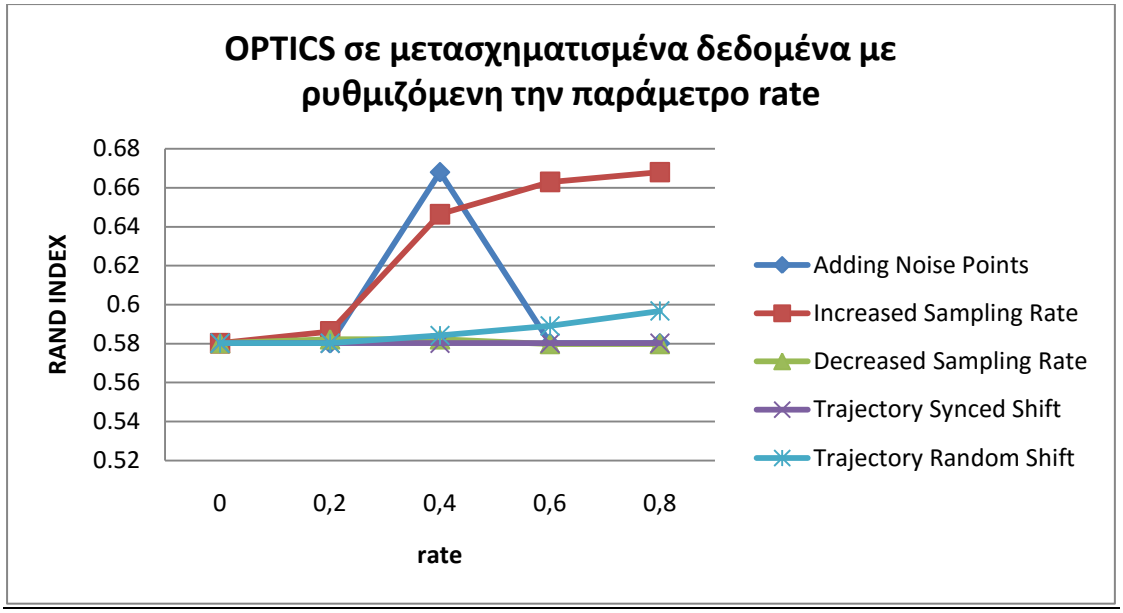


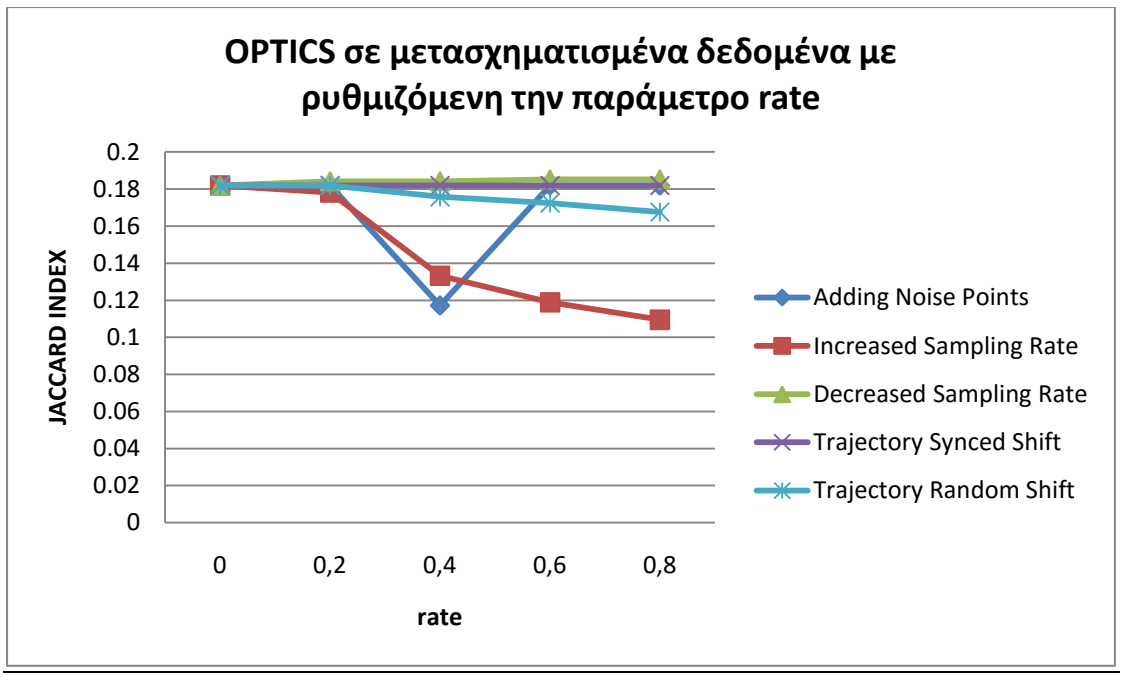


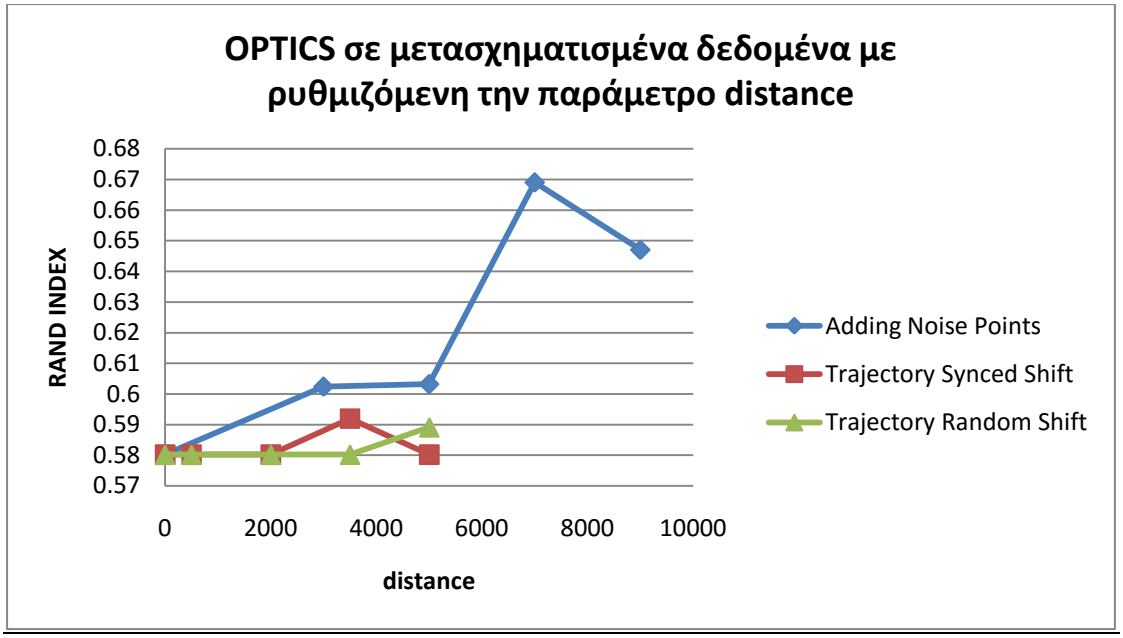
6.3.2 Εξωτερική αξιολόγηση ομαδοποίησης Optics

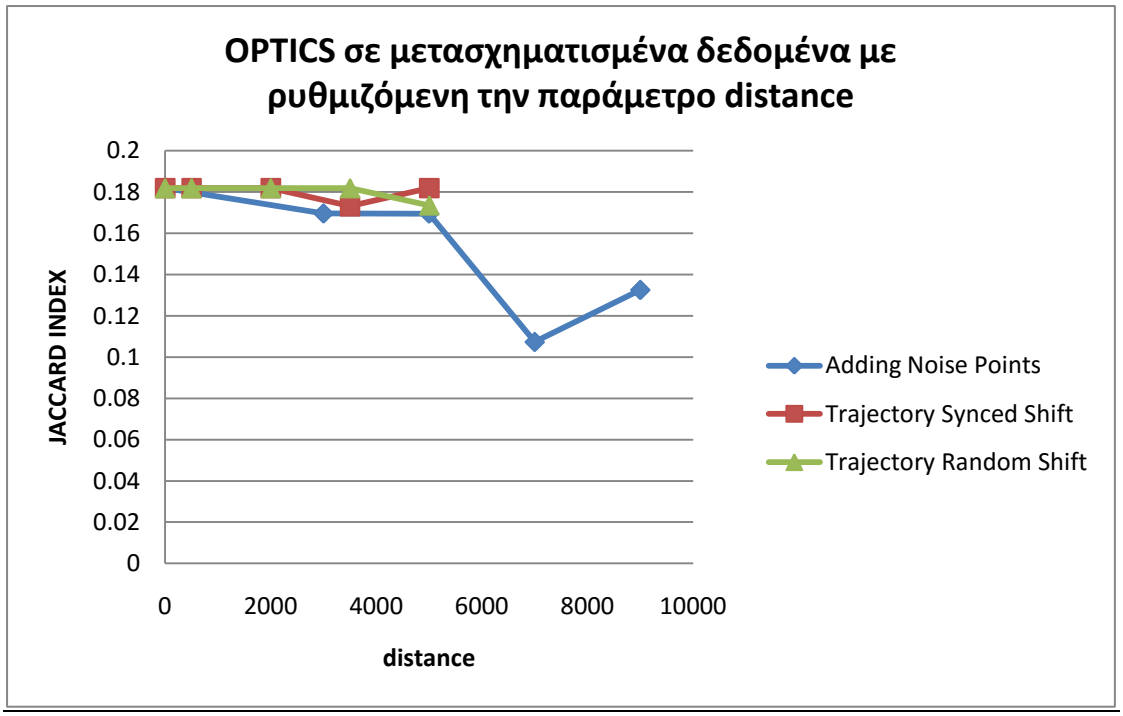
- Ευκλείδεια απόσταση

Με χρήση της ευκλείδειας απόστασης ο αλγόριθμος Optics δείχνει να είναι ευαίσθητος στην προσθήκη θορύβου. Ακόμη ευαίσθητος δείχνει να είναι και στον μετασχηματισμό της αύξησης του ρυθμού δειγματοληψίας. Ενώ αντίθετα στην μείωση του ρυθμού δειγματοληψίας ο αλγόριθμος Optics δείχνει να είναι ανθεκτικός. Τέλος, παρουσιάζεται να έχει αποδεκτή συμπεριφορά στον μετασχηματισμό της τυχαίας και της συγχρονισμένης μετατόπισης.


OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
RAND INDEX
0.68
0.66
0.64
0.62
0.6
0.58
0.56
0.54
0.52
0
0,2
0,4
0,6
0,8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

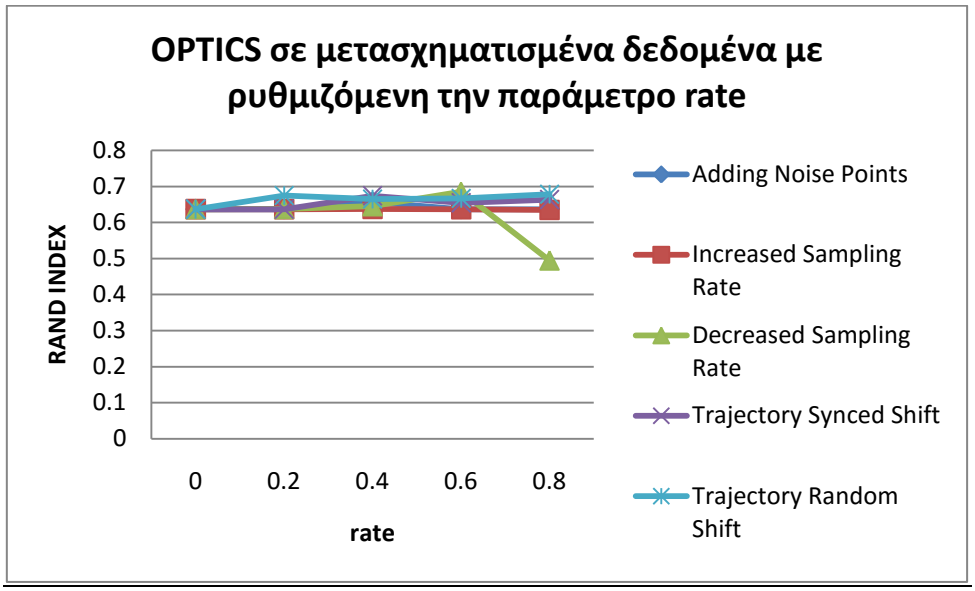

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0,2
0,4
0,6
0,8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

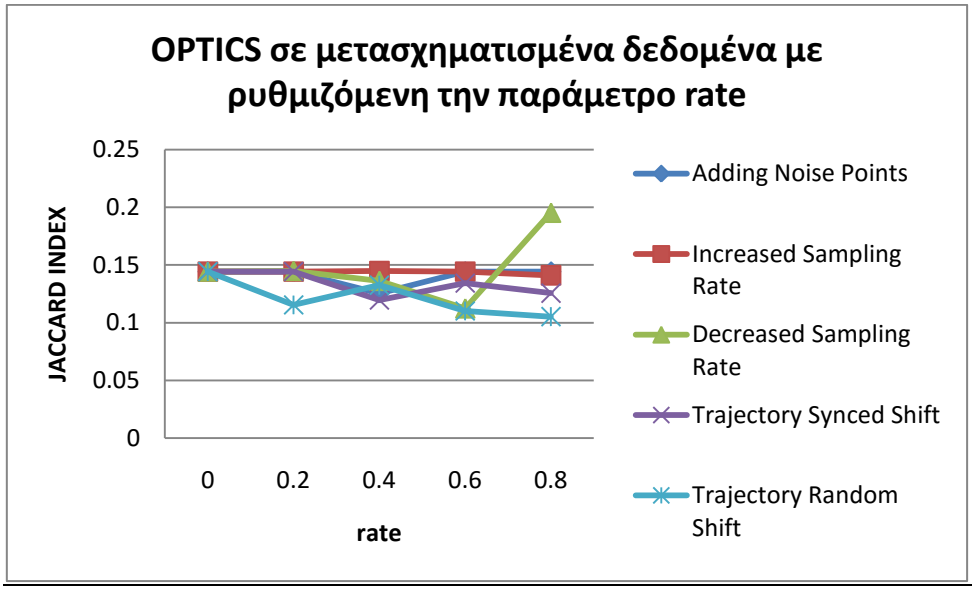


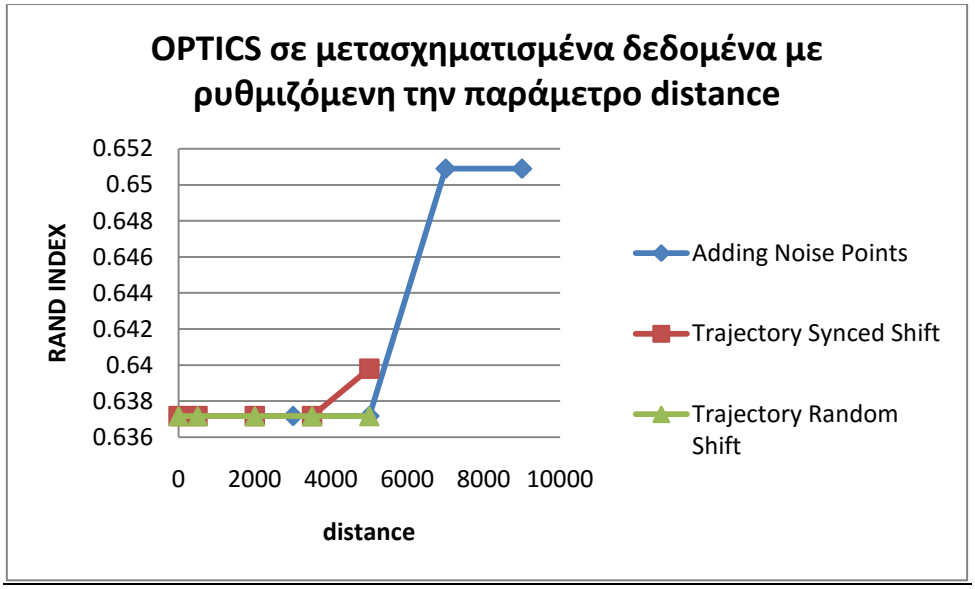


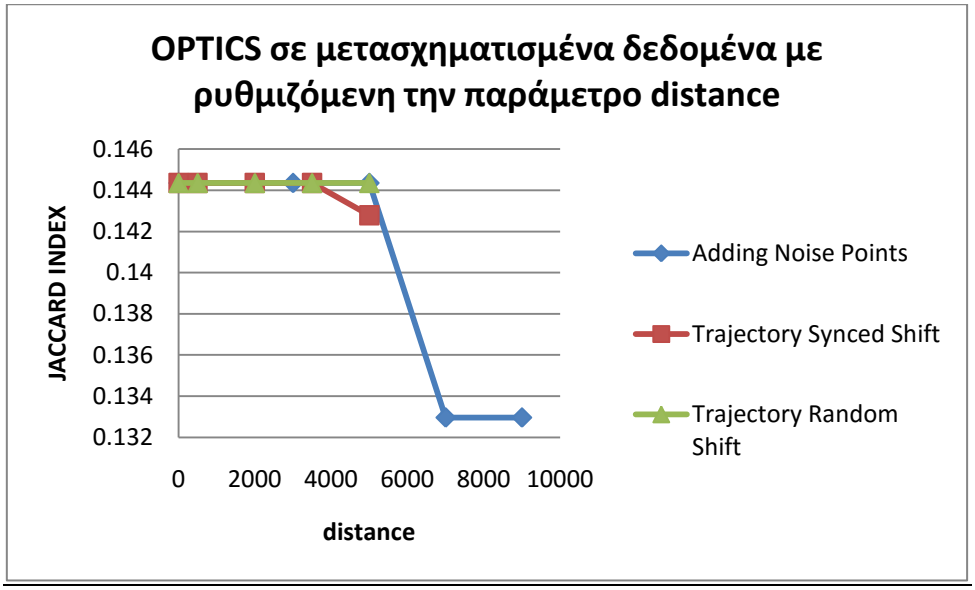
- Ευκλείδεια STARTEND

Με χρήση της απόστασης Euclideanstartend, ο αλγόριθμος Optics δείχνει να είναι ευαίσθητος στην προσθήκη θορύβου και στην μείωση του ρυθμού δειγματοληψίας. Στην αύξηση του ρυθμού δειγματοληψίας ο αλγόριθμος Optics φαίνεται να αντιδρά ικανοποιητικά αφού δείχνει να είναι ανθεκτικός σε αυτόν τον μετασχηματισμό. Επιπλέον, δείχνει να είναι ανθεκτικός στην τυχαία μετατόπιση και έχει αποδεκτή συμπεριφορά για τον μετασχηματισμό της συγχρονισμένης μετατόπισης.


OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
RAND INDEX
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

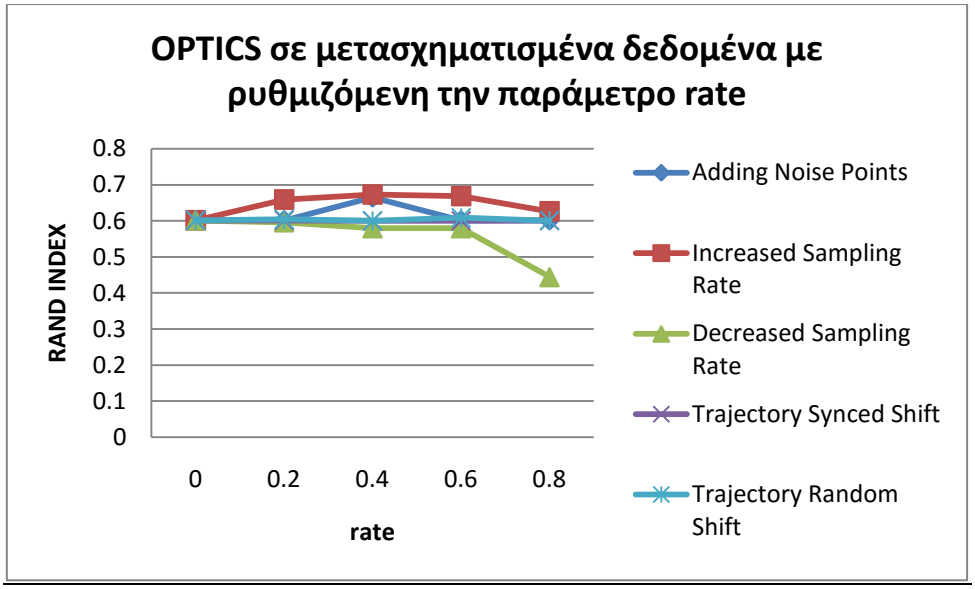

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

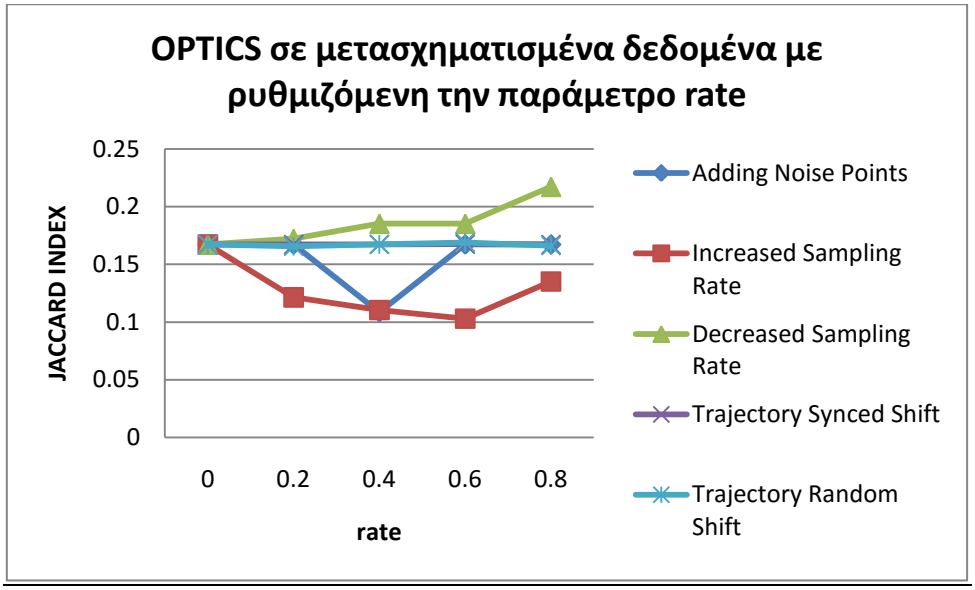


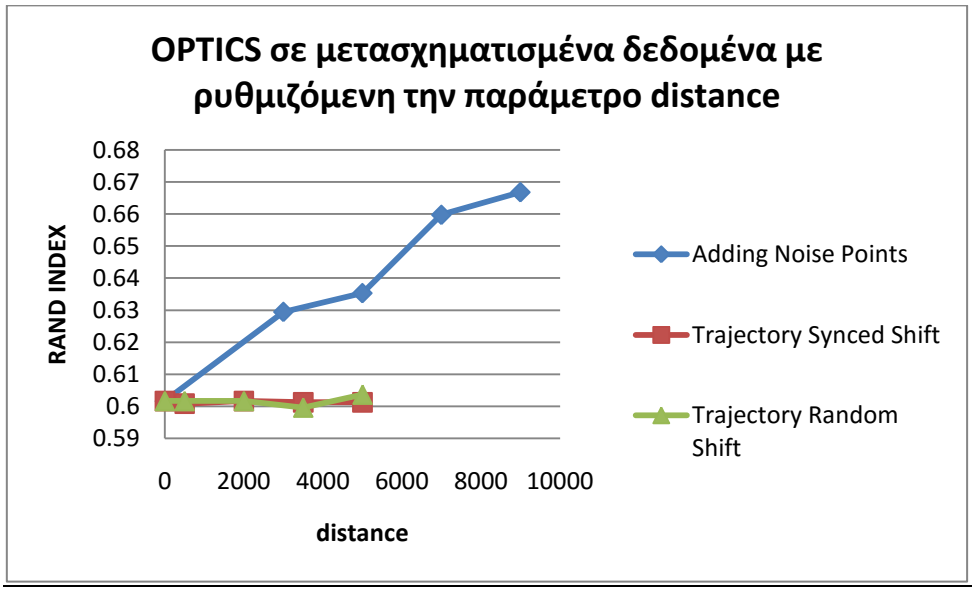


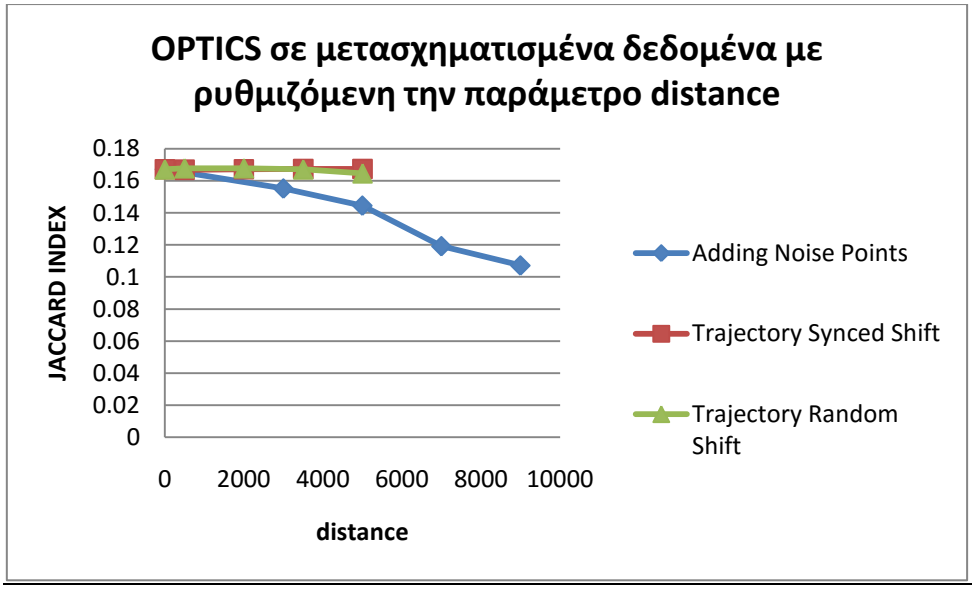
- Manhattan απόσταση

Με χρήση της απόστασης Manhattan, ο αλγόριθμος Optics δείχνει να είναι ευαίσθητος στην προσθήκη θορύβου. Ο αλγόριθμος Optics δείχνει να έχει ευαισθητη συμπεριφορά στην αύξηση του ρυθμού δειγματοληψίας. Επιπλέον δείχνει να επηρεάζεται από τη μείωση του ρυθμού δειγματοληψίας. Τέλος, δείχνει να είναι ανθεκτικός σε όλες τις περιπτώσεις των μετασχηματισμών της μετατόπισης.


OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
RAND INDEX
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

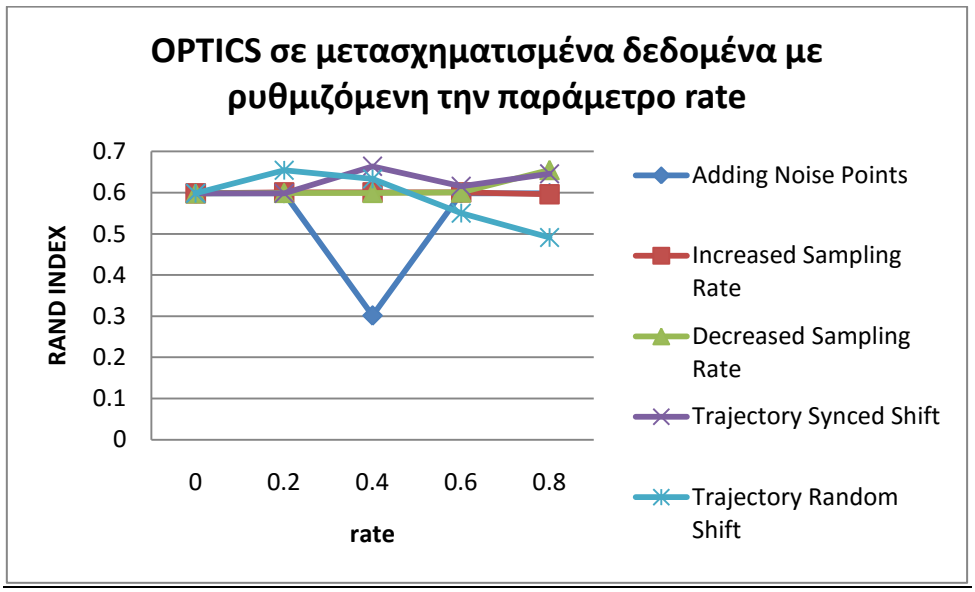

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

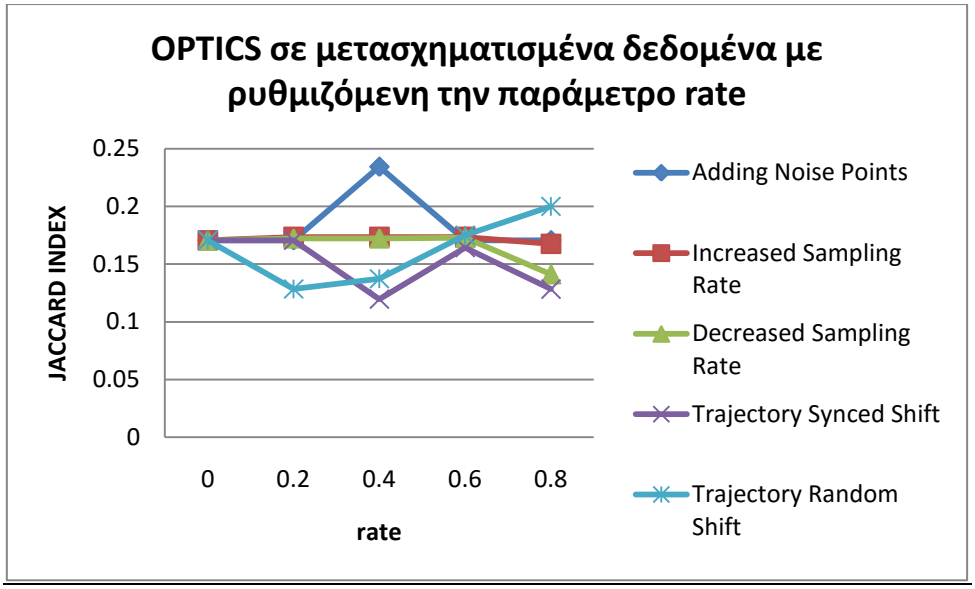

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
RAND INDEX
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0.61
0.6
0.59
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

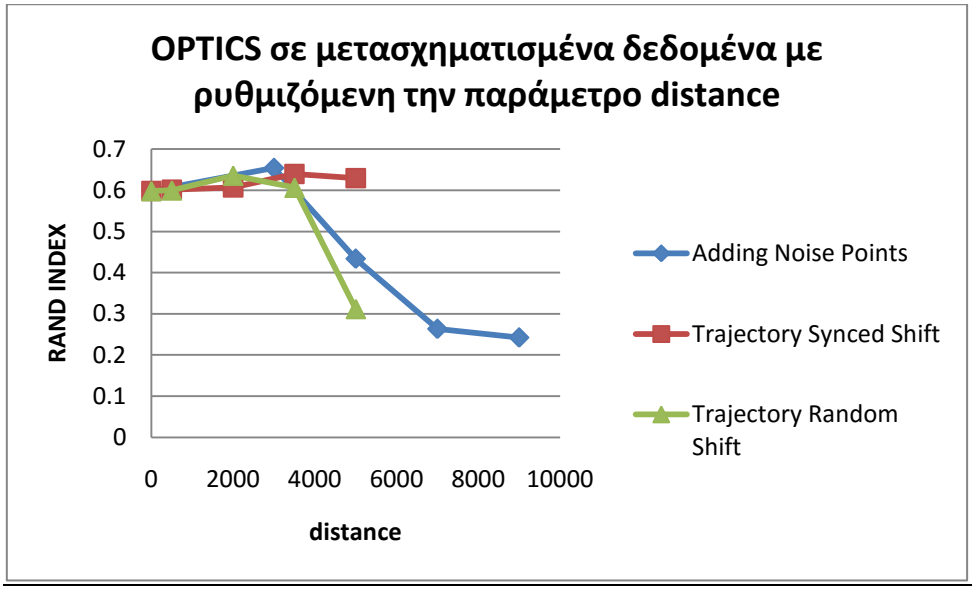


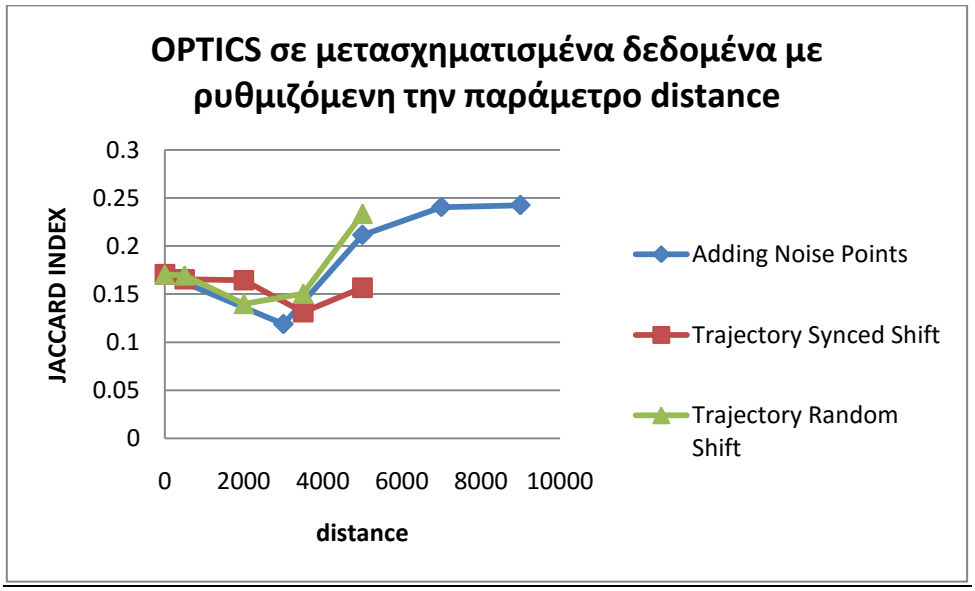
- Chebyshev απόσταση

Με χρήση της απόστασης Chebyshev, ο αλγόριθμος Optics δείχνει να είναι ευαίσθητος στην προσθήκη θορύβου και στην τυχαία μετατόπιση. Στην αύξηση του ρυθμού δειγματοληψίας ο αλγόριθμος Optics φαίνεται να αντιδρά ικανοποιητικά αφού δείχνει να είναι ανθεκτικός σε αυτόν τον μετασχηματισμό. Επιπλέον, στον μετασχηματισμό της μείωσης του ρυθμού δειγματοληψίας δείχνει να έχει αποδεκτή συμπεριφορά. Ομοίως, ο αλγόριθμος δείχνει να έχει και στην συγχρονισμένη μετατόπιση αποδεκτή συμπεριφορά.



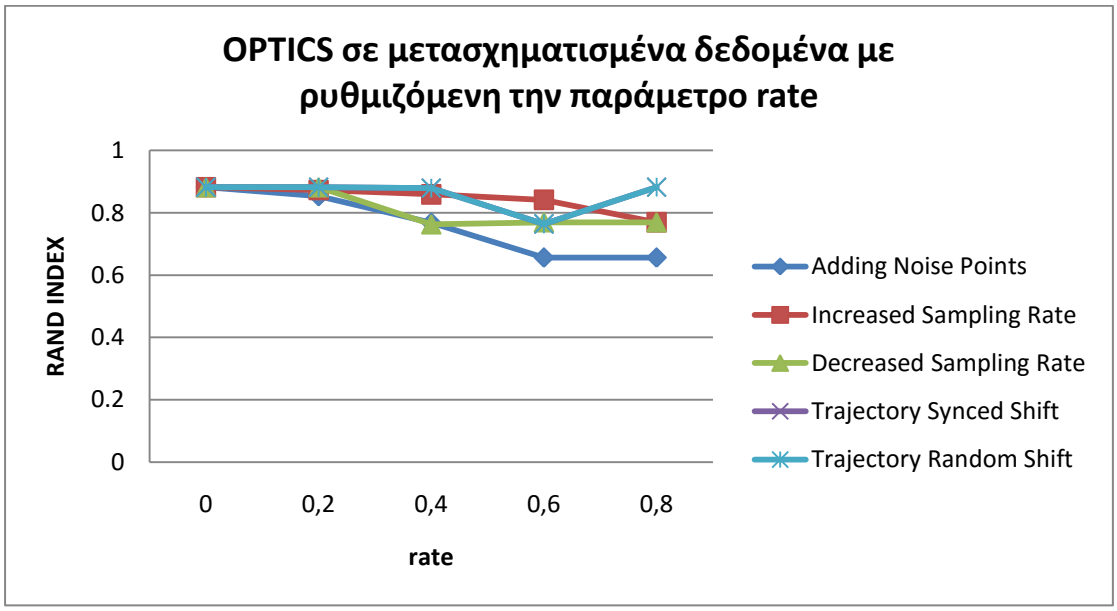

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

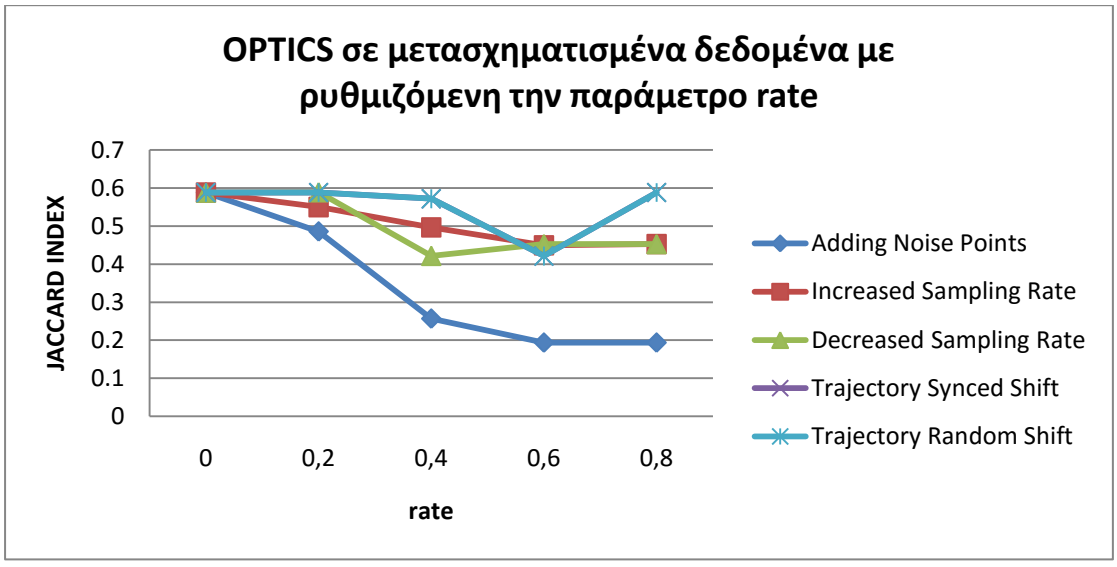

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
RAND INDEX
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

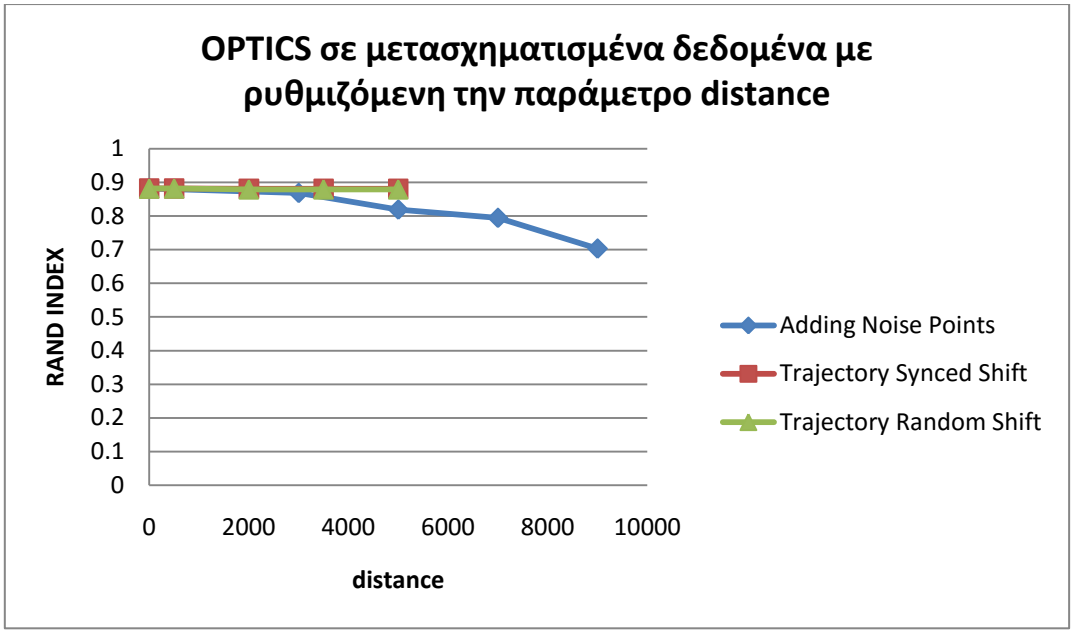

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
JACCARD INDEX
0.3
0.25
0.2
0.15
0.1
0.05
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

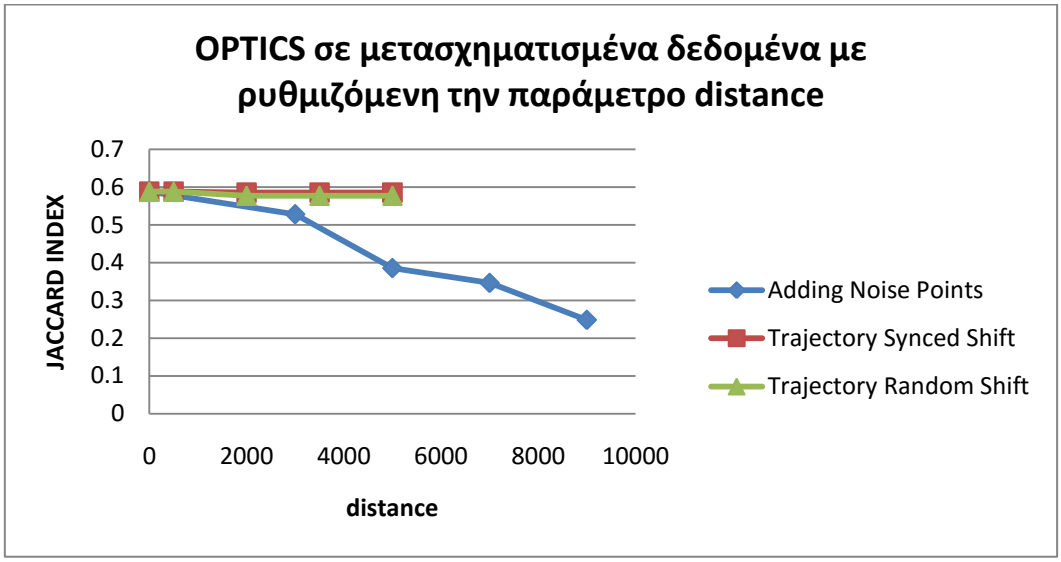
- Dynamic Time Warping

Ο αλγόριθμος Optics με την χρήση της απόστασης DTW φαίνεται να είναι ευαίσθητος στους μετασχηματισμούς της προσθήκης θορύβου, της μείωσης και της αύξησης του ρυθμού δειγματοληψίας. Τέλος, δείχνει να συμπεριφέρεται αποδεκτά για τους δυο μετασχηματισμούς της μετατόπισης.



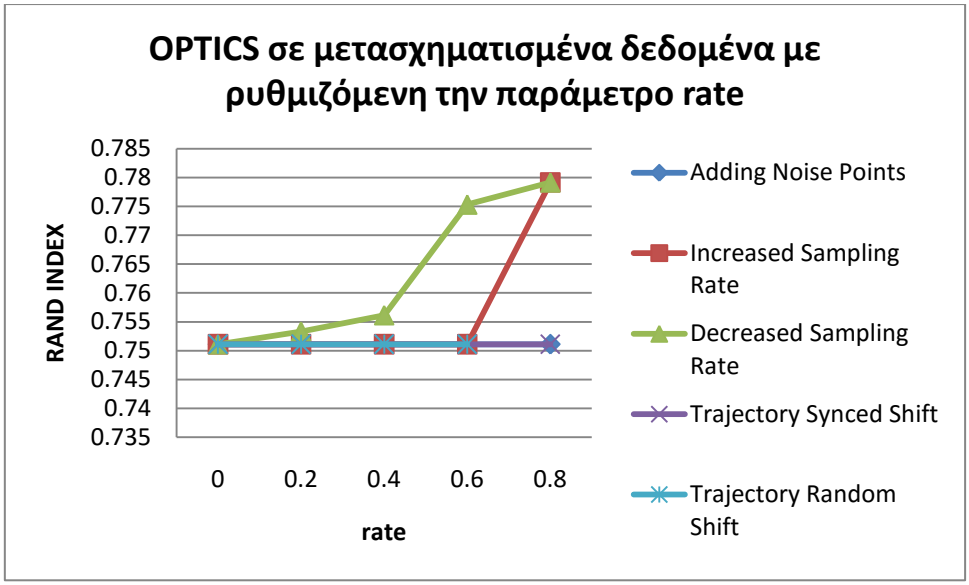


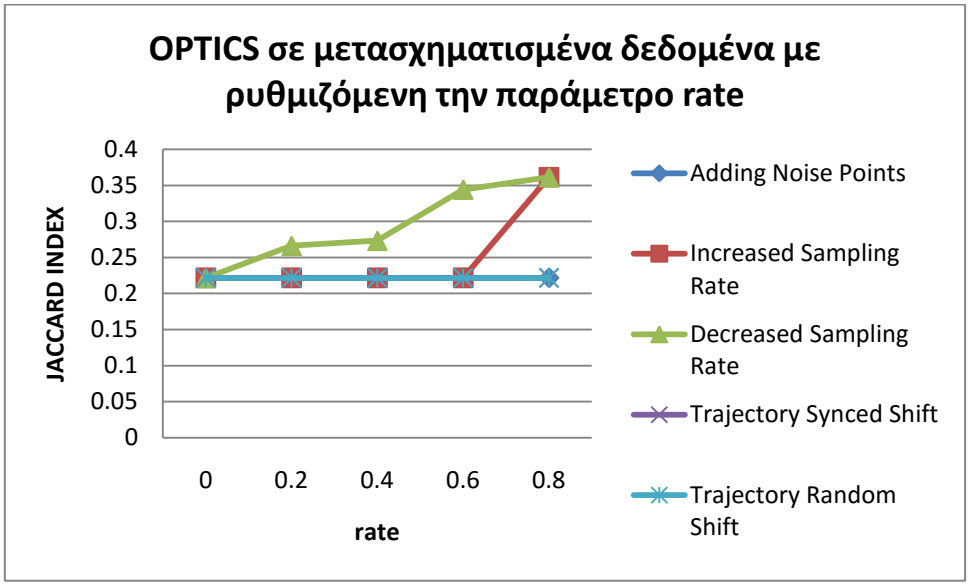


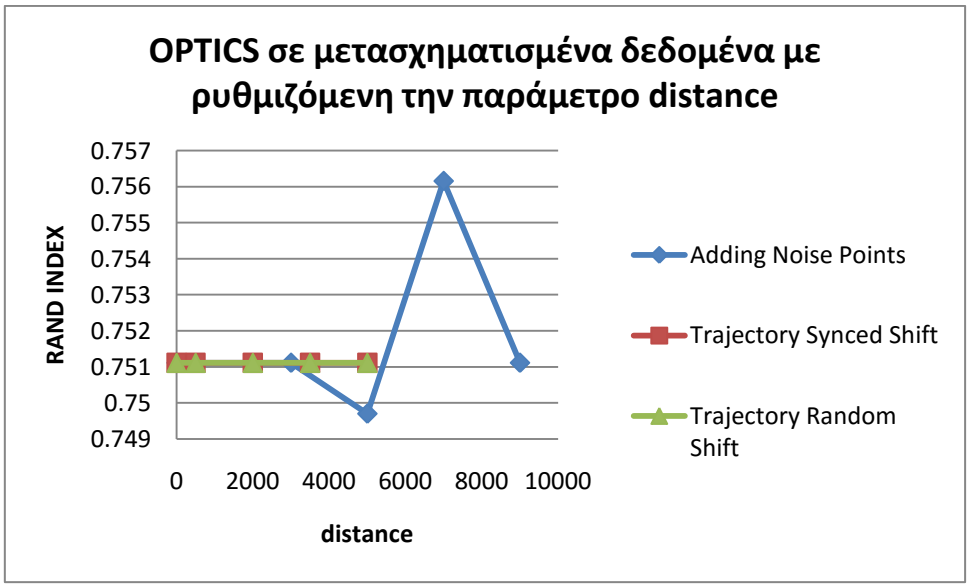


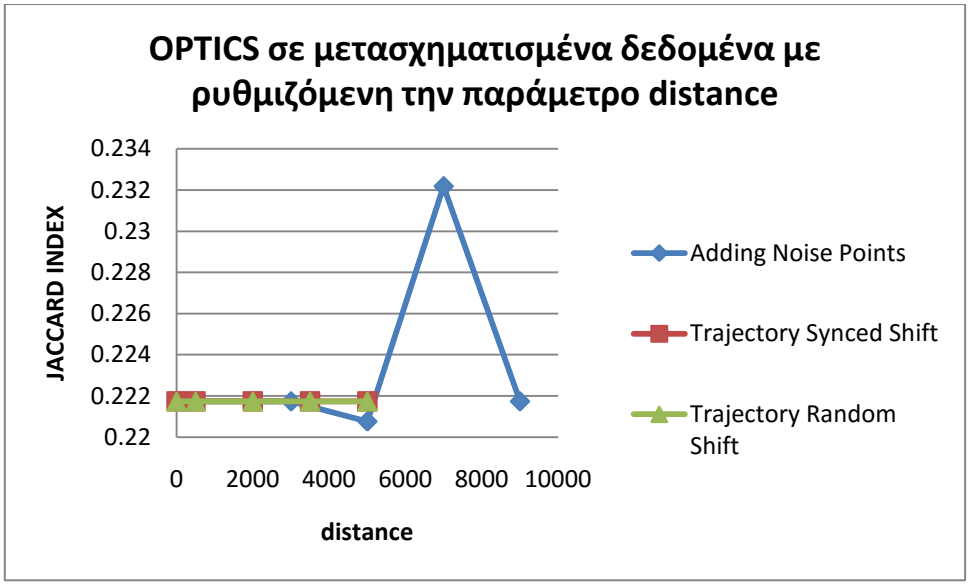
- Edit Distance on Real sequence

Ο αλγόριθμος Optics με τη χρήση της απόστασης EDR φαίνεται να είναι ευαίσθητος στον μετασχηματισμό της προσθήκης θορύβου. Ακόμη ευαίσθητος δείχνει και στους μετασχηματισμούς της αύξησης και της μείωσης του ρυθμού δειγματολειψίας. Αντίθετα, δείχνει να είναι ανθεκτικός στην τυχαία και στη συγχρονισμένη μετατόπιση.


OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
RAND INDEX
0.785
0.78
0.775
0.77
0.765
0.76
0.755
0.75
0.745
0.74
0.735
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

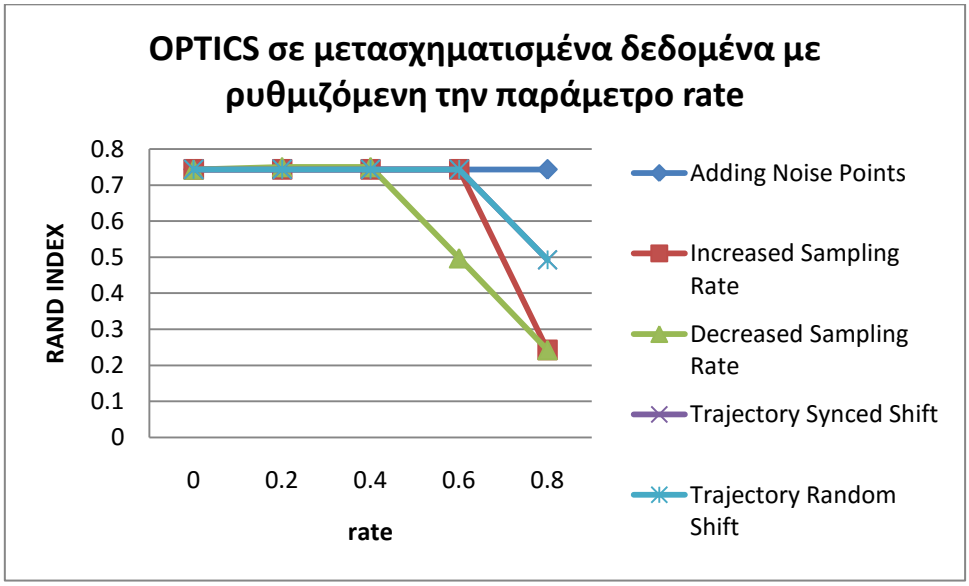

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

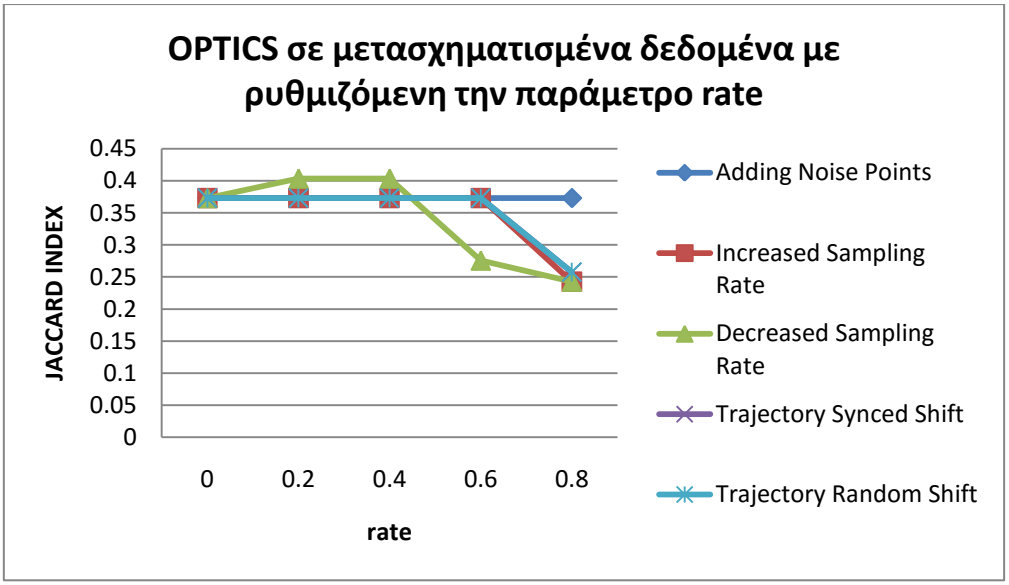

OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
RAND INDEX
0.757
0.756
0.755
0.754
0.753
0.752
0.751
0.75
0.749
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

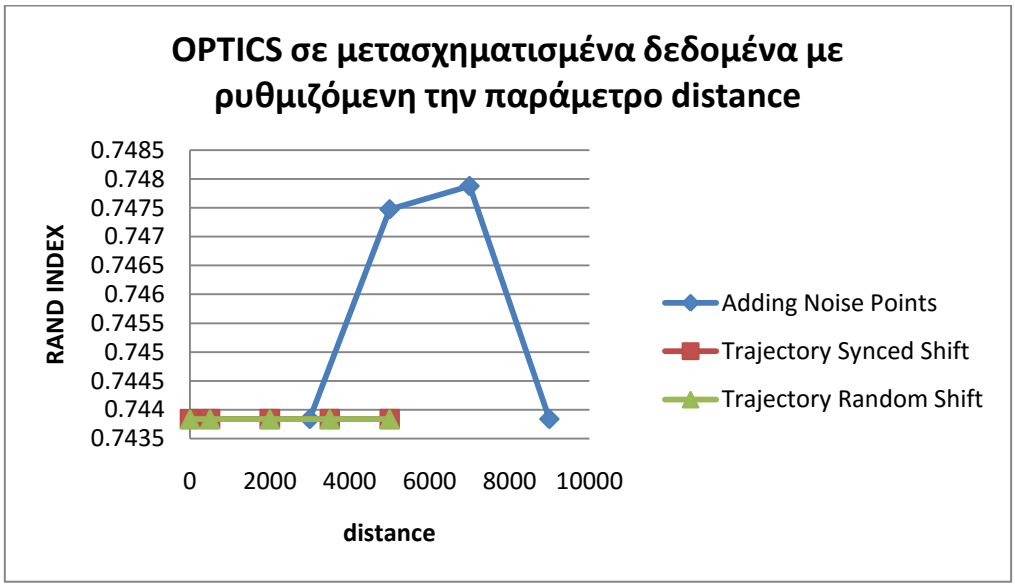


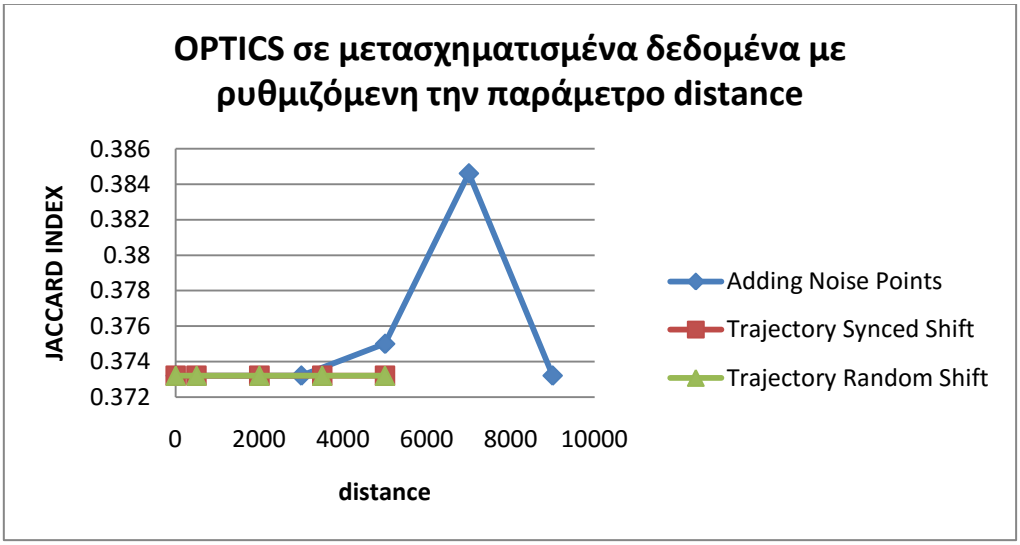
- Longest Common Subsequence

Ο αλγόριθμος Optics με τη χρήση της απόστασης LCSS φαίνεται να έχει αποδεκτή συμπεριφορά στο μετασχηματισμό της προσθήκης θορύβου και της αύξησης του ρυθμού δειγματοληψίας. Δείχνει ευαίσθητος στη μείωση του ρυθμού δειγματοληψίας, ενώ δείχνει να είναι ανθεκτικός στην τυχαία και στη συγχρονισμένη μετατόπιση.




OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift


OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
RAND INDEX
0.7485
0.748
0.7475
0.747
0.7465
0.746
0.7455
0.745
0.7445
0.744
0.7435
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift


OPTICS σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
JACCARD INDEX
0.386
0.384
0.382
0.38
0.378
0.376
0.374
0.372
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

6.4 Αξιολόγηση εγκυρότητας της ιεραρχικής ομαδοποίησης

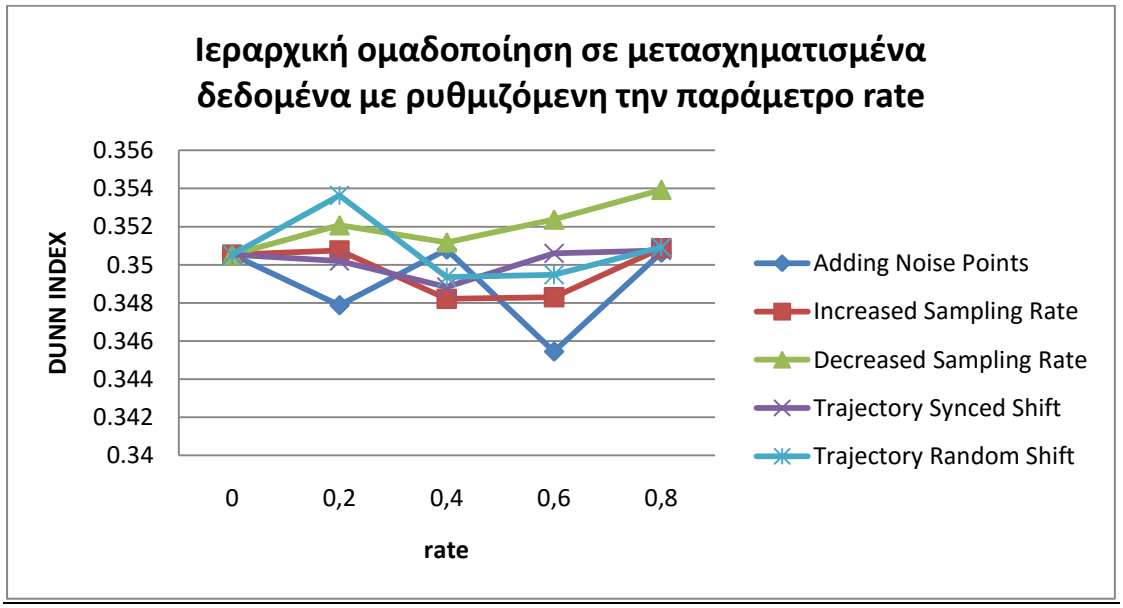
Στην ενότητα αυτή θα μελετήσουμε τη συμπεριφορά της ιεραρχικής ομαδοποίησης (μέθοδος Ward), κάνοντας χρήση εσωτερικών και εξωτερικών δεικτών αξιολόγησης της εγκυρότητας των συστάδων. Θεωρούμε κάθε φορά διαφορετικές συναρτήσεις απόστασης και διάφορα είδη μετασχηματισμών,

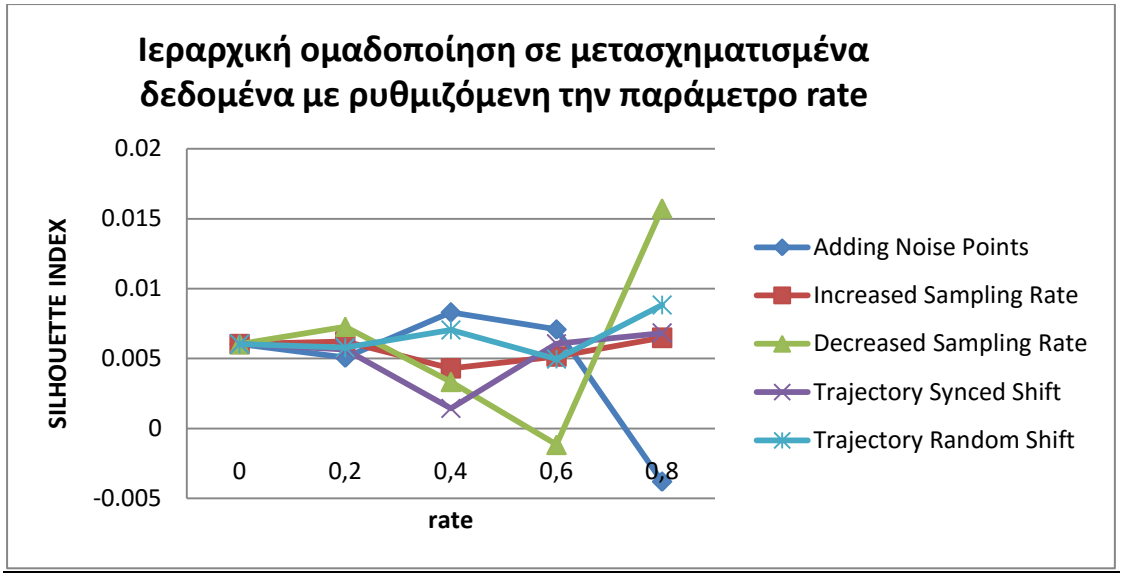
Για την υλοποίηση των μετασχηματισμών αρχικά σταθεροποιούμε την παράμετρο distance (=5.5km) και αλλάζουμε την τιμή της παραμέτρου rate από το 0.2 έως το 0.8, με βήμα=0.2. Έπειτα, σταθεροποιούμε την παράμετρο rate (=0.3) και αλλάζουμε την τιμή της παραμέτρου distance από 3 km έως 9 km, με βήμα=2 km για την προσθήκη θορύβου. Για τους μετασχηματισμούς της μετατόπισης σταθεροποιούμε την παράμετρο rate (=0.3) και αλλάζουμε την τιμή της παραμέτρου distance από το 0.5 km έως 5 km, με βήμα=1.5 km.

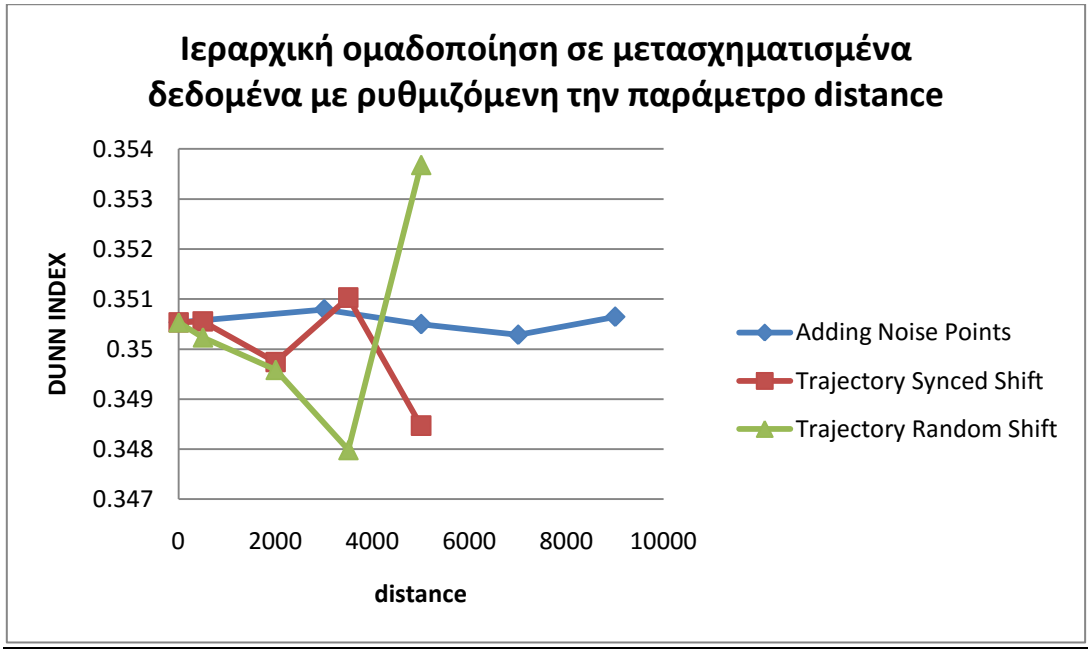
6.4.1 Εσωτερική αξιολόγηση ιεραρχικής ομαδοποίησης

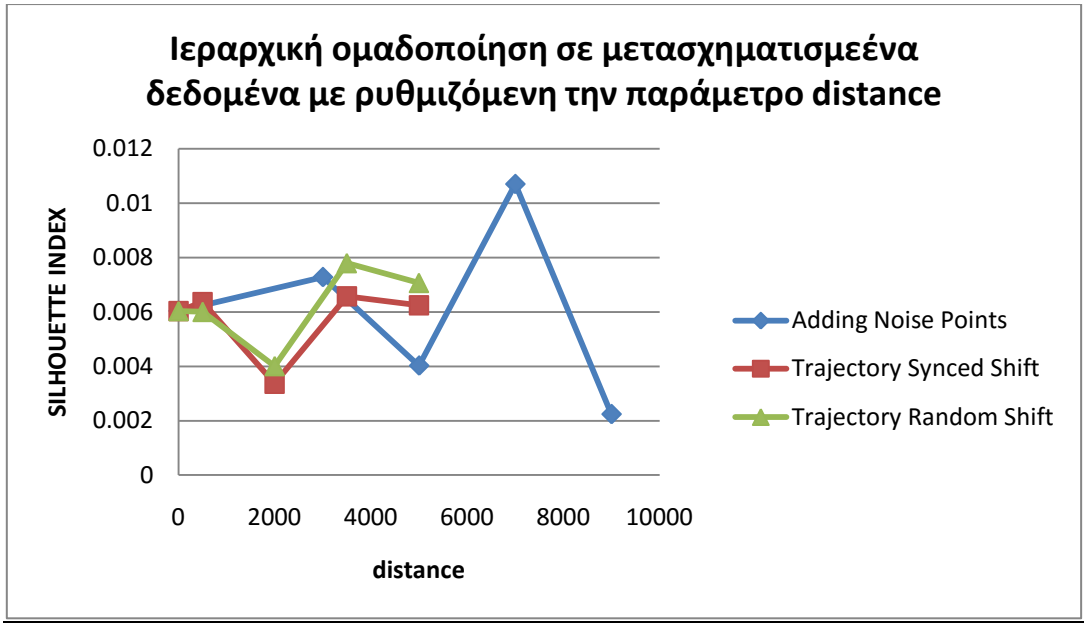
- Ευκλείδεια απόσταση

Με χρήση της ευκλείδειας απόστασης ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) δείχνει να είναι ευαίσθητος σε όλους τους μετασχηματισμούς.



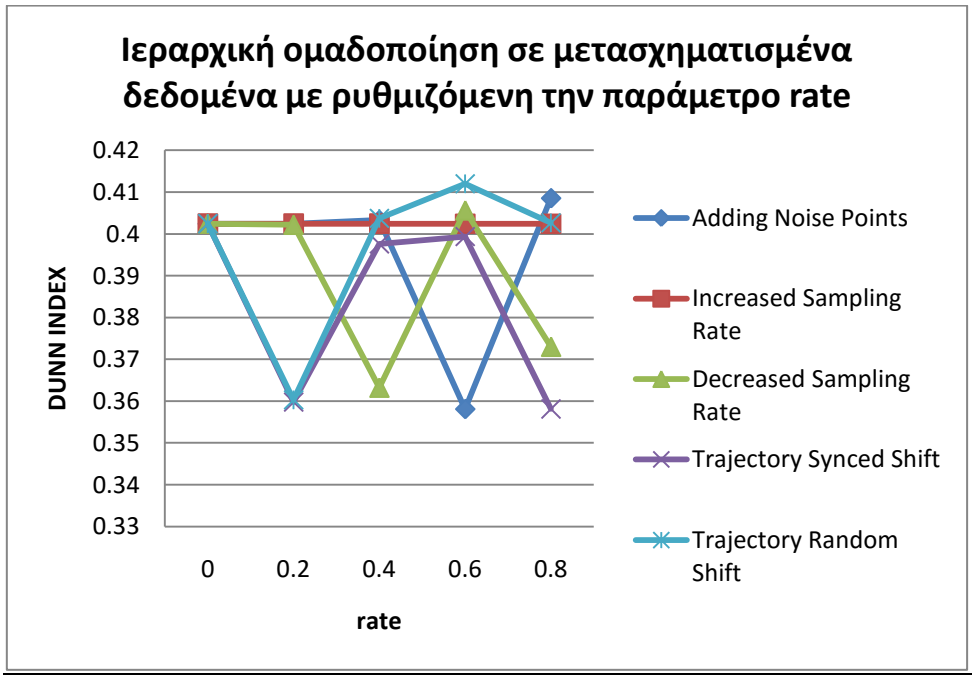

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUETTE INDEX
0.02
0.015
0.01
0.005
0
-0.005
0
0,2
0,4
0,6
0,8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

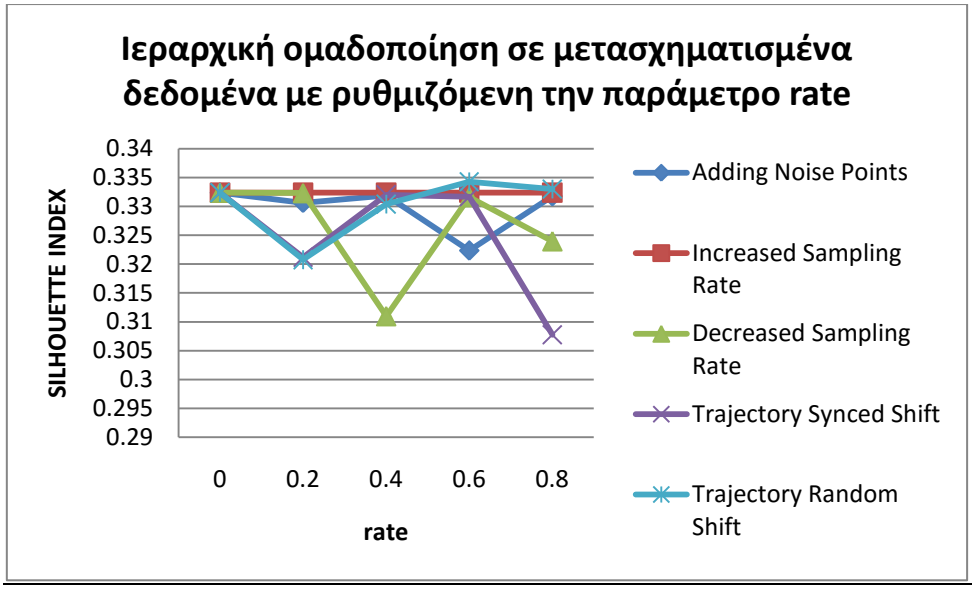

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
DUNN INDEX
0.354
0.353
0.352
0.351
0.35
0.349
0.348
0.347
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

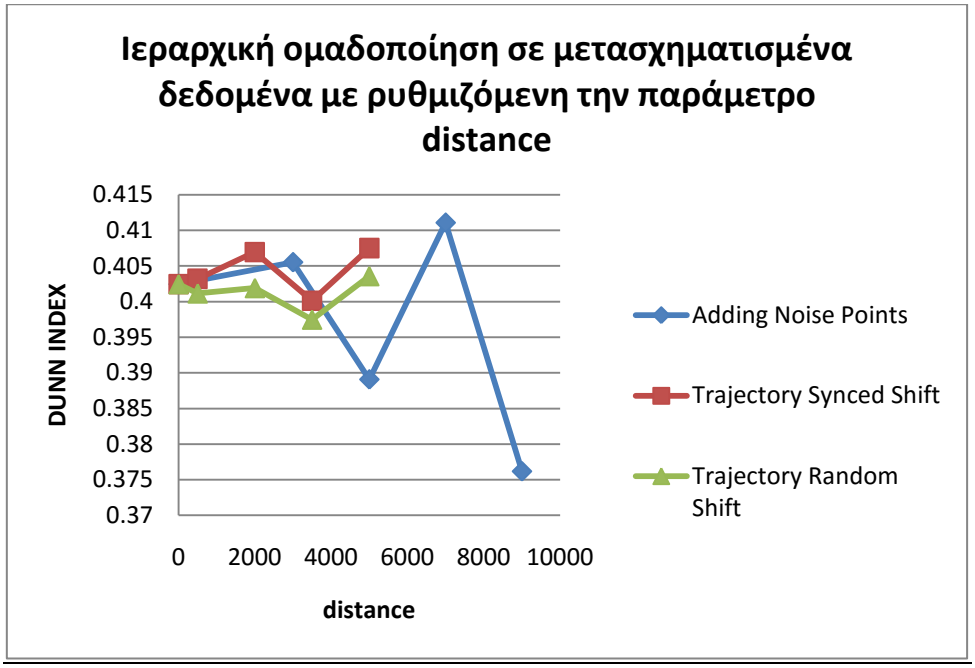


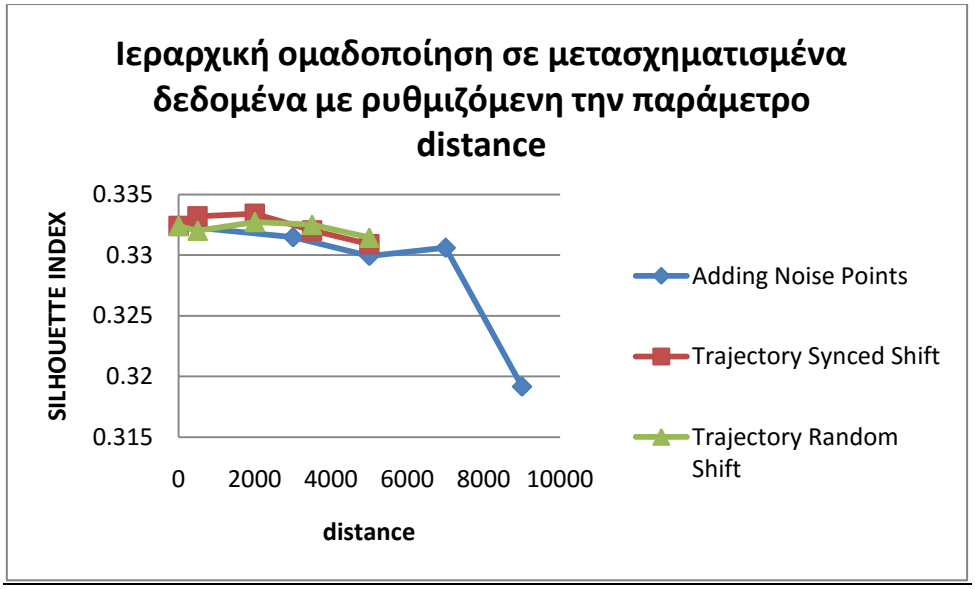
- Ευκλείδεια STARTEND

Με χρήση της απόστασης Euclideanstartend, ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) φαίνεται να είναι ευαίσθητος σε όλους τους μετασχηματισμούς εκτός από αυτόν της αύξησης του ρυθμού δειγματοληψίας στον οποίο παρουσιάζεται ανθεκτικός.



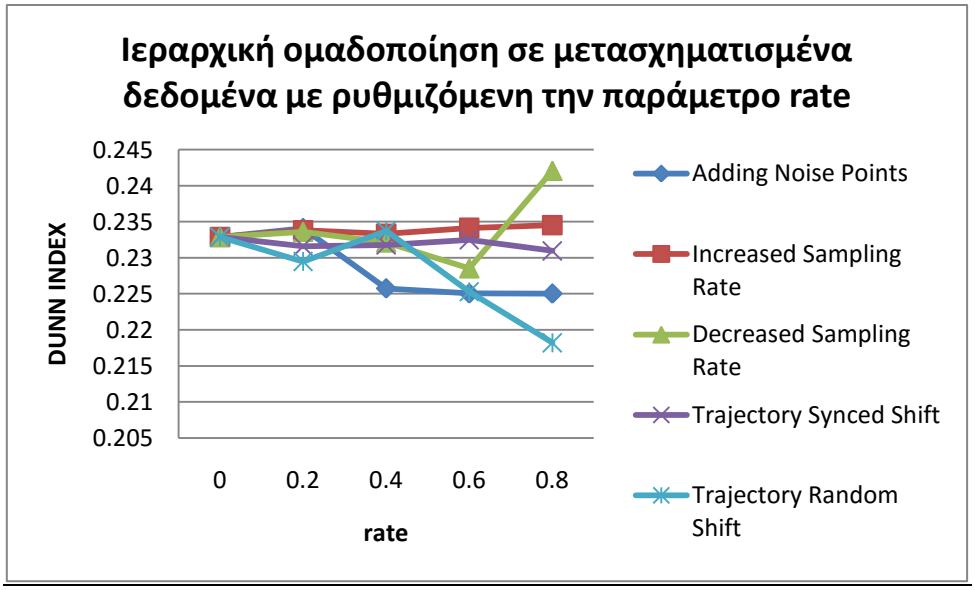

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUETTE INDEX
0.34
0.335
0.33
0.325
0.32
0.315
0.31
0.305
0.3
0.295
0.29
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

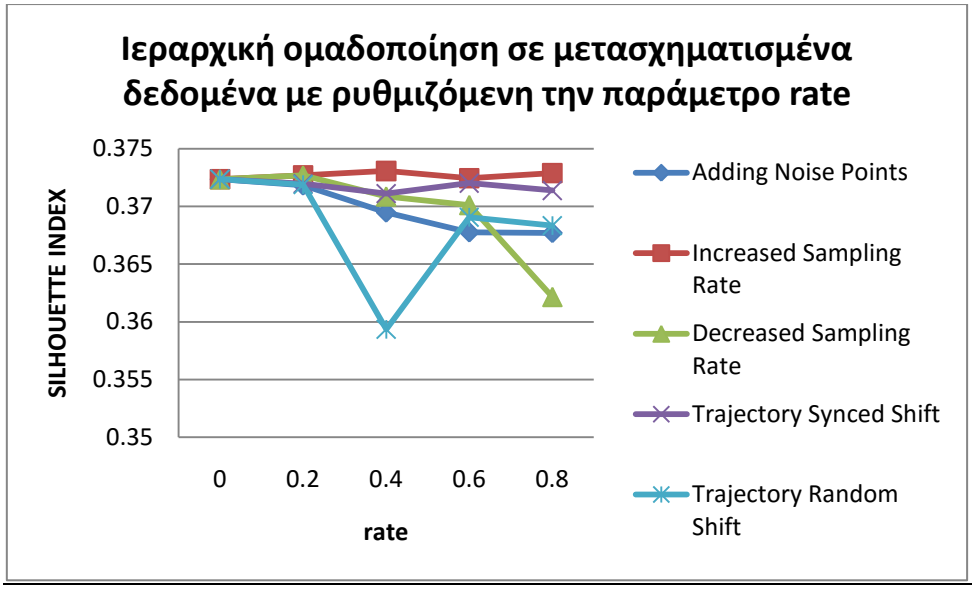

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
DUNN INDEX
0.415
0.41
0.405
0.4
0.395
0.39
0.385
0.38
0.375
0.37
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

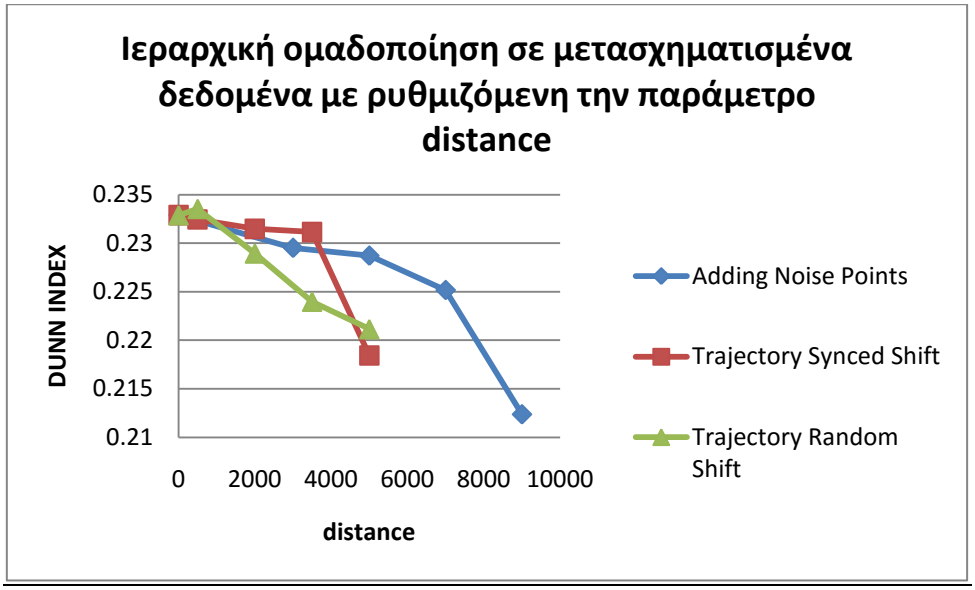


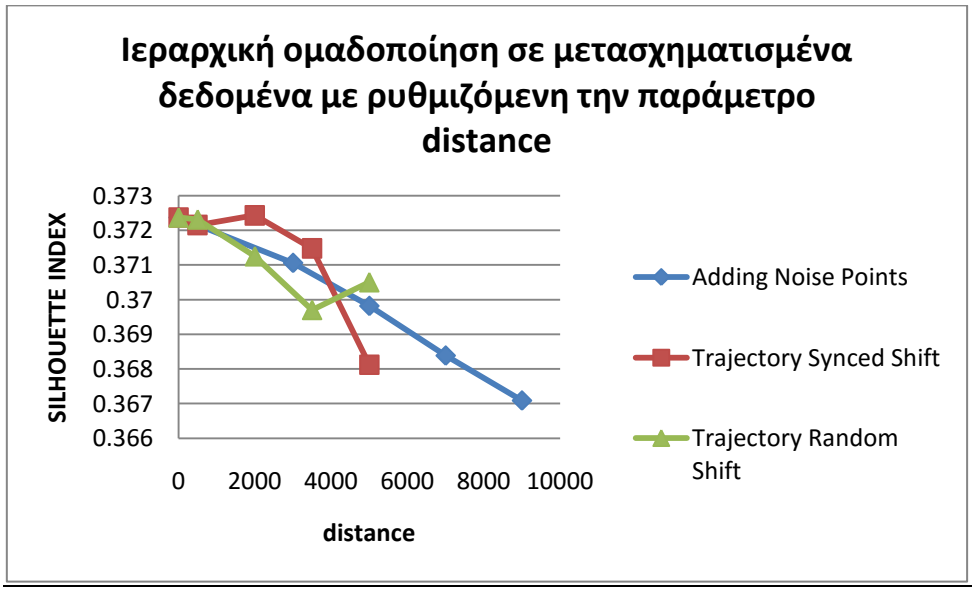
- Manhattan απόσταση

Με χρήση της απόστασης Manhattan, ο αλγόριθμος της Ιεραρχικής ομαδοποίησης παρουσιάζεται ευαίσθητος στην προσθήκη θορύβου, στην μείωση του ρυθμού δειγματοληψίας, στην τυχαία μετατόπιση και στον μετασχηματισμό της συγχρονισμένης μετατόπισης. Αντίθετα δείχνει να έχει αποδεκτή συμπεριφορά στην αύξηση του ρυθμού δειγματοληψίας.



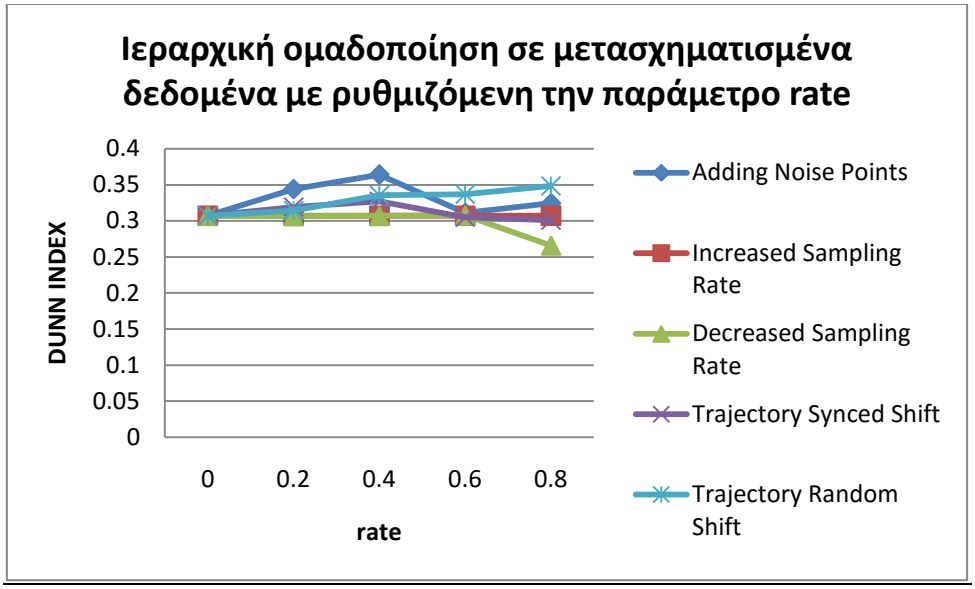

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUETTE INDEX
0.375
0.37
0.365
0.36
0.355
0.35
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

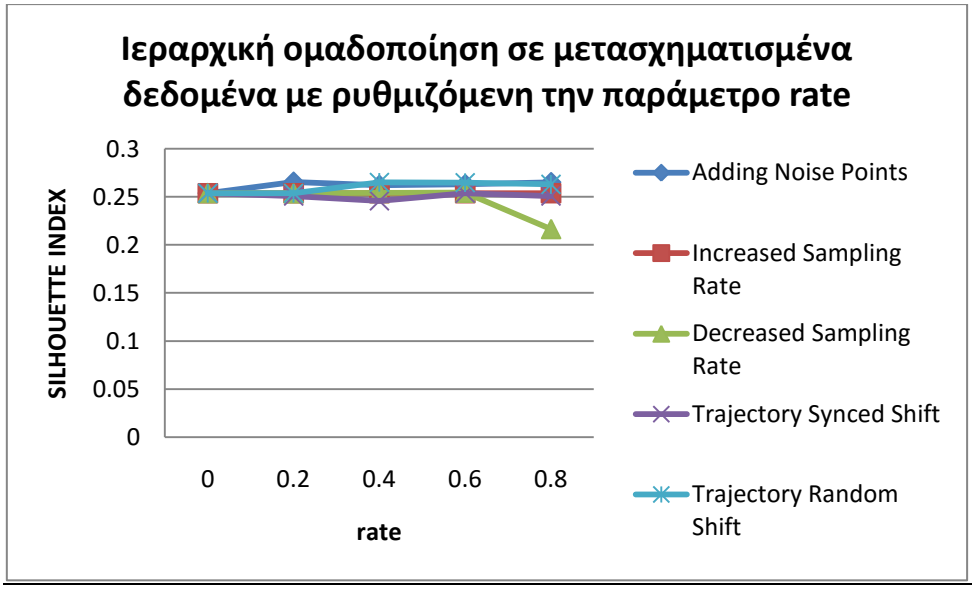

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
DUNN INDEX
0.235
0.23
0.225
0.22
0.215
0.21
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

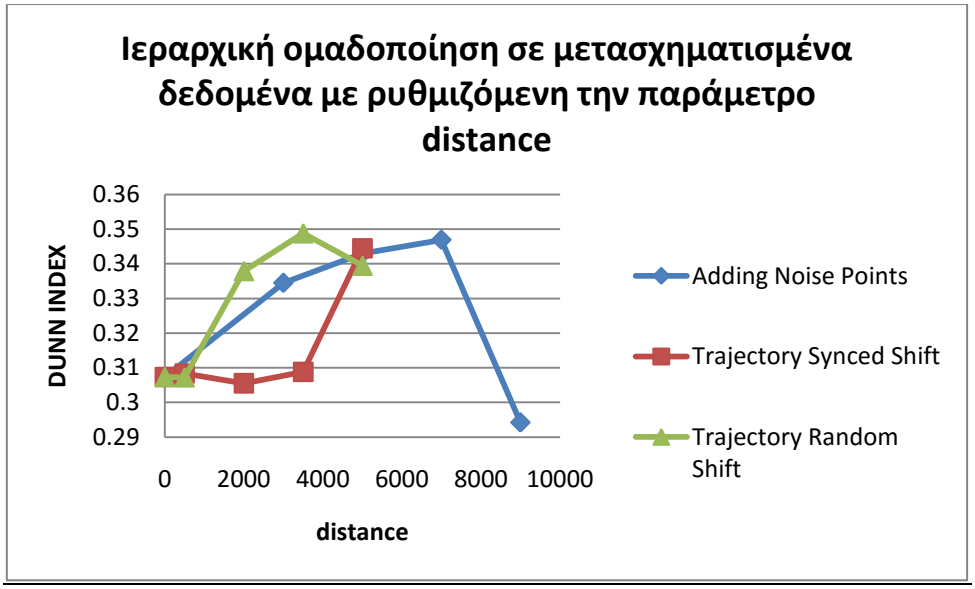

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
SILHOUETTE INDEX
0.373
0.372
0.371
0.37
0.369
0.368
0.367
0.366
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

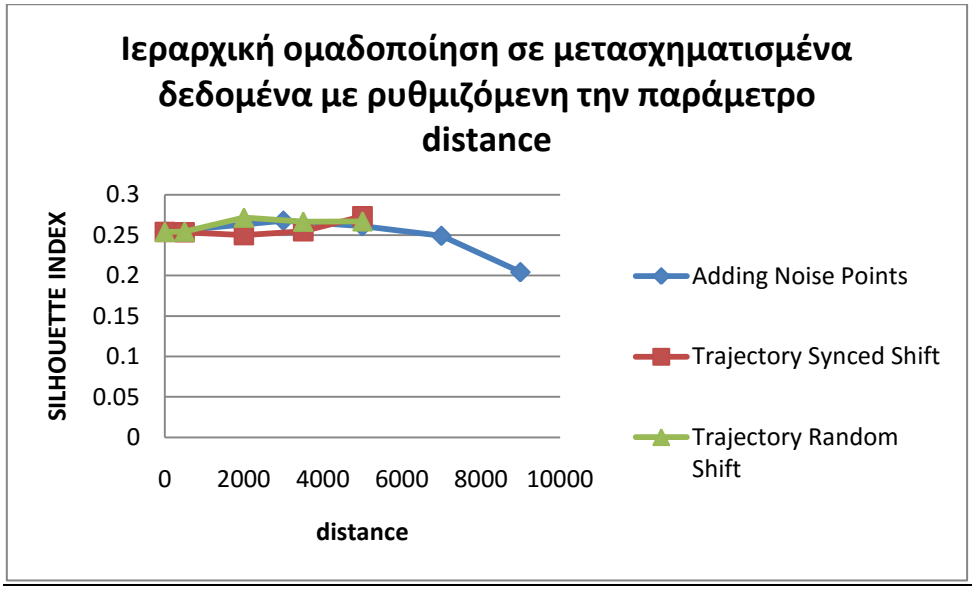
- Chebyshev απόσταση

Με χρήση της απόστασης Chebyshev, ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) φαίνεται να είναι ευαίσθητος στην προσθήκη θορύβου, στην συγχρονισμένη αλλά και στην τυχαία μετατόπιση ενώ αντίθετα φαίνεται να είναι ανθεκτικός στην αύξηση του ρυθμού δειγματοληψίας και να έχει αποδεκτή συμπεριφορά στην μείωση του ρυθμού δειγματοληψίας.



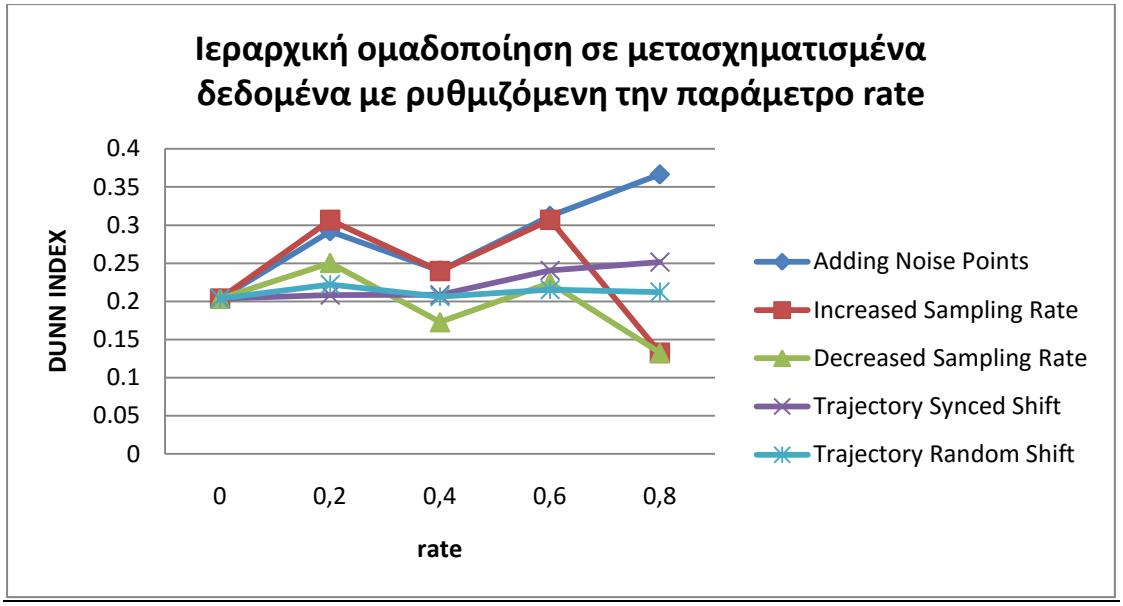


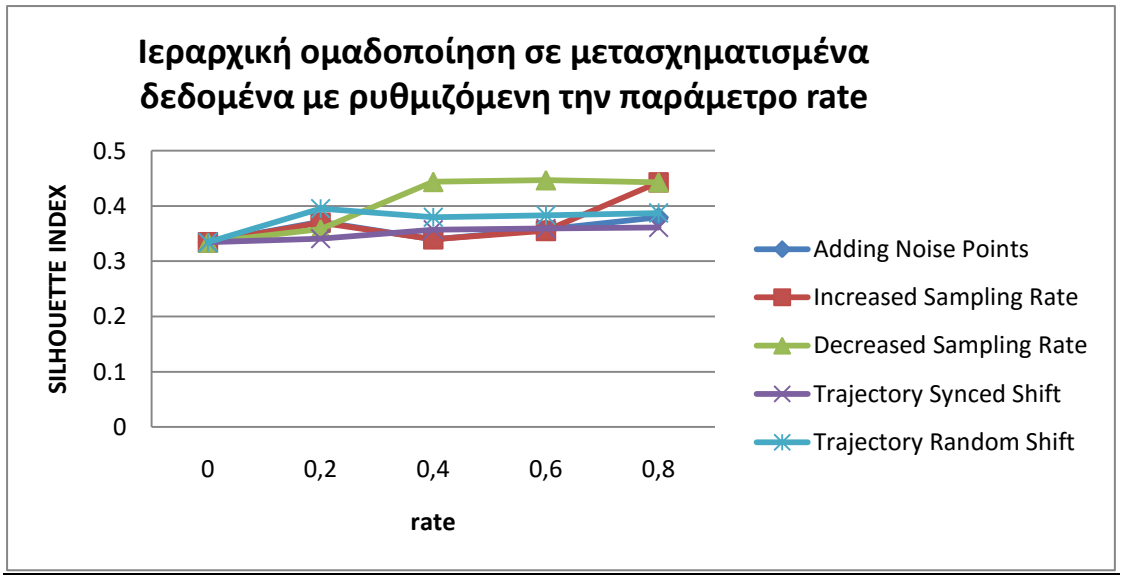


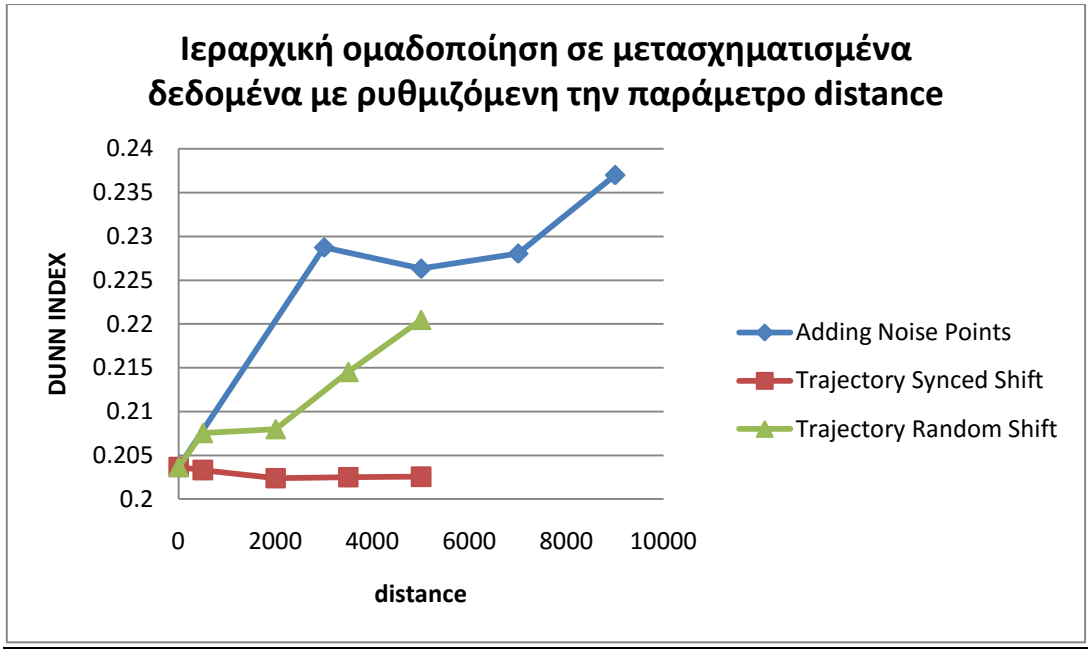


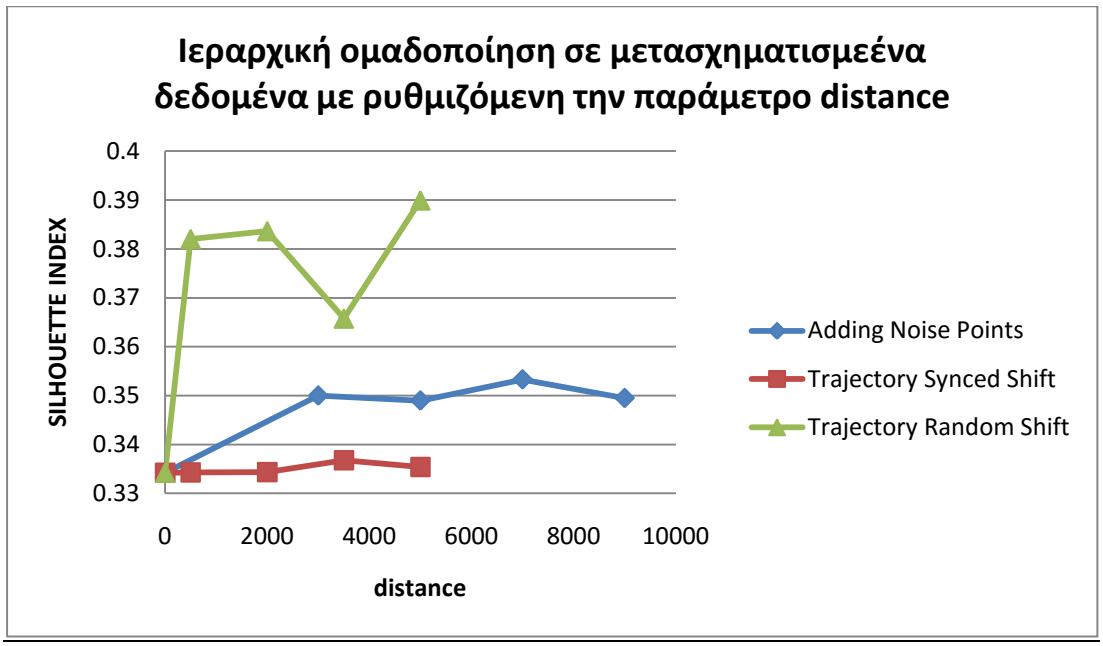
- Dynamic Time Warping

Ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) με τη χρήση της απόστασης DTW φαίνεται να είναι ευαίσθητος στην προσθήκη θορύβου. Παρουσιάζεται ευαίσθητος στη μείωση και στην αύξηση του ρυθμού δειγματοληψίας. Τέλος, δείχνει να είναι ανθεκτικός στην συγχρονισμένη και ευαίσθητος στην τυχαία μετατόπιση.


Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
DUNN INDEX
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0,2
0,4
0,6
0,8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

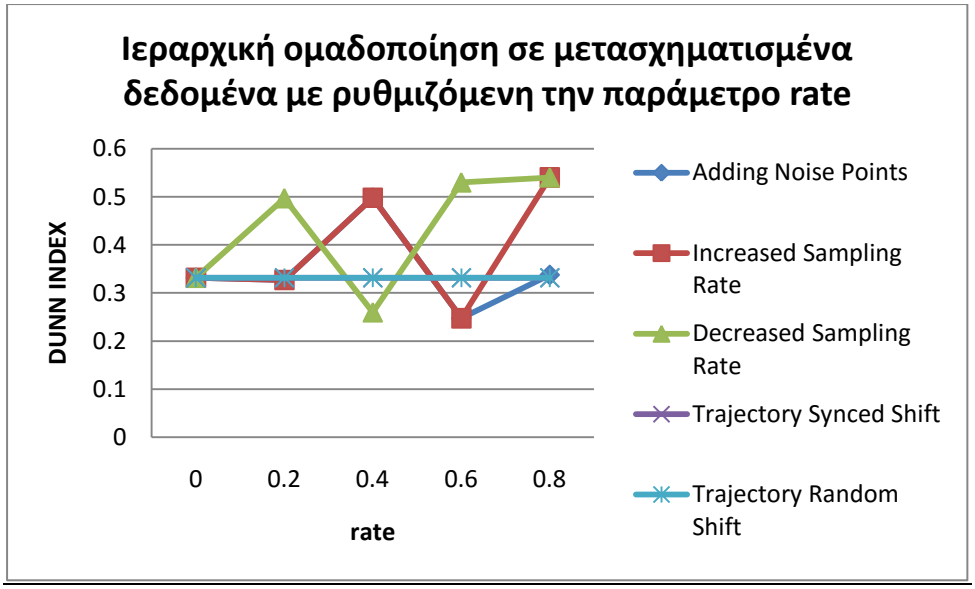

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUETTE INDEX
0.5
0.4
0.3
0.2
0.1
0
0
0,2
0,4
0,6
0,8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

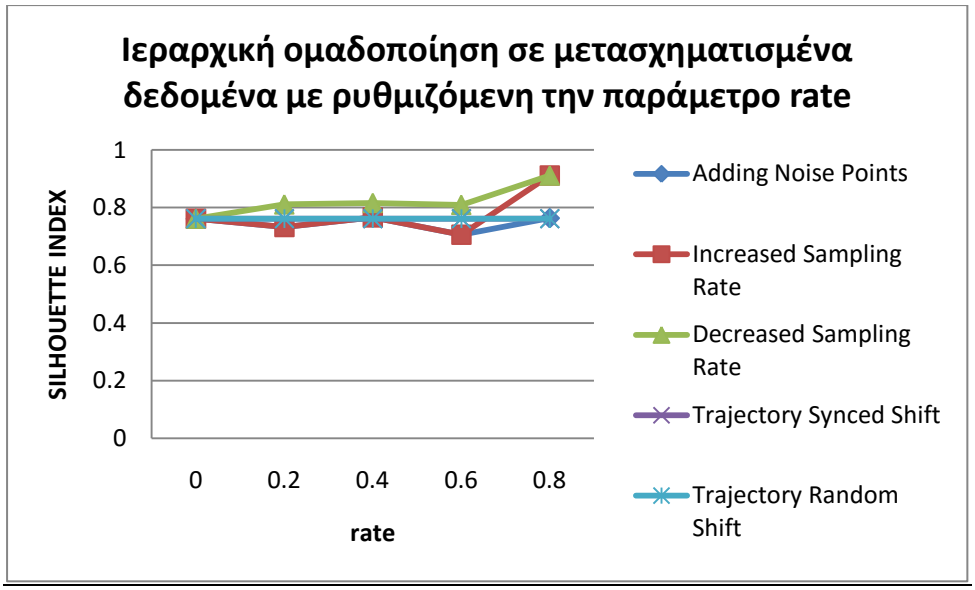


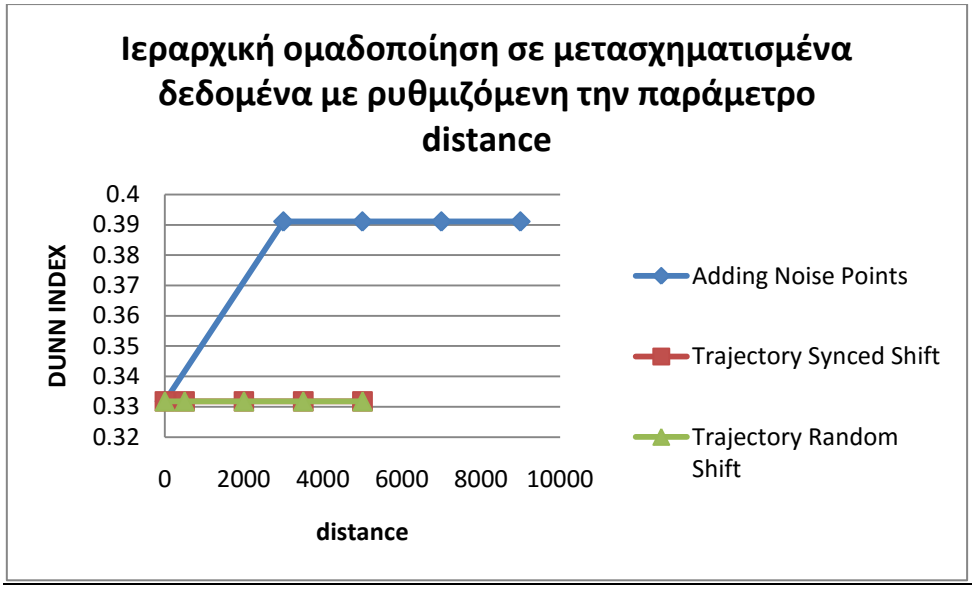


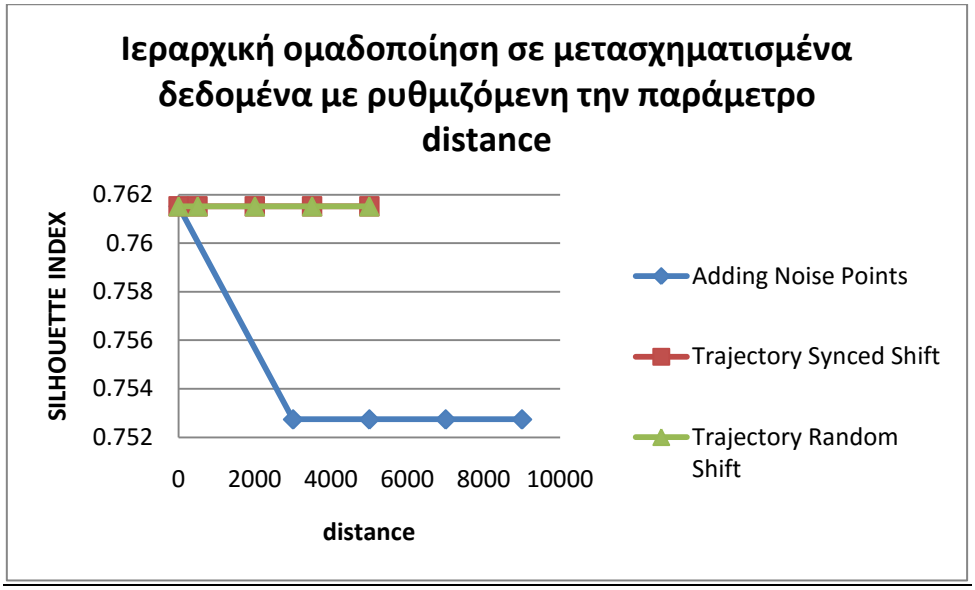
- Edit Distance on Real sequence

Ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) με τη χρήση της απόστασης EDR φαίνεται να είναι ευαίσθητος στους πρώτους τρείς μετασχηματισμούς, δηλαδή στην προσθήκη θορύβου, στην αύξηση και μείωση του ρυθμού δειγματοληψίας. Αντίθετα, παρουσιάζεται ανθεκτικός στη συγχρονισμένη και στην τυχαία μετατόπιση.


Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
DUNN INDEX
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

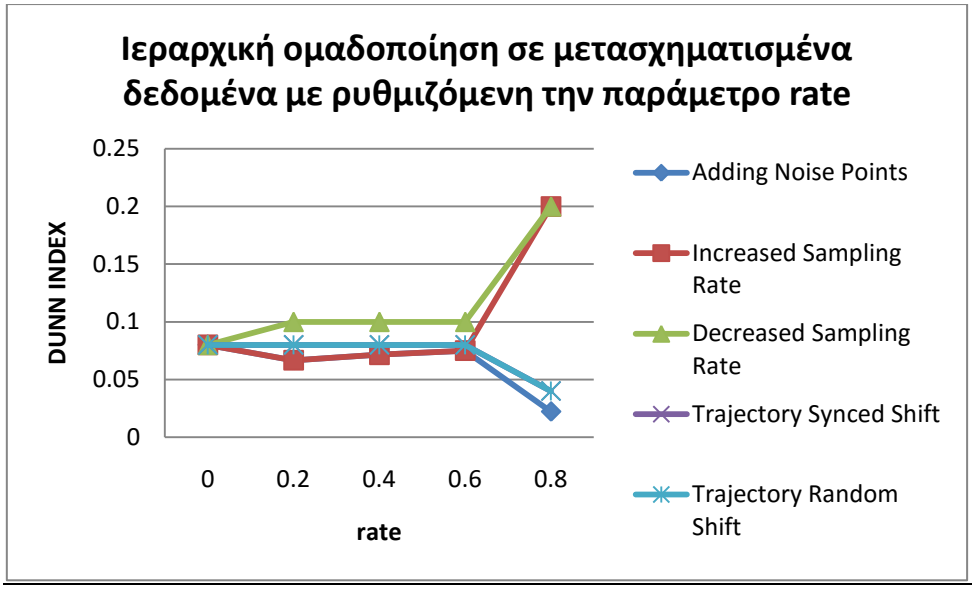

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUETTE INDEX
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

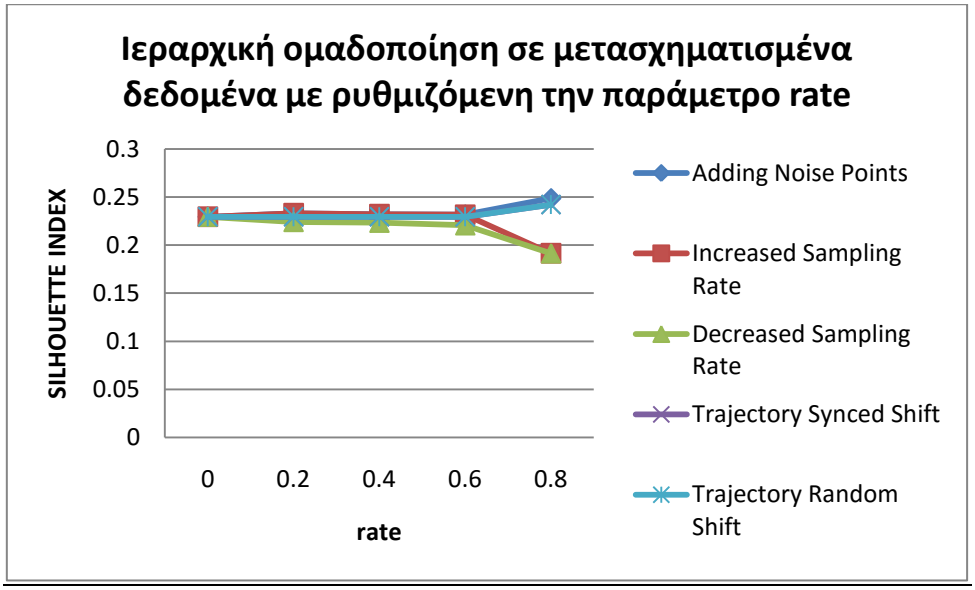


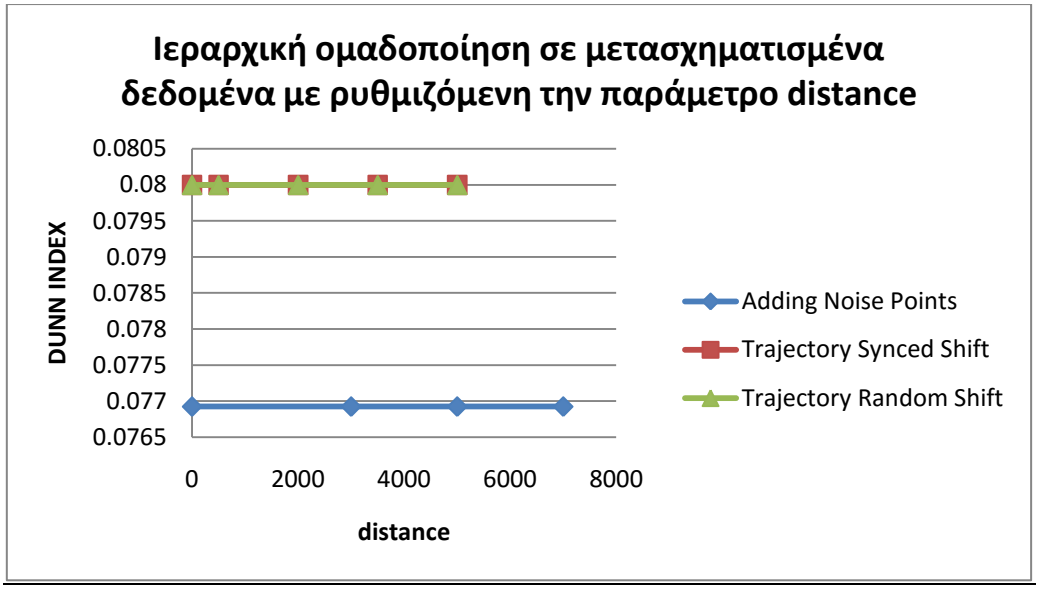


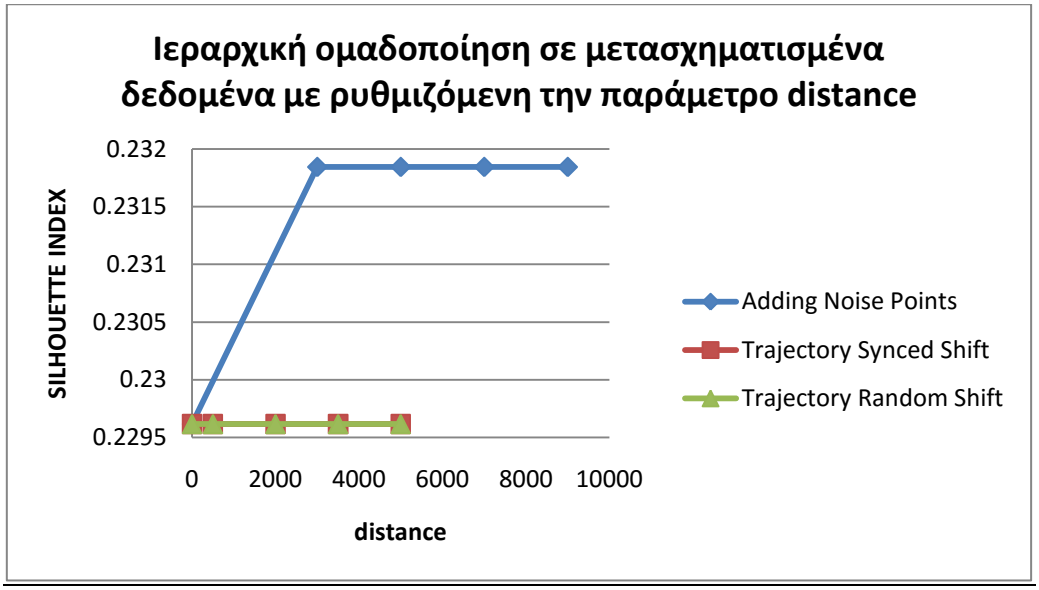
- Longest Common Subsequence

Ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) με τη χρήση της απόστασης LCSS φαίνεται να έχει αποδεκτή συμπεριφορά σε όλους τους μετασχηματισμούς, εκτός από τις περιπτώσεις όπου η παράμετρος rate λαμβάνει τη μέγιστη τιμή της.


Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
DUNN INDEX
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift


Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
SILHOUETTE INDEX
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift



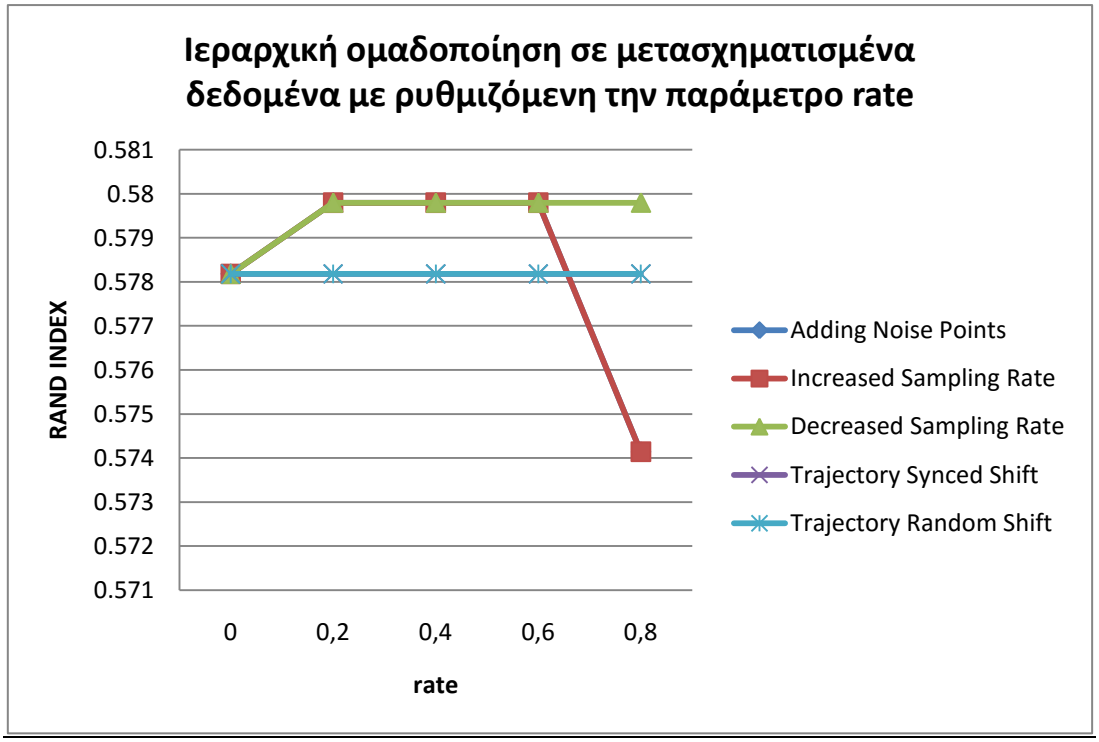


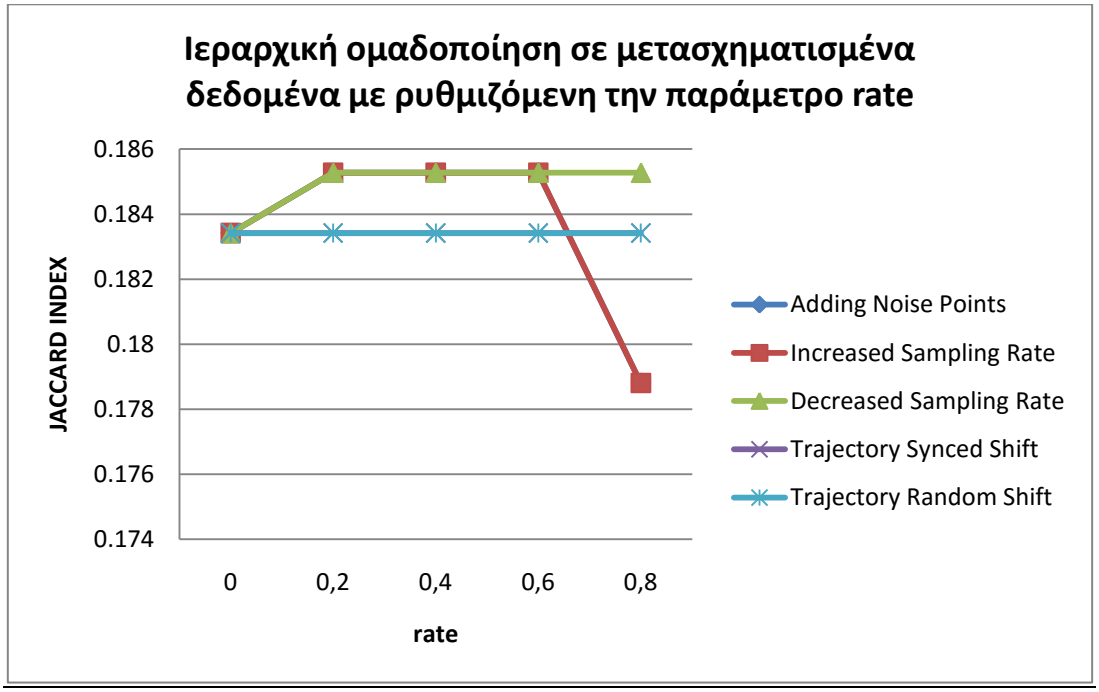
6.4.2 Εξωτερική αξιολόγηση ιεραρχικής ομαδοποίησης

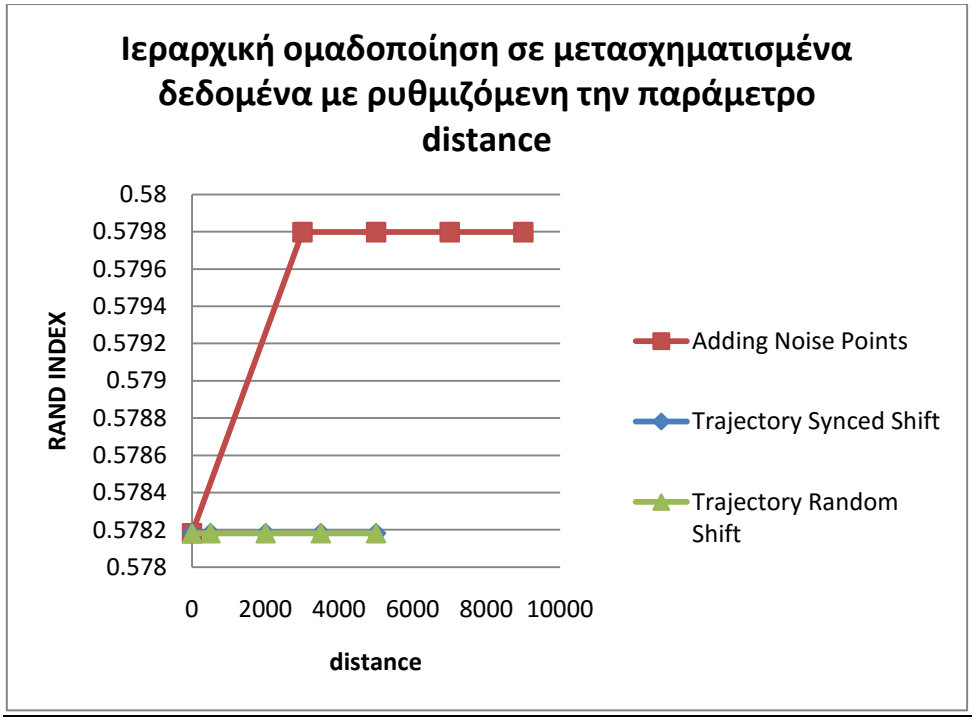
- Ευκλείδεια απόσταση

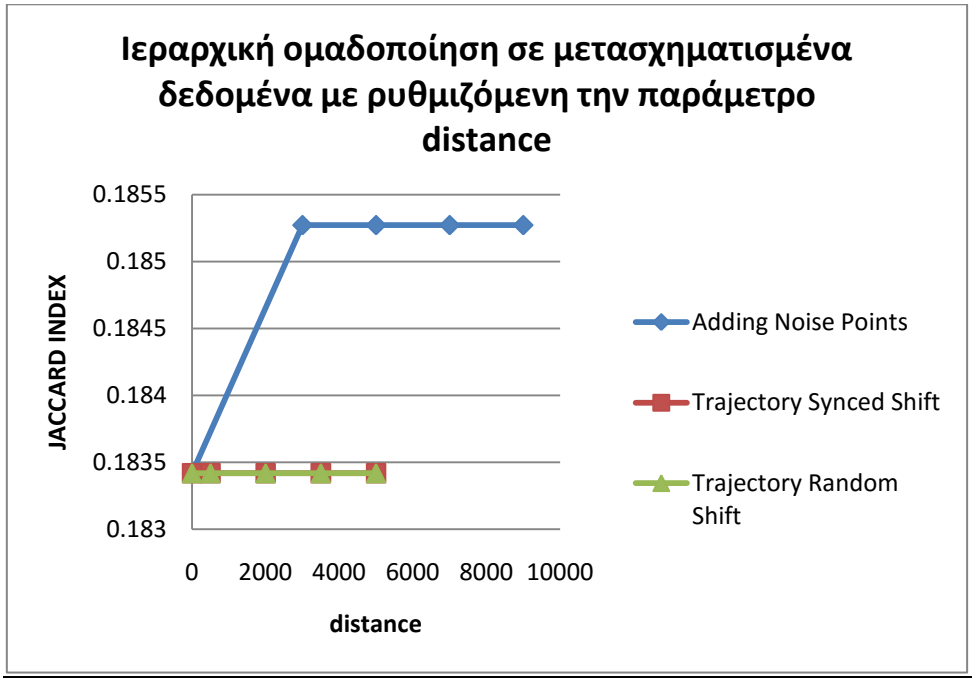
Με χρήση της ευκλείδειας απόστασης ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) δείχνει να είναι ευαίσθητος στην προσθήκη θορύβου, στον

μετασχηματισμό της αύξησης του ρυθμού δειγματοληψίας και στον μετασχηματισμό της μείωσης του ρυθμού δειγματοληψίας. Τέλος, ο αλγόριθμος της ιεραρχικής ομαδοποίησης παρουσιάζει στους δυο μετασχηματισμούς της μετατόπισης ανθεκτική συμπεριφορά.





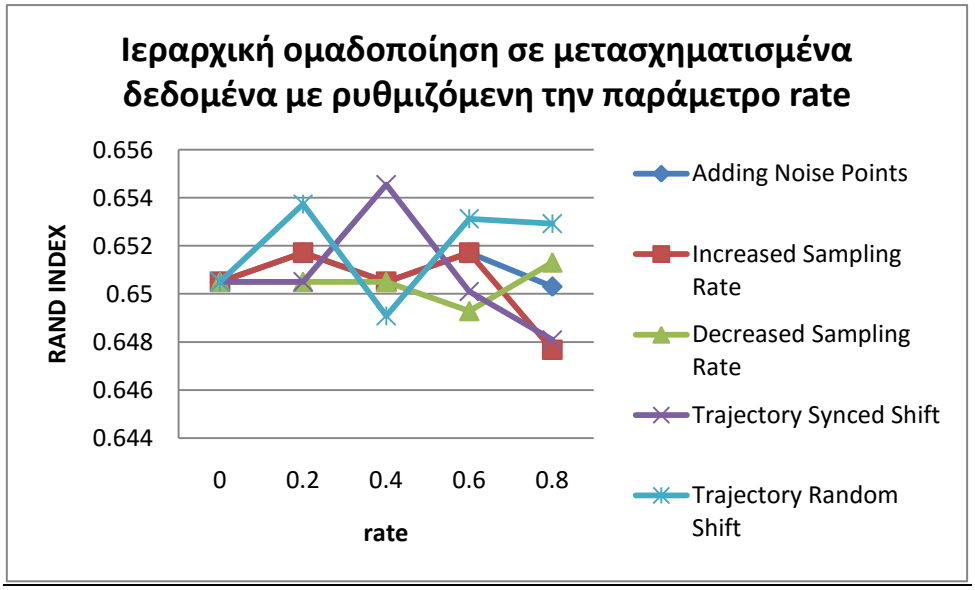


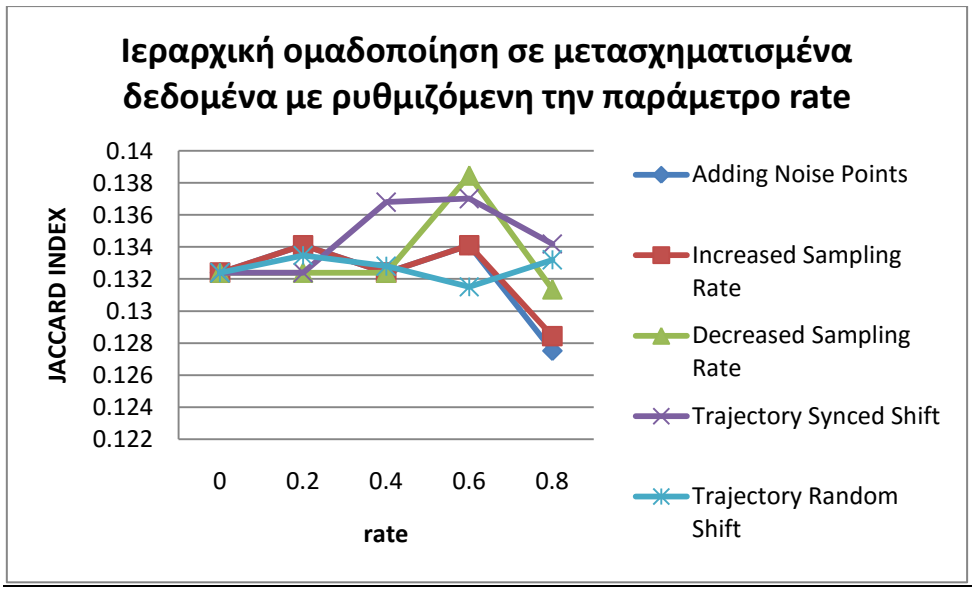


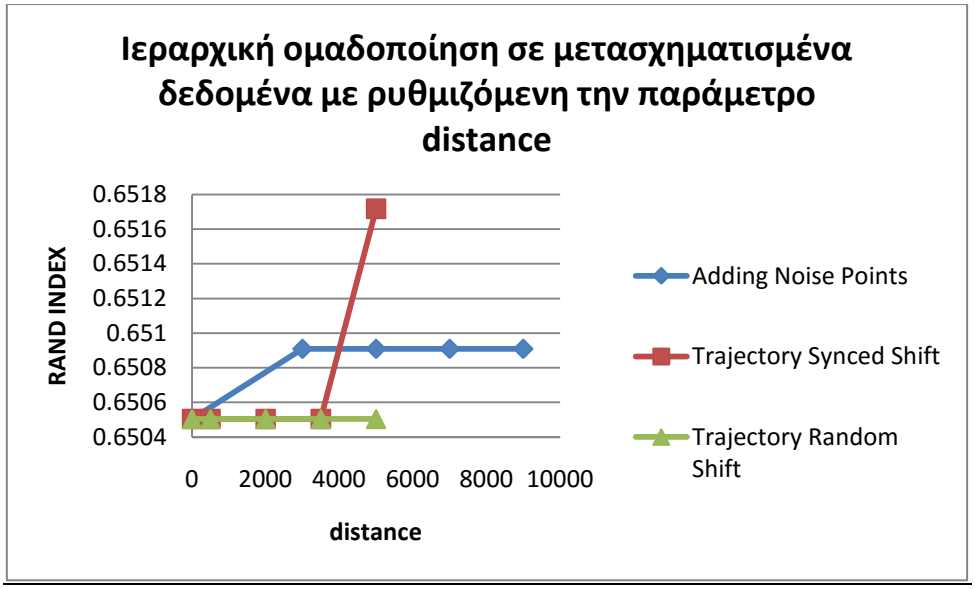
- Ευκλείδεια STARTEND

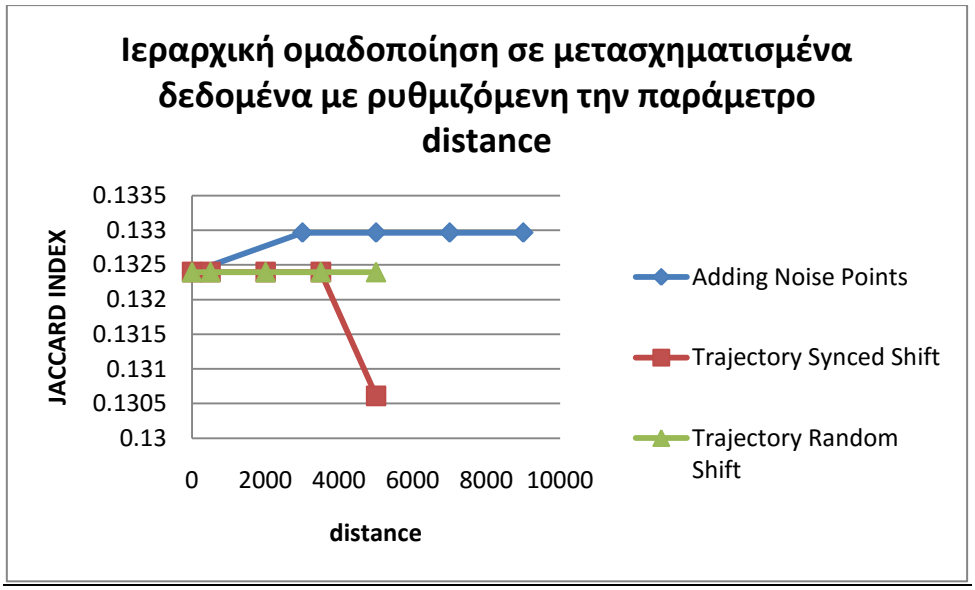
Με χρήση της απόστασης Euclideanstartend, ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) φαίνεται να είναι ευαίσθητος στην προσθήκη θορύβου. Εφαρμόζοντας τους μετασχηματισμούς της αύξησης και της μείωσης του ρυθμού δειγματοληψίας βλέπουμε ότι ο αλγόριθμος παρουσιάζεται ευαίσθητος και

στους δύο αυτούς τους μετασχηματισμούς. Τέλος, δείχνει να είναι ευαίσθητος και στους δυο μετασχηματισμούς της μετατόπισης με την παρατήρηση όμως ότι κατά την διαδικασία του μετασχηματισμού της τυχαίας μετατόπισης με ρυθμιζόμενη την παράμετρο της απόστασης ο αλγόριθμος παρουσιάζεται ανθεκτικός.



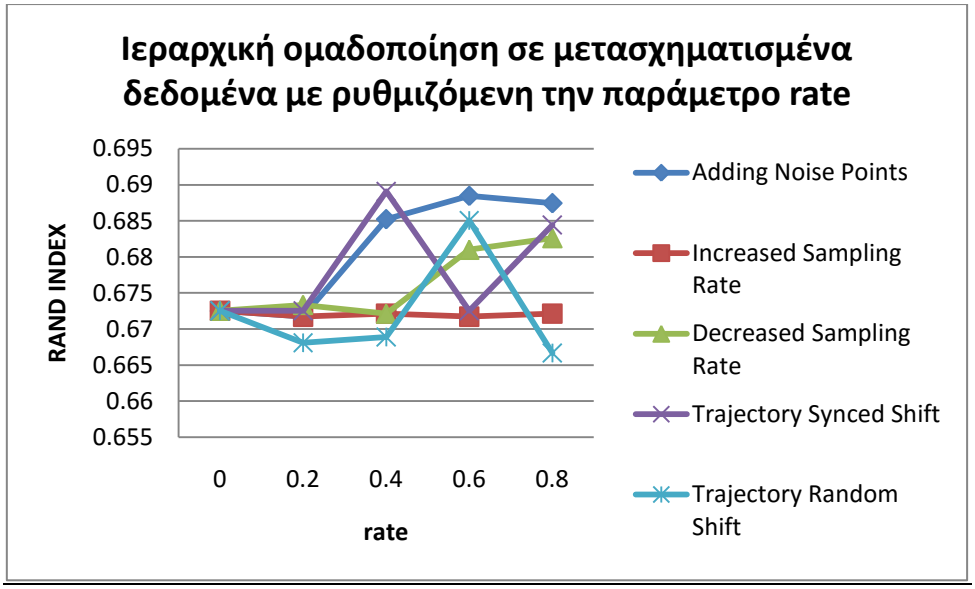


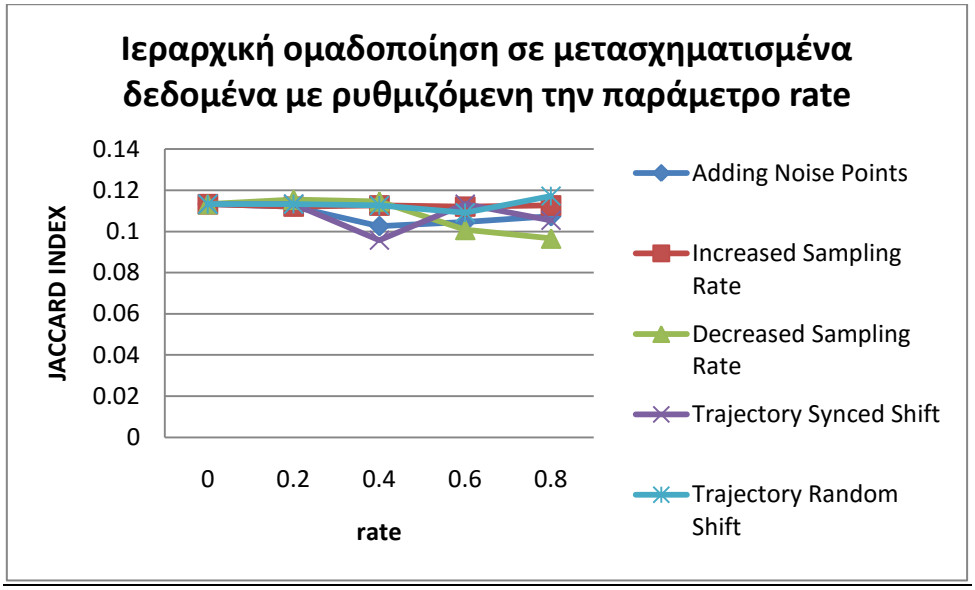


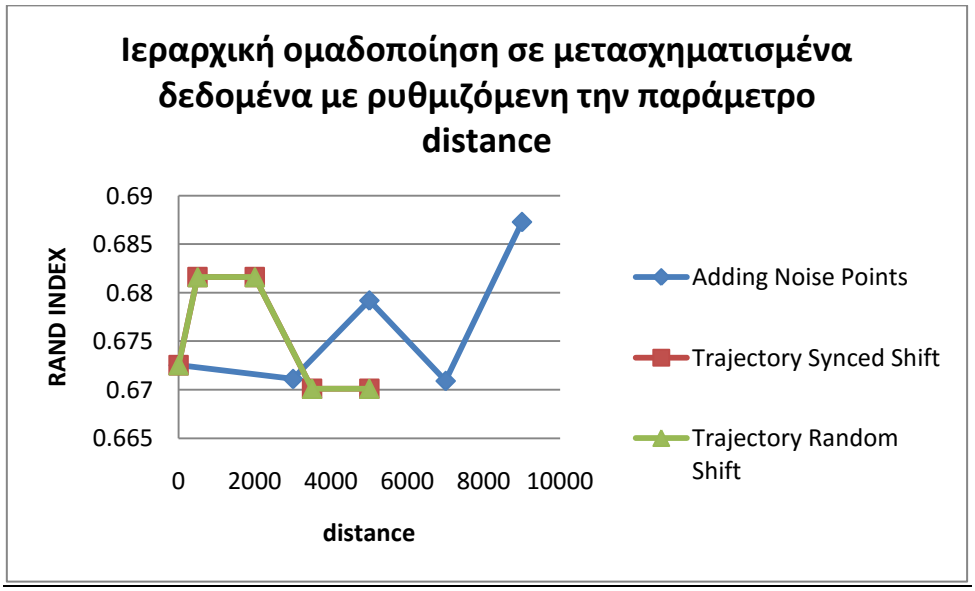


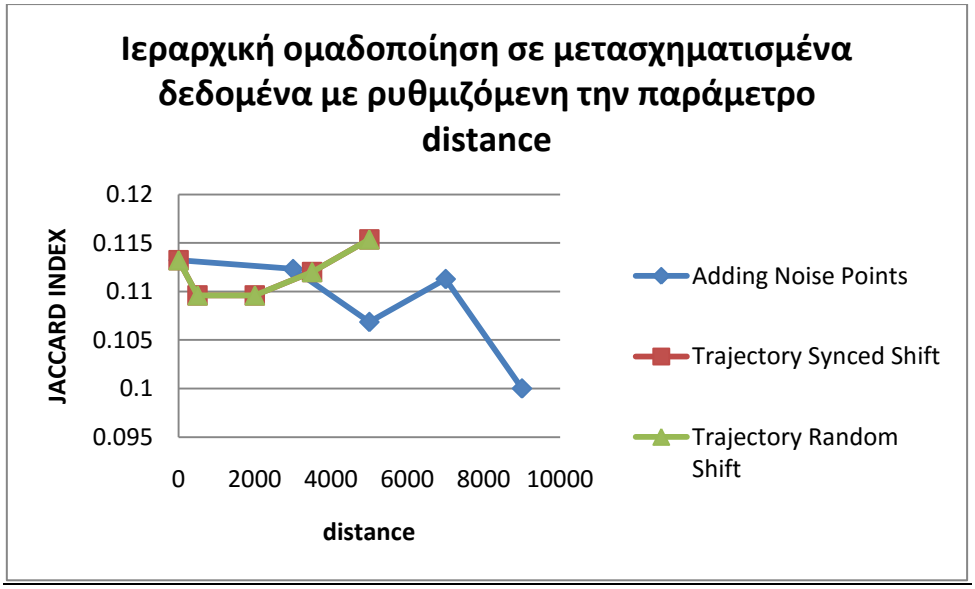
- Manhattan απόσταση

Με χρήση της απόστασης Manhattan, ο αλγόριθμος της Ιεραρχικής ομαδοποίησης δείχνει να επηρεάζεται σημαντικά από όλους τους μετασχηματισμούς εκτός από αυτόν της αύξησης του ρυθμού δειγματοληψίας που δείχνει να είναι ανθεκτικός.


Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
RAND INDEX
0.695
0.69
0.685
0.68
0.675
0.67
0.665
0.66
0.655
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

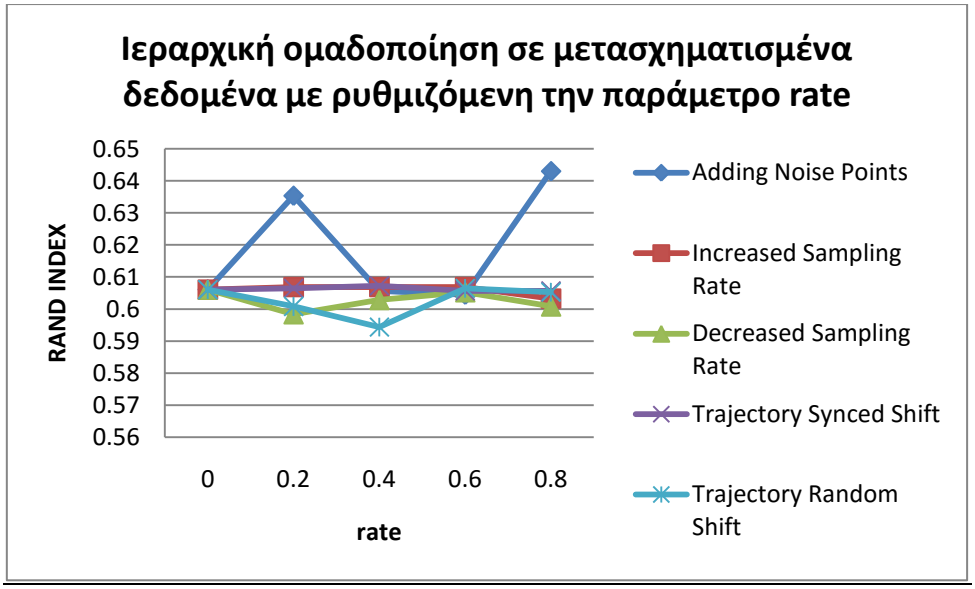

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

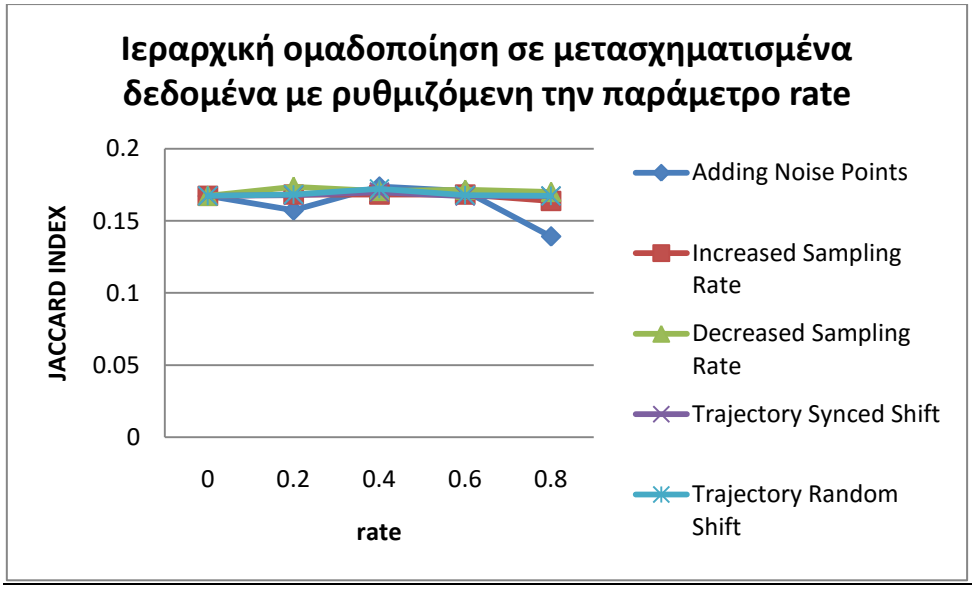

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
RAND INDEX
0.69
0.685
0.68
0.675
0.67
0.665
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

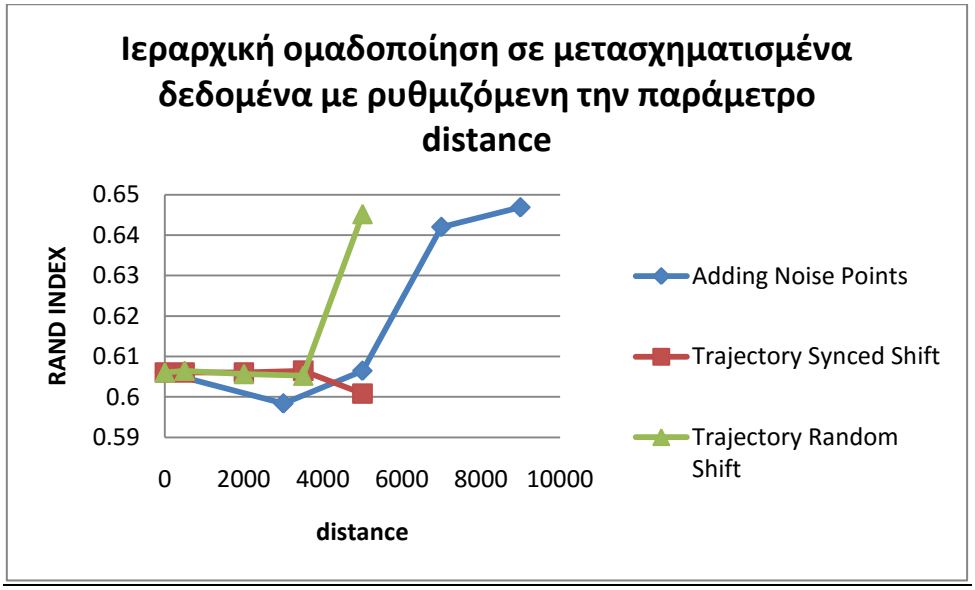


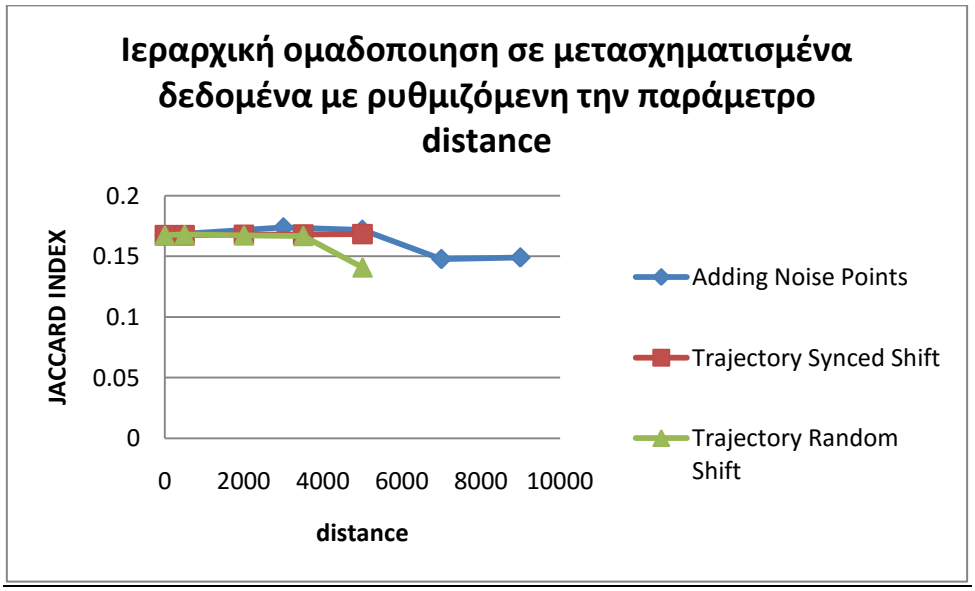
- Chebyshev απόσταση

Με χρήση της απόστασης Chebyshev, ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) φαίνεται να είναι ευαίσθητος στην προσθήκη θορύβου. Εφαρμόζοντας τους μετασχηματισμούς της αύξησης και της μείωσης του ρυθμού δειγματοληψίας βλέπουμε ότι ο αλγόριθμος είναι ανθεκτικός στην αύξηση του ρυθμού δειγματοληψίας και έχει αποδεκτή συμπεριφορά στην μείωση του ρυθμού δειγματοληψίας. Τέλος δείχνει να έχει αποδεκτή συμπεριφορά στον μετασχηματισμό της συγχρονισμένης μετατόπισης και να έχει ευαίσθητη συμπεριφορά στην τυχαία μετατόπιση.



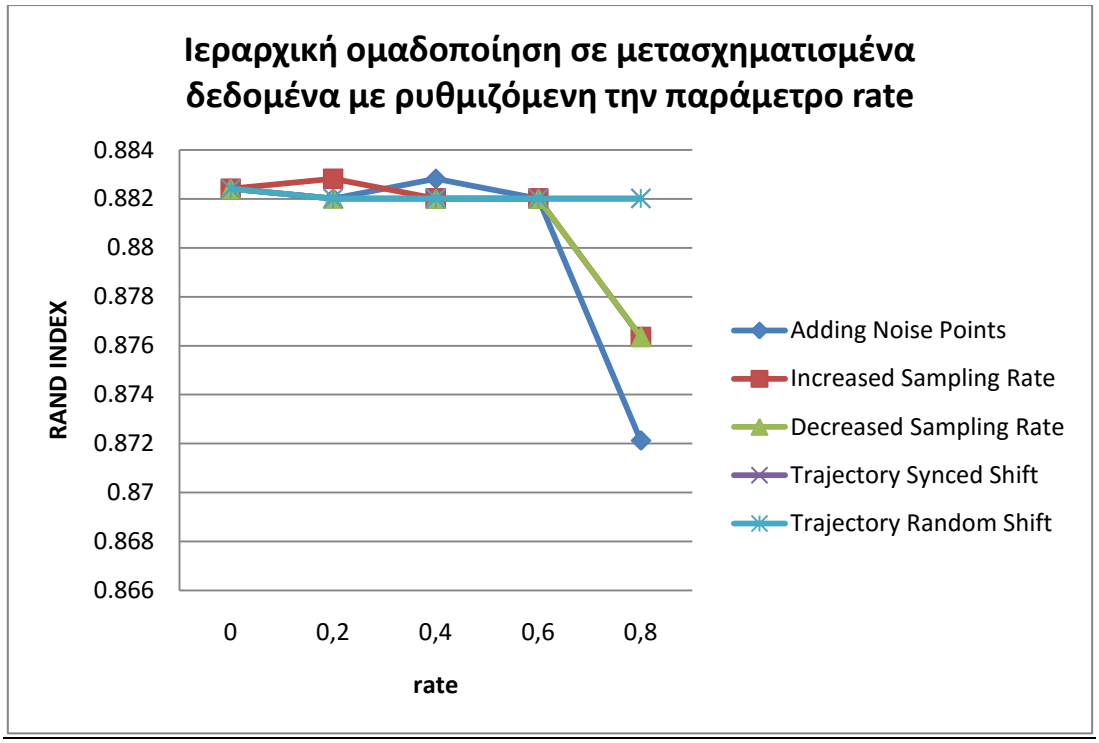

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

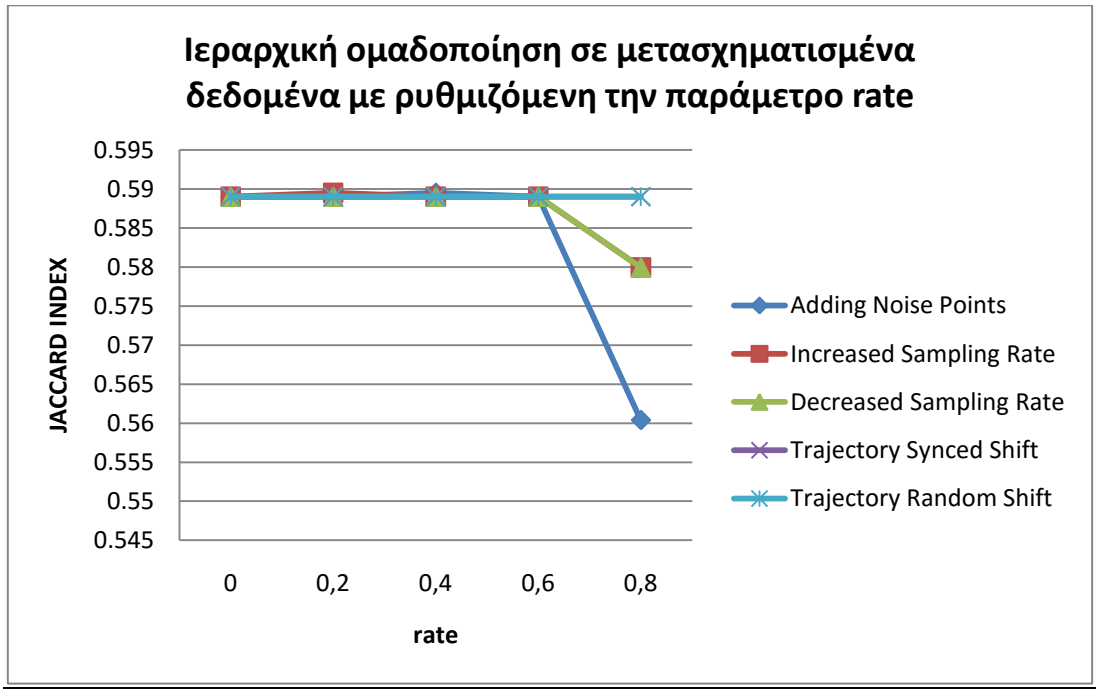

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
RAND INDEX
0.65
0.64
0.63
0.62
0.61
0.6
0.59
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

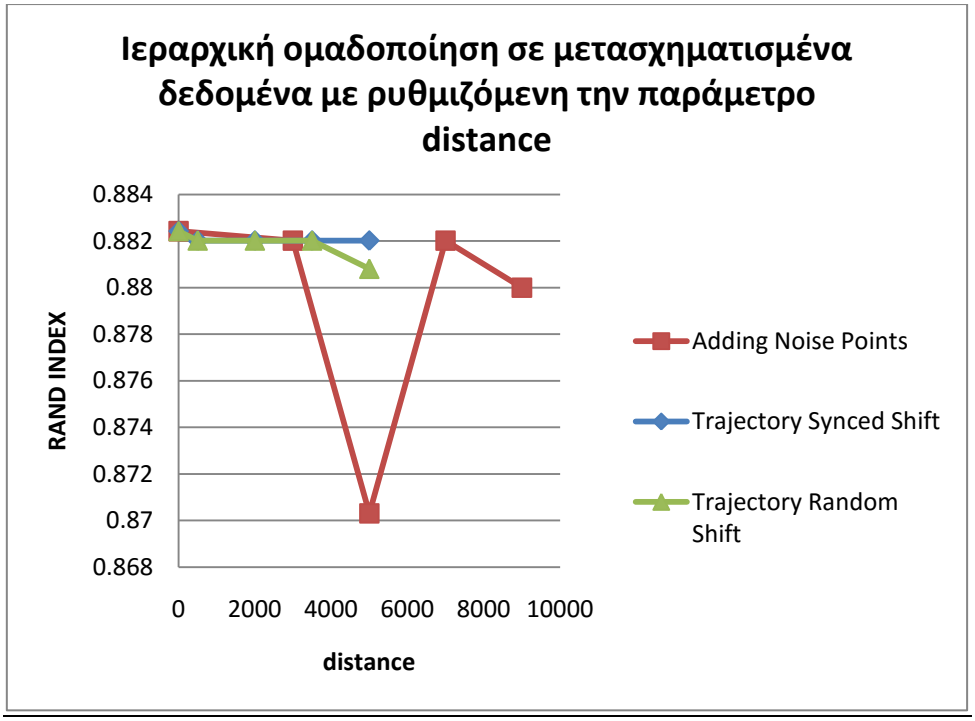

Ιεραρχική ομαδοποιηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
JACCARD INDEX
0.2
0.15
0.1
0.05
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

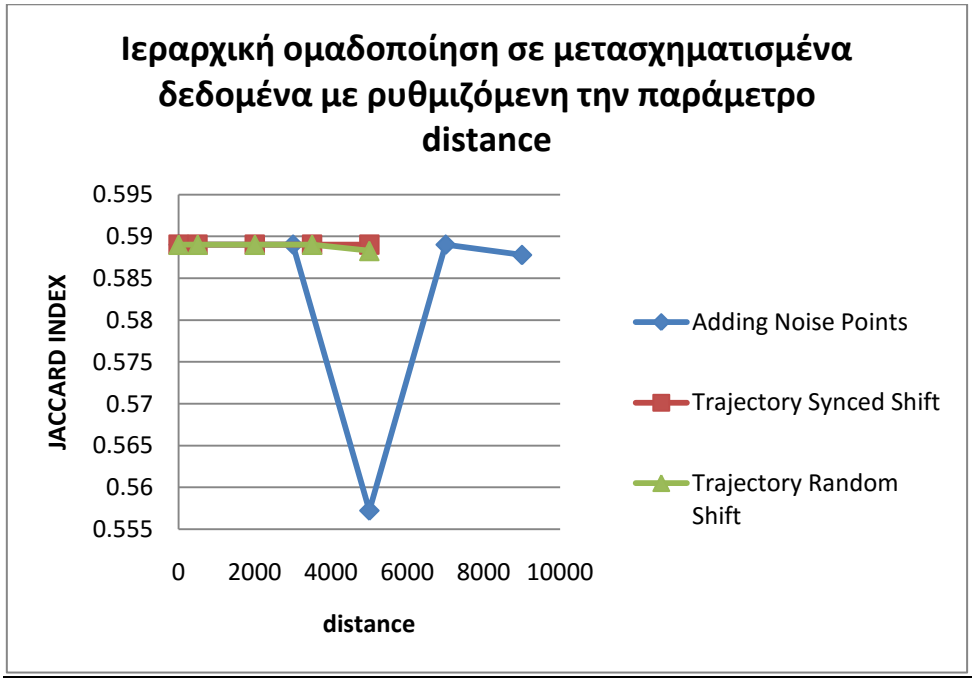
- Dynamic Time Warping

Ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) με τη χρήση της απόστασης DTW φαίνεται να είναι ευαίσθητος στην προσθήκη θορύβου. Παρουσιάζεται να έχει ευαίσθητη συμπεριφορά στη μείωση και στην αύξηση του ρυθμού δειγματοληψίας . Τέλος, δείχνει να είναι ανθεκτικός στην συγχρονισμένη και στην τυχαία μετατόπιση.



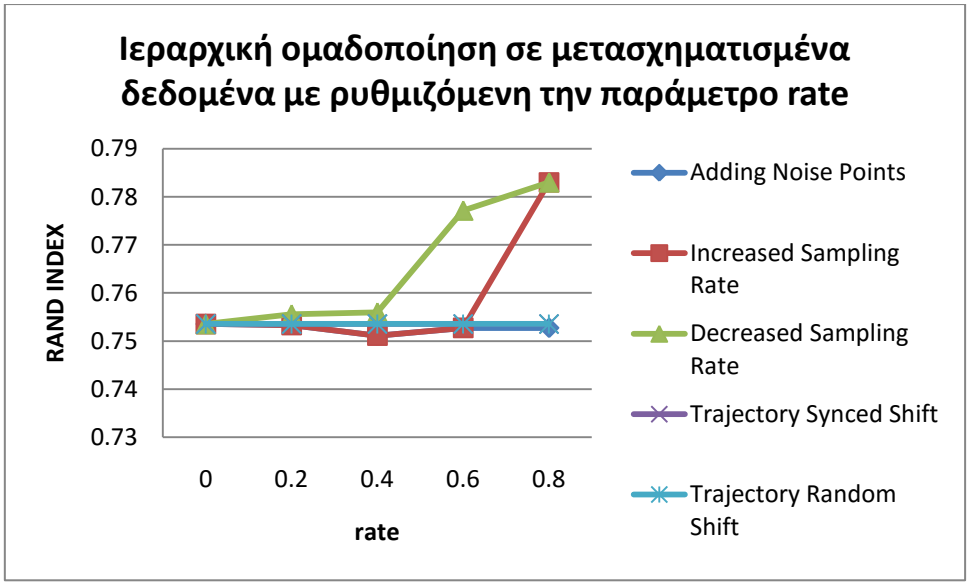


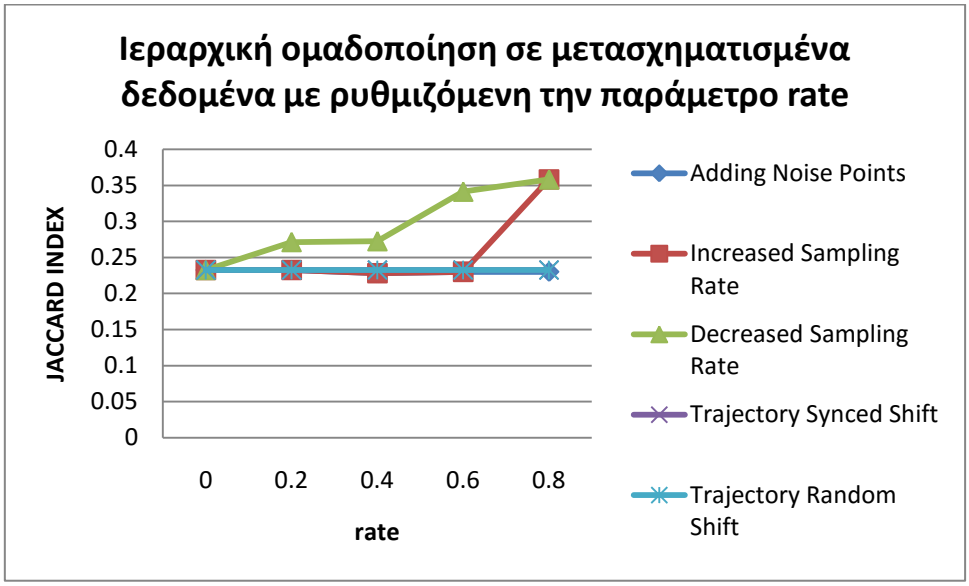


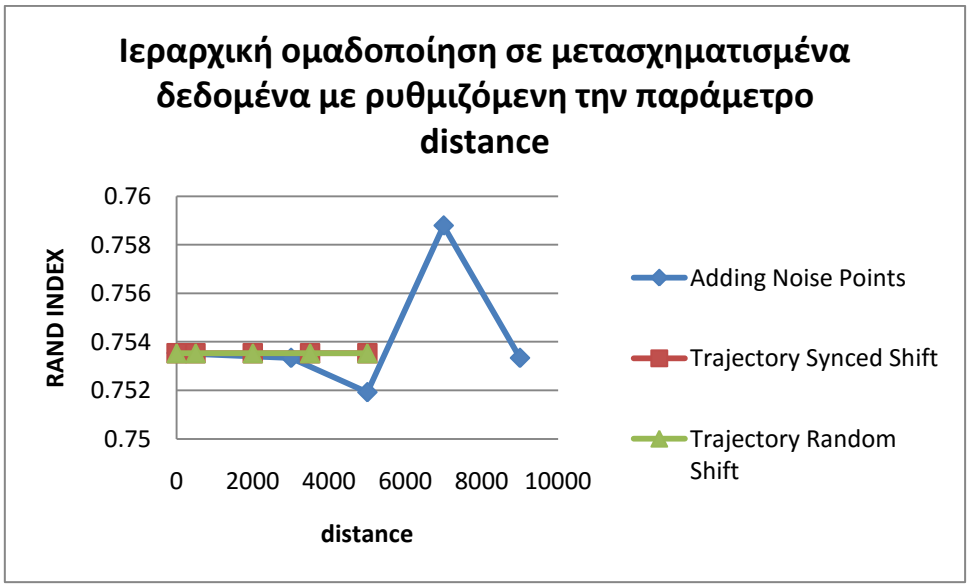


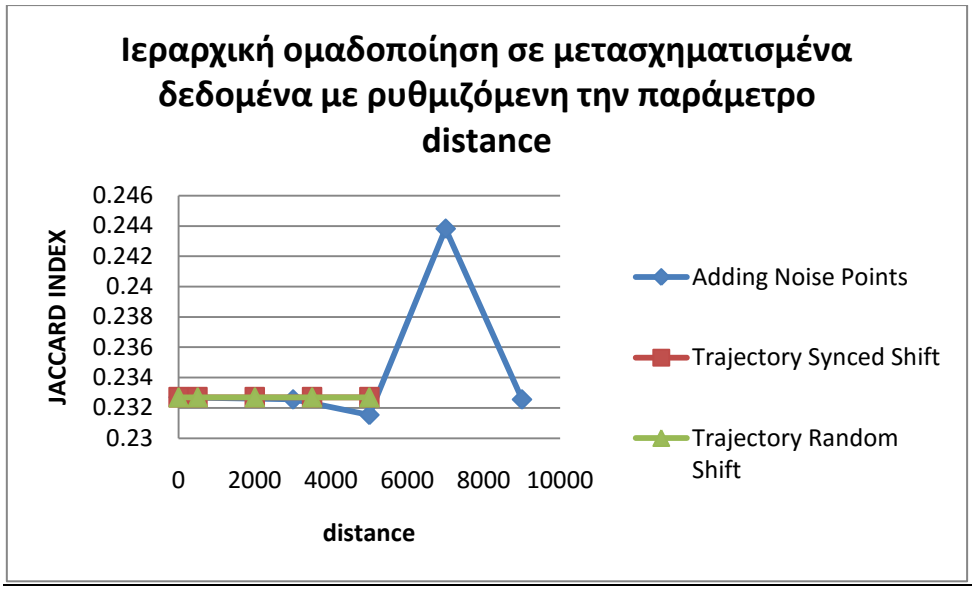
- Edit Distance on Real sequence

Ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) με τη χρήση της απόστασης EDR φαίνεται να είναι ευαίσθητος στον μετασχηματισμό της προσθήκης θορύβου. Ακόμη ευαισθητος δείχνει και στους μετασχηματισμούς της αύξησης και της μείωσης του ρυθμού δειγματολειψίας. Αντίθετα, δείχνει να είναι ανθεκτικός στην τυχαία και στη συγχρονισμένη μετατόπιση.


Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
RAND INDEX
0.79
0.78
0.77
0.76
0.75
0.74
0.73
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

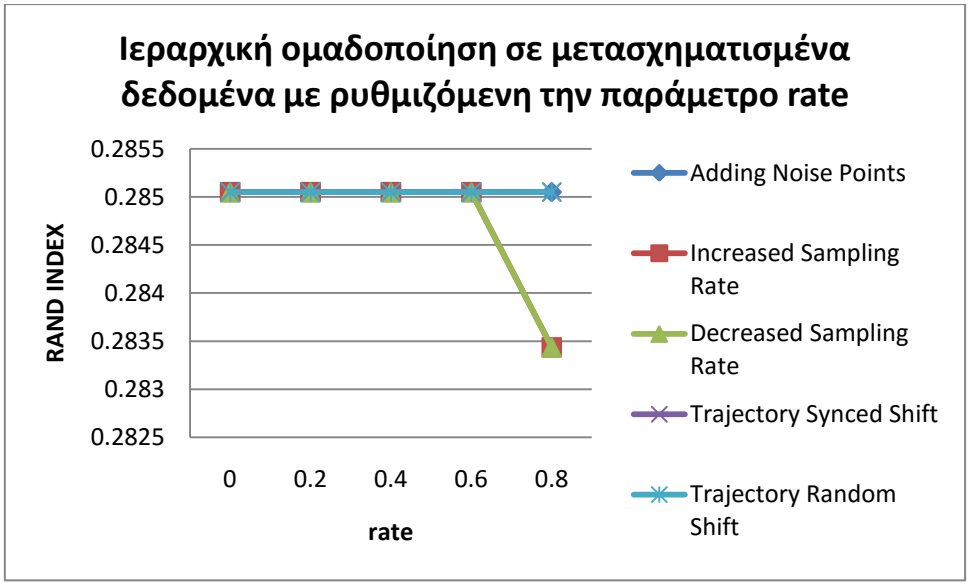

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift

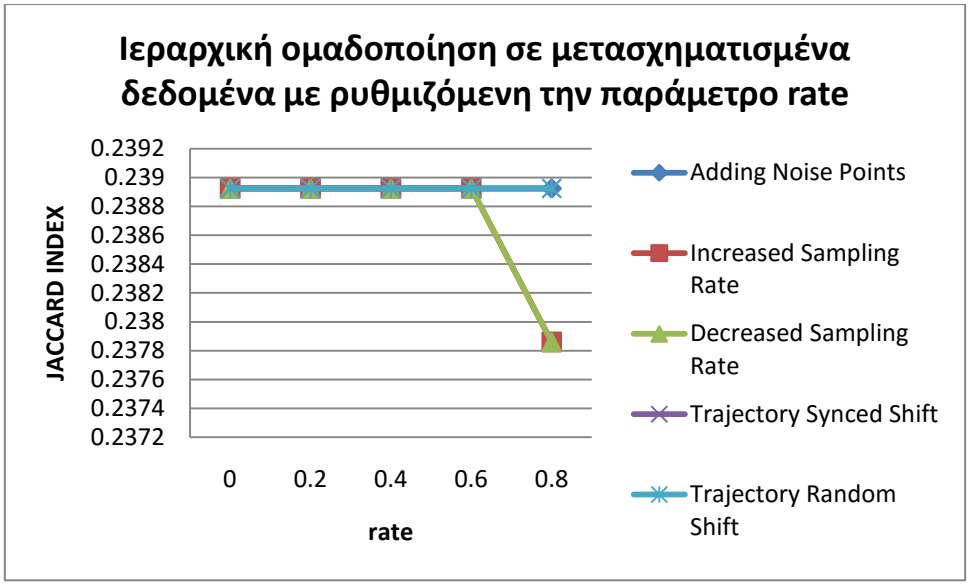

Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
RAND INDEX
0.76
0.758
0.756
0.754
0.752
0.75
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

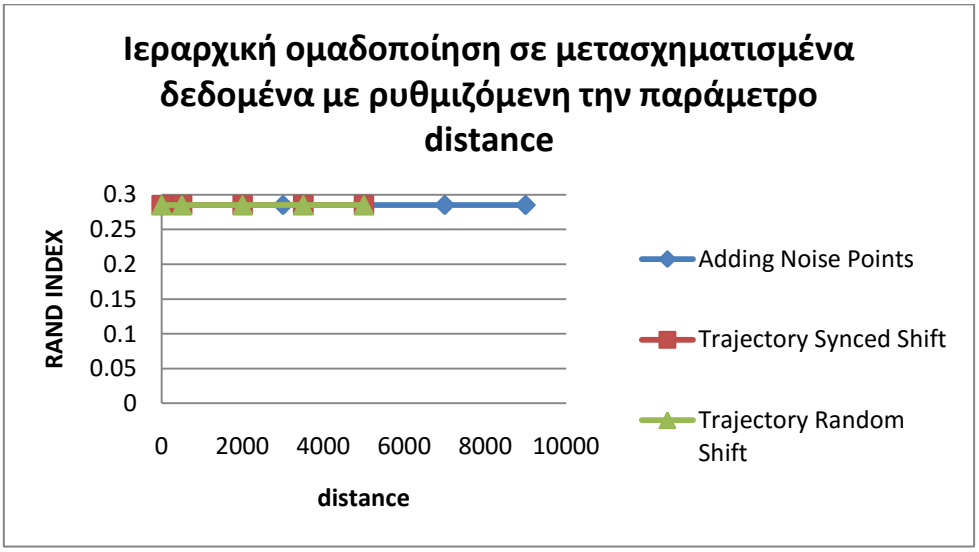


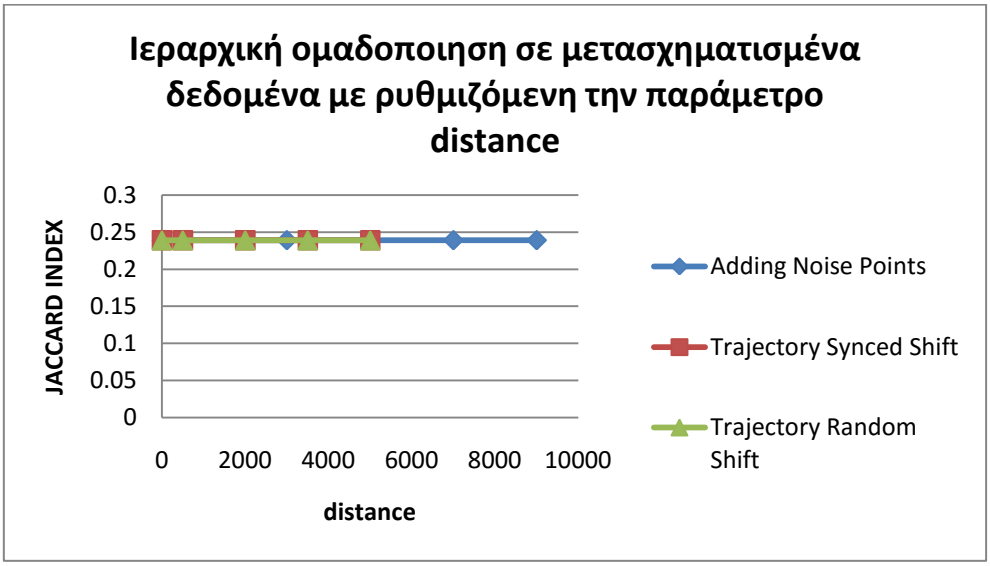
- Longest Common Subsequence

Ο αλγόριθμος της ιεραρχικής ομαδοποίησης (μέθοδος Ward) με τη χρήση της απόστασης LCSS φαίνεται να είναι ανθεκτικός γενικά σε όλους τους μετασχηματισμούς εκτός από την αύξηση και τη μείωση του ρυθμού δειγματοληψίας στους οποίους φαίνεται να έχει αποδεκτή συμπεριφορά και αυτό διότι βλέπουμε να επηρεάζεται όταν η παράμετρος rate παίρνει την μεγαλύτερη τιμή της.




Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο rate
JACCARD INDEX
0.2392
0.239
0.2388
0.2386
0.2384
0.2382
0.238
0.2378
0.2376
0.2374
0.2372
0
0.2
0.4
0.6
0.8
rate
Adding Noise Points
Increased Sampling Rate
Decreased Sampling Rate
Trajectory Synced Shift
Trajectory Random Shift


Ιεραρχική ομαδοποίηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
RAND INDEX
0.3
0.25
0.2
0.15
0.1
0.05
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift


Ιεραρχική ομαδοποιηση σε μετασχηματισμένα δεδομένα με ρυθμιζόμενη την παράμετρο distance
JACCARD INDEX
0.3
0.25
0.2
0.15
0.1
0.05
0
0
2000
4000
6000
8000
10000
distance
Adding Noise Points
Trajectory Synced Shift
Trajectory Random Shift

7. Συμπεράσματα

Στην παρούσα εργασία βασικός μας στόχος ήταν η μελέτη και η αξιολόγηση της εγκυρότητας των ομαδοποιήσεων που επιτυγχάνονται από τον αλγόριθμο optics και την ιεραρχική ομαδοποίηση με τη μέθοδο Ward, χρησιμοποιώντας κάθε φορά διαφορετικά μέτρα απόστασης (ομοιότητας) και διαφορετικούς δείκτες αξιολόγησης. Αναλυτικότερα, θεωρήσαμε τους δείκτες Dunn & Silhouette για την εσωτερική αξιολόγηση, και τους δείκτες Rand & Jaccard για την εξωτερική, αντίστοιχα. Συγκεκριμένα, θεωρήσαμε μετρικές συναρτήσεις απόστασης (Ευκλείδεια, Manhattan, Chebyshev, EuclideanSTARTEND), και μη μετρικές συναρτήσεις απόστασης, είτε βασισμένες στη δυναμική χρονική στρέβλωση (Dynamic Time Warping), είτε βασισμένες στην "επεξεργασία" της απόστασης (Edit Distance on Real sequence) ή βασισμένες στη μεγαλύτερη κοινή υποαλληλουχία (Longest Common Subsequence). Εφαρμόσαμε ποικίλους μετασχηματισμούς τροχιών (επαναδειγματολειψία, προσθήκη θορύβου και μετατόπιση σημείου) ελεγχόμενους από δυο παραμέτρους, τον ρυθμό και την απόσταση, σε πραγματικά και συνθετικά σύνολα δεδομένων τροχιών. Για κάθε μετασχηματισμό, αξιολογήσαμε την ομαδοποίηση του αρχικού συνόλου δεδομένων και των μετασχηματισμένων συνόλων, ανάλογα με την τιμή της παραμέτρου που "τρέχει".

Τα εξαγόμενα αποτελέσματα της εκτενούς πειραματικής μελέτης για τον αλγόριθμο optics και την ιεραρχική ομαδοποίηση παρουσιάζονται συγκεντρωτικά στους ακόλουθους Πίνακες 2-5, για κάθε έναν από τους μετασχηματισμούς που εφαρμόστηκαν, για την εσωτερική και εξωτερική αξιολόγηση, αντίστοιχα.

Πίνακας 2: Συγκεντρωτικά αποτελέσματα εσωτερικής αξιολόγησης ομαδοποίησης Optics

ΕΣΩΤΕΡΙΚΗ ΑΞΙΟΛΟΓΗΣΗ		Euclidean	Manhattan	Chebysev	Euclidean StartEnd	DTW	EDR	LCSS
OPTICS	Adding Noise Points	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive
	Increased Sampling rate	Sensitive	Sensitive	Robust	Robust	Fair	Sensitive	Fair
	Decreased Sampling rate	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Fair	Sensitive
	Trajectory Synced Shift	Robust	Robust	Sensitive	Sensitive	Robust	Robust	Fair
	Trajectory Random Shift	Robust	Fair	Sensitive	Sensitive	Sensitive	Robust	Fair

Πίνακας 3: Συγκεντρωτικά αποτελέσματα εξωτερικής αξιολόγησης ομαδοποίησης Optics

ΕΞΩΤΕΡΙΚΗ ΑΞΙΟΛΟΓΗΣΗ		Euclidean	Manhattan	Chebysev	Euclidean StartEnd	DTW	EDR	LCSS
OPTICS	Adding Noise Points	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Fair
	Increased Sampling rate	Sensitive	Sensitive	Robust	Robust	Sensitive	Sensitive	Fair
	Decreased Sampling rate	Robust	Sensitive	Fair	Sensitive	Sensitive	Sensitive	Fair
	Trajectory Synced Shift	Fair	Robust	Fair	Fair	Fair	Robust	Fair
	Trajectory Random Shift	Fair	Robust	Sensitive	Robust	Fair	Robust	Fair

Πίνακας 4: Συγκεντρωτικά αποτελέσματα εσωτερικής αξιολόγησης ιεραρχικής ομαδοποίησης

ΕΣΩΤΕΡΙΚΗ ΑΞΙΟΛΟΓΗΣΗ		Euclidean	Manhattan	Chebysev	Euclidean StartEnd	DTW	EDR	LCSS
Ιεραρχική Ομαδοποίηση	Adding Noise Points	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Fair
	Increased Sampling rate	Sensitive	Fair	Robust	Robust	Sensitive	Sensitive	Fair
	Decreased Sampling rate	Sensitive	Sensitive	Fair	Sensitive	Sensitive	Sensitive	Sensitive
	Trajectory Synced Shift	Sensitive	Sensitive	Sensitive	Sensitive	Robust	Robust	Robust
	Trajectory Random Shift	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Robust	Robust

Πίνακας 5: Συγκεντρωτικά αποτελέσματα εξωτερικής αξιολόγησης ιεραρχικής ομαδοποίησης

ΕΞΩΤΕΡΙΚΗ ΑΞΙΟΛΟΓΗΣΗ		Euclidean	Manhattan	Chebysev	Euclidean StartEnd	DTW	EDR	LCSS
Ιεραρχική Ομαδοποίηση	Adding Noise Points	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Sensitive	Robust
	Increased Sampling rate	Sensitive	Robust	Robust	Sensitive	Sensitive	Sensitive	Fair
	Decreased Sampling rate	Sensitive	Sensitive	Fair	Sensitive	Sensitive	Sensitive	Fair
	Trajectory Synced Shift	Robust	Sensitive	Fair	Sensitive	Robust	Robust	Robust
	Trajectory Random Shift	Robust	Sensitive	Sensitive	Sensitive	Robust	Robust	Robust

Τα συμπεράσματα τα οποία προέκυψαν από την πειραματική μελέτη για τον αλγόριθμο optics και την ιεραρχική ομαδοποίηση παρουσιάζονται συγκεντρωτικά ως ακολούθως.

Ο αλγόριθμος Optics και ο αλγόριθμος της ιεραρχικής ομαδοποίησης με τη μέθοδο Ward, βάσει των κριτηρίων εσωτερικής και εξωτερικής αξιολόγησης, φαίνεται να έχουν περισσότερο αποδεκτή συμπεριφορά στους μετασχηματισμούς τροχιών της επαναδειγματοληψίας, της προσθήκης θορύβου και της μετατόπισης σημείου, με χρήση των μη μετρικών μέτρων ομοιότητας, σε σύγκριση με τις μετρικές συναρτήσεις απόστασης.

Ο αλγόριθμος Optics και ο αλγόριθμος της ιεραρχικής ομαδοποίησης με τη μέθοδο Ward, βάσει των κριτηρίων εσωτερικής και εξωτερικής αξιολόγησης, φαίνεται να είναι περισσότερο εύρωστοι στους μετασχηματισμούς τροχιών της επαναδειγματοληψίας, της προσθήκης θορύβου και της μετατόπισης σημείου, με χρήση της μη μετρικής συνάρτησης ομοιότητας Longest Common Subsequence (LCSS) σε σύγκριση με τις υπόλοιπες μη μετρικές συναρτήσεις απόστασης. Αναλυτικότερα, οι αλγόριθμοι με χρήση της LCSS παρουσιάζουν ευρωστία στη μετατόπιση σημείου και στην αύξηση ρυθμού δειγματοληψίας, ευαισθησία στη μείωση του ρυθμού δειγματοληψίας, ενώ για την προσθήκη θορύβου τα αποτελέσματα δεν είναι ξεκάθαρα.

Ο αλγόριθμος Optics και ο αλγόριθμος της ιεραρχικής ομαδοποίησης με τη μέθοδο Ward, βάσει των κριτηρίων εσωτερικής και εξωτερικής αξιολόγησης, φαίνεται να είναι περισσότερο εύρωστοι στους μετασχηματισμούς τροχιών της επαναδειγματοληψίας, της προσθήκης θορύβου και της μετατόπισης σημείου, με χρήση της μετρικής συναρτήσης απόστασης Chebyshev, σε σύγκριση με τις υπόλοιπες μετρικές συναρτήσεις απόστασης. Αναλυτικότερα, οι αλγόριθμοι με χρήση της Chebyshev παρουσιάζουν ευαισθησία στην προσθήκη θορύβου και στη μετατόπιση σημείου, ευρωστία στην αύξηση του ρυθμού δειγματοληψίας και αποδεκτή συμπεριφορα στη μείωση του ρυθμού δειγματοληψίας.

Αξίζει να σημειωθεί ότι σε πολλές από τις υπό εξέταση περιπτώσεις προέκυψαν αρκετά χαμηλές τιμές των χρησιμοποιούμενων δεικτών αξιολόγησης, οι οποίες υποδεικνύουν χαμηλή ισχύ (κακής ποιότητας ομαδοποιήσεις). Αναφέρεται ενδεικτικά ότι για τη συνάρτηση απόστασης LCSS παρατηρήθηκαν ιδιαίτερα χαμηλοί δείκτες αξιολόγησης (Dunn=0.08, Silhouette=-0.35, Rand=0.08, Jaccard=0.022) με χρήση του αλγορίθμου Optics, ενώ αντίθετα η μέθοδος της ιεραρχικής ομαδοποίησης σε αυτές τις περιπτώσεις βρέθηκε να αποδίδει καλύτερα, γεγονός που είναι αντιφατικό με αυτό που συναντάμε στη διεθνή βιβλιογραφία. Για παράδειγμα, ενώ αναμένετο η LCSS να χειρίζεται καλύτερα τον μετασχηματισμό της προσθήκης θορύβου, αντ’αυτού βρέθηκε να είναι ευαίσθητη σε αυτόν.

Συμπερασματικά, βλέποντας πιο σφαιρικά τα αποτελέσματα τα οποία προέκυψαν από την εκτενή πειραματική μελέτη, διαπιστώθηκε ότι δεν μπορούν να εξαχθούν ασφαλή συμπεράσματα για το ποια από τις υπό εξέταση μεθόδους (Optics & ιεραρχική ομαδοποίηση με τη μέθοδο Ward) υπερείχε ξεκάθαρα όσον αφορά την ισχύ των ομαδοποιήσεών της σε σύγκριση με την άλλη. Πιθανοί λόγοι που μπορεί να συμβαίνει κάτι τέτοιο είναι ο προκαθορισμός των παραμέτρων εισόδου από τον πειραματιστή, η δομή των χρησιμοποιούμενων συνόλων δεδομένων και η επιλογή ενός μέρους του συνόλου δεδομένων για λόγους υπολογιστικού χρόνου.

Βιβλιογραφία

[1] Agrawal R, Faloutsos C, Swami A, (1993), Efficient similarity search in sequence databases. In Proceedings of FODO.

[2] Ankerst M, Breunig M.M., Kriegel H-P., Sander J, (1999), OPTICS: Ordering Points to Identify the Clustering Structure. ACM SIGMOD international conference on Management of data.ACM Press. pp. 49–60.

[3] Berndt J, Clifford J, (1996), Finding patterns in time series: A dynamic programming approach. In Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press.

[4] Bollobas B, Das G, Gunopulos D, Mannila H, (2001), Time-series similarity problems and well-separated geometric sets. Nordic Journal of Computing, 8(4), pp. 409-423.

[5] Chen L, Ng RT, (2004), On the marriage of lp-norms and edit distance. In Proceedings of VLDB.

[6] Chen L, Ozsu MT, Oria V, (2005), Robust and fast similarity search for moving object trajectories. In Proceedings of SIGMOD. Clarke, F. (1976), Optimal solutions to differential inclusions, Journal of Optimization Theory and Applications 19 (3), 469-478.

[7] Ciaccia P, Patella M, Zezula, P, (1997), M-tree: An efficient access method for similarity search in metric spaces, in ‘VLDB’97’, Morgan Kaufmann Publishers, Inc., pp. 426–435.

[8] Demsar J, (2006), Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research 7, 1–30.

[9] Halkidi M, Batistakis Y, Vazirgiannis M, (2002a), Cluster Validity Methods: Part I, SIGMOD Record, Volume 31, Issue 2, pp. 40-45.

[10] Halkidi M, Batistakis Y, Vazirgiannis M, (2002b), Cluster Validity Checking Methods: Part II, SIGMOD Record, Volume 31 Issue 3, pp. 19-27.

[11] Jonkery R, De Leve G, Van Der Velde J, Volgenant A, (1980), `Rounding symmetric traveling salesman problems with an asymmetric assignment problem', Operations Research pp. 623-627.

[12] Kallitzaki S, (2014), Hermes.chorochronos.org: A Web Portal for the Analysis of Moving Object Trajectories, Msc Thesis, Univeristy of Peiraeus.

[13] Keogh E, Pazzani M, (2000), Scaling up dynamic time warping for datamining applications, in `Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining',ACM, pp. 285-289.

[14] Nanni M, Pedreschi D, (2006), Time-focused clustering of trajectories of moving objects, Journal of Intelligent Information Systems, 27: 267–289.

[15] Pelekis N, Theodoridis Y, (2014), Mobility Data Management and Exploration, Springer.

[16] Pelekis N, Theodoridis Y, Vosinakis S, and Panayiotopoulos T, (2006), Hermes-A Framework for Location-Based Data Management, In the Proceedings of the 10th International Conference on Extending Database Technology (EDBT06), LNCS 3896, pp. 1130-1134, Munich, Germany,Springer.

[17] Richalet J, Rault A, Testud J, Papon J, (1978), Model predictive heuristic control: Applications to industrial processes, Automatica 14(5), 413-428.

[18] Sakurai Y, Yoshikawa M, Faloutsos C, (2005), Ftw: fast similarity search under the time warping distance, in `Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems', ACM, pp. 326-337.

[19] Sanderson A, Wong A, (1980), Pattern trajectory analysis of nonstationary multivariate data, Systems, Man and Cybernetics, IEEE Transactions on 10(7), 384-392.

[20] Takens F, (1980), `Motion under the inuence of a strong constraining force', Global theory of dynamical systems pp. 425-445.

[21] Vlachos M, Gunopulos D, Das G, (2002a), Rotation invariant distance measures for trajectories. In Proceedings of SIGKDD.

[22]Vlachos M, Gunopulos D, Kollios G, (2002b), Discovering similar multidimensional trajectories. In Proceedings of ICDE.

[23] Wang H, Su H, Zheng K, Sadiq S, Zhou X-F, (2013), An Effectiveness Study on Trajectory Similarity Measures, Proceedings of the Twenty-Fourth Australasian Database Conference (ADC 2013), CRPIT Volume 137, pp. 13-22, Adelaide, Australia.

[24] Yi B, Jagadish H, Faloutsos C, (1998), Efficient retrieval of similar time sequences under time warping, in `Data Engineering, 1998. Proceedings 14th International Conference on', IEEE, pp. 201-208.

[25] https://en.wikipedia.org/wiki/Cluster_analysis#Evaluation_and_assessment