ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Χρήση Τεχνικών Ομαδοποίησης στον Επιχειρηματικό Σχεδιασμό

Χρυσούλα Γ. Ρίνου

Διπλωματική Εργασία

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

Πειραιάς
Σεπτέμβριος 2021

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

Τα μέλη της Επιτροπής ήσαν:

- Καθηγητής Μ. Κούτρας (Επιβλέπων)
- Καθηγήτρια Γ. Βερροπούλου
- Αναπλ. Καθηγητής Ο. Τήνιος

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

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Application of cluster analysis techniques in business analytics

By

Chrysoula G. Rinou

MSc Dissertation

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

Piraeus, Greece
September 2021

Στους γονείς μου
Γιώργο και Βαλεντίνα

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της διπλωματικής εργασίας μου, κ. Μάρκο Κούτρα, για την γνώση, την έμπνευση αλλά και τις πολύτιμες συμβουλές που μου παρείχε σε όλη την διάρκεια των σπουδών μου στο Μεταπτυχιακό Πρόγραμμα, καθώς και την περίοδο της συγγραφής της παρούσας διπλωματικής εργασίας.

Θα ήθελα να ευχαριστήσω και τα μέλη της τριμελούς κ. Βερροπούλου και κ. Τήνιο για τις χρήσιμες παρατηρήσεις τους πάνω στην εργασία καθώς και στην παρουσίαση της.

Επιπλέον θα ήθελα επίσης να ευχαριστήσω τους καθηγητές του προγράμματος μεταπτυχιακών σπουδών για τις γνώσεις που μου μετέδωσαν ο καθένας ξεχωριστά. Φυσικά δεν θα μπορούσα να παραλείψω τους συμφοιτητές μου, οι οποίοι μου προσέφεραν ένα περιβάλλον στο οποίο μπόρεσα να αναπτυχθώ ακόμα περισσότερο .

Τέλος, ευχαριστώ τους γονείς μου για όλη την υποστήριξη τους και την ενθάρρυνση τόσο στις προπτυχιακές μου όσο και στις μεταπτυχιακές μου σπουδές.

Περίληψη

Ζούμε στην εποχή όπου τα δεδομένα μεγάλου όγκου (Big Data) υπάρχουν σε πολλούς τομείς, όπως και στον τομέα του επιχειρηματικού σχεδιασμού. Παράλληλα, το στρατηγικό πλαίσιο σε ένα επιχειρηματικό πλάνο είναι πολύ κρίσιμο για την επιτυχία του. Προς αυτή την κατεύθυνση, η Μηχανική Μάθηση (Machine Learning) μπορεί να συνεισφέρει στον αποδοτικότερο σχεδιασμό ενός επιχειρηματικού πλάνου. Μέσα από αλγορίθμους μηχανικής μάθησης μας επιτρέπεται η εξόρυξη γνώσης από δεδομένα, η οποία μας δίνει τη δυνατότητα για μια πιο ρεαλιστική απεικόνιση των δεδομένων που έχουμε.

Ένας βασικός πυλώνας της Μηχανικής Μάθησης είναι η μη επιβλεπόμενη μάθηση, όπου εκεί κυριαρχούν οι αλγόριθμοι ομαδοποίησης δεδομένων. Σε περιπτώσεις που έχουμε δεδομένα χωρίς ετικέτα (label) ένας τέτοιος αλγόριθμος μπορεί να αποτυπώσει καλύτερα τη δομή των δεδομένων αυτών έτσι ώστε να βοηθήσει στον καλύτερο σχεδιασμό ενός επιχειρηματικού πλάνου.

Στην παρούσα διπλωματική εργασία, γίνεται η μελέτη παραδοσιακών αλγορίθμων ομαδοποίησης και μια συγκριτική αξιολόγηση σε δεδομένα πραγματικού κόσμου που αφορούν προβλήματα επιχειρησιακού σχεδιασμού. Τα αποτελέσματα έδειξαν την υπεροχή των αλγορίθμων διαχωρισμού (partitioning) και αναδεικνύουν τα προβλήματα που προκύπτουν στην περίπτωση που εφαρμοσθεί ένας αλγόριθμος που δεν ταιριάζει σε συγκεκριμένο σύνολο δεδομένων και συγκεκριμένα χαρακτηριστικά.

Abstract

We are in the Big Data era which exists in many areas, as well as in the field of business planning. At the same time, the strategic framework in a business plan is very critical to its success. In this regard, Machine Learning can contribute to the more efficient design of a business plan.

Through machine learning algorithms we can extract knowledge from data, allowing for a more realistic data analysis. A key pillar of Machine Learning is unsupervised learning, where data clustering algorithms predominate. In cases where we have data without a label (or class), clustering techniques can better capture the structure of this data offering a better business plan design.

In this thesis, we compare traditional clustering algorithms in real world data related to business planning problems. The results showed the superiority of partitioning algorithms and highlight the limitations that arise when we apply the wrong algorithm to a case under study.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Γλωσσάρι

Όρος	Περιγραφή
Διαχωριστική Ομαδοποίηση	Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα -non-overlapping - υποσύνολα (ομάδες) τέτοιος ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο
Ιεραρχική Ομαδοποίηση	Ένα σύνολο από εμφωλευμένες (nested) ομάδες. Επιτρέπουμε σε μια ομάδα να έχει υπο-ομάδες οργανωμένες σε ένα ιεραρχικό δέντρο
Μηχανική Μάθηση	Είναι μια περιοχή της τεχνητής νοημοσύνης η οποία ασχολείται με την μελέτη αλγορίθμων που επιτρέπουν στους υπολογιστές να «μαθαίνουν» μέσα από τα δεδομένα τους.
Ομαδοποίηση	Είναι η διαδικασία της κατηγοριοποίησης των δεδομένων σε σύνολα ομοειδών αντικειμένων καλούμενα ομάδες

Εισαγωγή

Κάθε επιχείρηση έχει ως κύριο στόχο την επίτευξη των επιθυμητών αποτελεσμάτων αναφορικά με την ανάπτυξη και την επιλογή στρατηγικών σε ένα δεδομένο σενάριο στρατηγικής διαχείρισης. Για κάθε επιχείρηση, η επιλογή στρατηγικής είναι ένα σημαντικό έργο, από την απόφαση του οποίου εξαρτάται η περαιτέρω δραστηριότητά του. Επίσης, μια στρατηγική είναι ένα σύνολο κανόνων για τη λήψη αποφάσεων που καθοδηγούν έναν οργανισμό στις δραστηριότητές του. Κάθε μία από τις στρατηγικές, που επιλέγει μια επιχείρηση, οδηγεί σε διαφορετικά αποτελέσματα. Συνεπώς, η επιλογή της στρατηγικής θα πρέπει να γίνεται λαμβάνοντας υπόψη όλες τις διαθέσιμες πληροφορίες στο στάδιο της λήψης αποφάσεων. Όταν τα δεδομένα που έχει να διαχειριστεί η επιχείρηση είναι λίγα, τότε οι εναλλακτικές επιλογές δεν είναι πολλές, οπότε η πιθανότητα να δυσκολευτεί στη στρατηγική διαχείριση της επιχείρησης, είναι μικρή. Όταν όμως τα δεδομένα είναι πολλά, τότε γίνεται απαραίτητη η χρήση των υπολογιστικών μοντέλων.

Σε αυτή την κατεύθυνση είναι η Μηχανική Μάθηση, η οποία είναι στην ευρύτερη οικογένεια της Τεχνητής νοημοσύνης και της Υπολογιστικής Νοημοσύνης. Μέρος της μηχανικής μάθησης, αποτελεί η ομαδοποίηση δεδομένων, η οποία ανήκει στην κατηγορία μη επιβλεπόμενης μάθησης. Ως ομαδοποίηση ορίζεται η διαδικασία ομαδοποίησης παρατηρήσεων παρόμοιου είδους σε μικρότερες ομάδες εντός του μεγαλύτερου πληθυσμού. Έχει ευρεία εφαρμογή στην ανάλυση επιχειρήσεων. Η ανάλυση δεδομένων μέσω της ομαδοποίησης είναι ένα εργαλείο ανάλυσης δεδομένων που στοχεύει στην ταξινόμηση διαφορετικών αντικειμένων σε ομάδες με τρόπο που ο βαθμός συσχέτισης μεταξύ δύο αντικειμένων να είναι μέγιστος εάν ανήκουν στην ίδια ομάδα και ελάχιστος εάν όχι.

Η ανάγκη της ομαδοποίησης στον επιχειρηματικό σχεδιασμό γίνεται πιο έντονη όταν τα δεδομένα που έχουμε να διαχειριστούμε είναι χωρίς ταμπέλα ή σήμανση ή κλάση (label or class). Αυτό δε μας επιτρέπει να χρησιμοποιήσουμε μεθόδους επιβλεπόμενης μάθησης (κατηγοριοποίηση ή παλινδρόμηση), ωστόσο η ομαδοποίηση δεδομένων μπορεί να συνεισφέρει σημαντικά. Πέρα από την εύρεση ομάδων όπου μέσω αντιπροσώπων (κέντρα του cluster) έχουμε μια καλύτερη απεικόνιση των δεδομένων, μέσω των αλγορίθμων ομαδοποίησης μπορούμε να εντοπίσουμε πρότυπα (pattern recognition) και να ανιχνεύσουμε σημεία ανωμαλιών (outliers). Στην παρούσα εργασία γίνεται η μελέτη διαφόρων αλγορίθμων ομαδοποίησης σε πραγματικά δεδομένα επιχειρηματικού σχεδιασμού.

Πιο συγκεκριμένα στο 1ο κεφάλαιο περιγράφονται εισαγωγικές έννοιες του επιχειρηματικού και του στρατηγικού σχεδιασμού καθώς επίσης στο 2ο κεφάλαιο αναφέρονται τα σενάρια στον επιχειρηματικό σχεδιασμό. Στο επόμενο κεφάλαιο περιγράφονται οι τεχνικές αξιολόγησης κινδύνων στον επιχειρηματικό σχεδιασμό. Το 4ο κεφάλαιο ακολουθεί με τη βασική μας μελέτη, η οποία αφορά την εφαρμογή τεχνικών ομαδοποίησης στον επιχειρηματικό σχεδιασμό. Υλοποιούνται και συγκρίνονται γνωστοί αλγόριθμοι ομαδοποίησης σε δύο πραγματικά προβλήματα επιχειρηματικού σχεδιασμού. Στο τελευταίο κεφάλαιο αναλύονται και ερμηνεύονται τα αποτελέσματα καθώς επίσης δίνεται και η μελλοντική κατεύθυνση για την επέκταση της παρούσας πτυχιακής.

Κεφάλαιο 1. Επιχειρηματικός σχεδιασμός

1.1. Εισαγωγή

Ένα επιχειρηματικό σχέδιο είναι ένα γραπτό έγγραφο που περιγράφει λεπτομερώς τον τρόπο με τον οποίο μια επιχείρηση - συνήθως μια νεοσύστατη εταιρεία - καθορίζει τους στόχους της και πώς πρόκειται να τους επιτύχει. Ένα επιχειρηματικό σχέδιο καθορίζει έναν γραπτό χάρτη πορείας για την εταιρεία από εμπορική, οικονομική και επιχειρησιακή άποψη (βλ. Covello & Hazelgren, 2006).

Τα επιχειρηματικά σχέδια είναι σημαντικά έγγραφα που χρησιμοποιούνται για την προσέλκυση επενδύσεων προτού μια εταιρεία έχει δημιουργήσει ένα αποδεδειγμένο ιστορικό. Είναι επίσης ένας καλός τρόπος για τις εταιρείες να διατηρούν τον στόχο τους στο μέλλον. Αν και είναι ιδιαίτερα χρήσιμα για νέες επιχειρήσεις, κάθε εταιρεία πρέπει να έχει επιχειρηματικό σχέδιο. Στην ιδανική περίπτωση, το σχέδιο επανεξετάζεται και ενημερώνεται περιοδικά για να διαπιστωθεί εάν οι στόχοι έχουν επιτευχθεί ή έχουν αλλάξει και εξελιχθεί. Μερικές φορές, δημιουργείται ένα νέο επιχειρηματικό σχέδιο για μια καθιερωμένη επιχείρηση που έχει αποφασίσει να κινηθεί προς μια νέα κατεύθυνση (βλ. Anderson, (2020)).

Η έκταση του επιχειρηματικού σχεδίου ποικίλλει σε μεγάλο βαθμό από επιχείρηση σε επιχείρηση. Όλες οι πληροφορίες προτείνεται να έχουν μέγεθος από 15 έως 20 σελίδες. Εάν υπάρχουν κρίσιμα στοιχεία του επιχειρηματικού σχεδίου που καταλαμβάνουν πολύ χώρο - όπως αιτήσεις για διπλώματα ευρεσιτεχνίας - θα πρέπει να αναφέρονται στο κύριο σχέδιο και να περιλαμβάνονται ως παραρτήματα. Ποτέ δεν υπάρχουν δύο ίδια επιχειρηματικά σχέδια. Όλοι όμως έχουν τα ίδια στοιχεία. Παρακάτω είναι μερικά από τα κοινά και βασικά μέρη ενός επιχειρηματικού σχεδίου.

Σύμφωνα με τους Covello and Hazelgren (2006) εκτιμάται ότι πάνω από ένα εκατομμύριο νέες επιχειρήσεις ξεκινούν κάθε χρόνο στην Αμερική, από μικρής κλίμακας εγχώριες επιχειρήσεις μέχρι μεγάλες επιχειρήσεις που απαιτούν πολλά εκατομμύρια δολάρια αρχικού κεφαλαίου. Από όλες αυτές τις νέες επιχειρήσεις, αναμένεται ότι μόνο μία στις πέντε θα φτάσει στην πέμπτη επέτειο και θα επιτύχει τα αποτελέσματα που είχαν αρχικά προβλέψει κατά την εκκίνηση.

Αυτό αποτελεί ανησυχητικό στατιστικό στοιχείο καθώς τα δεδομένα αυτά αναφέρονται στην κατεξοχήν χώρα της ιδιωτικής πρωτοβουλία Αν και υπάρχουν πολλοί λόγοι για τους οποίους κάτι τέτοιο μπορεί να συμβεί ωστόσο, ο πιο κοινός λόγος τυγχάνει να είναι η αποτυχία σχεδιασμού των κινήσεων της επιχείρησης. Ένα επιχειρηματικό σχέδιο βοηθά τους

επιχειρηματίες και τους διευθυντές να σκεφτούν τις στρατηγικές τους, να εξισορροπήσουν τον ενθουσιασμό τους με τα γεγονότα και να αναγνωρίσουν τους περιορισμούς τους. Και φυσικά επιτρέπει την αποφυγή ενδεχομένως καταστροφικών σφαλμάτων όπως η ελλιπής κεφαλαιοποίηση, η δημιουργία αρνητικών ταμειακών ροών, η πρόσληψη λανθασμένων ατόμων, η επιλογή λανθασμένης τοποθεσίας, η υποτίμηση του ανταγωνισμού και η στόχευση λάθους αγοράς.

Ένα επιχειρηματικό σχέδιο που επιτυγχάνει, απαιτεί χρόνο και σύμφωνα με τους Covello and Hazelgren (2006) η κατασκευή ενός ολοκληρωμένου επιχειρηματικού σχεδίου απαιτεί 50 έως 150 ώρες ασχολίας συμπεριλαμβανομένης της έρευνας, της τεκμηρίωσης, της ανάλυσης και της αναθεώρησης. Οι επιχειρηματίες θα πρέπει να αρχίσουν να σχεδιάζουν τουλάχιστον έξι μήνες προτού ανοίξουν ή επεκτείνουν μια επιχείρηση.

Οι περισσότεροι επιχειρηματίες θα πρέπει να αφιερώσουν χρόνο στην εκκίνηση κατάρτισης ενός επιχειρηματικού σχεδίου πριν ακόμη την έναρξη των επιχειρηματικών τους δραστηριοτήτων, στην περίπτωση που εργάζονται σε κάποια άλλη επιχείρηση ή την συνέχιση των δραστηριοτήτων τους εάν του επιχειρηματικό σχέδιο αφορά την διεύρυνση των δραστηριοτήτων της υφιστάμενης επιχείρησης τους. Έξι μήνες θεωρούνται αρκετοί για την εστίαση τους στις νέες επιχειρηματικές ιδέες και έννοιες.

Η μέχρι στιγμής περιγραφή του επιχειρηματικού σχεδιασμού ίσως υπονοεί μια θεωρητική εργασία η οποία ενεργοποιείται μετά από ένα καταιγισμό ιδεών (brainstorming) και συνεχίζεται με την τακτοποίηση και εφαρμογή τους. Παρόλα αυτά, είναι μια αυστηρή και τυποποιημένη διαδικασία που περιλαμβάνει τις εξής ενέργειες (βλ. Covello & Hazelgren, (2006); Anderson, (2020))

- Την εκτίμηση των δυνατών σημείων και των αδυναμιών της προσπάθειας.
- Τη διευθέτηση στόχων. Οι στόχοι αυτοί θα πρέπει να είναι
    - Οικονομικοί π.χ. κέρδος επιχείρησης, απώλειες κλπ.
    - Προσωπικοί π.χ. το όραμα της εκκίνησης μιας επιχειρηματικής προσπάθειας και
    - Απόσυρσης. Αυτοί μπορούν να αναφέρονται στην απόσυρση από την ενεργό δράση για λόγους όπως η ηλικία, η επίτευξη ή μη των αναμενόμενων στόχων κ.α.

Όλες οι προηγούμενες παράμετροι θα πρέπει να ποσοτικοποιηθούν ώστε να γίνουν συγκρίσιμα μεγέθη και να εκτελεστούν ανάλογες αναλύσεις.

Οι κύριοι λόγοι για την κατασκευή ενός επιχειρηματικού σχεδίου μπορεί να είναι οι παραάτω τρείς (βλ. Horan (2008); Covello and Hazelgren (2006); Berry (2006))

i. *Εξασφάλιση χρηματοδότησης από επενδυτές.*

Δεδομένου ότι το περιεχόμενο του σχεδίου περιστρέφεται γύρω από το πώς οι επιχειρήσεις πετυχαίνουν, εξισορροπούν και αποκομίζουν κέρδος, ένα επιχειρηματικό σχέδιο χρησιμοποιείται κυρίως ως εργαλείο για την άντληση κεφαλαίων. Αυτό το έγγραφο είναι ένας τρόπος του επιχειρηματία για να δείξει στους πιθανούς επενδυτές ή δανειολήπτες πώς θα λειτουργήσει το κεφάλαιό τους και πώς θα βοηθήσει μια επιχείρηση να αναπτυχθεί. Όλες οι τράπεζες, οι επενδυτές και οι εταιρείες επιχειρηματικών κεφαλαίων θα θέλουν να δουν ένα επιχειρηματικό σχέδιο πριν παραδώσουν τα χρήματά τους και οι επενδυτές συνήθως θέλουν μια απόδοση 20-25% από το κεφάλαιο που επενδύουν σε μια επιχείρηση. Επομένως, αυτοί οι επενδυτές πρέπει να γνωρίζουν εάν - και πότε - θα κερδίσουν τα χρήματά τους πίσω (και έπειτα μερικοί). Επιπλέον, θα θελήσουν να διαβάσουν τη διαδικασία και τη στρατηγική για το πώς η επιχείρηση θα επιτύχει αυτούς τους οικονομικούς στόχους, όπου είναι το πλαίσιο που παρέχεται από τα σχέδια πωλήσεων, μάρκετινγκ και λειτουργίας.

ii. *Τεκμηρίωση της στρατηγικής και των στόχων μιας εταιρείας.*

Τα επιχειρηματικά σχέδια μπορούν να εκτείνονται σε δεκάδες ή και εκατοντάδες σελίδες, δίνοντας στους συντάκτες τους την ευκαιρία να εξηγήσουν ποιοι είναι οι στόχοι μιας επιχείρησης και πώς θα τους επιτύχει η επιχείρηση. Για να δείξουν στους πιθανούς επενδυτές ότι έχουν αντιμετωπίσει κάθε ερώτηση και σκέψη σε κάθε πιθανό σενάριο, οι επιχειρηματίες πρέπει να εξηγήσουν διεξοδικά τις στρατηγικές μάρκετινγκ, πωλήσεων και λειτουργίας τους - από την απόκτηση μιας φυσικής θέσης για την επιχείρηση έως την εξήγηση μιας τακτικής προσέγγισης για τη διείσδυση στο μάρκετινγκ. Αυτές οι διευκρινήσεις θα πρέπει ουσιαστικά να οδηγήσουν στην μεταβολή του στρατηγικού σχεδιασμού μιας επιχείρησης που θα πρέπει να υποστηρίζεται από προβλέψεις πωλήσεων και οικονομικές προβλέψεις, και ο συγγραφέας του επιχειρηματικού σχεδίου να είναι σε θέση να δώσει περισσότερες λεπτομέρειες για οτιδήποτε περιγράφεται στο σχέδιο αυτό.

1.2. Στρατηγικός σχεδιασμός

Ο επιχειρηματικός σχεδιασμός ίσως να φαίνεται περισσότερο ως έκθεση ιδεών για μια ιδέα που αφορά το πρώτο βήμα προς την εφαρμογή μιας ιδέας ή την εφαρμογή μιας καινοτομίας σε μια υπάρχουσα επιχείρηση. Για αυτό τον λόγο, και στην συνέχεια της εργασίας, ο επιχειρηματικός σχεδιασμός θα αναφέρεται και θα είναι συνώνυμος του στρατηγικού σχεδιασμού.

Ο στρατηγικός σχεδιασμός είναι η διαδικασία τεκμηρίωσης και καθορισμού της κατεύθυνσης μια επιχείρηση που αξιολογεί την παρούσα και εκτιμά την μελλοντική κατάστασή της (βλ. Vo (2020)). Το στρατηγικό σχέδιο αφορά την καταγραφή της αποστολής, του οράματος και των αξιών της επιχείρησης, και θέτει μακροπρόθεσμους στόχους και σχέδια δράσης που θα χρησιμοποιηθούν για την επίτευξη τους. Ένα καλογραμμένο στρατηγικό σχέδιο μπορεί να διαδραματίσει καθοριστικό ρόλο στην ανάπτυξη και την επιτυχία μιας επιχείρησης, διότι καθοδήγει και βοηθά στην ανταπόκριση στις επιχειρηματικές ευκαιρίες και προκλήσεις.

Αν και ουσιαστικά οι δύο έννοιες έχουν παρόμοια σημασία, υπάρχουν κάποιες σημαντικές διακρίσεις μεταξύ του επιχειρηματικού και του στρατηγικού σχεδιασμού που αξίζει να σημειωθούν (βλ. Brunings (2020)).

- Ένα στρατηγικό σχέδιο χρησιμοποιείται κυρίως για την εφαρμογή και τη διαχείριση της στρατηγικής κατεύθυνσης ενός υπάρχοντος οργανισμού. Ένα επιχειρηματικό σχέδιο χρησιμοποιείται συνήθως για την έναρξη μιας επιχείρησης, τη λήψη χρηματοδότησης ή την άμεση λειτουργία. Τα δύο σχέδια καλύπτουν επίσης διαφορετικά χρονικά πλαίσια. Ένα στρατηγικό σχέδιο καλύπτει γενικά μια περίοδο από 3 έως 5+ έτη, ενώ ένα επιχειρηματικό σχέδιο συνήθως δεν υπερβαίνει το ένα έτος.
- Ένα στρατηγικό σχέδιο είναι για καθιερωμένες επιχειρήσεις, οργανισμούς και ιδιοκτήτες επιχειρήσεων που είναι προσηλωμένοι για την ανάπτυξη της οργάνωσής τους. Ενώ ένα επιχειρηματικό σχέδιο θα μπορούσε να είναι για νέες επιχειρήσεις και επιχειρηματίες που είναι νεοσύστατες επιχειρήσεις.
- Ένα στρατηγικό σχέδιο χρησιμοποιείται για να παρέχει εστίαση, κατεύθυνση και δράση για να μεταφέρει την επιχείρηση από το σημείο που βρίσκεται τώρα στο

τελικό σημείο-στόχο. Ενώ ένα επιχειρηματικό σχέδιο χρησιμοποιείται για την παροχή μιας δομής ιδεών για τον αρχικό ορισμό της επιχείρησης.

- Ένα στρατηγικό σχέδιο είναι ζωτικής σημασίας για την ιεράρχηση των πόρων (χρόνος, χρήμα και άνθρωποι), για την αύξηση των εσόδων και την αύξηση της απόδοσης της επένδυσης. Ενώ ένα επιχειρηματικό σχέδιο είναι κρίσιμο εάν η επιχείρηση αναζητά χρηματοδότηση.
- Ένα στρατηγικό σχέδιο επικεντρώνεται στην οικοδόμηση ενός βιώσιμου ανταγωνιστικού πλεονεκτήματος και έχει φουτουριστικό χαρακτήρα. Ενώ ένα επιχειρηματικό σχέδιο χρησιμοποιείται για την εκτίμηση της βιωσιμότητας μιας επιχειρηματικής ευκαιρίας, και έχει πιο τακτικό χαρακτήρα.
- Ένα στρατηγικό σχέδιο χρησιμοποιείται για να κοινοποιήσει την κατεύθυνση του οργανισμού στο προσωπικό και τους ενδιαφερόμενους. Ωστόσο, ένα επιχειρηματικό σχέδιο χρησιμοποιείται για την παρουσίαση των ιδεών του επιχειρηματία σε μια τράπεζα.
- Μεγαλύτεροι οργανισμοί με πολλαπλές επιχειρηματικές μονάδες και μεγάλη ποικιλία προϊόντων ξεκινούν συχνά τη διαδικασία ετήσιου προγραμματισμού τους με στρατηγικό σχέδιο. Συχνά ακολουθούνται από τμήματα και σχέδια μάρκετινγκ που απορρέουν από το Στρατηγικό Σχέδιο. Οι μικρότερες εταιρείες και οι νεοσύστατες εταιρείες συνήθως χρησιμοποιούν μόνο ένα επιχειρηματικό σχέδιο για να αναπτύξουν όλες τις πτυχές της επιχείρησης σε χαρτί, να λάβουν χρηματοδότηση και μετά να ξεκινήσουν την επιχείρηση. Πολλές μικρότερες εταιρείες - συμπεριλαμβανομένων των νεοσύστατων επιχειρήσεων, δεν αναπτύσσουν ποτέ Στρατηγικό Σχέδιο.

Στην παρούσα εργασία, στην εξέταση των μεθόδων του επιχειρηματικού σχεδιασμού θεωρήθηκε ότι οι επιχειρήσεις δεν είναι νεοσύστατες και ότι βασικός σκοπός είναι η αύξηση του κέρδους και των δραστηριοτήτων της. Αυτή η θεώρηση ήταν αναγκαία καθώς μια νεοσύστατη επιχείρηση θα εστιάσει στην οικονομική βιωσιμότητά της μέσα από την άντληση κεφαλαίων από τρίτους και όχι μέσα από διάφορες δραστηριότητες.

Κεφάλαιο 2. Τα σενάρια στον επιχειρηματικό σχεδιασμό

2.1. Εισαγωγή

Καθώς ο ρυθμός των αλλαγών στην εποχή μας είναι αυξανόμενος (και ιδιαίτερα των τεχνολογικών αλλαγών), εμφανίζονται συνεχώς νέες τεχνολογίες, οι οποίες αντικαθιστούν παλαιότερες μεθόδους άμεσα (νέες μεθοδολογίες) ή έμμεσα π.χ. νέα προϊόντα υποστήριξης όπως η τεχνητή νοημοσύνη. Παρόμοια εμφανίζονται μη αναμενόμενοι ανταγωνιστές ενώ οι οικονομικές συνθήκες που επηρεάζουν την κερδοφορία γίνονται όλο και πιο απρόβλεπτες. Αυτός ο γρήγορος ρυθμός αλλαγής έχει καταστήσει όλο και πιο δύσκολη την πρόβλεψη του μέλλοντος, ειδικά τις μακροπρόθεσμες προβλέψεις.

Παρόλα αυτά τα στελέχη των επιχειρήσεων, θα πρέπει να λάβουν αποφάσεις για τους οργανισμούς τους που έχουν μακροπρόθεσμες επιπτώσεις. Έτσι, γίνεται σαφές ότι χρειάζονται νέες τεχνικές για να επιτρέπουν στις επιχειρήσεις και τα ανώτερα στελέχη τους να ενεργούν με αυτοπεποίθηση σε ένα γρήγορα μεταβαλλόμενο περιβάλλον.

Τα σενάρια είναι μια τεχνική που χρησιμοποιείται από επιχειρήσεις και κυβερνητικές οργανώσεις και μέσω αυτών γίνεται συστηματική διερεύνηση της αβεβαιότητας κατά εξέταση μακροπρόθεσμων σχεδίων. Η διαδικασία αυτή περιλαμβάνει την ανάπτυξη ενός συνόλου εναλλακτικών περιγραφών για το μέλλον που διαφέρουν σε βασικούς παράγοντες όπως η τεχνολογία, η συμπεριφορά και οι προσδοκίες των καταναλωτών κ.α. Ο στόχος της ανάπτυξης αυτών των πολλαπλών σεναρίων δεν είναι η βελτίωση της πιθανότητας συγκεκριμένης έκβασης αλλά η κατανόηση των κινητήριων δυνάμενων που επηρεάζουν μια μελλοντική κατάσταση. Με την κατανόηση και την αναγνώριση αυτών των κινητήριων δυνάμεων, αυξάνεται η ικανότητα των διαχειριστών να σχεδιάζουν ενναλακτικές προτάσεις ενός επιχειρηματικού σχεδίου.

2.2. Κατασκευή σεναρίων

Η ώθηση για την προέλευση και την ανάπτυξη σεναρίων ήταν οι αποτυχίες πολλών στρατηγικών αποφάσεων που βασίζονταν σε μελλοντικές προβλέψεις, ιδιαίτερα όταν οι προβλέψεις αυτές αφορούσαν ειδικά καθοριστικούς παράγοντες που επηρέαζαν τις συνέπειες αυτών των αποφάσεων. Από τη δεκαετία του '70 η ανάπτυξη σεναρίων έχει υιοθετηθεί από πολλές και μεγάλες εταιρείες. Η εταιρεία Shell ήταν πρωτοπόρος της προσέγγισης σεναρίων

για την λήψη στρατηγικών αποφάσεων. Ο Heiden (1999) επισημαίνει ότι από τη δεκαετία του '80 η εταιρεία Shell έχει παγιώσει την αξιολόγηση κάθε σημαντικού σχεδίου (project) με βάση το σύνολο των σεναρίων που αναπτύσσονται από τα στελέχη της. Από τότε τα σενάρια έχουν γίνει ένα σημαντικό εργαλείο της δημιουργικής ανάπτυξης της σκέψης κατά την εξέταση των δυνατοτήτων μιας μελλοντικής απόφασης-ανάπτυξης της εταιρείας. Ομοίως, τα σενάρια αποτέλεσαν ένα σημαντικό εργαλείο υποστήριξης αποφάσεων στρατηγικού σχεδιασμού.

Πλέον, η κατασκευή σεναρίων με σύγχρονες μεθόδους αποτελεί συνήθη διαδικασία για την εξασφάλιση της ευελιξίας του στρατηγικού σχεδίου αλλά και της ανάπτυξης αντανακλαστικών μιας επιχείρησης σε καταστάσεις, που υπό άλλες συνθήκες θα θεωρούταν απρόβλεπτες με αποτέλεσμα την ατελή αντιμετώπιση τους, εάν υπήρχε χρονικό περιθώριο για τέτοιου είδους ελιγμούς. Η επεξεργασία πληροφοριών επιχειρηματικού περιβάλλοντος, γνωστή ως Επιχειρηματική Ευφυΐα (Business Intelligence), πλέον καθιερώνεται και γίνεται αναγκαιότητα. Τα στρατηγικά σενάρια περιέχουν προσεγγίσεις στρατηγικού σχεδιασμού (Strategic Planning), διαχείρισης κινδύνων (Risk Management) και διαχειριστικής απόφασης (Managerial Decision). Ταυτόχρονα, αυξάνονται οι απαιτήσεις του διευθύνοντα συμβούλου (CEO) προς το προσωπικό του. Πλέον, εκτός από την εμπειρία και τη διαίσθηση απαιτείται ευχέρεια και ικανότητα των στελεχών να μπορούν να διαχειριστούν την πληροφορία.

2.3. Η έννοια του σεναρίου

Η έννοια των σεναρίων δεν είναι ενιαία. Οι διαφορές μεταξύ τους οφείλονται στον βαθμό χρήσης της χρήσης των διαφορών πτυχών-παραμέτρων στην λήψη αποφάσεων με βάση τα σενάρια, αλλά και την αρτιότητα κατασκευής των σεναρίων. Έτσι, για παράδειγμα δυο εταιρείες που έχουν αναπτύξει το ίδιο σενάριο μπορεί να έχουν διαφορετικό αποτέλεσμα στην περίπτωση εφαρμογής του, έχοντας διαφορετικούς τρόπους αντίδρασης στην περίπτωση εμφάνισης της κατάστασης που προβλέπει το σενάριο. Ουσιαστικά τα σενάρια αποτελούν εργαλείο του επιχειρηματικού σχεδιασμού και ασχολούνται με την εξέταση παραγόντων εξέλιξης μια τρέχουσας ή μελλοντικής επιχειρηματικής κίνησης προσπαθώντας παράλληλα να ελαχιστοποιήσουν την επίδραση από τις εσωτερικές διαδικασίες της επιχείρησης, δηλαδή την πιθανή αλλοίωση της έκβασης ή της πιθανότητας εμφάνισης ενός γεγονότος εξαιτίας κακών χειρισμών της εταιρείας. Αποτελούν δηλαδή ανεξάρτητα γεγονότα που εξετάζουν (ή

προσπαθούν να εξετάσουν) μόνο εξωτερικές επιδράσεις στην έκβαση μιας επιχειρηματικής κίνησης (βλ. Conway (2004)).

Σύμφωνα με τον Gugan (2008), τα σενάρια παρέχουν εναλλακτικές απόψεις για την έκβαση ενός γεγονότος. Προσδιορίζουν ορισμένα θεμελιώδη γεγονότα, βασικούς παράγοντες και τα κίνητρά τους και προσφέρουν διαφορετικές προοπτικές της μελλοντικής ανάπτυξης σε τοπικό ή/και παγκόσμιο επίπεδο. Συνεπώς, η ανάπτυξη και η εφαρμογή σεναρίων συμβάλλουν στην αναζήτηση τρόπων εμφάνισης μελλοντικών καταστάσεων, στην πιθανή έκβαση τους και στην αντιμετώπιση ή εκμετάλλευσή τους. Σύμφωνα με τους Schoemaker and Gunther (2002), τα σενάρια αντιπροσωπεύουν εσωτερικά συνεπείς εικόνες του μέλλοντος, οι οποίες βασίζονται σε μια συγκεκριμένη ομάδα αμοιβαία αλληλοσυνδεόμενων παραγόντων τόσο ποιοτικού όσο και ποσοτικού χαρακτήρα. Το σημείο εκκίνησης του σχηματισμού τους θα πρέπει να είναι η εξειδίκευση των όσων γνωρίζουμε για τη μελλοντική ανάπτυξη, οι τάσεις (το πιθανό) και οι αβεβαιότητες (το μη πιθανό). Κάθε σενάριο στη συνέχεια βασίζεται στη διασύνδεση αυτών των τάσεων και αβεβαιοτήτων. Σύμφωνα με τον Foster (1993) τα σενάρια αντιπροσωπεύουν μια συγκεκριμένη εικόνα του μέλλοντος που συνδυάζει ποιοτικά και ποσοτικά χαρακτηριστικά. Σύμφωνα με τους Pearson and Lyons (1999), ο σχηματισμός σεναρίων προέρχεται από παράγοντες που είναι τόσο αβέβαιοι και έχουν βασικό αντίκτυπο στο σύστημα. Σύμφωνα με τους Tessun and Hermann (1999), τα σενάρια προσδιορίζουν τις βασικές κινητήριες δυνάμεις της ανάπτυξης, συμπεριλαμβανομένων των αμοιβαίων σχέσεων τους, οι οποίες συνδέονται περαιτέρω με τις υπάρχουσες ευκαιρίες και κινδύνους. Ο Van der Heijden (1999), υποστηρίζει την ίδια έννοια για τον ορισμό των σεναρίων. Τα σενάρια, σε αντίθεση με τις συνήθεις μεθόδους πρόβλεψης, επικεντρώνονται στον εντοπισμό των ασυνεχειών στην ανάπτυξη και βοηθούν μια επιχείρηση να αντιμετωπίσει τις ξαφνικές αλλαγές και να συμβάλει αισθητά στην επιβίωσή της. Επιτρέπουν όχι μόνο την καλύτερη κατανόηση των πιθανών αδυναμιών μιας εταιρείας, αλλά συμβάλλουν επίσης στον βέλτιστο στρατηγικό προσανατολισμό της.

2.4. Ανάπτυξη σεναρίων για την υποστήριξη του στρατηγικού σχεδιασμού

Το κίνητρο για ανάπτυξη σεναρίων είναι το όραμα της εταιρείας. Ως στρατηγικό όραμα διαχείρισης θεωρείται ότι είναι μια ακριβής και δομημένη έκφραση της θέσης της εταιρείας σε έναν καθορισμένο μελλοντικό χρονικό ορίζοντα. Δεδομένου ότι η στρατηγική σκέψη

αντιπροσωπεύει ένα βασικό σημείο σχεδιασμού, αυτή η ανάπτυξη της σκέψης πρέπει να διατυπωθεί με λεπτομέρεια και με συνέπεια σε όλα επιμέρους στοιχεία της, χωρίς όμως να χάσει το νόημα στο σύνολό του. Ο ορίζοντας προγραμματισμού είναι συνήθως μεσοπρόθεσμος, αλλά ο προσδιορισμός αυτός εξαρτάται τόσο από τον χαρακτήρα της επιχείρησης (π.χ. η φαρμακευτική βιομηχανία, η ενεργειακή βιομηχανία ή η επενδυτική βιομηχανία έχουν μεγαλύτερους ορίζοντες σχεδιασμού) όσο και από την ανάπτυξη του οικονομικού κύκλου που επηρεάζει τη συνάφεια των προβλέψεων. Ακόμη και αν ο σχεδιασμός περιέχει όλες τις αρχικές πληροφορίες, πρέπει να έχει το απαραίτητο εύρος και βάθος ώστε να γίνει ένα εργαλείο για τον καθορισμό μακροπρόθεσμων στρατηγικών στόχων. Υπό την προϋπόθεση ότι για την επιτυχία της επεξεργασίας σεναρίων όλοι οι ενδιαφερόμενοι φορείς δεν θα έχουν αντιρρήσεις για το περιεχόμενο των σεναρίων και την εφαρμογή τους.

Η διαδικασία ανάπτυξης σεναρίων μπορεί να χωριστεί σε πέντε βασικά βήματα: (βλ. Fotr et al. (2014))

1. Προσδιορισμός των παραγόντων κινδύνου και προσδιορισμός της σημασίας τους.
2. Επιλογή βασικών κινδύνων που, σύμφωνα με τη γνώμη της εταιρείας, επηρεάζουν ουσιαστικά την εκπλήρωση των στρατηγικών στόχων.
3. Διαμόρφωση βασικών σεναρίων και δοκιμή της συνέπειάς τους.
4. Προσδιορισμός της πιθανότητας εμφάνισης σεναρίων.
5. Εκτέλεση ανάλυσης (έκθεσης) εντοπισμού ελλείψεων (Gap Analysis)[1]

2.5. Προσδιορισμός των παραγόντων κινδύνου και προσδιορισμός της σημασίας τους

Ο προσδιορισμός των κινδύνων είναι μια διαδικασία κατά την οποία καθορίζονται παράγοντες που μπορούν να επηρεάσουν έκβαση ενός επιχειρηματικού σχεδιασμού αρνητικά ή θετικά[2].

[1] Η αποτελεσματική αποτύπωση όλων των πιθανών κινδύνων και ελλείψεων που έχουν σχέση με τη λειτουργία των κτιριακών δομών μιας επιχειρήσεις και αφορούν την ασφάλεια των ανθρώπων της, προς αποφυγή πάσης φύσεως επικίνδυνων καταστάσεων.

[2] Είναι σημαντικό να τονίσουμε ότι δεν καταλαβαίνουμε τους κινδύνους μόνο ως αρνητικούς – απειλές, αλλά και ως θετικές – ευκαιρίες

Η ποιότητα και η αξιοπιστία των σεναρίων εξαρτώνται από την ποιότητα και την έκταση των πληροφοριών που συλλέγονται και αναλύονται. Η Επιχειρηματική Ευφυΐα (Business Intelligence) αντιπροσωπεύει μια σύνθετη μέθοδο που στοχεύει στην απόκτηση και ανάλυση υποστηρικτικών πληροφοριών σχετικά με το επιχειρηματικό περιβάλλον στο σύνολό του. Η φύση αυτής της μεθόδου είναι η επεξεργασία και η ανάλυση των διαθέσιμων και ακριβών δεδομένων, τα οποία περιλαμβάνουν κυβερνητικές πληροφορίες και έγγραφα, δικτυακούς τόπους, επιχειρηματικές παρουσιάσεις, διαφημίσεις, συνεντεύξεις, έρευνες, οικονομικές εκθέσεις, επαγγελματικές συναντήσεις, εκθέσεις, δηλώσεις διευθυντών εταιρειών και ομιλητών (ανταγωνιστές, προμηθευτές, διανομείς, πελάτες κ.λπ.) (βλ. Fotr et al. (2011))

.

Για την εφαρμογή της έννοιας της επιχειρηματικής ευφυΐας, χρησιμοποιούνται ποικίλες μεθοδολογικές προσεγγίσεις και εργαλεία τεχνικών συλλογής και ανάλυσης δεδομένων, όπως Swot Analysis, προφίλ ανταγωνιστών, συγκριτική αξιολόγηση, μοντελοποίηση της ανάπτυξης του περιβάλλοντος, οικονομικές αναλύσεις και αναλύσεις Κέρδους/Απώλειας. Συνήθως αφορούν αναδρομικές μελέτες που πραγματοποιούνται κυρίως μετά από σημαντικές συναλλαγές. Στόχος των αναλύσεων αυτών είναι να διαπιστωθεί κατά πόσον οι συναλλαγές αυτές ήταν επιτυχείς ή όχι και γιατί (βλ. Gray (2010)).

Λόγω των πολυάριθμων κινδύνων που πρέπει να εντοπίζονται συνήθως, είναι αναγκαίος ο περιορισμός σε συγκεκριμένους παράγοντες κινδύνου που παίζουν ρόλο στην ανάπτυξη σεναρίων. Επιπλέον, είναι απαραίτητο να αξιολογηθεί η σημασία και η βαρύτητα αυτών των παραγόντων. Ο πίνακας εκτίμησης κινδύνου ή η ανάλυση ευαισθησίας (υποθέτοντας ότι οι κίνδυνοι είναι ποσοτικά προσδιορισιμοί) μπορούν να χρησιμοποιηθούν ως υποστηρικτικά εργαλεία (βλ. Fotr and Souček (2011)). Τα εργαλεία αυτά περιγράφονται με μεγαλύτερη λεπτομέρεια στο επόμενο κεφάλαιο.

Είναι σημαντικό να επισημανθεί ότι οι παράγοντες κινδύνου μπορεί να δημιουργήσουν μια αλυσίδα γεγονότων (causal chain). Στο αριστερό άκρο αυτής της αλυσίδας βρίσκονται οι πηγές (αιτίες) του κινδύνου ή των κινδύνων που και στο δεξί οι επιπτώσεις τους. Για παράδειγμα, η πολεμική σύγκρουση στην Μέση Ανατολή μπορεί να θεωρηθεί το αίτιο (cause) της αύξησης των τιμών του αργού πετρελαίου. Η αύξηση των τιμών του αργού πετρελαίου έχει ως

αποτέλεσμα (occurrence) την αύξηση της τιμής της βενζίνης καθώς και την αύξηση των τιμών άλλων προϊόντων αργού πετρελαίου, τα οποία θεωρούνται πρώτες ύλες για την πετροχημική βιομηχανία. Η αύξηση της τιμής της βενζίνης συνεπάγεται την αύξηση των τιμών υπηρεσιών των εταιρειών μεταφορών (ή τη μείωση των κερδών τους, υποθέτοντας ότι δεν αυξάνουν τις τιμές των υπηρεσιών). Οι εν λόγω τιμές υπηρεσιών έχουν τότε αντίκτυπο (outcome) στο κόστος μεταφοράς των εταιρειών που καταναλώνουν αυτές τις υπηρεσίες κ.λπ.

Κατά την επεξεργασία σεναρίων στρατηγικού σχεδιασμού του αντίστοιχου θέματος, π.χ. μιας βιομηχανικής εταιρείας, είναι απαραίτητο να εργαστεί με παράγοντες κινδύνου που βρίσκονται σε κάθε επίπεδο της γραμμής παραγωγής και τα οποία μπορούν να μεταφερθούν σχετικά εύκολα στην μελέτη του στρατηγικού σχεδιασμού. Όσον αφορά τις εταιρείες μεταφορών, οι παράγοντες αυτοί μπορεί να είναι η τιμή του αργού πετρελαίου και η συναλλαγματική ισοτιμία του δολαρίου ΗΠΑ/Euro, δεδομένου ότι το δολάριο ΗΠΑ είναι ένα νόμισμα στο οποίο καθορίζονται οι τιμές του αργού πετρελαίου.

2.6. Διαμόρφωση αντιπροσωπευτικών σεναρίων και δοκιμή της συνέπειάς τους

Ο σχηματισμός σεναρίων βασίζεται στην εκτίμηση των παραγόντων κινδύνου. Ο αριθμός αυτών των παραγόντων πρέπει να περιοριστεί έτσι ώστε να μην είναι δυνατή η ύπαρξη πάρα πολλών σεναρίων. Ο υπερβολικός αριθμός σεναρίων αποτελεί εμπόδιο στην εφαρμογή τους. Η απλούστερη περίπτωση είναι η ύπαρξη των δύο σημαντικότερων δυαδικών κινδύνων, η οποία συνεπάγεται τέσσερα πιθανά σενάρια. Για την περιγραφή αυτών των σεναρίων, που αποτελούνται από δύο κινδύνους χρησιμοποιείται, ένας πίνακας σεναρίων (scenario matrix). Αυτός ο πίνακας επιτρέπει την ανάλυση της αλληλεπίδρασης μεταξύ των βασικών των παραμέτρων κινδύνων. Είναι συνήθης τακτική, ένα από τα σενάρια να αντιπροσωπεύει την πιο πιθανή έκβαση. Αυτό το σενάριο συνήθως ορίζεται ως βασικό σενάριο. Τα υπόλοιπα σενάρια βοηθούν στην ανάπτυξη λιγότερο πιθανών εναλλακτικών μελλοντικών καταστάσεων.

Είναι δυνατόν να καταλήξουμε στα ίδια αποτελέσματα μέσω δέντρων αποφάσεων (decision trees), πιθανοτήτων ή συμβάντων (βλ. Clemens (1990)), τα οποία περιγράφονται στο επόμενο κεφάλαιο. Το μειονέκτημα της χρήσης των δέντρων αποφάσεων, είναι ο χαμηλός αριθμός παραγόντων κινδύνων που μπορεί να εισέλθουν σε αυτά.

Στην πραγματικότητα, οι αναλυτές συνήθως εργάζονται με μερικά σενάρια που αντιπροσωπεύουν ορισμένες πραγματικές και πιθανές εξελίξεις βασικών παραγόντων κινδύνων. Στον στρατηγικό σχεδιασμό της εταιρείας εφαρμόζονται συνήθως τα ακόλουθα σενάρια:

1. Το αισιόδοξο σενάριο, όπου λαμβάνονται υπόψη άλλες υφιστάμενες ευκαιρίες που χρησιμοποιούνται από το εσωτερικό δυναμικό της εταιρείας. Η αισιόδοξη εξέλιξη σημαίνει ότι θα υπάρξει υπέρβαση των στόχων που έχουν τεθεί.
2. Το βασικό (το πιθανότερο) σενάριο βασίζεται στην πιο πιθανή εξέλιξη βασικών παραγόντων κινδύνου.
3. Το απαισιόδοξο σενάριο, όπου λαμβάνονται υπόψη οι περιστάσεις και οι τάσεις που προκύπτουν από απειλές που εντοπίζονται. Η εταιρεία δεν μπορεί συνήθως να αντιμετωπίσει αυτές τις απειλές από τις εσωτερικές της δυνατότητες και, κατά συνέπεια, υποτίθεται ότι οι στόχοι που καθορίζονται δεν θα εκπληρωθούν.

Ο έλεγχος σεναρίων ασχολείται με την εξέταση της συνέπειάς τους, πιο συγκεκριμένα με τον ορθολογισμό των υποθέσεων που έχουν επιλεγεί, καθώς και τη βιωσιμότητα των σεναρίων. Τα σενάρια υποβάλλονται σε κριτική λογική ανάλυση (logical analysis) με σκοπό να αξιολογηθεί το νόημά τους από την ομάδα των συντακτών τους. Εκτός από τη λογική ανάλυση, ακόμη και διαισθητικές προσεγγίσεις δεν αποκλείονται από τη διαδικασία δοκιμών. Όσο το σενάριο αποδεικνύεται ακριβές και αληθές, θα πρέπει να απαντηθεί το ερώτημα: «Ποιο είναι το αίτιο;» Το συνηθισμένο πρόβλημα είναι ότι μία ή περισσότερες υποθέσεις αποδεικνύονται μη ρεαλιστικές. Στην περίπτωση αυτή, είναι απαραίτητο να επανέλθουμε στην αρχή και να επαναπροσδιορίσουμε τις παραδοχές και μέσω της επαναληπτικής διαδικασίας να καταλήξουμε σε ένα στάδιο κατά το οποίο τα σενάρια που αναπτύσσονται είναι επαρκώς συνεπή.

Κεφάλαιο 3. Τεχνικές αξιολόγησης κινδύνων στον επιχειρηματικό σχεδιασμό

3.1. Εισαγωγή

Κατά τη δημιουργία ενός επαγγελματικού επιχειρηματικού σχεδίου, είναι σημαντικό να διασφαλιστεί ότι περιλαμβάνει τους κινδύνους και τις προκλήσεις που μπορεί να συναντήσει κατά την εκτέλεσή του. Αν και δεν είναι δυνατός ο εντοπισμός και η αντιμετώπιση όλων των πιθανών κινδύνων είναι σημαντική η επισήμανση των πιο σημαντικών. Αυτό θα βοηθήσει τη στην εύρεση τρόπων για την μετρίαση του πιθανού αντίκτυπου των κινδύνων στις επιχειρηματικές δραστηριότητες. Εκτός από τον εντοπισμό και τη συζήτηση των κινδύνων και των προκλήσεων, το σχέδιο πρέπει επίσης να περιλαμβάνει την ανάπτυξη στρατηγικών για την αντιμετώπισή τους. Τα σενάρια διαδραματίζουν σημαντικό ρόλο στην αξιολόγηση των στρατηγικών παραλλαγών και στην αξιολόγηση κινδύνου (βλ. O'Brien et al. (2007)).

Ένα καλό επιχειρηματικό σχέδιο θα ενισχύσει την αξιοπιστία, αυξάνοντας παράλληλα την εμπιστοσύνη που έχουν οι δυνητικοί επενδυτές στην επιχείρηση και τις χρηματοοικονομικές προβολές της. Είναι σημαντικό, όταν συζητάμε για τους πιθανούς επιχειρηματικούς κινδύνους, να υπάρχει ειλικρίνεια σχετικά με τους πιθανούς κινδύνους χωρίς την υποτίμηση ή την υπερβολική εκτίμηση των πιθανών κινδύνων. Η ανάλυση κινδύνου είναι ιδιαίτερα σημαντική για τις μικρές επιχειρήσεις και τις νεοσύστατες επιχειρήσεις που προσπαθούν να εξασφαλίσουν κεφάλαια για επέκταση ή για συνεχιζόμενες δραστηριότητες. Για την αποδοτικότερη ανάλυση κινδύνου, θα πρέπει να αναφέρονται οι εξής τέσσερεις παράγοντες

i. Προσδιορισμός επιχειρηματικών κινδύνων

Η διαδικασία ανάλυσης επιχειρηματικού κινδύνου ξεκινά με τον εντοπισμό των εξωτερικών και εσωτερικών απειλών που μπορούν να εμποδίσουν την επίτευξη των προγραμματισμένων αποτελεσμάτων. Οι απειλές μπορούν να ομαδοποιηθούν σε τρεις κατηγορίες, συγκεκριμένα «γενικοί επιχειρηματικοί κίνδυνοι» που αντιμετωπίζουν όλες οι εταιρείες, «συγκεκριμένοι κίνδυνοι για τη βιομηχανία» που επηρεάζουν τις επιχειρήσεις σε συγκεκριμένους κλάδους και «συγκεκριμένοι εταιρικοί κίνδυνοι» που αντιμετωπίζει η συγκεκριμένη εταιρεία. Οι κύριοι κίνδυνοι είναι εκείνοι που έχουν δυσμενείς επιπτώσεις στη ρευστότητα της εταιρείας, την οικονομική κατάσταση και τα προβλεπόμενα οικονομικά αποτελέσματα.

ii. Γενικοί επιχειρηματικοί κίνδυνοι

Οι περισσότερες επιχειρήσεις έχουν κοινούς γενικούς επιχειρηματικούς κινδύνους, αλλά οι επιπτώσεις ή η σημασία τους ποικίλλει ανάλογα με την εταιρεία. Οι νέες επιχειρήσεις ή οι νεοσύστατες επιχειρήσεις πρέπει να αποκτήσουν την απαιτούμενη εμπειρία στη διαχείριση του μάρκετινγκ, των επιχειρησιακών και άλλων ζητημάτων που θα προκύψουν. Ορισμένες πιθανές απειλές περιλαμβάνουν προβλήματα που μπορούν να αναπτυχθούν κατά τη διάρκεια του μάρκετινγκ, του ποιοτικού ελέγχου, της προώθησης, της διανομής και άλλων τομέων. Οι νεοσύστατες επιχειρήσεις και οι εταιρείες στα πρώτα στάδια πρέπει να προσελκύσουν και να οικοδομήσουν σχέσεις με πελάτες από τους ανταγωνιστές τους. Οι καθιερωμένες εταιρείες θα έχουν παρόμοια προβλήματα, αλλά ορισμένες είναι πιο ευάλωτες από άλλες.

iii. Βιομηχανικές προκλήσεις

Υπάρχουν προκλήσεις και κίνδυνοι που σχετίζονται με τη βιομηχανία και είναι σημαντικό για τις επιχειρήσεις να προσδιορίσουν τι είναι. Μια σημαντική πρόκληση είναι το ζήτημα του ανταγωνισμού στη βιομηχανία. Αν και αυτό μπορεί να είναι μια κοινή πρόκληση, το επιχειρηματικό σχέδιο πρέπει να εξετάζει τι μπορεί να κάνει η επιχείρηση για να ανταγωνιστεί αποτελεσματικά. Είναι σημαντική η ανάπτυξη στρατηγικών μάρκετινγκ ενώ και ο εντοπισμός των πλεονεκτημάτων και των αδυναμιών της επιχείρησης για την αντιμετώπιση του ανταγωνισμού.

iv. Ειδικοί κίνδυνοι για την εταιρεία

Υπάρχουν κίνδυνοι και αβεβαιότητες που σχετίζονται με διαφορετικές εταιρείες, συμπεριλαμβανομένων ζητημάτων του ανθρώπινου δυναμικού. Οι κίνδυνοι είναι διαφορετικοί ανάλογα με το στάδιο στο οποίο βρίσκεται η εταιρεία. Οι νεοσύστατες εταιρείες θα έχουν συχνά προβλήματα όταν πρόκειται για την απόκτηση αρχικού κεφαλαίου ή κεφαλαίου κίνησης, το οποίο επηρεάζει αποτελεσματικά τις δραστηριότητες. Υπάρχουν επίσης κίνδυνοι που σχετίζονται με τη δομή σταθερού κόστους της επιχείρησης και μπορεί να διαφέρουν ανάλογα με την εταιρεία. Ορισμένες εταιρείες έχουν υψηλό σταθερό κόστος λόγω των μεγάλων επενδύσεων σε εγκαταστάσεις και εξοπλισμό.

3.2. Πίνακας εκτίμησης κινδύνου

Ο πίνακας εκτίμησης κινδύνου είναι ένα εργαλείο αξιολόγησης από εμπειρογνώμονες της σημασίας του κινδύνου, η οποία βασίζεται σε δύο πτυχές. Το πρώτο είναι η πιθανότητα

εμφάνισης κινδύνου και το δεύτερο είναι η ισχύς του αντικτύπου της στην επιχείρηση (συνήθως με τη μορφή επιλεγμένου χρηματοοικονομικού δείκτη, ο οποίος μπορεί να είναι κέρδος, ταμειακές ροές κ.λπ.). Το αποτέλεσμα του πίνακα εκτίμησης κινδύνου είναι ο κατάλογος των βασικών κινδύνων που χαρακτηρίζονται τόσο από την υψηλή πιθανότητα εμφάνισης όσο και από την υψηλή σημασία του αντίκτυπου.

Η φύση της ανάλυσης ευαισθησίας είναι οι επιπτώσεις των ίσων σχετικών αποκλίσεων (π.χ. 10%) των παραγόντων κινδύνου (πωλήσεις, τιμή πώλησης, τιμές αγοράς πρώτων υλών, τιμές ενέργειας, συναλλαγματικές ισοτιμίες) από τις πιο πιθανές τιμές τους βάσει των βασικών κριτηρίων απόδοσης της εταιρείας. Το αποτέλεσμα της ανάλυσης είναι στη συνέχεια ο προσδιορισμός των παραγόντων στους οποίους το επιλεγμένο κριτήριο αξιολόγησης του στρατηγικού σχεδίου είναι εξαιρετικά ευαίσθητο. Στη συνέχεια, οι παράγοντες αυτοί διευρύνουν το σύνολο των βασικών κινδύνων.

3.3. Προσδιορισμός πιθανότητας σεναρίου

Σε περίπτωση σχετικά απλών σεναρίων είναι δυνατόν να προσδιοριστούν οι πιθανότητές τους. Κατά την εφαρμογή δέντρων πιθανότητας σε σενάρια, είναι απαραίτητο, ειδικά λόγω της συχνής εξάρτησης των μεμονωμένων κινδύνων, πρώτα απ 'όλα, να προσδιοριστεί η άνευ όρων κατανομή πιθανότητας των κινδύνων, η οποία απεικονίζεται από τον κόμβο ευκαιρία τοποθετείται στην αριστερή πλευρά του δέντρου. Είναι η ρίζα του δέντρου. Επιπλέον, η κατανομή των πιθανοτήτων των κινδύνων που απεικονίζονται από κόμβους που βρίσκονται στη δεξιά πλευρά της ρίζας του δέντρου έχει καθοριστεί σταδιακά. Αυτή η διαδικασία συνεχίζεται μέχρι να καταλήξουμε σε κινδύνους που απεικονίζονται από τον κόμβο που βρίσκεται στην άκρα δεξιά πλευρά του δέντρου. Η πιθανότητα κάθε σεναρίου υπολογίζεται στη συνέχεια ως πολλαπλασιασμός των τιμών πιθανότητας κινδύνου που βρίσκονται στον ίδιο κλάδο του δέντρου πιθανοτήτων. Οι πιθανότητες των κινδύνων συνήθως δεν έχουν το χαρακτήρα των αντικειμενικών πιθανοτήτων που καθορίζονται με βάση ιστορικά δεδομένα μέσω στατιστικών μεθόδων. Ασχολείται κατά κύριο τρόπο με υποκειμενικές πιθανότητες, οι οποίες βασίζονται στην εμπειρία, τη διαίσθηση και το ιστορικό πληροφοριών αυτών των εμπειρογνωμόνων (βλ. Courtney (2003)).

Κεφάλαιο 4. Εφαρμογή τεχνικών ομαδοποίησης στον επιχειρηματικό σχεδιασμό

4.1. Μέθοδοι Ομαδοποίησης

Σε αυτό το κεφάλαιο πραγματοποιήθηκε η εξέταση διαφόρων αλγορίθμων ομαδοποίησης σε πραγματικά σενάρια προβλήματα επιχειρηματικού σχεδιασμού. Πιο συγκεκριμένα, υλοποιήθηκαν και εφαρμόσθηκαν 4 παραδοσιακοί και διαδεδομένοι αλγόριθμοι, ο k-means, ο k-medoid, ο DBSCAN, ο συσσωρευτικό ιεραρχικός αλγόριθμος (Agglomerative) με το κριτήριο «μονού συνδέσμου» (single linkage) ως κριτήριο επιλογής για την ένωση των ομάδων (cluster). Οι εφαρμογές υλοποιήθηκαν στην R V4.0.3 και ο κώδικας που χρησιμοποιήθηκε, αντίστοιχα, παρουσιάζεται στο παράρτημα της εργασίας.

4.1.1. (Μη)-Επιβλεπόμενη Μάθηση

Πριν την περιγραφή των αλγορίθμων, περιγράφεται το πλαίσιο της μη επιβλεπόμενης μάθησης για να γίνει πιο κατανοητή η λειτουργία των αλγορίθμων. Όπως αναφέρθηκε και στην εισαγωγή, το αν τα δεδομένα μας έχουνε labels, ορίζει και την επιλογή του αλγορίθμου μηχανικής μάθησης. Κατηγοριοποίηση ή Παλινδρόμηση μπορούμε να κάνουμε στην περίπτωση που έχουμε labels, ενώ ομαδοποίηση ή οπτικοποίηση μπορούμε να εφαρμόσουμε όταν τα δεδομένα μας είναι χωρίς label. Αυτά ως γενικός κανόνας, διότι η ομαδοποίηση μπορεί να εφαρμοσθεί και σε δεδομένα με label τα οποία χρησιμοποιούνται ως κριτήριο για την αξιολόγηση της απόδοσής τους. Παρακάτω ορίζεται μαθηματικά η διαφορά τους.

Στην επιβλεπόμενη μάθηση, όσον αφορά την πρόβλεψη των τιμών μίας ή περισσότερων εξόδων ή μεταβλητών απόκρισης $Y = (Y_1, \dots, Y_m)$ για ένα δεδομένο σύνολο μεταβλητών εισόδου ή πρόβλεψης $X^T = (X_1, \dots, X_p)$, οι είσοδοι για την αντίστοιχη i περίπτωση προπόνησης, δηλώνονται με $x_i^T = (x_{i1}, \dots, x_{ip})$. Το y_i αφορά μια μέτρηση απόκρισης. Οι προβλέψεις βασίζονται στο εκπαιδευτικό δείγμα $(x_1, y_1), \dots, (x_N, y_N)$ προηγούμενων επιλυθέντων περιπτώσεων, όπου οι κοινές τιμές όλων των μεταβλητών είναι γνωστές. Αυτό ονομάζεται εποπτευόμενη μάθηση ή «μάθηση με έναν δάσκαλο». Κάτω από αυτή τη μεταφορά, ο «μαθητής» παρουσιάζει μια απάντηση $\hat{y}_i$ για κάθε x_i στο δείγμα κατάρτισης και ο επιβλέπων ή «δάσκαλος» παρέχει είτε τη σωστή απάντηση ή/και ένα σφάλμα που σχετίζεται με την απάντηση του μαθητή. Αυτό συνήθως χαρακτηρίζεται από κάποια συνάρτηση απώλειας

$L(y, \hat{y})$, για παράδειγμα, $L(y, \hat{y}) = (y - \hat{y})^2$. Εάν εκληφθεί ότι οι (X, Y) είναι τυχαίες μεταβλητές που αντιπροσωπεύονται από κάποια πυκνότητα πιθανότητας αρθρώσεων $\Pr(X, Y)$, τότε η εποπτευόμενη μάθηση μπορεί τυπικά να χαρακτηριστεί ως πρόβλημα εκτίμησης πυκνότητας όταν κάποιος ενδιαφέρεται για τον προσδιορισμό των ιδιοτήτων της υπό όρους πυκνότητας $\Pr(Y \mid X)$. Συνήθως οι ιδιότητες που έχουν το μεγαλύτερο ενδιαφέρον είναι οι παράμετροι «θέσης» μ που ελαχιστοποιούν το αναμενόμενο σφάλμα σε κάθε x,

$$\mu(x) = \underset{\theta}{\operatorname{argmin}} E_{Y|X} L(Y, \theta) \tag{4.1}$$

, όπου το θ αντιπροσωπεύει το $\hat{y}$ και το L το loss function. Στην ουσία η παραπάνω εξίσωση αναζητεί το ελάχιστο θ το οποίο θα κάνει προβλέψεις όσο το δυνατόν πιο κοντά στα πραγματικά label. Η πιθανότητα ορίζεται ως

$$\Pr(X, Y) = \Pr(Y \mid X) \cdot \Pr(X) \tag{4.2}$$

, $\Pr(X)$ είναι η οριακή πυκνότητα μόνο των τιμών X. Στην εποπτευόμενη μάθηση το $\Pr(X)$ συνήθως δεν απασχολεί άμεσα. Κάποιος ενδιαφέρεται κυρίως για τις ιδιότητες της υπό όρους πυκνότητας $\Pr(Y \mid X)$. Δεδομένου ότι το Y είναι συχνά χαμηλής διάστασης (συνήθως ένα), και ενδιαφέρει μόνο η θέση του $\mu(x)$, το πρόβλημα απλοποιείται σε μεγάλο βαθμό. Υπάρχουν όμως πολλές προσεγγίσεις για την επιτυχή αντιμετώπιση της εποπτευόμενης μάθησης σε διάφορα πλαίσια. Σε αυτό το κεφάλαιο αναφέρεται η μάθηση χωρίς επίβλεψη ή η "μάθηση χωρίς δάσκαλο". Σε αυτήν την περίπτωση υπάρχει ένα σύνολο N παρατηρήσεων $(x_1, x_2, \dots, x_N)$ ενός τυχαίου p -φορέα X που έχει πυκνότητα άρθρωσης $\Pr(X)$. Ο στόχος είναι να εξακριβωθούν άμεσα οι ιδιότητες αυτής της πυκνότητας πιθανότητας, χωρίς τη βοήθεια ενός επόπτη ή δασκάλου που παρέχει σωστές απαντήσεις ή βαθμό σφάλματος για κάθε παρατήρηση. Η διάσταση του X είναι μερικές φορές πολύ υψηλότερη από ό, τι στην εποπτευόμενη μάθηση και οι ιδιότητες του ενδιαφέροντος είναι συχνά πιο περίπλοκες από τις απλές εκτιμήσεις τοποθεσίας. Αυτοί οι παράγοντες μετριάζονται κάπως από το γεγονός ότι το X αντιπροσωπεύει όλες τις υπό εξέταση μεταβλητές άρα δεν απαιτείται να βγει κάποιο συμπέρασμα για το πώς αλλάζουν οι ιδιότητες του $\Pr(X)$ που εξαρτώνται από τις μεταβαλλόμενες τιμές ενός άλλου συνόλου μεταβλητών. Σε προβλήματα χαμηλών διαστάσεων (όπως το $p \leq 3$) και σύμφωνα με την έρευνα του (Silverman, 1986), υπάρχει μια ποικιλία αποτελεσματικών μη παραμετρικών μεθόδων για την άμεση εκτίμηση της πυκνότητας $\Pr(X)$ σε όλες τις τιμές X και την γραφική της αναπαράσταση. Λόγω του φαινομένου της κατάρας της διάστασης, αυτές οι μέθοδοι αποτυγχάνουν σε υψηλές διαστάσεις. Πρέπει να δοθεί βάση στην εκτίμηση μάλλον

ακατέργαστων παγκόσμιων μοντέλων, όπως μίγματα Gauss ή διάφορα απλά περιγραφικά στατιστικά που χαρακτηρίζουν το $\Pr(X)$. Γενικά, αυτές οι περιγραφικές στατιστικές επιχειρούν να χαρακτηρίσουν τιμές X ή συλλογές τέτοιων τιμών, όπου το $\Pr(X)$ είναι σχετικά μεγάλο. Τα κύρια συστατικά όπως, η πολυδιάστατη κλιμάκωση, οι αυτοοργανωτικοί χάρτες και οι κύριες καμπύλες, προσπαθούν να εντοπίσουν πολλαπλότητες χαμηλής διάστασης εντός του X-χώρου που αντιπροσωπεύουν υψηλή πυκνότητα δεδομένων. Αυτό παρέχει πληροφορίες σχετικά με τις συσχετίσεις μεταξύ των μεταβλητών και αν μπορούν ή όχι να θεωρηθούν συναρτήσεις ενός μικρότερου συνόλου «λανθάνων» μεταβλητών. Η ανάλυση συμπλεγμάτων προσπαθεί να βρει πολλαπλές κυρτές περιοχές του X-χώρου που περιέχουν λειτουργίες $\Pr(X)$. Αυτό μπορεί να φανερώσει εάν το $\Pr(X)$ μπορεί να αναπαρασταθεί από ένα μείγμα απλούστερων πυκνοτήτων που αντιπροσωπεύουν ξεχωριστούς τύπους ή κατηγορίες παρατηρήσεων. Ο συνδυασμός μοντελοποίησης έχει παρόμοιο στόχο. Οι κανόνες σύνδεσης επιχειρούν να κατασκευάσουν απλές περιγραφές (συνδετικοί κανόνες) που περιγράφουν περιοχές υψηλής πυκνότητας στην ειδική περίπτωση δεδομένων δυαδικής αξίας πολύ υψηλών διαστάσεων. Με την εποπτευόμενη μάθηση υπάρχει ένα σαφές μέτρο επιτυχίας ή έλλειψης αυτής, το οποίο μπορεί να χρησιμοποιηθεί για να κριθεί η επάρκεια σε συγκεκριμένες καταστάσεις και να συγκριθεί η αποτελεσματικότητα διαφορετικών μεθόδων σε διάφορες καταστάσεις. Η έλλειψη επιτυχίας μετριέται άμεσα με την αναμενόμενη απώλεια στην κοινή κατανομή $\Pr(X, Y)$. Αυτό μπορεί να εκτιμηθεί με διάφορους τρόπους, συμπεριλαμβανομένης της διασταυρούμενης επικύρωσης. Στο πλαίσιο της μάθησης χωρίς επίβλεψη, δεν υπάρχει τέτοιο άμεσο μέτρο επιτυχίας. Είναι δύσκολο να εξακριβωθεί η εγκυρότητα των συμπερασμάτων που προέρχονται από το αποτέλεσμα των περισσότερων αλγορίθμων μάθησης χωρίς επίβλεψη. Κάποιος πρέπει να καταφύγει σε ευρετικά επιχειρήματα όχι μόνο για την παρακίνηση των αλγορίθμων, όπως συμβαίνει συχνά στην εποπτευόμενη μάθηση, αλλά και για την κρίση ως προς την ποιότητα των αποτελεσμάτων. Αυτή η δυσάρεστη κατάσταση έχει οδηγήσει σε έντονο πολλαπλασιασμό των προτεινόμενων μεθόδων, καθώς η αποτελεσματικότητα είναι θέμα γνώμης και δεν μπορεί να επαληθευτεί άμεσα.

4.1.2. Ο αλγόριθμος k-means

Ο αλγόριθμος K-means είναι μία από τις πιο δημοφιλείς επαναληπτικές μεθόδους ομαδοποίησης. Προορίζεται για καταστάσεις στις οποίες όλες οι μεταβλητές είναι ποσοτικού τύπου και η τετραγωνισμένη Ευκλείδεια απόσταση

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 = \|x_i - x_i\|^2 \quad (4.3)$$

επιλέγεται ως μέτρο ανομοιότητας. Πρέπει να σημειωθεί ότι η σταθμισμένη Ευκλείδεια απόσταση μπορεί να χρησιμοποιηθεί επαναπροσδιορίζοντας τις τιμές x_{ij}. Η διασπορά εντός σημείου μπορεί να γραφτεί ως

$$\begin{aligned} W(C) \quad &\equiv \frac{1}{2}\sum_{k=1}^{K} \sum_{C(i)=kC(i')=k} \sum \|x_{\ i} - x_{i'}\|^2 \\ &= \sum_{k=1}^{K} \left(N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2 \right) \end{aligned} \quad (4.4)$$

όπου $\bar{x}_k = \left(\bar{x}_{1k}, \ldots, \bar{x}_{pk}\right)$ είναι το μέσο διάνυσμα που σχετίζεται με το αντίστοιχο k cluster και $N_k = \sum_{i=1}^{N} I(C(i) = k)$. Έτσι, το κριτήριο ελαχιστοποιείται με την ανάθεση των παρατηρήσεων N στα συμπλέγματα K με τέτοιο τρόπο ώστε μέσα σε κάθε cluster να ελαχιστοποιείται η μέση ανομοιότητα των παρατηρήσεων από τη μέση ομάδα, όπως ορίζεται από τα σημεία αυτού του συμπλέγματος. Ένας αλγόριθμος επαναληπτικής καθόδου για επίλυση της

$$C^* = \min_{C} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2 \quad (4.5)$$

μπορεί να ληφθεί σημειώνοντας ότι για οποιοδήποτε σύνολο παρατηρήσεων S

$$\bar{x}_S = \underset{m}{\operatorname{argmin}} \sum_{i \in S} \|x_i - m\|^2 \quad (4.6)$$

Ως εκ τούτου, μπορούμε να λάβουμε C^* λύνοντας το διευρυμένο πρόβλημα βελτιστοποίησης

$$\min_{C, \{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - m_k\|^2 \quad (4.7)$$

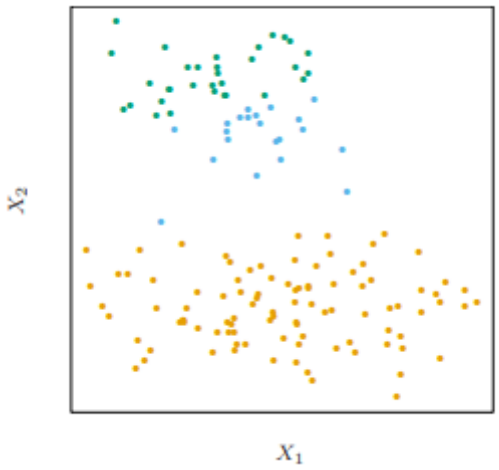
Αυτό μπορεί να ελαχιστοποιηθεί με μια εναλλακτική διαδικασία βελτιστοποίησης που δίνεται στον Αλγόριθμο 4.1.

Αλγόριθμος 4.1 Ομαδοποίηση K-means.

1. Για μια δεδομένη εκχώρηση συμπλέγματος C, η συνολική διακύμανση του συμπλέγματος (σύμφωνα με την εξίσωση 4.4), ελαχιστοποιείται σε σχέση με το $\{m_1, \dots, m_K\}$ αποδίδοντας τα τωρινά μέσα των τύπων που έχουν εκχωρηθεί (σύμφωνα με την εξίσωση 4.3).
2. Λαμβάνοντας υπόψη ένα τρέχον σύνολο μέσων $\{m_1, \dots, m_K\}$, η εξίσωση 4.4 ελαχιστοποιείται με την ανάθεση κάθε παρατήρησης στο πλησιέστερο (τρέχον) μέσο συμπλέγματος. Αυτό είναι,
$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} \|x_i - m_k\|^2 \tag{4.8}$$
3. Τα βήματα 1 και 2 επαναλαμβάνονται έως ότου δεν αλλάξουν οι εργασίες.

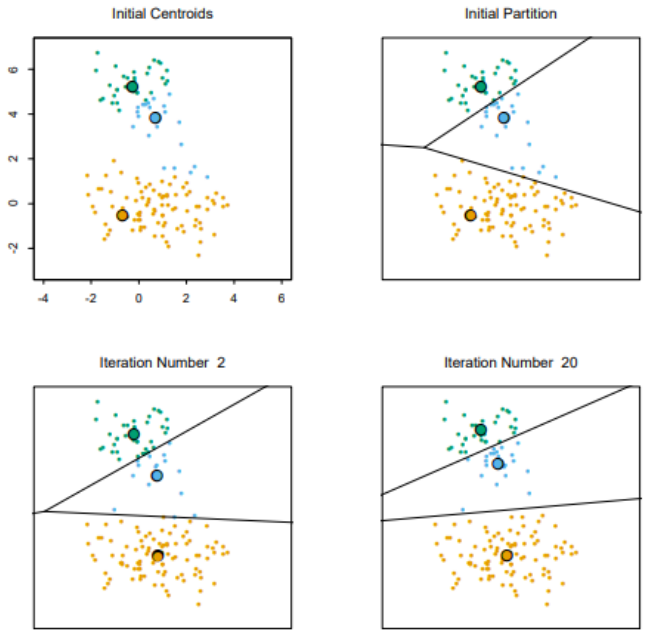
Κάθε ένα από τα βήματα 1 και 2 μειώνει την τιμή του κριτηρίου (εξίσωση 4.4), έτσι ώστε να εξασφαλίζεται η σύγκλιση. Ωστόσο, το αποτέλεσμα μπορεί να αντιπροσωπεύει ένα μη βέλτιστο τοπικό ελάχιστο. Ο αλγόριθμος των (Hartigan & Wong, 1979) πηγαίνει ένα βήμα παραπέρα και διασφαλίζει ότι δεν υπάρχει μεμονωμένη εναλλαγή παρατήρησης από τη μία ομάδα στην άλλη, που να μπορεί να μειώσει τον στόχο. Επιπλέον, κάποιος πρέπει να ξεκινήσει τον αλγόριθμο με πολλές διαφορετικές τυχαίες επιλογές για τα μέσα εκκίνησης και να επιλέξει τη λύση που έχει τη μικρότερη τιμή της αντικειμενικής συνάρτησης.

Το σχήμα 1 δείχνει μερικά προσομοιωμένα δεδομένα ομαδοποιημένα σε τρεις ομάδες μέσω του δημοφιλούς αλγορίθμου K-μέσων (K-means). Είναι προφανές ότι φαινομενικά βλέπουμε 2 ομάδες, συνεπώς η μία ομάδα σπάει σε 2 cluster, εφόσον η επιλογή του χρήστη είναι k=3.



Σχήμα 1. Προσομοιωμένα δεδομένα στο επίπεδο, ομαδοποιημένα σε τρεις κατηγορίες (που αντιπροσωπεύονται από πορτοκαλί, μπλε και πράσινο) με τον αλγόριθμο ομαδοποίησης K-means. (Πηγή: Friedman et al., 2001). Ως επιλογή έχει δοθεί το k=3, κάτι που δημιουργεί 3 ομάδες με αποτέλεσμα ένα από τα 2 φαινομενικά clusters να σπάει στα 2.

Το σχήμα 2 δείχνει μερικές από τις επαναλήψεις των K-means για τα προσομοιωμένα δεδομένα του σχήματος 1. Τα κέντρα μάζας απεικονίζονται με "O". Οι ευθείες δείχνουν την κατανομή των σημείων και κάθε τομέας είναι το σύνολο των σημείων που είναι πιο κοντά στο κάθε κέντρο μάζας. Αυτή η διαίρεση ονομάζεται διάγραμμα Voronoi (Voronoi diagram). Μετά από 20 επαναλήψεις, η διαδικασία συγκλίνει.



Σχήμα 2. Διαδοχικές επαναλήψεις του αλγορίθμου ομαδοποίησης K-means για τα προσομοιωμένα δεδομένα του σχήματος 4.2. (Πηγή: Friedman et al., 2001). Παρατηρούμε στο πρώτο σχήμα να γίνεται τυχαία τοποθέτηση στα 3 κέντρα (K=3). Στην ίδια εικόνα έχουμε και την ανάθεση

4.1.3. Ο αλγόριθμος k-medoid

Ο αλγόριθμος K-means είναι κατάλληλος όταν το μέτρο ανομοιότητας λαμβάνεται ως τετραγωνικό Ευκλείδειας απόστασης $D(x_i, x_{i'})$. Αυτό απαιτεί όλες οι μεταβλητές να είναι ποσοτικού τύπου. Επιπλέον, η χρήση τετραγωνικής Ευκλείδειας απόστασης ασκεί την μεγαλύτερη επιρροή στις μεγαλύτερες αποστάσεις. Αυτό προκαλεί έλλειψη ευρωστίας στη διαδικασία έναντι των πολύ υψηλών αποστάσεων που παράγουν πολύ μεγάλες αποστάσεις. Αυτοί οι περιορισμοί μπορούν να αρθούν σε βάρος του υπολογισμού. Το μόνο μέρος του αλγορίθμου K-means που υποθέτει τετραγωνισμένη ευκλείδεια απόσταση είναι το βήμα ελαχιστοποίησης. Οι εκπρόσωποι συμπλέγματος $\{m_1, \dots, m_K\}$, θεωρούνται ως τα μέσα των συμπλεγμάτων που έχουν εκχωρηθεί αυτήν τη στιγμή. Ο αλγόριθμος μπορεί να γενικευτεί για χρήση με αυθαίρετα καθορισμένες ομοιότητες $D(x_i, x_i)$ αντικαθιστώντας αυτό το βήμα με μια ρητή βελτιστοποίηση σε σχέση με $\{m_1, \dots, m_K\}$ στην εξίσωση 4.7. Στην πιο κοινή μορφή, τα κέντρα για κάθε cluster περιορίζονται ως μία από τις παρατηρήσεις που έχουν ανατεθεί στο cluster, όπως συνοψίζεται στον Αλγόριθμο 4.2. Αυτός ο αλγόριθμος υποθέτει δεδομένα χαρακτηριστικών, αλλά η προσέγγιση μπορεί επίσης να εφαρμοστεί σε δεδομένα που

περιγράφονται μόνο με πίνακες εγγύτητας. Δεν χρειάζεται να υπολογιστούν ρητά τα κέντρα ομάδων καθώς γίνεται απλή παρακολούθηση των δεικτών i_k^*. Η επίλυση της εξίσωσης 4.6 για κάθε προσωρινό cluster k απαιτεί ένα ποσό υπολογισμού ανάλογο με τον αριθμό των παρατηρήσεων που του έχουν ανατεθεί, ενώ για την επίλυση της εξίσωσης 4.6, ο υπολογισμός αυξάνεται σε $O(N_k^2)$. Δεδομένου ενός συνόλου "κέντρων" συμπλέγματος, $\{i_1, \dots, i_K\}$ για να αποκτηθούν νέες εργασίες

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} d_{ii_k^*}$$

απαιτείται υπολογισμός ανάλογος του $K \cdot N$ όπως και πριν. Έτσι, τα K-μεδοειδή (K-medoids) είναι πολύ πιο υπολογιστικά εντατικά συγκριτικά με τα K-means.

Αλγόριθμος 4.2 Ομαδοποίηση K-medoids.

1. Για μια δεδομένη ανάθεση συμπλέγματος C η παρατήρηση στο cluster εντοπίζεται ελαχιστοποιώντας τη συνολική απόσταση από άλλα σημεία σε αυτό το cluster:

$$i_k^* = \underset{\{i:C(i)=k\}}{\operatorname{argmin}} \sum_{C(i')=k} D(x_i, x_{i'}) \qquad (4.9)$$

   Τότε οι $m_k \equiv x_{i_k^*}, k = 1,2, \dots, K$ είναι οι τρέχουσες εκτιμήσεις των κέντρων των συμπλεγμάτων.

2. Λαμβάνοντας υπόψη ένα τρέχον σύνολο κέντρων συμπλέγματος $\{m_1, \dots, m_K\}$, ελαχιστοποιείται το συνολικό σφάλμα εκχωρώντας κάθε παρατήρηση στο πλησιέστερο (τρέχον) κέντρο συμπλέγματος:

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} D(x_i, m_k) \qquad (4.10)$$

3. Τα βήματα 1 και 2 επαναλαμβάνονται έως ότου δεν αλλάξουν οι εργασίες.

4.1.4. Αλγόριθμος Ιεραρχικής Συσσωρευτική Ομαδοποίησης

Τα αποτελέσματα της εφαρμογής αλγορίθμων ομαδοποίησης K-μέσων εξαρτώνται από την επιλογή του αριθμού των συμπλεγμάτων που θα αναζητηθούν και από την εκκίνηση της εκχώρησης διαμόρφωσης. Αντίθετα, οι ιεραρχικές μέθοδοι ομαδοποίησης δεν απαιτούν τέτοιες προδιαγραφές. Αντ' αυτού, απαιτούν από τον χρήστη να καθορίσει ένα μέτρο ανομοιότητας μεταξύ (ασύνδετων) ομάδων παρατηρήσεων, με βάση τα ζεύγη διαφορών μεταξύ των παρατηρήσεων στις δύο ομάδες. Όπως υποδηλώνει το όνομα, παράγουν ιεραρχικές αναπαραστάσεις στις οποίες οι ομάδες σε κάθε επίπεδο της ιεραρχίας δημιουργούνται με τη

συγχώνευση συμπλεγμάτων στο επόμενο χαμηλότερο επίπεδο. Στο χαμηλότερο επίπεδο, κάθε cluster περιέχει μία μόνο παρατήρηση. Στο υψηλότερο επίπεδο υπάρχει μόνο ένα cluster που περιέχει όλα τα δεδομένα. Οι στρατηγικές για την ιεραρχική ομαδοποίηση χωρίζονται σε δύο βασικά παραδείγματα: συσσωρευτική (από κάτω προς τα πάνω) και διαιρετική (από πάνω προς τα κάτω). Οι συσσωρευτικές στρατηγικές ξεκινούν από το κάτω μέρος και σε κάθε επίπεδο συγχωνεύουν αναδρομικά ένα επιλεγμένο ζεύγος ομάδων σε ένα μόνο cluster. Αυτό δημιουργεί μια ομαδοποίηση στο επόμενο υψηλότερο επίπεδο με ένα μικρό cluster. Το ζεύγος που επιλέχθηκε για συγχώνευση αποτελείται από τις δύο ομάδες με τη μικρότερη ομοιότητα μεταξύ των ομάδων. Οι μέθοδοι διαίρεσης ξεκινούν από την κορυφή και σε κάθε επίπεδο χωρίζουν αναδρομικά ένα από τα υπάρχοντα συμπλέγματα σε αυτό το επίπεδο σε δύο νέα συμπλέγματα. Η διάσπαση επιλέγεται για να δημιουργήσει δύο νέες ομάδες με τη μεγαλύτερη ανομοιότητα μεταξύ των ομάδων. Και με τα δύο παραδείγματα υπάρχουν επίπεδα $N-1$ στην ιεραρχία. Κάθε επίπεδο της ιεραρχίας αντιπροσωπεύει μια συγκεκριμένη ομαδοποίηση των δεδομένων σε ασύνδετες ομάδες παρατηρήσεων. Ολόκληρη η ιεραρχία αντιπροσωπεύει μια διατεταγμένη σειρά τέτοιων ομαδοποιήσεων. Εναπόκειται στον χρήστη να αποφασίσει ποιο επίπεδο (εάν υπάρχει) αντιπροσωπεύει πραγματικά μια «φυσική» ομάδα με την έννοια ότι οι παρατηρήσεις σε κάθε ομάδα της, είναι αρκετά πιο παρόμοιες μεταξύ τους παρά με παρατηρήσεις που έχουν εκχωρηθεί σε διαφορετικές ομάδες σε αυτό το επίπεδο. Ο αναδρομικός δυαδικός διαχωρισμός/συσσωμάτωση μπορεί να αναπαρασταθεί από ένα ριζωμένο δυαδικό δέντρο. Οι κόμβοι των δέντρων αντιπροσωπεύουν ομάδες. Ο ριζικός κόμβος αντιπροσωπεύει ολόκληρο το σύνολο δεδομένων. Οι τελικοί κόμβοι N αντιπροσωπεύουν ο καθένας μία από τις μεμονωμένες παρατηρήσεις (μεμονωμένες ομάδες). Κάθε μη τερματικός κόμβος ("γονέας") έχει δύο θυγατρικούς κόμβους. Για την ομαδοποίηση διαιρέσεων, οι δύο κόρες αντιπροσωπεύουν τις δύο ομάδες που προκύπτουν από τη διάσπαση του γονέα, ενώ στην συσσωρευτική ομαδοποίηση, οι κόρες αντιπροσωπεύουν τις δύο ομάδες που συγχωνεύθηκαν για να σχηματίσουν τον γονέα. Όλες οι συσσωρευτικές και μερικές μέθοδοι διαίρεσης (όταν τις βλέπουμε από κάτω προς τα πάνω) διαθέτουν μια ιδιότητα μονοτονίας. Δηλαδή, η ανομοιότητα μεταξύ συγχωνευμένων ομάδων είναι μονότονη και αυξανόμενη με το επίπεδο της συγχώνευσης. Έτσι, το δυαδικό δέντρο μπορεί να σχεδιαστεί έτσι ώστε το ύψος κάθε κόμβου να είναι ανάλογο με την τιμή της ανισότητας μεταξύ ομάδων μεταξύ των δύο κόρων του. Όλοι οι τελικοί κόμβοι που αντιπροσωπεύουν μεμονωμένες παρατηρήσεις σχεδιάζονται σε μηδενικό ύψος. Αυτός ο τύπος γραφικής απεικόνισης ονομάζεται δενδρόγραμμα. Ένα

δενδρόγραμμα παρέχει μια εξαιρετικά ερμηνεύσιμη πλήρη περιγραφή της ιεραρχικής ομαδοποίησης σε γραφική μορφή. Αυτός είναι ένας από τους κύριους λόγους για τη δημοτικότητα των ιεραρχικών μεθόδων ομαδοποίησης.

Συσσωρευτική ομαδοποίηση

Οι αλγόριθμοι συσσωρευτικής ομαδοποίησης ξεκινούν με κάθε παρατήρηση που αντιπροσωπεύει ένα απλό cluster. Σε κάθε ένα από τα βήματα $N-1$ τα δύο πλησιέστερα (λιγότερο διαφορετικά) συμπλέγματα συγχωνεύονται σε ένα μόνο cluster, παράγοντας ένα μικρό cluster στο επόμενο υψηλότερο επίπεδο. Επομένως, πρέπει να καθοριστεί ένα μέτρο ανομοιότητας μεταξύ δύο ομάδων (ομάδες παρατηρήσεων). Τα G και H αντιπροσωπεύουν δύο τέτοιες ομάδες. Η ανομοιότητα $d(G,H)$ μεταξύ G και H υπολογίζεται από το σύνολο των διαφορών παρατήρησης κατά ζεύγη $d_{ii'}$ όπου το ένα μέλος του ζεύγους i είναι στο G και το άλλο i' είναι στο H. Η συσσωρευτική ομαδοποίηση μονής σύνδεσης (Single linkage - SL) θεωρεί ότι η ομοιότητα μεταξύ των ομάδων είναι αυτή του πλησιέστερου (λιγότερο ανόμοιου) ζεύγους

$$d_{SL}(G,H) = \min_{\substack{i\in G \\ i\in H}} d_{ii'} \tag{4.11}$$

Επίσης, αυτό ονομάζεται συχνά η τεχνική του πλησιέστερου γείτονα (nearest-neighbor technique). Η συσσωρευτική ομαδοποίηση πλήρης σύνδεσης (Complete linkage - CL) θεωρεί ότι η ομοιότητα μεταξύ των ομάδων είναι αυτή του πιο μακρινού (πιο ανόμοιου) ζεύγους

$$d_{CL}(G,H) = \max_{i\in G} d_{ii'} \tag{4.12}$$

Η ομαδοποίηση του μέσου όρου ομάδας (Group average - GA) χρησιμοποιεί τη μέση ανομοιότητα μεταξύ των ομάδων

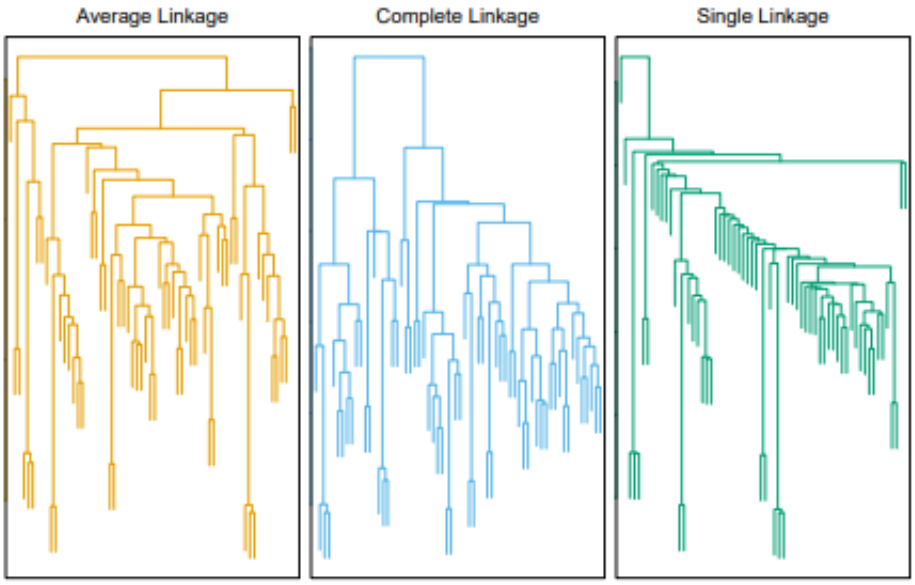
$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{i\in G} \sum_{i'\in H} d_{ii'} \tag{4.13}$$

Όπου N_G και N_H είναι ο αντίστοιχος αριθμός παρατηρήσεων σε κάθε ομάδα. Παρόλο που έχουν υπάρξει πολλές άλλες προτάσεις για τον καθορισμό της ανισότητας μεταξύ ομάδων στο πλαίσιο της συγκεντρωτικής ομαδοποίησης, οι τρεις παραπάνω είναι αυτές που χρησιμοποιούνται συχνότερα. Το σχήμα 3 δείχνει παραδείγματα και των τριών. Εάν οι διαφορές δεδομένων $\{d_{ii'}\}$ εμφανίζουν έντονη τάση ομαδοποίησης, με κάθε ένα από τα συμπλέγματα να είναι συμπαγές και να διαχωρίζεται καλά από τα άλλα, τότε και οι τρεις μέθοδοι παράγουν παρόμοια αποτελέσματα. Τα συμπλέγματα είναι συμπαγή εάν όλες οι παρατηρήσεις μέσα τους είναι σχετικά κοντά (μικρές ομοιότητες) σε σύγκριση με τις

παρατηρήσεις σε διαφορετικές ομάδες. Στο βαθμό που αυτό δεν συμβαίνει, τα αποτελέσματα θα διαφέρουν. Η μονή σύνδεση (σύμφωνα με την εξίσωση 4.9) απαιτεί μόνο από μια μονή ανομοιότητα $d_{ii'}$, $i \in G$ και $i' \in H$, να είναι μικρή για δύο ομάδες G και H (οι οποίες θεωρούνται κοντά μεταξύ τους), ανεξάρτητα από τις άλλες ομοιότητες παρατήρησης μεταξύ των ομάδων. Συνεπώς, θα έχει την τάση να συνδυάζει, σε σχετικά χαμηλά όρια, παρατηρήσεις που συνδέονται με μια σειρά στενών ενδιάμεσων παρατηρήσεων. Αυτό το φαινόμενο, που αναφέρεται ως αλυσίδα, συχνά θεωρείται ελάττωμα της μεθόδου. Τα συμπλέγματα που παράγονται με απλή σύνδεση μπορούν να παραβιάσουν την ιδιότητα να είναι "συμπαγής" ότι δηλαδή όλες οι παρατηρήσεις σε κάθε cluster τείνουν να μοιάζουν μεταξύ τους, με βάση τις παρεχόμενες διαφορές παρατήρησης $\{d_{ii'}\}$. Αν η διάμετρος D_G μιας ομάδας παρατηρήσεων οριστεί ως η μεγαλύτερη ανομοιότητα μεταξύ των μελών της

$$D_G = \max_{\substack{i \in G \\ i' \in G}} d_{ii'} \qquad (4.14)$$

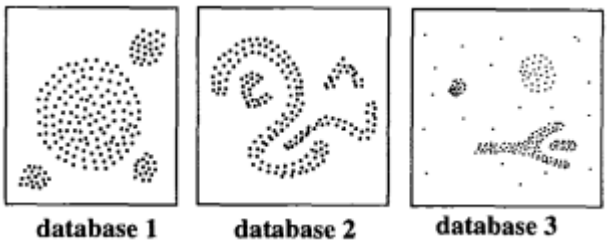
, τότε η απλή σύνδεση μπορεί να δημιουργήσει ομάδες με πολύ μεγάλες διαμέτρους. Η πλήρης σύνδεση (σύμφωνα με την εξίσωση 4.10) αντιπροσωπεύει το αντίθετο άκρο. Δύο ομάδες G και H θεωρούνται στενές μόνο εάν όλες οι παρατηρήσεις στην ένωση τους είναι σχετικά όμοιες. Θα τείνει να παράγει συμπαγείς ομάδες με μικρές διαμέτρους (σύμφωνα με την εξίσωση 4.12). Ωστόσο, μπορεί να δημιουργήσει ομάδες που παραβιάζουν την ιδιότητα της "εγγύτητας". Δηλαδή, οι παρατηρήσεις που έχουν ανατεθεί σε ένα cluster μπορεί να είναι πολύ πιο κοντά σε μέλη άλλων συμπλεγμάτων από ό, τι σε ορισμένα μέλη του δικού τους συμπλέγματος. Τέλος, η μέση ομαδοποίηση του συμπλέγματος αντιπροσωπεύει έναν συμβιβασμό μεταξύ των δύο άκρων της μονής και πλήρους σύνδεσης. Προσπαθεί να παράγει σχετικά συμπαγή σμήνη που είναι σχετικά μακριά μεταξύ τους. Ωστόσο, τα αποτελέσματά του εξαρτώνται από την αριθμητική κλίμακα στην οποία μετρούνται οι διαφορές παρατήρησης$d_{ii'}$. Η εφαρμογή ενός μονοτονικού αυστηρά αυξανόμενου μετασχηματισμού $h(\cdot)$ στο $d_{ii'}, h_{ii'} = h(d_{ii'})$, μπορεί να αλλάξει το αποτέλεσμα που παράγεται από την εξίσωση 4.13.



Σχήμα 3 - Δενδρογράμματα από συγκεντρωτική ιεραρχική ομαδοποίηση δεδομένων μικροσυστοιχίας ανθρώπινου όγκου. (Πηγή: Friedman et al., 2001).

4.1.5. Ο αλγόριθμος DBSCAN

Όταν εξετάζονται τα δείγματα των σημείων που απεικονίζονται στο σχήμα 4, διαπιστώνουμε ότι ένας διαχωριστικός αλγόριθμος όπως ο k-means δε θα έδινε τα σωστά αποτελέσματα. Ο κύριος λόγος είναι γιατί στις ομάδες υπάρχει μια τυπική πυκνότητα σημείων που είναι σημαντικά υψηλότερη από ό, τι έξω από το cluster. Επιπλέον, η πυκνότητα εντός των περιοχών θορύβου είναι χαμηλότερη από την πυκνότητα σε οποιοδήποτε από τα συμπλέγματα. Στη συνέχεια, γίνεται προσπάθεια να επισημοποιηθεί αυτή η διαισθητική έννοια των "συμπλεγμάτων" και του "θορύβου" σε μια βάση δεδομένων D σημείων κάποιου χώρου k-διάστασης S.



Σχήμα 4 - Δείγματα βάσεων δεδομένων. (Πηγή: Ester et al., 1996).

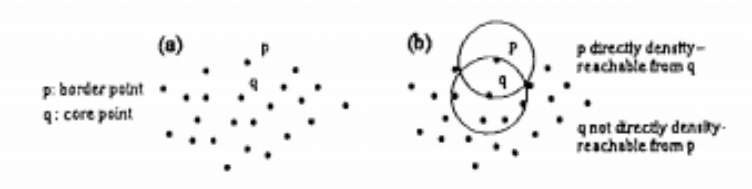
Πρέπει να σημειωθεί ότι τόσο η έννοια των συμπλεγμάτων όσο και ο αλγόριθμός χωρικής ομαδοποίησης εφαρμογών με θόρυβο με βάση την πυκνότητα (Density-based spatial clustering of applications with noise – DBSCAN), ισχύουν επίσης για Ευκλείδειους δισδιάστατους ή τρισδιάστατους χώρους (2D ή 3D) ως προς ορισμένους χώρους χαρακτηριστικών υψηλής διάστασης. Η βασική ιδέα είναι ότι για κάθε σημείο ενός συμπλέγματος η γειτονιά μιας δεδομένης ακτίνας πρέπει να περιέχει τουλάχιστον έναν ελάχιστο αριθμό σημείων, δηλαδή η πυκνότητα στη γειτονιά πρέπει να υπερβαίνει κάποιο όριο. Το σχήμα μιας γειτονιάς καθορίζεται από την επιλογή μιας συνάρτησης απόστασης για δύο σημεία p και q, που συμβολίζονται με dist (p, q). Για παράδειγμα, όταν χρησιμοποιείται η απόσταση Μανχάταν σε 2D χώρο, το σχήμα της γειτονιάς είναι ορθογώνιο. Επίσης, πρέπει να σημειωθεί ότι αυτή η προσέγγισή λειτουργεί με οποιαδήποτε συνάρτηση απόστασης, έτσι ώστε να μπορεί να επιλεγεί η κατάλληλη συνάρτηση για κάποια συγκεκριμένη εφαρμογή. Για σκοπούς σωστής απεικόνισης, όλα τα παραδείγματα θα βρίσκονται σε 2D χώρο χρησιμοποιώντας την Ευκλείδεια απόσταση.

Ορισμός 1: (Eps - γειτονιά ενός σημείου – Eps «Epsilon» - neighborhood of a point) Η γειτονιά Eps ενός σημείου p, που συμβολίζεται με $N_{Eps}(p)$, ορίζεται ως $N_{Eps}(p) = \{q \in D \mid \text{dist}(p, q) \leq Eps\}$. Μια αφελής προσέγγιση θα μπορούσε να απαιτήσει για κάθε σημείο σε ένα cluster να υπάρχουν τουλάχιστον ένας ελάχιστος αριθμός σημείων σε μια γειτονιά Eps αυτού του σημείου. Ωστόσο, αυτή η προσέγγιση αποτυγχάνει επειδή υπάρχουν δύο είδη σημείων σε ένα cluster, σημεία στο εσωτερικό του συμπλέγματος (σημεία πυρήνα) και σημεία στο όριο του συμπλέγματος (σημεία συνόρων). Γενικά, μια γειτονιά Eps ενός συνοριακού σημείου περιέχει σημαντικά λιγότερα σημεία από μια γειτονιά Eps ενός βασικού σημείου. Επομένως, θα πρέπει να οριστεί ο ελάχιστος αριθμός πόντων (Minimum points - MinPts) σε μία σχετικά χαμηλή τιμή με σκοπό να συμπεριληφθούν όλα τα σημεία που ανήκουν στο ίδιο cluster. Αυτή η τιμή, ωστόσο, δεν θα είναι χαρακτηριστική για το αντίστοιχο cluster - ιδιαίτερα όταν υπάρχει θόρυβος. Επομένως, απαιτείται για κάθε σημείο p σε ένα cluster C να υπάρχει ένα σημείο q στο C, έτσι ώστε το p να βρίσκεται μέσα στη γειτονιά Eps του q και το $N_{Eps}(q)$ να περιέχει τουλάχιστον σημεία MinPts. Ο ορισμός αυτός παρουσιάζεται παρακάτω.

Ορισμός 2: (άμεσα προσβάσιμο σε πυκνότητα) Ένα σημείο p είναι άμεσα προσβάσιμο σε πυκνότητα από ένα σημείο q σε σχέση με (wrt. - with respect to) Eps και MinPts, εάν

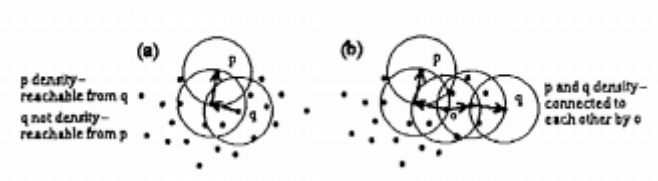
1. $p \in N_{Eps}(q)$
2. $N_{Ep}(q) | \geq MinPts$ (κατάσταση βασικού σημείου).

Προφανώς, η άμεση προσέγγιση πυκνότητας είναι συμμετρική για ζεύγη σημείων πυρήνα. Σε γενικές γραμμές, ωστόσο, δεν είναι συμμετρική εάν εμπλέκεται ένα κεντρικό σημείο και ένα οριακό σημείο. Το σχήμα 5 δείχνει την ασύμμετρη περίπτωση.



Σχήμα 5 - Βασικά και συνοριακά σημεία. (Πηγή: Ester et al., 1996).

Ορισμός 3: (προσβάσιμη σε πυκνότητα) Ένα σημείο p είναι προσβάσιμο σε πυκνότητα από ένα σημείο q wrt. Eps και MinPts εάν υπάρχει αλυσίδα σημείων $p_1, \ldots, p_n, p_1 = q, p_n = p$, τέτοια ώστε το p_{i+1} να είναι άμεσα προσβάσιμο σε πυκνότητα από το p_i. Η δυνατότητα προσέγγισης πυκνότητας είναι μια κανονική προέκταση της άμεσης δυνατότητας πυκνότητας. Αυτή η σχέση είναι μεταβατική, αλλά δεν είναι συμμετρική. Το σχήμα 6. απεικονίζει τις σχέσεις ορισμένων σημείων δειγμάτων και, ειδικότερα, την ασύμμετρη περίπτωση. Αν και δεν είναι συμμετρική γενικά, είναι προφανές ότι η πυκνότητα-εφικτότητας είναι συμμετρική για τα σημεία πυρήνα. Δύο συνοριακά σημεία του ίδιου συμπλέγματος C είναι πιθανό να μην είναι προσβάσιμα σε πυκνότητα το ένα από το άλλο επειδή η συνθήκη του βασικού σημείου μπορεί να μην ισχύει και για τα δύο. Ωστόσο, πρέπει να υπάρχει ένα κεντρικό σημείο στο C από το οποίο και τα δύο συνοριακά σημεία του C είναι προσβάσιμα σε πυκνότητα. Επομένως, εισάγεται η έννοια της πυκνότητας-συνδεσιμότητας που καλύπτει αυτή τη σχέση των συνοριακών σημείων.



Σχήμα 6 - Προσβάσιμο σε πυκνότητα και πυκνότητα-συνδεσιμότητα. (Πηγή: Ester et al., 1996).

Ορισμός 4: (συνδεδεμένη με πυκνότητα) Ένα σημείο p είναι η πυκνότητα που συνδέεται με ένα σημείο q wrt. Eps και MinPts εάν υπάρχει σημείο o τέτοιο ώστε και τα δύο, p και q να

είναι προσβάσιμα σε πυκνότητα από o wrt. Eps και MinPts. Η συνδεσιμότητα πυκνότητας είναι μια συμμετρική σχέση. Για σημεία προσβάσιμης πυκνότητας, η σχέση πυκνότητας-συνδεσιμότητας είναι επίσης ανακλαστική (βλ. Σχήμα 6). Τώρα, είναι εφικτό να οριστεί η έννοια του συμπλέγματος βάσει πυκνότητας. Διαισθητικά, ένα cluster ορίζεται ως ένα σύνολο συνδεδεμένων σημείων πυκνότητας που είναι το μέγιστο σε σχέση με πυκνότητα - εφικτότητα. Ο θόρυβος θα οριστεί σε σχέση με ένα δεδομένο σύνολο συμπλεγμάτων. Ο θόρυβος είναι απλά το σύνολο των σημείων στο D που δεν ανήκουν σε κανένα από τα συμπλέγματα του.

Ορισμός 5: (cluster) Έστω D μια βάση δεδομένων με σημεία. Ένα cluster C wrt. Eps και MinPts, είναι ένα μη κενό υποσύνολο του D που πληροί τις ακόλουθες προϋποθέσεις:

1. $\forall \mathrm{p}, \mathrm{q}$: εάν το $\mathrm{p} \in \mathrm{C}$ και το q είναι προσβάσιμα σε πυκνότητα από τα p wrt. Eps και MinPts, τότε $q \in C$. (Μεγιστότητα)
2. $\forall p, q \in C$: p συνδέεται με την πυκνότητα με τα q wrt. EPS και MinPts. (Συνδεσιμότητα)

Ορισμός 6: (θόρυβος) Έστω $C_1, \ldots, C_k$ τα συμπλέγματα της βάσης δεδομένων D wrt. με παραμέτρους Eps_i και $\mathrm{MinPts}_i, i = 1, \ldots, k$. Στη συνέχεια ορίζεται ο θόρυβος ως το σύνολο των σημείων στη βάση δεδομένων D που δεν ανήκουν σε κανένα cluster C_i , δηλαδή θόρυβος $= \{p \in D \mid \forall \mathrm{i}: \mathrm{p} \notin C_i\}$. Σημειώνεται ότι ένα cluster C wrt. Eps και MinPts περιέχει τουλάχιστον σημεία MinPts για τους ακόλουθους λόγους. Δεδομένου ότι το C περιέχει τουλάχιστον ένα σημείο p , το p πρέπει να είναι συνδεδεμένο με την πυκνότητα μέσω ενός σημείου o (το οποίο μπορεί να είναι ίσο με το p). Έτσι, τουλάχιστον το σημείο o πρέπει να πληροί την προϋπόθεση του βασικού σημείου και κατά συνέπεια, το γειτονικό Eps του o, να περιέχει τουλάχιστον σημεία MinPts.

Τα ακόλουθα λήμματα είναι σημαντικά για την επικύρωση της ορθότητας του αλγορίθμου ομαδοποίησης. Διαισθητικά, αναφέρουν τα εξής. Δεδομένων των παραμέτρων Eps και MinPts, μπορεί να ανακαλυφθεί ένα cluster σε μια προσέγγιση δύο βημάτων. Αρχικά, επιλέγεται ένα αυθαίρετο σημείο από τη βάση δεδομένων που πληροί την προϋπόθεση του βασικού σημείου ως σπόρου. Δεύτερον, γίνεται ανάκτηση όλων των σημείων που είναι προσβάσιμα σε πυκνότητα από τον σπόρο που λαμβάνει το cluster, το οποίο περιέχει τον σπόρο.

Λήμμα 1: Έστω ότι το p είναι ένα σημείο στο D και $\mathrm{IN}_{\mathrm{Eps}}(\mathrm{p})! \geq \mathrm{MinPts}$. Τότε το σύνολο $O = \{o \mid o \in D$ και o είναι προσβάσιμο σε πυκνότητα από το p wrt. Eps and MinPts$\}$ είναι ένα

cluster σε σχέση με τα Eps and MinPts. Δεν είναι προφανές ότι σε ένα cluster C wrt. Eps και MinPts καθορίζονται μοναδικά από οποιοδήποτε από τα βασικά του σημεία. Ωστόσο, κάθε σημείο στο C είναι προσβάσιμο σε πυκνότητα από οποιοδήποτε από τα βασικά σημεία του C και ως εκ τούτου, ένα cluster C περιέχει ακριβώς τα σημεία στα οποία η πυκνότητα είναι προσβάσιμη από ένα αυθαίρετο σημείο πυρήνα του C.

Λήμμα 2: Έστω ότι το C είναι ένα cluster wrt. Eps και MinPts και ας είναι το p οποιοδήποτε σημείο στο C με $\left|\mathrm{N}_{\mathrm{Eps}}(\mathrm{p})\right| \geq$ MinPts. Τότε το C ισούται με το σύνολο $O = \{$ {ο I ο και είναι προσβάσιμο σε πυκνότητα από τα p wrt. Eps και MinPts}.

DBSCAN: Χωρική ομαδοποίηση εφαρμογών με θόρυβο βάσει πυκνότητας

Σε αυτή την ενότητα, παρουσιάζεται ο αλγόριθμος DBSCAN ο οποίος έχει σχεδιαστεί για να ανακαλύπτει τα συμπλέγματα και το θόρυβο σε μια χωρική βάση δεδομένων σύμφωνα με τους ορισμούς 5 και 6. Στην ιδανική περίπτωση, θα πρέπει να έχουν μελετηθεί οι κατάλληλοι παράμετροι Eps και MinPts κάθε συμπλέγματος και τουλάχιστον ένα σημείο από το αντίστοιχο cluster. Στη συνέχεια, θα μπορούσαν να ανακτηθούν όλα τα σημεία που είναι προσιτά σε πυκνότητα από το δεδομένο σημείο χρησιμοποιώντας τις σωστές παραμέτρους.

Ο αλγόριθμος

Για να βρεθεί ένα cluster, το DBSCAN ξεκινά με ένα αυθαίρετο σημείο p και ανακτά όλα τα σημεία που είναι προσβάσιμα σε πυκνότητα από τα p wrt. Eps και MinPts. Εάν το p είναι ένα βασικό σημείο, αυτή η διαδικασία δίνει ένα cluster wrt. Eps και MinPts (βλ. Λήμμα 2). Εάν το p είναι ένα συνοριακό σημείο, κανένα σημείο δεν είναι προσβάσιμο σε πυκνότητα από το p και το DBSCAN επισκέπτεται το επόμενο σημείο της βάσης δεδομένων. Δεδομένου ότι χρησιμοποιούνται καθολικές τιμές για Eps και MinPts, το DBSCAN μπορεί σύμφωνα με τον ορισμό 5 να συγχωνεύσει δύο ομάδες σε ένα cluster, εάν δύο ομάδες διαφορετικής πυκνότητας είναι "κοντά" μεταξύ τους. Η απόσταση μεταξύ δύο συνόλων σημείων S_1 και S_2 ορίζεται ως dist $(\mathrm{S}_1,\ \mathrm{S}_2) = \min\{$ dist(p,q) | p ∈ S_1, q ∈ $\mathrm{S}_2\}$. Στη συνέχεια, δύο σύνολα σημείων που έχουν τουλάχιστον την πυκνότητα του λεπτότερου συμπλέγματος θα διαχωριστούν μεταξύ τους μόνο εάν η απόσταση μεταξύ των δύο συνόλων είναι μεγαλύτερη από Eps. Κατά συνέπεια, μια αναδρομική κλήση του DBSCAN μπορεί να είναι απαραίτητη για τις ομάδες που ανιχνεύθηκαν με υψηλότερη τιμή για MinPts. Αυτό, ωστόσο, δεν αποτελεί μειονέκτημα επειδή η επαναληπτική εφαρμογή του DBSCAN αποδίδει έναν κομψό και πολύ αποτελεσματικό βασικό

αλγόριθμο. Επιπλέον, η αναδρομική ομαδοποίηση των σημείων ενός συμπλέγματος είναι απαραίτητη μόνο υπό συνθήκες που μπορούν εύκολα να εντοπιστούν.

4.2. Προβλήματα Επιχειρηματικού Σχεδιασμού - Εταιρεία ενοικιάσεων ποδηλάτων

Οι εταιρείες κοινής χρήσης ποδηλάτων είναι πολύ διαδεδομένες στην Αμερική και με μεγάλη επισκεψιμότητα από το κοινό. Ωστόσο, έχει αρκετά προβλήματα ως προς την εξυπηρέτηση των πελατών. Είναι λογικό μιας και όταν η εξυπηρέτηση αφορά μια ολόκληρη πόλη είναι δύσκολο να βρεις το βέλτιστο τρόπο τοποθέτησης των ποδηλάτων και των περιπτέρων εξυπηρέτησης.

Το πρόβλημα της παρούσας εργασίας αφορά στην εταιρεία ενοικιάσεως ποδηλάτων Bike Sharing LLC (Gedron, 2016). Στόχος της εταιρείας είναι η κατασκευή έως τριών περιπτέρων εξυπηρέτησης του κοινού σε στρατηγικά σημεία της Ουάσιγκτον. Η ιδέα είναι η πρόσληψη υπαλλήλων για απασχόληση σε αυτά τα περίπτερα καθόλη την διάρκεια της ημέρας, παρέχοντας στους ποδηλάτες εξοπλισμό, χάρτες, νερό, σνακ, και gadgets. Επιπρόσθετες υπηρεσίες των υπάλληλων θα είναι και η καταχώρηση περιστασιακών χρηστών σε μόνιμους παρέχοντας πληροφορίες, απαντώντας σε ερωτήσεις των πελατών και δίνοντας τους οδηγίες. Τα δεδομένα τα αντλήσαμε από στην διαδεδομένη βάση Kaggle[3].

Με βάση αυτά τα δεδομένα θα πρέπει να γίνει διερεύνηση για την εύρεση καταλληλότερης θέσης τοποθέτησης των περιπτέρων όπως και η εύρεση του βέλτιστου αριθμού τους, ώστε να καλύπτουν επαρκώς την περιοχή. Τα δεδομένα που παρέχονται (bicycles.xlsx) για την επίλυση του προβλήματος, περιέχουν το γεωγραφικό μήκος και πλάτος 244 σταθμών της εταιρείας κατά μήκος της πόλης (Gerdon, 2016).

Η διοίκηση ζητά όχι μόνο της εύρεσης των καλύτερων τοποθεσιών αλλά και την επεξήγηση του γιατί θα πρέπει να τοποθετηθούν εκεί. Ο σκοπός του εμπεριέχει και την τοποθέτηση των περιπτέρων που έχουν την μεγαλύτερη πληθυσμιακή πυκνότητα, αλλά και την όσο το δυνατό μείωση των αποστάσεων μεταξύ τους και μεταξύ των σημείων διανομής ποδηλάτων. Το πρόβλημα αυτό ανήκει στην γενική κατηγορία προβλημάτων της επιχειρησιακής έρευνας που βασίζονται στον γραμμικό προγραμματισμό και έχουν πολλές εφαρμογές, όπως η τοποθέτηση κέντρων διανομής, κεντρικών αποθηκών.

[3] https://www.kaggle.com/

Η ίδια εταιρεία αντιμετωπίζει και δεύτερο πρόβλημα, το οποία σχετίζεται με την διερεύνηση των οικονομικών δεδομένων των υπαρχόντων πελατών. Τα δεδομένα αφορούν τις ηλικίες και το ετήσιο εισόδημα περισσότερων από 8.000 πελατών της εταιρείας (income.xlsx). Οι ηλικίες κυμαίνονται μεταξύ 18 και 89 ενώ το εισόδημα μεταξύ 233 και 178676 δολάρια. Η διοίκηση θέλει να γνωρίζει τα κοινά την δυνατή τμηματοποίηση αυτών με βάση τα κοινά χαρακτηριστικά τους. Για το σκοπό αυτό θα εξεταστούν διάφορες πιθανές ομάδες για να ανακαλυφθούν τυχόν πρότυπα (patterns). Επιπλέον θα εξεταστούν και θα συγκριθούν το μέγεθος της κάθε ομάδας, η διάμεσος, η μέγιστη και η ελάχιστη τιμή της ηλικίας και του εισοδήματος ανά ομάδα. Σε αυτά τα δύο σετ δεδομένων εφαρμόσθηκαν όλοι οι αλγόριθμοι ομαδοποίησης. Περισσότερες λεπτομέρειες αναφέρονται στο επόμενο υποκεφάλαιο.

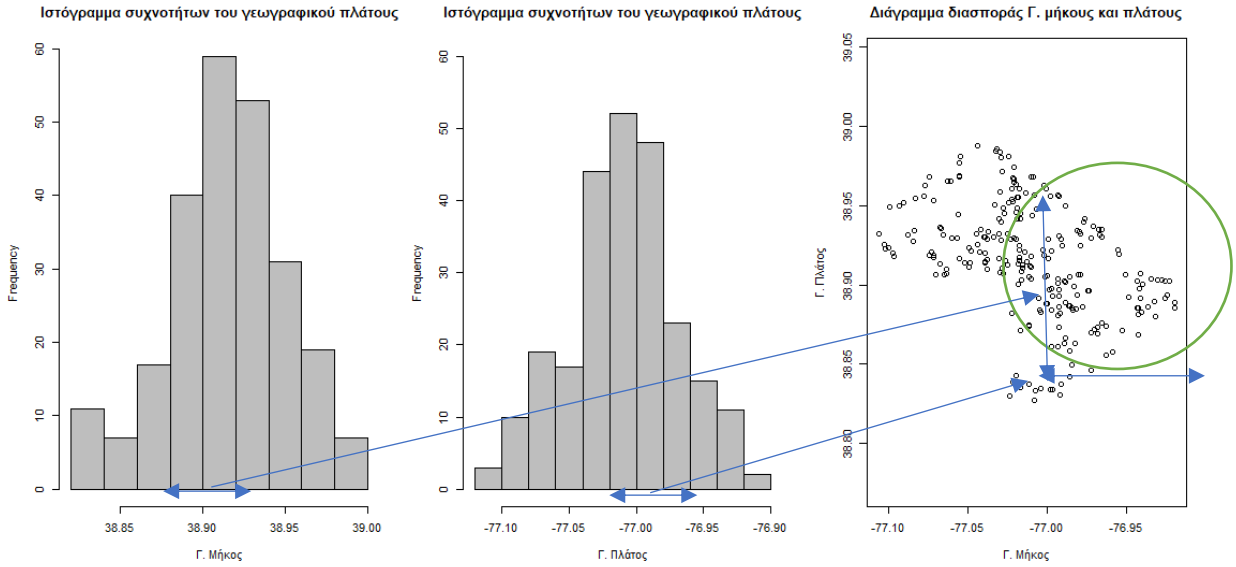
4.3. Αποτελέσματα

Σε αυτή την παράγραφο γίνεται η παρουσίαση των αποτελεσμάτων της επίλυσης του προβλήματος, ενώ ο κώδικας των υπολογισμών παρουσιάζεται στο παράρτημα της εργασίας. Αρχικά παρατηρούμε κάποια χαρακτηριστικά για το σετ δεδομένων του 1^{ου} προβλήματος, τα οποία αφορούν συγκεντρωτικές τιμές και κατανομές του γεωγραφικού μήκους και πλάτους (2 διαστάσεις) των θέσεων των 244 δειγμάτων μας. Τα αποτελέσματα του πίνακα 1 δείχνουν ότι δεν υπάρχει κάποια απούσα τιμή στα ζεύγη των συντεταγμένων των πιθανών θέσεων των περιπτέρων.

Πίνακας 1 - Χαρακτηριστικά του σετ δεδομένου του 1ου προβλήματος

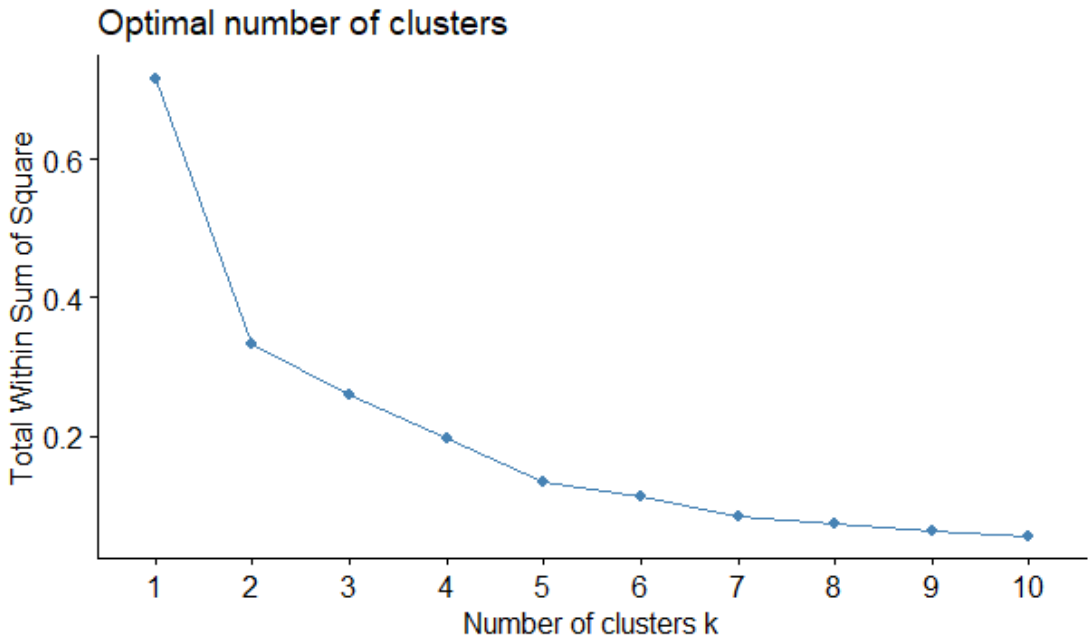
	Γ. Μήκος	**Γ. Πλάτος**
Ελάχιστη τιμή	38.83	-77.11
Πρώτο τεταρτημόριο	38.89	-77.03
Διάμεσος	38.92	-77.01
Μ.Τ.	38.91	-77.01
Τρίτο τεταρτημόριο	38.94	-76.99
Μέγιστη τιμή	38.99	-76.92

Μέσα από το ιστόγραμμα του γραφήματος (βλ. Σχήμα 7) παρατηρούμε την συγκέντρωση και συμμετρική κατανομή των δεδομένων γύρω από την μέση τιμή τους. Επιπλέον δεν παρατηρούνται κάποιες εμφανείς ομαδοποιήσεις από το διάγραμμα διασποράς.



Σχήμα 7 - Ιστογράμματα συχνοτήτων και διάγραμμα διασποράς του γεωγραφικού μήκους και πλάτους.

Μετέπειτα, εξετάζουμε το βέλτιστο αριθμό ομάδων στο πρόβλημα εύρεσης των κατάλληλων σημείων για περίπτερα εξυπηρέτησης. Ως αλγόριθμος ομαδοποίησης επιλέχθηκε ο k-means και ως κριτήριο αξιολόγησης το άθροισμα των τετραγώνων των αποστάσεων ανά ομάδα. Οι αποστάσεις είναι για κάθε σημείο από το κοντινότερο του κέντρο, δηλαδή την ομάδα που ανήκει. Στο σχήμα 8 φαίνεται το αποτέλεσμα της εύρεσης του βέλτιστου αριθμού περιπτέρων στο οποίο παρατηρούμε ότι με περισσότερες ομάδες το σφάλμα μειώνεται. Αυτό είναι λογικό, καθώς με περισσότερα περίπτερα, το κάθε ένα χρειάζεται να καλύπτει μικρότερο χώρο. Για παράδειγμα αν είχαμε όσα περίπτερα είναι και τα σημεία, το σφάλμα αυτό θα ήταν μηδέν μιας και θα είχαμε ένα περίπτερο σε κάθε σημείο. Κάτι το οποίο δεν είναι ρεαλιστικό. Στο σχήμα μας ενδιαφέρει κυρίως η απότομη αλλαγή της κλίσης του διαγράμματος το οποίο συμβαίνει από το K=1 στο K=2. Συνοπτικά, θα λέγαμε ότι τα δύο ή τρία περίπτερα είναι μια καλή επιλογή, δηλαδή K=2 ή K=3.



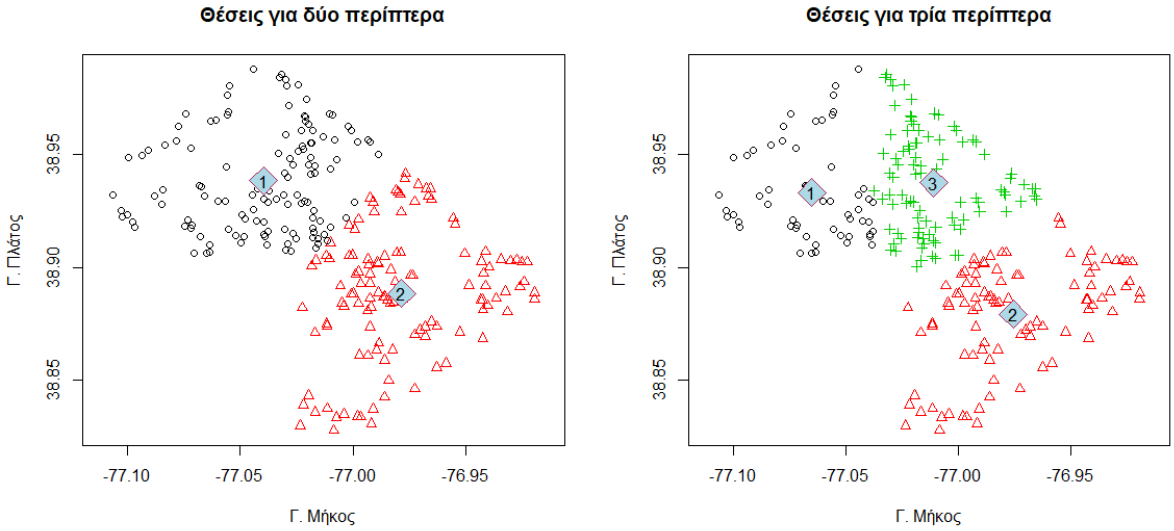
Σχήμα 8 - Εύρεση βέλτιστου αριθμού ομάδων για τον αλγόριθμο K-means

Αλγόριθμος K-means

Εφαρμόζοντας τον k-means για k=2 και k=3 (βλ. Σχήμα 9) ομάδες, παρατηρούμε ότι η ομαδοποίηση με 3 ομάδες, παράγει μεγαλύτερη ερμηνευτικότητα (διασπορά μεταξύ των ομάδων/συνολική διασπορά) από την περίπτωση των δύο ομάδων (Πίνακας 2).

Πίνακας 2 - Σύγκριση αποτελεσμάτων της μεθόδου k-means.

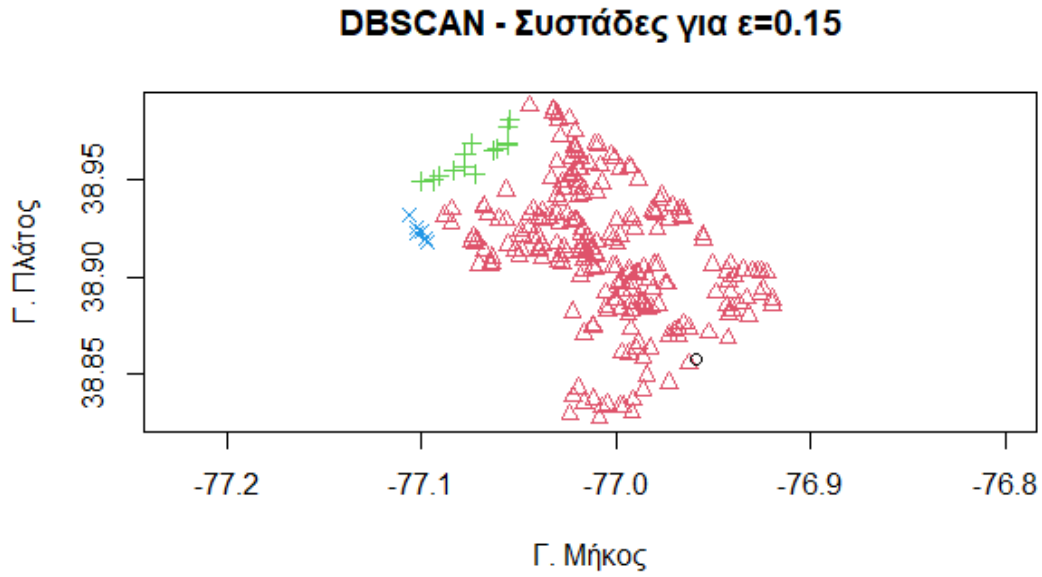
Ομάδες	Μέσες τιμές		**Ερμηνευτικότητα**
	Γ. Μήκος	**Γ. Πλάτος**	
2	38,93855	-77,03975	53,4%
	38,88838	-76,97846	
3	38,93327	-77,06502	63,7%
	38,87904	-76,97566	
	38,93765	-77,01089	



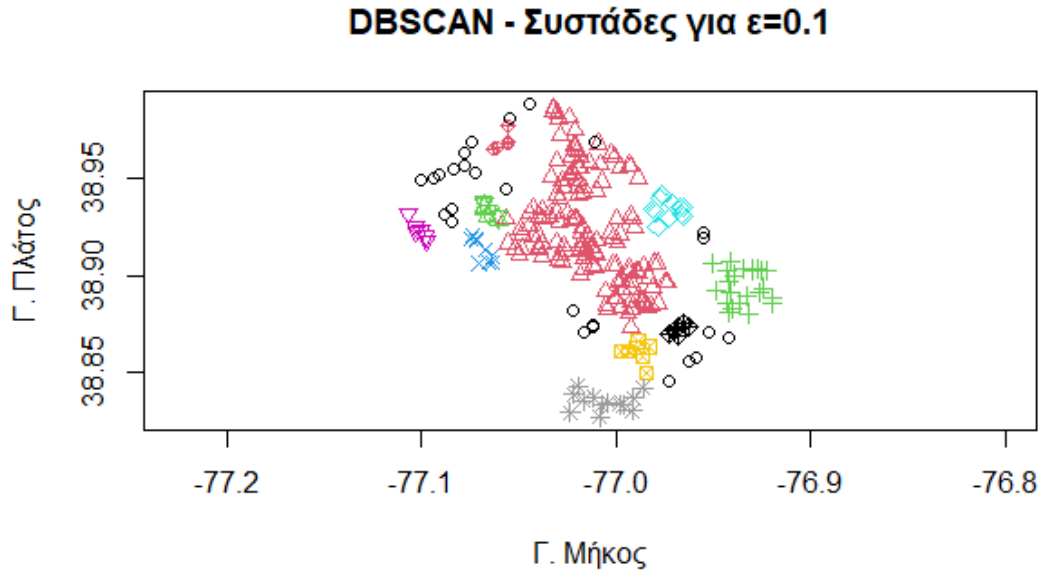
Σχήμα 9 - Αποτελέσματα ομαδοποίησης για k=2 και k=3

Αλγόριθμος DBSCAN

Ο αλγόριθμος DBSCAN έχει μια παράμετρο ε, η οποία ορίζει έμμεσα τον αριθμό των ομάδων. Στην Εικόνα 10 φαίνεται το αποτέλεσμα για ε = 0.15, όπου προκύπτουν 4 ομάδες και στην Εικόνα 11 για ε = 0.1, όπου προκύπτουν 11 ομάδες. Παρατηρούμε ότι αυτός ο αλγόριθμος δεν φαίνεται να είναι κατάλληλος για αυτό το σύνολο δεδομένων, καθώς σπάει σε ξεχωριστές ομάδες μόνο περιφερειακά σημεία. Μιας και αυτός ο αλγόριθμος βασίζεται στην πυκνότητα των σημείων, ουσιαστικά ανιχνεύει μια μεγάλη πυκνή ομάδα την οποία δεν μπορεί να διασπάσει ικανοποιητικά σε μικρότερες ομάδες.



Σχήμα 10 - Ομαδοποίηση με τον αλγόριθμο DBSCAN για ε = 0.15



Σχήμα 11 - Ομαδοποίηση με τον αλγόριθμο DBSCAN για ε = 0.1

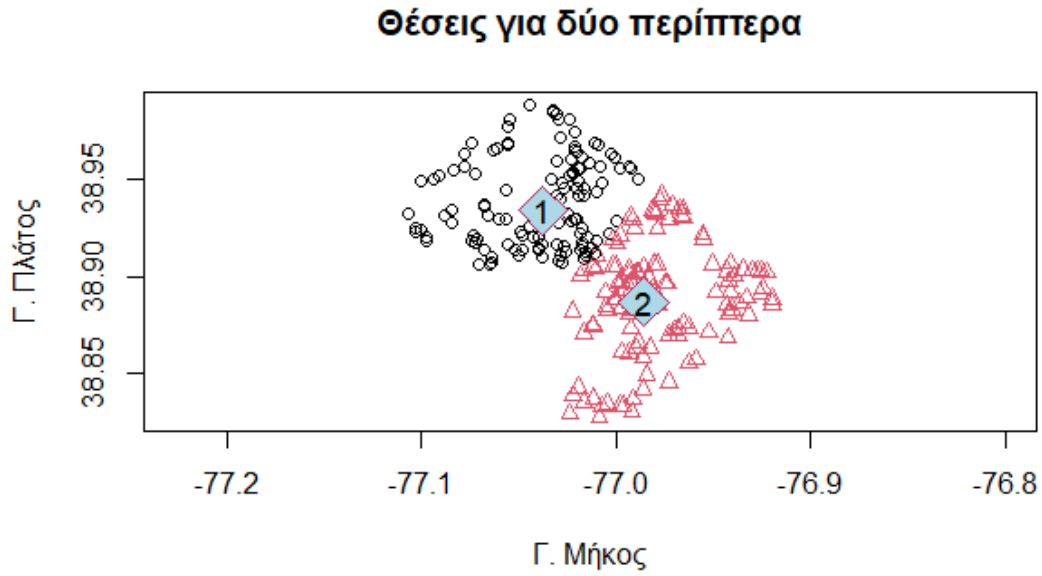
K-medoid

Αυτός ο αλγόριθμος λειτουργεί όπως ο K-means, αλλά επιλέγει ένα από τα υπάρχοντα σημεία ως κέντρο. Για K=2 ομάδες, στον πίνακα 3 φαίνονται οι συντεταγμένες των δύο

σημείων που επιλέχθηκαν ως κέντρα και στην Εικόνα (βλ. Σχήμα 12) το αποτέλεσμα της ομαδοποίησης.

Πίνακας 3 - Συντεταγμένες των δύο σημείων που επιλέχθηκαν ως κέντρα

Σημείο	**Γ. Μήκος**	**Γ. Πλάτος**
132	-77.03739	38.93387
113	-76.98574	38.88692

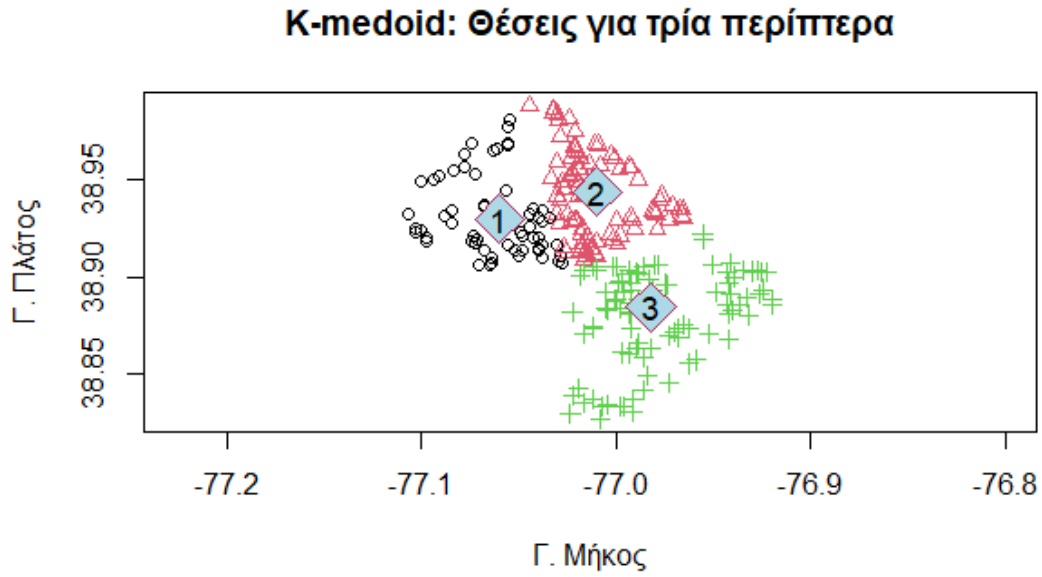


Σχήμα 12 - Ομαδοποίηση με τον αλγόριθμο K-medoids για K=2

Ομοίως για K=3, τα αποτελέσματα φαίνονται στον Πίνακα 4 και στην Εικόνα (βλ. Σχήμα 13).

Πίνακας 4 - Αποτελέσματα για k=3

Σημείο	**Γ. Μήκος**	**Γ. Πλάτος**
144	-77.06007	38.92952
190	-77.00982	38.94369
239	-76.98199	38.88445

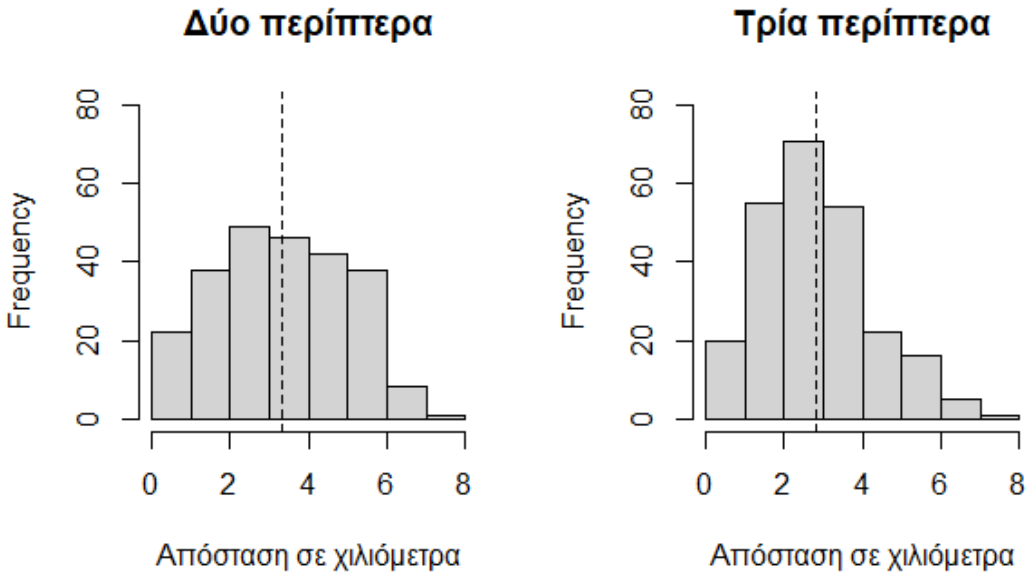


Σχήμα 13 - Ομαδοποίηση με τον αλγόριθμο K-medoids για K=3.

Στον Πίνακα 5 έχουμε τις μετρικές μέση και μέγιστη απόσταση σε χλμ. Παρατηρούμε ότι το αποτέλεσμα με 3 ομάδες είναι καλύτερο και με τις δύο μετρικές. Σε σύγκριση με τον K-means, επίσης έχουμε καλύτερα αποτελέσματα. Η κατανομή των αποστάσεων φαίνεται στην Εικόνα (βλ. Σχήμα 14).

Πίνακας 5 - Μετρικές μέση και μέγιστη απόσταση σε χλμ

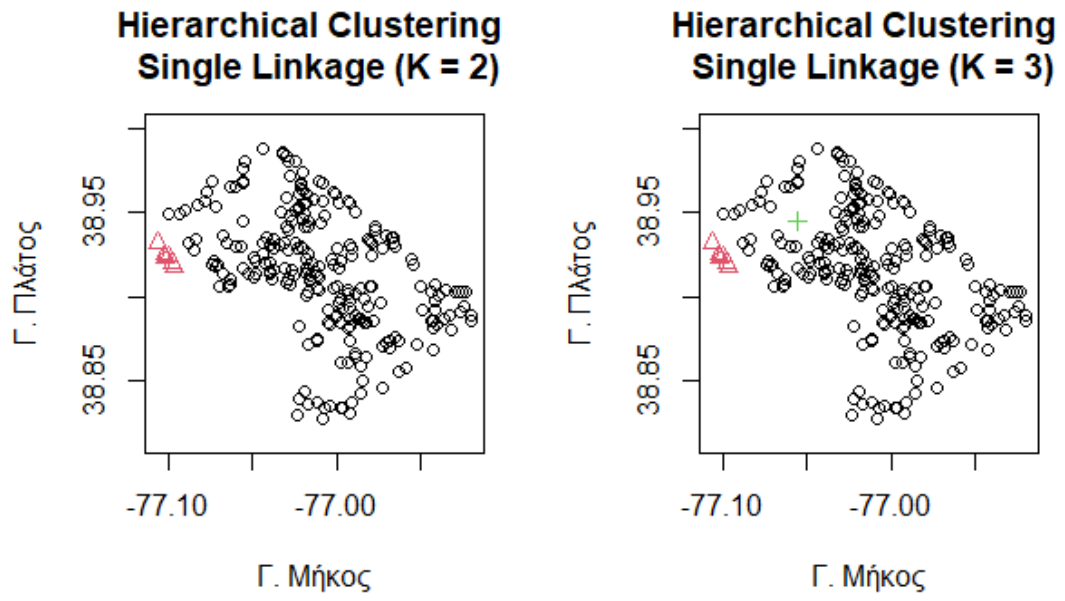
Μέτρο	**2 Ομάδες**	**3 Ομάδες**
Μέση Απόσταση	3.304	2.832
Διάμεση Απόσταση	3.239	2.690
Μέγιστη Απόσταση	7.160	7.076



Σχήμα 14 - Κατανομή αποστάσεων από κεντροειδή

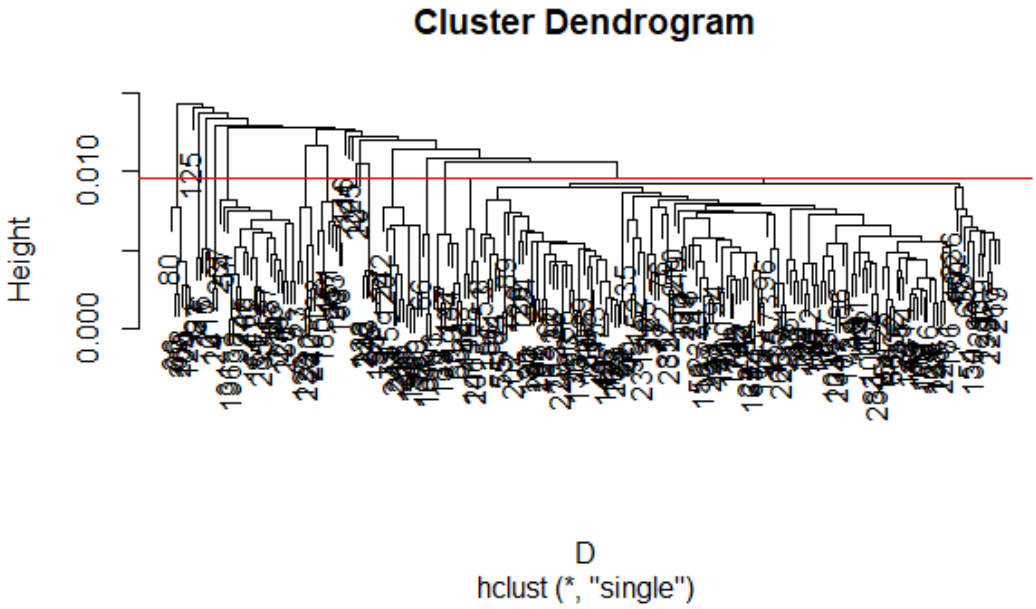
Ιεραρχικός αλγόριθμος – Single Linkage

Δοκιμάσαμε Ιεραρχικό αλγόριθμο με single linkage για K =2 και K = 3 ομάδες (βλ. Σχήμα 15 και 16 αντίστοιχα). Όπως και στην περίπτωση του DBSCAN, ο αλγόριθμος ξεχωρίζει σε διαφορετικές ομάδες μερικά σημεία που είναι πιο μακριά από τα υπόλοιπα, οπότε δεν είναι κατάλληλος για αυτά δεδομένα.



Σχήμα 15 - Ομαδοποίηση με ιεραρχικό αλγόριθμο με single linkage για K=2 και K=3

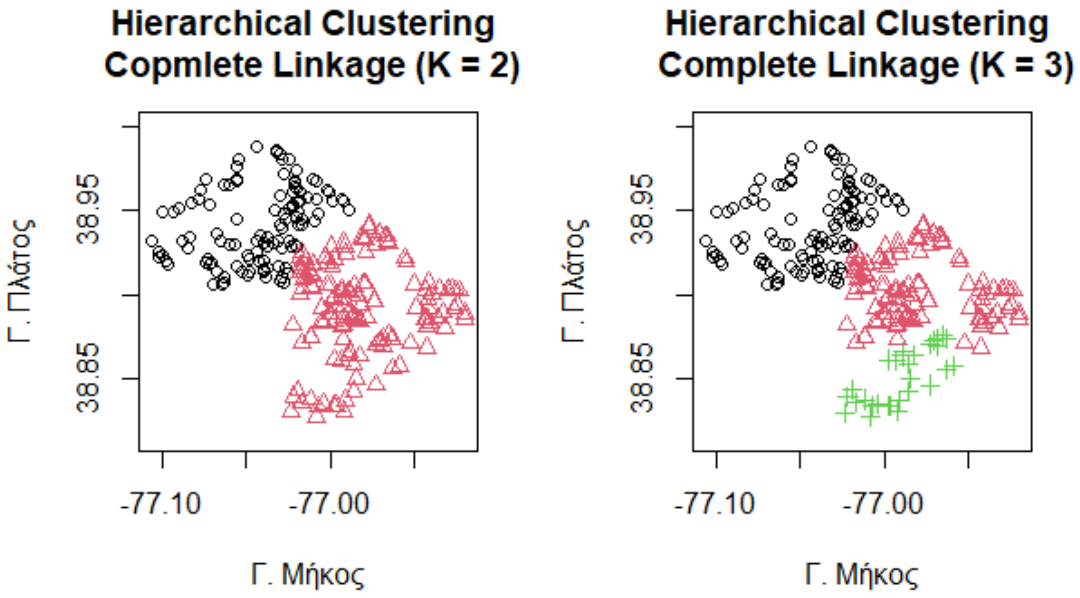
Από το δενδρόγραμμα βρίσκουμε ότι το μέγιστο ύψος δίνει 17 ομάδες, ωστόσο οι αποστάσεις είναι μικρές και ουσιαστικά δεν υπάρχει σαφής διαχωρισμός σε ομάδες.



Σχήμα 16 – Δενδρόγραμμα ομαδοποίησης με ιεραρχικό αλγόριθμο με single linkage

Ιεραρχικός αλγόριθμος – complete linkage

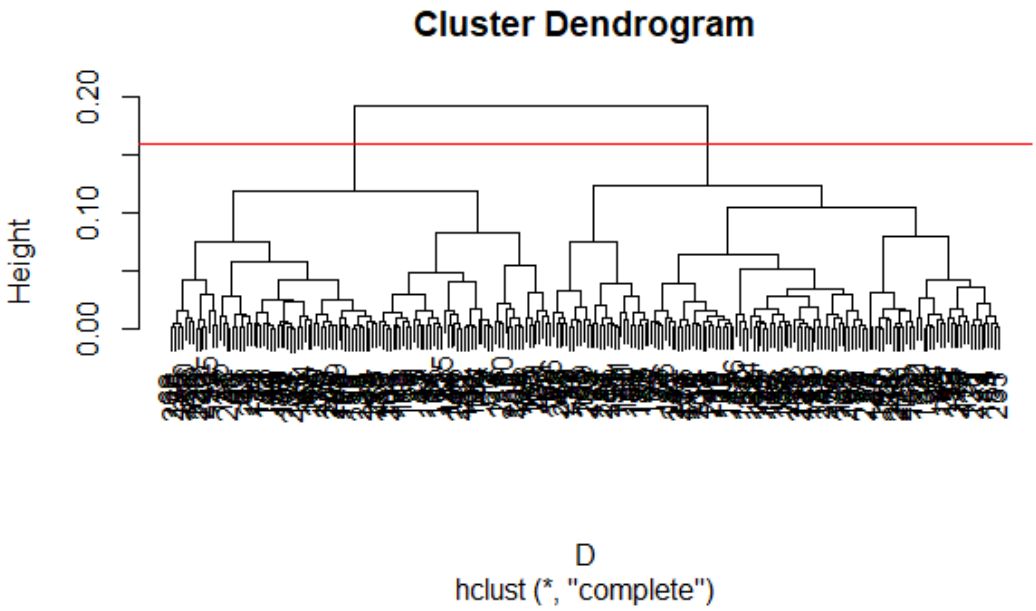
Ομοίως με πριν, δοκιμάσαμε με complete linkage (Σχήμα 17). Τώρα φαίνεται ότι όντως μπορεί να διαχωρίσει κατάλληλα τις ομάδες.



Σχήμα 17 - Ομαδοποίηση με ιεραρχικό αλγόριθμο με complete linkage για K=2 και K=3

Εικόνα 8. Ομαδοποίηση με ιεραρχικό αλγόριθμο με complete linkage για K=2 και K = 3.

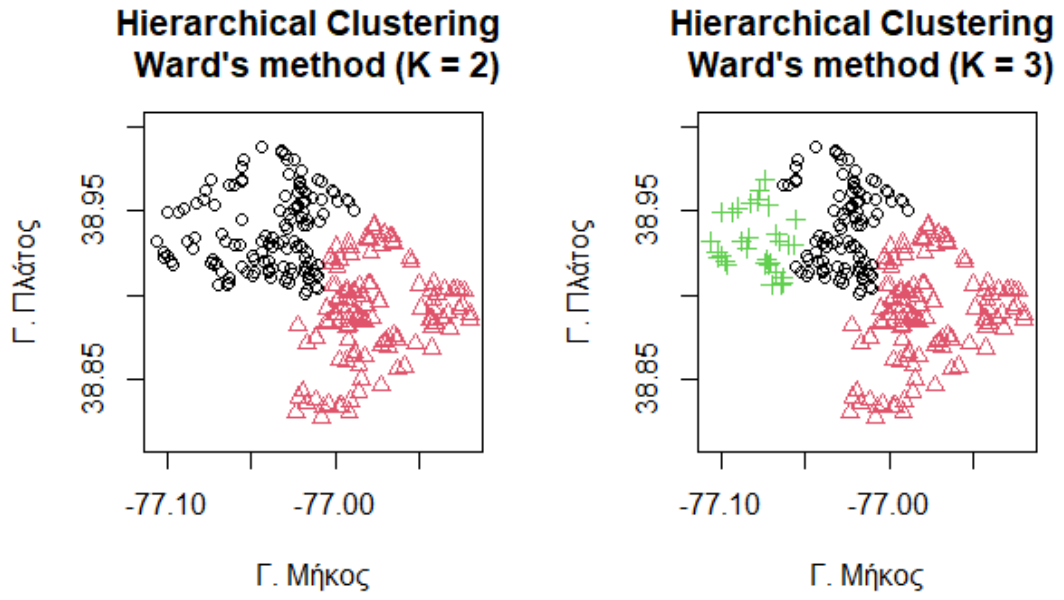
Από το δενδρόγραμμα βρίσκουμε ότι το μέγιστο ύψος δίνει 2 ομάδες



Σχήμα 18 - Δενδρόγραμμα ομαδοποίησης με ιεραρχικό αλγόριθμο με complete linkage

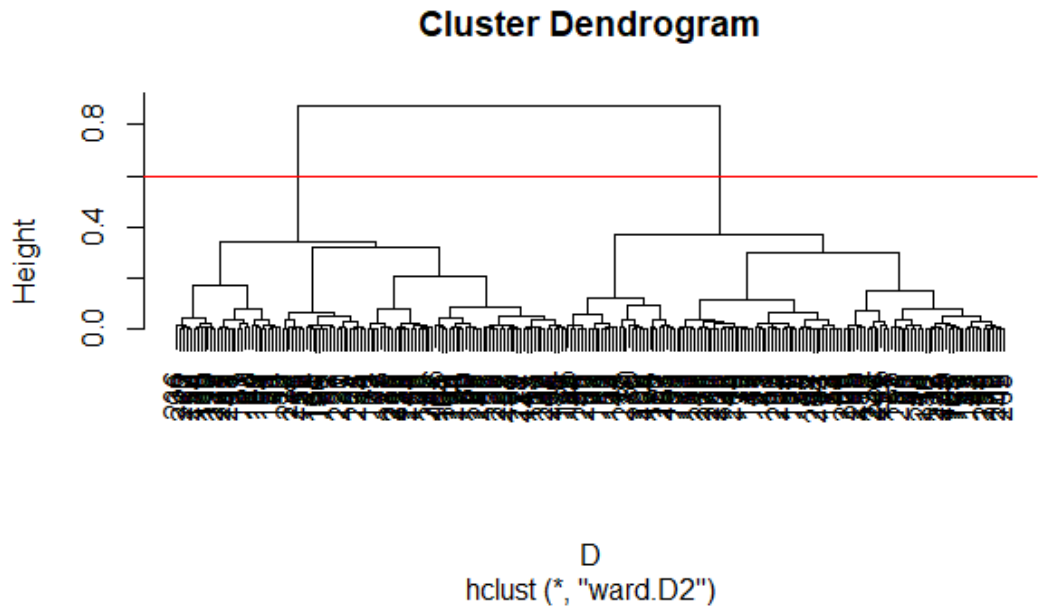
Ιεραρχικός αλγόριθμος – Ward method

Τέλος, δοκιμάσαμε ιεραρχικό αλγόριθμο με τη μέθοδο Ward (Σχήμα 19). Σε αυτή την περίπτωση, ο αλγόριθμος μπορεί να διαχωρίσει κατάλληλα τις ομάδες.



Σχήμα 19 - Ομαδοποίηση με ιεραρχικό αλγόριθμο με Ward για K=2 και K=3

Από το δενδρόγραμμα βρίσκουμε ότι το μέγιστο ύψος δίνει 2 ομάδες



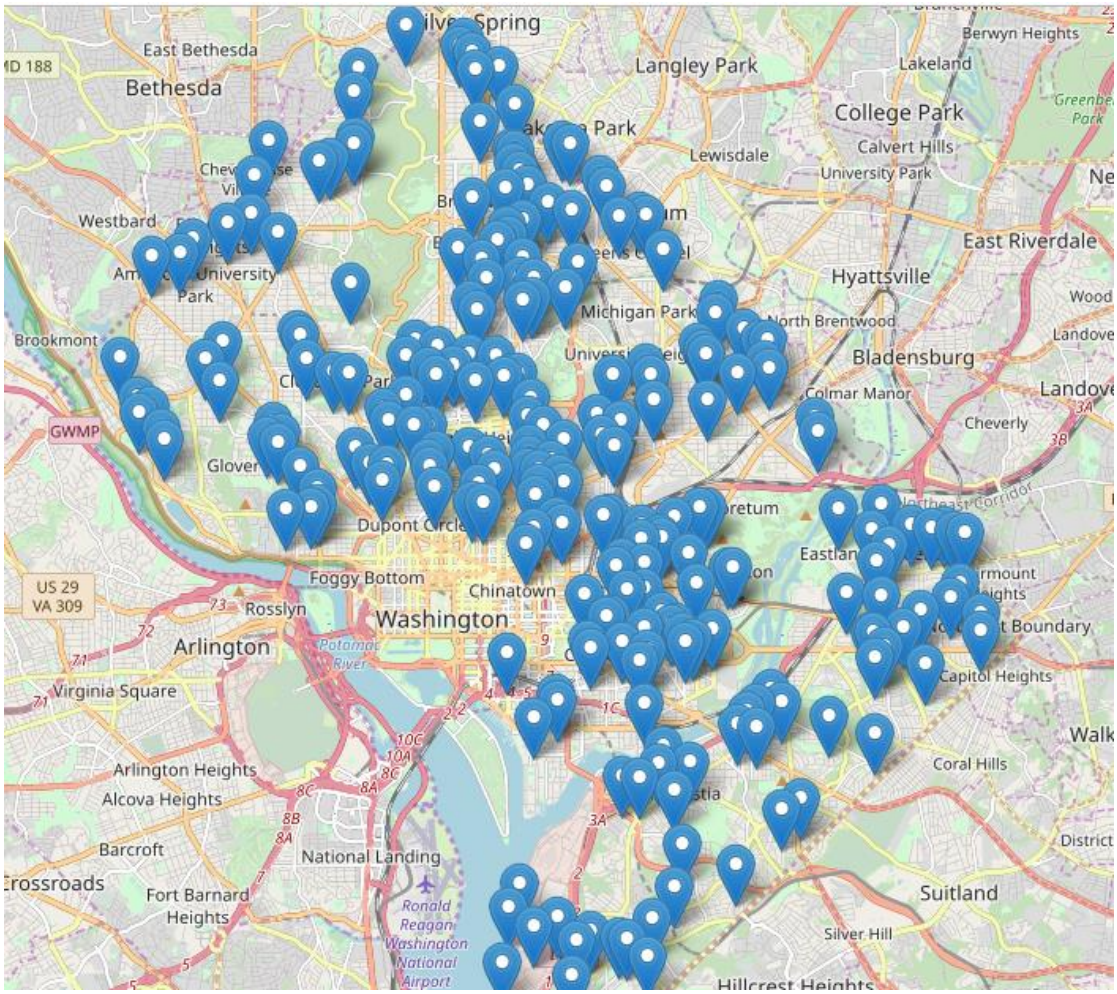
Σχήμα 20 - Δενδρόγραμμα ομαδοποίησης με ιεραρχικό αλγόριθμο με Ward

Όλες οι περιπτώσεις στις εφαρμογές των αλγορίθμων ομαδοποίησης αξιολογήθηκαν από το εσωτερικό μέτρο silhouette, για να εάν αποτιμήσουμε την καταλληλότητα ή ποιότητα της ομαδοποίησης. Το μέτρο silhouette (s) ελέγχει την ομοιότητα κάθε σημείου με τα σημεία που ανήκουν στο ίδιο cluster. Τα σημεία s(i) κοντά στο ένα είναι πολύ καλά ομαδοποιημένες, τα σημεία κοντά στο μηδέν σημαίνει ότι βρίσκονται μεταξύ δύο ομάδων και τα σημεία με αρνητικό s(i) πιθανότατα τοποθετούνται στο λάθος cluster. Παρατηρούμε ότι τα 2 cluster από τους αλγορίθμους k-means και k-medoid είναι η καλύτερη επιλογή.

Πίνακας 6 - Μέση μέτρηση της μετρικής silhouette για κάθε αλγόριθμο ομαδοποίησης

Μέθοδος	**Τιμή Silhouette**
Kmeans K = 2	0.448
Kmeans K = 3	0.405
DBSCAN (e = 0.15, 4 clusters)	-0.074
DBSCAN (e = 0.10, 11 clusters)	0.060
KMedoid K = 2	0.448
KMedoid K = 3	0.373
Hierarchical – Single Linkage K = 2	0.285
Hierarchical – Single Linkage K =3	-0.018
Hierarchical – Complete Linkage K = 2	0.436
Hierarchical – Complete Linkage K =3	0.352
Hierarchical – Ward’s method K = 2	0.446
Hierarchical – Ward’s method K =3	0.370

Καθώς το κοινό στο οποίο θα πρέπει να γίνει η παρουσίαση των αποτελεσμάτων δεν είναι σίγουρο ότι θα μπορέσει να εξάγει την απαραίτητη πληροφορία που θέλει να μεταδώσει ο αναλυτής, είναι σημαντικό ότι τα γραφήματα θε πρέπει να γίνουν πιο φιλικά για τον χρήστη. Για τον λόγο αυτό κρίθηκε αναγκαία μια πιο απλή παρουσίαση των σημείων διανομής με την βοήθεια του geo mapping δηλαδή της αναπαράστασης των σημείων σε πραγματικό χάρτη (Σχήμα 17)



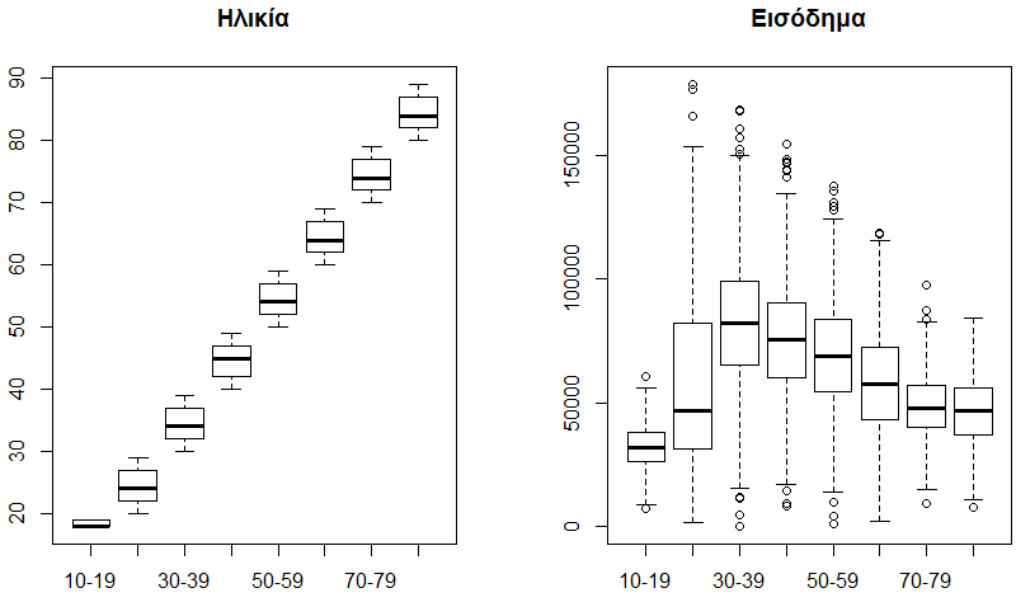
Σχήμα 21 - Γεωγραφική απεικόνιση των σημείων διανομής

Σχετικά με το δεύτερο πρόβλημα και με τα εισοδήματα των χρηστών των ποδηλάτων, ξεκινάμε με την εισαγωγική περιγραφή των δεδομένων. Τα αποτελέσματα (Σχήμα 18) δείχνουν ότι δεν υπάρχουν απούσες τιμές στο δείγμα, η κατηγοριοποίηση των ηλικιών έχει γίνει χωρίς σφάλματα όπως και ότι υπάρχει πιθανή σχέση μεταξύ ηλικίας και εισοδήματος. Με την εξέταση του συντελεστή συσχέτισης του Pearson βρέθηκε ότι όντως παρουσιάζεται ένα τέτοιο είδος σχέσης (ρ=-0,0599, Sig.<0.01) η οποία όμως είναι σχεδόν μηδενική και υπονοεί ότι η αύξηση της ηλικία του δείγματος έχει ως αποτέλεσμα την πιθανή εμφάνιση χαμηλού εισοδήματος. Η συσχέτιση αυτή γίνεται πιο ξεκάθαρη με την διαίρεση του δείγματος σε δύο κατηγορίες· μικρότερη των 40 ετών και μεγαλύτερη ή ίση των 40, με την βοήθεια του γραφήματος 4.10. αποτελέσματα δείχνουν ότι για ηλικίες μικρότερες των 40 ετών η συσχέτιση είναι θετική (ρ=0.5624) και για μεγαλύτερες των 40 ετών αρνητική (ρ=-0,4674). Δηλαδή,

αύξηση της ηλικίας, μέχρι τα 40 σημαίνει και αύξηση του εισοδήματος και αντίστροφα για μετά τα 40 έτη.

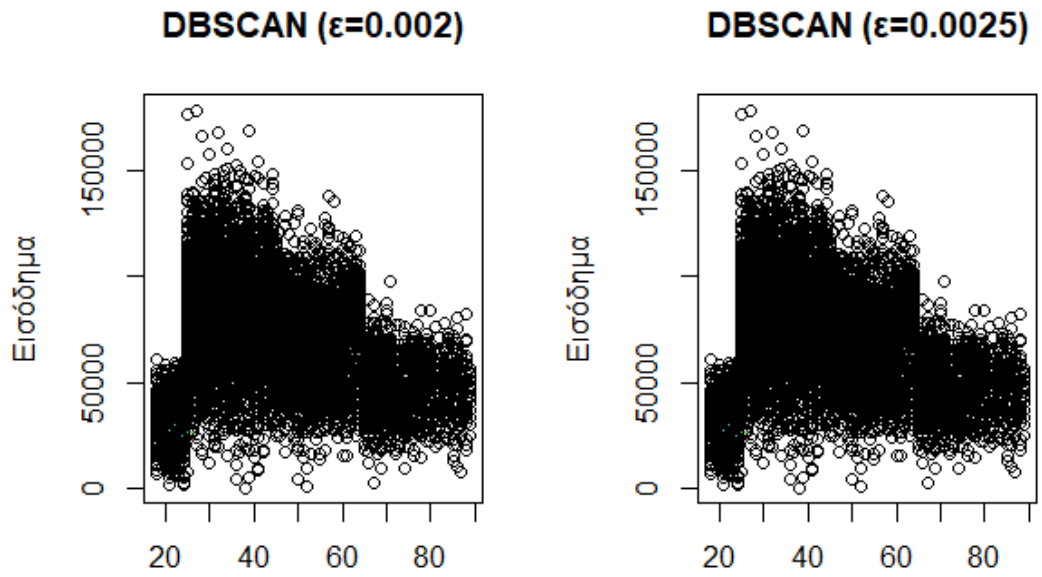
Πίνακας 7 - Βασικά μέτρα θέσης των μεταβλητών των δεδομένων του παραδείγματος ιεραρχικής ομαδοποίησης

	Ηλικία	**Εισόδημα**
Ελάχιστη τιμή	18,00	233
Πρώτο τεταρτημόριο	28,00	43792
Διάμεσος	39,00	65059
Μ.Τ.	42,85	66223
Τρίτο τεταρτημόριο	55,00	85944
Μέγιστη τιμή	89,00	178676



Σχήμα 22 – Θηκογράμματα (Boxplots) ηλικιών και εισοδήματος ανά ηλικιακή κατηγορία

Σχετικά με τους αλγορίθμους ομαδοποίησης, υλοποιήσαμε και εφαρμόσαμε τον k-means για K=5 και K=6. Για τον αλγόριθμο DBSCAN, για ε = 0,002, το αποτέλεσμα έχει 4 ομάδες. Ωστόσο τα μεγέθη των ομάδων είναι 8089, 5, 6, 5 , δηλαδή σχεδόν όλα τα σημεία είναι σε μια ομάδα. Οπότε και σε αυτή την περίπτωση τα δεδομένα με την κατανομή που παρουσιάζουν στο χώρο δεν ενδείκνυνται για ανίχνευση ομάδων με βάση την πυκνότητα.



Σχήμα 23 - Ο αλγόριθμος DBSCAN για διάφορες τιμές της παραμέτρου ε παράγει μια μεγάλη ομάδα η οποία περιέχει σχεδόν όλα τα σημεία, οπότε δεν διακρίνονται οι μικρότερες ομάδες

Τέλος, δοκιμάσαμε τον K-medoid για K=5 και K=6. Για τον K-medoid, τα κέντρα για K=5 και K=6 φαίνονται στους παρακάτω πίνακες. Τα κέντρα είναι όντως αντιπροσωπευτικά των δεδομένων, καθώς έχουμε διαφορετικές ηλικίες και επίπεδα εισοδήματος.

Πίνακας 8 – Αποτελέσματα για τον K-medoid, για K=5

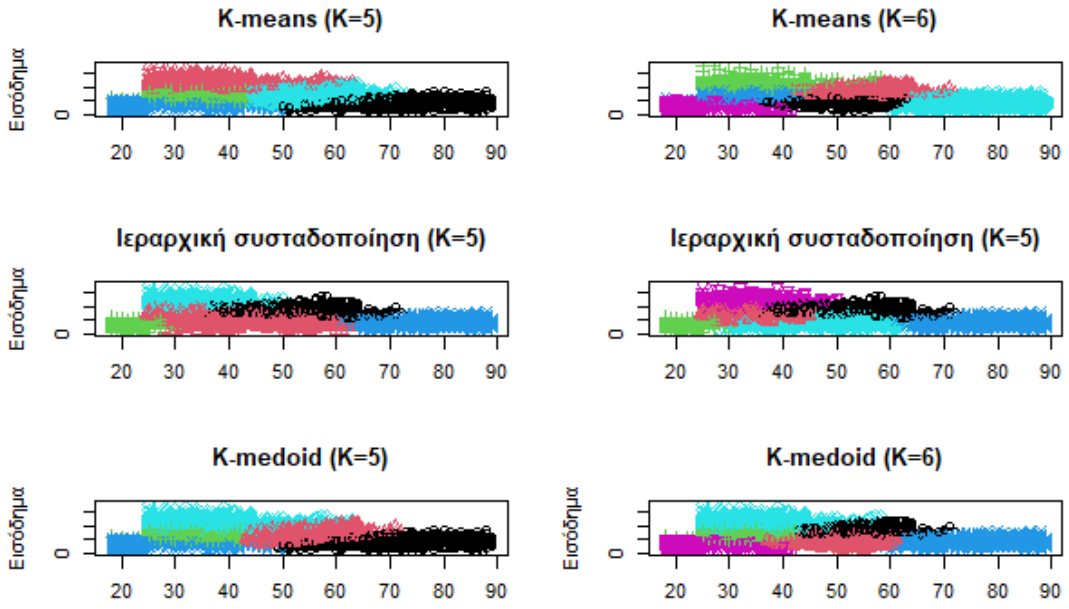
Σημείο	**Ηλικία**	**Εισόδημα**
7616	71	45514
7328	53	70025
6153	33	70394
7369	22	32658
6759	35	102456

Πίνακας 9 - Αποτελέσματα για τον K-medoid, για K=6

Σημείο	**Ηλικία**	**Εισόδημα**
7436	54	81177
396	50	52559
2388	33	71610

7740	73	46754
5198	34	103673
7369	22	32658

Το αποτέλεσμα όλων των αλγορίθμων φαίνεται παρακάτω.



Σχήμα 24 - Σύγκριση αλγορίθμων ομαδοποίησης

Κεφάλαιο 5. Συζήτηση - Συμπεράσματα

Σκοπός αυτής της μελέτης ήταν η θεωρητική και εμπειρική εξέταση εφαρμογής μεθόδων ομαδοποίησης στον επιχειρηματικό σχεδιασμό. Επιπλέον, έγινε αναφορά και στα αποτελέσματα εφαρμογής αυτών των μεθόδων με την βοήθεια ρεαλιστικών παραδειγμάτων. Οι επιχειρήσεις συχνά χαρακτηρίζονται από ανομοιογενείς περιορισμούς, οι οποίοι αποτελούν παράγοντα επιβράδυνσης ή αναστολής στις επιχειρηματικές προσδοκίες ή την ανάπτυξη τους. Συχνά απροσδόκητες μεταβολές επηρεάζουν τις επιχειρήσεις, νεοσύστατες και μη, με αποτέλεσμα να διακόψουν τη σειρά των επιχειρηματικών τους δραστηριοτήτων.

Πιο συγκεκριμένα, στόχος της εργασίας ήταν να μελετήσουμε διάφορους παραδοσιακούς και γνωστούς αλγόριθμους ομαδοποίησης πάνω σε πραγματικά προβλήματα επιχειρηματικού σχεδιασμού. Υλοποιήσαμε και εφαρμόσαμε τέσσερις αλγόριθμους ομαδοποίησης. Τον k-means, k-medoids, dbscan και τον ιεραρχικό αλγόριθμο συσσωρευτικής ομαδοποίησης με κριτήριο επιλογής τον μόνο σύνδεσμο.

Η εξέταση των μεθόδων ομαδοποίησης επιβεβαίωσε την χρησιμότητα εφαρμογής αυτών των μεθόδων. Επίσης επιβεβαίωσε την αναγκαιότητα συνεχούς ελέγχου της διαδικασίας αλλά και πιθανής ενδιάμεση ανθρώπινης παρέμβασης όπου αυτό κριθεί απαραίτητο. Το βασικό νόημα και τα πλεονεκτήματα που παρέχει αυτού του είδους η αυτοματοποίηση στον επιχειρησιακό σχεδιασμό βρίσκεται στην βελτιστοποίηση των μεθόδων που χρησιμοποιούνται για τον υπολογισμό των αποτελεσμάτων. Αυτού του είδους η ευελιξία επιτρέπει και την αντίστοιχη ευελιξία στην κατασκευή, παραλλαγή και εκτίμηση των πιθανών εκβάσεων των εξεταζόμενων σεναρίων. Έτσι για παράδειγμα, με την υποστήριξη των νέων τεχνολογιών, τα σενάρια μπορούν να αλλάξουν ακόμη και στον ενδιάμεσο της εξέλιξης τους και να αναδιατυπωθούν ή και να ακυρωθούν υπέρ ενός άλλου ποιο ευέλικτου ή ρεαλιστικού σεναρίου. Αντίθετα, όπως είδαμε και στην εφαρμογή των νευρωνικών δικτύων, κάτι τέτοιο μπορεί να σημαίνει μεγάλα χρονικά διαστήματα ή ακόμη και τη ανάγκη χρήσης περισσότερων οικονομικών και ανθρώπινων πόρων.

Για μια σωστή εφαρμογή των τεχνικών αυτών είναι αναγκαία η κριτική ικανότητα του στελέχους ή του υπαλλήλου που την εφαρμόζει και ερμηνεύει. Με αυτό τον τρόπο μπορεί να δει πέρα από τους αριθμούς και να κρίνει με βάση την κοινή λογική ή ακόμη και την διαίσθηση, κάτι που ο υπολογιστής δεν μπορεί να βοηθήσει. Η απλοποίηση των εφαρμογών βρίσκεται πλέον σε τέτοιο βαθμό ώστε να μπορούν να εκτελεστούν και μέσα από δωρεάν πακέτα όπως το R. Παρόλα αυτά το συγκεκριμένο λογισμικό απαιτεί σημαντική γνώση μαθηματικών,

μεθόδων στατιστικής ανάλυσης και προγραμματισμού υπολογιστών. Είναι πολύ εύκολο ένας οικονομικός αναλυτής να εξάγει ανακριβή αποτελέσματα με την απλοϊκή εφαρμογή και ερμηνεία αυτών των μεθόδων.

Ως παράδειγμα θα μπορούσε να αναφερθεί η εφαρμογή των περιπτέρων και των σημείων διανομής των ποδηλάτων. Μικρή αντίληψη του προβλήματος και χαμηλή ικανότητα εξαγωγής αποτελεσμάτων θα οδηγούσε στην δημιουργία ενός τρίτου περιπτέρου που θα συνεπαγόταν αυξημένα λειτουργικά έξοδα αλλά κα άλλες απώλειες οικονομικές και μη. Ωστόσο, με βάση το μέτρο αξιολόγησης silhouette η καλύτερη ομαδοποίηση προκύπτει στις περιπτώσεις k-means και k-medoids για K=2. Τα εσωτερικά μέτρα εγκυρότητας στηρίζονται σε κριτήρια που προκύπτουν από τα ίδια τα δεδομένα, όπως οι αποστάσεις των σημείων από το cluster που ανήκει ή οι αποστάσεις μεταξύ των cluster κοκ. Επιλέγουμε ένα εσωτερικό μέτρο μιας και δεν έχουμε label για τα δεδομένα μας και δε μπορούμε να εξάγουμε συμπέρασμα στο κατά πόσο οι ομάδες ανήκουν σε μεγάλο ποσοστό σε μια κατηγορία (class) του σετ δεδομένου μας. Συνεπώς, είναι ένας παράγοντας που μας καθορίζει το «σωστό» clustering ωστόσο για την τελική μας επιλογή θα πρέπει να λαμβάνονται και άλλοι παράγοντες υπόψη.

Με βάση τα προηγούμενα συμπεράσματα είναι απολύτως ασφαλής ο ισχυρισμός ότι η εφαρμογή των μεθόδων ομαδοποίησης σε επιχειρήσεις για την εκπόνηση ενός επιχειρηματικού σχεδίου είναι εύκολες στην υλοποίηση τους, ανέξοδες και εφικτές. Παρόλα αυτά, η επιφυλακτικότητα, η έλλειψη επαρκούς στελέχωσης η άγνοια αυτών των μεθόδων και των πλεονεκτημάτων τους, μπορεί να αποτελέσουν ανασταλτικούς παράγοντες για μια επιχείρηση ώστε να προχωρήσει προς αυτή την κατεύθυνση.

Από τα αποτελέσματα φάνηκε ότι k-means και k-medoids δουλεύουνε καλύτερα και παράγουν καλύτερα αποτελέσματα. Πιο συγκεκριμένα στην περίπτωση της επιλογής της θέσης των περιπτέρων φαίνεται ότι έχουν καλύτερα αποτελέσματα μιας και τα δεδομένα αυτά χρειάζονται διαχωριστικούς αλγόριθμους και όχι αλγορίθμους που βασίζονται στην πυκνότητα. Είναι γραμμικώς διαχωρίσιμα τα clusters που θέλουμε και ένας density-based αλγόριθμος όπως ο DBSCAN εύκολα μπορεί να μπερδευτεί και να μας οδηγήσει σε λανθασμένα cluster.

Ο k-medoids βγάζει παρόμοια αποτελέσματα με τον k-means, κάτι που είναι λογικό μιας και έχουνε ακριβώς την ίδια με τη διαφορά ότι ο k-medoids ως κέντρο ομάδας έχει κάποιο σημείο του σετ δεδομένων το οποίο ορίζει ως αντιπρόσωπο. Πρακτικά, θα είχαν διαφορετικά αποτελέσματα, αν τα δεδομένα μας είχαν ακραίες τιμές (πχ. σημεία θορύβου). Σε αυτή την περίπτωση θα ήταν προτιμότερο να χρησιμοποιηθεί ο k-medoid, ο οποίος λόγω των

αντιπροσώπων ως κέντρα, δεν επηρεάζεται από τις ακραίες τιμές. Τέτοια δεδομένα είναι συνήθως από βιολογικά ή ιατρικά πειράματα ή από τιμές από αισθητήρες, εικόνα, βίντεο κτλ. Εδώ το πρόβλημα μας είναι γραμμικώς διαχωρισμένο, χωρίς ακραίες τιμές και για το λόγο αυτό και οι 2 διαχωριστικοί αλγόριθμοι έχουν σχεδόν ίδια αποτελέσματα.

Ο DBSCAN θα βοηθούσε αν είχαμε σημεία outliers τα οποία θα εμπόδιζαν την σωστή ομαδοποίηση από αλγορίθμους όπως ο k-means. Στο παράδειγμα μας με το 1^{ο} πρόβλημα, μπερδεύεται και κάποια σημεία πολύ κοντά μεταξύ τους τα βάζει σε μια ομάδα, τοποθετώντας όλα τα υπόλοιπα σε ένα cluster. Τα δεδομένα μας δεν είναι για να εφαρμοστεί αλγόριθμος βασισμένος στην πυκνότητα γιατί έχει κάποιες τοπικές γειτονιές που έχουν αρκετά πυκνά σημεία και άλλα σημεία τα οποία δεν έχουνε. Συνεπώς εύκολα μπορεί να μπερδευτεί αυτός ο αλγόριθμος. Στο πρόβλημα μας μας ενδιαφέρουν οι αποστάσεις και η εύκολη πρόσβαση από όλους, συνεπώς ταιριάζει καλύτερα ένας διαχωριστικός αλγόριθμος. Το ίδιο και ο ιεραρχικός αλγόριθμος, δεν παράγει καλά αποτελέσματα κάτι που είναι λογικό μιας και στην περίπτωση για κ = 2 λίγα σημεία τα κατατάσσει σε μία ομάδα και όλα τα υπόλοιπα στο άλλο cluster. Αυτό συμβαίνει διότι τα σημεία αυτά ήταν πολύ κοντά και ενώθηκαν για να φτιαχτεί το cluster με αποτέλεσμα να μείνουν πολλά σημεία μαζί σε ένα cluster (τα υπόλοιπα). Να σημειωθεί εδώ ότι ο ιεραρχικός με single-linkage ενώνει ομάδες με την ελάχιστη απόσταση μεταξύ των ομάδων που ορίζεται από την ελάχιστη απόσταση μεταξύ των σημείων τους.

Καταλήγουμε στο συμπέρασμα ότι τα δεδομένα του πραγματικού προβλήματος με βάση την κατανομή τους, τη δομή τους και τα χαρακτηριστικά τους χρήζουν ενός αλγορίθμου διαχωριστικής ομαδοποίησης, όπως οι k-means και k-medoids.

Παράρτημα - Κώδικας R

Code1.R – 1ο Πρόβλημα

```
#############################################################################
# Εφαρμογή 3.1. Μέθοδοι ομαδοποίησης #
#############################################################################

# Ομαδοποίηση με την μέθοδο K-means Clustering (Gerdon, 2016)
# Διερεύνηση των δεδομένων
# install.packages("readxl") # Εγκατάσταση βιβλιοθήκης
library(readxl) # Φόρτωση βιβλιοθήκης
stations<-read_excel(file.choose(), col_names = TRUE)
summary(stations)
par(mfrow = c(1, 3))
hist(stations$latitude, main="Ιστόγραμμα συχνοτήτων του γεωγραφικού πλάτους", col = 'gray',xlab="Γ. Μήκος")
hist(stations$longitude, main="Ιστόγραμμα συχνοτήτων του γεωγραφικού πλάτους",xlab="Γ. Πλάτος", ylim = c(0, 60),
    col = 'gray')
plot(stations$longitude, stations$latitude, main="Διάγραμμα διασποράς Γ. μήκους και πλάτους", asp = 1,xlab="Γ.
Μήκος",ylab="Γ. Πλάτος")
par(mfrow = c(1, 1))

# Εφαρμογή της συνάρτησης kmeans()
set.seed(123) #Απαλοιφή τυχαιότητας
two <- kmeans(stations, 2)
three <- kmeans(stations, 3)
# Εξαγωγή αποτελεσμάτων
three
two$centers
two$size
clus <- cbind(stations, clus2 = two$cluster,
          clus3 = three$cluster)
head(clus)



# Ανάπτυξη υπόθεσης Business case
par(mfrow = c(1, 2))
plot(clus$longitude, clus$latitude, col = two$cluster, asp = 1,
    pch = two$cluster, main = "Θέσεις για δύο περίπτερα",
    xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")
points(two$centers[ ,2], two$centers[ ,1], pch = 23,
     col = 'maroon', bg = 'lightblue', cex = 3)
text(two$centers[ ,2], two$centers[ ,1], cex = 1.1,
    col = 'black', attributes(two$centers)$dimnames[[1]])
plot(clus$longitude, clus$latitude, col = three$cluster, asp = 1,
```

```
      pch = three$cluster, main = "Θέσεις για τρία περίπτερα",
      xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")
  points(three$centers[ ,2], three$centers[ ,1],
       pch = 23, col = 'maroon', bg = 'lightblue', cex = 3)
  text(three$centers[ ,2], three$centers[ ,1], cex = 1.1,
      col = 'black', attributes(three$centers)$dimnames[[1]])


  par(mfrow = c(1, 1))
  hybrid <- cbind(clus, hybrid_shape = rep(0, dim(clus)[1]))
  for (e in 1:dim(hybrid[1])[1]) {
   if (hybrid[e, 3] == hybrid[e, 4]) {
    hybrid[e, 5] <- hybrid[e, 3]
   }
   if (hybrid[e, 3] != hybrid[e, 4]) {
    hybrid[e, 5] <- hybrid[e ,3] + 15
   }
  }
  plot(hybrid$longitude, hybrid$latitude, col = two$cluster,
      main = "Συνένωση: Ομαδοποίηση δύο περιπτέρων σε περιοχή τριών ομάδων", pch = hybrid$hybrid_shape,
cex =
       1.1,
      xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος", asp = 1)
  points(three$centers[1:2, 2], three$centers[1:2, 1],
       pch = 23, col = 'maroon', bg = 'lightblue', cex = 3)
  text(three$centers[1:2, 2], three$centers[1:2, 1], cex = 1.1,
      col = 'black', attributes(two$centers)$dimnames[[1]])
  #rm(hybrid, stations, e)

  # ------------------------------------------------------
  # Επιπλεόν υπολογισμοί
  # ------------------------------------------------------
  # Δημιουργία πινακα δεομένων με αποστάσεις
  compare <- cbind(clus, dist2 = rep(0, dim(clus)[1]),
                   dist3 = rep(0, dim(clus)[1]))
  # -------
  # Τριγωνομετρικοί υπολογισμοί αποστάσεων
  distance2 <- function(lat, long, clus_id) {
   acos(sin(lat * pi / 180) * sin(two$centers[clus_id, 1] * pi / 180) +
        cos(lat * pi / 180) * cos(two$centers[clus_id, 1] * pi / 180) *
        cos(two$centers[clus_id, 2] * pi / 180 - long * pi / 180)) *
     6371 #in km

  }
  distance3 <- function(lat, long, clus_id) {
   acos(sin(lat * pi / 180)*sin(three$centers[clus_id, 1] * pi / 180) +
        cos(lat * pi / 180) * cos(three$centers[clus_id, 1] * pi / 180) *
        cos(three$centers[clus_id, 2] * pi / 180 - long * pi / 180)) *
     6371 #in km
  }
```

```
# -------
for (e in 1:dim(compare[1])[1]) {
  compare[e, 5] <- distance2(compare[e, 1], compare[e, 2], compare[e, 3])
  compare[e, 6] <- distance3(compare[e, 1], compare[e, 2], compare[e, 4])
}

if(!require("dplyr")) install.packages("dplyr")
suppressMessages(suppressWarnings(library(dplyr)))

compare <- cbind(compare, hybrid = rep(0, dim(compare)[1]))
for (e in 1:dim(compare[1])[1]) {
  compare[e, 7] <- distance3(compare[e, 1], compare[e, 2], compare[e, 3])
}
compare <- mutate(compare, temp_increase = (hybrid - dist3))

print(mean(compare[,5]))
print(mean(compare[,6]))
print(median(compare[,5]))
print(median(compare[,6]))
print(max(compare[,5]))
print(max(compare[,6]))

print(mean(compare[,7]))
print(median(compare[,7]))
print(max(compare[,7]))


# -------
par(mfrow = c(1, 3))
hist(compare[ ,5], ylim = c(0, 80), xlim = c(0, 8), col = "lightgray",
    xlab = "Απόσταση σε χιλιόμετρα", main = "Δύο περίπτερα")
abline(v = mean(compare[ ,5]), lty = "dashed")
hist(compare[ ,6], ylim = c(0, 80), xlim = c(0, 8), col = "lightgray",
    xlab = "Απόσταση σε χιλιόμετρα", main = "Τρία περίπτερα")
abline(v = mean(compare[ ,6]), lty = "dashed")
hist(compare[ ,7], ylim = c(0, 80), xlim = c(0, 8), col = "lightgray",
    xlab = "Απόσταση σε χιλιόμετρα", main = "Δύο περίπτερα (Συνένωση)")
abline(v = mean(compare[ ,7]), lty = "dashed")
par(mfrow = c(1, 1))
summary(compare) #to indicate max
hist(compare$temp_increase, breaks = 4, xlab = "Απόσταση σε χιλιόμετρα",
    main = "Αύξηση της απόστασης: Δημιουργία δύο περιπτέρων σε μελλοντικές θέσεις", col =
"gray",ylab="Συχνότητα")
bins <- as.data.frame(table(cut(compare$temp_increase, breaks = c(-1:5))))
text(seq(-.5, 4.5, 1), 50, cex = 1.1, col = 'black', bins[ ,2])
increase <- filter(compare, temp_increase > 0)
increase <- increase[ ,-c(1:7)]

summary(increase)
text(2, 80, paste("Μέση αύξηση:", round(mean(increase) *
```

```
                                0.62137119, 2), "μίλια")) # .62 mi/km

# GEO Mapping
library(magrittr)
library(leaflet)
leaflet() %>%
  addTiles() %>%
  addMarkers(data = stations, ~longitude, ~latitude)

######
#check number of clusters
#install.packages('factoextra')

library(factoextra)
fviz_nbclust(stations, kmeans, method = "wss")

#Use DBSCAN

#install.packages("dbscan")
library(dbscan)

eps = 0.015 #this results in 4 clusters

res <- dbscan(stations, eps)
N = unique(res$cluster)
clusters = res$cluster + 1 #in res, cluster number starts from 0
dbscan_clusters1 = clusters

plot(stations$longitude, stations$latitude, col = clusters, asp = 1,
     pch = clusters, main = "DBSCAN - Ομάδες για ε=0.15",
     xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")


eps = 0.01 #this results in 11 clusters
res <- dbscan(stations, eps)
N = unique(res$cluster)
clusters = res$cluster + 1 #in res, cluster number starts from 0
dbscan_clusters2 = clusters

plot(stations$longitude, stations$latitude, col = clusters, asp = 1,
     pch = clusters, main = "DBSCAN - Ομάδες για ε=0.1",
     xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")


### K-medoid
#install.packages("kmed")
library(kmed)

D = dist(stations, method = "euclidean")
K = 2
```

```
res2 = fastkmed(D, K)

plot(stations$longitude, stations$latitude, col = res2$cluster, asp = 1,
    pch = res2$cluster, main = "K-medoid: Θέσεις για δύο περίπτερα",
    xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")
points(stations$longitude[res2$medoid], stations$latitude[res2$medoid], pch = 23,
      col = 'maroon', bg = 'lightblue', cex = 3)
text(stations$longitude[res2$medoid], stations$latitude[res2$medoid], cex = 1.1,
    col = 'black', attributes(res2$cluster)$names[1:K])


print(cbind(res2$medoid, stations$longitude[res2$medoid], stations$latitude[res2$medoid]))

K = 3
res3 = fastkmed(D, K)

plot(stations$longitude, stations$latitude, col = res3$cluster, asp = 1,
    pch = res3$cluster, main = "K-medoid: Θέσεις για τρία περίπτερα",
    xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")
points(stations$longitude[res3$medoid], stations$latitude[res3$medoid], pch = 23,
      col = 'maroon', bg = 'lightblue', cex = 3)
text(stations$longitude[res3$medoid], stations$latitude[res3$medoid], cex = 1.1,
    col = 'black', attributes(res3$cluster)$names[1:K])

print(cbind(res3$medoid,stations$longitude[res3$medoid], stations$latitude[res3$medoid]))



# Τριγωνομετρικοί υπολογισμοί αποστάσεων
distMedoid <- function(lat, long, center, clus_id) {
  acos(sin(lat * pi / 180) * sin(center[clus_id,1] * pi / 180) +
       cos(lat * pi / 180) * cos(center[clus_id,1] * pi / 180) *
       cos(center[clus_id,2] * pi / 180 - long * pi / 180)) *
    6371 #in km
}

# -------
# Δημιουργία πινακα δεομένων με αποστάσεις
clus_m <- cbind(stations, clus2 = res2$cluster,
            clus3 = res3$cluster)

compare <- cbind(clus_m, dist2 = rep(0, dim(clus_m)[1]),
              dist3 = rep(0, dim(clus_m)[1]))
centers2 = cbind(stations$latitude[res2$medoid], stations$longitude[res2$medoid])
centers3 = cbind(stations$latitude[res3$medoid], stations$longitude[res3$medoid])
for (e in 1:dim(compare[1])[1]) {
  compare[e, 5] <- distMedoid(compare[e, 1], compare[e, 2], centers2, compare[e, 3])
  compare[e, 6] <- distMedoid(compare[e, 1], compare[e, 2], centers3, compare[e, 4])
}
print(mean(compare[,5]))
```

```
print(mean(compare[,6]))
print(median(compare[,5]))
print(median(compare[,6]))
print(max(compare[,5]))
print(max(compare[,6]))


par(mfrow = c(1, 2))
hist(compare[ ,5], ylim = c(0, 80), xlim = c(0, 8), col = "lightgray",
    xlab = "Απόσταση σε χιλιόμετρα", main = "Δύο περίπτερα")
abline(v = mean(compare[ ,5]), lty = "dashed")
hist(compare[ ,6], ylim = c(0, 80), xlim = c(0, 8), col = "lightgray",
    xlab = "Απόσταση σε χιλιόμετρα", main = "Τρία περίπτερα")
abline(v = mean(compare[ ,6]), lty = "dashed")


##hierarchical clustering
D = dist(stations, method = "euclidean")
hier_clust <- hclust(D, method = 'single')

d = hier_clust$height[2:243] - hier_clust$height[1:242]
d_max = which.max(d)
#226: 17 clusters

par(mfrow = c(1,1))
plot(hier_clust)
abline(h=hier_clust$height[d_max], col='red')

h_clusters2 <- cutree(hier_clust, k = 2)
h_clusters3 <- cutree(hier_clust, k = 3)

plot(stations$longitude, stations$latitude, col = h_clusters2, asp = 1,
    pch = h_clusters2, main = "Hierarchical Clustering \n Single Linkage (K = 2)",
    xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")

plot(stations$longitude, stations$latitude, col = h_clusters3, asp = 1,
    pch = h_clusters3, main = "Hierarchical Clustering \n Single Linkage (K = 3)",
    xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")



hier_clust <- hclust(D, method = 'complete')

d = hier_clust$height[2:243] - hier_clust$height[1:242]
which.max(d)
#242: last distance is max

par(mfrow = c(1,1))
plot(hier_clust)
abline(h=0.16, col='red')
```

```
h_clusters2_c <- cutree(hier_clust, k = 2)
h_clusters3_c <- cutree(hier_clust, k = 3)

plot(stations$longitude, stations$latitude, col = h_clusters2_c, asp = 1,
    pch = h_clusters2_c, main = "Hierarchical Clustering \n Copmlete Linkage (K = 2)",
    xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")

plot(stations$longitude, stations$latitude, col = h_clusters3_c, asp = 1,
    pch = h_clusters3_c, main = "Hierarchical Clustering \n Complete Linkage (K = 3)",
    xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")


hier_clust <- hclust(D, method = 'ward.D2')

d = hier_clust$height[2:243] - hier_clust$height[1:242]
which.max(d)
#242: last distance is max

par(mfrow = c(1,1))
plot(hier_clust)
abline(h=0.6, col='red')


h_clusters2_w <- cutree(hier_clust, k = 2)
h_clusters3_w <- cutree(hier_clust, k = 3)

plot(stations$longitude, stations$latitude, col = h_clusters2_w, asp = 1,
    pch = h_clusters2_w, main = "Hierarchical Clustering \n Ward's method (K = 2)",
    xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")

plot(stations$longitude, stations$latitude, col = h_clusters3_w, asp = 1,
    pch = h_clusters3_w, main = "Hierarchical Clustering \n Ward's method (K = 3)",
    xlab = "Γ. Μήκος", ylab = "Γ. Πλάτος")

#silhouette metric
library(cluster)
si = silhouette(two$cluster,dist=D)
print(mean(si[,3]))
si = silhouette(three$cluster,dist=D)
print(mean(si[,3]))
hybrid_cluster = hybrid$hybrid_shape
hybrid_cluster[hybrid_cluster==1] = 16
si = silhouette(hybrid_cluster,dist=D)
print(mean(si[,3]))


si = silhouette(dbscan_clusters1,dist=D)
print(mean(si[,3]))
```

```
si = silhouette(dbscan_clusters2,dist=D)
print(mean(si[,3]))

si = silhouette(res2$cluster,dist=D)
print(mean(si[,3]))
si = silhouette(res3$cluster,dist=D)
print(mean(si[,3]))

si = silhouette(h_clusters2,dist=D)
print(mean(si[,3]))
si = silhouette(h_clusters3,dist=D)
print(mean(si[,3]))

si = silhouette(h_clusters2_c,dist=D)
print(mean(si[,3]))
si = silhouette(h_clusters3_c,dist=D)
print(mean(si[,3]))

si = silhouette(h_clusters2_w,dist=D)
print(mean(si[,3]))
si = silhouette(h_clusters3_w,dist=D)
print(mean(si[,3]))
```

Code2.R – 2ο Πρόβλημα

```
##############################################################################
# Εφαρμογή 3.2. Μέθοδοι ιεραρχικής ταξινόμησης #
##############################################################################

# Διερεύνση των δεδομένων
market<-read_excel(file.choose(), col_names = TRUE)
attach(market)
as.factor(bin)
str(market)
summary(market)
par(mfrow = c(1, 2))
boxplot(market$age ~ market$bin, main = "Ηλικία")
boxplot(market$inc ~ market$bin, main = "Εισόδημα")
par(mfrow = c(1, 1))

cor.test(market$age, market$inc)
library(dplyr)
f<-filter(market, age < 40)
cor.test(f$age, f$inc)
f<-filter(market, age >= 40)
cor.test(f$age, f$inc)

set.seed(789)
three <- kmeans(market[, 2:3], 3)
plot(market$age, market$inc, col = three$cluster, xlab = 'Ηλικία',
    ylab = 'Εισόδημα', main = 'K-means χωρίς κλίμακα')
points(three$centers[, 1], three$centers[, 2],
pch = 23, col = 'maroon', bg = 'lightblue', cex = 3)
text(three$centers[, 1], three$centers[, 2], cex = 1.1,
    col = 'black', attributes(three$centers)$dimnames[[1]])
#rm(three)

market$age_scale <- as.numeric(scale(market$age))
market$inc_scale <- as.numeric(scale(market$inc))
set.seed(789)
three <- kmeans(market[, 4:5], 3)
plot(market$age_scale, market$inc_scale, col=three$cluster,
    xlab = 'Ηλικία', ylab = 'Εισόδημα',
    main = 'K-means με κλίμακα')
points(three$centers[, 1], three$centers[, 2],
      pch = 23, col = 'maroon', bg = 'lightblue', cex = 3)
text(three$centers[, 1], three$centers[, 2], cex = 1.1,
    col = 'black', attributes(three$centers)$dimnames[[1]])
```

```
#rm(three)

# Εκτέλεση συνάρτησης hclust()
set.seed(456)
hc_mod <- hclust(dist(market[, 4:5]), method = "ward.D2")
# Απεικόνιση αποτελεσμάτων
dend <- as.dendrogram(hc_mod)
if(!require("dendextend")) install.packages("dendextend")
suppressMessages(suppressWarnings(library(dendextend)))
dend_six_color <- color_branches(dend, k = 6)
plot(dend_six_color, leaflab = "none", horiz = TRUE,
    main = "Δενδρόγραμμα Ηλικίας και Εισοδήματος", xlab = "Απόσταση")
abline(v = 37.5, lty = 'dashed', col = 'blue')
str(cut(dend, h = 37.5)$upper)
rm(dend_six_color)
set.seed(456)
two <- kmeans(market[, 4:5], 2)
three <- kmeans(market[, 4:5], 3)
four <- kmeans(market[, 4:5], 4)
five <- kmeans(market[, 4:5], 5)
six <- kmeans(market[, 4:5], 6)
seven <- kmeans(market[, 4:5], 7)
eight <- kmeans(market[, 4:5], 8)
nine <- kmeans(market[, 4:5], 9)
ten <- kmeans(market[, 4:5], 10)
# Υπολογισμός μοντέλων
optimize <- data.frame(clusters = c(2:10), wss = rep(0, 9))


optimize[1, 2] <- as.numeric(two$tot.withinss)
optimize[2, 2] <- as.numeric(three$tot.withinss)
optimize[3, 2] <- as.numeric(four$tot.withinss)
optimize[4, 2] <- as.numeric(five$tot.withinss)
optimize[5, 2] <- as.numeric(six$tot.withinss)
optimize[6, 2] <- as.numeric(seven$tot.withinss)
optimize[7, 2] <- as.numeric(eight$tot.withinss)
optimize[8, 2] <- as.numeric(nine$tot.withinss)
optimize[9, 2] <- as.numeric(ten$tot.withinss)
plot(optimize$wss ~ optimize$clusters, type = "b",
    ylim = c(0, 12000), ylab = 'Άροισμα τετραγωνικού σφάλματος εντός των ομάδων',
    main = 'Εύρεση βέλτιστου αριθμού ομάδων με βάση το σφάλμα',
    xlab = 'Αριθμός ομάδων', pch = 17, col = 'black')
rm(optimize)
three$size; four$size; five$size; six$size; seven$size
rm(two, three, four, seven, eight, nine, ten)
```

```
market$clus5 <- five$cluster
dend_five <- cutree(dend, k = 5)
market$dend5 <- dend_five
market$clus6 <- six$cluster
dend_six <- cutree(dend, k = 6)
market$dend6 <- dend_six


#install.packages("dbscan")
library(dbscan)

eps = 0.002 #this results in 4 clusters
#eps = 0.0025 #this results in 8 clusters

res <- dbscan(market[, 4:5], eps)
N = unique(res$cluster)
print(length(N))

dbscan_clusters1 = res$cluster + 1 #in res, cluster number starts from 0
tabulate(dbscan_clusters1)

par(mfrow = c(1, 2))
plot(market$age, market$inc, col = dbscan_clusters1,
    pch = dbscan_clusters1, main = "DBSCAN (ε=0.002)",
    xlab = "Ηλικία", ylab = "Εισόδημα")

eps = 0.0025 #this results in 8 clusters

res <- dbscan(market[, 4:5], eps)
N = unique(res$cluster)
print(length(N))

dbscan_clusters2 = res$cluster + 1 #in res, cluster number starts from 0
tabulate(dbscan_clusters2)


plot(market$age, market$inc, col = dbscan_clusters2,
    pch = dbscan_clusters2, main = "DBSCAN (ε=0.0025)",
    xlab = "Ηλικία", ylab = "Εισόδημα")


#kmedoids
D = dist(market[,4:5], method = "euclidean")
K = 5
res2 = fastkmed(D, K)
```

```
centers5= cbind(res2$medoid, market$age[res2$medoid], market$inc[res2$medoid])
print(centers5)

K = 6
res3 = fastkmed(D, K)

centers6 = cbind(res3$medoid, market$age[res3$medoid], market$inc[res3$medoid])
print(centers6)




# Επιλογή μοντέλου
par(mfrow = c(1, 1))
par(mfrow = c(3, 2), mar = c(3, 4, 4, 2) + 0.1)
plot(market$age, market$inc, col = five$cluster,ylab = 'Εισόδημα',
    pch = five$cluster, xlab = '', main = 'K-means (K=5)')
plot(market$age, market$inc, col = six$cluster, xlab = '',
    ylab = '', pch = six$cluster, main = 'K-means (K=6)')

plot(market$age, market$inc, col = market$dend5, xlab = '', ylab = 'Εισόδημα',
    pch = market$dend5, main = 'Ιεραρχική ομαδοποίηση k = 5 ')
plot(market$age, market$inc, col = market$dend6, xlab = '', ylab = '',
    pch = market$dend6, main = 'Ιεραρχική ομαδοποίηση k = 6')

plot(market$age, market$inc,  col = res2$cluster,
    pch = res2$cluster, main = "K-medoid (K=5)",
    xlab = "Ηλικία", ylab = "Εισόδημα")

plot(market$age, market$inc,  col = res3$cluster,
    pch = res3$cluster, main = "K-medoid (K=6)",
    xlab = "Ηλικία", ylab = "Εισόδημα")



#silhouette metric
library(cluster)
si = silhouette(five$cluster,dist=D)
print(mean(si[,3]))
si = silhouette(six$cluster,dist=D)
print(mean(si[,3]))


si = silhouette(dbscan_clusters1,dist=D)
```

```
print(mean(si[,3]))
si = silhouette(dbscan_clusters2,dist=D)
print(mean(si[,3]))

si = silhouette(market$dend5,dist=D)
print(mean(si[,3]))
si = silhouette(market$dend6,dist=D)
print(mean(si[,3]))

si = silhouette(res2$cluster,dist=D)
print(mean(si[,3]))
si = silhouette(res3$cluster,dist=D)
print(mean(si[,3]))


par(mfrow = c(1, 1), mar = c(5, 4, 4, 2) + 0.1)




# Αποτελέσματα
if(!require("dplyr")) install.packages("dplyr")
suppressMessages(suppressWarnings(library(dplyr)))
labels <- as.data.frame(market %>%
                        group_by(dend6) %>%
                        summarise(avg_age = median(age), avg_inc = median(inc)))
plot(market$age, market$inc, col = market$dend6,
    pch = market$dend6 - 1, xlab = "Ηλικία", ylab = "Εισόδημα",
    main = 'Ιεραρχική ομαδοποίηση για την επισήμανση ομαδών Marketing\n (Με επισήμανση των διαμέσων
της ηλικίας
και του εισοδήματος ανά ομάδα)')
points(labels[ ,2], labels[ ,3], pch = 21, col = 'maroon',
     bg = 'white', cex = 3)
text(labels[, 2], labels[, 3], cex = 1.1, col = 'black',
    labels[, 1])
market %>% group_by(dend6) %>% summarise(ClusterSize = n())
market %>% group_by(dend6) %>%
  summarise(min_age = min(age), med_age = median(age),
            max_age = max(age), med_inc = median(inc),
            min_inc = min(inc), max_inc = max(inc))
```

Βιβλιογραφία

Anderson, S., (2020). Business plan. Ανακτήθηκε από: https://www.investopedia.com/terms/b/business-plan.asp

Berkeley. (2011). Cluster analysis. Ανακτήθηκε από: https://www.stat.berkeley.edu/~s133/cluster2a.html

Berry, T., (2006) Hurdle The Book on Business Planning: A Step-by-step Guide to Creating a Thorough, Concrete and Concise Busines Plan. USA: Palo Alto.

Botto, L. and Mastroiacovo, P. (2018). Triple surveillance: a proposal for an integrated strategy to support and accelerate birth defect prevention: Triple surveillance for birth defect prevention. *Annals of the New York Academy of Sciences*. 1414. 126-136.

Brunings, J. (2020). What is the Difference Between a Business Plan and a Strategic Plan?. Ανακτήθηκε από: https://onstrategyhq.com/resources/what-is-the-difference-between-a-business-plan-and-a-strategic-plan/

Clemens, P.L. (1990). Event Tree Analysis, 1990. *Dostupné z:* Ανακτήθηκε από: http://www.fault-tree.net/papers/clemens-event-tree.pdf

Conway, Maree. (2004). Scenario Planning: An Innovative Approach to Strategy Development. Ανακτήθηκε από: https://www.researchgate.net/publication/242707360_Scenario_Planning_An_Innovative_Approach_to_Strategy_Development

Courtney, H. (2003) Decision Driven Scenarios for Assessing Four Levels of Uncertainty, *Strategy and Leadership* 31, (1), 14-22.

Covello, J.A. and Hazelgren, B.J., (2006). *The Complete Book of Business Plans: Simple Steps to Writing Powerful Business Plans*. Sourcebooks Inc: USA

Ester, M., Kriegel, H. P., & Xu, X. (1995). *A database interface for clustering in large spatial databases*. Inst. für Informatik.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).

Etaati, l. (2017). *Advance analytics with power bi and r.* New zealand: radacad.

Foster, M.J., (1993). Scenario Planning for Small Business. *Long Range Planning*, 26 (1), Pergamon Press Ltd.

Fotr, J. and Souček, I. (2011) *Investiční rozhodování a řízení projektů.* Praha: Grada Publishing.

Fotr, J., Spacek, M., Soucek, I. and Vacík, E. (2014). Scenarios and their application in strategic planning. *E+M Ekonomie a Management*. 17. 118-135.

Fotr, J., Spacek, M., Soucek, I. and Vacík, E., Hájek, S. (2011). *Tvorba strategie a strategické plánování. Teorie a praxe.* Praha: Grada Publishing.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

García, J. A., Fdez-Valdivia, J., Cortijo, F. J., & Molina, R. (1994). A dynamic approach for clustering data.

Gedron, J. (2016). *Introduction to r for business intelligence*. Uk: packt publishing.

Gray, P. (2010). Competitive Intelligence. *Business Intelligence Journal*, 15 (4), 31-37

Gugan, A. (2008). *Successful Scenario Planning*. JISC infoNet.

Han, J. (2011). *Data mining: concepts and techniques*. Waltham, ma: Morgan Kaufmann Publishers.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100-108.

Hejden, Van der, K. (1999). The Art of Maverick Thinking. *Scenario and Strategy Planning*, 1, (1), 19-23.

Horan, J. (2004). *The One Page Business Plan: Start with a Vision, Build a Company*!. UK : Capstone.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Lantz, B. (2013). *Machine learning with r*. Birmingham: Packt publishing

Ng, R. T., & Han, J. (1994, September). E cient and E ective clustering methods for spatial data mining. In *Proceedings of VLDB* (pp. 144-155).

O'Brien, F., Meadows, M. and Murtland, M. (2007). Creating and using scenarios – Exploring alternative possible futures and their impact on strategic decisions. In: O'Brien FA and Dyson RG (eds.). *Supporting Strategy: Frameworks, Methods and Models. Chichester: John Wiley and Sons Ltd*., pp. 211-247.

Pearson, I. and Lyons, M. (1999). Re-evaluation In: An Age of Uncertainty. Scenario and Κύρκος, Ε. (2015). *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*. [ηλεκτρ. Βιβλ.] Αθήνα: σύνδεσμος ελληνικών ακαδημαϊκών βιβλιοθηκών. Ανακτήθηκε από: http://hdl.handle.net/11419/1226

Piech, C., (χ.χ.). K Means. Ανακτήθηκε από: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

R- Bloggers (2020). Ανακτήθηκε από https://www.r-bloggers.com/

Schoemaker, P.J.H. and Gunther, R. E. (2002). *Profiting from Uncertainty. Strategies for Succeeding No Matter What the Future Brings*. New York: The Free Press

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Tessun, F. and Hermann, A. (1999). *Harnessing Potential Future. Scenario and Strategy Planning*, 1 (1), 8-12.

Vo, E., (2020). An Introduction to Strategic Planning. Ανακτήθηκε από: https://sba.thehartford.com/business-management/what-is-strategic-planning/