ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΘΟΔΟΙ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΑΝΑΛΥΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΣΕ ΤΡΑΠΕΖΙΚΑ ΔΕΔΟΜΕΝΑ

Ονοματεπώνυμο: Ορλώφ Μαρία
Επιβλέπων καθηγητής: Κουκουβίνος Χρήστος

ΑΘΗΝΑ 2014

Περίληψη
Η ομαδοποίηση όμοιων πελατών και αντικειμένων είναι μία θεμελιώδης δραστηριότητα του μάρκετινγκ που διευκολύνει την επικοινωνία. Λόγω του ότι οι εταιρείες δεν μπορούν να συνδεθούν με κάθε πελάτη τους ξεχωριστά, πρέπει να χωρίσουν τις αγορές σε ομάδες καταναλωτών και πελατών με όμοιες ανάγκες και επιθυμίες.
Στην παρούσα εργασία μελετήθηκαν δύο μέθοδοι πολυμεταβλητής στατιστικής ανάλυσης που στοχεύουν στην ομαδοποίηση των παρατηρήσεων έτσι ώστε να είναι ευκολότερη η περαιτέρω μελέτη τους: η ανάλυση κατά συστάδες και η διακριτική ανάλυση.
Στο πρώτο κεφάλαιο περιγράφεται η ανάλυση κατά συστάδες, η δημιουργία δηλαδή ομάδων έτσι ώστε να ελαχιστοποιούνται οι διαφορές των αντικειμένων που ανήκουν στην ίδια ομάδα και να μεγιστοποιούνται οι διαφορές μεταξύ αντικειμένων που ανήκουν σε διαφορετικές ομάδες. Περιγράφονται οι ιεραρχικές μέθοδοι, η μέθοδος k-means και η ομαδοποίηση σε δύο βήματα καθώς και τα μέτρα ομοιότητας ή διαφοράς που μπορούμε να χρησιμοποιήσουμε για τη σύγκριση αντικειμένων.
Στο δεύτερο κεφάλαιο περιγράφεται η διακριτική ανάλυση. Η διακριτική ανάλυση χρησιμοποιείται για να καθορίσει ποιες μεταβλητές διαχωρίζουν τα δεδομένα σε δύο ή περισσότερες αμοιβαία αποκλειόμενες ομάδες και σε ποια συγκεκριμένη ομάδα ανήκει ένα αντικείμενο βάσει των χαρακτηριστικών του. Με τη βοήθεια της διακριτικής ανάλυσης μπορούμε να ταξινομήσουμε μία νέα παρατήρηση, να προβλέψουμε δηλαδή σε ποια ομάδα θα ανήκει με βάση κάποια χαρακτηριστικά της χρησιμοποιώντας την συνάρτηση ταξινόμησης που προκύπτει από την ανάλυση.
Τέλος, στο τρίτο κεφάλαιο γίνονται δύο εφαρμογές σε τραπεζικά δεδομένα. Η πρώτη είναι μία εφαρμογή της ανάλυσης κατά συστάδες και συγκεκριμένα της ανάλυσης με τη μέθοδο διαμέρισης k-means. Η δεύτερη είναι μία εφαρμογή της διακριτικής ανάλυσης (ομαδοποίηση και ταξινόμηση).

Abstract
Grouping similar customers or products is a fundamental marketing activity that makes communication easier. Companies cannot connect to each customer individually, so they have to divide markets into groups of consumers and customers with similar needs and desires.
In the present study, two methods of multivariate statistical analysis, aiming at grouping observations so that it is easier to study them further, were examined: cluster analysis and discriminant analysis.
In the first chapter the cluster analysis is described, namely the creation of groups so that the differences between objects belonging to the same group are minimized and the differences between objects belonging to different groups are maximized. There is a description of hierarchical methods, the k-means method and two-step clustering and measures of similarity or dissimilarity that can be used to compare objects.
In the second chapter the discriminant analysis is described. The discriminant analysis is used to determine which variables separate the data into two or more mutually exclusive groups and to which specific group an object belongs, based on its characteristics. Using discriminant analysis we can classify a new observation, in other words predict which group it belongs to using the sorting functions resulting from the analysis.
Finally, in the third chapter are two applications on a banking dataset. The first one is an application of cluster analysis, specifically the k-means partitioning method. The second is an application of discriminant analysis (clustering and classification).

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χ. Κουκουβίνο για την συνεργασία, αφού μου έδωσε την ευκαιρία να μελετήσω ένα τόσο ενδιαφέρον αντικείμενο. Ακόμη, την υποψήφια διδάκτορα κ. Χ. Παρπούλα για την καθοδήγηση και τις παρατηρήσεις της. Επίσης οφείλω ένα μεγάλο ευχαριστώ στην κ. Ειρήνη Λυγκώνη για την πολύτιμη βοήθεια και τις συμβουλές της.

Τέλος, θα ήθελα να ευχαριστήσω στην οικογένεια μου για την υποστήριξη καθ' όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας.

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφάλαιο 1: Ανάλυση κατά συστάδες

1.1 Εισαγωγή

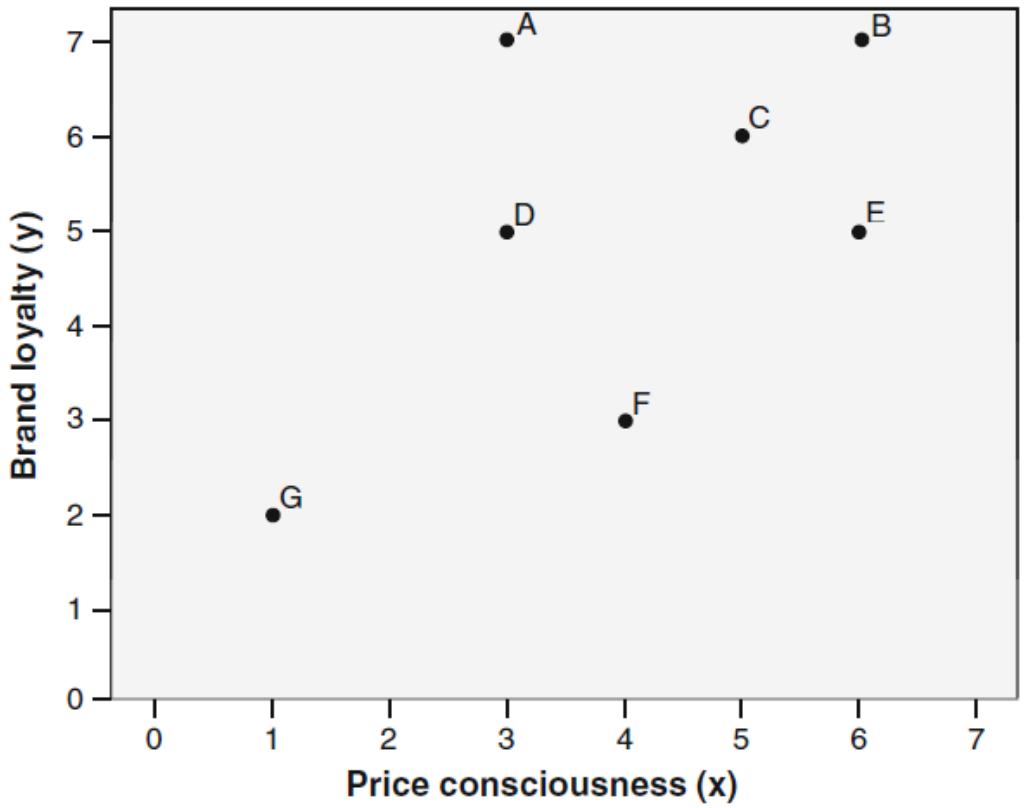
Η ομαδοποίηση όμοιων πελατών και προϊόντων είναι μία θεμελιώδης δραστηριότητα του μάρκετινγκ. Χρησιμοποιείται, εμφανώς, στην κατάτμηση της αγοράς. Λόγω του ότι οι εταιρείες δεν μπορούν να συνδεθούν με όλους τους πελάτες τους, πρέπει να χωρίσουν τις αγορές σε ομάδες καταναλωτών και πελατών με όμοιες ανάγκες και επιθυμίες. Οι εταιρείες μπορούν τότε να στοχεύσουν σε κάθε μία από αυτές τις ομάδες τοποθετώντας τον εαυτό τους μέσα σε μία μοναδική ομάδα (όπως για παράδειγμα η Ferrari στην αγορά των σπορ οχημάτων τελευταίας τεχνολογίας). Ενώ οι ερευνητές της αγοράς συχνά σχηματίζουν τις ομάδες βασιζόμενοι σε πρακτικά πεδία, την πρακτική της βιομηχανίας και την σοφία, η ανάλυση κατά συστάδες επιτρέπει στις ομάδες να σχηματιστούν με βάση δεδομένα που είναι λιγότερο εξαρτώμενα από την υποκειμενικότητα.
Η ομαδοποίηση των πελατών είναι μία βασική εφαρμογή της ανάλυσης κατά συστάδες, αλλά μπορεί να χρησιμοποιηθεί σε διαφορετικές εφαρμογές όπως η εκτίμηση τυπικών μονοπατιών σε supermarket (Larson, 2005) ή στρατηγικές branding που προκύπτουν από τους εργοδότες (Moroco and Uncles, 2009).

1.2 Κατανόηση της ανάλυσης κατά συστάδες

Η ανάλυση κατά συστάδες είναι μία βολική μέθοδος για να εντοπίζουμε ομογενείς ομάδες αντικειμένων που ονομάζονται συστάδες. Τα αντικείμενα (ή οι περιπτώσεις ή παρατηρήσεις) σε μία συγκεκριμένη συστάδα μοιράζονται πολλά χαρακτηριστικά, αλλά διαφέρουν πολύ από αντικείμενα που δεν ανήκουν στη συστάδα αυτή.
Ας προσπαθήσουμε να κατανοήσουμε τη διαδικασία της ανάλυσης κατά συστάδες κοιτώντας ένα απλό παράδειγμα. Ας υποθέσουμε ότι μας ενδιαφέρει η ομαδοποίηση της βάσης των πελατών για να καταφέρουμε να τους στοχεύσουμε μέσω των στρατηγικών τιμολόγησης.
Το πρώτο βήμα είναι να αποφασίσουμε ποια χαρακτηριστικά θα χρησιμοποιήσουμε ώστε να δημιουργήσουμε τις ομάδες πελατών. Με άλλα λόγια, πρέπει να αποφασίσουμε ποιες μεταβλητές θα συμπεριληφθούν στην ανάλυση κατά συστάδες. Για παράδειγμα, μπορούμε να ομαδοποιήσουμε τους πελάτες μιας αγοράς χρησιμοποιώντας τις μεταβλητές συνείδηση τιμών (price conciousness (x)) και πίστη στη μάρκα (brand loyalty (y)). Αυτές οι δύο μεταβλητές μπορούν να μετρηθούν σε μία κλίμακα από 1 ως 7 με τις μεγαλύτερες τιμές να δείχνουν μεγαλύτερη συνείδηση τιμής ή εμπιστοσύνη στη φίρμα. Οι τιμές από τις απαντήσεις εφτά ερωτηθέντων φαίνονται στον πίνακα 1 και το διάγραμμα διασποράς 1.

Πελάτης	A	B	C	D	E	F	G
x	3	6	5	3	6	4	1
y	7	7	6	5	5	3	2

Πίνακας 1



Εικόνα 1: Διάγραμμα Διασποράς

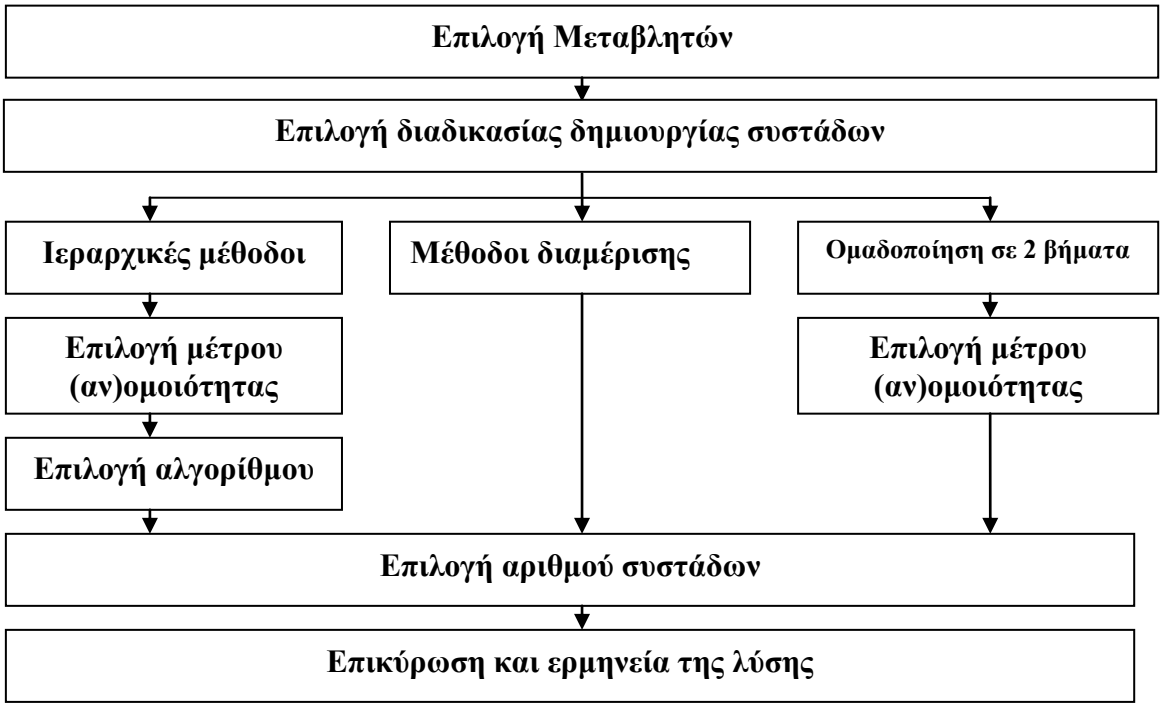
Ο σκοπός της ανάλυσης κατά συστάδες είναι να αναγνωρίσουμε ομάδες αντικειμένων (σε αυτή την περίπτωση πελατών) που είναι αρκετά όμοια όσον αφορά στις παραπάνω μεταβλητές και να τα κατατάξουμε σε συστάδες. Αφού αποφασίσουμε για τις μεταβλητές της ανάλυσης, πρέπει να αποφασίσουμε για τη διαδικασία ομαδοποίησης για να δημιουργήσουμε τις συστάδες. Το βήμα αυτό είναι κρίσιμο για την ανάλυση, αφού διαφορετικές διαδικασίες απαιτούν διαφορετικές αποφάσεις πριν την ανάλυση. Υπάρχει μία πληθώρα διαφορετικών προσεγγίσεων και πολύ μικρή καθοδήγηση στο ποια προσέγγιση να χρησιμοποιήσουμε στην πράξη. Θα μελετήσουμε τις πιο γνωστές προσεγγίσεις στην έρευνα αγοράς, αφού μπορούν να υπολογιστούν εύκολα με χρήση του SPSS. Οι προσεγγίσεις αυτές είναι: ιεραρχικές μέθοδοι, διαχωριστικές μέθοδοι (συγκεκριμένα k-means) και ομαδοποίηση σε 2 βήματα (two-step clustering), που χοντρικά είναι ένας συνδυασμός των δύο πρώτων μεθόδων. Καθεμιά από αυτές τις διαδικασίες ακολουθεί μία διαφορετική προσέγγιση ομαδοποιώντας τα πιο όμοια αντικείμενα σε μία συστάδα και καθορίζοντας τη θέση κάθε αντικειμένου στην συστάδα. Με άλλα λόγια, ενώ ένα αντικείμενο σε μία συγκεκριμένη συστάδα θα πρέπει να είναι όσο πιο όμοιο γίνεται με τα άλλα αντικείμενα της ίδιας συστάδας, ομοίως θα πρέπει να είναι όσο πιο διακριτό γίνεται από τα αντικείμενα των άλλων συστάδων.

Αλλά πώς μετράμε την ομοιότητα; Κάποιες προσεγγίσεις – και ιδιαίτερα οι ιεραρχικές μέθοδοι - απαιτούν να συγκεκριμενοποιήσουμε πόσο όμοια ή διαφορετικά είναι τα αντικείμενα έτσι ώστε να προσδιορίσουμε τις διαφορετικές συστάδες. Τα περισσότερα πακέτα λογισμικού υπολογίζουν ένα μέτρο ομοιότητας ή ανομοιότητας εκτιμώντας την απόσταση μεταξύ ζευγών αντικειμένων. Τα αντικείμενα με μικρές αποστάσεις μεταξύ τους είναι πιο όμοια, ενώ τα αντικείμενα με μεγαλύτερες αποστάσεις είναι πιο ανόμοια.

Ένα σημαντικό πρόβλημα στην εφαρμογή της ανάλυσης κατά συστάδες είναι η απόφαση του αριθμού των συστάδων που θα πρέπει να προκύψουν από τα δεδομένα. Το ερώτημα αυτό μελετάται στο επόμενο βήμα της ανάλυσης. Κάποιες φορές όμως γνωρίζουμε ήδη

τον αριθμό των ομάδων που πρέπει να προκύψουν από τα δεδομένα. Για παράδειγμα, αν μας ζητούσαν να εξακριβώσουμε ποια χαρακτηριστικά ξεχωρίζουν τους ανθρώπους που ψωνίζουν συχνά από αυτούς που δεν ψωνίζουν συχνά, χρειαζόμαστε δύο διαφορετικές συστάδες. Όμως συνήθως δεν γνωρίζουμε τον ακριβή αριθμό των συστάδων και τότε αντιμετωπίζουμε ένα δίλημμα. Από την μία πλευρά, θέλουμε όσο το δυνατόν λιγότερες συστάδες για να είναι ευκολότερες στην κατανόηση και επεξεργασία. Από την άλλη, η ύπαρξη πολλών συστάδων μας επιτρέπει να εντοπίσουμε περισσότερες ομάδες και πιο ανεπαίσθητες διαφορές μεταξύ των ομάδων. Σε μία ακραία περίπτωση, μπορούμε να αντιμετωπίσουμε κάθε άτομο ξεχωριστά (one-to-one marketing) για να ικανοποιήσουμε τις διάφορες ανάγκες των καταναλωτών με τον καλύτερο δυνατό τρόπο. Παραδείγματα μίας τέτοιας στρατηγικής είναι το Mongolian Shoe BBQ της Puma (www.mongolianshoebbq.puma.com) και το Nike ID της Nike, με το οποίο οι πελάτες μπορούν να προσαρμόσουν πλήρως ένα ζευγάρι παπούτσια στο γούστο τους μέσω μιας απτικής και διαδραστικής εμπειρίας δημιουργίας παπουτσιών. Από την άλλη, το κόστος που συνοδεύει μία τέτοια στρατηγική ίσως είναι απαγορευτικά υψηλό σε πολλά επιχειρηματικά πλαίσια. Έτσι, πρέπει να διασφαλίσουμε ότι οι ομάδες είναι αρκετά μεγάλες ώστε τα προγράμματα στοχευμένου μάρκετινγκ να είναι κερδοφόρα. Συνεπώς, πρέπει να αντιμετωπίσουμε έναν συγκεκριμένο βαθμό ετερογένειας μέσα στη συστάδα, πράγμα που κάνει τα προγράμματα αυτά λιγότερο αποτελεσματικά.

Στο τελευταίο βήμα, πρέπει να ερμηνεύσουμε τη λύση καθορίζοντας και ονομάζοντας τις αποκτηθείσες συστάδες. Αυτό μπορεί να επιτευχθεί εξετάζοντας τις μέσες τιμές των μεταβλητών της ανάλυσης ή εντοπίζοντας επεξηγηματικές μεταβλητές για να καθορίσουμε τις συστάδες. Τελικά, τα διευθυντικά στελέχη πρέπει να μπορούν να αναγνωρίσουν πελάτες σε κάθε ομάδα με βάση εύκολα υπολογίσιμες μεταβλητές. Αυτό το τελευταίο βήμα απαιτεί επίσης να αξιολογήσουμε την σταθερότητα και εγκυρότητα της λύσης της ανάλυσης. Η εικόνα 2 δείχνει τα βήματα της ανάλυσης κατά συστάδες, που αναλύονται παρακάτω.



Εικόνα 2: Τα βήματα της ανάλυσης κατά συστάδες

1.3 Διεξαγωγή Ανάλυσης Κατά Συστάδες

1.3.1 Επιλογή Μεταβλητών της Ανάλυσης

Στην αρχή της διαδικασίας της δημιουργίας των συστάδων, πρέπει να επιλέξουμε τις κατάλληλες μεταβλητές για την ανάλυση κατά συστάδες. Παρόλο που αυτή η επιλογή είναι υψίστης σημασίας, σπάνια αντιμετωπίζεται έτσι και αντί για αυτό, ένα μείγμα από ένστικτο και διαθεσιμότητα δεδομένων καθοδηγούν τις περισσότερες αναλύσεις στην πρακτική του μάρκετινγκ. Όμως, εσφαλμένες υποθέσεις μπορεί να οδηγήσουν σε ακατάλληλες ομάδες αγοράς και συνεπώς σε ατελείς στρατηγικές μάρκετινγκ. Έτσι λοιπόν, πρέπει να είμαστε πολύ προσεκτικοί όταν επιλέγουμε τις μεταβλητές.

Υπάρχουν πολλά είδη μεταβλητών για την ανάλυση κατά συστάδες και μπορούν να ταξινομηθούν στις εξής κατηγορίες: από τη μία στις γενικές (ανεξάρτητες των προϊόντων, υπηρεσιών ή συνθηκών) και τις ειδικές (σχετιζόμενες και με τον πελάτη και με το προϊόν, την υπηρεσία και/ή την συγκεκριμένη συνθήκη), και από την άλλη στις παρατηρήσιμες (άμεσα μετρήσιμες) και τις μη παρατηρίσιμες (που έχουν συναχθεί). Ο πίνακας 2 δείχνει διάφορους τύπους και παραδείγματα μεταβλητών.

	Γενικές	Ειδικές
Παρατηρήσιμες (άμεσα μετρήσιμες	Πολιτιστικές, γεωγραφικές, δημογραφικές	Κατάσταση χρήστη, συχνότητα χρήσης, πίστη στη μάρκα ή στο κατάστημα
Μη παρατηρήσιμες	Ψυχογραφικές, αξίες, προσωπικότητα, τρόπος ζωής	Πλεονεκτήματα, αντιλήψεις, στάσεις, προθέσεις, προτιμήσεις

Πίνακας 2: Είδη και παραδείγματα μεταβλητών ομαδοποίησης

Οι τύποι των μεταβλητών που χρησιμοποιούνται στην ανάλυση κατά συστάδες δίνουν διαφορετικές ομάδες και έτσι, επηρεάζουν τις στρατηγικές που στοχεύουν σε ομάδες. Τις τελευταίες δεκαετίες, η προσοχή έχει στραφεί από τις πιο παραδοσιακές γενικές μεταβλητές προς τις μη παρατηρίσιμες μεταβλητές. Οι τελευταίες γενικά προσφέρουν καλύτερη καθοδήγηση για αποφάσεις πάνω στην αποτελεσματική προδιαγραφή των οργάνων του μάρκετινγκ. Είναι κοινώς αποδεκτό ότι οι ομάδες που καθορίζονται μέσω συγκεκριμένων παρατηρίσιμων μεταβλητών είναι συνήθως περισσότερο ομογενείς και οι καταναλωτές τους ανταποκρίνονται με συνέπεια σε ενέργειες μάρκετινγκ. (Wedel and Kamakura 2000). Όμως, οι καταναλωτές σε αυτές τις ομάδες είναι συχνά δύσκολο να καθοριστούν μέσω εύκολα υπολογίσιμων μεταβλητών, όπως τα δημογραφικά στοιχεία. Αντίστροφα, οι ομάδες που έχουν καθοριστεί μέσω γενικά παρατηρίσιμων μεταβλητών συνήθως ξεχωρίζουν εξαιτίας της αναγνωρισιμότητάς τους αλλά συχνά τους λείπει μία μοναδική δομή απόκρισης. Συνεπώς, οι ερευνητές συχνά συνδυάζουν διαφορετικές μεταβλητές (για παράδειγμα πολλαπλά χαρακτηριστικά τρόπου ζωής με δημογραφικά χαρακτηριστικά), εκμεταλλευόμενοι τα δυνατά σημεία της κάθε μεταβλητής.

Σε μερικές περιπτώσεις, η επιλογή των μεταβλητών είναι προφανής από τη φύση του προβλήματος. Για παράδειγμα, ένα διοικητικό πρόβλημα σχετικά με τις εταιρικές επικοινωνίες θα έχει ένα αρκετά καλά ορισμένο σύνολο μεταβλητών, που θα περιλαμβάνει υποψήφιους όπως ευαισθητοποίηση, στάσεις, αντιλήψεις και συνήθειες των μέσων. Όμως, αυτό δεν συμβαίνει πάντα και οι ερευνητές πρέπει να επιλέξουν μέσα από ένα σύνολο υποψήφιων μεταβλητών.

Όποιες μεταβλητές και να επιλεγούν, είναι σημαντικό να επιλέξουμε αυτές που δίνουν μία σαφή διαφοροποίηση μεταξύ των ομάδων όσον αφορά σε ένα συγκεκριμένο διαχειριστικό

στόχο. Πιο συγκεκριμένα, η εγκυρότητα των κριτηρίων είναι ειδικού ενδιαφέροντος, δηλαδή στο βαθμό στον οποίο οι ανεξάρτητες μεταβλητές σχετίζονται με μία ή περισσότερες εξαρτημένες μεταβλητές που δεν συμπεριλαμβάνονται στην ανάλυση. Δεδομένης αυτής της σχέσης, θα πρέπει να υπάρχουν σημαντικές διαφορές μεταξύ των εξαρτημένων μεταβλητών μέσα στις συστάδες. Αυτές οι συσχετίσεις μπορεί να είναι αιτιολογικές ή όχι, αλλά είναι απαραίτητο οι μεταβλητές να διαχωρίζουν σημαντικά τις εξαρτημένες μεταβλητές. Οι μεταβλητές κριτηρίων συνήθως έχουν να κάνουν με κάποιο είδος συμπεριφοράς, όπως η πρόθεση αγοράς ή η συχνότητα χρήσης.

Γενικά θα πρέπει να αποφεύγουμε να χρησιμοποιούμε μία πληθώρα μεταβλητών, αφού αυτό αυξάνει τις πιθανότητες οι μεταβλητές να μην είναι πλέον ανόμοιες. Αν υπάρχει μεγάλος βαθμός συγγραμμικότητας μεταξύ των μεταβλητών, δεν είναι αρκετά κατάλληλες στην αναγνώριση διακριτών ομάδων της αγοράς. Αν χρησιμοποιούνται μεταβλητές με υψηλή συσχέτιση στην ανάλυση κατά συστάδες, συγκεκριμένες πτυχές που αναλύονται από αυτές τις μεταβλητές θα εμφανίζονται περισσότερο στην λύση της ανάλυσης. Γι' αυτό, απόλυτες συσχετίσεις πάνω από 0.90 είναι πάντα προβληματικές. Για παράδειγμα, αν έπρεπε να προσθέσουμε άλλη μία μεταβλητή με όνομα προτίμηση μάρκας στην ανάλυσή μας, προφανώς θα κάλυπτε το ίδιο κομμάτι με την πίστη στη μάρκα. Έτσι, η ιδέα του να θέλει κάποιος να αγοράζει προϊόντα μίας μάρκας θα παρουσιαζόταν σε υπερβολικό βαθμό στην ανάλυση αφού η διαδικασία της ανάλυσης κατά συστάδες δεν διαφοροποιείται μεταξύ των μεταβλητών από εννοιολογική άποψη. Οι ερευνητές συχνά χειρίζονται αυτό το πρόβλημα εφαρμόζοντας ανάλυση κατά συστάδες to the observation factor scores derived from a previously carried out factor analysis. Όμως, σύμφωνα με τους Dolnicar και Grun (2009), συτή η προσέγγιση μπορεί να οδηγήσει σε διάφορα προβλήματα:

1. Τα δεδομένα είναι προ-επεξεργασμένα και οι συστάδες καθορίζονται με βάση μετασχηματισμένες τιμές, κι όχι τις αυθεντικές πληροφορίες, πράγμα που οδηγεί σε διαφορετικά αποτελέσματα.
2. Στην ανάλυση παραγόντων, η λύση δεν εξηγεί ένα συγκεκριμένο μέρος της μεταβλητής, έτσι αφήνουμε πληροφορίες εκτός ανάλυσης πριν καθορίσουμε ή κατασκευάσουμε τις ομάδες.
3. Η παράλειψη μεταβλητών με χαμηλό φορτίο σε όλους τους παράγοντες σημαίνει ότι πιθανόν οι πιο σημαντικές πληροφορίες για τον καθορισμό εξειδικευμένων ομάδων να έχουν αγνοηθεί, καθιστώντας αδύνατο τον καθορισμό τέτοιων ομάδων.
4. Η ερμηνείες των συστάδων που βασίζονται στις αυθεντικές μεταβλητές γίνονται αμφισβητήσιμες δεδομένου ότι οι ομάδες κατασκευάστηκαν χρησιμοποιώντας παραγοντικά αποτελέσματα.

Αρκετές μελέτες έχουν δείξει ότι η προσέγγιση αυτή μειώνει την επιτυχία ανάκτησης των τμημάτων. Συνεπώς, είναι προτιμότερο να μειώσουμε τον αριθμό των αντικειμένων πριν το ερωτηματολόγιο, κρατώντας ένα λογικό αριθμό σχετικών, μη-περιττών ερωτήσεων που πιστεύουμε ότι διαφοροποιούν σωστά τις ομάδες. Πρέπει όμως αν έχουμε αμφιβολίες για τη δομή των δεδομένων, η παραπάνω προσέγγιση μπορεί να είναι καλύτερη επιλογή από το να παραλείψουμε αντικείμενα που μπορεί εννοιολογικά να είναι απαραίτητα.

Επιπλέον, πρέπει να λάβουμε υπόψη το μέγεθος του δείγματος. Πρώτα και κύρια, αυτό συνδέεται με θέματα διαχειριστικού ενδιαφέροντος αφού τα μεγέθη των ομάδων πρέπει να είναι ουσιώδη για να διασφαλίσουν ότι τα στοχευμένα προγράμματα μάρκετινγκ θα είναι κερδοφόρα. Από μία στατιστική άποψη, κάθε επιπλέον μεταβλητή χρειάζεται μία μεγάλη αύξηση στις παρατηρήσεις για να εξασφαλιστεί η εγκυρότητα των αποτελεσμάτων. Δυστυχώς, δεν υπάρχει κοινώς αποδεκτός κανόνας για το ελάχιστο μέγεθος δείγματος ή τη σχέση μεταξύ των αντικειμένων και του αριθμού των μεταβλητών που χρησιμοποιούνται.

Σε ένα σχετικό μεθοδολογικό πλαίσιο, ο Formann (1984) προτείνει μέγεθος δείγματος 2^m , όπου m ο αριθμός των μεταβλητών. Αυτό μπορεί να δώσει μόνο πρόχειρη καθοδήγηση.

Παρόλα αυτά, πρέπει να προσέξουμε την σχέση μεταξύ των αντικειμένων και των μεταβλητών. Για παράδειγμα, δεν φαίνεται λογικό να ομαδοποιήσουμε σε συστάδες δέκα αντικείμενα χρησιμοποιώντας δέκα μεταβλητές. Η ανάλυση κατά συστάδες θα δώσει πάντα αποτέλεσμα όσο και να ναι το πλήθος των μεταβλητών ή το μέγεθος του δείγματος. Τελικά, η επιλογή των μεταβλητών εξαρτάται πάντα από εννοιολογικές επιρροές, όπως η διαθεσιμότητα δεδομένων ή οι πηγές που θα μπορέσουμε να βρούμε επιπλέον δεδομένα. Οι ερευνητές μάρκετινγκ συχνά προσπερνούν το γεγονός ότι η επιλογή των μεταβλητών είναι στενά συνδεδεμένη με την ποιότητα των δεδομένων. Μόνο εκείνες οι μεταβλητές που εξασφαλίζουν ότι μπορούν να χρησιμοποιηθούν δεδομένα υψηλής ποιότητας, θα πρέπει να συμπεριληφθούν στην ανάλυση. Αυτό είναι πολύ σημαντικό αν μία λύση ομαδοποίησης πρέπει να είναι διοικητικά χρήσιμη. Επιπλέον, τα δεδομένα είναι υψηλής ποιότητας αν οι ερωτήσεις έχουν ισχυρή θεωρητική βάση, δεν έχουν «μολυνθεί» από την κούραση των ερωτηθέντων ή το στυλ της απάντησης, είναι πρόσφατες κι έτσι αντικατοπτρίζουν την τρέχουσα κατάσταση της αγοράς (Dolnicar και Lazarevski 2009). Τέλος, οι απαιτήσεις άλλων διοικητικών λειτουργιών μέσα στον οργανισμό παίζουν πρωταρχικό ρόλο. Οι πωλήσεις και η διανομή μπορούν να έχουν μεγάλη επιρροή στο σχεδιασμό των τμημάτων αγοράς. Συνεπώς, πρέπει να έχουμε υπόψη ότι η συμφωνία υποκειμενικότητας και κοινής λογικής θα επηρεάσουν πάντα την επιλογή των μεταβλητών της ανάλυσης.

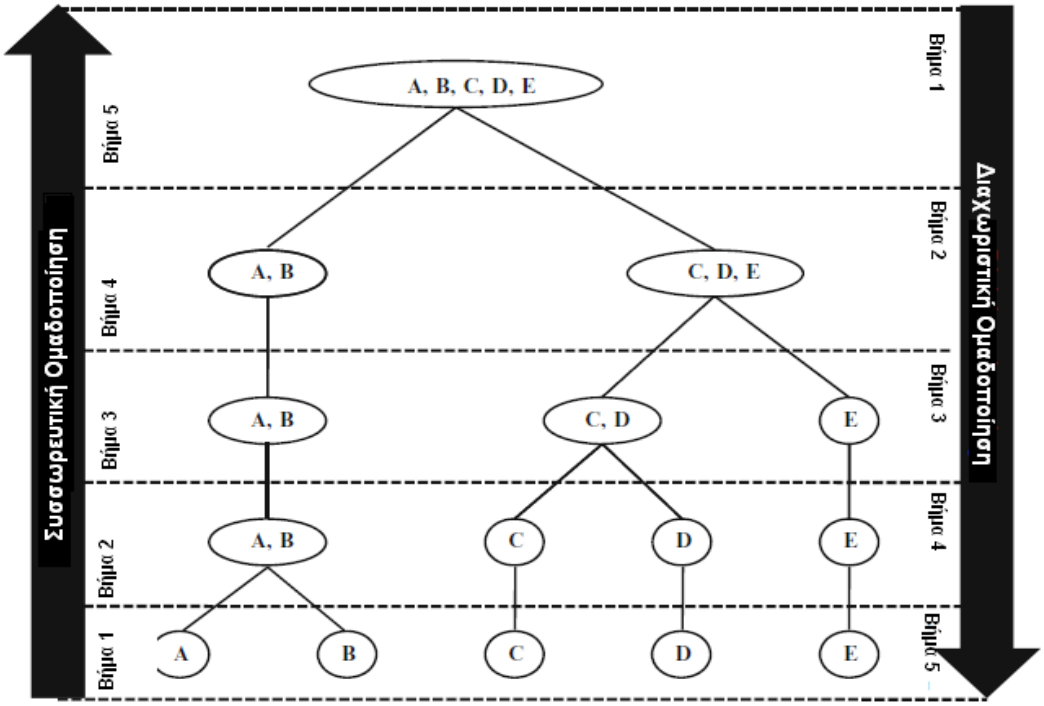
1.3.2 Απόφαση για τη Διαδικασία Δημιουργίας Συστάδων

Επιλέγοντας μία συγκεκριμένη διαδικασία, καθορίζουμε πώς θα σχηματιστούν οι συστάδες. Αυτό περιλαμβάνει πάντα τη βελτιστοποίηση κάποιου είδους κριτηρίου, όπως η ελαχιστοποίηση της διασποράς μέσα σε μία συστάδα (δηλαδή την ολική διασπορά των μεταβλητών για τα αντικείμενα μιας συγκεκριμένης συστάδας), ή η μεγιστοποίηση της απόστασης μεταξύ των αντικειμένων ή των συστάδων. Η διαδικασία μπορεί να αντιμετωπίσει το ερώτημα πώς να καθοριστεί η ομοιότητα ή η διαφορά μεταξύ των αντικειμένων σε μία καινούργια συστάδα που δημιουργείται και των αντικειμένων που παραμένουν στο σύνολο των δεδομένων.
Υπάρχουν πολλές διαδικασίες για τη δημιουργία συστάδων και επίσης πολλοί τρόποι ταξινόμησης τους (για παράδειγμα επικαλυπτόμενες - με μη-επικαλυπτόμενες, μονοτροπικές - πολυτροπικές, εξαντλητικές - μη εξαντλητικές). Ένας πρακτικός διαχωρισμός είναι η διαφοροποίηση μεταξύ ιεραρχικών και διαχωριστικών μεθόδων (ιδιαίτερα η διαδικασία k-means), που αναλύονται παρακάτω. Επίσης η δημιουργία συστάδων σε δύο βήματα (two-step clustering) συνδυάζει τις αρχές των ιεραρχικών μεθόδων και διαχωριστικών μεθόδων και έχει πρόσφατα τραβήξει αυξανόμενη προσοχή από την πρακτική της έρευνας αγοράς.

1.3.3 Ιεραρχικές μέθοδοι

Οι ιεραρχικές μέθοδοι δημιουργίας συστάδων χαρακτηρίζονται από την δομή δένδρου που δημιουργείται στην πορεία της ανάλυσης. Οι περισσότερες ιεραρχικές τεχνικές ανήκουν σε μια κατηγορία που ονομάζεται συσσωρευτική ομαδοποίηση. Σε αυτή την κατηγορία, οι συστάδες σχηματίζονται διαδοχικά από τα αντικείμενα. Αρχικά, αυτό το είδος διαδικασίας ξεκινάει με κάθε αντικείμενο να αντιπροσωπεύει μία ξεχωριστή συστάδα. Αυτές οι συστάδες μετά συγχωνεύονται με τη σειρά σύμφωνα με την ομοιότητά τους. Πρώτα οι δύο πιο όμοιες (αυτές με τη μικρότερη απόσταση μεταξύ τους) συγχωνεύονται για να σχηματίσουν μία νέα συστάδα στην βάση της ιεραρχίας. Στο επόμενο βήμα, ένα άλλο ζευγάρι συστάδων συγχωνεύεται και συνδέεται με ένα μεγαλύτερο επίπεδο ιεραρχίας

κ.ο.κ. Αυτό επιτρέπει μία ιεραρχία συστάδων να δημιουργηθεί από τη βάση προς τα πάνω. Στην εικόνα 3, φαίνεται πως λειτουργεί αυτή η διαδικασία.



Εικόνα 3: Συσσωρευτική και διαχωριστική ομαδοποίηση

Μία ιεραρχία συστάδων μπορεί να δημιουργηθεί και από την κορυφή προς τα κάτω. Σε αυτή την διαχωριστική δημιουργία συστάδων (divisive clustering), όλα τα αντικείμενα είναι αρχικά συγχωνευμένα σε μία συστάδα, που σταδιακά χωρίζεται. Η εικόνα 3 δείχνει αυτή τη διαδικασία. Και στις δύο διαδικασίες, μία συστάδα που ανήκει σε μεγαλύτερο επίπεδο πάντα περικλείει όλες τις συστάδες των κατώτερων επιπέδων. Αυτό σημαίνει ότι αν ένα αντικείμενο τοποθετηθεί σε μία συγκεκριμένη συστάδα, δεν υπάρχει πιθανότητα να επανατοποθετηθεί σε άλλη. Αυτή είναι μία σημαντική διαφορά μεταξύ αυτού του είδους δημιουργίας συστάδων και των διαχωριστικών μεθόδων.
Οι διαχωριστικές διαδικασίες χρησιμοποιούνται σπάνια στην έρευνα αγοράς. Επικεντρωνόμαστε έτσι στις διαδικασίες συσσωρευτικής ομαδοποίησης. Υπάρχουν πολλά είδη τέτοιων διαδικασιών. Πρέπει όμως πρώτα να ορίσουμε πώς θα μετρηθούν οι ομοιότητες και οι διαφορές μεταξύ των αντικειμένων.

1.3.4 Επιλογή Μέτρου Ομοιότητας ή Διαφοράς

Υπάρχουν πολλά μέτρα για να εκφράσουν πόσο δύο αντικείμενα είναι όμοια ή διαφέρουν. Ένας άμεσος τρόπος να εκτιμήσουμε την εγγύτητα δύο αντικειμένων είναι να σχεδιάσουμε το ευθύγραμμο τμήμα που τα ενώνει. Για παράδειγμα, στην εικόνα 1, βλέπουμε ότι το μήκος του ευθύγραμμου τμήματος B-C είναι μικρότερο από το B-G. Αυτό το είδος απόστασης είναι γνωστό σαν Ευκλείδεια απόσταση και είναι το πιο συνηθισμένο είδος που χρησιμοποιείται όταν πρόκειται για ανάλυση δεδομένων αναλογικών ή σε κλίμακα διαστημάτων. Για να χρησιμοποιήσουμε μία ιεραρχική διαδικασία δημιουργίας συστάδων, πρέπει να εκφράσουμε μαθηματικά τις αποστάσεις αυτές. Λαμβάνοντας υπόψη τα δεδομένα του πίνακα 1, μπορούμε εύκολα να υπολογίσουμε την Ευκλείδεια απόσταση

μεταξύ των πελατών B και C (d(B,C)), με βάση τις δύο μεταβλητές x και y χρησιμοποιώντας τον ακόλουθο τύπο:

$$d_{Euclidean}(B,C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

Η Ευκλείδεια απόσταση είναι η τετραγωνική ρίζα του αθροίσματος των τετραγώνων των διαφορών των τιμών κάθε μεταβλητής. Χρησιμοποιώντας τα δεδομένα του πίνακα 1, λαμβάνουμε:

$$d_{Euclidean}(B,C) = \sqrt{(6-5)^2 + (7-6)^2} = \sqrt{2} = 1.414$$

Η απόσταση ανταποκρίνεται στο μήκος του ευθύγραμμου τμήματος που ενώνει τα αντικείμενα B και C. Σε αυτή την περίπτωση, χρησιμοποιήσαμε δύο μεταβλητές αλλά μπορούμε εύκολα να προσθέσουμε περισσότερες στον παραπάνω τύπο. Όμως, κάθε επιπλέον μεταβλητή θα προσθέσει μία διάσταση στο πρόβλημα (για παράδειγμα αν έχουμε έξι μεταβλητές θα έχουμε έξι διαστάσεις), καθιστώντας αδύνατο να αναπαρασταθεί η λύση γραφικά. Ομοίως, μπορούμε να υπολογίσουμε την απόσταση μεταξύ των B και G:

$$d_{Euclidean}(B,G) = \sqrt{(6-1)^2 + (7-2)^2} = \sqrt{50} = 7.071$$

Με τον ίδιο τρόπο μπορούμε να υπολογίσουμε την απόσταση για κάθε ζευγάρι αντικειμένων. Όλες αυτές οι αποστάσεις εκφράζονται συνήθων μέσω ενός πίνακα αποστάσεων. Σε αυτό τον πίνακα, τα μη διαγώνια στοιχεία εκφράζουν τις αποστάσεις μεταξύ των ζευγών των αντικειμένων, ενώ τα διαγώνια στοιχεία είναι μηδέν (η απόσταση ενός αντικειμένου από τον εαυτό του είναι 0). Στο παράδειγμα, ο πίνακας έχει διαστάσεις 8x8 με τις γραμμές και τις στήλες να αντιπροσωπεύουν τα αντικείμενα (πελάτες). Ο πίνακας είναι συμμετρικός και αφού τα διαγώνια στοιχεία είναι μηδέν, αρκεί να κοιτάξουμε είτε τα πάνω είτε τα κάτω μη διαγώνια στοιχεία.

Αντικείμενα	**A**	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	3	0					
C	2.236	1.414	0				
D	2	3.606	2.236	0			
E	3.606	2	1.414	3	0		
F	4.123	4.472	3.162	2.236	2.828	0	
G	5.385	7.071	5.657	3.606	5.831	3.162	0

Πίνακας 3: Πίνακας Ευκλείδειων Αποστάσεων

Υπάρχουν επίσης και εναλλακτικά μέτρα απόστασης. Η απόσταση city-block χρησιμοποιεί το άθροισμα των απόλυτων διαφορών των μεταβλητών. Συχνά ονομάζεται Manhattan metric αφού μοιάζει με το περπάτημα μεταξύ δύο σημείων σε μια πόλη όπως το Manhattan της Νέας Υόρκης, όπου η απόσταση ισούται με τον αριθμό των τετραγώνων με κατεύθυνση Βόρεια-Νότια και Ανατολικά-Δυτικά. Χρησιμοποιώντας αυτή την απόσταση προκύπτει το εξής:

$$d_{City-block}(B,C) = |x_B - x_C| + |y_B - y_C| = |6-5| + |7-6| = 2$$

Ενώ ο πίνακας αποστάσεων θα είναι ο πίνακας 4.

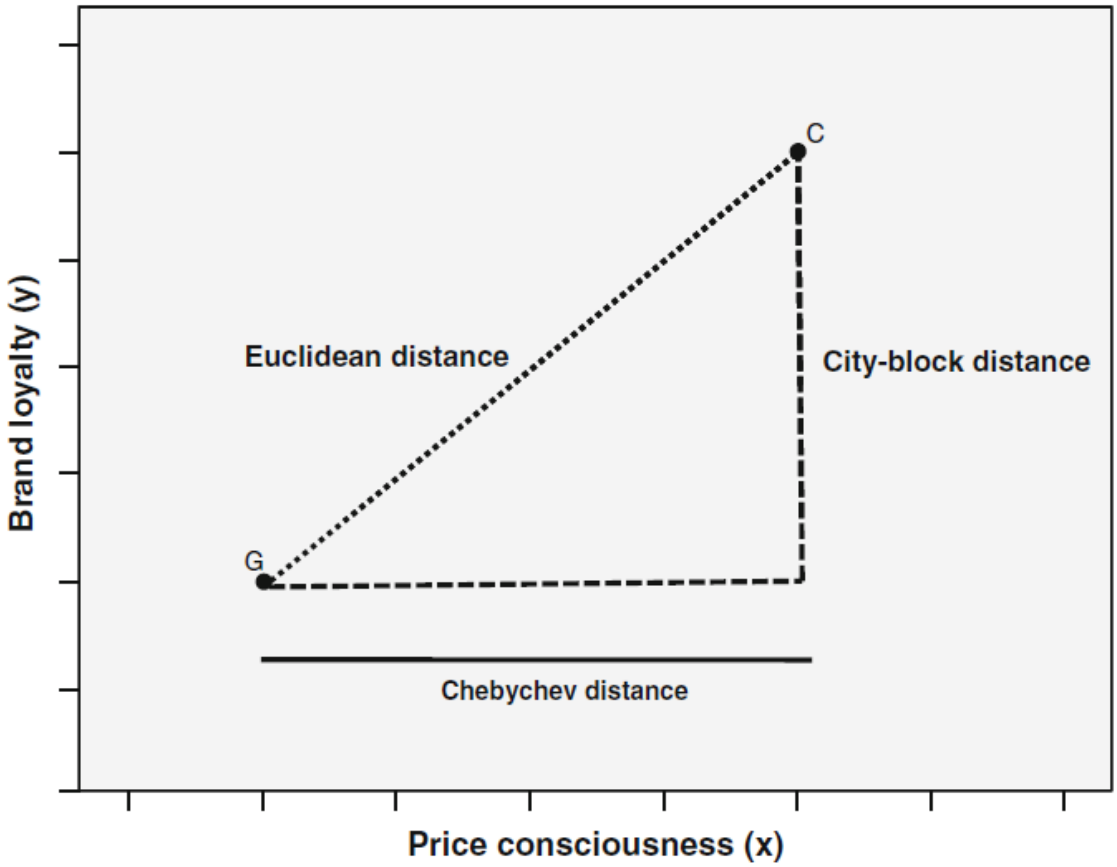
Αντικείμενα	A	B	C	D	E	F	G
A	0						
B	3	0					
C	3	2	0				
D	2	5	3	0			
E	5	2	2	3	0		
F	5	6	4	3	4	0	
G	7	10	8	5	8	4	0

Πίνακας 4: Πίνακας Αποστάσεων city-block

Τέλος, όταν εργαζόμαστε με μετρικά (ή αριθμητικά) δεδομένα, πολλές φορές χρησιμοποιείται η απόσταση Chebychev, που είναι το μέγιστο της απόλυτης διαφοράς των τιμών των μεταβλητών. Για τους πελάτες B και C, η απόσταση αυτή είναι:

$$d_{Chebychev}(B,C) = \max(|\ x_B - x_C\ |, |\ y_B - y_C\ |) = \max(|\ 6-5\ |, |\ 7-6\ |) = 1$$

Η εικόνα 4 δείχνει την σχέση μεταξύ των τριών αυτών αποστάσεων για τους πελάτες C και G του παραδείγματος.



Εικόνα 4: Μέτρα Απόστασης

Υπάρχουν κι άλλα μέτρα απόστασης όπως οι αποστάσεις Angular, Canberra ή Mahalanobis. Σε πολλές περιπτώσεις το τελευταίο είναι επιθυμητό αφού αντισταθμίζει την συγγραμμικότητα μεταξύ των μεταβλητών, αλλά δεν υπάρχει στο SPSS.

Σε πολλά προβλήματα ανάλυσης, οι μεταβλητές μετρούνται σε διαφορετική κλίμακα. Αυτό θα συνέβαινε αν προσθέταμε για παράδειγμα άλλη μία μεταβλητή που θα αντιπροσωπεύει το εισόδημα των πελατών και θα το μετρούσαμε σε κλίμακα 1 ως 15. Επειδή η διασπορά της μεταβλητής αυτής θα ήταν πολύ μεγαλύτερη από των x και y (αφού μετρήθηκαν σε κλίμακα 1-7) τα αποτελέσματα της ανάλυσης προφανώς θα παραμορφώνονταν. Μπορούμε να λύσουμε αυτό το πρόβλημα προσαρμόζοντας τα δεδομένα πριν την ανάλυση.

Υπάρχουν πολλές μέθοδοι προσαρμογής των δεδομένων, όπως η απλή z προσαρμογή που τοποθετεί τις μεταβλητές σε νέα κλίμακα ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1. Στις περισσότερες περιπτώσεις όμως η προσαρμογή κατά εύρος (πχ σε εύρος 0 σε 1 ή -1 σε 1) λειτουργεί καλύτερα.

Ένας άλλος τρόπος προσαρμογής των δεδομένων είναι χρησιμοποιώντας τη συσχέτιση μεταξύ των αντικειμένων αντί για τα μέτρα απόστασης. Για παράδειγμα, ας υποθέσουμε έναν πελάτη με τιμή της μεταβλητής x ίση με 2 και της y ίση με 3, έναν δεύτερο πελάτη με τιμή της x ίση με 5 και της y ίση με 6 και έναν τρίτο πελάτη με τιμή της x ίση με 3 και της y ίση με 3. Οι 3 αποστάσεις (Ευκλείδεια, city-block και Chebychev) θα έδειχναν ότι ο πρώτος πελάτης είναι πιο όμοιος με τον τρίτο απ' ότι με τον δεύτερο. Παρόλα αυτά, κάποιος θα μπορούσε να πει ότι είναι πιο όμοιος με τον δεύτερο αφού και οι δύο έχουν δώσει υψηλότερη τιμή στη μεταβλητή y απ' ότι στην x. Αυτό μπορεί να ελεγχθεί υπολογίζοντας τη συσχέτιση μεταξύ δύο διανυσμάτων τιμών σαν μέτρο ομοιότητας (υψηλοί συντελεστές συσχέτισης δείχνουν μεγάλο βαθμό ομοιότητας). Συνεπώς, η ομοιότητα δεν καθορίζεται πλέον με βάση τη διαφορά στις απαντήσεις αλλά με βάση την ομοιότητα στα προφίλ απαντήσεων. Η χρήση της συσχέτισης είναι επίσης ένας τρόπος έμμεσης προσαρμογής των δεδομένων.

Το αν θα χρησιμοποιήσουμε συσχέτιση ή ένα από τα μέτρα απόστασης εξαρτάται από το αν πιστεύουμε ότι το σχετικό μέγεθος των μεταβλητών ενός αντικειμένου (που ευνοεί τη συσχέτιση) έχει περισσότερη σημασία από το σχετικό μέγεθος κάθε μεταβλητής σε όλα τα αντικείμενα (που ευνοεί την απόσταση). Όμως είναι γενικά προτιμότερο να χρησιμοποιούμε τη συσχέτιση όταν εφαρμόζουμε διαδικασίες δημιουργίας συστάδων που είναι ευαίσθητες σε ακραίες παρατηρήσεις.

Ενώ τα μέτρα απόστασης που έχουν παρουσιαστεί ως τώρα μπορούν να χρησιμοποιηθούν για δεδομένα κανονικής κλίμακας, η εφαρμογή τους σε ονομαστικά ή δυαδικά δεδομένα, δεν έχει νόημα. Σ' αυτό το είδος ανάλυσης, είναι προτιμότερο να επιλέξουμε ένα μέτρο ομοιότητας που θα εκφράζει το βαθμό στον οποίο οι τιμές των μεταβλητών ανήκουν στην ίδια κατηγορία. Αυτοί οι ονομαζόμενοι συντελεστές αντιστοιχίας μπορούν να πάρουν διάφορες μορφές αλλά βασίζονται στο ίδιο σχέδιο που φαίνεται στον πίνακα 5.

		Αντικείμενο 1	
		Αριθμός μεταβλητών κατηγορίας 1	Αριθμός μεταβλητών κατηγορίας 2
Αντικείμενο 2	Αριθμός μεταβλητών κατηγορίας 1	a	b
	Αριθμός μεταβλητών κατηγορίας 2	c	d

Πίνακας 5: Σχέδιο κατανομής για τους συντελεστές αντιστοιχίας

Βασιζόμενοι στο σχέδιο του πίνακα 5, μπορούμε να υπολογίσουμε διάφορους τέτοιους συντελεστές, όπως τον απλό συντελεστή αντιστοιχίας:

$$SM = \frac{a+d}{a+b+c+d}$$

Αυτός ο συντελεστής είναι χρήσιμος όταν και οι θετικές και οι αρνητικές τιμές περιέχουν έναν ίσο βαθμό πληροφορίας. Για παράδειγμα, το φύλο είναι μια συμμετρική μεταβλητή επειδή ο αριθμός των ανδρών και των γυναικών δίνουν έναν ίσο βαθμό πληροφορίας.
Ας δούμε ένα παράδειγμα υποθέτοντας ότι έχουμε ένα σύνολο δεδομένων με τρεις δυαδικές μεταβλητές: φύλο (άνδρας = 1, γυναίκα =2), πελάτης (πελάτης = 1, όχι πελάτης = 2), και εισόδημα (χαμηλό = 1 , υψηλό = 2). Το πρώτο αντικείμενο είναι ένας άντρας, όχι πελάτης με υψηλό εισόδημα ενώ το δεύτερο αντικείμενο είναι μια γυναίκα, όχι πελάτης με υψηλό εισόδημα. Σύμφωνα με το σχέδιο του πίνακα 5, a = b = 0, c=1 και d=2, με τον απλό συντελεστή αντιστοιχίας να παίρνει την τιμή 0.667 .
Δύο άλλοι τύποι τέτοιων συντελεστών, που δεν εξισώνουν την κοινή απουσία ενός χαρακτηριστικού με ομοιότητα και μπορεί ως εκ τούτου να έχει μεγαλύτερη αξία στις μελέτες κατάτμησης, είναι οι συντελεστές Jaccard (JC) και Russel και RAO (RR). Ορίζονται ως εξής:

$$JC = \frac{a}{a+b+c}$$

$$RR = \frac{a}{a+b+c+d}$$

Αυτοί οι συντελεστές χρησιμοποιούνται - όπως τα μέτρα απόστασης - για να προσδιορίσουν μια λύση. Υπάρχουν πολλοί ακόμα συντελεστές αντιστοιχίας, όπως Yule, Kulczynski ή Ochiai, αλλά αφού οι περισσότερες εφαρμογές της ανάλυσης κατά συστάδες βασίζονται σε μετρικά ή αριθμητικά δεδομένα, δεν χρειάζεται περαιτέρω ανάλυση.
Για ονομαστικές μεταβλητές με περισσότερες από δυο κατηγορίες πρέπει πάντα να μετατρέπουμε την κατηγορηματική μεταβλητή σ' ένα σύνολο δυαδικών μεταβλητών, ώστε να μπορούμε να χρησιμοποιήσουμε τους παραπάνω συντελεστές. Όταν έχουμε αριθμητικά δεδομένα πρέπει πάντα να χρησιμοποιούμε μέτρα απόστασης όπως η Ευκλείδεια απόσταση. Παρόλο που η χρήση συντελεστών θα ήταν εφικτή και - από μια αυστηρά στατιστική άποψη - καταλληλότερη, θα παραβλέπαμε πληροφορίες της μεταβλητής στην ακολουθία των κατηγοριών. Στο τέλος, ένας αποκρινόμενος που δείχνει ότι είναι πολύ πιστός σε μια μάρκα, θα είναι πιο κοντά με κάποιον που είναι λιγότερο πιστός απ' ότι με κάποιον που δεν είναι καθόλου πιστός. Επιπλέον τα μέτρα απόστασης αντιπροσωπεύουν καλύτερα την ιδέα της εγγύτητας, που είναι θεμελιώδης στην ανάλυση κατά συστάδες.
Τα περισσότερα σύνολα δεδομένων περιλαμβάνουν μεταβλητές που μετρούνται σε πολλές κλίμακες. Για παράδειγμα, ένα ερωτηματολόγιο έρευνας αγοράς μπορεί να ρωτάει για το εισόδημα του πελάτη, αξιολογήσεις προϊόντων και για την μάρκα που αγόρασε τελευταία φορά. Έτσι, έχουμε μεταβλητές που μετρούνται σε αναλογική, τακτική και ονομαστική κλίμακα. Πως μπορούμε να ενσωματώσουμε ταυτόχρονα αυτές τις μεταβλητές σε μια ανάλυση; Δυστυχώς, αυτό το πρόβλημα δεν μπορεί να λυθεί εύκολα και στην πραγματικότητα πολλοί ερευνητές της αγοράς απλά αγνοούν το επίπεδο της κλίμακας. Αντί γι' αυτό χρησιμοποιούν ένα από τα μέτρα απόστασης. Παρόλο που αυτή η προσέγγιση μπορεί να αλλάξει ελαφρά τα αποτελέσματα συγκρινόμενα με εκείνα όπου έχουμε χρησιμοποιήσει συντελεστές αντιστοιχίας, δεν θα έπρεπε να απορριφθεί. Η ανάλυση κατά συστάδες είναι κυρίως μια διερευνητική τεχνική, της οποίας τα αποτελέσματα παρέχουν μια πρόχειρη καθοδήγηση για διοικητικές αποφάσεις. Παρά το

γεγονός αυτό, υπάρχουν διάφορες διαδικασίες που επιτρέπουν την ταυτόχρονη ενσωμάτωση αυτών των μεταβλητών σε μια ανάλυση.
Πρώτον, μπορούμε να υπολογίσουμε διακριτούς πίνακες αποστάσεων για κάθε ομάδα μεταβλητών, δηλαδή ένα πίνακα αποστάσεων, βασιζόμενο για παράδειγμα, σε μεταβλητές αριθμητικής κλίμακας και έναν άλλον βασιζόμενο σε ονομαστικές μεταβλητές. Στη συνέχεια μπορούμε απλά να υπολογίσουμε το σταθμισμένο αριθμητικό μέσο των αποστάσεων και να χρησιμοποιήσουμε αυτόν τον πίνακα μέσων αποστάσεων ως εισαγόμενα δεδομένα για την ανάλυση κατά συστάδες. Όμως, τα βάρη πρέπει να έχουν καθοριστεί εκ των προτέρων ενώ τα ακατάλληλα μπορούν να οδηγήσουν σε μια μεροληπτική αντιμετώπιση των διαφορετικών τύπων μεταβλητών. Επιπλέον ο υπολογισμός και η διαχείριση των πινάκων απόστασης δεν είναι ασήμαντα. Χρησιμοποιώντας το συντακτικό του SPSS πρέπει να προσθέσουμε την υπο-εντολή MATRIX χειροκίνητα, που εξάγει τον αρχικό πίνακα αποστάσεων σ' ένα νέο αρχείο δεδομένων.
Δεύτερον, μπορούμε να διχοτομήσουμε όλες τις μεταβλητές και να εφαρμόσουμε τους συντελεστές που αναφέρθηκαν παραπάνω. Στην περίπτωση των μετρικών μεταβλητών, αυτό θα περιλαμβάνει τον προσδιορισμό των κατηγοριών (παράδειγμα: χαμηλό, μέτριο και υψηλό εισόδημα) και την μετατροπή τους σε σύνολο δυαδικών μεταβλητών. Στις περισσότερες περιπτώσεις όμως ο προσδιορισμός των κατηγοριών είναι μάλλον αυθαίρετος και, όπως αναφέρθηκε νωρίτερα, αυτή η διαδικασία θα μπορούσε να οδηγήσει σε μια σοβαρή απώλεια πληροφοριών.

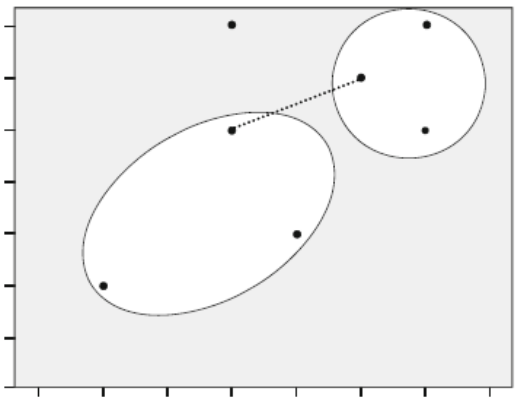
Λαμβάνοντας αυτά υπόψη, θα 'πρεπε να αποφύγουμε τον συνδυασμό αριθμητικών και ονομαστικών μεταβλητών σε μια ανάλυση κατά συστάδες, αλλά αν αυτό δεν είναι εφικτό, η διαδικασία δημιουργίας συστάδων σε δυο βήματα (two-step clustering procedure) αποτελεί μια πολύτιμη εναλλακτική. Τέλος, η επιλογή του μέτρου ομοιότητας (ή ανομοιότητας) δεν είναι εξαιρετικά κρίσιμη για την ανάκτηση της βασικής δομής της συστάδας. Από αυτήν την άποψη, η επιλογή του αλγορίθμου δημιουργίας συστάδων είναι πολύ πιο σημαντική.

1.3.5 Επιλογή Αλγορίθμου Δημιουργίας Συστάδων

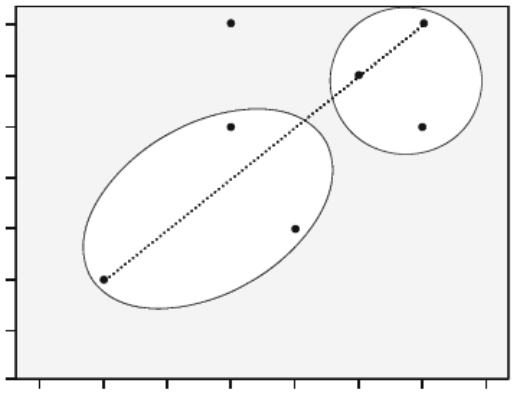
Αφού επιλέξουμε το μέτρο απόστασης ή ομοιότητας, πρέπει να αποφασίσουμε ποιον αλγόριθμο δημιουργίας συστάδων θα εφαρμόσουμε. Υπάρχουν πολλές συσωρευτικές διαδικασίες και μπορούν να διακριθούν από τον τρόπο που ορίζουν την απόσταση από μια νεοσυσταθείσα συστάδα σ' ένα συγκεκριμένο αντικείμενο ή σε άλλες συστάδες στη λύση. Οι πιο δημοφιλείς συσωρευτικές διαδικασίες δημιουργίας συστάδων περιλαμβάνουν τα ακόλουθα:

- *Ενιαία σύνδεση* (κοντινότερος γείτονας): Η απόσταση μεταξύ δυο συστάδων ανταποκρίνεται στην μικρότερη απόσταση μεταξύ δυο οποιωνδήποτε αντικειμένων στις δυο συστάδες.
- *Πλήρης σύνδεση* (μακρύτερος γείτονας): Η αντίθετη προσέγγιση στην ενιαία σύνδεση υποθέτει ότι η απόσταση μεταξύ δυο συστάδων βασίζεται στην μεγαλύτερη απόσταση μεταξύ δυο οποιωνδήποτε αντικειμένων στις δυο συστάδες.
- *Μέση σύνδεση*: Η απόσταση μεταξύ δυο συστάδων ορίζεται ως η μέση απόσταση μεταξύ όλων των ζευγαριών των αντικειμένων των δυο συστάδων.
- *Κεντροειδής*: Σ' αυτήν την προσέγγιση, πρώτα υπολογίζουμε το γεωμετρικό κέντρο (centroid) κάθε συστάδας. Η απόσταση μεταξύ δυο συστάδων ισούται με την απόσταση μεταξύ των δυο γεωμετρικών κέντρων.

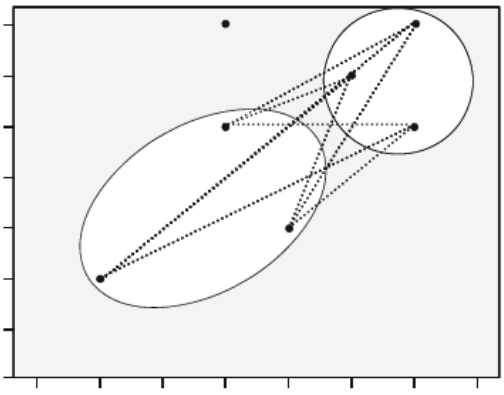
Οι εικόνες 5-8 παρουσιάζουν αυτές τις διαδικασίες σύνδεσης για δυο τυχαία σχεδιασμένες συστάδες.



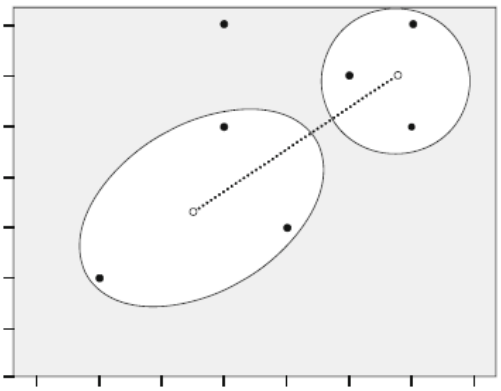
Εικόνα 5: Ενιαία Σύνδεση



Εικόνα 6: Πλήρης Σύνδεση



Εικόνα 7: Μέση Σύνδεση



Εικόνα 8: Κεντροειδής

Κάθε ένας από αυτούς τους αλγόριθμους σύνδεσης μπορεί να παράγει διαφορετικά αποτελέσματα όταν χρησιμοποιηθεί στο ίδιο σύνολο δεδομένων, αφού καθένας έχει συγκεκριμένες ιδιότητες. Αφού ο αλγόριθμος ενιαίας σύνδεσης βασίζεται σε ελάχιστες αποστάσεις, τείνει να δημιουργεί μια μεγάλη συστάδα, με τις άλλες συστάδες να περιέχουν μόνο ένα ή λίγα αντικείμενα η κάθε μία. Μπορούμε να εκμεταλλευτούμε αυτό το "εφφέ αλυσίδας" για να εντοπίσουμε ακραίες τιμές, αφού αυτές θα συγχωνευθούν με τα εναπομείναντα αντικείμενα - συνήθως σε πολύ μεγάλες αποστάσεις - στα τελευταία βήματα της ανάλυσης. Γενικά ο αλγόριθμος ενιαίας σύνδεσης θεωρείται ο ποιο ευέλικτος αλγόριθμος. Αντίστροφα, η μέθοδος της πλήρους σύνδεσης επηρεάζεται πολύ απ' τις ακραίες τιμές, αφού βασίζεται σε μέγιστες αποστάσεις. Οι συστάδες που παράγονται με αυτήν την μέθοδο είναι πιθανό να είναι αρκετά πυκνές. Οι τελευταίοι δυο αλγόριθμοι (Μέση σύνδεση, Κεντροειδής) τείνουν να παράγουν συστάδες με μικρή διασπορά εντός συστάδας και όμοια μεγέθη. Όμως, και οι δυο διαδικασίες επηρεάζονται από τις ακραίες τιμές αν και όχι τόσο όσο η πλήρης σύνδεση.

Μια άλλη αρκετά χρησιμοποιούμενη προσέγγιση στην ιεραρχική δημιουργία συστάδων είναι η μέθοδος του Ward. Αυτή η προσέγγιση δεν συνδυάζει τα δυο πιο όμοια αντικείμενα διαδοχικά. Αντίθετα, συνδυάζονται εκείνα τα αντικείμενα των οποίων η συγχώνευση αυξάνει την ολική διασπορά εντός συστάδας στο μικρότερο δυνατό βαθμό. Αν περιμένουμε συστάδες περίπου ίσου μεγέθους και το σύνολο δεδομένων δεν περιλαμβάνει ακραίες τιμές, πρέπει να χρησιμοποιούμε την μέθοδο του Ward.

Για να καταλάβουμε καλύτερα πως λειτουργεί ένας αλγόριθμος δημιουργίας συστάδων, ας εξετάσουμε κάποια απ' τα βήματα υπολογισμών της διαδικασίας ενιαίας σύνδεσης. Ξεκινάμε κοιτώντας τον αρχικό πίνακα (Ευκλείδειων) αποστάσεων (πίνακας 3). Στο πρώτο βήμα τα δυο αντικείμενα που παρουσιάζουν την μικρότερη απόσταση στον πίνακα, συγχωνεύονται. Πάντα συγχωνεύουμε εκείνα τα αντικείμενα με την μικρότερη απόσταση ανεξαρτήτως διαδικασίας (Ενιαία είτε Πλήρης σύνδεση). Όπως βλέπουμε, αυτό συμβαίνει σε δυο ζεύγη αντικειμένων, το ζεύγος B-C (d(B,C) = 1.414), και το ζεύγος C-E (d(C,E) = 1.414). Στο επόμενο βήμα θα δούμε ότι δεν έχει διαφορά ποια συγχώνευση θα κάνουμε πρώτη, οπότε συνεχίζουμε σχηματίζοντας μια νέα συστάδα χρησιμοποιώντας τα αντικείμενα B και C.

Ύστερα φτιάχνουμε ένα νέο πίνακα αποστάσεων λαμβάνοντας υπόψη τον κανόνα απόστασης της ενιαίας σύνδεσης. Σύμφωνα με αυτόν τον κανόνα, η απόσταση, για παράδειγμα, του αντικειμένου A από την νεοσυσταθείσα συστάδα είναι το ελάχιστο των αποστάσεων d(A,B) και d(A,C). Αφού η απόσταση d(A,C) είναι μικρότερη από την d(A,B), η απόσταση του A από την συστάδα ισούται με d(A,C) = 2.236. Υπολογίζουμε επίσης τις αποστάσεις από την συστάδα [B,C] μέχρι όλα τα άλλα αντικείμενα (D, E, F, G) και απλά αντιγράφουμε τις υπόλοιπες αποστάσεις - όπως την απόσταση d(E,F) - που η προηγούμενη διαδικασία δεν έχει επηρεάσει. Αυτή η διαδικασία παράγει τον πίνακα 6.

Αντικείμενα	A	B, C	D	E	F	G
A	0					
B, C	2.236	0				
D	2	2.236	0			
E	3.606	1.414	3	0		
F	4.123	3.162	2.236	2.828	0	
G	5.385	5.657	3.606	5.831	3.162	0

Πίνακας 6: Πίνακας αποστάσεων μετά το πρώτο βήμα ομαδοποίησης (ενιαία σύνδεση)

Συνεχίζοντας την διαδικασία δημιουργίας συστάδων, απλά επαναλαμβάνουμε το τελευταίο βήμα, συγχωνεύοντας τα αντικείμενα που έχουν την μικρότερη απόσταση στον νέο πίνακα

αποστάσεων (σε αυτήν την περίπτωση η συστάδα [B,C] και το αντικείμενο E) και υπολογίζουμε την απόσταση από αυτήν την συστάδα σ' όλα τα άλλα αντικείμενα. Το αποτέλεσμα αυτού του βήματος περιγράφεται στον πίνακα 7, ενώ τα τρία επόμενα στους πίνακες 8-10.

Αντικείμενα	A	B, C, E	D	F	G
A	0				
B, C, E	2.236	0			
D	2	2.236	0		
F	4.123	2.828	2.236	0	
G	5.385	5.657	3.606	3.162	0

Πίνακας 7: Πίνακας αποστάσεων μετά το δεύτερο βήμα ομαδοποίησης (ενιαία σύνδεση)

Αντικείμενα	A, D	B, C, E	F	G
A, D	0			
B, C, E	2.236	0		
F	2.236	2.828	0	
G	3.606	5.657	3.162	0

Πίνακας 8: Πίνακας αποστάσεων μετά το τρίτο βήμα ομαδοποίησης (ενιαία σύνδεση)

Αντικείμενα	A, B, C, D, E	F	G
A, B, C, D, E	0		
F	2.236	0	
G	3.606	3.162	0

Πίνακας 9: Πίνακας αποστάσεων μετά το τέταρτο βήμα ομαδοποίησης (ενιαία σύνδεση)

Αντικείμενα	A, B, C, D, E, F	G
A, B, C, D, E, F	0	
G	3.162	0

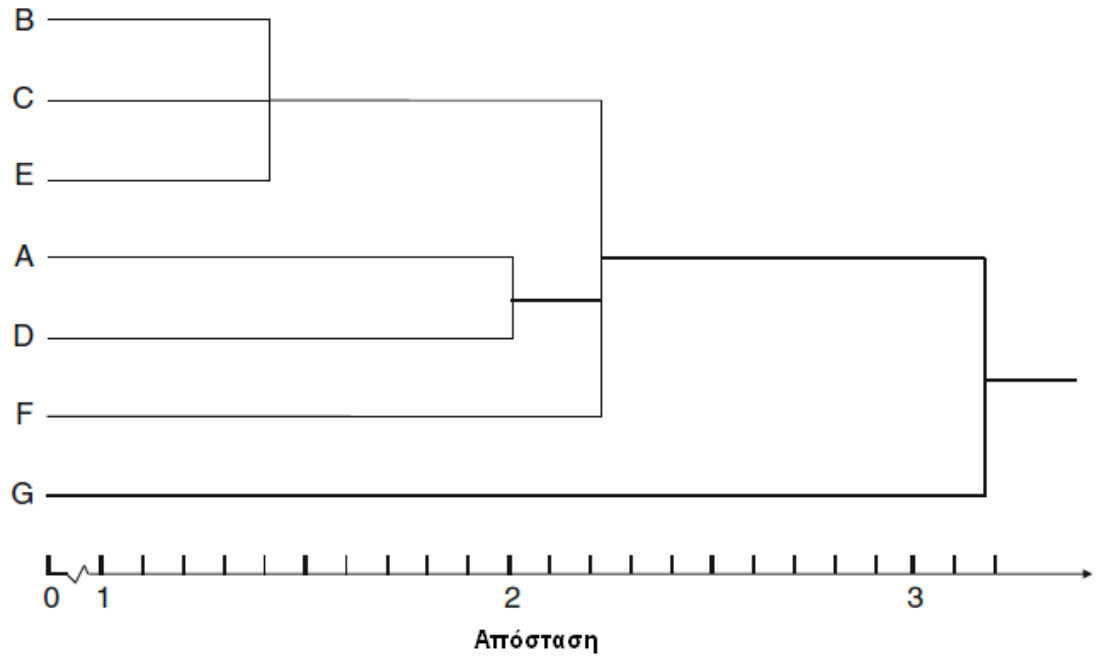
Πίνακας 10: Πίνακας αποστάσεων μετά το πέμπτο βήμα ομαδοποίησης (ενιαία σύνδεση)

Ακολουθώντας την διαδικασία της ενιαίας σύνδεσης, τα τελευταία βήματα περιλαμβάνουν την δημιουργία της συστάδας [A,B,C,D,E,F] και το αντικείμενο G σε απόσταση 3.162.
Ένας κοινός τρόπος για να οπτικοποιήσουμε την διαδικασία της ανάλυσης κατά συστάδες είναι σχεδιάζοντας ένα δενδρόγραμμα, που απεικονίζει το επίπεδο της απόστασης στο οποίο υπήρξε συνδυασμός αντικειμένων και συστάδων (εικόνα 9).
Διαβάζουμε το δενδρόγραμμα από τα αριστερά προς τα δεξιά για να δούμε σε ποια απόσταση έχουν συνδυαστεί τα αντικείμενα. Για παράδειγμα, σύμφωνα με τους παραπάνω υπολογισμούς, τα αντικείμενα B, C και E συνδυάστηκαν σε απόσταση 1.414 .

1.3.6 Επιλογή του Αριθμού των Συστάδων

Ένα σημαντικό ερώτημα που δεν έχουμε διευθετήσει είναι πως θα αποφασίσουμε για τον αριθμό των συστάδων που προκύπτουν απ' τα δεδομένα. Δυστυχώς οι ιεραρχικές μέθοδοι παρέχουν πολύ περιορισμένη καθοδήγηση για την λήψη αυτής της απόφασης. Ο μόνος

σημαντικός δείκτης σχετίζεται με τις αποστάσεις στις οποίες έχουν συνδυαστεί τα αντικείμενα. Όπως και στο γράφημα της ανάλυσης κατά παράγοντες, μπορούμε να ψάξουμε μια λύση στην οποία ένας επιπλέον συνδυασμός συστάδων ή αντικειμένων θα προέρχεται σε μια αυξημένη σε μεγάλο βαθμό απόσταση. Έτσι προκύπτει το ερώτημα ποια απόσταση είναι μεγάλη.
Ένας πιθανός τρόπος να λυθεί αυτό το πρόβλημα είναι, να κάνουμε γράφημα του αριθμού των συστάδων (άξονας x) σε σχέση με την απόσταση στην οποία συνδυάζονται τα αντικείμενα ή οι συστάδες (άξονας y). Μετά, χρησιμοποιώντας αυτό το γράφημα ψάχνουμε για την διακριτική διακοπή (elbow). Το SPSS δεν παράγει αυτόματα αυτό το γράφημα - πρέπει να χρησιμοποιήσουμε τις αποστάσεις που μας παρέχει για να σχεδιάσουμε ένα ραβδόγραμμα χρησιμοποιώντας ένα κοινό πρόγραμμα όπως το microsoft excel.
Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε το δενδρόγραμμα που ουσιαστικά περιέχει τις ίδιες πληροφορίες. Το SPSS παρέχει ένα δενδρόγραμμα, όμως αυτό διαφέρει ελαφρά από αυτό της εικόνας 9. Συγκεκριμένα, το SPSS αλλάζει την κλίμακα των αποστάσεων σ' ένα εύρος από 0-25, όπου το τελευταίο βήμα συγχώνευσης σε μια λύση μιας συστάδας συμβαίνει σε απόσταση 25. Η αλλαγή κλίμακας συχνά επιμηκύνει τα βήματα συγχώνευσης, κάνοντας έτσι τις διακοπές, που συμβαίνουν σε αυξημένη απόσταση, πιο εμφανείς.



Εικόνα 9: Δενδρόγραμμα

Παρά το γεγονός αυτό, αυτός ο κανόνας απόφασης που βασίζεται στις αποστάσεις δεν λειτουργεί πολύ καλά σε όλες τις περιπτώσεις. Είναι συχνά δύσκολο να εντοπίσουμε που συμβαίνει πραγματικά η διακοπή. Αυτό συμβαίνει και στο παραπάνω παράδειγμα. Κοιτώντας το δενδρόγραμμα μπορούμε να δικαιολογήσουμε μια λύση δυο συστάδων ([A,B,C,D,E,F] και [G]), καθώς και μια λύση 5 συστάδων ([B,C,E] , [A], [D], [F], [G]).
Η έρευνα έχει προτείνει διάφορες άλλες διαδικασίες για τον καθορισμό του αριθμού των συστάδων σ' ένα σύνολο δεδομένων. Ειδικότερα, το κριτήριο αναλογίας διασποράς (varriants ratio criterion- VRC) των Calinski και Harabasz (1974) έχει αποδειχτεί ότι λειτουργεί καλά σε πολλές περιπτώσεις. Για μια λύση με n αντικείμενα και k τμήματα, το κριτήριο δίνεται από την σχέση:

$$VRC_k = \left(SS_B / (k-1)\right) / \left(SS_W / (n-k)\right),$$

όπου SS_B είναι το άθροισμα των τετραγώνων μεταξύ των τμημάτων και SS_W είναι το άθροισμα των τετραγώνων εντός των τμημάτων. Το κριτήριο φαίνεται γνωστό, αφού δεν είναι άλλο από την τιμή F μιας ανάλυσης διασποράς με έναν παράγοντα (one-way ANOVA) με το k να αντιπροσωπεύει τα επίπεδα παραγόντων. Συνεπώς το VRC μπορεί να υπολογιστεί εύκολα χρησιμοποιώντας το SPSS, παρόλο που δεν είναι άμεσα διαθέσιμο στα δεδομένα εξόδου των διαδικασιών δημιουργίας συστάδων.
Για να προσδιορίσουμε τελικά τον κατάλληλο αριθμό τμημάτων, υπολογίζουμε το ω_k για τη λύση κάθε τμήματος σύμφωνα με τον τύπο:

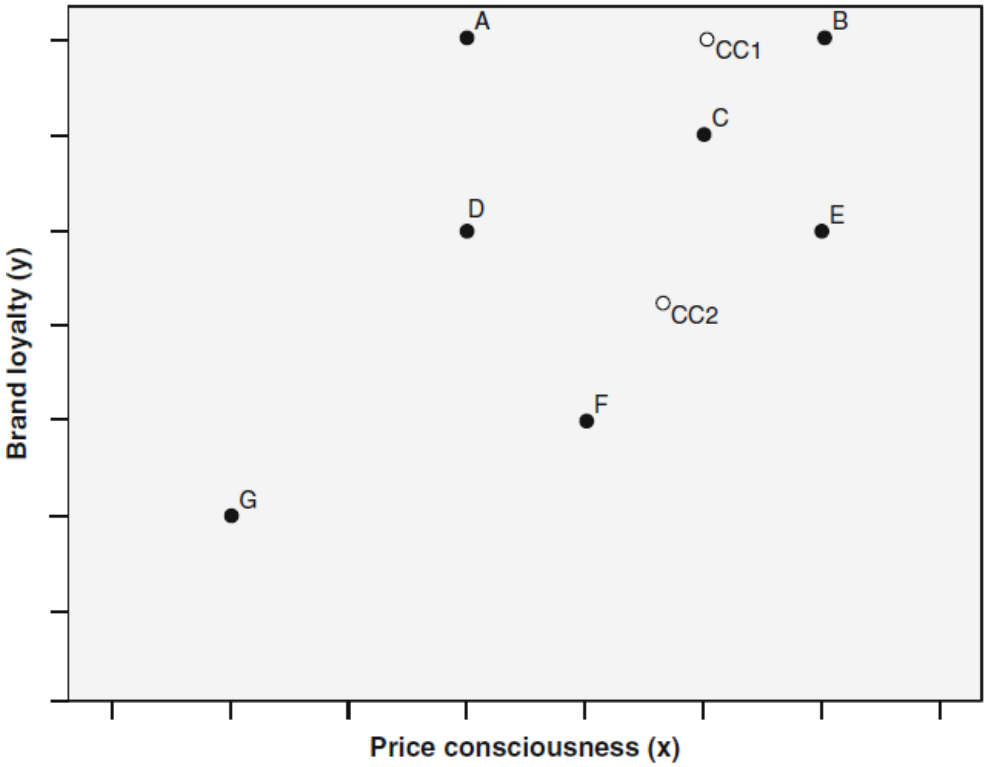
$$\omega_k = \left(VRC_{k+1} - VRC_k\right) - \left(VRC_k - VRC_{k-1}\right).$$

Στο επόμενο βήμα, επιλέγουμε τον αριθμό των τμημάτων που ελαχιστοποιεί την τιμή του ω_k. Εξαιτίας του όρου VRC_{k-1}, ο ελάχιστος αριθμός συστάδων που μπορεί να επιλεγεί είναι τρία, που είναι ένα μειονέκτημα του κριτηρίου, περιορίζοντας έτσι την εφαρμογή του στην πράξη.
Συνολικά, τα δεδομένα μπορούν συχνά να παρέχουν πρόχειρη καθοδήγηση όσον αφορά στον αριθμό των συστάδων που θα επιλέξουμε. Συνεπώς πρέπει να επαναφερθούμε σε πρακτικές εκτιμήσεις. Ενίοτε μπορεί να έχουμε εκ των προτέρων γνώσεις ή μια θεωρία στην οποία μπορούμε να βασίσουμε την επιλογή μας. Όμως πρώτα και κύρια, πρέπει να εξασφαλίσουμε ότι τα αποτελέσματα είναι ερμηνεύσιμα και εποικοδομητικά. Όχι μόνο πρέπει ο αριθμός των συστάδων να είναι αρκετά μικρός για να εξασφαλίζεται η διαχείρισή τους αλλά και κάθε τμήμα πρέπει να είναι αρκετά μεγάλο για να δικαιολογεί στρατηγική προσοχή.

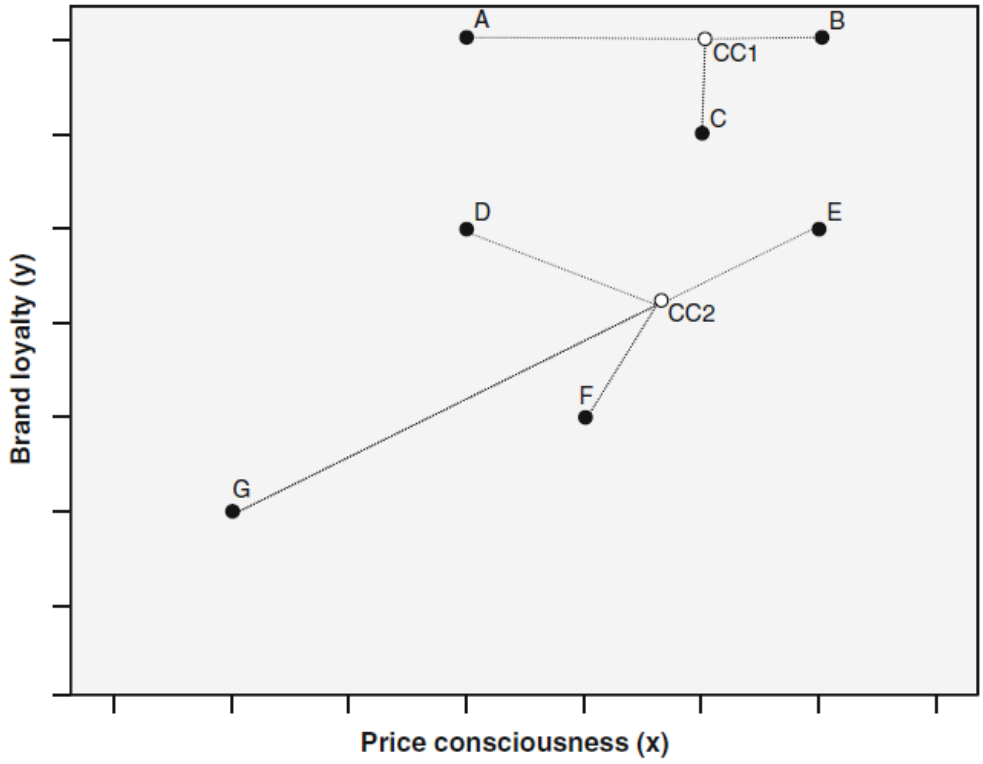
1.3.7 Μέθοδοι Διαμέρισης: k-means

Μια άλλη σημαντική ομάδα διαδικασιών δημιουργίας συστάδων είναι οι μέθοδοι διαμέρισης (ή διαχωριστικές). Όπως στις ιεραρχικές μεθόδους, υπάρχει μια σειρά από διαφορετικούς αλγόριθμους. Από αυτούς η διαδικασία k-means είναι ο σημαντικότερος για την έρευνα αγοράς. Ο αλγόριθμος της k-means ακολουθεί μια τελείως διαφορετική ιδέα από τις ιεραρχικές μεθόδους που αναφέρθηκαν παραπάνω. Αυτός ο αλγόριθμος δεν βασίζεται σε μέτρα απόστασης όπως η Ευκλείδεια ή η απόσταση city-block, αλλά χρησιμοποιεί την διασπορά εντός συστάδας σαν μέτρο για τη δημιουργία ομογενών συστάδων. Συγκεκριμένα, η διαδικασία στοχεύει στην κατάτμηση των δεδομένων μ' ένα τέτοιο τρόπο ώστε να ελαχιστοποιείται η διασπορά εντός συστάδας. Συνεπώς δεν χρειάζεται να αποφασίσουμε για ένα μέτρο απόστασης στο πρώτο βήμα της ανάλυσης.
Η διαδικασία δημιουργίας συστάδων ξεκινάει εκχωρώντας τυχαία αντικείμενα σ' έναν αριθμό συστάδων (Αυτό ισχύει μόνο για τον αυθεντικό σχεδιασμό του αλγορίθμου. Το SPSS δεν επιλέγει τυχαία τα κέντρα.). Τα αντικείμενα, στη συνέχεια εκχωρούνται εκ νέου διαδοχικά σε άλλες συστάδες, έτσι ώστε να ελαχιστοποιηθεί η διασπορά εντός συστάδας, που ουσιαστικά είναι η (τετραγωνισμένη) απόσταση κάθε παρατήρησης από το κέντρο της συστάδας. Αν η ανακατανομή ενός αντικειμένου σε μια άλλη συστάδα ελαττώνει την διασπορά εντός της συστάδας, το αντικείμενο εκχωρείται στη συστάδα αυτή.
Με τις ιεραρχικές μεθόδους, ένα αντικείμενο παραμένει σε μια συστάδα, όταν έχει εκχωρηθεί σε αυτή, αλλά με την k-means, οι σχέσεις μεταξύ των συστάδων μπορούν να αλλάξουν στη πορεία της διαδικασίας δημιουργίας τους. Συνεπώς, η k-means δεν χτίζει μια ιεραρχία, γι' αυτό κι αυτή η προσέγγιση συχνά καλείται μη ιεραρχική.

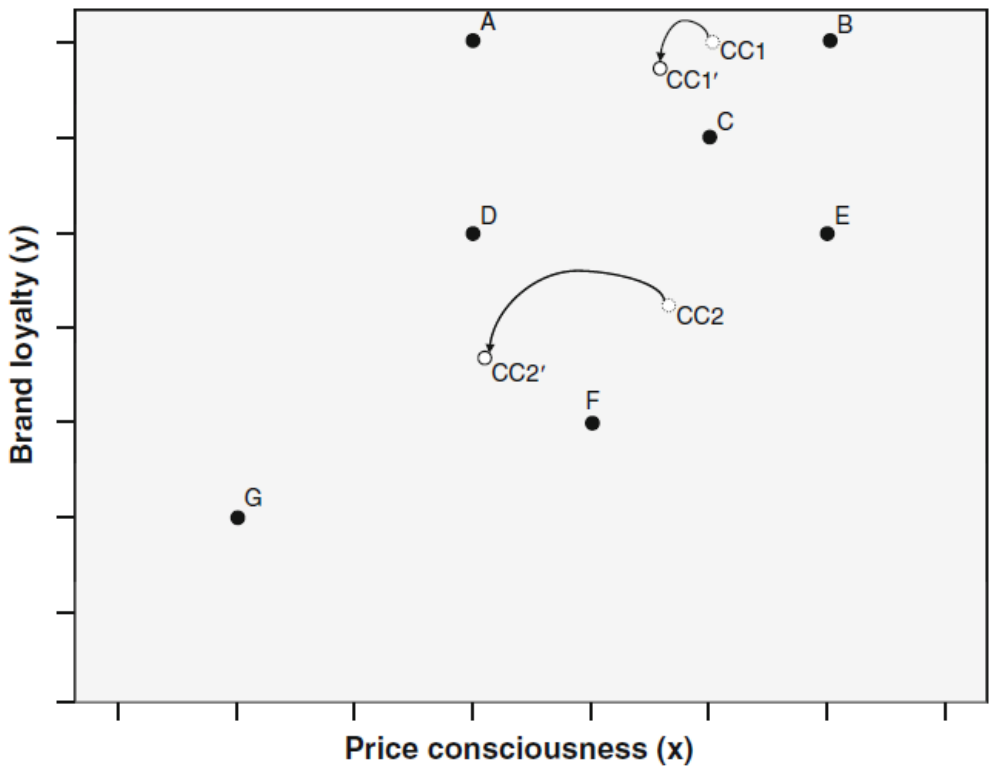
Για την καλύτερη κατανόηση της προσέγγισης, ας δούμε πως λειτουργεί στην πράξη. Οι εικόνες 10-13 παρουσιάζουν την διαδικασία k-means.



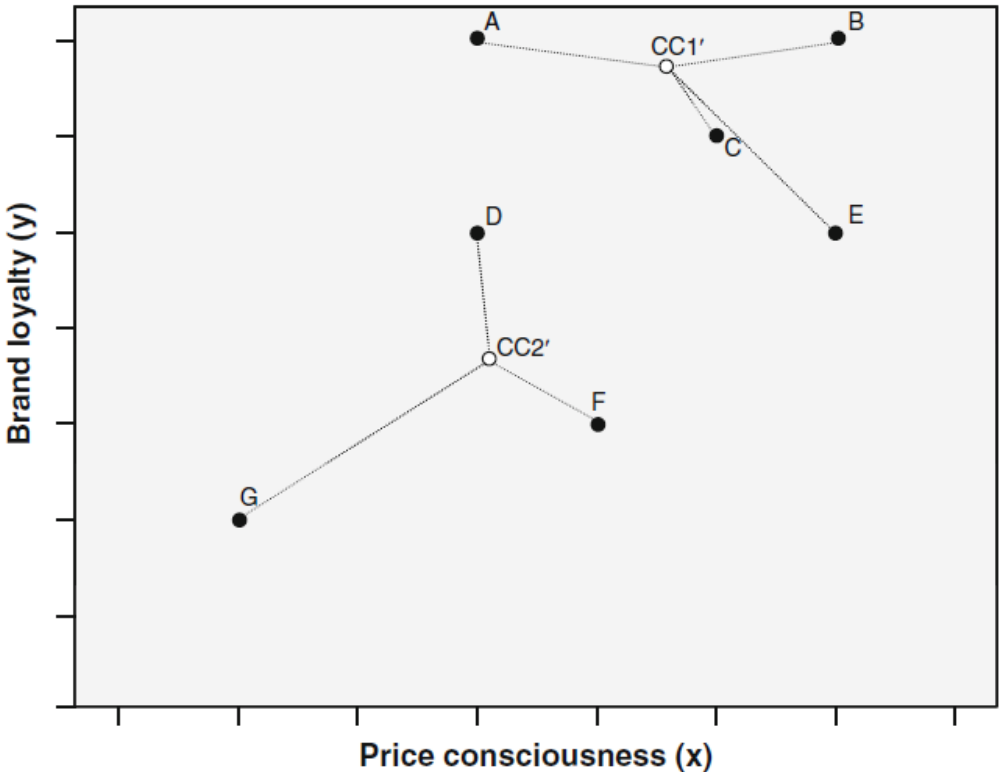
Εικόνα 10: Η διαδικασία k-means (1^{ο} βήμα)



Εικόνα 11: Η διαδικασία k-means (2^{ο} βήμα)



Εικόνα 12: Η διαδικασία k-means (3ο βήμα)



Εικόνα 13: Η διαδικασία k-means (4ο βήμα)

Πριν την ανάλυση πρέπει να αποφασίσουμε για τον αριθμό των συστάδων. Ο πελάτης θα μπορούσε για παράδειγμα να μας πει πόσα τμήματα χρειάζονται ή μπορεί να ξέρουμε από προηγούμενη έρευνα. Βασιζόμενος σε αυτές τις πληροφορίες, ο αλγόριθμος επιλέγει τυχαία ένα κέντρο για κάθε συστάδα (βήμα 1). Στο παράδειγμα, δυο κέντρα έχουν εισαχθεί τυχαία, που αντιπροσωπεύονται από τα σημεία CC1 (πρώτη συστάδα) και CC2 (δεύτερη συστάδα) στην εικόνα 10. Μετά από αυτό (βήμα 2), υπολογίζονται οι Ευκλείδειες αποστάσεις από τα κέντρα των συστάδων προς κάθε αντικείμενο. Στη συνέχεια κάθε αντικείμενο εκχωρείται στη συστάδα, από της οποίας το κέντρο απέχει λιγότερο. Στο παράδειγμα (εικόνα 11), τα αντικείμενα A, B και C εκχωρούνται στην πρώτη συστάδα, ενώ τα αντικείμενα D, E, F και G εκχωρούνται στη δεύτερη. Έτσι έχουμε την αρχική διαμέριση του αντικειμένου σε δυο συστάδες.
Στο τρίτο βήμα υπολογίζουμε το γεωμετρικό κέντρο κάθε συστάδας βασιζόμενοι στην αρχική διαμέριση. Αυτό γίνεται υπολογίζοντας τις μέσες τιμές των αντικειμένων που ανήκουν στη συστάδα, για κάθε μια από τις μεταβλητές x και y. Όπως βλέπουμε στην εικόνα 12 και τα δυο κέντρα των συστάδων μετατοπίζονται σε νέες θέσεις (CC1΄ για την πρώτη και CC2΄ για την δεύτερη συστάδα).
Στο τέταρτο βήμα, οι αποστάσεις κάθε αντικειμένου στα επανατοποθετημένα κέντρα των συστάδων υπολογίζονται και τα αντικείμενα εκχωρούνται ξανά σε μια συγκεκριμένη συστάδα με βάση την ελάχιστη απόστασή τους από τα κέντρα των συστάδων. Αφού η θέση των κέντρων των συστάδων έχει αλλάξει, μπορεί να οδηγηθούμε σε διαφορετική λύση από την αρχική διαμέριση. Αυτό ισχύει και στο παράδειγμα, αφού το αντικείμενο E είναι τώρα - σε αντίθεση με την αρχική διαμέριση - πιο κοντά στο κέντρο της πρώτης συστάδας απ' ότι της δεύτερης. Συνεπώς το αντικείμενο αυτό εκχωρείται τώρα στην πρώτη συστάδα (εικόνα 13). Η διαδικασία k-means επαναλαμβάνει το τρίτο βήμα και υπολογίζει ξανά τα κέντρα των νεοσυσταθέντων συστάδων. Με άλλα λόγια τα βήματα 3 και 4 επαναλαμβάνονται μέχρι να φτάσουμε σ' ένα προκαθορισμένο αριθμό επαναλήψεων ή να επιτευχθεί σύγκλιση (δηλαδή να μην υπάρχουν αλλαγές στις συστάδες).
Γενικά, η k-means είναι ανώτερη από τις ιεραρχικές μεθόδους, αφού επηρεάζεται λιγότερο από τις ακραίες τιμές και την παρουσία άσχετων μεταβλητών. Επιπλέον, η k-means μπορεί να εφαρμοσθεί σε πολύ μεγάλα σύνολα δεδομένων, αφού η διαδικασία έχει λιγότερες υπολογιστικές απαιτήσεις από τις ιεραρχικές μεθόδους. Για μέγεθος δείγματος πάνω από 500 είναι προτιμότερη η χρήση της k-means, ειδικά αν χρησιμοποιούνται πολλές μεταβλητές. Από μια αυστηρά στατιστική άποψη, η k-means πρέπει να χρησιμοποιείται μόνο για δεδομένα σε διάστημα ή σε αναλογική κλίμακα, αφού η διαδικασία βασίζεται στις Ευκλείδειες αποστάσεις. Όμως η διαδικασία χρησιμοποιείται και σε αριθμητικά δεδομένα, παρόλο που μπορεί να υπάρχουν κάποιες αλλοιώσεις.
Ένα πρόβλημα που συνδέεται με την εφαρμογή της k-means σχετίζεται με το γεγονός ότι ο ερευνητής πρέπει να προκαθορίσει τον αριθμό των συστάδων που θα κρατήσει από τα δεδομένα. Αυτό κάνει την k-means λιγότερο ελκυστική σε κάποιους και εμποδίζει ακόμη την τακτική εφαρμογή της στην πράξη. Ένας εναλλακτικός τρόπος αντιμετώπισης που πολλοί ερευνητές της αγοράς χρησιμοποιούν τακτικά είναι να εφαρμόσουν μια ιεραρχική διαδικασία για να καθορίσουν τον αριθμό των συστάδων και μετά την k-means. Αυτό επιτρέπει επίσης στον χρήστη να βρει αρχικές τιμές για τα αρχικά κέντρα, αντιμετωπίζοντας ένα δεύτερο πρόβλημα, που σχετίζεται με την ευαισθησία της διαδικασίας στην αρχική ταξινόμηση.

1.3.8 Δημιουργία Συστάδων σε δύο βήματα (two-step clustering)

Έχουμε ήδη αναφερθεί στο θέμα της ανάλυσης μεταβλητών μετρημένων σε διαφορετικές κλίμακες. Η ανάλυση κατά συστάδες σε δυο βήματα που αναπτύχθηκε από τον Chiu

(2001), έχει σχεδιαστεί ειδικά για να διαχειρίζεται αυτό το πρόβλημα. Όπως η k-means, η διαδικασία αυτή μπορεί να ανταπεξέλθει αποτελεσματικά σε πολύ μεγάλα σύνολα δεδομένων.
Το όνομα της μεθόδου υποδεικνύει ότι ο αλγόριθμος βασίζεται σε μια προσέγγιση δύο σταδίων: Στο πρώτο στάδιο ο αλγόριθμος αναλαμβάνει μια διαδικασία που μοιάζει αρκετά με τον αλγόριθμο της k-means. Βασιζόμενοι σ' αυτά τα αποτελέσματα η διαδικασία των δύο βημάτων διεξάγει μια τροποποιημένη ιεραρχική συσωρευτική διαδικασία δημιουργίας συστάδων, που συνδυάζει τα αντικείμενα διαδοχικά για να παράξει ομογενείς συστάδες. Αυτό γίνεται χτίζοντας ένα δένδρο, του οποίου τα "φύλλα" αντιπροσωπεύουν διακριτά αντικείμενα στο σύνολο των δεδομένων. Η διαδικασία μπορεί να διαχειριστεί κατηγορηματικές και συνεχείς μεταβλητές ταυτόχρονα και προσφέρει στον χρήστη τη δυνατότητα να συγκεκριμενοποιήσει τον αριθμό των συστάδων, καθώς και τον μέγιστο αριθμό των συστάδων ή να επιτρέψει στην τεχνική να επιλέξει αυτόματα τον αριθμό των συστάδων με βάση κριτήρια στατιστικής εκτίμησης. Ομοίως, η διαδικασία καθοδηγεί την απόφαση, πόσες συστάδες θα διατηρηθούν από τα δεδομένα υπολογίζοντας μέτρα καταλληλότητας, όπως το κριτήριο πληροφορίας του Akaike (AIC) ή το κριτήριο πληροφορίας Bayes (BIC). Επιπλέον, η διαδικασία υποδεικνύει τη σημασία κάθε μεταβλητής για τη δημιουργία μιας συγκεκριμένης συστάδας. Αυτά τα επιθυμητά χαρακτηριστικά κάνουν την κάπως λιγότερο δημοφιλή διαδικασία των δύο βημάτων μια καλή εναλλακτική στις παραδοσιακές μεθόδους.

1.3.9 Εγκυρότητα και Ερμηνεία της Λύσης

Πριν ερμηνεύσουμε τη λύση της ανάλυσης πρέπει να εκτιμήσουμε τη σταθερότητα και την εγκυρότητά της. Η σταθερότητα εκτιμάται χρησιμοποιώντας διαφορετικές διαδικασίες δημιουργίας συστάδων στα ίδια δεδομένα και ελέγχοντας αν αυτές δίνουν τα ίδια αποτελέσματα. Στην ιεραρχική παραγωγή συστάδων, μπορούμε ομοίως να χρησιμοποιήσουμε διαφορετικά μέτρα απόστασης. Παρόλα αυτά συμβαίνει συχνά τα αποτελέσματα να αλλάζουν ακόμα κι αν η λύση είναι επαρκής. Το πόση διασπορά πρέπει να επιτρέψουμε πριν αμφισβητήσουμε τη σταθερότητα της λύσης είναι θέμα επιλογής. Μια άλλη συνήθης προσέγγιση είναι να χωρίσουμε το σύνολο των δεδομένων στα δύο και να αναλύσουμε μετά κάθε υποσύνολο χωριστά χρησιμοποιώντας τις ίδιες ρυθμίσεις παραμέτρων. Μετά συγκρίνουμε τα γεωμετρικά κέντρα των συστάδων των δύο λύσεων. Εάν αυτά δεν διαφέρουν σημαντικά μπορούμε να υποθέσουμε ότι η ολική λύση έχει έναν υψηλό βαθμό σταθερότητας. Όταν χρησιμοποιούμε ιεραρχικές μεθόδους αξίζει τον κόπο να αλλάξουμε την σειρά των αντικειμένων στο σύνολο των δεδομένων και να εκτελέσουμε ξανά την ανάλυση για να ελέγξουμε τη σταθερότητα των αποτελεσμάτων. Τα αποτελέσματα δεν πρέπει φυσικά να εξαρτώνται από την σειρά των αντικειμένων. Αν αυτό συμβαίνει πρέπει να εξακριβώσουμε αν κάποιες προφανείς ακραίες τιμές επηρεάζουν τα αποτελέσματα της αλλαγής στη σειρά.
Η αξιολόγηση της αξιοπιστίας της λύσης είναι στενά συνδεδεμένη με τα παραπάνω, αφού η αξιοπιστία αναφέρεται στο βαθμό στον οποίο η λύση είναι σταθερή διαχρονικά. Αν τα τμήματα αλλάζουν γρήγορα τη σύνθεσή τους ή τα μέλη της την συμπεριφορά τους, οι στοχευμένες στρατηγικές είναι πιθανό να μην πετύχουν. Ως εκ τούτου, ένας συγκεκριμένος βαθμός σταθερότητας είναι αναγκαίος για τη διασφάλιση ότι οι στρατηγικές μάρκετινγκ μπορούν να υλοποιηθούν και να παράγουν ικανοποιητικά αποτελέσματα. Αυτό μπορεί να αξιολογηθεί με αναθεώρηση και αναπαραγωγή των αποτελεσμάτων της ομαδοποίησης σε ένα μεταγενέστερο χρονικό σημείο.
Για την επικύρωση της λύσης, θα πρέπει να αξιολογηθεί η εγκυρότητα των κριτηρίων της. Στην έρευνα, μπορούμε να επικεντρωθούμε σε μεταβλητές κριτηρίων που έχουν μία

θεωρητικά βασιζόμενη σχέση με της μεταβλητές της ομαδοποίησης, αλλά δεν συμπεριλήφθηκαν στην ανάλυση. Στην έρευνα αγοράς, οι μεταβλητές κριτηρίων συνήθως αφορούν σε διαχειριστικά αποτελέσματα, όπως οι πωλήσεις ανά άτομο ή η ικανοποίηση. Αν αυτές οι μεταβλητές κριτηρίων διαφέρουν σημαντικά, μπορούμε να συμπεράνουμε ότι οι συστάδες είναι διακριτές ομάδες με εγκυρότητα κριτηρίων.

Για να κρίνουμε την εγκυρότητα, θα πρέπει επίσης να αξιολογήσουμε την εγκυρότητα προσώπων και, αν είναι δυνατόν την εγκυρότητα πραγματογνωμοσύνης. Ενώ αρχικά εξετάζουμε την εγκυρότητα κριτηρίων όταν επιλέγουμε τις μεταβλητές ομαδοποίησης, καθώς και σε αυτό το τελευταίο στάδιο της διαδικασίας ανάλυσης, η αξιολόγηση της εγκυρότητας προσώπων είναι μάλλον μια διαδικασία και όχι ένα μεμονωμένο γεγονός. Το κλειδί για την επιτυχή κατάτμηση είναι να εξετάζουμε κριτικά τα αποτελέσματα διαφορετικών αναλύσεων κατά συστάδες (για παράδειγμα με τη χρήση διαφορετικών αλγορίθμων στα ίδια δεδομένα) από την άποψη διοικητικού ενδιαφέροντος. Αυτό υπογραμμίζει το διερευνητικό χαρακτήρα της μεθόδου. Τα παρακάτω κριτήρια βοηθούν στην επιλογή αξιολόγησης μιας λύσης ομαδοποίησης (Dibb 1999, Tonks 2009, Kotler και Keller 2009).

Ουσιαστικά: Τα τμήματα είναι μεγάλα και αρκετά επικερδή ώστε να εξυπηρετούν.

Προσβάσιμα: Τα τμήματα μπορούν να προσπελαστούν και να εξυπηρετηθούν αποτελεσματικά, που απαιτεί να χαρακτηρισθούν μέσω παρατηρήσιμων μεταβλητών.

Διαχωρίσιμα: Τα τμήματα μπορούν να διακριθούν εννοιολογικά και να ανταποκριθούν με διαφορετικό τρόπο σε διαφορετικά προγράμματα μάρκετινγκ.

Πρακτικά: Αποτελεσματικά προγράμματα μπορούν να δημιουργηθούν για να προσελκύσουν και να εξυπηρετήσουν τα τμήματα.

Σταθερά: Μόνο τα τμήματα που είναι σταθερά στο χρόνο μπορούν να παρέχουν την αναγκαία βάση για μια επιτυχημένη στρατηγική μάρκετινγκ.

Φειδωλά: Για να έχει νόημα στη διαχείριση, μόνο ένα μικρό σύνολο από σημαντικές συστάδες θα πρέπει να προσδιοριστεί.

Αναγνωρίσιμα: Για να εξασφαλιστεί αποδοχή από τη διοίκηση, η σύνθεση των τμημάτων πρέπει να είναι κατανοητή.

Σχετικά: Τα τμήματα πρέπει να είναι συναφή όσον αφορά τις αρμοδιότητες και τους στόχους της εταιρείας.

Συμπαγή: Τα τμήματα παρουσιάζουν υψηλό βαθμό ομοιογένειας στο εσωτερικό των τμημάτων και ετερογένειας μεταξύ των τμημάτων.

Συμβατότητα: Τα αποτελέσματα της τμηματοποίησης ανταποκρίνονται στις απαιτήσεις άλλων διοικητικών λειτουργιών.

Το τελευταίο βήμα κάθε ανάλυσης κατά συστάδες είναι η ερμηνεία των συστάδων.

Η ερμηνεία των συστάδων πάντα περιλαμβάνει την εξέταση των γεωμετρικών κέντρων των συστάδων, που είναι οι μέσες τιμές των μεταβλητών ομαδοποίησης όλων των αντικειμένων μιας συγκεκριμένης συστάδας. Αυτό το βήμα είναι υψίστης σημασίας, αφού η ανάλυση ρίχνει φως στο αν τα τμήματα είναι εννοιολογικά διακριτά. Μόνο αν συγκεκριμένες συστάδες φαίνονται σημαντικά διαφορετικές σημαίνει ότι σύμφωνα με αυτές τις μεταβλητές μπορούν να διακριθούν – από άποψη δεδομένων τουλάχιστον. Αυτό μπορεί να διαπιστωθεί εύκολα συγκρίνοντας τις συστάδες με ανεξάρτητα δείγματα t-test ή ανάλυση διασποράς (ANOVA).

Χρησιμοποιώντας αυτές τις πληροφορίες, μπορούμε επίσης να προσπαθήσουμε να καταλήξουμε σε ένα χαρακτηριστικό όνομα ή ετικέτα για κάθε συστάδα. Δηλαδή, ένα που να αντανακλά επαρκώς τα αντικείμενα της συστάδας, Αυτό είναι συνήθως ένα πολύ δύσκολο έργο. Επιπλέον, οι μεταβλητές ομαδοποίησης είναι συχνά μη παρατηρήσιμες, που δημιουργεί άλλο ένα πρόβλημα. Πώς μπορούμε να επιλέξουμε σε ποιο τμήμα πρέπει να τοποθετηθεί ένα αντικείμενο αν τα μη παρατηρίσιμα χαρακτηριστικά του, όπως τα

στοιχεία της προσωπικότητας, οι προσωπικές αξίες και ο τρόπος ζωής είναι άγνωστα; Ωστόσο, αυτό δεν είναι εφικτό στις περισσότερες περιπτώσεις και γι' αυτό οι ερευνητές προσπαθούν να προσδιορίσουν παρατηρήσιμες μεταβλητές που να αντικατοπτρίζουν καλύτερα την κατάτμηση των αντικειμένων.
Αν είναι δυνατόν, για παράδειγμα, να προσδιοριστούν οι δημογραφικές μεταβλητές που οδηγούν σε έναν αρκετά όμοιο διαχωρισμό όπως αυτός που προέκυψε απ' την κατάτμηση, τότε είναι εύκολο να εκχωρηθεί ένα νέο αντικείμενο σε ένα συγκεκριμένο τμήμα βάσει αυτών των δημογραφικών χαρακτηριστικών. Αυτές οι μεταβλητές μπορούν να χρησιμοποιηθούν στη συνέχεια για το χαρακτηρισμό συγκεκριμένων τμημάτων, μία ενέργεια που ονομάζεται συνήθως profiling.
Για παράδειγμα, υποθέτουμε ότι χρησιμοποιήσαμε ένα σύνολο στοιχείων για να εκτιμήσουμε τις αξίες των ερωτηθέντων και μάθαμε ότι ένα συγκεκριμένο τμήμα περιλαμβάνει τους ερωτηθέντες που εκτιμούν την αυτοπραγμάτωση, την απόλαυση της ζωής και μια αίσθηση ολοκλήρωσης, ενώ αυτό δεν συμβαίνει σε άλλο τμήμα. Αν ήμασταν σε θέση να προσδιορίσουμε επεξηγηματικές μεταβλητές όπως το φύλο ή η ηλικία, που διακρίνουν επαρκώς αυτά τα τμήματα τότε θα μπορούσαμε να εκχωρήσουμε και ένα νέο άτομο βασιζόμενοι στις παραμέτρους αυτών των παρατηρήσιμων μεταβλητών των οποίων τα χαρακτηριστικά μπορεί να είναι ακόμα άγνωστα.
Στον πίνακα 11 βλέπουμε τα βήματα που σχετίζονται με την ομαδοποίηση σε 2 στάδια.

Θεωρία	Ενέργεια
Ερευνητικό πρόβλημα	
Αναγνώριση ομογενών ομάδων αντικειμένων σε έναν πληθυσμό	
Επιλογή μεταβλητών ομαδοποίησης που θα χρησιμοποιηθούν για να δημιουργήσουν τα τμήματα	Επιλέγουμε σχετικές μεταβλητές που δυνητικά παρουσιάζουν υψηλούς βαθμούς ισχύος κριτηρίων όσον αφορά έναν συγκεκριμένο διοικητικό στόχο.
Απαιτήσεις	
Επαρκές μέγεθος δείγματος	Βεβαιωνόμαστε ότι η σχέση μεταξύ των αντικειμένων και των μεταβλητών ομαδοποίησης είναι λογική (πρόχειρη κατεύθυνση: ο αριθμός των παρατηρήσεων πρέπει να είναι τουλάχιστον 2^m, όπου m είναι ο αριθμός των μεταβλητών). Βεβαιωνόμαστε ότι το μέγεθος του δείγματος είναι αρκετά μεγάλο ώστε να εγγυάται ουσιαστικά τμήματα.
Χαμηλά επίπεδα συγγραμμικότητας μεταξύ των μεταβλητών	► Ανάλυση ► Συσχέτιση ► Ελέγχουμε τη διασπορά Καταργούμε ή αντικαθιστούμε τις μεταβλητές με υψηλή συσχέτιση (συντελεστής συσχέτισης > 0.90).
Προδιαγραφές	
Επιλογή διαδικασίας ομαδοποίησης	Αν υπάρχει περιορισμένος αριθμός αντικειμένων στο σύνολο των δεδομένων ή δεν γνωρίζουμε τον αριθμό των συστάδων: ► Ανάλυση ► Ταξινόμηση ► Ιεραρχική ομαδοποίηση Αν υπάρχουν πολλές παρατηρήσεις (>500) στο σύνολο των δεδομένων και γνωρίζουμε εκ τον προτέρων τον αριθμό των συστάδων: ► Ανάλυση ► Ταξινόμηση ► k-means

(συνέχεια)

Θεωρία	Ενέργεια
	Αν υπάρχουν πολλές παρατηρήσεις στο σύνολο των δεδομένων και οι μεταβλητές ομαδοποίησης έχουν μετρηθεί σε διαφορετικές κλίμακες: ► Ανάλυση ► Ταξινόμηση ►Two-step cluster
Επιλογή μέτρου ομοιότητας ή διαφοράς (μόνο ιεραρχική ομαδοποίηση και ομαδοποίηση σε 2 στάδια)	*Ιεραρχικές μέθοδοι:* ► Ανάλυση ► Ταξινόμηση ► Ιεραρχική ομαδοποίηση ► Μέτρο Ανάλογα με το επίπεδο της κλίμακας, επιλέγουμε το μέτρο. Μετατρέπουμε τις μεταβλητές με πολλαπλές κατηγορίες σε ένα σύνολο δυαδικών μεταβλητών και χρησιμοποιούμε συντελεστές αντιστοιχίας. Αν χρειάζεται τυποποιούμε τις μεταβλητές (σε εύρος από 0 ως 1 ή από -1 ως 1). *Ομαδοποίηση σε 2 στάδια:* ► Ανάλυση ► Ταξινόμηση ► Two-step cluster ► Μέτρο απόστασης Χρησιμοποιούμε Ευκλείδειες αποστάσεις όταν όλες οι μεταβλητές είναι συνεχείς. Για μεικτές μεταβλητές, χρησιμοποιούμε λογαριθμική πιθανότητα (log-likelihood).
Επιλογή αλγορίθμου ομαδοποίησης (μόνο ιεραρχική ομαδοποίηση)	► Ανάλυση ► Ταξινόμηση ► Ιεραρχική ομαδοποίηση ► Μέθοδος ομαδοποίησης Χρησιμοποιούμε τη μέθοδο του Ward αν περιμένουμε συστάδες ίδιου μεγέθους και δεν έχουμε ακραίες τιμές. Κατά προτίμηση χρησιμοποιούμε ενιαία σύνδεση, για να ανιχνεύσουμε και τις ακραίες τιμές.
Επιλογή αριθμού συστάδων	*Ιεραρχική ομαδοποίηση:* Εξετάζουμε το δενδρόγραμμα: ► Ανάλυση ► Ταξινόμηση ► Ιεραρχική ομαδοποίηση ► Γραφήματα ► Δενδρόγραμμα Σχεδιάζουμε ένα γράφημα (χρησιμοποιώντας για παράδειγμα το Microsoft Excel) βασιζόμενοι στους συντελεστές στο χρονοδιάγραμμα της συσσώρευσης. Υπολογίζουμε το VRC χρησιμοποιώντας ανάλυση διασποράς: ► Ανάλυση ► Σύγκριση μέσων ►One-way ANOVA Μετακινούμε τη μεταβλητή μέλους της συστάδας στο πλαίσιο **Factor** και τις μεταβλητές ομαδοποίησης στο πλαίσιο **Dependent List.** Υπολογίζουμε το VRC για κάθε τμήμα και συγκρίνουμε τις τιμές. *k-means:* Κάνουμε ιεραρχική ανάλυση κατά συστάδες και επιλέγουμε τον αριθμό των τμημάτων βασιζόμενοι σε ένα δενδρόγραμμα ή ένα γράφημα. Χρησιμοποιούμε αυτές τις

(συνέχεια)

Θεωρία	Ενέργεια
	πληροφορίες για να εκτελέσουμε την k-means με k συστάδες. Υπολογίζουμε το VRC χρησιμοποιώντας ανάλυση διασποράς (ANOVA): ► Ανάλυση ► Ταξινόμηση ► k-means συστάδες ► Πίνακας ANOVA ► Υπολογισμός VRC για κάθε τμήμα και σύγκριση τιμών. *Ομαδοποίηση σε 2 στάδια:* Καθορίζουμε τον μέγιστο αριθμό συστάδων: ► Ανάλυση ► Ταξινόμηση ► Ομαδοποίηση σε 2 στάδια ► Αριθμός συστάδων Κάνουμε ξεχωριστές αναλύσεις χρησιμοποιώντας το AIC και εναλλακτικά το BIC ως κριτήρια ομαδοποίησης: ► Ανάλυση ► Ταξινόμηση ► Ομαδοποίηση σε 2 στάδια ► Κριτήριο ομαδοποίησης Εξετάζουμε τα δεδομένα εξόδου της ομαδοποίησης.
Επικύρωση και ερμηνεία της λύσης	
Αξιολόγηση της σταθερότητας της λύσης	Εκτελούμε ξανά την ανάλυση χρησιμοποιώντας διαφορετικές διαδικασίες, αλγορίθμους ή μέτρα απόστασης. Χωρίζουμε τα σύνολα δεδομένων στα δύο και υπολογίζουμε τα γεωμετρικά κέντρα των μεταβλητών ομαδοποίησης. Συγκρίνουμε τα κέντρα αυτά για σημαντικές διαφορές (πχ. με t-test ανεξάρτητων δειγμάτων ή ανάλυση διασποράς με έναν παράγοντα) Αλλάζουμε τη σειρά των αντικειμένων στο σύνολο των δεδομένων (μόνο στην ιεραρχική ομαδοποίηση).
Αξιολόγηση της αξιοπιστίας της λύσης	Αναπαράγουμε την ανάλυση χρησιμοποιώντας ένα ξεχωριστό, νέο σύνολο δεδομένων.
Αξιολόγηση της εγκυρότητας της λύσης	*Εγκυρότητα κριτηρίων:* Αξιολογούμε αν υπάρχουν σημαντικές διαφορές μεταξύ των τμημάτων όσον αφορά σε μία ή περισσότερες μεταβλητές κριτηρίων *Εγκυρότητα προσώπων:* Τα τμήματα πρέπει να είναι ουσιαστικά, προσβάσιμα, διαχωρίσιμα, σταθερά, φειδωλά, αναγνωρίσιμα και σχετικά. Πρέπει να παρουσιάζουν μεγάλο βαθμό ομογένειας εντός τμήματος και ετερογένειας μεταξύ τμημάτων. Τα αποτελέσματα της ομαδοποίησης πρέπει να συμφωνούν με τις απαιτήσεις άλλων διοικητικών λειτουργιών.
Ερμηνεία της λύσης	Εξετάζουμε τα γεωμετρικά κέντρα των συστάδων και εκτιμούμε αν αυτά διαφέρουν σημαντικά το ένα απ᾽ το άλλο (πχ με t-test ή ανάλυση διασποράς

(συνέχεια)

Θεωρία	Ενέργεια
	Προσδιορίζουμε ονόματα ή ετικέτες για κάθε συστάδα και την χαρακτηρίζουμε μέσω παρατηρίσιμων μεταβλητών, αν χρειάζεται (profiling).

Πίνακας 11: Πίνακας βημάτων ανάλυσης με SPSS

Ενώ κάποιες φορές οι εταιρείες αναπτύσσουν δικά τους τμήματα αγοράς, συχνά χρησιμοποιούν τυποποιημένα τμήματα, που καθορίζονται με γνώμονα τις αγοραστικές τάσεις, τις συνήθειες και τις ανάγκες των πελατών και έχουν σχεδιαστεί ειδικά για χρήση από πολλά προϊόντα της αγοράς. Μία από τις πιο δημοφιλείς προσεγγίσεις είναι το σύστημα κατάτμησης του τρόπου ζωής PRIZM που αναπτύχθηκε από την Claritas Inc., μία κορυφαία εταιρεία έρευνας αγοράς. Το PRIZM καθορίζει σε ποιο τμήμα θα ανήκει κάθε νοικοκυριό των ΗΠΑ, από 66 τμήματα που διαχωρίζονται με βάση τα δημογραφικά στοιχεία και τη συμπεριφορά, έτσι ώστε να βοηθήσει τους εμπόρους να διακρίνουν τις προτιμήσεις, τον τρόπο ζωής και την αγοραστική συμπεριφορά αυτών των πελατών.
Στην ιστοσελίδα της εταιρείας (http://www.claritas.com/MyBestSegments/Default.jsp) μπορούμε να δούμε τα προφίλ των τμημάτων. Ένα παράδειγμα ενός τμήματος είναι το «Γκρίζα Δύναμη, που περιλαμβάνει άτομα μεσαίας τάξης, ιδιοκτήτες κατοικίας στα προάστια που προτιμούν να ζουν στο σπίτι τους απ' ότι σε εγκαταστάσεις για συνταξιούχους. Η Γκρίζα Δύναμη αντικατοπτρίζει αυτή την τάση, ένα τμήμα ηλικιωμένων, μεσαίας κλίμακας ατόμων ή ζευγαριών που ζουν σε ήσυχο και άνετο περιβάλλον.

Κεφάλαιο 2: Διακριτική Ανάλυση

2.1 Εισαγωγή

Η διακριτική ανάλυση είναι μία στατιστική τεχνική με βάση την παλινδρόμηση, που χρησιμοποιείται για να καθορίσει ποιες μεταβλητές διαχωρίζουν τα δεδομένα σε δύο ή περισσότερες ομάδες και σε ποια συγκεκριμένη ομάδα ανήκει ένα στοιχείο δεδομένων βάσει των χαρακτηριστικών του. Η διάκριση των ομάδων γίνεται με βάση κάποιες ανεξάρτητες μεταβλητές και οι ομάδες είναι αμοιβαία αποκλειόμενες. Η τεχνική αυτή διαφέρει από άλλες τεχνικές ομαδοποίησης, όπως η ανάλυση κατά συστάδες στο ότι οι ομάδες απ' τις οποίες έχουμε να επιλέξουμε πρέπει να είναι γνωστές εκ των προτέρων, αφού επιθυμούμε να προβλέψουμε την ταξινόμηση μιας νέας παρατήρησης κι όχι να χωρίσουμε τις παρατηρήσεις σε ομάδες. Η αρχή στην οποία στηρίζεται η διακριτική ανάλυση είναι ο γραμμικός συνδυασμός των μεταβλητών, έτσι ώστε η τοποθέτηση ενός αντικειμένου σε μία ομάδα να γίνεται κατά τον βέλτιστο τρόπο. Οι μεταβλητές που χρησιμοποιούνται για την διάκριση των ομάδων ονομάζονται διακριτικές μεταβλητές, ενώ οι γραμμικοί τους συνδυασμοί διακριτικές συναρτήσεις. Οι συναρτήσεις αυτές δίνονται από τη σχέση:

$$D_i = d_{i_1} Z_1 + d_{i_2} Z_2 + ... + d_{i_k} Z_k$$

όπου D_i είναι ο διακριτικός βαθμός της κανονικοποιημένης μεταβλητής D στη συνάρτηση, i, d είναι οι τυποποιημένοι διακριτικοί συντελεστές (στάθμισης) και Z οι τυποποιημένες τιμές των k διακριτικών μεταβλητών. Οι συναρτήσεις είναι διαμορφωμένες έτσι ώστε να μεγιστοποιείται η διάκριση μεταξύ των ομάδων.
Η τεχνική αυτή δημιουργήθηκε από τον Ronald Fisher (1936) για να λύσει το πρόβλημα της ταξινόμησης των φυτών.
Έχοντας υπολογίσει τις παραπάνω συναρτήσεις μπορούμε να προχωρήσουμε στους αντικειμενικούς σκοπούς της μεθόδου, την ανάλυση και την ταξινόμηση. Με την ανάλυση παρέχεται σημαντικός αριθμός στατιστικών εργαλείων, όπως στατιστικοί έλεγχοι, που είναι αναγκαίοι για την ερμηνεία των αποτελεσμάτων. Στη συνέχεια η διακριτική ανάλυση εφαρμόζεται σαν τεχνική ταξινόμησης, μετά τον προσδιορισμό των μεταβλητών που επιτρέπουν την διάκριση των παρατηρήσεων σε ομάδες σ' έναν ικανοποιητικό βαθμό.
Οι τεχνικές ταξινόμησης μπορούν να βρουν εφαρμογή σε διάφορους τομείς όπως η ταξινόμηση ενός αρχαιολογικού ευρήματος σε μία από δύο ή περισσότερες φυλές ή χρονικές περιόδους, η ταξινόμηση ασθενών ανάλογα με την πιθανότητα επιτυχημένης ανάρρωσης (πλήρης, μερική, καθόλου) από μία ασθένεια, η ταξινόμηση ατόμων με βάση τα φυσιολογικά χαρακτηριστικά τους ή η ταξινόμηση μιας ομάδας εργαζομένων ανάλογα με τις ανάγκες επιμόρφωσής τους.

2.2 Κατανόηση της διακριτικής ανάλυσης

2.2.1 Υπολογιστική Προσέγγιση

Υπολογιστικά, η διακριτική ανάλυση είναι παρόμοια με την ανάλυση διακύμανσης (ANOVA). Για παράδειγμα, ας υποθέσουμε ότι μετράμε το ύψος σε ένα τυχαίο δείγμα αποτελούμενο από 50 ενήλικες και 50 παιδιά. Τα παιδιά, κατά μέσο όρο, δεν είναι τόσο ψηλά όσο οι ενήλικες, και αυτή η διαφορά θα πρέπει να αντικατοπτρίζεται στη διαφορά των μέσων τιμών (για τη μεταβλητή Ύψος). Ως εκ τούτου, η μεταβλητή Ύψος μας επιτρέπει να διακρίνουμε μεταξύ των ενηλίκων και των παιδιών: αν ένα άτομο είναι ψηλό, τότε είναι πιθανότερο να είναι ενήλικας, ενώ αν ένα άτομο δεν είναι ψηλό, τότε είναι πιθανότερο να είναι παιδί.

Μπορούμε να γενικεύσουμε αυτό το σκεπτικό σε ομάδες και τις μεταβλητές που είναι λιγότερο "ασήμαντες". Για παράδειγμα, ας υποθέσουμε ότι έχουμε δύο ομάδες των αποφοίτων λυκείου: εκείνοι που επιλέγουν να συνεχίσουν στο πανεπιστήμιο μετά την αποφοίτηση και όσοι δεν το κάνουν. Θα μπορούσε να μετρηθεί η πρόθεση των μαθητών να συνεχίσει στο πανεπιστήμιο ένα έτος πριν από την αποφοίτηση. Εάν οι μέσοι όροι για τις δύο ομάδες (αυτούς που πραγματικά πήγαν στο πανεπιστήμιο και εκείνων που δεν πήγαν) είναι διαφορετικοί, τότε μπορούμε να πούμε ότι η πρόθεση να σπουδάσει στο πανεπιστήμιο όπως δηλώθηκε ένα έτος πριν από την αποφοίτηση μας επιτρέπει να κάνουμε διάκριση ανάμεσα σε εκείνους που είναι υποψήφιοι να μπουν στο πανεπιστήμιο και σε εκείνους που δεν είναι. Η πληροφορία αυτή θα μπορούσε να χρησιμοποιηθεί από τους συμβούλους επαγγελματικού προσανατολισμού ώστε να παρέχουν την κατάλληλη καθοδήγηση στους αντίστοιχους μαθητές.

Η βασική ιδέα της διακριτικής ανάλυσης είναι να καθορίσει αν ομάδες διαφέρουν όσον αφορά την μέση τιμή μιας μεταβλητής, και στη συνέχεια να χρησιμοποιήσει την μεταβλητή αυτή για την πρόβλεψη σε ποια ομάδα θα ανήκει μία παρατήρηση (πχ μία καινούρια περίπτωση).

2.2.2 Ανάλυση Διακύμανσης

Με αυτή τη μορφή, το πρόβλημα διακριτικής ανάλυσης μπορεί να επαναδιατυπωθεί ως ένα πρόβλημα ανάλυσης διακύμανσης με έναν παράγοντα (ANOVA). Συγκεκριμένα, μπορεί να θέλουμε να μάθουμε αν δύο ή περισσότερες ομάδες είναι διαφορετικές ή όχι από κάθε άλλη σε σχέση με την μέση τιμή μιας συγκεκριμένης μεταβλητής. Εάν οι μέσοι για μια μεταβλητή είναι σημαντικά διαφορετικοί σε διαφορετικές ομάδες, τότε μπορούμε να πούμε ότι αυτή η μεταβλητή συμβάλλει στη διάκριση μεταξύ των δύο ομάδων.

Στην περίπτωση μιας μόνο μεταβλητής, ο τελικός έλεγχος στατιστικής σημαντικότητας για να αποφασίσουμε αν η μεταβλητή συμβάλλει στη διάκριση των ομάδων είναι ο έλεγχος F. Το στατιστικό F υπολογίζεται ως ο λόγος της διακύμανσης των δεδομένων μεταξύ των ομάδων προς τη συγκεντρωτική διακύμανση εντός των ομάδων. Αν η διακύμανση μεταξύ των ομάδων είναι σημαντικά μεγαλύτερη τότε πρέπει να υπάρχουν σημαντικές διαφορές μεταξύ των μέσων,

2.2.3 Πολλαπλές μεταβλητές

Συνήθως περιλαμβάνονται πολλές μεταβλητές σε μία μελέτη προκειμένου να διαπιστωθεί ποιες συμβάλλουν στη διάκριση μεταξύ των ομάδων. Σε αυτή την περίπτωση, έχουμε έναν πίνακα των ολικών διακυμάνσεων και συνδυακυμάνσεων. Ομοίως έχουμε έναν πίνακα των συγκεντρωτικών διακυμάνσεων και συνδυακυμάνσεων εντός των ομάδων. Μπορούμε να συγκρίνουμε αυτούς τους πίνακες μέσω πολυμεταβλητών ελέγχων F έτσι ώστε να

διαπιστώσουμε αν υπάρχουν στατιστικά σημαντικές διαφορές (όσον αφορά όλες τις μεταβλητές) μεταξύ των ομάδων Αυτή η διαδικασία είναι ίδια με την πολυμεταβλητή ανάλυση διακύμανσης (MANOVA). Έτσι κι εδώ, πρώτα εκτελούμε πολυμεταβλητό έλεγχο και αν έχουμε στατιστική σημαντικότητα, προχωράμε για να δούμε ποιες από τις μεταβλητές έχουν στατιστικά σημαντικές διαφορές στους μέσους όλων των ομάδων. Έτσι, ακόμα κι αν οι υπολογισμοί με πολλαπλές μεταβλητές είναι πιο περίπλοκοι, το βασικό επιχείρημα εξακολουθεί να ισχύει, δηλαδή ότι ψάχνουμε τις μεταβλητές που συμβάλλουν στη διάκριση των ομάδων, όπως προκύπτει από τις παρατηρούμενες διαφορές των μέσων.

2.3 Βηματική Διακριτική Ανάλυση

Πιθανώς η πιο κοινή εφαρμογή της διακριτικής ανάλυσης είναι να συμπεριλάβει πολλά μέτρα στην ανάλυση, έτσι ώστε να καθορίσει αυτά που θα διακρίνουν τις ομάδες. Για παράδειγμα ένας σύμβουλος επαγγελματικού προσανατολισμού που ενδιαφέρεται για την πρόβλεψη των επιλογών περεταίρω εκπαίδευσης των αποφοίτων Λυκείου, κατά πάσα πιθανότητα θα συμπεριλάβει όσο το δυνατόν περισσότερα μέτρα της προσωπικότητας, των κινήτρων επιτευγμάτων ή των επιδόσεων στο σχολείο, προκειμένου να διαπιστώσει ποιο ή ποια προσφέρουν την καλύτερη πρόβλεψη.
Με άλλα λόγια θέλουμε να δημιουργήσουμε ένα μοντέλο για το πώς μπορούμε να προβλέψουμε σε ποια ομάδα ανήκει ένα στοιχείο.
Η διακριτικές συναρτήσεις βρίσκονται σταδιακά: Πρώτα βρίσκουμε την D_1, ετσι ώστε να μεγιστοποιείται η διαφορά των τιμών της διακριτικής συνάρτησης μεταξύ των ομάδων. Μετά βρίσκουμε την D_2, που θα είναι ορθογώνια στην D_1 και έτσι ώστε πάλι η διαφορά των τιμών της διακριτικής συνάρτησης μεταξύ των ομάδων να μεγιστοποιείται και ούτω καθεξής.

2.3.1 Προς-τα-εμπρός βηματική ανάλυση

Στη βηματική (σταδιακή) διακριτική ανάλυση, ένα μοντέλο διαχωρισμού κατασκευάζεται βήμα-βήμα. Συγκεκριμένα, σε κάθε βήμα όλες οι μεταβλητές εξετάζονται και αξιολογούνται για να προσδιοριστεί ποια θα συμβάλλουν περισσότερο στη διάκριση των ομάδων. Η μεταβλητή τότε συμπεριλαμβάνεται στο μοντέλο και η διαδικασία ξεκινά και πάλι.
Η σειρά με την οποία οι ανεξάρτητες μεταβλητές εισέρχονται στο υπόδειγμα δεν υποδηλώνει κατ' ανάγκη και τη σειρά σπουδαιότητάς τούς. Αυτό οφείλεται στις συσχετίσεις πού υπάρχουν μεταξύ των μεταβλητών, με συνέπεια να μοιράζονται τη διακριτική ικανότητα, γεγονός πού μπορεί να οδηγήσει ακόμη και στον αποκλεισμό σημαντικών από άποψη αυτοδυναμίας μεταβλητών, όχι όμως σημαντικών όταν αυτές συνδυαστούν με άλλες ως προς τη διακριτική ικανότητα.

2.3.2 Προς-τα-πίσω βηματική ανάλυση

Μπορούμε επίσης να ξεκινήσουμε με όλες τις μεταβλητές να συμπεριλαμβάνονται στο μοντέλο και στη συνέχεια, σε κάθε βήμα, η μεταβλητή που συμβάλλει λιγότερο στην πρόβλεψη των μελών των ομάδων να αποβάλλεται. Έτσι, ως αποτέλεσμα μιας επιτυχούς διακριτικής ανάλυσης, θα κρατήσουμε στο μοντέλο μόνο τις μεταβλητές που έχουν μεγαλύτερη συμμετοχή στη διάκριση των ομάδων.

2.3.3 Κριτήρια για την Επιλογή Μεταβλητών

Στατιστικό Λ του Wilks: Κατά τη διαδικασία της σταδιακής επιλογής των ανεξάρτητων μεταβλητών μπορούν να χρησιμοποιηθούν διάφορα κριτήρια, όπως αυτό της ελαχιστοποίησης του στατιστικού Λ του Wilks. Σύμφωνα με αυτό, σε κάθε βήμα επιλέγουμε τη μεταβλητή που δίνει τη μικρότερη τιμή στο Λ. Η σημαντικότητα της μεταβολής του Λ με την είσοδο μιας μεταβλητής εξετάζεται μέσω ελέγχου F:

$$F_{\mu\varepsilon\tau\alpha\beta o\lambda\acute{\eta}} = \left[\frac{n-g-k}{g-1}\right]\left[\frac{1-\Lambda_{k+1}/\Lambda_k}{\Lambda_{k+1}/\Lambda_k}\right]$$

όπου

- n είναι ο αριθμός των παρατηρήσεων,
- g ο αριθμός των ομάδων,
- k ο αριθμός των ανεξάρτητων μεταβλητών,
- Λ_p η τιμή του Λ του Wilks πριν την είσοδο της μεταβλητής και
- Λ_{p+1} η τιμή του Λ του Wilks μετά την είσοδο της μεταβλητής.

Παρόλο που η προσθήκη επιπλέον μεταβλητών συμβάλλει στη μείωση της τιμής Λ, τα επίπεδα στατιστικής σημαντικότητας δεν ελαττώνονται αναγκαστικά, γιατί αυτό εξαρτάται από την τιμή του Λ και από τον αριθμό των ανεξάρτητων μεταβλητών στο υπόδειγμα.

Κριτήριο V του Rao: Αποτελεί γενικό δείκτη απόστασης μεταξύ των ομάδων. Επιλέγεται η μεταβλητή που συμβάλλει περισσότερο στην αύξηση της τιμής του V όταν προστίθεται στις μεταβλητές που είναι ήδη στην εξίσωση. Το V του Rao, γνωστό και ως ίχνος των Lawley-Hotelling, δίνεται από τη σχέση:

$$V = (n-g)\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}^{*}\sum_{m=1}^{R} n_m\left(\overline{X_{im}} - \overline{X}\right)\left(\overline{X_{jm}} - \overline{X_j}\right)$$

όπου

- k είναι ο αριθμός των μεταβλητών στο μοντέλο,
- g ο αριθμός των ομάδων, n_m το μέγεθος του δείγματος στην ομάδα m,
- X_{im} είναι ο μέσος της μεταβλητής i στην ομάδα m,
- X_i είναι ο μέσος της μεταβλητής i για το σύνολο των ομάδων και
- w_{ij}^{*} στοιχείο του αντίστροφου πίνακα συνδιακυμάνσεων εντός των ομάδων.

Η διαφορά στο V αυξάνεται με την αύξηση των διαφορών μεταξύ των κεντροειδών των ομάδων.

Κριτήριο D^2 της απόστασης Mahalanobis: Αποτελεί ένα γενικευμένο μέτρο της απόστασης μεταξύ των ομάδων. Η απόσταση μεταξύ δύο ομάδων α και b δίνεται από τη σχέση:

$$V = (n-g)\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}^{*}\left(\overline{X_{ia}} - \overline{X_{ib}}\right)\left(\overline{X_{ja}} - \overline{X_{jb}}\right)$$

όπου:

- k είναι ο αριθμός των μεταβλητών στο μοντέλο,
- g ο αριθμός των ομάδων, n_m το μέγεθος του δείγματος στην ομάδα m,
- $X_{i\alpha}$ και X_{ib} είναι ο μέσος της μεταβλητής i στις ομάδες α και b,
- $X_{j\alpha}$ και $X_{j\beta}$ είναι ο μέσος της μεταβλητής i στις ομάδες α και b και

- w_{ij}^{*} στοιχείο του αντίστροφου πίνακα συνδιακυμάνσεων εντός των ομάδων.

Σύμφωνα με αυτό το κριτήριο, αφού υπολογιστούν οι αποστάσεις Mahalanobis όλων των ζευγών ομάδων, εισέρχεται η μεταβλητή που έχει τη μεγαλύτερη τιμή D^2 για τις δύο εγγύτερες ομάδες.

Κριτήριο μεγιστοποίησης μικρότερης F τιμής: Αφορά τον έλεγχο της υπόθεσης ότι οι δύο μέσοι είναι ίσοι. Βασίζεται στην απόσταση Mahalanobis και δίνεται από τη σχέση:

$$F = \frac{(n-1-k)n_1 n_2}{k(n-2)(n_1+n_2)} \cdot D_{ab}^2$$

Επιλέγεται η μεταβλητή με την μεγαλύτερη τιμή F. Το κριτήριο αυτό ταυτίζεται με το D^2 μόνο στην περίπτωση διάκρισης σε ομάδες ίδιου μεγέθους.

Έλεγχος F για εισαγωγή ή αφαίρεση μεταβλητής: Η σταδιακή διαδικασία καθοδηγείται από τις τιμές του ελέγχου F. Η τιμή F για μία μεταβλητή υποδεικνύει τη στατιστική της σημαντικότητα στη διάκριση των ομάδων, δηλαδή είναι ένα μέτρο του βαθμού στον οποίο μια μεταβλητή συνεισφέρει στην πρόβλεψη των μελών της ομάδας.

Μία κοινή παρερμηνεία των αποτελεσμάτων της βηματικής διακριτικής ανάλυσης είναι να λάβει τα επίπεδα στατιστικής σημαντικότητας εκ πρώτης όψεως. Από φύσεως, οι διαδικασίες επωφελούνται από την ευκαιρία επειδή επιλέγουν τις μεταβλητές που θα εισαχθούν στο μοντέλο έτσι ώστε να αποφέρουν μέγιστη διάκριση. Έτσι όταν χρησιμοποιούμε τη βηματική προσέγγιση πρέπει να έχουμε υπόψη ότι τα επίπεδα σημαντικότητας δεν αντανακλούν το πραγματικό ποσοστό σφάλματος, δηλαδή υπάρχει πιθανότητα να απορρίψουμε εσφαλμένα την H_0 (μηδενική υπόθεση: δεν υπάρχει διάκριση μεταξύ των ομάδων).

2.4 Ερμηνεία διακριτικής συνάρτησης δύο ομάδων

Στην περίπτωση που έχουμε δύο ομάδες η διακριτική ανάλυση είναι ανάλογη με την πολλαπλή παλινδρόμηση. Η διακριτική ανάλυση δύο ομάδων είναι γνωστή και ως γραμμική διακριτική ανάλυση Fisher. Αν κωδικοποιήσουμε τις δύο ομάδες ως 1 και 2 και χρησιμοποιήσουμε αυτή τη μεταβλητή ως εξαρτημένη μεταβλητή σε μια ανάλυση πολλαπλής παλινδρόμησης, θα προκύψουν αποτελέσματα ανάλογα με εκείνα της διακριτικής ανάλυσης. Γενικά, στην περίπτωση των δύο ομάδων έχουμε μια γραμμική εξίσωση της μορφής:

$$D = a + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_n \cdot X_n$$

όπου α είναι μια σταθερά και b_i οι διακριτικοί συντελεστές. Η ερμηνεία των αποτελεσμάτων είναι απλή και ακολουθεί τη λογική της πολλαπλής παλινδρόμησης: Οι μεταβλητές με τους μεγαλύτερους συντελεστές είναι εκείνες που συμβάλλουν περισσότερο στην πρόβλεψη των μελών των ομάδων. Με άλλα λόγια οι τιμές των συντελεστών εκφράζουν τη συνεισφορά των αντίστοιχων μεταβλητών στην διακριτική ικανότητα της D.

2.5 Διακριτικές συναρτήσεις για Πολλαπλές Ομάδες

Όταν υπάρχουν περισσότερες από δύο ομάδες, τότε μπορούμε να εκτιμήσουμε περισσότερες από μία διακριτικές συναρτήσεις. Αν για παράδειγμα, έχουμε τέσσερις ομάδες θα μπορούσαμε να εκτιμήσουμε μία συνάρτηση που θα διακρίνει την ομάδα 1 από τις 2, 3 και 4, μία δεύτερη συνάρτηση που θα διακρίνει την ομάδα 2 από τις 3 και 4 και μία τρίτη που θα διακρίνει την ομάδα 2 από την 3. Οι συντελεστές σε αυτές τις συναρτήσεις ερμηνεύονται όπως παραπάνω. Θα μπορούσαμε να έχουμε μόνο δύο συναρτήσεις και να αγνοούσαμε την τρίτη αν δεν μας έδινε παραπάνω πληροφορία, αφού δεν θα ήταν θεωρητικής ή πρακτικής σημασίας.
Κατά την εκτέλεση μιας πολλαπλής διακριτικής ανάλυσης, δεν πρέπει να καθορίσουμε πως θα συνδυαστούν οι ομάδες έτσι ώστε να σχηματιστούν διαφορετικές διακριτικές συναρτήσεις αλλά μάλλον θα μπορεί να προσδιοριστεί αυτόματα κάποιος βέλτιστος συνδυασμός μεταβλητών έτσι ώστε η πρώτη συνάρτηση να παρέχει τη μεγαλύτερη διάκριση μεταξύ των ομάδων, η δεύτερη ομάδα την δεύτερη μεγαλύτερη και ούτω καθεξής. Οι διακριτικές συναρτήσεις παρέχουν δηλαδή φθίνουσα διάκριση. Επιπλέον οι συναρτήσεις θα είναι ανεξάρτητες ή ορθογώνιες, δηλαδή η συνεισφορές τους στη διάκριση των ομάδων δεν θα επικαλύπτονται. Υπολογιστικά, εκτελούμε μία κανονική ανάλυση συσχέτισης, η οποία θα καθορίσει τις διαδοχικές συναρτήσεις. Ο μέγιστος αριθμός των συναρτήσεων είναι ίσος με τον αριθμό των ομάδων ελαττωμένο κατά μία μονάδα ή με τον αριθμό των μεταβλητών της ανάλυσης, όποιο είναι μικρότερο. Δηλαδή:

$$\#(\sigma\upsilon\nu\alpha\rho\acute{\eta}\sigma\varepsilon\iota\varsigma) = \min(g-1, k)$$

όπου g ο αριθμός των ομάδων και k ο αριθμός των μεταβλητών. Στη διακριτική ανάλυση κάθε ομάδα θεωρείται ως σημείο και κάθε διακριτική συνάρτηση χαρακτηρίζεται από μια διάσταση που δίνει την θέση της ομάδας σε σχέση με τις υπόλοιπες.
Η συσχέτιση των διακριτικών συναρτήσεων μπορεί να γίνει μέσω των χαρακτηριστικών ριζών τους, των συντελεστών κανονικοποιημένης συσχέτισης και του στατιστικού ελέγχου Λ του Wilks. Η διαδικασία μπορεί να σταματήσει όποτε διαπιστωθεί ότι η επιπλέον πληροφορία είναι ασήμαντη. Επειδή δεν υπάρχει κανόνας που να ορίζει που θα σταματήσει η διαδικασία δημιουργίας των διακριτικών συναρτήσεων, αποφασίζει ο ερευνητής.

Εκτίμηση συντελεστών διακριτικής συνάρτησης
Παρόλο που τα περιγραφικά στατιστικά μέτρα και οι απλοί έλεγχοι μας δίνουν κάποιες βασικές πληροφορίες για την κατανομή των τιμών των ανεξάρτητων μεταβλητών εντός των ομάδων της εξαρτημένης μεταβλητής και μας βοηθούν να αναγνωρίσουμε χαρακτηριστικά μεταξύ των ομάδων, είναι επιβεβλημένη η από κοινού ανάλυση των μεταβλητών κι όχι κάθε μιας ξεχωριστά. Με αυτόν τον τρόπο καταφέρνουμε να συμπεριλάβουμε στην ανάλυση σημαντικές πληροφορίες σχετικά με τις ενδεχόμενες σχέσεις μεταξύ των ανεξάρτητων μεταβλητών.

Ερμηνεία των διακριτικών συναρτήσεων
Όπως και πριν θα έχουμε τους συντελεστές b για κάθε μεταβλητή σε κάθε διακριτική συνάρτηση, και ερμηνεύονται ως συνήθως: όσο μεγαλύτερος είναι ο τυποποιημένος συντελεστής, τόσο μεγαλύτερη είναι η συνεισφορά της αντίστοιχης μεταβλητής στη διάκριση μεταξύ των μεταβλητών. Ωστόσο, αυτοί οι συντελεστές δεν μας λένε μεταξύ ποιων από τις ομάδες κάνουν διάκριση οι αντίστοιχες συναρτήσεις. Αυτό μπορούμε να το

προσδιορίσουμε κοιτάζοντας τους μέσους για τις συναρτήσεις σε όλες τις ομάδες. Μπορούμε επίσης να έχουμε μια εικόνα διάκρισης των ομάδων σχεδιάζοντας το γράφημα των τιμών για τις διακριτικές συναρτήσεις.

Πίνακας παραγοντικής δομής
Ένας άλλος τρόπος να καθοριστούν ποιες μεταβλητές ορίζουν μία συγκεκριμένη διακριτική συνάρτηση είναι να εξετάσουμε την παραγοντική δομή. Οι συντελεστές της παραγοντικής δομής είναι οι συσχετίσεις μεταξύ των μεταβλητών στο μοντέλο και των διακριτικών συναρτήσεων.

Σημασία των διακριτικών συναρτήσεων
Μπορούμε να δοκιμάσουμε τον αριθμό των συναρτήσεων που προσθέτουν σημαντικά στην διάκριση μεταξύ ομάδων. Μόνο εκείνες που είναι στατιστικά σημαντικές πρέπει να χρησιμοποιηθούν για την ερμηνεία, ενώ οι μη-σημαντικές συναρτήσεις θα πρέπει να αγνοηθούν.
Η σημασία της διακριτικής συνάρτησης D_j μπορεί να κριθεί και από το συντελεστή

$$R_j = \sqrt{\frac{\lambda_j}{1+\lambda_j}}$$

της κανονικής συσχέτισης της συνάρτησης αυτής με τη μεταβλητή καθορισμού των ομάδων. Η ποσότητα R_j^2 εκφράζει το ποσοστό της διακύμανσης που υπάρχει στη D_j και που ερμηνεύεται από τις ομάδες. Το λ_j είναι η ιδιοτιμή που αντιστοιχεί στη διακριτική συνάρτηση D_j και είναι ιδιοτιμή του πίνακα $W^{-1}B$, όπου W είναι ο πίνακας διακυμάνσεων-συνδιακυμάνσεων εντός των ομάδων και B είναι ο πίνακας διακυμάνσεων-συνδιακυμάνσεων μεταξύ των ομάδων. Οι ιδιοτιμές λ_j έχουν τοποθετηθεί σε φθίνουσα σειρά. Η ιδιοτιμή αποτελεί δείκτη της αποτελεσματικότητας της διακριτικής συνάρτησης από τη μία και της επιλογής του αριθμού των διακριτικών συναρτήσεων από την άλλη. Μεγαλύτερη τιμή της ιδιοτιμής σημαίνει καλύτερη διακριτική συνάρτηση.
Το άθροισμα όλων των ιδιοτιμών αποτελεί μέτρο εκτίμησης της ολικής διακύμανσης των διακριτικών μεταβλητών ενώ το πηλίκο

$$\frac{\lambda_j}{\sum_{i=1}^{k}\lambda_j}$$

εκφράζει το ποσοστό της ολικής διακύμανσης που ερμηνεύεται από την D_j.

2.6 Υποθέσεις της Διακριτικής Ανάλυσης

2.6.1 Κανονική Κατανομή

Υποτίθεται ότι τα δεδομένα αντιπροσωπεύουν ένα δείγμα μιας πολυμεταβλητής κανονικής κατανομής. Μπορούμε να εξετάσουμε κατά πόσο οι μεταβλητές κατανέμονται κανονικά ή όχι με ιστογράμματα ή κατανομές συχνοτήτων. Με την τεχνική του «φυλλογραφήματος» κάθε μεταβλητής μπορούμε να ελέγξουμε αν η κατανομή είναι συμμετρική ή αν υπάρχουν ακραίες τιμές. Αν κάποια παρουσιάζει ασυμμετρία έχουμε μια ένδειξη για τη μη κανονικότητα της πολλαπλής κατανομής.
Μια άλλη τεχνική είναι η εξής: αφού διατάξουμε τις τιμές τις μεταβλητής σε αύξουσα σειρά, να τις συγκρίνουμε με τις θεωρητικές της κανονικής κατανομής. Αν τα σημεία των ζευγών βρίσκονται σε ευθεία τότε η κατανομή της μεταβλητής θα είναι κανονική. Επίσης

μέσω του διαγράμματος των καταλοίπων απόκλισης από την κανονικότητα (διαφορά μεταξύ παρατηρούμενων και θεωρητικών τιμών) μπορούμε να ελέγξουμε τη γραμμικότητα του διαγράμματος της κανονικής κατανομής με βάση τις πιθανότητες. Για να είναι κανονική μια ανεξάρτητη μεταβλητή πρέπει τα σημεία του διαγράμματος αυτού να βρίσκονται κοντά σε μια ευθεία γραμμή, η οποία προσεγγίζει την μηδενική τιμή.
Ωστόσο, οι παραβιάσεις της υπόθεσης κανονικότητας δεν είναι συνήθως «μοιραίες», με την έννοια ότι οι έλεγχοι σημαντικότητας που προκύπτουν εξακολουθούν να είναι αξιόπιστες. Εκτός από γραφήματα μπορούν να χρησιμοποιηθούν συγκεκριμένοι έλεγχοι κανονικότητας. Σε περίπτωση μη κανονικότητας μπορούμε αν θέλουμε να μετασχηματίσουμε τις τιμές.

2.6.2 Ομοιογένεια διακυμάνσεων – συνδυακυμάνσεων

Υποτίθεται ότι οι πίνακες διακυμάνσεων – συνδυακυμάνσεων είναι ομογενείς στις ομάδες. Μικρές αποκλίσεις δεν είναι τόσο σημαντικές, ωστόσο, πριν από την αποδοχή των τελικών αποτελεσμάτων μιας σημαντικής μελέτης καλό θα είναι να επανεξετάσουμε τις διακυμάνσεις εντός των ομάδων και τους πίνακες συσχέτισης. Συγκεκριμένα ένας πίνακας σκέδασης μπορεί να είναι πολύ χρήσιμος για το σκοπό αυτό. Σε περίπτωση αμφιβολιών, μπορούμε να δοκιμάσουμε να εκτελέσουμε εκ νέου την ανάλυση εξαιρώντας μία ή δύο ομάδες που παρουσιάζουν λιγότερο ενδιαφέρον. Αν τα συνολικά αποτελέσματα διατηρούνται, τότε μάλλον δεν υπάρχει πρόβλημα. Μπορούμε επίσης να χρησιμοποιήσουμε διάφορους ελέγχους για να εξετάσουμε κατά πόσο παραβιάζεται αυτή η υπόθεση στα δεδομένα.
Ένας έλεγχος είναι με το στατιστικό Μ του Box, που κάνει χρήση των οριζουσών των ατομικών πινάκων διακυμάνσεων – συνδιακυμάνσεων και του συνδυασμένου κοινού πίνακα ενώ η σημαντικότητα του Μ στηρίζεται στα στατιστικά F ή χ^2 . Η μηδενική υπόθεση (ότι οι πίνακες είναι ίσοι) απορρίπτεται αν η στατιστική σημαντικότητα του Μ είναι μικρή. Αυτό δεν αποτελεί απόδειξη αλλά μάλλον ένδειξη για την διαφορά των πινάκων, αφού αν το δείγμα είναι μεγάλο η στατιστική σημαντικότητα μπορεί να είναι μικρή ακόμα και αν οι πίνακες δεν διαφέρουν και τόσο. Ακόμη αυτός ο έλεγχος επηρεάζεται πολύ από τις αποκλίσεις του συνόλου των μεταβλητών από την κανονικότητα. Δηλαδή σε περίπτωση που δεν έχουμε κανονικότητα ο έλεγχος μπορεί να θεωρήσει άνισους τους πίνακες.

2.6.3 Συσχετίσεις μεταξύ μέσων τιμών και διασπορών

Η συσχέτιση των μέσων των μεταβλητών των ομάδων με τις διασπορές (ή τις τυπικές αποκλίσεις) αποτελούν τη μεγαλύτερη απειλή για την εγκυρότητα των ελέγχων σημαντικότητας. Διαισθητικά, όταν υπάρχει μεγάλη μεταβλητότητα σε μια ομάδα με ιδιαίτερα υψηλές μέσες τιμές σε κάποιες μεταβλητές, τότε αυτές οι μέσες τιμές δεν είναι αξιόπιστες. Ωστόσο, οι συνολικοί έλεγχοι σημαντικότητας βασίζονται σε συγκεντρωτικές διακυμάνσεις, δηλαδή στη μέση διακύμανση όλων των ομάδων. Έτσι, οι έλεγχοι σημαντικότητας των σχετικά μεγαλύτερων μέσων (με τις μεγάλες διακυμάνσεις) θα βασίζονταν στις σχετικά μικρότερες συγκεντρωτικές διακυμάνσεις, οδηγώντας εσφαλμένα σε στατιστική σημαντικότητα. Στην πράξη, αυτό μπορεί να προκύψει αν μία ομάδα της μελέτης περιλαμβάνει μερικές υπερβολικά ακραίες τιμές, που έχουν μεγάλο αντίκτυπο στον μέσο και αυξάνουν τη μεταβλητότητα. Για να αποφευχθεί αυτό το πρόβλημα, πρέπει να ελέγχουμε τα περιγραφικά στατιστικά μέτρα, δηλαδή τους μέσους και τις τυπικές αποκλίσεις για μια τέτοια συσχέτιση.
Ένας έλεγχος της απουσίας συσχέτισης είναι ο έλεγχος σφαιρικότητας του Bartlett. Ο έλεγχος αυτός, με τη χρήση του χ^2, εξετάζει την υπόθεση ότι ο πίνακας συσχετίσεων είναι

ταυτοτικός. Απόρριψη της υπόθεσης (α>0.05) σημαίνει ότι οι ανεξάρτητες μεταβλητές δεν συσχετίζονται και συνεπώς το μοντέλο της διακριτικής ανάλυσης είναι καλά προσαρμοσμένο στα δεδομένα.

2.6.4 Βαθμός ανοχής

Κάνοντας χρήση του βαθμού ανοχής που δίνεται από τη σχέση 1- R_i^2, όπου R_i^2 είναι ο συντελεστής πολλαπλού προσδιορισμού μιας ανεξάρτητης μεταβλητής i με τις υπόλοιπες ανεξάρτητες, δεν εισάγουμε στην ανάλυση μεταβλητές με μικρές τιμές ανοχής, οι οποίες δηλώνουν ότι η i μεταβλητή αποτελεί γραμμικό συνδυασμό των άλλων. Επίσης αν η είσοδος μιας μεταβλητής ανατρέπει την ανοχή μιας ήδη υπάρχουσας μεταβλητής τότε η μεταβλητή αυτή δεν εισέρχεται στην ανάλυση. Γενικά όταν μια μεταβλητή είναι σχεδόν τελείως περιττή τότε ο βαθμός ανοχής θα προσεγγίζει το 0.

2.7 Ταξινόμηση

Η διακριτική ανάλυση, εκτός από εργαλείο στατιστικής ανάλυσης και ερμηνείας αποτελεσμάτων, είναι εξαιρετική τεχνική ταξινόμησης παρατηρήσεων, δηλαδή ταυτοποίησης μιας παρατήρησης (να ανήκει σε μία ή άλλη ομάδα-ταξινομική κατηγορία). Άλλη χρήση της διακριτικής ταξινόμησης αφορά τον έλεγχο της ακρίβειας των διακριτικών συναρτήσεων πού προκύπτουν. Η ταξινόμηση επιτυγχάνεται μέσω της χρήσης συναρτήσεων ταξινόμησης, μίας για κάθε ομάδα.

2.7.1 A priori και a posteriori προβλέψεις

Αν εκτιμούμε, βασιζόμενοι σε ένα σύνολο δεδομένων, τις διακριτικές συναρτήσεις που κάνουν τον καλύτερο διαχωρισμό των ομάδων και στη συνέχεια χρησιμοποιήσουμε τα ίδια δεδομένα για να αξιολογήσουμε πόσο ακριβής είναι η πρόβλεψή μας, τότε εκμεταλλευόμαστε τις πιθανότητες. Γενικά, πάντα θα παίρνουμε μία χειρότερη ταξινόμηση όταν προβλέπουμε περιπτώσεις που δεν χρησιμοποιήσαμε για την εκτίμηση της διακριτικής συνάρτησης. Με άλλα λόγια οι εκ των υστέρων προβλέψεις (η πρόβλεψη για κάτι που ξέρουμε ότι έχει συμβεί) είναι πάντα καλύτερες από τις εκ των προτέρων. Ως εκ τούτου, δεν πρέπει να βασίζουμε την εμπιστοσύνη μας όσον αφορά την σωστή ταξινόμηση των μελλοντικών παρατηρήσεων στο ίδιο σύνολο δεδομένων από το οποίο προήλθαν οι διακριτικές συναρτήσεις, αλλά μάλλον είναι απαραίτητο να συλλέξουμε νέα στοιχεία για να επικυρώσουμε τη χρησιμότητα των διακριτικών συναρτήσεων.

2.7.2 Συναρτήσεις Ταξινόμησης

Οι συναρτήσεις ταξινόμησης δεν πρέπει να συγχέονται με τις διακριτικές συναρτήσεις. Οι συναρτήσεις ταξινόμησης μπορούν να χρησιμοποιηθούν για να καθορίσουν σε ποια ομάδα είναι πιθανότερο να ανήκει μία παρατήρηση. Κάθε συνάρτηση μας επιτρέπει να υπολογίσουμε βαθμούς ταξινόμησης για κάθε παρατήρηση κάθε ομάδας, εφαρμόζοντας τον τύπο:

$$C_i = c_{i_0} + c_{i_1} V_1 + c_{i_2} V_2 + \ldots + c_{i_p} V_p$$

όπου C_i είναι ο βαθμός ταξινόμησης για την ομάδα i, c_{i_0} η σταθερά, c_{i_j} οι συντελεστές ταξινόμησης και V είναι οι αρχικές τιμές των διακριτικών μεταβλητών. Για καθεμιά ομάδα υπάρχει και μια ξεχωριστή εξίσωση. Υπάρχουν αρκετοί τρόποι υπολογισμού συναρτήσεων

ταξινόμησης, μερικοί από τους οποίους βασίζονται στις αρχικές τιμές των διακριτικών βαθμών. Συχνά γίνεται χρήση του κανόνα των πιθανοτήτων του Bayes για την εκτίμηση εκ των προτέρων της συμμετοχής δεδομένης παρατήρησης σε κάποια ομάδα.
Τέλος, μια άλλη ομάδα συντελεστών, οι συντελεστές γραμμικής διακριτικής συνάρτησης του Fisher, γνωστοί και ως ταξινομικοί συντελεστές για καθεμιά ομάδα ξεχωριστά, μπορούν να χρησιμοποιηθούν άμεσα για σκοπούς ταξινόμησης. Για κάθε ομάδα υπολογίζονται συντελεστές ταξινόμησης και κάθε παρατήρηση εντάσσεται στην ομάδα στην οποία έχει τη μεγαλύτερη τιμή. Τα αποτελέσματα της ταξινόμησης είναι ταυτόσημα με αυτά που προκύπτουν με τη χρησιμοποίηση των συντελεστών των κανονικοποιημένων διακριτικών συναρτήσεων.
Αν θεωρήσουμε ότι έχουμε πολυμεταβλητή κανονική κατανομή, οι βαθμοί ταξινόμησης μπορούν να εκφραστούν σε πιθανότητες ως προς το βαθμό που συμμετέχουν σε μία ομάδα. Ο κανόνας, με βάση τον οποίο μια παρατήρηση εντάσσεται σε μια ομάδα με τη μεγαλύτερη βαθμολογία, είναι ισοδύναμος με την ένταξη της παρατήρησης σε μια ομάδα όπου έχει τη μεγαλύτερη πιθανότητα συμμετοχής.

2.7.3 Ο κανόνας Bayes

Σύμφωνα με τον κανόνα αυτόν, η πιθανότητα που έχει μια παρατήρηση με συγκεκριμένο διακριτικό βαθμό να ενταχθεί στην ομάδα i εκτιμάται με το μαθηματικό τύπο:

$$P(G_i / D) = \frac{P(D / G_i)P(G_i)}{\sum_{i=1}^{R} P(D / G_i)P(G_i)}$$

όπου

- D είναι ο διακριτικός βαθμός
- $P(G_i)$ είναι η εκ των προτέρων πιθανότητα - εκτιμητής πιθανοφάνειας ότι μια παρατήρηση εντάσσεται σε συγκεκριμένη ομάδα, όταν δεν είναι διαθέσιμη καμία σχετική πληροφορία. Συνήθως εκφράζεται ως ποσοστό του αριθμού των παρατηρήσεων που ανήκουν σε δεδομένη ομάδα στο σύνολο των παρατηρήσεων ή, όταν οι ομάδες είναι ισάριθμες ή δεν είναι γνωστή η πιθανότητα συμμετοχής σε κάποια ομάδα, με βάση τη μαθηματική πιθανότητα του Laplace (περί ίσων πιθανοτήτων σε όλες τις ομάδες).
- $P(D/G_i)$ είναι η δεσμευμένη πιθανότητα απόκτησης διακριτικού βαθμού D, δεδομένου ότι η παρατήρηση ανήκει σε συγκεκριμένη ομάδα και
- $P(G_i/D)$ είναι η εκ των υστέρων πιθανότητα, η τιμή της οποίας προκύπτει με την εφαρμογή του παραπάνω μαθηματικού τύπου.

Άρα για να ταξινομηθεί μία παρατήρηση με βάση τον διακριτικό βαθμό D εντάσσεται στην ομάδα στην οποία η εκ των υστέρων πιθανότητα έχει τη μεγαλύτερη τιμή.

2.7.4 Ταξινόμηση παρατηρήσεων

Αφού υπολογίσουμε τους βαθμούς ταξινόμησης για μία παρατήρηση, είναι εύκολο να αποφασίσουμε πώς να την ταξινομήσουμε. Γενικά τοποθετούμε την παρατήρηση στην ομάδα για την οποία έχει προκύψει ο μεγαλύτερος βαθμός ταξινόμησης.

2.7.5 Απόσταση Mahalanobis και ταξινόμηση

Η απόσταση Mahalanobis είναι ένα μέτρο απόστασης μεταξύ δύο σημείων στο χώρο που ορίζεται από δύο ή περισσότερες μεταβλητές. Στην περίπτωση των δύο ασυσχέτιστων μεταβλητών μπορούμε να σχεδιάσουμε τα σημεία (παρατηρήσεις) σε ένα διδιάστατο γράφημα και η απόσταση Mahalanobis ταυτίζεται με την Ευκλείδεια απόσταση. Στην περίπτωση περισσότερων από τρεις μεταβλητών δεν μπορούμε να σχεδιάσουμε την απόσταση σε γράφημα και η Ευκλείδεια απόσταση δεν αποτελεί κατάλληλο μέτρο ενώ η απόσταση Mahalanobis εξηγεί ικανοποιητικά τους συσχετισμούς.

Για κάθε ομάδα του δείγματος, μπορούμε να καθορίσουμε τη θέση του σημείου που αντιπροσωπεύει τη μέση τιμή όλων των μεταβλητών στον πολυδιάστατο χώρο που ορίζεται από τις διακριτικές μεταβλητές. Αυτά τα σημεία ονομάζονται κέντρα βάρους των ομάδων ή κεντροειδή. Για κάθε παρατήρηση μπορούμε να υπολογίσουμε την απόσταση Mahalanobis της παρατήρησης από κάθε κεντροειδές. Η παρατήρηση ταξινομείται έτσι ώστε να ανήκει στην ομάδα που είναι πιο κοντά, δηλαδή εκεί που ελαχιστοποιείται η απόσταση Mahalanobis.

2.7.6 Μεταγενέστερες πιθανότητες ταξινόμησης

Χρησιμοποιώντας την απόσταση Mahalanobis για την ταξινόμηση, μπορούμε να αντλήσουμε πιθανότητες. Η πιθανότητα να ανήκει μία παρατήρηση σε μία συγκεκριμένη ομάδα είναι βασικά ανάλογη με την απόσταση Mahalanobis από το κέντρο της ομάδας (δεν είναι ακριβώς ανάλογη επειδή υποθέτουμε πολυμεταβλητή κανονική κατανομή γύρω από κάθε κέντρο). Επειδή υπολογίζουμε τη θέση κάθε παρατήρησης από την προηγούμενη γνώση μας για τις τιμές αυτής τις παρατήρησης στις μεταβλητές του μοντέλου, αυτές οι πιθανότητες καλούνται μεταγενέστερες πιθανότητες. Συνοψίζοντας, η μεταγενέστερη πιθανότητα είναι η πιθανότητα, με βάση τη γνώση μας για τις τιμές των άλλων μεταβλητών, να ανήκει η αντίστοιχη παρατήρηση σε μια συγκεκριμένη ομάδα.

2.7.7 A priori πιθανότητες ταξινόμησης

Μερικές φορές, γνωρίζουμε εκ των προτέρων ότι υπάρχουν περισσότερες παρατηρήσεις σε μία ομάδα σε σχέση με τις υπόλοιπες. Έτσι η a priori πιθανότητα να ανήκει μία παρατήρηση σε αυτή την ομάδα είναι μεγαλύτερη. Για παράδειγμα, αν γνωρίζουμε ότι 60% των παρατηρήσεων ανήκουν στην ομάδα 1, 25% στην ομάδα 2 και 15% στην ομάδα 3, τότε θα προσαρμόσουμε αναλόγως την πρόβλεψή μας: εκ των προτέρων είναι πιο πιθανό μία παρατήρηση να ανήκει στην ομάδα 1 από ότι στις δύο άλλες. Μπορούμε να καθορίσουμε διαφορετικές a priori πιθανότητες, που αργότερα θα χρησιμοποιηθούν για την ανάλογη προσαρμογή της ταξινόμησης των παρατηρήσεων.

Στην πράξη, πρέπει να αναρωτηθούμε αν ο άνισος αριθμός των παρατηρήσεων σε διαφορετικές ομάδες στο δείγμα είναι μια αντανάκλαση της πραγματικής κατανομής του πληθυσμού ή αν είναι μόνο το (τυχαίο) αποτέλεσμα της διαδικασίας δειγματοληψίας. Στην πρώτη περίπτωση, θα ορίσουμε τις a priori πιθανότητες να είναι ανάλογες με τα μεγέθη των ομάδων στο δείγμα, ενώ στη δεύτερη θα καθορίσουμε τις a priori πιθανότητες ώστε να είναι ίσες σε κάθε ομάδα. Η προδιαγραφή διαφορετικών a priori πιθανοτήτων μπορεί να επηρεάσει σε μεγάλο βαθμό την ακρίβεια της πρόβλεψης.

2.7.8 Περίληψη της πρόβλεψης

Ένας κοινός τρόπος για να καθορίσουμε πόσο καλά προβλέπουν τη συμμετοχή σε μία ομάδα οι τρέχουσες συναρτήσεις ταξινόμησης είναι ο πίνακας ταξινόμησης. Ο πίνακας αυτός δείχνει τον αριθμό των παρατηρήσεων που έχουν ταξινομηθεί σωστά, καθώς κι εκείνες που ταξινομήθηκαν εσφαλμένα.

2.7.9 Εκτίμηση του βαθμού εσφαλμένης ταξινόμησης

Έλεγχος της ακρίβειας των διακριτικών συναρτήσεων μπορεί να γίνει όταν με την ταξινόμηση των αρχικών παρατηρήσεων είμαστε σε θέση να βεβαιωθούμε πόσες από αυτές ταξινομούνται σωστά και με βάση τις ανεξάρτητες μεταβλητές πού έχουν χρησιμοποιηθεί.

Ένα υπόδειγμα συνήθως είναι προσαρμοσμένο καλύτερα στο δείγμα των παρατηρήσεων από το οποίο έχει προέλθει απ' ότι θα ήταν σε ένα άλλο δείγμα του ίδιου πληθυσμού. Συνεπώς, το ποσοστό των ορθά ταξινομημένων μέσω της διακριτικής ανάλυσης παρατηρήσεων αποτελεί έναν καλό εκτιμητή του γεγονότος αυτού στον πληθυσμό των παρατηρήσεων,

Υπάρχουν αρκετοί, τρόποι για την απόκτηση του καλύτερου εκτιμητή του βαθμού εσφαλμένης ταξινόμησης. Εάν το δείγμα είναι αρκετά μεγάλο ώστε να χωριστεί σε δύο τμήματα, το ένα τμήμα μπορεί να χρησιμοποιηθεί για τον υπολογισμό της διακριτικής συνάρτησης και το άλλο για τον έλεγχό της. Εφόσον μάλιστα δεν χρησιμοποιούνται οι ίδιες παρατηρήσεις για την εκτίμηση και τον έλεγχο της συνάρτησης, το παρατηρούμενο λάθος στο τμήμα πού χρησιμοποιείται για τον έλεγχο πρέπει να αντανακλά καλύτερα την αποτελεσματικότητα της συνάρτησης. Όμως, η μέθοδος αυτή απαιτεί μεγάλα μεγέθη δειγμάτων και δεν κάνει σωστή χρήση όλης της διαθέσιμης πληροφορίας.

Μια άλλη τεχνική απόκτησης ενός βελτιωμένου εκτιμητή του βαθμού εσφαλμένης ταξινόμησης είναι η τεχνική της «διασταυρούμενης αξιολόγησης» για τον περιορισμό της «αισιόδοξης εκτίμησης» της επιτυχίας ταξινόμησης. Κατά την τεχνική αυτήν, κάθε παρατήρηση ταξινομείται σε μια ομάδα σύμφωνα με τη γραμμική συνάρτηση ταξινόμησης επί του συνόλου των παρατηρήσεων πλην της παρατήρησης πού είναι υπό ταξινόμηση. Έτσι, αφού η παρατήρηση πού ταξινομείται δεν περιλαμβάνεται στον υπολογισμό της συνάρτησης, ο βαθμός της παρατηρούμενης εσφαλμένης ταξινόμησης είναι ο λιγότερο επηρεασμένος εκτιμητής της σωστής ταξινόμησης. Θα πρέπει να σημειωθεί, ότι εάν στόχος είναι η δημιουργία υποδείγματος για ταξινόμηση μελλοντικών παρατηρήσεων, η χρησιμοποίηση εκτιμητών με βάση την τεχνική της διασταυρούμενης αξιολόγησης είναι σε μεγάλο βαθμό υπεραισιόδοξη.

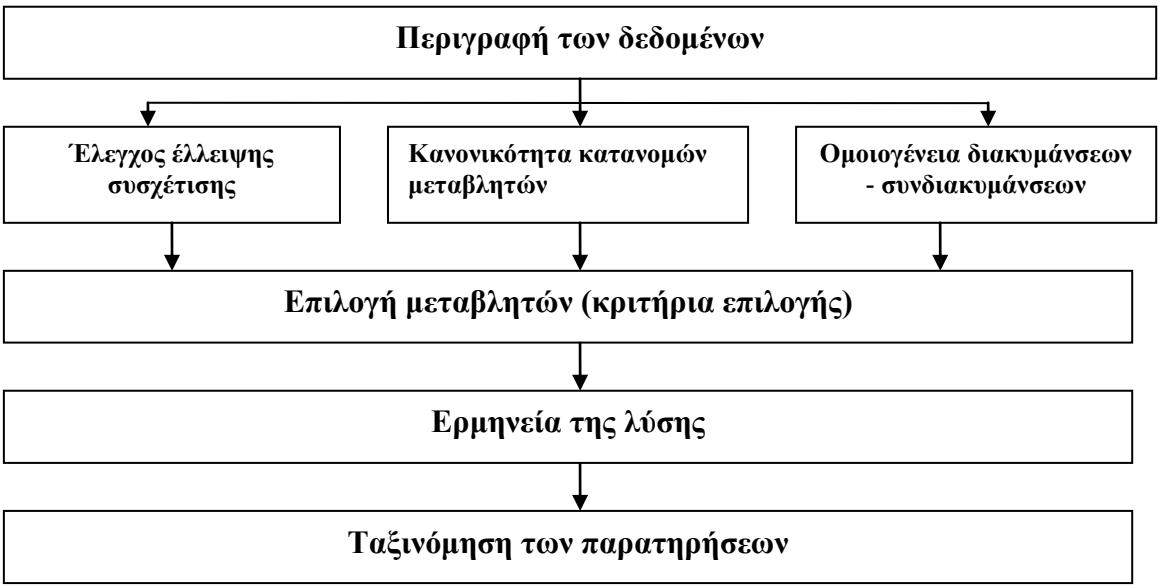
Όταν μία από τις δύο ομάδες είναι πολύ μικρότερου μεγέθους από την άλλη, μπορεί να επιτευχθεί υψηλός βαθμός ορθής ταξινόμησης ακόμη κι όταν οι περισσότερες από τις παρατηρήσεις της μειοψηφούσας ομάδας ταξινομούνται εσφαλμένα. Το επιθυμητό αποτέλεσμα δεν είναι, βέβαια, η ελαχιστοποίηση του βαθμού της ολικής εσφαλμένης ταξινόμησης, αλλά η αναγνώριση των περισσότερων παρατηρήσεων της μικρότερης ομάδας.

Για παράδειγμα, ελέγχοντας ασθενείς για τη μόλυνσή τούς από τον ιό του AIDS, το σφάλμα ταξινόμησής τούς θα είναι πολύ μικρό αφού λίγοι στην πραγματικότητα έχουν προσβληθεί από τον ιό. Ωστόσο, το αποτέλεσμα της ταξινόμησης είναι όχι ιδιαίτερης αξίας, μια και ο στόχος είναι η αναγνώριση των προσβεβλημένων ατόμων.

Τα αποτελέσματα της ταξινόμησης των παρατηρήσεων με διαφορετικές μεθόδους για την αναγνώριση των «μειοψηφουσών» ομάδων μπορούν να διερευνηθούν με την ταξινόμηση όλων των παρατηρήσεων με βάση τούς διακριτικούς βαθμούς.

2.8 Βήματα Διακριτικής Ανάλυσης

Στην παρακάτω εικόνα παρουσιάζονται τα στάδια της διακριτικής ανάλυσης:



Εικόνα 14: Τα βήματα της Διακριτικής Ανάλυσης

2.8.1 Περιγραφή των δεδομένων

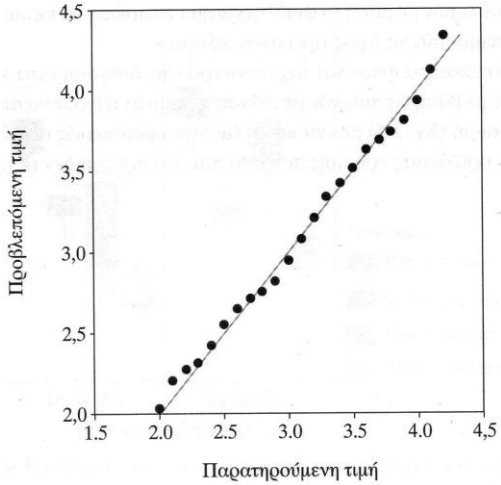
Το πρώτο βήμα είναι η παράθεση των δεδομένων και η περιγραφή στατιστικών μέτρων, όπως η μέση τιμή και η τυπική απόκλιση, τα οποία δίνουν μια γενική εικόνα του υλικού που χρησιμοποιούμε στην έρευνα.

2.8.2 Έλεγχος έλλειψης συσχέτισης των ανεξάρτητων μεταβλητών

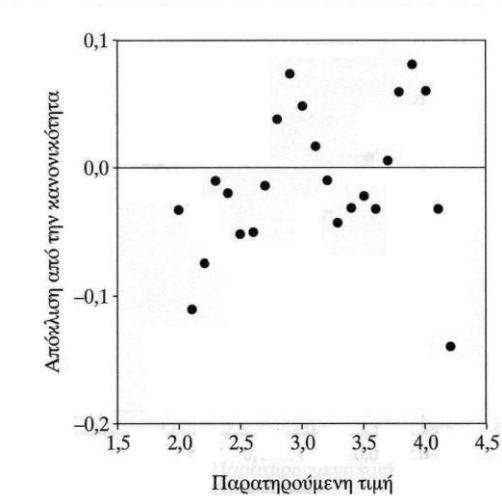
Διαπιστώνουμε, για το σύνολο των δεδομένων, με βάση τις τιμές των συντελεστών συσχέτισης των μεταβλητών του συγκεντρωτικού πίνακα συσχετίσεων εντός των ομάδων, αν υπάρχει συσχέτιση μεταξύ των μεταβλητών. Μπορούμε να χρησιμοποιήσουμε την ορίζουσα των σφαλμάτων του πίνακα συσχετίσεων (μηδενική τιμή σημαίνει απουσία συσχέτισης) και τον έλεγχο σφαιρικότητας του Bartlett.

2.8.3 Έλεγχος κανονικότητας κατανομών των μεταβλητών

Ο έλεγχος κανονικότητας γίνεται μέσω φυλλογραφήματος, διαγραμμάτων κανονικής κατανομής με βάση τις πιθανότητες, όπου τα σημεία τείνουν να ακολουθούν ευθεία γραμμή και μέσω διαγραμμάτων απόκλισης από την κανονικότητα όπου τα σημεία τείνουν να κατανέμονται ομοιόμορφα σε ζώνη πάνω και κάτω από οριζόντια γραμμή που προσεγγίζει την μηδενική τιμή. Στις εικόνες 15-16 φαίνονται δύο παραδείγματα.



Εικόνα 15: Διάγραμμα κανονικότητας των τιμών μιας μεταβλητής



Εικόνα 16: Διάγραμμα απόκλισης από την κανονικότητα μιας μεταβλητής

2.8.4 Ομοιογένεια διακυμάνσεων-συνδιακυμάνσεων

Ελέγχουμε την ομοιογένεια των πινάκων διακυμάνσεων-συνδιακυμάνσεων των τιμών των ανεξάρτητων μεταβλητών στις ομάδες με απλούς ελέγχους όπως ο Μ του Box.

2.8.5 Ερμηνεία της λύσης

Βρίσκουμε τους συντελεστές των διακριτικών συναρτήσεων και μέσω αυτών τη σημασία κάθε μεταβλητής σε μια συνάρτηση.

2.8.6 Επιλογή μεταβλητών

Με χρήση της βηματικής επιλογής μεταβλητών και διαφόρων κριτηρίων, όπως της ελαχιστοποίησης του Λ του Wilks και του V του Rao επιλέγουμε τις διακριτικές μεταβλητές, αυτές δηλαδή που θα εισέλθουν στην διακριτική συνάρτηση.

2.8.7 Ταξινόμηση

Χρησιμοποιώντας τις συναρτήσεις ταξινόμησης εντάσσουμε κάθε παρατήρηση στην ομάδα εκείνη στην οποία έχει τη μεγαλύτερη τιμή, δηλαδή το μεγαλύτερο βαθμό ταξινόμησης. Επίσης η ταξινόμηση μας δείχνει και το βαθμό των εσφαλμένων προβλέψεων, οπότε μπορούμε να αποφανθούμε για την αποτελεσματικότητα της διακριτικής συνάρτησης.

Κεφάλαιο 3: Εφαρμογές

3.1 Περιγραφή Δεδομένων

Το σύνολο των δεδομένων που χρησιμοποιήθηκε για τις εφαρμογές της ανάλυσης κατά συστάδες και της διακριτικής ανάλυσης αποτελείται από 200.701 παρατηρήσεις και το χρονικό διάστημα παρατήρησης είναι από τον Ιανουάριο του 2011 έως και τον Σεπτέμβριο του 2011. Οι παρατηρήσεις αναφέρονται σε πελάτες που έχουν στην κατοχή τους κάποιο τραπεζικό προϊόν, όπως κάρτα ή δάνειο και έχουν καθυστερήσει την πληρωμή περισσότερο από 180 ημέρες. Στην περίπτωση αυτή λέμε ότι το προϊόν για το οποίο δεν έχει πληρώσει τη δόση ο πελάτης έχει διαγραφεί (written-off). Για παράδειγμα, στην περίπτωση που το προϊόν είναι μια πιστωτική κάρτα, ο πελάτης δεν θα μπορεί να τη χρησιμοποιήσει. Οι μεταβλητές στο σύνολο των δεδομένων είναι:

- ***GB***: 1 = ο πελάτης θα αποπληρώσει ποσοστό μικρότερο ή ίσο του 2% του ποσού οφειλής, 0 = θα πληρώσει περισσότερο από 2% του ποσού οφειλής
- ***WO-amount***: Συνολικό ποσό οφειλής
- ***Sum_Paym***: (Άθροισμα πληρωμών τους τελευταίους 3 μήνες)-(Άθροισμα πληρωμών τους προηγούμενους 3 μήνες)
- ***Incr_Bal***: Μέγιστος αριθμός διαδοχικών αυξήσεων οφειλής τους τελευταίους 12 μήνες
- ***last_paym***: αριθμός μηνών από την τελευταία πληρωμή ή επαναφοράς μετά από διαγραφή
- ***sum_delq_months***: Άθροισμα buckets τους τελευταίους 12 μήνες
- ***months_since_wo***: αριθμός μηνών από την πρώτη διαγραφή
- ***property_flag***: ιδιοκτησία (ναι ή όχι)
- ***occupation***: επάγγελμα (κωδικοί)
- ***Marital_status***: Οικογενειακή κατάσταση
- ***Income***: Εισόδημα
- ***ZIP_area***: Περιοχή κατοικίας

Στην μεταβλητή GB η τιμή 1 σημαίνει ότι ο πελάτης θα έχει ποσοστό επαναφοράς (πληρωμής μετά τη διαγραφή) λιγότερο από 2%. Σε άλλη περίπτωση η τιμή είναι 0. Η σημασία της μεταβλητής *Sum_Paym* είναι ότι αν η τιμή της είναι θετική τότε η πληρωμές που έχουν γίνει πιο πρόσφατα είναι μεγαλύτερες από αυτές που έγιναν νωρίτερα. Στην μεταβλητή *sum_delq_months* αναφέρεται ο όρος bucket. Με τον όρο αυτό εννοούμε την ομάδα ημερών καθυστέρησης της πληρωμής. Για παράδειγμα αν ένας πελάτης έχει καθυστερήσει 60 μέρες την πληρωμή τότε bucket=3. Στο σύνολο των δεδομένων του προβλήματος η τιμή είναι μεγαλύτερη του 6. Δηλαδή έχει ήδη γίνει διαγραφή. Στον παρακάτω πίνακα μπορούμε να δούμε τα buckets ανάλογα με την καθυστέρηση σε ημέρες.

Bucket	**Καθυστέρηση πληρωμής (ημέρες)**
0	0
1	1-29
2	30-59
3	60-89
4	90-119
5	120-149
6	150-179
7	180-209
...	…

Πίνακας 12: Buckets – καθυστέρηση

Ο πίνακας 13 δείχνει τη μέση τιμή και την τυπική απόκλιση για κάθε μεταβλητή, καθώς και τη μέγιστη και ελάχιστη τιμή για κάθε μεταβλητή.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
GB	200701	0	1	,90	,294
WO_amount	200701	,04	91524,45	13329,4645	13821,20761
Sum_Paym	200700	-57990,00	42538,40	-256,6456	1222,86947
Incr_Bal	200701	0	11	2,30	2,637
last_paym	200701	0	12	6,10	2,635
sum_delq_months	200701	0	108	72,70	27,444
months_since_wo	200701	0	32	12,56	8,464
property_flag	200701	0	1	,53	,491
income	200701	,00	316956,00	792,6038	3706,84965
Valid N (listwise)	200700				

Πίνακας 13: Περιγραφή δεδομένων

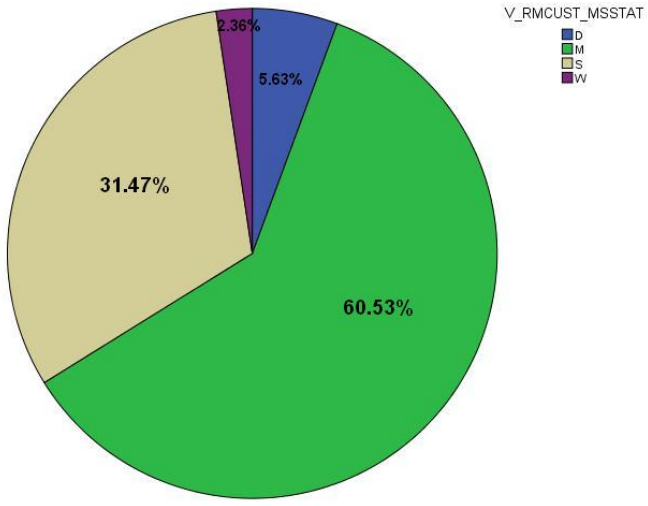
Στον πίνακα 14 βλέπουμε τις συχνότητες των «κακών» (θα αποπληρώσουν το πολύ το 2% του χρέους τους) και των «καλών» πελατών. Παρατηρούμε ότι οι παρατηρήσεις που βρίσκονται στην πρώτη ομάδα («καλοί») είναι 19115 και αποτελούν το 9.5% του συνόλου των παρατηρήσεων ενώ αυτοί που βρίσκονται στην δεύτερη είναι 181586, δηλαδή το 90.5%

GB

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	19115	9,5	9,5	9,5
	1	181586	90,5	90,5	100,0
	Total	200701	100,0	100,0	

Πίνακας 14: Συχνότητες της μεταβλητής *GB*

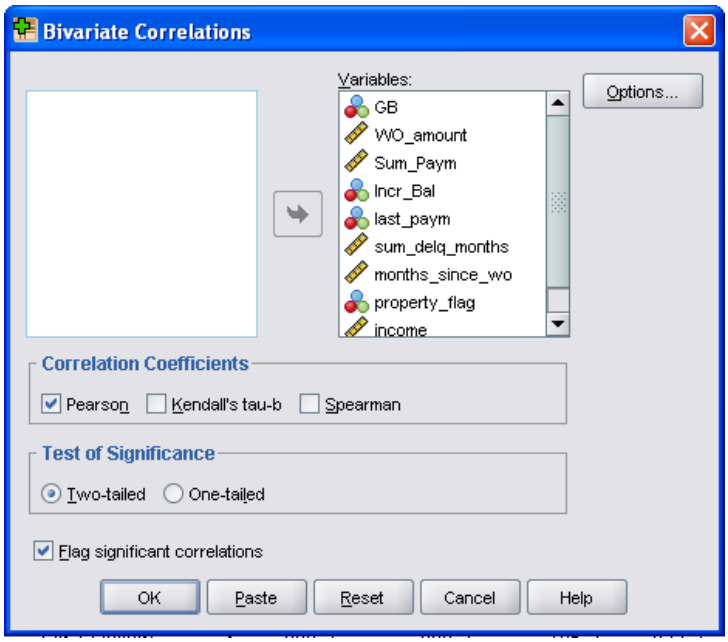
Η εικόνα δείχνει την οικογενειακή κατάσταση των πελατών σε ένα κυκλικό διάγραμμα (S – ελεύθερος, M – παντρεμένος, D – χωρισμένος, W – χήρος).



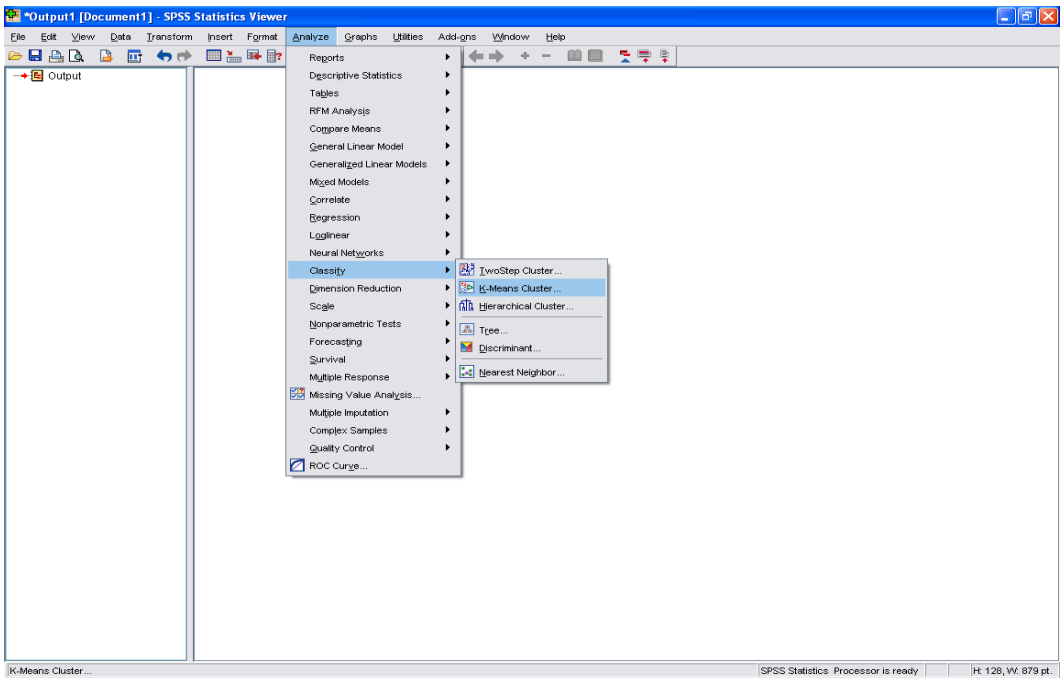
3.2 Ανάλυση κατά συστάδες

3.2.1 Ανάλυση

Αρχικά κατασκευάστηκε ένας πίνακας συσχετίσεων των μεταβλητών. Χρησιμοποιήθηκε ο συντελεστής συσχέτισης του Pearson. Από τον πίνακα βλέπουμε ότι η μεταβλητή *WO_amount* έχει υψηλή συσχέτιση με την *Sum_Paym*. Ομοίως η *Incr_Bal* με τις *sum_dlq_months*, *Sum_Paym* και *months_since_wo* και η *property_flag* με την *last_paym*. Άρα δεν θα χρησιμοποιηθούν στην ανάλυση κατά συστάδες οι μεταβλητές *WO_amount*, *Incr_Bal* και *property_flag*.

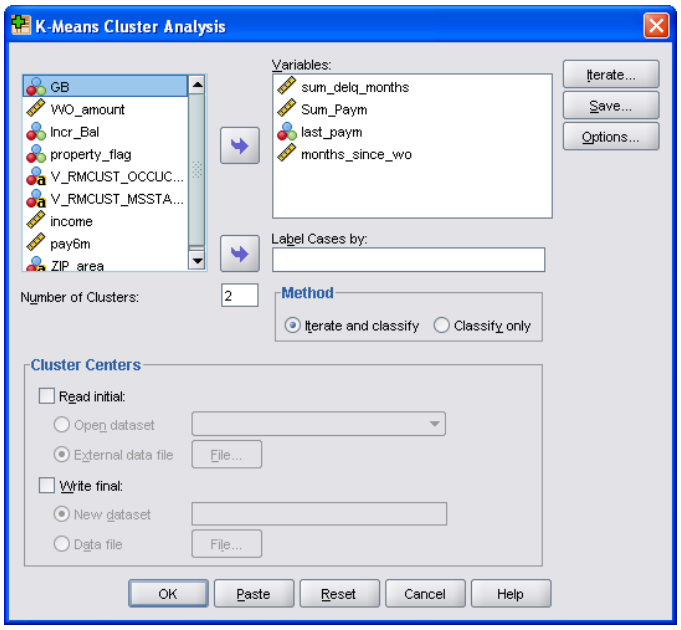


Επίσης η μεταβλητή *income* απομακρύνεται επειδή υπάρχουν αρκετές ελλειπούσες τιμές (missing values) και επηρεάζει την τελική επιλογή των συστάδων.

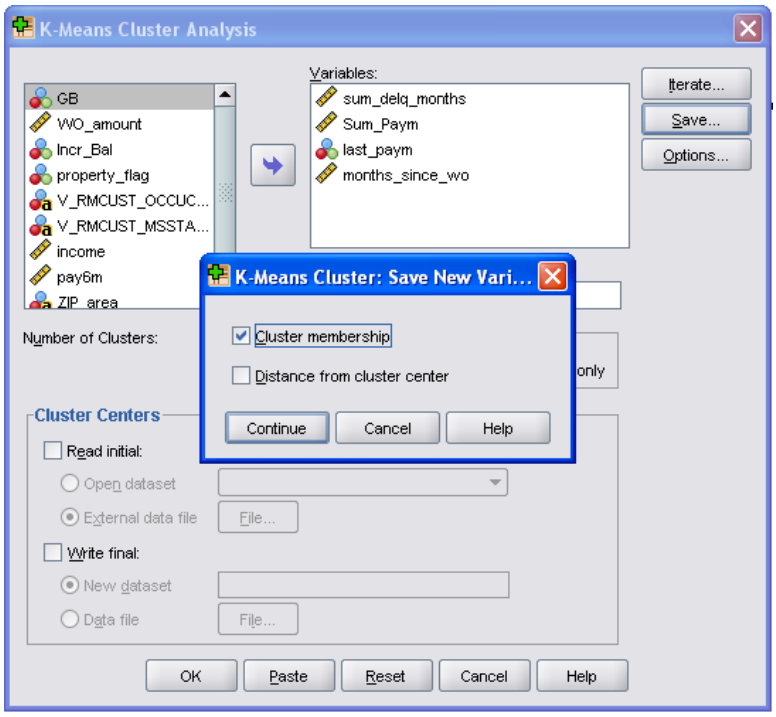


Η μέθοδος που χρησιμοποιήθηκε για την ανάλυση κατά συστάδες είναι η k-means, ενώ για την επιλογή του κατάλληλου αριθμού των συστάδων χρησιμοποιήθηκε το κριτήριο VRC (Variance Ratio Criterion) όπως περιγράφεται αναλυτικά παρακάτω:

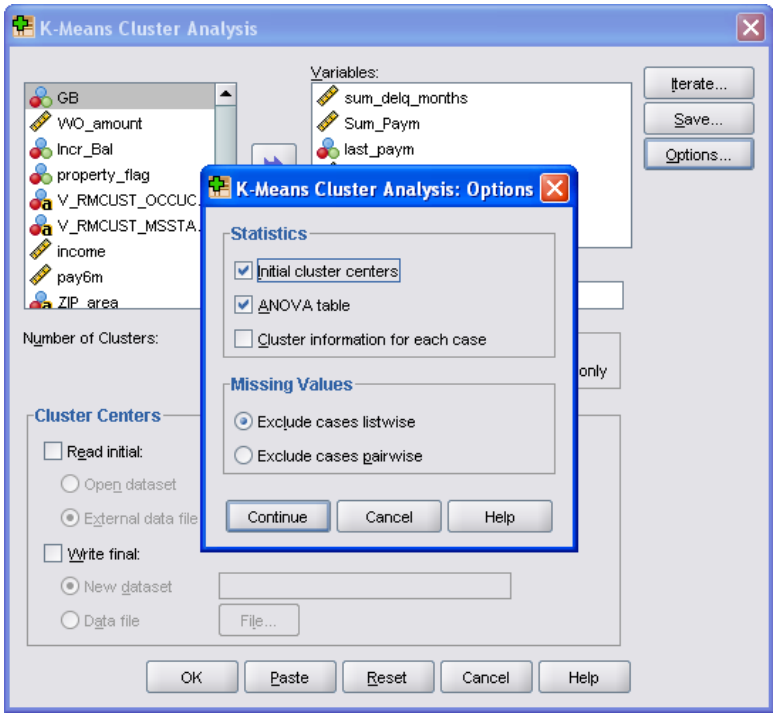
- ***Για k = 2***



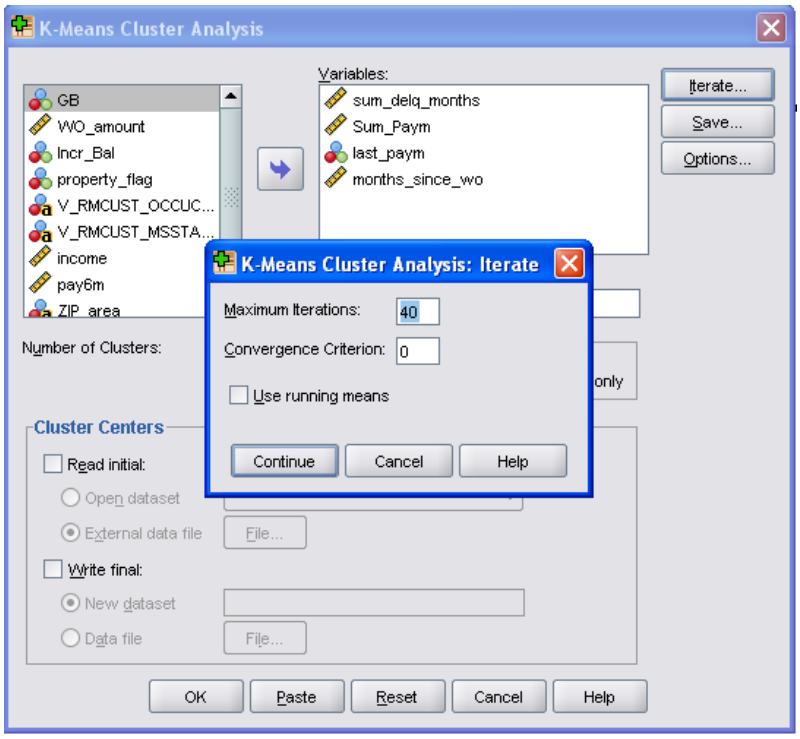
Οι μεταβλητές που εισάγουμε στην ανάλυση (*Sum_paym, last_paym, months_since_wo* και *sum_dlq_months*) προκύπτουν μετά από επαναλήψεις της μεθόδου, αφού οι υπόλοιπες μεταβλητές δεν επηρεάζουν ουσιαστικά τα αποτελέσματα, ενώ στο πεδίο Number of Clusters γράφουμε τον αριθμό 2.
Στο πλαίσιο Save μας δίνεται η δυνατότητα να δημιουργήσουμε μία νέα μεταβλητή στην οποία καταγράφεται η συστάδα όπου θα ανήκει τελικά η κάθε παρατήρηση.



Και τέλος στο πλαίσιο Options επιλέγουμε να κατασκευαστεί ο πίνακας ανάλυσης διασποράς των μεταβλητών.



Για να επιτευχθεί σύγκλιση στον εντοπισμό των κέντρων στο πλαίσιο Iterate της K-means αυξάνουμε τον αριθμό στο πεδίο iterations.



Ο πίνακας 15 είναι ο πίνακας ανάλυσης διασποράς που προκύπτει από την ανάλυση κατά συστάδες για k = 2.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
sum_delq_months	283597,841	1	751,754	200698	377,248	,000
Sum_Paym	1,030E11	1	982124,222	200698	104891,896	,000
last_paym	551,177	1	6,938	200698	79,445	,000
months_since_wo	3538,284	1	71,624	200698	49,401	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Πίνακας 15: Ανάλυση διασποράς για τη μέθοδο k-means με k = 2.

- ***Για k = 3, 4 και 5***

Με αλλαγή του αριθμού των συστάδων σε 3, 4 και 5 προκύπτουν οι πίνακες 16-18. Στον πίνακα Iteration History των δεδομένων εξόδου φαίνεται σε ποιο βήμα έχουμε σύγκλιση σε κάθε περίπτωση. Για παράδειγμα για k = 2 έχουμε σύγκλιση μετά από 9 επαναλήψεις, ενώ για k = 3 μετά από 18 επαναλήψεις.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
sum_delq_months	281137,412	2	750,369	200697	374,665	,000
Sum_Paym	7,359E10	2	762091,705	200697	96561,686	,000
last_paym	5133,858	2	6,889	200697	745,173	,000
months_since_wo	7975,136	2	71,562	200697	111,443	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Πίνακας 16: Ανάλυση διασποράς για τη μέθοδο k-means με k = 3.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
sum_delq_months	1546644,935	3	730,055	200696	2118,531	,000
Sum_Paym	6,924E10	3	460414,635	200696	150388,981	,000
last_paym	3533,577	3	6,888	200696	513,015	,000
months_since_wo	242966,587	3	68,010	200696	3572,504	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Πίνακας 17: Ανάλυση διασποράς για τη μέθοδο k-means με k = 4.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
sum_delq_months	1173511,363	4	729,789	200695	1608,014	,000
Sum_Paym	5,509E10	4	397544,693	200695	138564,299	,000
last_paym	4538,379	4	6,850	200695	662,511	,000
months_since_wo	179205,696	4	68,071	200695	2632,641	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Πίνακας 18: Ανάλυση διασποράς για τη μέθοδο k-means με k = 5.

Στον πίνακα 19 βλέπουμε τα αρχικά κέντρα των μεταβλητών των τριών κλάσεων, τα οποία επιλέγονται αυθαίρετα. Κάθε παρατήρηση τοποθετείται στην συστάδα από της οποίας το κέντρο απέχει λιγότερο και τα κέντρα υπολογίζονται εκ νέου. Η μέθοδος λειτουργεί έτσι ώστε να ελαχιστοποιεί τη διαφορετικότητα εντός των συστάδων και να την μεγιστοποιεί μεταξύ των συστάδων.

Initial Cluster Centers

	Cluster		
	1	2	3
sum_delq_months	9	46	56
Sum_Paym	-7739,00	-57990,00	42538,40
last_paym	9	9	3
months_since_wo	9	2	14

Πίνακας 19: Αρχικά κέντρα των συστάδων.

Όπως αναφέρθηκε πριν οι επαναλήψεις που εκτέλεσε η μέθοδος μέχρι να καταλήξει στα τελικά κέντρα ήταν 18 και φαίνεται στον πίνακα 20. Στην τελευταία επανάληψη έχουμε σύγκλιση αφού τα κέντρα έχουν πολύ μικρή ή μηδενική μεταβολή. Η διαδικασία σταματά και προκύπτουν τα τελικά κέντρα που φαίνονται στον πίνακα 21.

Iteration History[a]

	Change in Cluster Centers		
Iteration	1	2	3
1	7484,496	12979,997	18398,757
2	,268	7447,440	5214,074
3	,617	5903,208	2767,526
4	1,012	4440,105	2062,025
5	,932	3007,701	1527,356
6	,063	1466,370	964,543
7	,210	1253,649	677,583
8	,353	868,283	400,308
9	,338	570,887	212,339
10	,053	324,305	191,317
11	,029	243,002	145,384
12	,065	135,129	59,315
13	,118	100,282	11,785
14	,067	98,703	35,251
15	,234	192,322	23,301
16	,279	216,361	23,140
17	,093	60,763	,000
18	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 18. The minimum distance between initial centers is 50251,014.

Πίνακας 20: Επαναλήψεις της k-means.

Final Cluster Centers

	Cluster		
	1	2	3
sum_delq_months	73	42	47
Sum_Paym	-250,71	-18681,50	9824,40
last_paym	6	7	1
months_since_wo	13	9	7

Πίνακας 21: Τελικά κέντρα των συστάδων.

Συγκρίνοντας τα αρχικά και τα τελικά κέντρα παρατηρούμε ότι τις μικρότερες μεταβολές τις έχουμε στην $2^{η}$ συστάδα.
Από τον πίνακα ανάλυσης διασποράς (πίνακας 16) χρησιμοποιούμε τις F-τιμές ως ένδειξη για τη συμβολή μιας μεταβλητής στη διάκριση μεταξύ των ομάδων. Τα επίπεδα σημαντικότητας του πίνακα ανάλυσης διασποράς δεν μπορούν να χρησιμοποιηθούν για να ελέγξουμε την υπόθεση ισότητας των μέσων των μεταβλητών, οπότε τα αποτελέσματα του πίνακα αυτού έχουν καθαρά περιγραφικό χαρακτήρα και δείχνουν το ρόλο που παίζουν οι μεταβλητές στη δημιουργία των συστάδων. Αυτό συμβαίνει διότι οι ομάδες σχηματίζονται σκόπιμα σύμφωνα με την απόσταση μεταξύ τους στον πολυδιάστατο χώρο, δηλαδή η υπόθεση για την τυχαιότητα των παρατηρήσεων στις διάφορες δεν ικανοποιείται. Από τον πίνακα βλέπουμε ότι και οι τέσσερις μεταβλητές είναι σημαντικές αφού *Sig* < 0.05.
Στον πίνακα 22 βλέπουμε τον αριθμό των παρατηρήσεων που ανήκουν σε κάθε συστάδα καθώς και τον συνολικό τους αριθμό. Παρατηρούμε ότι η πρώτη συστάδα περιέχει τον μεγαλύτερο αριθμό παρατηρήσεων από τις 3.

Number of Cases in each Cluster

Cluster	1	199961,000
	2	303,000
	3	436,000
Valid		200700,000
Missing		1,000

Πίνακας 22: Αριθμός παρατηρήσεων σε κάθε συστάδα.

Το κριτήριο λόγου διασποράς (VRC) δίνεται από τη σχέση:

$$VRC_k = \left(SS_B / \left(K-1\right)\right) / \left(SS_W / \left(N-K\right)\right)$$

όπου SS_B η συνολική διασπορά μεταξύ συστάδων και SS_W η συνολική διασπορά εντός των συστάδων. Το VRC_k προκύπτει εύκολα αν υπολογίσουμε το άθροισμα των F από τους πίνακες ανάλυσης διασποράς (ANOVA). Για να καθορίσουμε τον αριθμό των συστάδων υπολογίζουμε το ω_k για κάθε k χρησιμοποιώντας τη σχέση:

$$\omega_k = \left(VRC_{k+1} - VRC_k\right) - \left(VRC_k - VRC_{k-1}\right)$$

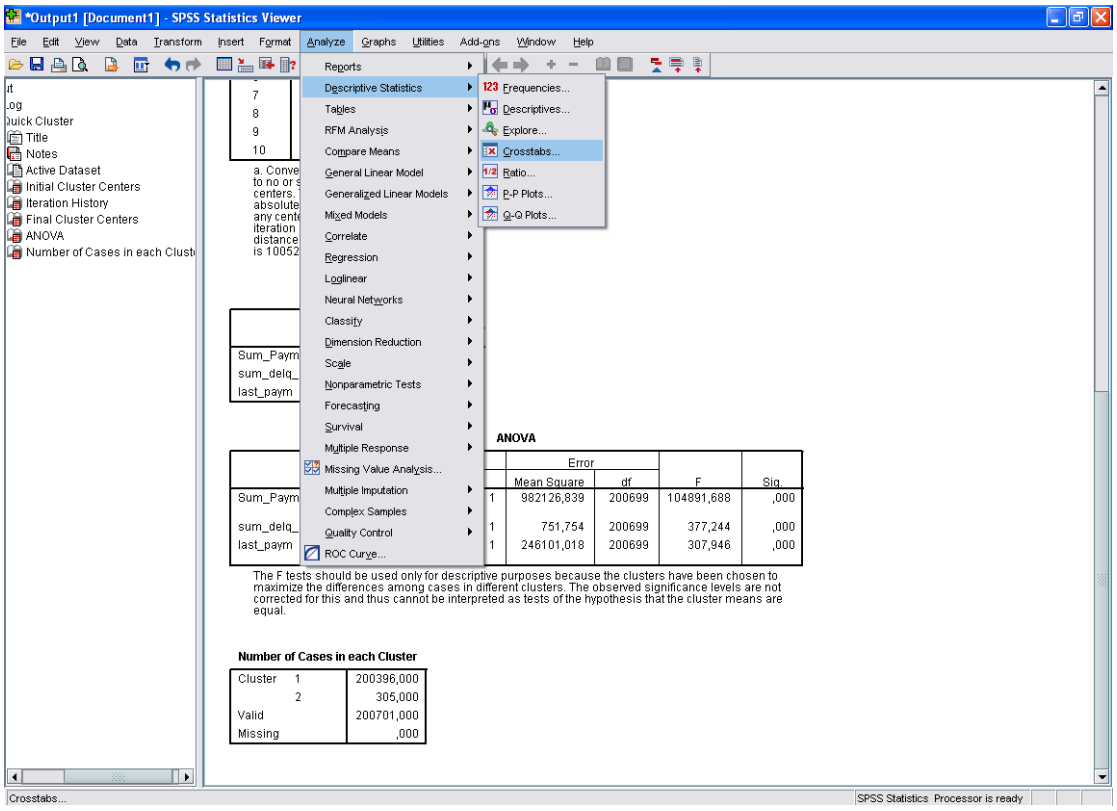
και επιλέγουμε το k εκείνο που ελαχιστοποιεί την (απόλυτη) τιμή του ω_k. Στον πίνακα 23 βλέπουμε τις τιμές των VRC_k και ω_k για τις τιμές του k.

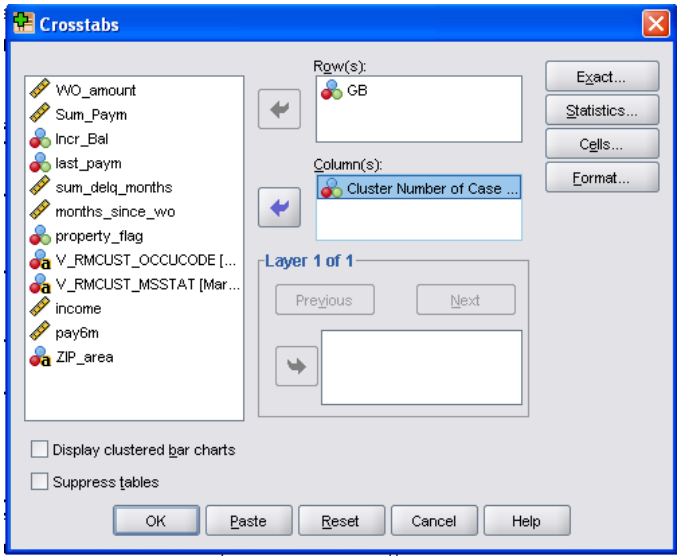
k	**VRC_k**	**ω_k**
2	105397.990	
3	97792.968	66405.085
4	156593.031	71925.68
5	143467.465	

Πίνακας 23: VRC_k και ω_k για k = 2 έως 5

Ο αριθμός των συστάδων θα είναι τελικά 3.
Αυτό προκύπτει και με την χρήση πινάκων διασταύρωσης (Crosstabs) της μεταβλητής GB με την μεταβλητή που δημιουργείται κάθε φορά μετά την εκτέλεση της K-means και δείχνει σε ποια συστάδα ανήκει κάθε παρατήρηση (Cluster Number of Case QCL_1). Υπολογίζουμε το ποσοστό

των «κακών» πελατών σε κάθε συστάδα. Τα ποσοστά αυτά πρέπει να διαφέρουν όσο το δυνατόν περισσότερο.





Ο πίνακας που προκύπτει από την διασταύρωση της μεταβλητής GB με την μεταβλητή που προέκυψε από την k-means για k = 2 είναι ο πίνακας 24 ενώ τα ποσοστά που προκύπτουν είναι 82.95% για την $1^{η}$ συστάδα και 90.48% για τη $2^{η}$.

GB * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case		
		1	2	Total
GB	0	52	19062	19114
	1	253	181333	181586
Total		305	200395	200700

Πίνακας 24: Διασταύρωση μεταβλητών GB - QCL_1

Ο πίνακας που προκύπτει από την διασταύρωση της μεταβλητής GB με την μεταβλητή που προέκυψε από την k-means για k = 3 είναι ο πίνακας 25 ενώ τα ποσοστά που προκύπτουν είναι 90.58% για την $1^{η}$ συστάδα, 82.84% για τη $2^{η}$ και 49.77% για την $3^{η}$.

GB * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case			
		1	2	3	Total
GB	0	18843	52	219	19114
	1	181118	251	217	181586
Total		199961	303	436	200700

Πίνακας 25: Διασταύρωση μεταβλητών GB - QCL_2

Ο πίνακας που προκύπτει από την διασταύρωση της μεταβλητής GB με την μεταβλητή που προέκυψε από την k-means για k = 4 είναι ο πίνακας 26 ενώ τα ποσοστά που προκύπτουν είναι 93.01% για την $1^{η}$ συστάδα, 50% για τη $2^{η}$, 90.4% για την $3^{η}$ και 83.06% για την $4^{η}$.

GB * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case				
		1	2	3	4	Total
GB	0	926	212	17945	31	19114
	1	12317	212	168905	152	181586
Total		13243	424	186850	183	200700

Πίνακας 26: Διασταύρωση μεταβλητών GB - QCL_3

Τέλος, ο πίνακας που προκύπτει από την διασταύρωση της μεταβλητής GB με την μεταβλητή που προέκυψε από την k-means για k = 5 είναι ο πίνακας 27 ενώ τα ποσοστά που προκύπτουν είναι 90.48% για την $1^{η}$ συστάδα, 83.06% για τη $2^{η}$, 93.0% για την $3^{η}$, 45.90% για την $4^{η}$ και 45.05% για την $5^{η}$.

GB * Cluster Number of Case Crosstabulation

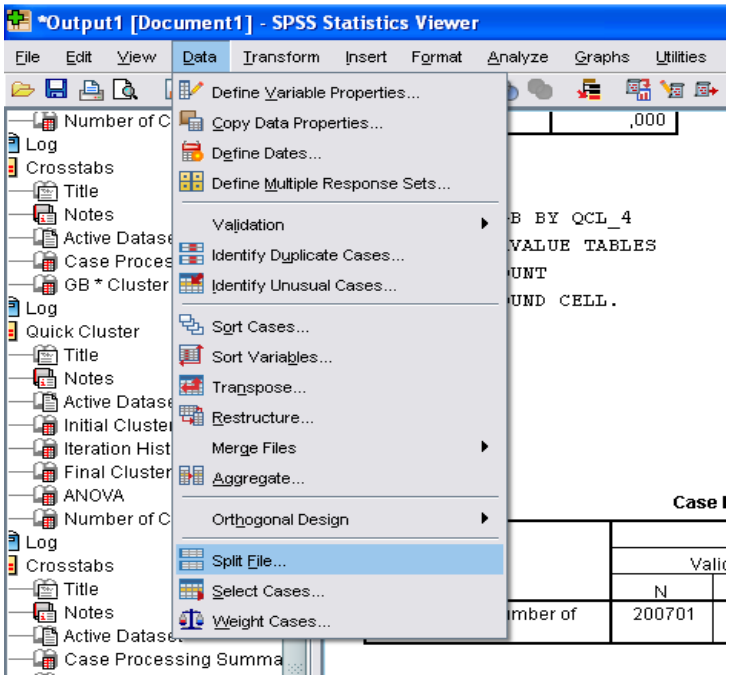
Count

		Cluster Number of Case					
		1	2	3	4	5	Total
GB	0	17776	31	907	350	50	19114
	1	169041	152	12055	297	41	181586
Total		186817	183	12962	647	91	200700

Πίνακας 27: Διασταύρωση μεταβλητών GB - QCL_4

Τα ποσοστά πρέπει να έχουν όσο το δυνατόν μεγαλύτερη απόσταση γι' αυτό η βέλτιστη λύση είναι για k = 3, αφού για παράδειγμα στην περίπτωση k = 5 τα ποσοστά για την $1^{η}$ και $3^{η}$ συστάδα δεν διαφέρουν σημαντικά, και ομοίως τα ποσοστά για την $4^{η}$ και $5^{η}$ συστάδα.

3.2.2 Ανάλυση αποτελεσμάτων



Για να δούμε το προφίλ των πελατών σε κάθε συστάδα χωρίζουμε το σύνολο των δεδομένων με βάση τη μεταβλητή QCL_2 που δείχνει σε ποια συστάδα ανήκει κάθε παρατήρηση. Στη συνέχεια φτιάχνουμε πίνακες με τα χαρακτηριστικά κάθε μεταβλητής της ανάλυσης σε κάθε συστάδα. Οι πίνακες 28-30 δείχνουν τη μέση τιμή, την ελάχιστη και τη μέγιστη τιμή καθώς και την τυπική απόκλιση κάθε μεταβλητής.

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Sum_Paym	199961	-9444,00	4759,27	-250,7075	710,33789
last_paym	199961	0	12	6,11	2,626
sum_delq_months	199961	0	108	72,80	27,395
months_since_wo	199961	0	32	12,58	8,465
Valid N (listwise)	199961				

a. Cluster Number of Case = 1

Πίνακας 28: Στατιστικά μέτρα για την $1^{η}$ συστάδα

Στην πρώτη συστάδα η μέση τιμή της μεταβλητής *Sum_Paym* είναι -250.71 με ελάχιστη τιμή -9444 και μέγιστη 4795. Δηλαδή κατά μέσο όρο οι πελάτες που βρίσκονται σε αυτή την συστάδα έχουν πληρώσει τους τελευταίους 3 μήνες λιγότερα χρήματα από αυτά που είχαν πληρώσει νωρίτερα. Η μέση τιμή της μεταβλητής *last_paym* είναι 6.11 ενώ της sum_delq_months είναι 72.80. Τέλος, ο μέσος αριθμός μηνών από την πρώτη διαγραφή είναι 12.58. Σε αυτή τη συστάδα βρίσκονται δηλαδή οι πελάτες που έχουν «κακή» συμπεριφορά, όσον αφορά στην πληρωμή των οφειλών τους και υπάρχει ελάχιστη πιθανότητα πληρωμής.

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Sum_Paym	303	-57990,00	-9499,31	-18681,5040	10859,68804
last_paym	303	0	12	7,44	3,427
sum_delq_months	303	0	108	42,44	29,417
months_since_wo	303	0	30	9,12	7,594
Valid N (listwise)	303				

a. Cluster Number of Case = 2

Πίνακας 29: Στατιστικά μέτρα για την 2η συστάδα

Στη δεύτερη συστάδα η μέση τιμή της μεταβλητής *Sum_Paym* είναι -18681.50 ενώ η ελάχιστη -57990 και η μέγιστη -9499, δηλαδή οι πελάτες της ομάδας αυτής είχαν κάνει μεγαλύτερες πληρωμές πριν 6 μήνες απ' ότι το τελευταίο τρίμηνο. Η μέση τιμή της μεταβλητής *sum_delq_months* είναι 42.44 ενώ της *last_paym* 7.44. Οι καθυστερήσεις στις πληρωμές είναι πολύ μικρότερες από αυτές της προηγούμενης συστάδας, ενώ και ο μέσος αριθμός μηνών από την πρώτη διαγραφή είναι μικρότερος. Στην συστάδα αυτή ανήκουν οι πελάτες που έχουν «μέτρια συμπεριφορά» όσον αφορά στην πληρωμή των οφειλών τους και υπάρχει κάποια πιθανότητα πληρωμής.

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Sum_Paym	436	4801,00	42538,40	9824,4045	6147,23652
last_paym	436	0	3	1,38	,993
sum_delq_months	436	0	108	47,24	24,870
months_since_wo	436	0	28	7,25	6,403
Valid N (listwise)	436				

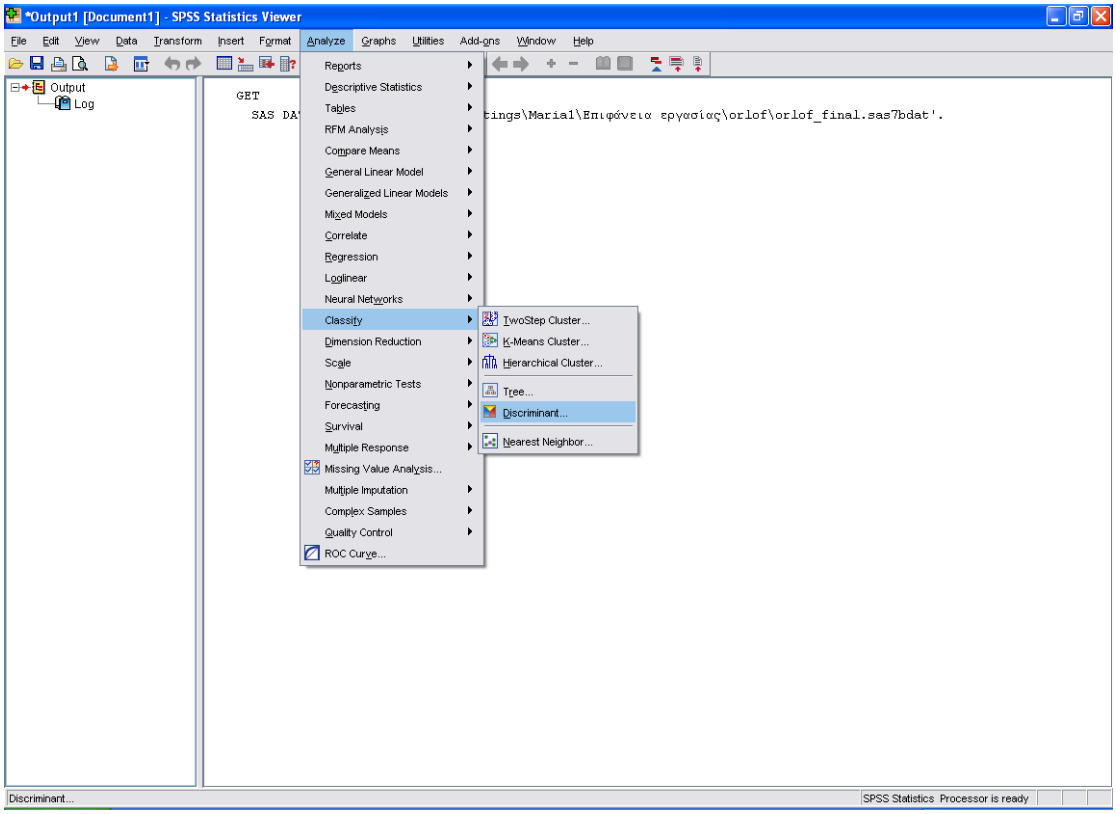
a. Cluster Number of Case = 3

Πίνακας 30: Στατιστικά μέτρα για την 3η συστάδα

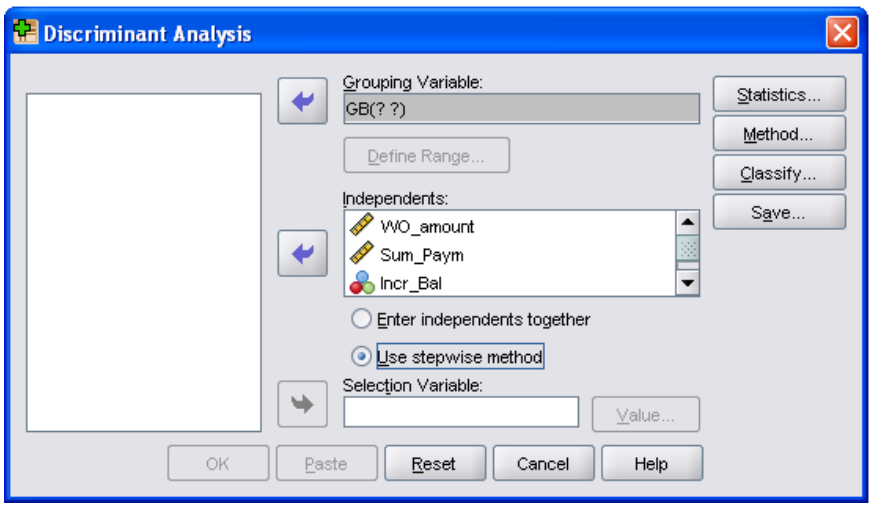
Τέλος, στην 3η συστάδα η μέση τιμή της μεταβλητής *Sum_Paym* είναι 9824.40 με ελάχιστη τιμή 4801 και μέγιστη 42538.40. Παρατηρούμε ότι στη συστάδα αυτή όλοι οι πελάτες έχουν πληρώσει περισσότερα το τελευταίο τρίμηνο σε σχέση με το προηγούμενο. Ενώ κατά μέσο όρο οι πελάτες έχουν αρκετά υψηλό αριθμό buckets, η στιγμή πρώτης διαγραφής είναι πιο πρόσφατη από τις δύο προηγούμενες ομάδες. Οι πελάτες αυτής της συστάδας έχουν «καλύτερη συμπεριφορά», όσον αφορά στην πληρωμή των οφειλών τους και υπάρχει πιθανότητα πληρωμής.

3.3 Διακριτική Ανάλυση

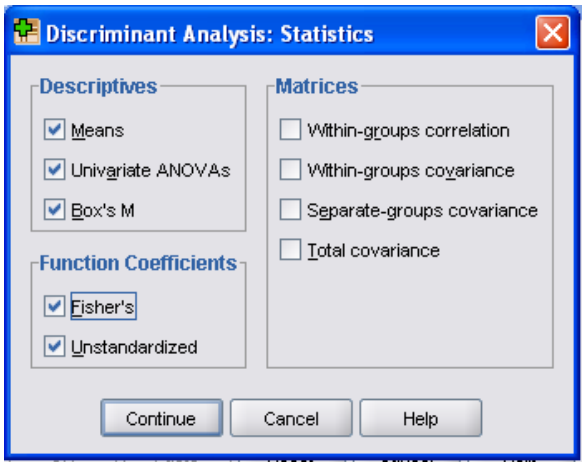
3.3.1 Ανάλυση

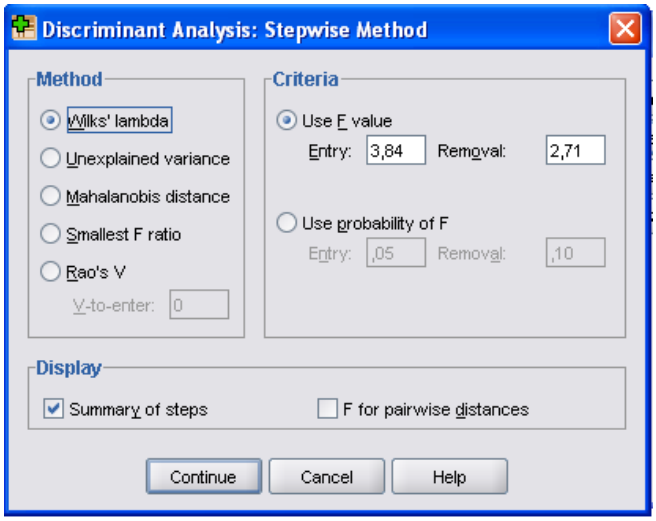


Στο παράθυρο διαλόγου της Διακριτικής Ανάλυσης εισάγουμε την μεταβλητή GB στο πλαίσιο «Μεταβλητή ομαδοποίησης» και τις υπόλοιπες στο πλαίσιο των ανεξάρτητων μεταβλητών και επιλέγουμε τη χρήση βηματικής μεθόδου (Use Stepwise Method).

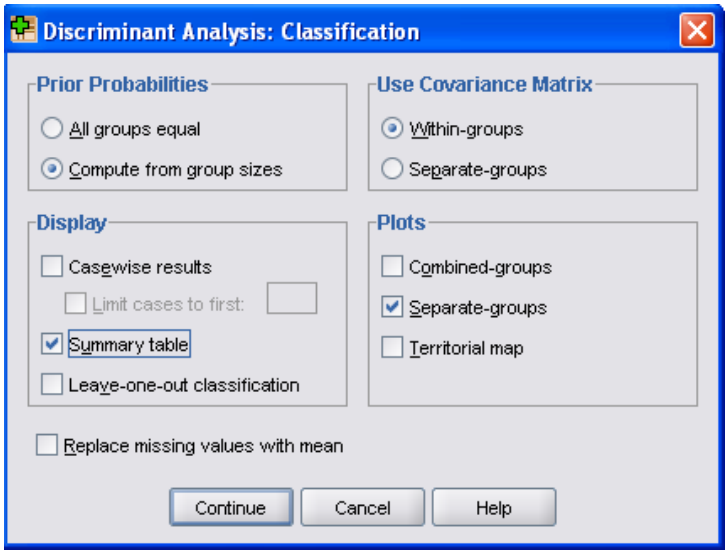


Στο παράθυρο Statistics επιλέγουμε Means για να πάρουμε περιγραφικά στατιστικά και Univariate ANOVAs για να κατασκευαστεί ένας πίνακας ανάλυσης διασποράς. Επίσης επιλέγουμε τους συντελεστές της διακριτικής συνάρτησης. Στο παράθυρο Method επιλέγουμε το στατιστικό Λ του Wilks (Wilks' lamda).

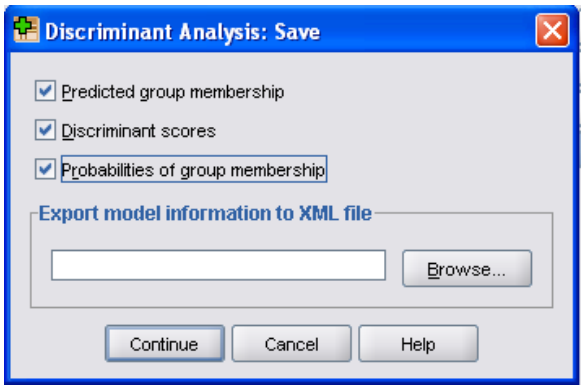




Στο παράθυρο της ταξινόμησης επιλέγουμε να υπολογιστούν οι εκ των προτέρων πιθανότητες από το μέγεθος των ομάδων καθώς επίσης και την δημιουργία ενός περιληπτικού πίνακα που θα δείχνει πόσο επιτυχημένη είναι η ταξινόμηση.



Τέλος, για να χρησιμοποιήσουμε την διακριτική ανάλυση για να προβλέψουμε σε ποια ομάδα θα ανήκει μία παρατήρηση, στο παράθυρο Save επιλέγουμε Predicted group membership για να δούμε σε ποια ομάδα προβλέπεται να ανήκει η παρατήρηση, Discriminant scores για να πάρουμε την τιμή της διακριτικής συνάρτησης για κάθε παρατήρηση και Probabilities of group membership για να δημιουργηθούν δύο νέες μεταβλητές που μας δίνουν την πιθανότητα να ανήκει μία παρατήρηση στην πρώτη ομάδα και την πιθανότητα να ανήκει στη δεύτερη.



Στον πίνακα 31 βλέπουμε τη μέση τιμή κάθε μεταβλητής σε κάθε ομάδα χωριστά.

Group Statistics

GB		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
0	WO_amount	10334,5681	10577,74353	19114	19114,000
	Sum_Paym	-16,6901	1702,37612	19114	19114,000
	Incr_Bal	2,5400	2,33329	19114	19114,000
	last_paym	2,4057	3,02055	19114	19114,000
	sum_delq_months	61,1066	25,13120	19114	19114,000
	months_since_wo	8,6741	7,57615	19114	19114,000
	property_flag	,6585	,47372	19114	19114,000
	income	973,2058	3281,08706	19114	19114,000
1	WO_amount	13644,7285	14082,40029	181586	181586,000
	Sum_Paym	-281,9037	1158,04945	181586	181586,000
	Incr_Bal	2,2717	2,66585	181586	181586,000
	last_paym	6,4881	2,26354	181586	181586,000
	sum_delq_months	73,9233	27,39198	181586	181586,000
	months_since_wo	12,9699	8,44884	181586	181586,000
	property_flag	,5198	,49114	181586	181586,000
	income	773,5977	3748,36591	181586	181586,000
Total	WO_amount	13329,4798	13821,24033	200700	200700,000
	Sum_Paym	-256,6456	1222,86947	200700	200700,000
	Incr_Bal	2,2973	2,63716	200700	200700,000
	last_paym	6,0993	2,63450	200700	200700,000
	sum_delq_months	72,7027	27,44382	200700	200700,000
	months_since_wo	12,5608	8,46410	200700	200700,000
	property_flag	,5330	,49120	200700	200700,000
	income	792,6077	3706,85846	200700	200700,000

Πίνακας 31: Στατιστικά των δύο ομάδων

Παρατηρούμε ότι ενώ η μέση τιμή όλων των παρατηρήσεων για τη μεταβλητή *Sum_Paym* είναι -256.65, η μέση τιμή της μεταβλητής αυτής είναι μόλις -16.69 για την πρώτη ομάδα (*GB* = 0) και -281.90 για την δεύτερη ομάδα (*GB* = 1). Επίσης η μεταβλητή *last_paym* έχει μέση τιμή 6.10 για όλες τις παρατηρήσεις αλλά για την πρώτη ομάδα η τιμή αυτή είναι 2.40 ενώ για τη δεύτερη 6.49.
Δηλαδή όπως φαίνεται απ' τον πίνακα παρόλο που ο αριθμός των διαδοχικών αυξήσεων του χρέους μέσα στους τελευταίους 12 μήνες δεν διαφέρει πολύ ανάμεσα στις δύο ομάδες, οι πελάτες της δεύτερης ομάδας έχουν πληρώσει λιγότερες οφειλές το τελευταίο τρίμηνο σε σχέση με 6 μήνες πριν, ενώ έχει μεσολαβήσει και πολύ μεγαλύτερο χρονικό διάστημα από την τελευταία πληρωμή. Επίσης έχουν μεγαλύτερο μέσο άθροισμα buckets, ενώ μεσολαβεί και περισσότερο χρονικό διάστημα από τη στιγμή πρώτης διαγραφής. Η δεύτερη ομάδα δηλαδή αποτελείται από πελάτες που - σε συνδυασμό με το μικρότερο μέσο εισόδημα - δείχνουν μία τάση προς το να μην πληρώσουν τελικά τις οφειλές τους.

Στον πίνακα 32 βλέπουμε ότι όλες οι τιμές του στατιστικού Λ του Wilks είναι πολύ κοντά στο 1. Για παράδειγμα βλέπουμε ότι υπάρχει σημαντική διαφορά (p<0.001)της τελευταίας πληρωμής (*last_paym*) των πελατών της πρώτης ομάδας από αυτήν των πελατών της δεύτερης.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
WO_amount	,995	996,874	1	200698	,000
Sum_Paym	,996	816,732	1	200698	,000
Incr_Bal	,999	179,160	1	200698	,000
last_paym	,793	52359,766	1	200698	,000
sum_delq_months	,981	3844,070	1	200698	,000
months_since_wo	,978	4555,884	1	200698	,000
property_flag	,993	1387,792	1	200698	,000
income	1,000	50,158	1	200698	,000

Πίνακας 32: Έλεγχοι ισότητας μέσων

Ο πίνακας 33 δείχνει τις τιμές των συντελεστών της διακριτικής συνάρτησης. Από τις τιμές μπορούμε να εντοπίσουμε τις μεταβλητές που συνεισφέρουν περισσότερο στη διάκριση, αφού όσο μεγαλύτερη είναι η (απόλυτη) τιμή του συντελεστή τόσο μεγαλύτερη είναι η συνεισφορά της αντίστοιχης μεταβλητής στη διάκριση. Άρα εδώ η μεταβλητή που συμβάλλει περισσότερο είναι η *last_paym* και η αμέσως επόμενη είναι η *months_since_wo.* Η μεταβλητή που συμβάλλει λιγότερο είναι η *income*. Αυτό μπορούμε να το δούμε και στον πίνακα των μεταβλητών που εισέρχονται στην ανάλυση (Παράρτημα: Variables in the Analysis) αφού είναι η μεταβλητή που δεν εισέρχεται.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
WO_amount	,139
Sum_Paym	-,083
Incr_Bal	-,153
last_paym	,951
sum_delq_months	,089
months_since_wo	,178
property_flag	-,077

Πίνακας 33: Τυποποιημένοι συντελεστές διακριτικής συνάρτησης.

Στον πίνακα δομής (πίνακας 34) παίρνουμε ένα διαφορετικό μέτρο της συνεισφοράς κάθε μεταβλητής στην διακριτική συνάρτηση. Οι μεταβλητές είναι τοποθετημένες με τη σειρά ανάλογα με το μέγεθος συνεισφοράς. Το αρνητικό πρόσημο δείχνει ότι η συγκεκριμένη μεταβλητή συσχετίζεται αρνητικά με την τιμή της συνάρτησης ενώ το θετικό πρόσημο το αντίθετο. Για παράδειγμα η μεταβλητή *Sum_paym* έχει αρνητική συσχέτιση με την τιμή της συνάρτησης. Αυτό συμβαίνει επειδή όσο περισσότερο έχει πληρώσει ένας πελάτης πρόσφατα τόσο πιο πιθανό είναι να ανήκει στους «καλούς» πελάτες (*GB*=0).

Structure Matrix

	Function
	1
last_paym	,927
months_since_wo	,273
sum_delq_months	,251
property_flag	-,151
WO_amount	,128
Sum_Paym	-,116
Incr_Bal	-,054
income[a]	-,022

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

a. This variable not used in the analysis.

Πίνακας 34: Πίνακας δομής

Στον πίνακα 35 βλέπουμε τη μέση τιμή της διακριτικής συνάρτησης για κάθε ομάδα. Η μέση τιμή για την πρώτη ομάδα είναι αρνητική ενώ για τη δεύτερη θετική, έτσι ώστε η συνάρτηση να κάνει διάκριση ανάμεσα στις δύο ομάδες πελατών.

Functions at Group Centroids

GB	Function 1
0	-1,699
1	,179

Unstandardized canonical discriminant functions evaluated at group means

Πίνακας 35: Μέση τιμή συνάρτησης για κάθε ομάδα

3.3.2 Ταξινόμηση

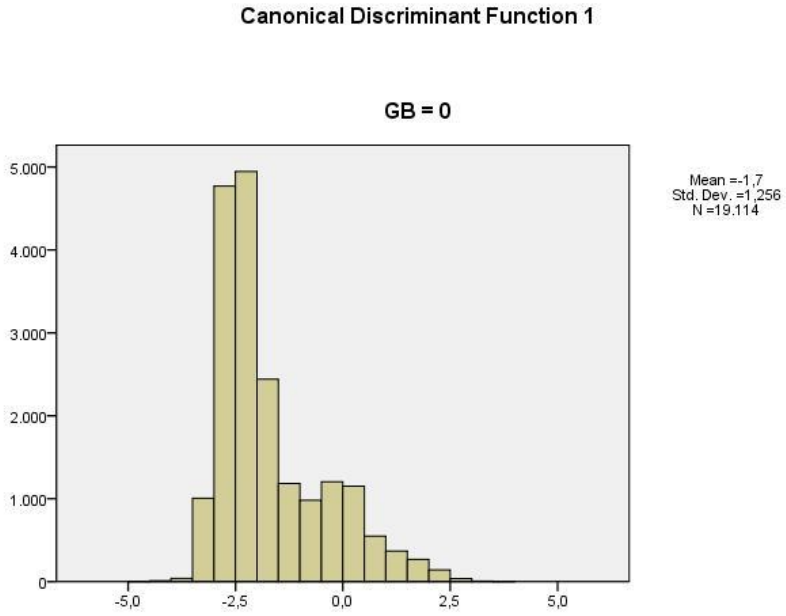
Οι εκ των προτέρων πιθανότητες να ανήκει μία παρατήρηση σε μία από τις δύο ομάδες φαίνονται στον πίνακα 36. Μία παρατήρηση δηλαδή έχει μεγαλύτερη πιθανότητα να ανήκει στη δεύτερη ομάδα απ' ότι στην πρώτη.

Prior Probabilities for Groups

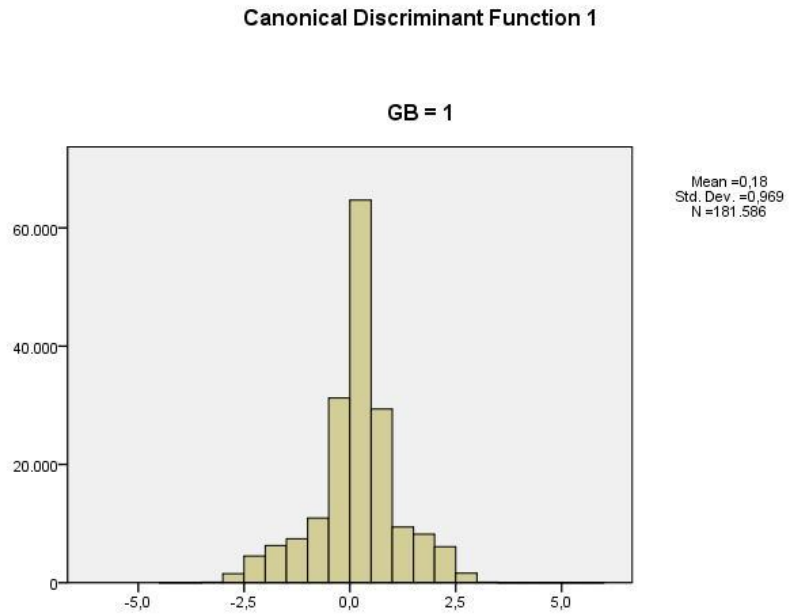
GB	Prior	Cases Used in Analysis	
		Unweighted	Weighted
0	,095	19114	19114,000
1	,905	181586	181586,000
Total	1,000	200700	200700,000

Πίνακας 36: Εκ των προτέρων πιθανότητες για κάθε ομάδα

Οι εικόνες δείχνουν τα γραφήματα της συνάρτησης για κάθε ομάδα. Παρατηρούμε ότι η διασπορά των τιμών της συνάρτησης διαφέρει μεταξύ των πελατών της πρώτης ομάδας και αυτών της δεύτερης.



Εικόνα: Γράφημα διακριτικής συνάρτησης για την 1η ομάδα



Εικόνα: Γράφημα διακριτικής συνάρτησης για τη 2η ομάδα

Στο παράθυρο με τα δεδομένα έχουν δημιουργηθεί κάποιες καινούριες μεταβλητές.

dataset_1.sav [DataSet1] - SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : Dis_1 Visible:

	Marital_status	income	pay6m	ZIP_area	Dis_1	Dis1_1	Dis1_2	Dis2_2
7		1000,00	99,24	East Athens	0	-2,43415	0,70934	0,29066
8		1167,00	0,00	Pireaus Area	0	-2,41757	0,70288	0,29712
9		1000,00	0,00	East Athens	0	-2,40845	0,69929	0,30071
10		0,00	2188,65	Pireaus Area	1	1,60664	0,00124	0,99876
11		0,00	2188,65	Pireaus Area	1	2,01035	0,00058	0,99942
12		0,00	2188,65	Pireaus Area	1	2,41406	0,00027	0,99973
13		0,00	0,00	West Athens	1	0,50250	0,00974	0,99026
14		0,00	0,00	West Athens	1	0,52380	0,00936	0,99064
15		0,00	0,00	West Athens	1	0,54510	0,00900	0,99100
16		0,00	0,00	West Athens	1	0,56640	0,00865	0,99135
17		0,00	0,00	West Athens	1	0,58769	0,00831	0,99169
18		0,00	2188,65	Pireaus Area	1	0,79621	0,00563	0,99437
19		0,00	2188,65	Pireaus Area	1	1,19992	0,00265	0,99735
20		0,00	0,00	Thessalia	0	-2,21761	0,61907	0,38093
21		1000,00	200,00	East Athens	1	-1,53287	0,31000	0,69000
22		1000,00	200,00	East Athens	1	-1,92199	0,48264	0,51736
23		0,00	0,00	Pireaus Area	1	0,56969	0,00859	0,99141
24		2000,00	481,04	Attica Region	0	-1,99195	0,51547	0,48453
25		0,00	0,00	Pireaus Area	1	-1,41800	0,26584	0,73416
26		0,00	0,00	Pireaus Area	1	0,93177	0,00437	0,99563
27		2857,00	0,00	West Athens	1	-0,15703	0,03281	0,96719
28		2857,00	0,00	West Athens	1	0,28944	0,01446	0,98554
29		0,00	0,00	East Athens	1	-1,00396	0,14267	0,85733
30		0,00	1121,25	Thessaloniki	1	-1,86274	0,45494	0,54506

Εικόνα: Δεδομένα – καινούριες μεταβλητές

Οι μεταβλητές είναι:

- *Dis_1:* Δίνει την προβλεπόμενη ομάδα όπου θα ανήκει η παρατήρηση.
- *Dis1_1:* Δημιουργείται από τη βηματική μέθοδο και δίνει την τιμή της διακριτικής συνάρτησης.
- *Dis1_2:* Προβλεπόμενη πιθανότητα να ανήκει η παρατήρηση στην πρώτη ομάδα.
- *Dis2_2:* Προβλεπόμενη πιθανότητα να ανήκει η παρατήρηση στη δεύτερη ομάδα.

Ο πίνακας 37 δίνει τις τιμές των συντελεστών της συνάρτησης ταξινόμησης για κάθε ομάδα. Η ταξινόμηση μιας παρατήρησης γίνεται ως εξής: υπολογίζεται η τιμή της συνάρτησης ταξινόμησης της παρατήρησης (διακριτικός βαθμός) για κάθε ομάδα και επιλέγουμε την ομάδα για την οποία έχει προκύψει η μεγαλύτερη τιμή.

Classification Function Coefficients

	GB	
	0	1
WO_amount	-4,647E-5	-2,750E-5
Sum_Paym	,000	,000
Incr_Bal	1,214	1,106
last_paym	,170	,931
sum_delq_months	,091	,097
months_since_wo	,300	,340
property_flag	2,984	2,687
(Constant)	-8,915	-10,708

Fisher's linear discriminant functions

Πίνακας 37: Συντελεστές συνάρτησης ταξινόμησης

Τέλος, στον πίνακα των αποτελεσμάτων της ταξινόμησης (πίνακας 38) δίνονται πληροφορίες για την καταλληλότητα της διακριτικής συνάρτησης. Στον πίνακα αυτό γίνεται διασταύρωση της μεταβλητής GB σε σχέση με την ταξινόμηση που προκύπτει από την διακριτική συνάρτηση. Σε 11056 περιπτώσεις που ανήκουν στην πρώτη ομάδα έγινε σωστή πρόβλεψη, ενώ σε 8058 λανθασμένη. Και για τη δεύτερη ομάδα έγινε σε 174981 περιπτώσεις σωστή πρόβλεψη και σε 6605 λανθασμένη. Η λανθασμένη πρόβλεψη στην περίπτωση που μία παρατήρηση ανήκει στην πρώτη ομάδα οφείλεται πιθανώς στο μικρό πλήθος παρατηρήσεων που ανήκει σε αυτήν σε σχέση με το συνολικό πλήθος παρατηρήσεων. Για τον ίδιο λόγο (μεγάλος αριθμός παρατηρήσεων στην $2^{η}$ ομάδα) είναι τόσο ακριβής η πρόβλεψη στην περίπτωση που μία παρατήρηση ανήκει στην δεύτερη ομάδα. Συνολικά, όπως δείχνει και ο πίνακας, έχει γίνει σωστή ταξινόμηση 92.7% των παρατηρήσεων.

Classification Results[a]

			Predicted Group Membership		
		GB	0	1	Total
Original	Count	0	11056	8058	19114
		1	6605	174981	181586
	%	0	57,8	42,2	100,0
		1	3,6	96,4	100,0

a. 92,7% of original grouped cases correctly classified.

Πίνακας 38: Αποτελέσματα ταξινόμησης

3.4 Συμπεράσματα

Στο δείγμα των 200.701 παρατηρήσεων η ανάλυση κατά συστάδες δημιούργησε 3 συστάδες οι οποίες χαρακτηρίζουν τους πελάτες βάσει της συμπεριφοράς τους όσον αφορά στα χρέη που έχουν στην τράπεζα. Η πρώτη ομάδα αποτελείται από τον μεγαλύτερο αριθμό πελατών και είναι εκείνη της οποίας οι πελάτες έχουν ελάχιστη πιθανότητα αποπληρωμής τουλάχιστον του 2% του χρέους τους. Έχουν περάσει περισσότεροι μήνες από την πρώτη διαγραφή απ' ότι στις δύο άλλες συστάδες ενώ έχουν και πολύ μεγάλο άθροισμα buckets μέσα στους τελευταίους 12 μήνες. Οι πελάτες της δεύτερης συστάδας έχουν «μέτρια» συμπεριφορά, καθώς το άθροισμα των buckets τον τελευταίο χρόνο είναι πολύ μικρότερο από εκείνο της προηγούμενης ομάδας όπως και ο μέσος αριθμός των μηνών που έχουν περάσει από την πρώτη διαγραφή. Τέλος, οι πελάτες της τρίτης συστάδας είναι εκείνοι με την «καλύτερη» συμπεριφορά αφού οι πληρωμές που έχουν κάνει το τελευταίο τρίμηνο είναι μεγαλύτερες από του προηγούμενου τριμήνου, έχουν μικρότερο μέσο πλήθος μηνών από την πρώτη διαγραφή και όλοι έχουν πληρώσει κάποια δόση μέσα στους τελευταίους 3 μήνες, σε αντίθεση με τους πελάτες των προηγούμενων συστάδων. Οι σημαντικές μεταβλητές για την ανάλυση ήταν οι *Sum_Paym, last_paym, sum_delq_months* και *months_since_wo*.

Στην διακριτική ανάλυση επιλέξαμε να γίνει ο διαχωρισμός με βάση τη μεταβλητή *GB,* δηλαδή η πρώτη ομάδα περιλαμβάνει τους πελάτες για τους οποίους ισχύει *GB* = 0, δηλαδή θα αποπληρώσουν ποσοστό μικρότερο ή ίσο του 2% του χρέους τους, ενώ η δεύτερη εκείνους για τους οποίους ισχύει *GB* = 1 και θα πληρώσουν μεγαλύτερο ποσοστό του χρέους τους. Τα αποτελέσματα της ανάλυσης θέτουν την μεταβλητή *last_paym* ως τη μεταβλητή που συμβάλλει περισσότερο στη διάκριση μεταξύ των ομάδων, ενώ ακολουθούν οι *monhs_since_wo* και *sum_delq_months*. Η μεταβλητή *income* δεν εισάγεται στη συνάρτηση ενώ από αυτές που εισάγονται τη μικρότερη συμβολή στη διάκριση παρέχει η *Incr_Bal*. Και σε αυτή την ανάλυση, δηλαδή, οι ίδιες μεταβλητές συμβάλλουν περισσότερο στη διάκριση μεταξύ των ομάδων με μόνη διαφορά την *Sum_Paym* που η συμβολή της είναι μεγαλύτερη στην ανάλυση κατά συστάδες, ενώ στην διακριτική ανάλυση η συμβολή της είναι πολύ μικρή. Επίσης η συνάρτηση ταξινόμησης που έδωσε η διακριτική ανάλυση είναι αρκετά ακριβής αφού ταξινομήθηκε σωστά το 92.7% των παρατηρήσεων. Το μεγάλο αυτό ποσοστό οφείλεται κυρίως στους πελάτες για τους οποίους ισχύει *GB* = 1 ($2^{η}$ ομάδα) αφού στην περίπτωσή τους έγινε σωστή ταξινόμηση του 96.4% των παρατηρήσεων ενώ στην ομάδα όπου *GB* = 0 το ποσοστό της σωστής ταξινόμησης είναι μόλις 57.8%. Λόγω του μεγάλου πλήθους παρατηρήσεων που ανήκουν στην πρώτη ομάδα σε σχέση με αυτό της δεύτερης ομάδας, το χαμηλό ποσοστό της σωστής ταξινόμησης στη δεύτερη δεν επηρεάζει σημαντικά το συνολικό ποσοστό. Θα μπορούσαμε να πούμε ότι η διακριτική ανάλυση συμπεριφέρεται καλύτερα από την ανάλυση κατά συστάδες, αφού η διάκριση των ομάδων που προκύπτουν είναι πιο σαφής. Παρόλα αυτά μπορούν και οι δύο να μας βοηθήσουν εξίσου στην ταξινόμηση μίας νέας παρατήρησης, η ανάλυση κατά συστάδες υπολογίζοντας τις αποστάσεις της παρατήρησης από τα κέντρα των συστάδων και επιλέγοντας τη συστάδα από το κέντρο της οποίας η απόσταση είναι μικρότερη και η διακριτική ανάλυση υπολογίζοντας την τιμή της συνάρτησης ταξινόμησης (διακριτικός βαθμός) για κάθε ομάδα και επιλέγοντας εκείνη με την μεγαλύτερη τιμή.

Παράρτημα

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
GB	200701	0	1	,90	,294
WO_amount	200701	,04	91524,45	13329,4645	13821,20761
Sum_Paym	200700	-57990,00	42538,40	-256,6456	1222,86947
Incr_Bal	200701	0	11	2,30	2,637
last_paym	200701	0	12	6,10	2,635
sum_delq_months	200701	0	108	72,70	27,444
months_since_wo	200701	0	32	12,56	8,464
property_flag	200701	0	1	,53	,491
income	200701	,00	316956,00	792,6038	3706,84965
pay6m	200701	,00	41157,45	135,1199	859,35424
Valid N (listwise)	200700				

Cluster Analysis – K-means, k=2

Initial Cluster Centers

	Cluster	
	1	2
sum_delq_months	46	56
Sum_Paym	-57990,00	42538,40
last_paym	9	3
months_since_wo	2	14

Iteration History[a]

Iteration	Change in Cluster …	
	1	2
1	42002,914	42762,691
2	634,444	1,151
3	510,279	,884
4	296,531	,497
5	257,039	,425
6	267,560	,433
7	427,901	,658
8	240,091	,359
9	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 9. The minimum distance between initial centers is 100528,401.

Final Cluster Centers

	Cluster	
	1	2
sum_delq_months	42	73
Sum_Paym	-18620,93	-228,70
last_paym	7	6
months_since_wo	9	13

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
sum_delq_months	283597,841	1	751,754	200698	377,248	,000
Sum_Paym	1,030E11	1	982124,222	200698	104891,896	,000
last_paym	551,177	1	6,938	200698	79,445	,000
months_since_wo	3538,284	1	71,624	200698	49,401	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	305,000
	2	200395,000
Valid		200700,000
Missing		1,000

Crosstabulation GB - Cluster Number of Case QCL_1

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
GB * Cluster Number of Case	200700	100,0%	1	,0%	200701	100,0%

GB * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case		
		1	2	Total
GB	0	52	19062	19114
	1	253	181333	181586
Total		305	200395	200700

Cluster Analysis – K-means, k=3

Initial Cluster Centers

	Cluster		
	1	2	3
sum_delq_months	9	46	56
Sum_Paym	-7739,00	-57990,00	42538,40
last_paym	9	9	3
months_since_wo	9	2	14

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	7484,496	12979,997	18398,757
2	,268	7447,440	5214,074
3	,617	5903,208	2767,526
4	1,012	4440,105	2062,025
5	,932	3007,701	1527,356
6	,063	1466,370	964,543
7	,210	1253,649	677,583
8	,353	868,283	400,308
9	,338	570,887	212,339
10	,053	324,305	191,317
11	,029	243,002	145,384
12	,065	135,129	59,315
13	,118	100,282	11,785
14	,067	98,703	35,251
15	,234	192,322	23,301
16	,279	216,361	23,140
17	,093	60,763	,000
18	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 18. The minimum distance between initial centers is 50251,014.

Final Cluster Centers

	Cluster		
	1	2	3
sum_delq_months	73	42	47
Sum_Paym	-250,71	-18681,50	9824,40
last_paym	6	7	1
months_since_wo	13	9	7

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
sum_delq_months	281137,412	2	750,369	200697	374,665	,000
Sum_Paym	7,359E10	2	762091,705	200697	96561,686	,000
last_paym	5133,858	2	6,889	200697	745,173	,000
months_since_wo	7975,136	2	71,562	200697	111,443	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	199961,000
	2	303,000
	3	436,000
Valid		200700,000
Missing		1,000

Crosstabulation GB - Cluster Number of Case QCL_2

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
GB * Cluster Number of Case	200700	100,0%	1	,0%	200701	100,0%

GB * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case			Total
		1	2	3	
GB	0	18843	52	219	19114
	1	181118	251	217	181586
Total		199961	303	436	200700

Cluster Analysis – K-means, k=4

Initial Cluster Centers

	Cluster			
	1	2	3	4
sum_delq_months	0	56	20	46
Sum_Paym	-25377,06	42538,40	8434,73	-57990,00
last_paym	12	3	2	9
months_since_wo	16	14	1	2

Iteration History[a]

Iteration	Change in Cluster Centers			
	1	2	3	4
1	10286,719	9703,241	8663,820	8588,443
2	1773,354	9266,047	1,278	4391,557
3	1577,608	4924,015	,270	3080,912
4	1358,736	2579,645	1,718	1172,881
5	1149,083	1999,989	2,052	1612,561
6	845,590	1496,818	2,112	500,605
7	672,874	964,543	2,531	,000
8	669,853	677,583	3,765	,000
9	538,521	400,308	4,154	,000
10	482,753	212,339	4,354	1080,489
11	461,871	90,433	5,295	957,015
12	513,951	150,034	6,741	2059,847
13	477,973	84,287	8,212	1705,021
14	427,671	47,611	9,228	1593,642
15	334,742	11,824	9,092	922,247
16	308,599	,000	9,422	1485,132
17	246,320	,000	8,851	1178,488
18	201,139	,000	7,757	1406,728
19	185,324	,000	8,007	1243,964
20	132,354	11,824	6,450	798,936
21	113,723	23,771	6,557	296,012
22	83,258	,000	5,114	173,519
23	55,637	,000	3,675	,000
24	42,513	,000	2,811	57,956
25	22,975	,000	1,584	,000
26	12,254	,000	,846	,000
27	8,057	,000	,559	,000
28	5,888	,000	,410	,000
29	9,741	,000	,683	,000
30	4,911	,000	,347	,000
31	1,805	,000	,127	,000
32	5,050	,000	,357	,000
33	1,619	,000	,115	,000
34	,426	,000	,030	,000
35	,000	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 35. The minimum distance between initial centers is 32612,976.

Final Cluster Centers

	Cluster			
	1	2	3	4
sum_delq_months	56	47	74	42
Sum_Paym	-2364,28	9965,30	-107,52	-23684,07
last_paym	6	1	6	7
months_since_wo	5	7	13	9

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
sum_delq_months	1546644,935	3	730,055	200696	2118,531	,000
Sum_Paym	6,924E10	3	460414,635	200696	150388,981	,000
last_paym	3533,577	3	6,888	200696	513,015	,000
months_since_wo	242966,587	3	68,010	200696	3572,504	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	13243,000
	2	424,000
	3	186850,000
	4	183,000
Valid		200700,000
Missing		1,000

Crosstabulation GB - Cluster Number of Case QCL_3

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
GB * Cluster Number of Case	200700	100,0%	1	,0%	200701	100,0%

GB * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case				
		1	2	3	4	Total
GB	0	926	212	17945	31	19114
	1	12317	212	168905	152	181586
Total		13243	424	186850	183	200700

Cluster Analysis – K-means, k=5

Initial Cluster Centers

	Cluster				
	1	2	3	4	5
sum_delq_months	51	46	63	70	56
Sum_Paym	-6845,22	-57990,00	-31300,00	17633,75	42538,40
last_paym	10	9	11	1	3
months_since_wo	3	2	5	17	14

Iteration History[a]

Iteration	Change in Cluster Centers				
	1	2	3	4	5
1	6582,389	7853,450	4739,039	7710,849	6133,447
2	5,605	1784,101	6612,495	992,852	5437,750
3	3,786	3002,667	3872,877	860,179	3107,996
4	3,738	1116,454	2277,885	619,815	2938,718
5	2,813	2595,019	1680,422	425,403	1493,381
6	3,263	882,106	1347,742	394,551	1140,771
7	3,737	1612,561	1215,232	224,325	343,641
8	4,218	500,605	953,976	185,942	445,067
9	3,964	,000	722,924	176,175	416,786
10	4,075	,000	604,074	144,011	566,994
11	4,677	,000	568,021	85,543	181,502
12	4,361	1080,489	491,194	88,162	175,026
13	5,157	957,015	477,219	65,682	251,787
14	6,619	1873,507	524,511	60,216	239,495
15	8,021	1733,963	496,804	54,971	228,028
16	9,030	1338,725	437,808	8,751	,000
17	9,143	816,164	354,475	,000	,000
18	8,466	1005,494	287,185	4,374	,000
19	8,268	1204,759	248,950	4,377	,000
20	8,055	1327,233	218,783	,000	,000
21	7,199	1559,879	185,828	,000	,000
22	7,045	1111,578	159,250	,000	,000
23	6,948	422,701	130,729	8,741	,000
24	5,648	296,012	97,856	,000	,000
25	4,233	173,519	68,968	,000	,000
26	3,141	57,956	48,250	4,369	,000
27	2,708	,000	40,016	4,381	,000
28	1,430	,000	20,824	,000	,000
29	,722	,000	10,500	,000	,000
30	,325	,000	4,721	,000	,000
31	,419	,000	5,849	4,382	,000
32	,150	,000	1,941	4,393	,000
33	,140	,000	2,026	,000	,000
34	,104	,000	1,494	,000	,000
35	,091	,000	1,316	,000	,000
36	,018	,000	,263	,000	,000
37	,006	,000	,088	,000	,000
38	,000	,000	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 38. The minimum distance between initial centers is 24454,783.

Final Cluster Centers

	Cluster				
	1	2	3	4	5
sum_delq_months	74	42	56	50	52
Sum_Paym	-115,56	-23684,07	-2388,53	5571,39	19438,06
last_paym	6	7	6	1	1
months_since_wo	13	9	5	8	8

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
sum_delq_months	1173511,363	4	729,789	200695	1608,014	,000
Sum_Paym	5,509E10	4	397544,693	200695	138564,299	,000
last_paym	4538,379	4	6,850	200695	662,511	,000
months_since_wo	179205,696	4	68,071	200695	2632,641	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	186817,000
	2	183,000
	3	12962,000
	4	647,000
	5	91,000
Valid		200700,000
Missing		1,000

Crosstabulation GB - Cluster Number of Case QCL_4

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
GB * Cluster Number of Case	200700	100,0%	1	,0%	200701	100,0%

GB * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case					
		1	2	3	4	5	Total
GB	0	17776	31	907	350	50	19114
	1	169041	152	12055	297	41	181586
Total		186817	183	12962	647	91	200700

Descriptives of split data

Cluster 1

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Sum_Paym	199961	-9444,00	4759,27	-250,7075	710,33789
last_paym	199961	0	12	6,11	2,626
sum_delq_months	199961	0	108	72,80	27,395
months_since_wo	199961	0	32	12,58	8,465
Valid N (listwise)	199961				

a. Cluster Number of Case = 1

Cluster 2

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Sum_Paym	303	-57990,00	-9499,31	-18681,5040	10859,68804
last_paym	303	0	12	7,44	3,427
sum_delq_months	303	0	108	42,44	29,417
months_since_wo	303	0	30	9,12	7,594
Valid N (listwise)	303				

a. Cluster Number of Case = 2

Cluster 3

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Sum_Paym	436	4801,00	42538,40	9824,4045	6147,23652
last_paym	436	0	3	1,38	,993
sum_delq_months	436	0	108	47,24	24,870
months_since_wo	436	0	28	7,25	6,403
Valid N (listwise)	436				

a. Cluster Number of Case = 3

Διακριτική Ανάλυση

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		200700	100,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	1	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Total	1	,0
Total		200701	100,0

Group Statistics

GB		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
0	WO_amount	10334,5681	10577,74353	19114	19114,000
	Sum_Paym	-16,6901	1702,37612	19114	19114,000
	Incr_Bal	2,5400	2,33329	19114	19114,000
	last_paym	2,4057	3,02055	19114	19114,000
	sum_delq_months	61,1066	25,13120	19114	19114,000
	months_since_wo	8,6741	7,57615	19114	19114,000
	property_flag	,6585	,47372	19114	19114,000
	income	973,2058	3281,08706	19114	19114,000
1	WO_amount	13644,7285	14082,40029	181586	181586,000
	Sum_Paym	-281,9037	1158,04945	181586	181586,000
	Incr_Bal	2,2717	2,66585	181586	181586,000
	last_paym	6,4881	2,26354	181586	181586,000
	sum_delq_months	73,9233	27,39198	181586	181586,000
	months_since_wo	12,9699	8,44884	181586	181586,000
	property_flag	,5198	,49114	181586	181586,000
	income	773,5977	3748,36591	181586	181586,000
Total	WO_amount	13329,4798	13821,24033	200700	200700,000
	Sum_Paym	-256,6456	1222,86947	200700	200700,000
	Incr_Bal	2,2973	2,63716	200700	200700,000
	last_paym	6,0993	2,63450	200700	200700,000
	sum_delq_months	72,7027	27,44382	200700	200700,000
	months_since_wo	12,5608	8,46410	200700	200700,000
	property_flag	,5330	,49120	200700	200700,000
	income	792,6077	3706,85846	200700	200700,000

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
WO_amount	,995	996,874	1	200698	,000
Sum_Paym	,996	816,732	1	200698	,000
Incr_Bal	,999	179,160	1	200698	,000
last_paym	,793	52359,766	1	200698	,000
sum_delq_months	,981	3844,070	1	200698	,000
months_since_wo	,978	4555,884	1	200698	,000
property_flag	,993	1387,792	1	200698	,000
income	1,000	50,158	1	200698	,000

Box's Test of Equality of Covariance Matrices

Log Determinants

GB	Rank	Log Determinant
0	7	45,586
1	7	45,023
Pooled within-groups	7	45,165

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		17782,094
F	Approx.	634,992
	df1	28
	df2	3,841E9
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Stepwise Statistics

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda							
						Exact F			
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	last_paym	,793	1	1	200698,000	52359,766	1	200698,000	,000
2	months_ since_wo	,776	2	1	200698,000	28903,824	2	200697,000	,000
3	WO_amount	,772	3	1	200698,000	19738,082	3	200696,000	,000
4	Incr_Bal	,770	4	1	200698,000	14990,380	4	200695,000	,000
5	property_flag	,769	5	1	200698,000	12067,039	5	200694,000	,000
6	Sum_Paym	,768	6	1	200698,000	10101,423	6	200693,000	,000
7	sum_delq_ months	,767	7	1	200698,000	8710,172	7	200692,000	,000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 16.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	last_paym	1,000	52359,766	
2	last_paym	,999	52069,793	,978
	months_since_wo	,999	4320,877	,793
3	last_paym	,999	52061,813	,972
	months_since_wo	,998	4423,866	,789
	WO_amount	,999	1092,273	,776
4	last_paym	,913	50682,647	,964
	months_since_wo	,503	886,153	,773
	WO_amount	,976	1316,054	,775
	Incr_Bal	,470	577,252	,772
5	last_paym	,905	49025,552	,957
	months_since_wo	,502	872,357	,772
	WO_amount	,944	1502,204	,775
	Incr_Bal	,469	543,846	,771
	property_flag	,952	287,945	,770
6	last_paym	,900	48001,766	,952
	months_since_wo	,500	931,704	,772
	WO_amount	,921	1295,468	,773
	Incr_Bal	,466	592,626	,770
	property_flag	,951	298,182	,769
	Sum_Paym	,924	210,396	,769
7	last_paym	,893	46565,596	,945
	months_since_wo	,482	719,007	,770
	WO_amount	,785	715,068	,770
	Incr_Bal	,461	504,978	,769
	property_flag	,948	266,027	,768
	Sum_Paym	,900	287,969	,768
	sum_delq_months	,745	278,777	,768

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	WO_amount	1,000	1,000	996,874	,995
	Sum_Paym	1,000	1,000	816,732	,996
	Incr_Bal	1,000	1,000	179,160	,999
	last_paym	1,000	1,000	52359,766	,793
	sum_delq_months	1,000	1,000	3844,070	,981
	months_since_wo	1,000	1,000	4555,884	,978
	property_flag	1,000	1,000	1387,792	,993
	income	1,000	1,000	50,158	1,000
1	WO_amount	1,000	1,000	990,941	,789
	Sum_Paym	,991	,991	34,834	,793
	Incr_Bal	,949	,949	3565,311	,779
	sum_delq_months	,999	,999	2424,914	,784
	months_since_wo	,999	,999	4320,877	,776
	property_flag	,992	,992	239,583	,792
	income	1,000	1,000	14,057	,793
2	WO_amount	,999	,998	1092,273	,772
	Sum_Paym	,958	,958	319,835	,775
	Incr_Bal	,481	,481	354,286	,775
	sum_delq_months	,883	,883	798,896	,773
	property_flag	,989	,989	129,679	,776
	income	,999	,999	4,590	,776
3	Sum_Paym	,929	,929	150,933	,772
	Incr_Bal	,470	,470	577,252	,770
	sum_delq_months	,777	,777	318,859	,771
	property_flag	,953	,953	321,307	,771
	income	,999	,998	7,850	,772
4	Sum_Paym	,924	,467	200,163	,769
	sum_delq_months	,766	,463	229,538	,769
	property_flag	,952	,469	287,945	,769
	income	,999	,470	10,154	,770
5	Sum_Paym	,924	,466	210,396	,768
	sum_delq_months	,764	,463	201,208	,768
	income	,994	,469	4,075	,769
6	sum_delq_months	,745	,461	278,777	,767
	income	,994	,466	3,940	,768
7	income	,993	,461	2,131	,767

Wilks' Lambda

Step	Number of Variables	Lambda	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	1	,793	1	1	200698	52359,766	1	200698,000	,000
2	2	,776	2	1	200698	28903,824	2	200697,000	,000
3	3	,772	3	1	200698	19738,082	3	200696,000	,000
4	4	,770	4	1	200698	14990,380	4	200695,000	,000
5	5	,769	5	1	200698	12067,039	5	200694,000	,000
6	6	,768	6	1	200698	10101,423	6	200693,000	,000
7	7	,767	7	1	200698	8710,172	7	200692,000	,000

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,304[a]	100,0	100,0	,483

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,767	53241,603	7	,000

Standardized Canonical Discriminant Function Coefficients

	Function
	1
WO_amount	,139
Sum_Paym	-,083
Incr_Bal	-,153
last_paym	,951
sum_delq_months	,089
months_since_wo	,178
property_flag	-,077

Structure Matrix

	Function
	1
last_paym	,927
months_since_wo	,273
sum_delq_months	,251
property_flag	-,151
WO_amount	,128
Sum_Paym	-,116
Incr_Bal	-,054
income[a]	-,022

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

a. This variable not used in the analysis.

Canonical Discriminant Function Coefficients

	Function
	1
WO_amount	,000
Sum_Paym	,000
Incr_Bal	-,058
last_paym	,405
sum_delq_months	,003
months_since_wo	,021
property_flag	-,158
(Constant)	-2,914

Unstandardized coefficients

Functions at Group Centroids

	Function
GB	1
0	-1,699
1	,179

Unstandardized canonical discriminant functions evaluated at group means

Classification Statistics

Classification Processing Summary

Processed		200701
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	1
Used in Output		200700

Prior Probabilities for Groups

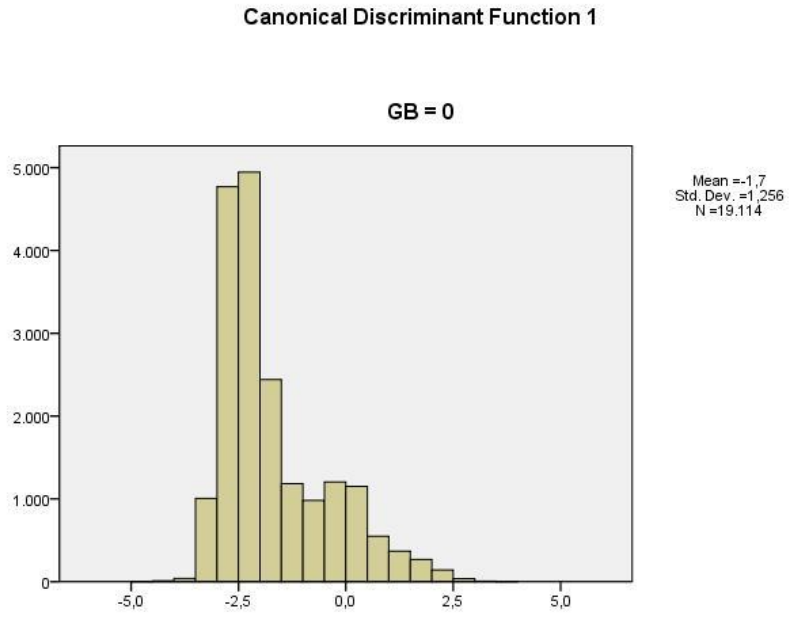
		Cases Used in Analysis	
GB	Prior	Unweighted	Weighted
0	,095	19114	19114,000
1	,905	181586	181586,000
Total	1,000	200700	200700,000

Classification Function Coefficients

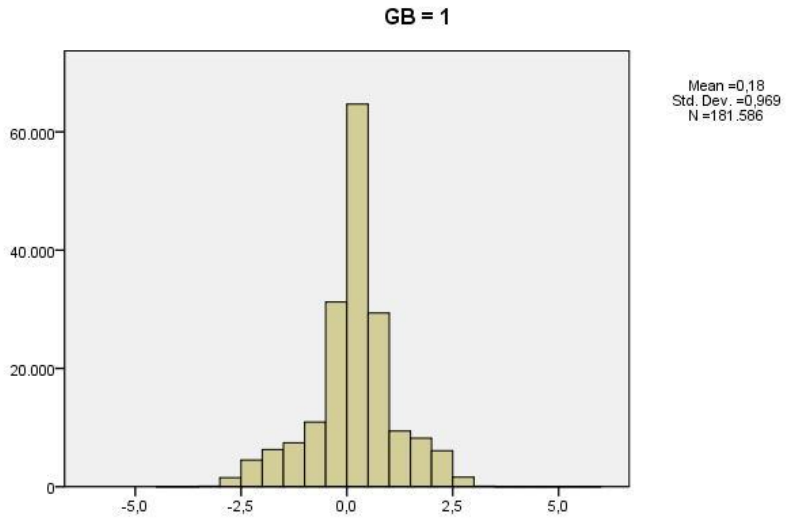
	GB	
	0	1
WO_amount	-4,647E-5	-2,750E-5
Sum_Paym	,000	,000
Incr_Bal	1,214	1,106
last_paym	,170	,931
sum_delq_months	,091	,097
months_since_wo	,300	,340
property_flag	2,984	2,687
(Constant)	-8,915	-10,708

Fisher's linear discriminant functions

Separate-Groups Graphs



Canonical Discriminant Function 1



Classification Results[a]

			Predicted Group Membership		
		GB	0	1	Total
Original	Count	0	11056	8058	19114
		1	6605	174981	181586
	%	0	57,8	42,2	100,0
		1	3,6	96,4	100,0

a. 92,7% of original grouped cases correctly classified.

Βιβλιογραφία

[1] Andrews RL, Currim IS (2003) *Recovering and profiling the true segmentation structure in markets: an empirical investigation.* Int J Res Mark 20(2):177–192

[2] Arabie P, Hubert L (1994) *Cluster analysis in marketing research*. In: Bagozzi RP (ed) *Advanced methods in marketing research*. Blackwell, Cambridge, pp 160–189

[3] Bishop CM (2006) *Pattern recognition and machine learning*. Springer, Berlin

[4] Brace, N, R. Kemp, R. Snelgar (2009) *SPSS for Psychologists*, 4th ed., UK: Palgrave Macmillan

[5] Calinski, T, Harabasz J (1974) *A dendrite method for cluster analysis*. Commun Stat Theory Methods 3(1):1–27

[6] Chiu T, Fang D, Chen J, Wang Y, Jeris C (2001) *A robust and scalable clustering algorithm for mixed type attributes in large database environment*. In: Proceedings of the 7th ACM SIGKDD international conference in knowledge discovery and data mining, Association for Computing Machinery, San Francisco, CA, pp 263–268

[7] Cooley, W., P. R. Lohnes (1971) *Multivariate Data Analysis.* New York: Wiley, ch. 9

[8] Dibb S (1999) *Criteria guiding segmentation implementation: reviewing the evidence*. J Strateg Mark 7(2):107–129

[9] Dolnicar S (2003) *Using cluster analysis for market segmentation – typical misconceptions, established methodological weaknesses and some recommendations for improvement*. Australas J Mark Res 11(2):5–12

[10] Dolnicar S, Grun B (2009) *Challenging "factor-cluster segmentation".* J Travel Res 47(1):63–71

[11] Dolnicar S, Lazarevski K (2009) *Methodological reasons for the theory/practice divide in market segmentation*. J Mark Manage 25(3–4):357–373

[12] Fisher, R. A. (1936) *The use of Multiple Measurements in Taxonomic Problems,* Annals of Eugenics, 7 (part 2): 179-188

[13] Formann AK (1984) *Die Latent-Class-Analyse: Einfuhrung in die Theorie und Anwendung*. Beltz, Weinheim

[14] Kaufman L, Rousseeuw PJ (2005) *Finding groups in data. An introduction to cluster analysis.* Wiley, Hoboken, NY

[15] Kohonen T (1982) *Self-organized formation of topologically correct feature maps*. Biol Cybern 43 (1):59–69

[16] Kotler P, Keller KL (2009) *Marketing management*, 13th edn. Pearson Prentice Hall, Upper Saddle River, NJ

[17] Geer J. P. van de (1971) *Introduction to Multivariate Analysis for the Social Sciences,* San Francisco: W. H. Freeman

[18] Klecka, W. R. (1980) *Discriminant Analysis,* Beverly Hills, California: Sage Publications Inc.

[19] Larson JS, Bradlow ET, Fader PS (2005) *An exploratory look at supermarket shopping paths*. Int J Res Mark 22(4):395–414

[20] McLachlan GJ, Peel D (2000) *Finite mixture models*. Wiley, New York, NY

[21] Milligan GW, Cooper M (1985) *An examination of procedures for determining the number of clusters in a data set*. Psychometrika 50(2):159–179

[22] Milligan GW, CooperM (1988) *A study of variable standardization*. J Classification 5(2):181–204

[23] Moroko L, Uncles MD (2009) *Employer branding and market segmentation*. J Brand Manage

[24] Morrison, D. F. (1976) *Multivariate Statistical Methods*, 2nd ed., New York: McGraw-Hill
[25] Okazaki S (2006) *What do we know about Mobile Internet Adopters? A Cluster Analysis.* Inf Manage 43(2):127–141
[26] Punji G Stewart DW (1983) *Cluster analysis in marketing research: review and suggestions for application.* J Mark Res 20(2):134–148
[27] Sheppard A (1996) *The sequence of factor analysis and cluster analysis: differences in segmentation and dimensionality through the use of raw and factor scores," tourism analysis.* Tourism Anal 1(Inaugural Volume):49–57
[28] Tatsuoka, M. M. (1981) *Multivariate Analysis*, New York: Wiley
[29] Tonks DG (2009) *Validity and the design of market segments*. J Mark Manage
[30] Wedel M, Kamakura WA (2000) *Market segmentation: conceptual and methodological foundations*, 2nd edn. Kluwer, Boston, NE
[31] Κουκουβίνος, Χ. (2005) *Γραμμικά Μοντέλα και Σχεδιασμοί,* Εκδόσεις ΕΜΠ
[32] Σιάρδος, Γ. (2005) *Μέθοδοι Πολυμεταβλητής Στατιστικής Ανάλυσης,* Μέρος 2ο, 3η έκδοση, Εκδόσεις Σταμούλης