



Πολυτεχνείο Κρήτης
Σχολή Μηχανικών Παραγωγής και Διοίκησης

Διπλωματική εργασία του Ιδομενέα Ε. Καμηλάκη

**“Ανάλυση ελλιπών δεδομένων σε
έρευνες ικανοποίησης πελατών:
Μια εφαρμογή στον κλάδο της υγείας”**

Επιβλέπων Καθηγητής: Ευάγγελος Γρηγορούδης

Χανιά 2015

Πίνακας Περιεχομένων

Περιεχόμενα	i
Περίληψη	iv
1. Εισαγωγή	1
1.1 Γενικά για το πρόβλημα των ελλιπών δεδομένων	1
1.2 Οι αιτίες του προβλήματος	2
1.3 Εξέλιξη αντιμετώπισης-ιστορική αναδρομή	4
1.4 Δομή και στόχοι της εργασίας	5
2. Ανάλυση Προβλήματος Ελλιπών δεδομένων.....	8
2.1 Στατιστικές έρευνες	8
2.2 Έρευνες ικανοποίησης πελατών	9
2.3 Μεθοδολογία έρευνας ικανοποίησης πελατών	13
2.4 Παραγόμενα δεδομένα και τύποι δεδομένων έρευνας	15
2.5 Αιτίες έλλειψης δεδομένων σε έρευνες ικανοποίησης	17
2.6 Μηχανισμοί έλλειψης δεδομένων	20
3. Αντιμετώπιση του προβλήματος	24
3.1 Αρχική αντιμετώπιση	24
3.2 Παραδοσιακές μέθοδοι	24
3.2.1 Μέθοδος Διαγραφής	25
3.2.2 Διαδικασίες στάθμισης	26
3.2.3 Μέθοδοι καταλογισμού	27
3.3 Σύγχρονες μέθοδοι	30
3.3.1 Πολλαπλός Καταλογισμός (Multiple Imputation)	30
3.3.2 Εκτίμηση Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation) ...	31
3.3.3 Μπεϋσιανές Μέθοδοι (Bayesian Methods)	32
4. Ανάπτυξη και αξιολόγηση αλγόριθμου hot deck	33
4.1 Θεωρητικό υπόβαθρο	33
4.2 Αρχική Προσέγγιση Υπολογισμού Απόστασης	35
4.3 Απόσταση Ordinal Δεδομένων (Μέθοδος Walesiak)	37
4.4 Ερώτηση Ολικής Ικανοποίησης	41
4.5 Κωδικοποίηση Δεδομένων	44

4.6 Εισαγωγή Συντελεστή Καταλογισμού	49
4.7 Λειτουργία Αλγορίθμου Hot Deck	56
4.8 Προγραμματισμός σε Java	58
4.9 Λειτουργία προγράμματος	60
5. Έλεγχος και Εφαρμογή αλγόριθμου hotdeck	69
5.1 Έλεγχος και επαλήθευση αλγορίθμου	69
5.2 Εφαρμογή σε έρευνα ικανοποίησης	74
5.2.1 Γενικά στοιχεία έρευνας	75
5.2.2 Απόδοση τιμών και εφαρμογή μεθόδου MUSA	77
5.2.3 Αποτελέσματα	85
5.2.4 Συμπεράσματα	91
6. Επίλογος	94
6.1 Σύνοψη	94
6.2 Μελλοντικές επεκτάσεις	97
ΠΑΡΑΡΤΗΜΑΤΑ	100
Παράρτημα Ι “ΒΙΒΛΙΟΓΡΑΦΙΑ”	100
Παράρτημα ΙΙ “Confusion Tables τεκμηρίωσης αλγόριθμου Hot Deck”	104
Παράρτημα ΙΙΙ “Ερωτηματολόγιο Έρευνας Νοσοκομείου”	107
Παράρτημα ΙV “ Αποτελέσματα ικανοποίησης ασθενών από τις υπηρεσίες του ΤΕΠ νοσοκομείου ανά κριτήριο ”	110
Παράρτημα V “Προγραμματισμός αλγόριθμου Hot Deck για Ordinal Data σε γλώσσα Java”	120

Περίληψη

Η έλλειψη δεδομένων είναι ένα συνηθισμένο πρόβλημα σε στατιστικές έρευνες, το οποίο μειώνει τη δυνατότητα των ερευνητών να εφαρμόσουν τις συνήθεις στατιστικές μεθόδους. Η παρούσα εργασία ασχολείται με το πρόβλημα των ελλιπών δεδομένων και συγκεκριμένα με την αντιμετώπιση του προβλήματος σε έρευνες ικανοποίησης πελατών.

Η μέθοδος hot deck imputation επιχειρεί να δώσει λύση στο πρόβλημα των ελλιπών δεδομένων μέσω εύρεσης κατάλληλων τιμών από τα υπόλοιπα ερωτηματολόγια της έρευνας, οι οποίες θα τοποθετηθούν στη θέση των ελλείψεων. Ο συνήθης τρόπος προσδιορισμού της βέλτιστης περίπτωσης από την οποία θα γίνει η απόδοση τιμής, είναι η μέτρηση κάποιου δείκτη απόστασης σύμφωνα με τον οποίο εντοπίζονται τα πιο “κοντινά” ερωτηματολόγια.

Η εργασία ασχολείται με την ιδιαίτερη περίπτωση των ερευνών ικανοποίησης πελατών όπου τα παραγόμενα δεδομένα είναι τύπου ordinal και δεν υπάρχει κάποιος ευρέως αποδεκτός τρόπος υπολογισμού αποστάσεων. Έτσι, επιλέχθηκε η μέθοδος υπολογισμού αποστάσεων ordinal δεδομένων Walesiak, ως κύριος άξονας πάνω στον οποίο αναπτύχθηκε ο αλγόριθμος υλοποίησης της μεθόδου hot deck. Φυσικά η μέθοδος υπολογισμού Walesiak τροποποιήθηκε στην πορεία ανάπτυξης του αλγόριθμου, προκειμένου να συμπεριληφθεί ο παράγοντας των ελλιπών στοιχείων στον υπολογισμό των αποστάσεων.

Η ανάπτυξη του αλγόριθμου ολοκληρώθηκε με την ενσωμάτωση του σε πρόγραμμα γλώσσας Java, το οποίο αναπτύχθηκε για τις ανάγκες της εργασίας και κάνει εύκολη τη χρήση της μεθόδου Hot deck σε έρευνες ικανοποίησης πελατών.

Μετά την επαλήθευση ορθής λειτουργίας του αλγόριθμου, το πρόγραμμα εκτελέστηκε σε δεδομένα πραγματικής έρευνας με ελλιπή δεδομένα, που αφορά την ικανοποίηση των ασθενών για τις υπηρεσίες που παρέχονται από το ΤΕΠ νοσοκομείου στην Αττική.

Τέλος, εφαρμόστηκε η πολυκριτήρια μέθοδος MUSA (MUlticriteria Satisfaction Analysis) στα αποτελέσματα που απέδωσε το πρόγραμμα και παρουσιάστηκαν τα βασικά συμπεράσματα, καταδεικνύοντας με αυτό τον τρόπο τη σημαντικότητα εφαρμογής της μεθόδου hot deck σε έρευνες με ελλιπή δεδομένα.

1. Εισαγωγή

1.1 Γενικά για το πρόβλημα των ελλιπών δεδομένων

Από τα βασικότερα χαρακτηριστικά του ανθρωπίνου είδους είναι η έμφυτη περιέργεια για παρατήρηση του περιβάλλοντα χώρου γύρω του και η αλληλεπίδραση με τα αντικείμενα που περιέχει. Η ανάγκη επίλυσης σημαντικών προβλημάτων αλλά και η καλύτερη εκμετάλλευση των πόρων που διατίθενται στον άνθρωπο, προϋποθέτουν την πλήρη κατανόηση και σωστή αντίληψη του περιβάλλοντος. Η πολυπλοκότητα της σημερινής κοινωνίας και η ανάγκη πλήρους εκμετάλλευσης των διαθέσιμων πόρων, απαιτούν ανάπτυξη μεθόδων οι οποίες συλλέγουν τα παρεχόμενα στοιχεία και εξάγουν ασφαλή συμπεράσματα για επεξεργασία.

Η επιστήμη η οποία ασχολείται με τη συλλογή, ταξινόμηση, επεξεργασία, παρουσίαση, ανάλυση και ερμηνεία διαφόρων δεδομένων με απώτερο στόχο την εξαγωγή ασφαλών συμπερασμάτων για λήψη ορθών αποφάσεων είναι η Στατιστική. Στη σύγχρονη κοινωνία του ανθρώπου η Στατιστική έχει εφαρμογή σε ένα ευρύτατο πεδίο επιστημών που περιλαμβάνει τις θετικές, τις εφαρμοσμένες, τις ανθρωπιστικές και τις κοινωνικές επιστήμες.

Η εφαρμογή της κλασικής στατιστικής μεθόδου περιλαμβάνει τη συλλογή και την πλήρη ύπαρξη δεδομένων από το περιβάλλον της έρευνας. Η πληρότητα των δεδομένων είναι μια απαίτηση η οποία δεν μπορεί να εξασφαλιστεί πάντα, για λόγους οι οποίοι μπορεί να οφείλονται σε πλήθος παραγόντων σχεδιασμού, ανθρωπίνου δυναμικού, εξοπλισμού κ.α.

Η απώλεια των δεδομένων εγείρει ερωτήματα για τη χρήση των μερικώς συμπληρωμένων περιπτώσεων, διότι πολλές φορές δημιουργούνται προβλήματα αλλοίωσης της πραγματικότητας ή προκύπτει η ανάγκη επανάληψης της έρευνας. Το πρόβλημα είναι τόσο αναγνωρισμένο στη στατιστική έρευνα και έχει αναλυθεί σε τέτοιο βαθμό, ώστε του έχει αποδοθεί ο όρος ελλιπή δεδομένα ή missing data.

1.2 Οι αιτίες του προβλήματος

Η έλλειψη των δεδομένων είναι ένα πρόβλημα το οποίο μπορεί να δημιουργηθεί σε όλα τα στάδια της έρευνας και για πλήθος διαφορετικών λόγων, τυχαίων ή μη. Οι σύγχρονες έρευνες προσπαθούν να αμβλύνουν το πρόβλημα αυτό αναλύοντας τους λόγους για τους οποίους μπορεί να εμφανιστεί πριν την έναρξη εκτέλεσης τους, χρησιμοποιώντας νέες τεχνικές και νέο εξοπλισμό. Παρόλα αυτά έχει αποδειχθεί ότι μια πλήρης στατιστική έρευνα είναι ένα πολύ σπάνιο φαινόμενο, ιδιαίτερα όταν μέσα σε αυτή υπεισέρχεται ο ανθρώπινος παράγοντας.

Η αναζήτηση των λόγων για τους οποίους εμφανίζονται ελλιπή δεδομένα στις στατιστικές έρευνες έχει επιφέρει την ανάγκη δημιουργίας διαχωρισμού των περιπτώσεων στις οποίες εμφανίζονται. Έτσι, μπορούμε να κατηγοριοποιήσουμε τους λόγους για τους οποίους μπορεί να προκύψει έλλειψη δεδομένων σε:

- λανθασμένου σχεδιασμού
- μη αξιόπιστου εξοπλισμού
- λανθασμένης εκτέλεσης της έρευνας
- μη διάθεσης εξοπλισμού ή ανθρωπίνου δυναμικού για κάποιο χρονικό διάστημα και
- εσκεμμένης απώλειας των δεδομένων.

Η κατηγοριοποίηση αυτή είναι προφανώς υποκειμενική και ο καθένας θα μπορούσε να διαφοροποιήσει τις κατηγορίες που αναφέρθηκαν, αλλά σε γενικές γραμμές όλες οι αιτίες έλλειψης των δεδομένων μπορούν να προσαρτηθούν σε κάποια από τις παραπάνω.

Ο λανθασμένος σχεδιασμός μιας έρευνας μπορεί να δημιουργήσει πολύ εύκολα συνθήκες απώλειας δεδομένων. Πολλές φορές, κατά τη διεξαγωγή μιας έρευνας εμφανίζονται προβλήματα τα οποία ο σχεδιαστής πρέπει να έχει προβλέψει και να προσφέρει άμεσες και αποτελεσματικές λύσεις. Είναι προφανές ότι δεν είναι δυνατή η

πρόβλεψη επίλυσης όλων των προβλημάτων, καθώς δεν υπάρχουν ούτε τέλειοι ερευνητές για τη συλλογή των δεδομένων, ούτε τέλειος εξοπλισμός για την εκτέλεση πειραματικών διατάξεων, υπολογισμών και καταγραφής των δεδομένων. Η αναζήτηση του βέλτιστου τρόπου διεξαγωγής της έρευνας επιφέρει αύξηση του απαιτούμενου κόστους, αναγκάζοντας το σχεδιαστή της έρευνας σε συμβιβασμό για πρόβλεψη επίλυσης μόνο των συνηθέστερων και σημαντικότερων προβλημάτων. Παραδείγματα λανθασμένου σχεδιασμού σε έρευνες είναι η απαίτηση συλλογής πολλαπλών στοιχείων από ένα άτομο ταυτόχρονα, η λάθος διατύπωση ερωτήσεων σε ερωτηματολόγια αξιολόγησης, η προσθήκη νέων μεταβλητών κατά τη διάρκεια εκτέλεσης της έρευνας κ.α.

Ο μη αξιόπιστος εξοπλισμός είναι άλλος ένας παράγοντας ο οποίος μπορεί να συμβάλλει στη δημιουργία έλλειψης δεδομένων, καθώς μπορεί ο ίδιος ο εξοπλισμός να είναι δύσχρηστος ή τα όργανα του να είναι δυσανάγνωστα, η λειτουργία του να γίνεται διακοπτόμενα, να “χάνονται” τα δεδομένα λόγω μη αποθήκευσης κ.α. Σε αυτή την κατηγορία ο χρήστης - ερευνητής μπορεί να αντιλαμβάνεται το πρόβλημα αλλά είτε δεν έχει το χρόνο να αντιδράσει, είτε δεν υπάρχει εναλλακτικός εξοπλισμός για χρήση.

Η λανθασμένη εκτέλεση της έρευνας οφείλεται κατά κύριο λόγο στον ανθρώπινο παράγοντα, δηλαδή το προσωπικό το οποίο εκτελεί την έρευνα. Ο ερευνητής μπορεί να μην πάρει κάποιες μετρήσεις λόγω απροσεξίας, μπορεί να απορρίπτει δεδομένα λόγω λανθασμένης αντίληψης τη δεδομένη στιγμή, να μην είναι σωστά εκπαιδευμένος ή να μην έχει τα ιδιαίτερα χαρακτηριστικά που απαιτούνται για την έρευνα, π.χ. σε έρευνες που απαιτούν συνέντευξη ατόμων κ.α.

Η μη διάθεση δεδομένων για κάποιο χρονικό διάστημα μπορεί να οφείλεται, τόσο στον εξοπλισμό και το περιβάλλον, όσο και στο ανθρώπινο δυναμικό που συμμετέχει στην έρευνα και συνεπάγεται διακοπή της συλλογής μέρους ή εξ ολοκλήρου των στοιχείων ενώ η έρευνα συνεχίζεται. Μη διάθεση δεδομένων είναι δυνατόν να προκύψει όταν υπάρχει βλάβη στον εξοπλισμό, διακοπή των παρεχόμενων πρώτων υλών ή αρχικών δεδομένων, μεταβολή του

τεχνικού και φυσικού περιβάλλοντος, απουσία κάποιων ερευνητών από την ομάδα εργασίας, απουσία ατόμων που λαμβάνουν μέρος στην έρευνα ως ερωτώμενοι, ασθενείς κ.α.

Τέλος **εσκεμμένη απώλεια των δεδομένων** μπορεί να προκύψει από καταστροφή των συλλεχθέντων δεδομένων, αδιαφορία κατά την εκτέλεση της έρευνας, κόπωση ή απροθυμία των συμμετεχόντων στην έρευνα να απαντήσουν σε ορισμένες ερωτήσεις. Ο συγκεκριμένος παράγοντας εμφανίζεται κυρίως στις έρευνες όπου επιλέγεται ακατάλληλο προσωπικό για τη διεξαγωγή της ή σε έρευνες που χρησιμοποιείται προσωπική ή τηλεφωνική συνέντευξη, χωρίς να έχει γίνει από την αρχή κατανοητό το μέγεθος και η φύση της.

1.3 Εξέλιξη αντιμετώπισης - ιστορική αναδρομή

Τα ελλιπή δεδομένα μειώνουν την ικανότητα του στατιστικού αναλυτή και των χρησιμοποιούμενων μεθόδων να προσεγγίσουν το πρόβλημα με αξιόπιστα αποτελέσματα. Στη σημερινή εποχή όπου είναι εύκολη η απόκτηση αξιόπιστου εξοπλισμού και ο έλεγχος καλής λειτουργίας είναι ευκολότερος, το πρόβλημα εμφανίζεται σε μεγαλύτερη κλίμακα σε επιστήμες που ασχολούνται με τον άνθρωπο και εντάσσεται η υποκειμενικότητα στη διεξαγωγή της έρευνας π.χ. σε επιστήμες όπως η κοινωνιολογία, η βιολογία, η ιατρική, η ψυχολογία κ.α.

Οι κλασικές μέθοδοι που χρησιμοποιήθηκαν αρχικά στη στατιστική έρευνα, δεν περιλάμβαναν τρόπους αντιμετώπισης του προβλήματος καθώς στηρίζονταν σε πλήρεις συλλογές δεδομένων. Για το λόγο αυτό, αναπτύχθηκαν διάφορες μέθοδοι με σκοπό την επεξεργασία δεδομένων με ελλιπή στοιχεία και την διαμόρφωση τους, ώστε ο ερευνητής να έχει διαθέσιμες μόνο πλήρης συλλογές δεδομένων.

Το πρόβλημα των ελλιπών δεδομένων απασχολεί συστηματικά τους επιστήμονες από τις αρχές του 20^{ου} αιώνα, όπου έγιναν τα πρώτα βήματα εξεύρεσης κάποιας λύσης. Μέχρι τη δεκαετία του '70 οι λύσεις

που επικρατούσαν σχετίζονταν με την διαγραφή των ερωτηματολογίων με ελλιπή στοιχεία ή τη συμπλήρωση τους με κάποια απλή μέθοδο όπως αυτή της μέσης τιμής.

Από το 1987 ο Donald B. Rubin ασχολήθηκε συστηματικά με το πρόβλημα των ελλιπών δεδομένων και εξέδωσε εργασία όπου αναλύοντας τους μηχανισμούς ανάπτυξης τους και καταλήγοντας σε διάφορα συμπεράσματα, πρότεινε τρόπους ταξινόμησης και χειρισμού τους. Αυτή ήταν η πρώτη απόπειρα εξεύρεσης αξιόπιστης λύσης στο πρόβλημα και τα συμπεράσματα της εργασίας χρησιμοποιούνται έως σήμερα.

Την ίδια χρονολογία οι Little και Rubin εκδίδουν βιβλίο όπου στην επίλυση τους προβλήματος χρησιμοποιείται ο αλγόριθμος EM (Expectation – Maximization), ο οποίος είχε για πρώτη φορά διατυπωθεί το 1977 από τους Dempster, Laird και Rubin. Η εν λόγω πρόταση οδήγησε στην ανάπτυξη της μεθόδου εκτίμησης μέγιστης πιθανότητας (Full Information Maximum Likelihood ή FIML), η οποία χρησιμοποιείται σε πολλά προβλήματα με ελλιπή δεδομένα ακόμα και σήμερα. Αυτά τα συμπεράσματα αν και βασίζονται σε συγκεκριμένα μοντέλα με σχετική δυσκολία εφαρμογής, αντιμετωπίστηκαν ως πολύ σημαντικές και εντυπωσιακές ανακαλύψεις.

Λίγο αργότερα, ο Rubin αναπτύσσει την μέθοδο του πολλαπλού καταλογισμού (Multiple Imputation), η οποία χρησιμοποιήθηκε ευρέως το επόμενο χρονικό διάστημα με την βελτίωση της υπολογιστικής δύναμης των Υ/Η.

Από τις αρχές της δεκαετίας του 90 εμφανίζεται μια νέα μέθοδος, η Μπεϋσιανή Προσομοίωση (Bayesian Simulation), ενώ ταυτόχρονα εμφανίζονται προγράμματα που χρησιμοποιούν τις προηγούμενες τεχνικές και μάλιστα διανέμονται δωρεάν.

1.4 Δομή και στόχοι της εργασίας

Στόχος της παρούσας εργασίας είναι η δημιουργία προγράμματος το οποίο θα χρησιμοποιεί αλγόριθμο της μεθόδου hot deck imputation,

για συμπλήρωση των ελλιπών στοιχείων σε έρευνες ικανοποίησης πελατών. Για την εύρεση του βέλτιστου ερωτηματολογίου-δότη της ελλιπούς τιμής, γίνεται χρήση της μεθόδου Walesiak όπου καθιστά εφικτή τη μέτρηση της απόστασης δεδομένων τύπου ordinal. Μετά την ολοκλήρωση του προγράμματος εκτελείται έλεγχος ορθής λειτουργίας και στη συνέχεια δοκιμαστική εφαρμογή σε πραγματική έρευνα ικανοποίησης πελατών.

Αρχικά γίνεται ανάλυση του προβλήματος των ελλιπών δεδομένων μέσω κατανόησης της λειτουργίας των ερευνών ικανοποίησης πελατών. Προσεγγίζεται δηλαδή η εκτέλεση των εν λόγω ερευνών και αναλύεται η μεθοδολογία, τα παραγόμενα δεδομένα και οι αιτίες για τις οποίες μπορεί να δημιουργηθεί έλλειψη. Επίσης γίνεται διαχωρισμός των μηχανισμών δημιουργίας του φαινομένου με βάση την τυχαιότητα ή μη εμφάνισης του.

Στη συνέχεια, παρουσιάζονται οι προσπάθειες αντιμετώπισης του προβλήματος από τους ερευνητές που ασχολήθηκαν με αυτό διαχρονικά. Γίνεται ανάλυση όλων των μεθόδων που έχουν χρησιμοποιηθεί κατά καιρούς από τους αναλυτές ώστε να έχουν στη διάθεση τους μόνο πλήρως συμπληρωμένα ερωτηματολόγια, από τις απλές μεθόδους διαγραφής έως τις σύγχρονες μεθόδους μοντελοποίησης.

Ακολουθεί η ανάπτυξη αλγόριθμου ο οποίος χρησιμοποιεί τη μέθοδο hot deck imputation, δηλαδή τον καταλογισμό σε ένα ελλιπές στοιχείο κάποιας υπάρχουσας τιμή από άλλο “όμοιο” ερωτηματολόγιο. Για την προσαρμογή του αλγόριθμου σε έρευνες ικανοποίησης πελατών όπου τα παραγόμενα δεδομένα είναι τύπου ordinal, έγινε χρήση της μεθόδου Walesiak όπου γίνεται υπολογισμός της απόστασης, δηλαδή του βαθμού ομοιότητας, διαφορετικών σετ ordinal δεδομένων. Για την εφαρμογή του αλγόριθμου σε έρευνες ικανοποίησης πελατών καθορίστηκε συγκεκριμένη κωδικοποίηση δεδομένων, ελλιπών τιμών και εισήχθη συντελεστής που χρησιμεύει στον έλεγχο αξιόπιστου καταλογισμού τιμών. Η ανάπτυξη του αλγόριθμου συνοδεύεται από αρκετά παραδείγματα για να γίνει περισσότερο κατανοητός ο τρόπος

με τον οποίο αντιμετωπίστηκαν τα διάφορα προβλήματα που εμφανίστηκαν και οριστικοποιήθηκε η τελική μορφή του.

Στο ίδιο κεφάλαιο παρουσιάζεται η μορφή και η λειτουργία του προγράμματος που δημιουργήθηκε με βάση τον αλγόριθμο hot deck σε γλώσσα προγραμματισμού java. Επίσης, παρουσιάζεται η τεκμηρίωση ορθής λειτουργίας του αλγόριθμου με εφαρμογή του προγράμματος σε πραγματική έρευνα με πλήρη δεδομένα.

Τελικά γίνεται εφαρμογή του προγράμματος σε πραγματική έρευνα ικανοποίησης πελατών με ελλιπή δεδομένα ώστε να συμπληρωθούν τα μερικώς συμπληρωμένα ερωτηματολόγια. Τα πλέον πλήρη ερωτηματολόγια αξιοποιούνται και αναλύονται με τη σύγχρονη μέθοδο πολυκριτήριας ανάλυσης MUSA. Γίνεται μια σύντομη παρουσίαση της μεθόδου και παρουσιάζονται τα αποτελέσματα με εφαρμογή της στην παραπάνω έρευνα.

2. Ανάλυση Προβλήματος Ελλιπών δεδομένων

2.1 Στατιστικές έρευνες

Όλες οι σύγχρονες επιστήμες στηρίζονται στο πείραμα και τη στατιστική έρευνα για τη θεμελίωση των θεωριών, κανόνων και προτάσεων που τις διέπουν. Η στατιστική έρευνα είναι το εργαλείο που δίνει στον επιστήμονα τη δυνατότητα να αποδείξει τις θεωρίες του, να βγάλει συμπεράσματα ή να οδηγηθεί σε νέες έρευνες χρησιμοποιώντας απλουστευμένες μορφές, δείκτες και διαγράμματα των αποτελεσμάτων.

Για αυτό το λόγο η στατιστική έρευνα έχει χρησιμοποιηθεί από την αρχαιότητα σε διάφορους πολιτισμούς και σε μεγάλο εύρος τομέων της ανθρώπινης δραστηριότητας. Η όλη δομή της σύγχρονης κοινωνίας εν γένει έχει στηριχθεί σε αποτελέσματα οργανωμένων ή όχι στατιστικών ερευνών. Η εκπαίδευση, η οικονομία, η πολιτική, ο στρατός, η υγεία και πολλοί άλλοι τομείς είναι παραδείγματα χρήσης αποτελεσμάτων στατιστικών ερευνών για την καλύτερη οργάνωση και τη λειτουργία τους.

Σε πολλές περιπτώσεις σήμερα, η στατιστική έρευνα έχει γίνει απαραίτητο στοιχείο για την οργάνωση και τη λειτουργία του κράτους, οργανισμών και επιχειρήσεων. Χωρίς αυτήν, είναι αδύνατη η οποιαδήποτε λήψη απόφασης και σχεδιασμού στρατηγικής χωρίς τη λήψη υψηλού ρίσκου. Η γνώση και κατανόηση του περιβάλλοντος λειτουργίας ενός οργανισμού είναι τα δεδομένα που παίζουν καθοριστικό ρόλο στη βιωσιμότητα του, μέσω σωστών αποφάσεων και χάραξης πολιτικής.

Δεν είναι λοιπόν παράδοξο το φαινόμενο δημιουργίας οργανισμών, φορέων ή τμημάτων σε όλες τις πτυχές της κοινωνίας, με αποκλειστική αρμοδιότητα την διεξαγωγή στατιστικών ερευνών και την παρουσίαση κατανοητών αποτελεσμάτων. Χρήση σύνθετων ή απλούστερων στατιστικών τεχνικών γίνεται καθημερινά από διεθνείς – διακρατικούς οργανισμούς, κρατικές υπηρεσίες, τμήματα επιχειρήσεων

ή ακόμα και απλούς ανθρώπους, άσχετους με οποιαδήποτε οργανωμένη έρευνα κατά την προσπάθεια επίλυσης ατομικών προβλημάτων.

Σαν ορισμός, στατιστική είναι ο κλάδος των εφαρμοσμένων επιστημών που έχει σαν αντικείμενο τη συλλογή, οργάνωση, περιγραφή και περιληπτική παρουσίαση διάφορων πληροφοριών και ακόμη τη διερεύνηση και ανάλυση των πληροφοριών αυτών για την εξαγωγή συμπερασμάτων.

Η διεξαγωγή μιας στατιστικής έρευνας συνήθως περιλαμβάνει τις παρακάτω κύριες φάσεις:

- **Προετοιμασία**, όπου γίνεται ο καθορισμός του σκοπού της έρευνας, δίνεται ο ορισμός των όρων που θα χρησιμοποιηθούν και προσδιορίζονται οι πηγές και τα μέσα συλλογής δεδομένων
- **Εκτέλεση**, όπου γίνεται η συλλογή των δεδομένων με βάση μια προκαθορισμένη μέθοδο
- **Επεξεργασία δεδομένων**, όπου γίνεται η παρουσίαση των δεδομένων σε κατάλληλη μορφή και η ανάλυση με κάποιο στατιστικό πακέτο
- **Παρουσίαση αποτελεσμάτων**, όπου παρουσιάζονται τα σχετικά συμπεράσματα που προκύπτουν από την έρευνα

2.2 Έρευνες ικανοποίησης πελατών

Στη στατιστική έρευνα μπορεί να αποδοθεί πλήθος χαρακτηρισμών, ανάλογα με το επιστημονικό πεδίο χρήσης της. Έτσι μπορεί να αναφερθεί ως οικονομική, μηχανική, ψυχολογική, κοινωνική, βιομηχανική, δημογραφική κ.α.

Ένα λοιπόν από τα πολλά πεδία χρήσης της στατιστικής είναι η βιομηχανική έρευνα, βασικός στόχος της οποίας είναι η αναζήτηση των τάσεων της αγοράς. Μέσω των αξιόπιστων στατιστικών αποτελεσμάτων μια επιχείρηση δύναται να καθορίσει πολιτικές για ανάπτυξη, βελτίωση,

διατήρηση ή κατάργηση προϊόντων και υπηρεσιών με στόχο την κατά το δυνατόν μέγιστη ικανοποίηση των πελατών της. Η αξιόπιστη στατιστική έρευνα είναι ένα σημαντικό εργαλείο, μέσω του οποίου μπορεί να αποφασισθεί μια κερδοφόρα επένδυση, σε μια δεδομένη χρονική περίοδο και σε συγκεκριμένο τόπο εφαρμογής της.

Η διεξαγωγή μιας βιομηχανικής έρευνας έχει ως κύριο σκοπό την αποτύπωση ικανοποίησης των πελατών μιας επιχείρησης, δηλαδή της πρόθεσης επαναγοράς των προϊόντων – υπηρεσιών που προσφέρει. Με αυτό τον τρόπο χαράσσονται στρατηγικές πολιτικές τόσο για τη διατήρηση των πελατών της επιχείρησης, όσο και για προσέλκυση νέων.

Η λέξη ικανοποίηση (Satisfaction) ως ετυμολογική έννοια προέρχεται από τη λατινική λέξη satis, που σημαίνει αρκετά και τη λέξη facere που σημαίνει κάνω ή φτιάχνω. Ωστόσο, δεν έχει συμφωνηθεί ένας κοινά αποδεκτός ορισμός για την ‘ικανοποίηση πελατών’, παρά την εκτεταμένη έρευνα και τις διαφορετικές προτάσεις ανά τους ερευνητές σε όλο τον κόσμο. Ο Oliver (1997) αναφέρει ότι “*όλοι γνωρίζουν τι είναι η ‘ικανοποίηση’ μέχρι τη στιγμή που θα τους ζητηθεί να δώσουν ένα ορισμό. Τότε φαίνεται ότι κανείς δεν ξέρει*”. Ακόμα και ως προς την ορολογία λοιπόν υπάρχει ασυμφωνία αφού η ικανοποίηση πελατών στην βιβλιογραφία αναφέρεται με διάφορους όρους όπως Ικανοποίηση καταναλωτή (Consumer Satisfaction), Ικανοποίηση πελάτη (Customer Satisfaction) ή απλά Ικανοποίηση (Satisfaction).

Έτσι, για τον ορισμό της ικανοποίησης του πελάτη έχουν προταθεί κατά καιρούς πλήθος ορισμών. Μερικοί από αυτούς φαίνονται παρακάτω:

«...Η ικανοποίηση είναι μια γνωστική κατάσταση του πελάτη, όσον αφορά στην επαρκή ή ανεπαρκή ανταμοιβή του για τις θυσίες και τις προσπάθειες που έχει καταβάλλει...» (Howard and Sheth, 1969)

«...Η ικανοποίηση είναι μια συναισθηματική αντίδραση στις εμπειρίες του πελάτη, οι οποίες σχετίζονται είτε με συγκεκριμένα προϊόντα και υπηρεσίες, είτε με τις διαδικασίες αγοράς, είτε ακόμα με

συγκεκριμένα χαρακτηριστικά του πελάτη αυτού...» (Westbrook and Reilly, 1983)

«...Η ικανοποίηση είναι το αποτέλεσμα της αγοράς και χρήσης ενός προϊόντος ή μιας υπηρεσίας, το οποίο απορρέει από τη σύγκριση του πελάτη ανάμεσα στην ανταμοιβή και το κόστος αγοράς, λαμβάνοντας υπόψη τις προσδοκώμενες επιπτώσεις...» (Churchill and Suprenant, 1982)

«...Η ικανοποίηση είναι μια διαδικασία αξιολόγησης, η οποία βασίζεται στο κατά πόσο η συγκεκριμένη εμπειρία ήταν τόσο καλή όσο ο πελάτης πίστευε ότι θα είναι...» (Hunt, 1977)

«...Η ικανοποίηση είναι μια διαδικασία αξιολόγησης, η οποία εξετάζει αν η συγκεκριμένη επιλογή είναι συμβατή με τις προγενέστερες πεποιθήσεις του πελάτη...» (Engel and Black, 1982)

«...Η ικανοποίηση είναι η αντίδραση του καταναλωτή στη διαδικασία αξιολόγησης, η οποία εξετάζει τις ασυμφωνίες μεταξύ προγενέστερων προσδοκιών και του πραγματικού επιπέδου απόδοσης του προϊόντος, όπως γίνεται αντιληπτό από τον καταναλωτή μετά τη χρήση του...» (Tse and Wilton, 1988)

«...Η ικανοποίηση είναι ένα μέτρο για το πόσο το προσφερόμενο ολικό προϊόν ή υπηρεσία εκπληρώνει τις προσδοκίες του πελάτη...» (Oliver 1996, Hill 1996, Gerson 1993, και Varna 1997).

Η γνώση της ικανοποίησης πελατών οδηγεί σε διαφορετική διαχείριση των ανθρωπίνων πόρων, εξοπλισμού και ανάπτυξης τεχνογνωσίας από οποιοδήποτε φορέα παροχής υπηρεσιών-προϊόντων. Είναι κατανοητό λοιπόν, ότι η μέτρηση της ικανοποίησης του πελάτη είναι ένα απαραίτητο εργαλείο για το σχεδιασμό και την εφαρμογή μιας αποτελεσματικής πολιτικής ποιότητας και διαδικασιών σε μια επιχείρηση ή οργανισμό.

Οι έρευνες ικανοποίησης πελατών κατηγοριοποιούνται σε ποιοτικές και ποσοτικές έρευνες, ανάλογα με τα αντικείμενα και τους στόχους της έρευνας.

Οι ποιοτικές έρευνες ικανοποίησης στοχεύουν στην καταγραφή αναλυτικών πληροφοριών για το αντικείμενο της έρευνας, δεν έχουν προκαθορισμένη μορφή και έχουν περιορισμένο πλήθος δείγματος. Το πλεονέκτημα αυτού του είδους έρευνας είναι η εις βάθος ανάλυση συμπεριφοράς του πελάτη, μέσω καταγραφής πληροφοριών που δεν είναι δυνατό να αποτυπωθούν σε ένα απλό ερωτηματολόγιο. Για αυτό το λόγο οι ποιοτικές έρευνες ικανοποίησης διεξάγονται είτε μέσω προσωπικών ή ομαδικών συνεντεύξεων όπου ο ερευνητής συντονίζει και καταγράφει τις συζητήσεις, είτε μέσω παρατήρησης των πελατών από εξειδικευμένο προσωπικό.

Οι ποσοτικές έρευνες ικανοποίησης αντίθετα, στοχεύουν στην καταγραφή συγκεκριμένων αξιόπιστων στατιστικών δεδομένων και την εξαγωγή μετρήσιμων δεικτών όπου παρουσιάζουν την ικανοποίηση των πελατών. Για αυτό το λόγο στις ποσοτικές έρευνες χρησιμοποιείται ένα αυστηρά δομημένο ερωτηματολόγιο μικρής έκτασης με κλειστές ερωτήσεις, ώστε να είναι δυνατή η εφαρμογή του σε μεγάλο πλήθος πελατών. Τα δεδομένα στη συνέχεια αναλύονται με μοντέλα ποσοτικής ανάλυσης και τα αποτελέσματα παρουσιάζονται μέσω κατάλληλα επιλεγμένων δεικτών και διαγραμμάτων. Τα πλεονεκτήματα των ποσοτικών ερευνών ικανοποίησης είναι η αξιοπιστία γενίκευσης των αποτελεσμάτων λόγω μεγάλου στατιστικού δείγματος και η ευκολία διεξαγωγής τους. Για αυτό το λόγο, οι εν λόγω έρευνες μπορούν να χρησιμοποιούνται τόσο ως τρόπος μέτρησης μεταβολής της εικόνας των πελατών σε βάθος χρόνου, όσο και ως εργαλείο ελέγχου της ανταπόκρισης των πελατών σε αλλαγές που πραγματοποιούνται.

Τα σημαντικότερα πλεονεκτήματα μιας έρευνας ικανοποίησης πελατών σύμφωνα με μελέτες που έχουν γίνει στο παρελθόν (Czarnecki 1999, Dutca 1995, Naumann and Giel 1995) μπορεί να συνοψιστεί στα παρακάτω σημεία:

- Εφόσον η διαδικασία μέτρησης αποτελεί μια συνεχή και αποτελεσματική προσπάθεια, βελτιώνεται η επικοινωνία με το σύνολο των πελατών.

- Δίνεται η ευκαιρία στην επιχείρηση να αποκτήσει εικόνα για το κατά πόσο οι προσφερόμενες υπηρεσίες ανταποκρίνονται στις προσδοκίες των πελατών.
- Γίνεται εύκολος ο εντοπισμός των κρίσιμων διαστάσεων της ικανοποίησης για τις οποίες πρέπει να αναζητηθούν τρόποι βελτίωσης.
- Προσδιορίζονται τα σημαντικότερα πλεονεκτήματα και μειονεκτήματα σε σχέση με τον ανταγωνισμό, πάντα σύμφωνα με τις απόψεις των πελατών.
- Η παρουσίαση της ικανοποίησης πελατών μπορεί να αποτελέσει κίνητρο επαύξησης της παραγωγικότητας και της ποιότητας της εργασίας των εργαζομένων.

2.3 Μεθοδολογία έρευνας ικανοποίησης πελατών

Η έρευνα ικανοποίησης πελατών, ως υποκατηγορία της στατιστικής έρευνας, έχει όλα τα βασικά στάδια λειτουργίας της που είναι ο σχεδιασμός, η εκτέλεση, η επεξεργασία και η παρουσίαση των αποτελεσμάτων. Για την αποτύπωση της πραγματικής εικόνας – ικανοποίησης που έχουν οι πελάτες για το προϊόν μιας επιχείρησης, είναι απαραίτητη η ομαλή εκτέλεση όλων των σταδίων. Οποιαδήποτε ατέλεια στη λειτουργία ενός σταδίου της έρευνας έχει σίγουρο αντίκτυπο στην παρουσίαση των αποτελεσμάτων.

Όπως είναι κατανοητό, τα περισσότερα προβλήματα που μπορεί να αλλοιώσουν την σωστή απόδοση της πραγματικότητας μπορούν να προβλεφθούν, μειωθούν ή και να εξαλειφθούν από τη φάση του σχεδιασμού της έρευνας. Ο ερευνητής ή η ομάδα ερευνητών που διεξάγουν την έρευνα πρέπει να είναι βέβαιοι ευέλικτοι σε οποιοδήποτε πρόβλημα παρουσιαστεί στη πορεία εκτέλεσης της έρευνας και δεν έχει προβλεφθεί στη φάση σχεδιασμού και να το αντιμετωπίζουν είτε μέσω τροποποιήσεων τμημάτων της έρευνας, είτε μέσω προβλέψεων των αλλοιώσεων που μπορεί να επέλθουν και αποδοχής τους με το ανάλογο ρίσκο.

Στο αρχικό στάδιο λοιπόν, καθορίζονται αναλυτικά οι στόχοι της έρευνας, ορίζονται οι κεντρικές έννοιες οι οποίες θα χρησιμοποιηθούν για την επικοινωνία των ερευνητών με τους πελάτες και μέσω προκαταρκτικής έρευνας προσδιορίζεται η συμπεριφορά των πελατών της επιχείρησης. Όταν συλλεχθούν όλα τα στοιχεία, είναι πλέον δυνατόν να καθοριστούν τα κριτήρια του μοντέλου που θα χρησιμοποιηθεί στην έρευνα, η μορφή διεξαγωγής της έρευνας και το πλήθος και η κλίμακα των ερωτήσεων που θα αποτελέσουν το ερωτηματολόγιο.

Είναι προφανές ότι στο αρχικό στάδιο ελλοχεύουν και οι περισσότεροι κίνδυνοι αποτυχίας της έρευνας, αφού ο λανθασμένος σχεδιασμός οποιουδήποτε τμήματος μπορεί να επιφέρει σημαντικές αποκλίσεις στα αποτελέσματα από την πραγματικότητα. Είναι απαραίτητο οι όροι, τα κριτήρια, η μορφή της έρευνας αλλά και το ερωτηματολόγιο που θα χρησιμοποιηθεί να εναρμονίζονται με τους στόχους που έχουν τεθεί από την αρχή για τα αποτελέσματα της έρευνας. Σε αυτό το στάδιο, είναι δυνατό να εφαρμοστεί δοκιμή του ερωτηματολογίου σε μια πιλοτική έρευνα ώστε να αντιμετωπιστούν πιθανά προβλήματα που θα προκύψουν πριν την έναρξη της κύριας έρευνας.

Το επόμενο στάδιο αφορά την καθ' αυτή διεξαγωγή της έρευνας. Σε αυτή τη φάση συντάσσεται το ερωτηματολόγιο, ορίζονται με σαφήνεια ο τόπος διεξαγωγής, το δείγμα των πελατών και οι ερευνητές οι οποίοι θα συμμετάσχουν στην έρευνα. Στη συνέχεια οργανώνεται η εκτέλεση της έρευνας, συλλέγονται τα ερωτηματολόγια με τις απαντήσεις των πελατών και δημιουργείται η βάση δεδομένων με τα αποτελέσματα.

Στο τρίτο στάδιο γίνεται η επεξεργασία και η ανάλυση των δεδομένων της έρευνας. Σε αυτό το στάδιο πριν την έναρξη των αναλύσεων, είναι πιθανό να εκτελεστεί μια αρχική επεξεργασία των συλλεχθέντων ερωτηματολογίων, ώστε να χρησιμοποιηθούν μόνο τα κατάλληλα συμπληρωμένα για τους σκοπούς της έρευνας. Απαιτείται για το σκοπό αυτό, η χρήση κατάλληλης μεθόδου που θα αποφανθεί

για την αποδοχή ή την απόρριψη των μη πλήρως συμπληρωμένων ερωτηματολογίων στους υπολογισμούς των τελικών αποτελεσμάτων. Στη συνέχεια, χρησιμοποιούνται οι κλασικές στατιστικές μέθοδοι για να υπολογιστούν ορισμένοι βασικοί δείκτες και πιθανώς κάποιο εξειδικευμένο πρόγραμμα υλοποίησης πολυκριτήριων μεθόδων ανάλυσης δεδομένων.

Στο τελικό στάδιο εκτιμώνται τα αποτελέσματα και επιλέγεται η μορφή παρουσίασης τους. Στην τελική μορφή παρουσίασης δίνονται και οι πιθανές προτάσεις βελτίωσης της ικανοποίησης των πελατών σε κάθε τομέα. Οι προτάσεις βελτίωσης μπορεί να αφορούν βελτίωση των χαρακτηριστικών που εμφανίζουν τη λιγότερη ικανοποίηση στους πελάτες, διατήρηση ή ισχυροποίηση των χαρακτηριστικών που εμφανίζουν μεγάλη ικανοποίηση στους πελάτες, ή αλλαγή της πολιτικής διάθεσης των προϊόντων ανάλογα με την ικανοποίηση των διάφορων ομάδων πελατών που προέκυψαν από την έρευνα.

2.4 Παραγόμενα δεδομένα και τύποι δεδομένων έρευνας

Ανάλογα με το είδος της έρευνας ικανοποίησης, τους στόχους και τα προσδοκώμενα αποτελέσματα, εμφανίζονται και διάφοροι τύποι παραγόμενων δεδομένων. Μια πρώτη κατηγοριοποίηση παραγόμενων δεδομένων στις έρευνες ικανοποίησης είναι σε ποσοτικά ή ποιοτικά δεδομένα, ανάλογα με την απαίτηση παρουσίασης των αποτελεσμάτων σε αριθμούς ή εκφράσεις. Η εξέλιξη κατηγοριοποίησης των εμφανιζόμενων δεδομένων σε έρευνες ικανοποίησης έχει προταθεί από τον Stevens (1946) με τέσσερις τύπους δεδομένων που αναφέρονται ως "nominal," "ordinal," "interval," and "ratio", προσπαθώντας έτσι να καλύψει όλες τις κατηγορίες ποσοτικών και ποιοτικών δεδομένων.

Ο nominal ή ονομαστικός τύπος δεδομένων χρησιμοποιείται για να εκφράσει μη μετρήσιμα δεδομένα τα οποία χωρίζονται σε κατηγορίες και δεν μπορούν να ταξινομηθούν με κάποιο μέγεθος. Για παράδειγμα ονομαστική κλίμακα δεδομένων μπορεί να είναι μια κλίμακα χρωμάτων, εθνικοτήτων, φύλλων κλπ. Στα ονομαστικά

δεδομένα προφανώς δεν είναι δυνατή οποιαδήποτε τέλεση πράξεων εκτός της ισότητας και της ανισότητας. Αυτός ο τύπος δεδομένων χρησιμοποιείται κυρίως στις ποιοτικές έρευνες ικανοποίησης, αφού μπορεί να εκφράσει πληροφορίες χαρακτηριστικών του προϊόντος που δε μετρούνται ή δεν μπορούν να ταξινομηθούν με βάση κάποια κλίμακα μεγέθους.

Ο ordinal ή βαθμωτός τύπος δεδομένων χρησιμοποιείται για να εκφράσει μη μετρήσιμα δεδομένα, τα οποία όμως είναι δυνατόν να ταξινομηθούν με βάση κάποιο χαρακτηριστικό τους. Ordinal κλίμακα δεδομένων είναι για παράδειγμα μια κλίμακα τύπου λίγο-μέτρια-πολύ ή ακόμα κρύο-χλιαρό-ζεστό. Στα ordinal δεδομένα είναι δυνατή η τέλεση πράξεων τόσο ισότητας - ανισότητας όσο και μεγαλύτερου - μικρότερου με τους τελεστές "<" και ">". Έτσι, σε αυτού του τύπου δεδομένα επιτρέπεται η κατάταξη από το μικρότερο στο μεγαλύτερο, αφού πάντα είναι δυνατή η έκφραση "μεγαλύτερο" από ή "μικρότερο από". Για την επεξεργασία, είναι δυνατή η αυθαίρετη κωδικοποίηση σε αριθμητικά δεδομένα αλλά εμπεριέχει υψηλό κίνδυνο σφαλμάτων στις πράξεις και τις αναλύσεις και πρέπει να γίνεται με προσοχή. Στα ordinal δεδομένα δεν υφίσταται η έκφραση κάποιας μετρήσιμης διαφοράς μεταξύ τους, παρά μόνο η ταξινόμηση με βάση τη διαφορά της θέσης τους στην κλίμακα.

Στα interval ή κλιμακωτά δεδομένα υπάρχει αυθαίρετη μονάδα μέτρησης η οποία χρησιμοποιείται για να δημιουργήσει μια κλίμακα ίσων αποστάσεων μεταξύ των επιπέδων κατάταξης. Ορίζεται δηλαδή μια κωδικοποίηση των δυνατών καταστάσεων με ίσες αποστάσεις μεταξύ τους. Ένα παράδειγμα interval κλίμακας είναι οι δυνατές απαντήσεις σε μια ερώτηση ικανοποίησης σε μια κλίμακα από το 1 έως το 5. Σε αυτά τα δεδομένα συνήθως δίνεται ο ορισμός των ακραίων τιμών π.χ. στις τιμές 1 και 5 παραπάνω "καθόλου" και "πολύ" αντίστοιχα, και ο ερωτώμενος καλείται να εκφράσει την απάντηση του με βάση το πόσο κοντά σε κάποιο από τα δύο άκρα βρίσκεται. Σε αυτή την κλίμακα δεδομένων παρόλο που υπάρχει έκφραση σε αριθμούς δεν υπάρχει το σημείο αφετηρίας 0 με λογική υπόσταση, αφού ουσιαστικά

η κωδικοποίηση γίνεται αυθαίρετα και δεν ανταποκρίνεται σε κάποιο μετρήσιμο φυσικό μέγεθος. Ομοίως, δεν μπορεί να γίνει μέτρηση της απόστασης του χαρακτηριστικού μεταξύ των δεδομένων αφού η κλίμακα έχει οριστεί αυθαίρετα, δηλαδή δεν μπορούμε να πούμε ότι η θερμοκρασία 40°C είναι δύο φορές πιο ζεστή από τη θερμοκρασία 20°C αφού το 0°C δεν έχει την έννοια του απόλυτου 0 που υπάρχει στην κλίμακα Kelvin.

Τέλος, ο **ratio ή αναλογικός** τύπος δεδομένων εκφράζει όλα τα μετρήσιμα δεδομένα τα οποία έχουν φυσική υπόσταση. Σε αυτά μπορούν να εφαρμοστούν όλες οι στατιστικές αναλυτικές μέθοδοι για την παραγωγή αποτελεσμάτων και συμπερασμάτων. Στην ratio κλίμακα δεδομένων υπάρχει η έννοια του μοναδικού και σαφώς ορισμένου 0 και για αυτό χρησιμοποιείται σε μαθηματικές, φυσικές και μηχανικές επιστήμες. Παραδείγματα ratio δεδομένων είναι η μάζα, το μήκος, η γωνία, η ενέργεια κλπ.

Η κατάταξη κατηγοριών των τύπων δεδομένων του Stevens (1946) δεν έχει τύχει ευρείας αποδοχής για ολόκληρο το φάσμα των επιστημών και έχουν προταθεί και άλλες μέθοδοι κατάταξης. Μια από αυτές έχει προταθεί από τον Σίσκο, 1986 και η οποία αναφέρει την κατάταξη των δεδομένων σε numerical/metric, ordinal και nominal. Ουσιαστικά είναι μια όμοια κατάταξη, όπου τα interval και τα ratio δεδομένα ομαδοποιούνται σε numerical/metric δηλαδή αριθμητικά, καλύπτοντας το φάσμα των επιλογών των δύο ομάδων. Αυτή η κατηγοριοποίηση δεδομένων καλύπτει τις δυνατές περιπτώσεις για την εφαρμογή στις έρευνες ικανοποίησης πελατών οι οποίες μπορεί να είναι ο βαθμός ικανοποίησης, το επίπεδο καταναλωτικής πίστης, η πρόθεση επαναγοράς του προϊόντος κ.α.

2.5 Αιτίες έλλειψης δεδομένων σε έρευνες ικανοποίησης

Μια ιδανική έρευνα ικανοποίησης πελατών όπου έχει ακολουθηθεί η μεθοδολογία των τεσσάρων φάσεων προσεχτικά, απαρτίζεται από κατανοητό και απλό ερωτηματολόγιο και εκτελείται

από σωστά επιλεγμένους ερευνητές έχει ως αποτέλεσμα την συλλογή πλήρως συμπληρωμένων ερωτηματολογίων. Στην πραγματικότητα όμως, πολύ λίγοι επιστήμονες κατορθώνουν να πραγματοποιήσουν έρευνα με πλήρη δεδομένα, καθώς στις περισσότερες περιπτώσεις υπάρχουν απώλειες κατά τη φάση της συλλογής ή της επεξεργασίας.

Φυσικά, όσο μεγαλύτερη έκταση έχει μια έρευνα και όσο περισσότερα είναι τα αναμενόμενα δεδομένα, τόσο πιο πιθανή είναι η εμφάνιση του προβλήματος. Η έλλειψη δεδομένων σε μια έρευνα μειώνει τη δυνατότητα των αναλυτών να εξάγουν ασφαλή συμπεράσματα για το σύνολο, αφού πολλές φορές το δείγμα των δεδομένων καταλήγει να είναι πολύ μικρότερο από το αναμενόμενο. Επίσης, είναι πιθανό να υπάρξουν στατιστικές αποκλίσεις μέσω μείωσης των συντελεστών συσχέτισης, μεταβολής των δεικτών κεντρικής τάσης και των δεικτών διασποράς.

Έτσι, αναγκαστικά είτε επαναλαμβάνεται τμήμα της έρευνας με συνέπειες στον χρόνο και το κόστος διεξαγωγής της έρευνας, είτε γίνεται αποδοχή των ερωτηματολογίων με ελλιπή δεδομένα στην ανάλυση, με σαφή κίνδυνο το δείγμα να μην αντιπροσωπεύει το σύνολο.

Η απώλεια δεδομένων μπορεί να προκύψει στο δεύτερο και το τρίτο στάδιο της έρευνας, δηλαδή τόσο στη φάση της συλλογής όσο και στη φάση της επεξεργασίας. Οι λόγοι που μπορεί να οδηγήσουν σε ερωτηματολόγια με ελλιπή στοιχεία είναι πολλοί και μπορεί να προκύψουν είτε από λάθη σχεδιασμού και εκτέλεσης της έρευνας, είτε από τυχαίους παράγοντες οι οποίοι δεν είναι δυνατόν να προβλεφθούν και να αποφευχθούν. Μπορεί λοιπόν να γίνει μια αυθαίρετη κατηγοριοποίηση των λόγων για τους οποίους εμφανίζονται ελλιπή δεδομένα σε έρευνες ικανοποίησης σε: λόγους λανθασμένου σχεδιασμού, λόγους που οφείλονται στους εκτελούντες την έρευνα, λόγους που οφείλονται στους συμμετέχοντες στην έρευνα και τυχαίους λόγους.

Ο λανθασμένος σχεδιασμός της έρευνας μπορεί να επιφέρει πλήθος προβλημάτων ένα από τα οποία μπορεί να είναι η δημιουργία

βάσης δεδομένων με ελλιπή στοιχεία. Η λανθασμένη επιλογή ερωτήσεων, η επιλογή ακατάλληλου τρόπου για την εκτέλεση της έρευνας (π.χ. τηλεφωνική έρευνα, προσωπική συνέντευξη, ομαδική συνέντευξη κ.α.) και εν γένει η αδυναμία πρόβλεψης των παραγόντων που μπορεί να επηρεάσουν το αποτέλεσμα (καιρικές συνθήκες κατά την εκτέλεση, χρονικό πλαίσιο ανάλογα με το περιεχόμενο της έρευνας κ.α.) φέρουν άμεσο αντίκτυπο στον όγκο των συλλεχθέντων δεδομένων.

Για παράδειγμα, είναι σχεδόν σίγουρη η απώλεια δεδομένων σε έρευνες με μακροσκελή ή δυσνόητα ερωτηματολόγια, σε έρευνες που εκτελούνται σε ακατάλληλους χώρους (όπως σε περιβάλλον έντονου θορύβου, με έκθεσης σε καιρικά φαινόμενα κ.α.) ή σε τηλεφωνικές έρευνες που εκτελούνται σε ακατάλληλες ώρες.

Ομοίως, οι ερευνητές που καλούνται να συλλέξουν τα δεδομένα, φέρουν καθοριστικό ρόλο στην ομαλή εκτέλεση της έρευνας και στη συλλογή πλήρως συμπληρωμένων ερωτηματολογίων. Είναι σημαντικό, στη φάση σχεδιασμού να λαμβάνεται υπόψη η επιλογή των κατάλληλων προσώπων για την εκτέλεση της έρευνας, αφού άτομα αδιάφορα για την έκβαση της έρευνας, αγενή ως προς την προσέγγιση των συμμετεχόντων ή ανεκπαίδευτα για το ρόλο που καλούνται να παίξουν, είναι δεδομένο ότι θα συγκεντρώσουν πλήθος ερωτηματολογίων με ελλιπή στοιχεία. Για το σκοπό αυτό, είναι απαραίτητο οι εκτελούντες την έρευνα να έχουν τα σωστά κίνητρα ώστε να αποτελέσουν όχι απλά το μέσο εκτέλεσης της έρευνας, αλλά να συμβάλουν με τη συμπεριφορά και το ζήλο τους στη συμπλήρωση των ερωτηματολογίων από τους συμμετέχοντες στο μέγιστο βαθμό.

Η βέλτιστη δυνατή προετοιμασία της έρευνας παρόλα αυτά δεν διασφαλίζει ότι όλοι οι συμμετέχοντες είναι πρόθυμοι να απαντήσουν σε όλες τις απαντήσεις. Έτσι σε πολλές περιπτώσεις, ακόμα και σε ιδανικές συνθήκες εκτέλεσης της έρευνας, οι ερωτηθέντες είναι αυτοί οι οποίοι λόγω της ιδιαιτερότητας της προσωπικότητάς τους θα αποτελέσουν την αιτία συλλογής ερωτηματολογίων με ελλιπή δεδομένα.

Οι πιθανές περιπτώσεις για τις οποίες ένα άτομο μπορεί να μην απαντήσει σε μια ερώτηση, ενώ αρχικά έχει δηλώσει προθυμία να συμμετάσχει στην έρευνα είναι δύο, η απροθυμία για απάντηση ή η κατά λάθος παράληψη της ερώτησης. Η πιο συχνή περίπτωση είναι η πρώτη όπου ο ερωτώμενος είτε δεν θέλει να απαντήσει διότι η συγκεκριμένη ερώτηση είναι θέμα ταμπού και αφορά προσωπικές πληροφορίες που αποφεύγει να δώσει, είτε δεν κατανοεί πλήρως το περιεχόμενο της ερώτησης και την προσπερνά, είτε βιάζεται και παραλείπει επιλεγμένες ερωτήσεις ή ολόκληρο τμήμα του ερωτηματολογίου.

Τέλος, πάντοτε υπάρχουν οι τυχαίοι λόγοι απώλειας δεδομένων που δεν μπορούν να προβλεφθούν αλλά μπορεί να διαδραματίσουν σημαντικό ρόλο στην εξέλιξη της έρευνας. Κάποια παραδείγματα είναι η δημιουργία βλάβης στον εξοπλισμό του ερευνητή, η αλλαγή των περιβαλλοντολογικών συνθηκών, προβλήματα στη μεταφορά των δεδομένων κ.α.

Μια ιδιαίτερη περίπτωση είναι οι μακροχρόνιες έρευνες όπου για πλήθος αιτιών μπορεί να προκύψει έλλειψη δεδομένων, όπως απομάκρυνση των ερευνητών για προσωπικούς λόγους ή λόγους υγείας, μεταβολή των συνθηκών κάτω από τις οποίες εκτελείται η έρευνα, μεταβολή σε δυναμικούς τομείς του αντικειμένου της έρευνας όπου πλέον κάποιες ερωτήσεις καθίστανται ανενεργές κ.α.

2.6 Μηχανισμοί έλλειψης δεδομένων

Παρατηρείται λοιπόν ότι οι λόγοι έλλειψης δεδομένων σε μια έρευνα ενώ είναι δυνατό να προβλεφθούν και να ληφθούν μέτρα αντιμετώπισης του προβλήματος, είναι σχεδόν αδύνατο να εξαλειφθούν. Οι συνέπειες της δημιουργίας βάσης δεδομένων με ελλιπή στοιχεία κάνουν επιτακτική την ανάγκη ανάλυσης του προβλήματος και προσπάθειας επίλυσης του. Εφόσον οι ερευνητές αποδέχονται την εκτίμηση ότι έχουν ληφθεί τα απαραίτητα μέτρα για

μείωση του φαινομένου γίνεται μια πρώτη προσπάθεια ποιοτικής ανάλυσης.

Για την αποτελεσματικότερη αντιμετώπιση του προβλήματος, είναι απαραίτητο ο ερευνητής να έχει κατανοήσει σαφώς το μέγεθος του μέσω της πλήρους εικόνας των δεδομένων που συλλέχθηκαν, του ποσοστού των ελλιπών στοιχείων, της κατανομής τους κ.α. Για αυτό το λόγο τα συλλεχθέντα δεδομένα τοποθετούνται σε πίνακα όπου στις γραμμές του παρουσιάζονται οι συμμετέχοντες – μονάδες – περιπτώσεις και στις στήλες οι ερωτήσεις – μεταβλητές – κριτήρια. Κάθε κενό στοιχείο του πίνακα αντιπροσωπεύει έλλειψη απάντησης στη συγκεκριμένη ερώτηση. Με αυτό τον τρόπο ο ερευνητής μπορεί να προσδιορίσει ακόμα και οπτικά εάν τα ελλιπή δεδομένα συγκεντρώνονται σε κάποια ερώτηση ή σε συγκεκριμένα άτομα, δηλαδή εάν υπάρχει κάποιος μηχανισμός που να επηρεάζει τα δεδομένα της έρευνας και αναλόγως να αποφασίσει τον τρόπο επεξεργασίας τους.

Με τον όρο μηχανισμός ελλιπών στοιχείων εννοείται η παρουσία κάποιας αιτίας ή αιτιών που έχει ως αποτέλεσμα την εμφάνιση κενών στο σύνολο των δεδομένων. Οι πιθανοί μηχανισμοί που μπορεί να επηρεάζουν την έλλειψη δεδομένων χωρίζονται σε τελείως τυχαία έλλειψη, έλλειψη υπό προϋποθέσεις και σε έλλειψη εξ αιτίας συστηματικών λόγων. Η κατηγοριοποίηση των μηχανισμών έλλειψης δεδομένων εισήχθη από τον Donald B. Rubin (1987) και στην διεθνή ορολογία έχουν επικρατήσει ως: Missing Completely At Random (MCAR), Missing At Random (MAR) και Missing Not At Random (MNAR).

MCAR μηχανισμός έλλειψης υφίσταται όταν η έλλειψη δεν εξαρτάται από τις τιμές ούτε των ελλιπών, ούτε των παρατηρούμενων στοιχείων. Ουσιαστικά ο μηχανισμός MCAR απαιτεί η έλλειψη να είναι αποτέλεσμα εντελώς τυχαίας διαδικασίας, ώστε τόσο οι περιπτώσεις με ελλιπείς τιμές, όσο και οι περιπτώσεις χωρίς ελλιπείς τιμές να είναι τυχαία υποσύνολα του συνόλου των περιπτώσεων της έρευνας.

Υπάρχει βέβαια πιθανότητα η έλλειψη να είναι μηχανισμού MCAR ενώ σχετίζεται με κάποια μεταβλητή της έρευνας, όταν δεν

σχετίζεται με το σύνολο των παρατηρούμενων περιπτώσεων. Αυτή η περίπτωση αναφέρεται ως “στοιχειώδεις ή essential MCAR” (John W. Graham, 2012) και μπορεί να εμφανιστεί π.χ. όταν ασθενής που λαμβάνει μέρος σε ιατρική έρευνα, μετακομίζει λόγω δουλειάς και διακόπτει τη συμμετοχή του. Σε αυτή την περίπτωση, ενώ η έλλειψη σχετίζεται με τη μεταβλητή “τόπος διαμονής”, δεν σχετίζεται με το αντικείμενο της έρευνας και ο μηχανισμός έλλειψης είναι MCAR.

Σε αντίθεση με τον MCAR μηχανισμό όπου υπάρχει κοινή αντίληψη από τους ερευνητές για τον ορισμό του, στον **MAR** μηχανισμό υπάρχει ασάφεια και διαφορετικές απόψεις, αφού όλοι κατανοούν ότι η έλλειψη είναι αποτέλεσμα μιας τυχαίας διαδικασίας αλλά όχι εντελώς τυχαίας. Έτσι, ένας σχετικά ευρέως αποδεκτός ορισμός του MAR είναι όταν η έλλειψη εξαρτάται από το σύνολο παρατηρούμενων στοιχείων αλλά δεν εξαρτάται από τα ίδια τα ελλιπή δεδομένα. Σε αυτή την περίπτωση οι αιτίες της έλλειψης είναι γνωστές, ενδεχομένως να μπορούν να μετρηθούν και ο ερευνητής έχει δυνατότητα να τις ελέγξει αλλάζοντας παραμέτρους εκτέλεσης της έρευνας.

Κλασικό παράδειγμα είναι η συμπλήρωση μακροσκελούς ερωτηματολογίου σε ορισμένο χρονικό περιθώριο. Όσοι συμμετέχοντες διαβάζουν γρήγορα θα απαντήσουν σε όλες τις ερωτήσεις, ενώ αντίθετα οι πιο αργοί δεν θα προλάβουν. Ο ερευνητής γνωρίζοντας το δείγμα του πλήθους που συμμετέχει, μπορεί να ελέγξει την έλλειψη των δεδομένων μεταβάλλοντας το χρονικό περιθώριο συμπλήρωσης του ερωτηματολογίου. Έτσι, ενώ η έλλειψη εξαρτάται από τις παρατηρούμενες τιμές των παρατηρούμενων στοιχείων (ενδεχομένως το επίπεδο μόρφωσης), δεν εξαρτάται από τα ίδια τα ελλιπή στοιχεία.

Τέλος **MNAR** μηχανισμό έχουμε όταν οι αιτίες της έλλειψης σχετίζονται με τα ελλιπή στοιχεία. Η έλλειψη δηλαδή μπορεί να εξαρτάται είτε από το σύνολο των δεδομένων της έρευνας, είτε από μια συγκεκριμένη μεταβλητή της έρευνας, είτε από μια μεταβλητή που σχετίζεται με την έρευνα. Σε αυτή την περίπτωση οι αιτίες της έλλειψης δεν είναι δυνατό να γίνουν γνωστές και να μετρηθούν εκ των προτέρων.

Κλασικό παράδειγμα είναι οι ελλείψεις που παρουσιάζονται σε ερωτηματολόγια που περιέχουν ερωτήσεις για προσωπικά δεδομένα όπως εισόδημα, ηλικία κλπ. Η έλλειψη σε αυτή την περίπτωση οφείλεται στις εν λόγω μεταβλητές και οι αιτίες δεν μπορεί να μετρηθούν, ώστε να χρησιμοποιηθούν για ανάλυση από τον ερευνητή.

Η λεπτή διαφορά στον ορισμό των MAR και MNAR μηχανισμών, έχει προκαλέσει αντικρουόμενες απόψεις μεταξύ των επιστημόνων που ασχολούνται με τη στατιστική επιστήμη και πλήθος διαφορετικών ερμηνειών. Πολλοί μάλιστα ισχυρίζονται ότι ο ερευνητής δεν είναι ποτέ δυνατόν να έχει πρόσβαση σε όλες τις αιτίες ή συσχετίσεις που προκαλούν την έλλειψη, παρά μόνο εάν ακολουθήσει νέα έρευνα μετά την επεξεργασία της πρώτης στο ίδιο πλήθος. Για αυτό το λόγο έχει εκφραστεί η άποψη ότι είναι πολύ δύσκολο ο ερευνητής να γνωρίζει επακριβώς εάν ο μηχανισμός έλλειψης των δεδομένων του είναι MAR, παρά μόνο με τη μορφή υποθέσεων (Schafer and Graham 2002, Collins et al. 2001).

3. Αντιμετώπιση του προβλήματος

3.1 Αρχική αντιμετώπιση

Το πρόβλημα έχει αναγνωριστεί από τα πρώτα βήματα συλλογής δεδομένων μέσω έρευνας και έχει απασχολήσει σε τέτοιο βαθμό τους ερευνητές, ώστε πλέον να αποτελεί κομμάτι της στατιστικής επιστήμης με τον τίτλο Missing Data. Πολλοί ερευνητές έχουν απασχοληθεί με την επίλυση του προβλήματος και έχουν αναπτύξει διάφορες μεθόδους οι οποίες έχουν χρησιμοποιηθεί και εξελιχθεί έως σήμερα.

Η αρχική αντιμετώπιση του ζητήματος στηρίζεται στην απλή λογική ότι ένα κενό σε μια βάση δεδομένων είναι δυνατό, είτε να συμπληρωθεί με μια υποτιθέμενη τιμή, είτε να απαλειφθεί εντελώς από τους υπολογισμούς ώστε να μην επηρεαστούν τα αποτελέσματα. Με βάση αυτή τη λογική έχουν αναπτυχθεί τρεις κατηγορίες μεθόδων, οι μέθοδοι Διαγραφής (Deletion Methods), Στάθμισης (Weighting Methods) και Καταλογισμού (Imputation Methods). Οι παραπάνω παρουσιάζονται συχνά και ως «παραδοσιακές» ή «ειδικές» μέθοδοι επειδή χρησιμοποιήθηκαν σε μεγάλο βαθμό στο παρελθόν αλλά η χρήση τους έχει αναγνωριστεί ως ανεπαρκής στη σύγχρονη έρευνα.

Για αυτό το λόγο, στη σημερινή εποχή έχουν αναπτυχθεί οι λεγόμενες μοντέρνες μέθοδοι (Modern Methods), που στηρίζονται στην εφαρμογή συγκεκριμένων μοντέλων. Χρησιμοποιούνται σε μεγάλο βαθμό από τους σύγχρονους ερευνητές λόγω απαίτησης λιγότερων και ρεαλιστικότερων προϋποθέσεων από τις παραδοσιακές μεθοδολογίες ωστόσο, απαιτούν ιδιαίτερη προσοχή στην εφαρμογή των μοντέλων αλλά και μεγαλύτερη υπολογιστική ισχύ.

3.2 Παραδοσιακές μέθοδοι

Οι παραδοσιακές μέθοδοι αντιμετώπισης των ελλিপών δεδομένων έχουν ως βασική αρχή τη συμπλήρωση μιας πλήρους βάσης δεδομένων η οποία να μπορεί να αναλυθεί με τις κλασικές στατιστικές

μεθόδους, είτε μέσω διαγραφής, είτε μέσω συμπλήρωσης των κενών. Αυτή η αντιμετώπιση δεν θεωρείται αποτελεσματική καθώς σωστά αποτελέσματα υπάρχουν μόνο όταν υπάρχουν συγκεκριμένες προϋποθέσεις και κυρίως ο μηχανισμός της έλλειψης να είναι MCAR ή MAR. Παρακάτω παρουσιάζονται συνοπτικά οι κυριότερες μέθοδοι αυτής της κατηγορίας.

3.2.1 Μέθοδος Διαγραφής

Η μέθοδος διαγραφής αποτελεί τον πιο απλό τρόπο διαχείρισης των ελλিপών δεδομένων μέσω διαγραφής των κενών και της δημιουργίας πλήρους βάσης δεδομένων. Με βάση αυτή την λογική έχουν αναπτυχθεί δύο υπομέθοδοι, η μέθοδος πλήρως συμπληρωμένων περιπτώσεων (Complete Case Analysis - Listwise) και η μέθοδος διαθέσιμης περιπτωσιολογικής ανάλυσης (Available Case Analysis - Pairwise).

Η **Listwise** μέθοδος έχει χρησιμοποιηθεί κατά κόρον στο παρελθόν λόγω της απλότητας εφαρμογής της από τους ερευνητές και έχει ενσωματωθεί σε πολλές στατιστικές εφαρμογές. Η μέθοδος αυτή αφορά την διαγραφή όλων των περιπτώσεων που έχουν έστω και μια αναπάντητη ερώτηση, έτσι ώστε να σχηματιστεί μικρότερος πίνακας δεδομένων χωρίς κενά. Η χρήση της μεθόδου είναι ικανοποιητική σε μηχανισμούς έλλειψης MCAR, αλλά δίνει πολύ αμφιλεγόμενα αποτελέσματα σε MAR και ιδιαίτερα MNAR περιπτώσεις. Το κύριο μειονέκτημα της είναι η μείωση της ακρίβειας των αποτελεσμάτων λόγω μείωσης του προσδοκώμενου δείγματος και η πιθανότητα δημιουργίας αποκλίσεων. Είναι χαρακτηριστικό ότι όσο περισσότερες είναι οι μεταβλητές, τόσο μεγαλύτερη είναι και η μείωση του όγκου των διαθέσιμων δεδομένων (Rubin - Little 2002), ενώ όπως παρουσιάζουν οι Kim και Curry (1977), το 59% των δεδομένων σε μια έρευνα μπορεί να χαθεί εάν σε σύνολο δεδομένων με πέντε μεταβλητές χαθεί τυχαία το 10% των δεδομένων από κάθε μεταβλητή.

Η **Pairwise** μέθοδος είναι μια εναλλακτική μέθοδος διαγραφής που αναπτύχθηκε για να περιορίσει το φαινόμενο απώλειας μεγάλου όγκου δεδομένων της Listwise και χρησιμοποιήθηκε επίσης ευρέως σε πολλές στατιστικές εφαρμογές. Σύμφωνα με τους Kim και Curry (1977), η εν λόγω μέθοδος είναι μια ελκυστική εναλλακτική λύση όταν υπάρχει μικρός αριθμός ελλειπόντων δεδομένων σε κάθε μεταβλητή σε σχέση με το συνολικό μέγεθος δείγματος, καθώς και μεγάλος αριθμός μεταβλητών. Σε αντίθεση με την προηγούμενη μέθοδο, χρησιμοποιούνται αρχικά όλες οι διαθέσιμες περιπτώσεις και δημιουργούνται συσχετίσεις μεταξύ των μεταβλητών. Στη συνέχεια, κατά την ανάλυση διαγράφονται κάθε φορά μόνο οι περιπτώσεις που έχουν κενά σε μια από τις συσχετίσεις μεταβλητών που χρησιμοποιούνται. Έτσι υπολογίζονται στατιστικές παράμετροι για διαφορετικά υποσύνολα κάθε φορά, δημιουργώντας προβλήματα όσον αφορά τους υπολογισμούς των κοινών παραμέτρων του αρχικού συνόλου, ιδιαίτερα σε έρευνες με μεγάλο όγκο δεδομένων. Η μέθοδος Pairwise, όπως και η Listwise, δίνει ικανοποιητικά αποτελέσματα για μηχανισμούς MCAR αλλά αναξιόπιστα για μηχανισμούς MAR και MNAR.

3.2.2 Διαδικασίες στάθμισης

Οι διαδικασίες στάθμισης (Weighting Procedures) είναι ένα πακέτο μεθόδων οι οποίες έχουν ως σκοπό τη μείωση της απόκλισης των αποτελεσμάτων, μέσω εισαγωγής βαρών στα δεδομένα του προβλήματος. Η εισαγωγή βαρών γίνεται για να ληφθεί υπόψη η διαφορετική πιθανότητα παρατήρησης κάποιας τιμής ανάλογα με την εξεταζόμενη περίπτωση κενής απάντησης. Ουσιαστικά με αυτό τον τρόπο, τα ελλιπή δεδομένα αντιμετωπίζονται σαν μέρος του σχεδιασμού της έρευνας.

Η μέθοδος περιλαμβάνει τρία στάδια, όπου στο πρώτο παράγεται η βάση βαρών με γνώμονα τη διαφορετική πιθανότητα επιλογής των στοιχείων του δείγματος, στο δεύτερο στάδιο γίνεται προσαρμογή των βαρών σύμφωνα με κάποια μέθοδο όπως την προσαρμογή

σταθμισμένων ομάδων (Weighting Class Adjustment) και στο τρίτο στάδιο ρυθμίζονται περαιτέρω τα βάρη των παρατηρούμενων τιμών ώστε να γίνει εκτίμηση ορισμένων τιμών του συνόλου του πληθυσμού.

Η διαδικασία της στάθμισης είναι εύκολα εφαρμόσιμη όταν η μορφή των ελλিপών στοιχείων είναι μονότονη ή μονομεταβλητή, αλλά αντιθέτως είναι πολύ δύσκολα εφαρμόσιμη σε τυχαία μορφή έλλειψης, καθώς πρέπει να υπολογιστεί διαφορετικό σύνολο βαρών για κάθε μεταβλητή. Οι διαδικασίες στάθμισης χρησιμοποιούνται σε συνδυασμό με την Listwise μέθοδο διαγραφής και δίνουν ικανοποιητικά αποτελέσματα σε μηχανισμούς έλλειψης MCAR και MAR.

3.2.3 Μέθοδοι καταλογισμού

Οι μέθοδοι καταλογισμού βασίζονται στην λογική εύρεσης κατάλληλων τιμών για συμπλήρωση των κενών που έχουν προκύψει σε ένα πρόβλημα ελλিপών δεδομένων και αποσκοπούν στη δημιουργία πλήρους σετ δεδομένων, το οποίο να μπορεί να αναλυθεί με τις κλασικές στατιστικές μεθόδους. Η φιλοσοφία των μεθόδων αυτών στηρίζεται στο γεγονός ότι οι περισσότερες έρευνες εκτελούνται προς αναζήτηση μεγεθών που αφορούν το σύνολο του πληθυσμού και άρα οι μεμονωμένες περιπτώσεις δεν είναι ο αυτοσκοπός της έρευνας. Έτσι, σκοπός του καταλογισμού δεν είναι η αναζήτηση των βέλτιστων πιθανών προβλέψεων για κάθε περίπτωση, αλλά η αντικατάσταση του κενού με πιθανές τιμές που δεν θα αλλοιώσουν τα στατιστικά μεγέθη του πληθυσμού (Little & Rubin, 2002).

Η αναζήτηση της κατάλληλης τιμής γίνεται με βάση την υπόθεση ότι στα κενά ενός ερωτηματολογίου είναι δυνατή η τοποθέτηση μιας απάντησης που είναι πιθανό να είχε δοθεί από τον ερωτώμενο. Η αναζήτηση κατάλληλης τιμής γίνεται είτε μέσα από τις άλλες απαντήσεις που έχουν δοθεί για τη συγκεκριμένη μεταβλητή, είτε από πληροφορίες που προκύπτουν από την κάθε περίπτωση. Γίνεται δηλαδή αντικατάσταση του κενού με μια τιμή που προκύπτει από άλλα

δεδομένα της έρευνας, για αυτό το λόγο οι μέθοδοι καταλογισμού αναφέρονται και ως μέθοδοι αντικατάστασης.

Τα πλεονεκτήματα των μεθόδων καταλογισμού είναι η ευκολία χρήσης τους και η διατήρηση του μεγέθους του δείγματος. Η βέλτιστη δυνατή επίλυση του προβλήματος θα είχε ως αποτέλεσμα τη συμπλήρωση όλων των κενών, χωρίς να επηρεαστούν τα στατιστικά μεγέθη, δίνοντας τη δυνατότητα στον ερευνητή να χρησιμοποιήσει όλα τα ερωτηματολόγια στην ανάλυση. Όμως, η αντικατάσταση των κενών απαιτεί ιδιαίτερη προσοχή γιατί μπορεί να προκαλέσει μεγάλες αποκλίσεις στα μεγέθη του δείγματος από την πραγματική εικόνα.

Έχουν αναπτυχθεί πολλές μέθοδοι καταλογισμού από τους ερευνητές, οι οποίες μπορεί να καταταχθούν με βάση τον τρόπο αντικατάστασης των ελλιπών στοιχείων σε Μεθόδους μέσης τιμής (Mean Based Methods), Μεθόδους παλινδρόμησης (Regression Methods) και Μεθόδους Hot Deck.

Οι μέθοδοι μέσης τιμής επιλέγουν την συμπλήρωση των κενών με κάποια μέση τιμή που προκύπτει από τις παρατηρούμενες τιμές για αυτή τη μεταβλητή. Ο τρόπος επιλογής της τιμής αυτής μπορεί να διαφέρει ανά περίπτωση και μπορεί να γίνει, είτε με επιλογή της μέσης τιμής από το σύνολο όλων των τιμών της συγκεκριμένης μεταβλητής, είτε με επιλογή κάποιας μέσης τιμής από ένα υποσύνολο των παρατηρούμενων τιμών.

Ο δεύτερος τρόπος αντικατάστασης προϋποθέτει την ανάλυση και τον διαχωρισμό των περιπτώσεων σε ομάδες χρησιμοποιώντας βοηθητικές μεταβλητές, όπου υπολογίζεται μέση τιμή για κάθε ομάδα. Έτσι, όταν παρατηρείται ελλιπές στοιχείο, τοποθετείται η μέση τιμή με βάση την ομάδα στην οποία ανήκει η μεταβλητή. Η εφαρμογή των μεθόδων μέσης τιμής δίνει σωστές εκτιμήσεις μόνο σε MCAR μηχανισμούς έλλειψης, αλλά οι υπολογισμοί λοιπών στατιστικών μεγεθών όπως της διασποράς και της συνδιακύμανσης δεν μπορούν να θεωρούνται έγκυροι λόγω υποεκτίμησης της μεταβλητότητας των τιμών που τοποθετούνται.

Οι μέθοδοι παλινδρόμησης παράγουν τιμές προερχόμενες από ανάλυση παλινδρόμησης για κάθε μια μεταβλητή, η οποία στηρίζεται μόνο στις παρατηρούμενες τιμές της ίδιας μεταβλητής. Η διαδικασία πραγματοποιείται σε δύο στάδια όπου αρχικά εκτιμώνται οι σχέσεις μεταξύ όλων των μεταβλητών και στη συνέχεια χρησιμοποιούνται οι συντελεστές παλινδρόμησης για την εκτίμηση των ελλিপών στοιχείων. Οι εν λόγω μέθοδοι παράγουν πολύ καλά αποτελέσματα για MCAR μηχανισμούς και ικανοποιητικά καλά για MAR. Παρόλα αυτά παραμένει το πρόβλημα υποεκτίμησης της διασποράς όπως και στις μεθόδους μέσης τιμής.

Οι μέθοδοι Hot Deck πήραν το όνομα τους από τις στοίβες διάτρητων καρτών αποθήκευσης πληροφοριών που βρίσκονταν υπό επεξεργασία (hot stack) και υποδηλώνει ότι ο δέκτης μιας τιμής – πληροφορίας παίρνει την τιμή από ένα πλήθος (στοίβα) πιθανών δοτών. Στη στατιστική επιστήμη και στο πρόβλημα των ελλিপών δεδομένων, η μέθοδος hot deck υλοποιείται καλύπτοντας τα κενά της βάσης δεδομένων δανείζοντας τιμές από άλλη “όμοια” περίπτωση που παρατηρείται τιμή για την ίδια μεταβλητή.

Για να επιτευχθεί αυτό, οι εν λόγω μέθοδοι χρησιμοποιούνται σε συνδυασμό με μεθόδους διαγραφής. Αρχικά διαχωρίζονται μόνο οι περιπτώσεις που παρατηρείται τιμή στη συγκεκριμένη μεταβλητή και στη συνέχεια αναζητείται περίπτωση με παρόμοια χαρακτηριστικά στις υπόλοιπες μεταβλητές. Η περίπτωση που βρεθεί να έχει την μεγαλύτερη ομοιότητα με αυτή που περιέχει την ελλιπή τιμή επιλέγεται ως δότης της νέας τιμής. Το πλεονέκτημα της μεθόδου είναι ότι η τιμή που τοποθετείται στο κενό είναι πραγματική και με αυτό τον τρόπο μειώνεται το πρόβλημα υποεκτίμησης της διασποράς των προηγούμενων μεθόδων. Παρόλα αυτά απαιτείται μεγάλη προσοχή από τον ερευνητή κατά τον καθορισμό των κριτηρίων για την εύρεση των όμοιων περιπτώσεων.

Οι μέθοδοι hot deck αν και έχουν εφαρμοστεί ευρέως στην πράξη, ακόμα και από μεγάλους φορείς στατιστικών ερευνών (π.χ. U.S. Census Bureau) δεν έχουν υποστηριχθεί από ισχυρό θεωρητικό

υπόβαθρο, σε αντιστοιχία με τις άλλες μεθόδους καταλογισμού, για την υλοποίησή τους. Το κύριο πρόβλημα με τη θεωρία αυτή, έγκειται στο ότι δεν υπάρχει ένας αντικειμενικός τρόπος υπολογισμού της βέλτιστης περίπτωσης από την οποία θα γίνει ο καταλογισμός, δηλαδή δεν υπάρχουν κοινώς αποδεκτά κριτήρια που να ορίζουν την ομοιότητα των περιπτώσεων, αλλά στηρίζεται στην υποκειμενικότητα του ερευνητή που θα τα επιλέξει. Λόγω όμως της ελκυστικότητας και της απλής εφαρμογής, οι μέθοδοι αυτοί έχουν χρησιμοποιηθεί κατά κόρον στο παρελθόν και έχουν δοθεί ονόματα ανάλογα με τον τρόπο υλοποίησης εύρεσης της κατάλληλης τιμής, όπως τυχαία hot deck, διαδοχική hot deck, ιεραρχική hot deck, κ.α.

3.3 Σύγχρονες μέθοδοι

Οι σύγχρονες μέθοδοι θεωρούνται ανώτερες από τις παραδοσιακές διότι παράγουν εκτιμητές των παραμέτρων με αποδεκτά τυπικά σφάλματα. Οι εν λόγω μέθοδοι βασίζονται σε συγκεκριμένα μοντέλα και παρόλο που παρουσιάζουν δυσκολία στην εφαρμογή τους, αποτελούν πολύ σημαντική εξέλιξη στο πρόβλημα των ελλιπών δεδομένων. Παρακάτω παρουσιάζονται συνοπτικά οι κυριότερες μέθοδοι αυτής της κατηγορίας.

3.3.1 Πολλαπλός Καταλογισμός (Multiple Imputation)

Ο πολλαπλός καταλογισμός είναι ουσιαστικά μια διαδικασία καταλογισμού που συνδυάζει τόσο μεθόδους αντικατάστασης όσο και μεθόδους στηριζόμενες σε μοντέλα. Η διαδικασία όπως έχει περιγραφεί και από τον Rubin (1996), εκτελείται σε τρία βήματα.

Αρχικά (Imputation Step) δημιουργούνται σύνολα πιθανών τιμών με χρήση μοντέλων παλινδρόμησης, τα οποία είναι δυνατό να χρησιμοποιηθούν σε κάθε ελλιπή τιμή κάθε μεταβλητής. Έχει προταθεί μεγάλος αριθμός μοντέλων απόδοσης ανάλογα με το είδος των μεταβλητών της έρευνας, τα οποία χρησιμοποιούνται σε επαναληπτικές

διαδικασίες έως ότου δημιουργηθούν τα σύνολα καταλογισμού. Με βάση τα σύνολα αυτά, συμπληρώνονται οι ελλιπείς τιμές ώστε να παραχθούν όλα τα πιθανά πλήρως συμπληρωμένα σύνολα της βάσης δεδομένων.

Στη συνέχεια (Analysis Step) κάθε ένα από τα παραχθέντα πλήρη σύνολα αναλύεται με χρήση συνηθισμένων στατιστικών μεθόδων. Έτσι για κάθε σύνολο τιμών εκτιμώνται οι απαραίτητες στατιστικές παράμετροι και τα τυπικά σφάλματα.

Στο τελευταίο βήμα (Combination Step) γίνεται συνδυασμός των αποτελεσμάτων με σκοπό την παραγωγή μιας κοινής εκτίμησης. Για το σκοπό αυτό έχουν προταθεί διάφοροι τρόποι συνδυασμού, από τις οποίες πολύ συχνά χρησιμοποιείται η μέθοδος μονοδιάστατης παραμέτρου του Rubin (1987).

Το βασικό πλεονέκτημα του πολλαπλού καταλογισμού είναι ότι αντιμετωπίζει επιτυχώς το πρόβλημα των υπόλοιπων μεθόδων καταλογισμού, όπου η λύση θεωρείται γνωστή και προβλέψιμη από τη αρχή χωρίς καμία αβεβαιότητα, δημιουργώντας μια πιο έγκυρη πρόβλεψη η οποία έχει προέλθει μετά από ανάλυση και συνδυασμό αποτελεσμάτων. Το βασικό μειονέκτημα της μεθόδου είναι ότι απαιτεί πολλές πράξεις για την δημιουργία και ανάλυση όλων των πιθανών συνόλων, πρόβλημα το οποίο στις μέρες μας έχει ξεπεραστεί με την αύξηση της υπολογιστικής ισχύος για αυτό και είναι συχνή η χρήση της σε πολλά σύγχρονα προγράμματα.

3.3.2 Εκτίμηση Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation)

Η εκτίμηση μέγιστης πιθανοφάνειας είναι στατιστική μέθοδος αναζήτησης των τιμών των παραμέτρων ενός μοντέλου και χρησιμοποιείται ανεξάρτητα από την έλλειψη ή όχι δεδομένων. Κύριος άξονας της διαδικασίας είναι ο προσδιορισμός των στατιστικών παραμέτρων όπως ο μέσος ή η διασπορά, χωρίς να είναι γνωστές οι τιμές ολόκληρου του πληθυσμού παρά μόνο ενός δείγματος. Αυτό

επιτυγχάνεται αρχικά με τον καθορισμό ενός μοντέλου που χαρακτηρίζει τα δεδομένα του δείγματος και στη συνέχεια υπολογίζονται οι τιμές των παραμέτρων που μεγιστοποιούν τη συνάρτηση πιθανότητας.

Στο πρόβλημα των ελλιπών δεδομένων ακολουθείται η ίδια διαδικασία και οι ελλιπείς τιμές εκτιμώνται μετά την εκτίμηση των παραμέτρων του μοντέλου που χρησιμοποιείται. Για να επιτευχθεί αυτό έχει δημιουργηθεί ο αλγόριθμος EM (Expectation Maximization), ο οποίος μέσω επαναληπτικών διαδικασιών δίνει εκτιμήσεις για τις ελλιπείς τιμές και τις τρέχουσες εκτιμώμενες παραμέτρους μέχρι να υπάρξει σύγκλιση. Η μέθοδος αυτή θεωρείται πολύ αξιόπιστη για μηχανισμούς έλλειψης MCAR και MAR αλλά απαιτεί μεγάλα δείγματα πληθυσμού για να δώσει καλά αποτελέσματα και στις μέρες μας χρησιμοποιείται από πολλά σύγχρονα προγράμματα.

3.3.3 Μπεϋσιανές Μέθοδοι (Bayesian Methods)

Οι μέθοδοι αυτές χρησιμοποιούνται τα τελευταία χρόνια κυρίως σε περιπτώσεις που το δείγμα είναι μικρό και δεν είναι δυνατή η χρήση της μέγιστης πιθανοφάνειας. Δίνει πολύ ακριβή αποτελέσματα σε μηχανισμούς έλλειψης MCAR και MAR και σχετικά καλά για αποτελέσματα MNAR.

Οι Μπεϋσιανές μέθοδοι εφαρμόζουν επαναληπτικές διαδικασίες όπου τα ελλιπή στοιχεία αντικαθίστανται από εκτιμήσεις με συγκεκριμένο τρόπο, ανάλογα το μοντέλο που επιλέγεται (Data Augmentation, Gibb's Sampler) μέχρι να υπάρξει σύγκλιση. Η σύγκλιση μπορεί να είναι πολύ χρονοβόρα και απαιτεί μεγάλη υπολογιστική ισχύ, αλλά δίνουν μεγαλύτερη ακρίβεια σε σχέση με τις υπόλοιπες μεθόδους, ιδιαίτερα σε σχέση αντιμετώπισης προβλημάτων με περιορισμένο δείγμα.

4. Ανάπτυξη και αξιολόγηση αλγόριθμου

4.1 Θεωρητικό υπόβαθρο

Όπως έχει αναφερθεί στο πρώτο κεφάλαιο, στόχος της παρούσας εργασίας είναι η ανάπτυξη αλγορίθμου hot deck imputation για συμπλήρωση των ελλιπών δεδομένων σε έρευνες με ordinal data και στη συνέχεια ο προγραμματισμός του σε κατάλληλη γλώσσα. Η φιλοσοφία της μεθόδου hot deck αποτελεί τη συμπλήρωση των κενών στοιχείων μέσω εύρεσης “όμοιων” ερωτηματολογίων που έχουν συμπληρωμένη την εν λόγω απάντηση.

Για την παρούσα εργασία στο εξής ορίζεται ο πίνακας αποτελεσμάτων έρευνας με ordinal δεδομένα $X(n,m)$, όπου n το πλήθος των περιπτώσεων - ερωτηματολογίων και m το πλήθος των μεταβλητών-κριτηρίων. Το κάθε στοιχείο του πίνακα x_{ij} αντιπροσωπεύει την απάντηση του i ερωτώμενου στο j κριτήριο.

Ένας απλός αλγόριθμος που αντιπροσωπεύει την μεθοδολογία hot deck είναι ο παρακάτω:

Εισαγωγή από τον χρήστη στοιχείων έρευνας n ερωτηματολογίων με m ερωτήσεις-κριτήρια (Πίνακας $X(n,m)$)

Εάν βρεις κενό στην απάντηση j του i ερωτηματολόγιου ($x_{ij} = \text{"κενό"}$)

Βρες "όμοιο" ερωτηματολόγιο k χωρίς κενό στην απάντηση j ($x_{kj} \neq \text{"κενό"}$)

Κάνε αντικατάσταση του κενού της απάντησης j του i με την τιμή της απάντησης j του k ($x_{ij} \rightarrow x_{kj}$).

Παρακάτω παρατίθεται παράδειγμα χρήσης του αλγόριθμου σε υποτιθέμενη έρευνα σε τρία άτομα με πέντε ερωτήσεις.

Πίνακας 4.1: Παράδειγμα έρευνας 3x5 με ένα κενό

	Απάντηση 1	Απάντηση 2	Απάντηση 3	Απάντηση 4	Απάντηση 5
Άτομο 1	Πάρα Πολύ	Πάρα Πολύ	-	Μέτρια	Πάρα Πολύ
Άτομο 2	Πολύ	Μέτρια	Πολύ	Πολύ	Πολύ
Άτομο 3	Πολύ Λίγο	Πολύ Λίγο	Πολύ Λίγο	Μέτρια	Πολύ Λίγο

Η μέθοδος hot deck υλοποιώντας τον παραπάνω αλγόριθμο, μετά την εισαγωγή των δεδομένων και την αναγνώριση του κενού στο στοιχείο x_{13} , θα αναζητήσει άλλο άτομο που παρουσιάζει “όμοια” συμπεριφορά στις απαντήσεις του (προφανώς χωρίς να έχει κενό στην εν λόγω μεταβλητή) για να συμπληρώσει την τιμή της μεταβλητής 3.

Οπτικά και μόνο διακρίνεται ότι τα άτομα 1 και 2 έχουν χαρακτηριστικά γενικώς θετικά για τα αντικείμενα της έρευνας, ενώ το άτομο 3 έχει αρκετά αρνητική άποψη. Άρα ο αλγόριθμος της μεθόδου θα αναγνωρίσει μεγαλύτερη ομοιότητα στις περιπτώσεις 1 και 2 και η ελλιπής τιμή θα συμπληρωθεί από την αντίστοιχη του ατόμου 2 ($x_{13} \rightarrow$ “Πολύ”).

Η μεθοδολογία αυτή προσπαθεί να αποφύγει τα προβλήματα των κλασικών μεθόδων αντικατάστασης μέσω αναζήτησης της “ομοιότητας” δύο ερωτηματολογίων. Προφανώς, η μέτρηση της “ομοιότητας” δεν μπορεί να γίνει οπτικά όπως στο παράδειγμα 4.1, αλλά απαιτείται εφαρμογή μεθόδων μέτρησης κάποιων δεικτών. Με μια απλή ταξινόμηση των δεικτών γίνεται εφικτή η εύρεση των πιθανών ερωτηματολογίων από τα οποία θα γίνει η αντικατάσταση.

Όπως προαναφέρθηκε, δεν υπάρχουν κοινώς αποδεκτοί δείκτες στη θεωρία της μεθόδου που να καλύπτουν όλες τις περιπτώσεις και ο ερευνητής καλείται να προσδιορίσει το βέλτιστο δυνατό τρόπο κάθε

φορά. Παρόλα αυτά, ένας κλασικός δείκτης που χρησιμοποιείται για την εύρεση όμοιων ερωτηματολογίων είναι η απόσταση των δεδομένων του συνόλου.

Η μέτρηση της απόστασης σε σύνολα αριθμητικών δεδομένων είναι εύκολη διαδικασία με πλήθος απλούστερων ή πιο σύνθετων μεθόδων που μπορεί να χρησιμοποιηθούν. Κλασική μέθοδος μέτρησης απόστασης αριθμητικών δεδομένων (numerical/metric) σε ίδια κλίμακα είναι η χρήση της Ευκλείδειας απόστασης. Η Ευκλείδεια απόσταση σε δύο διανύσματα \vec{a}, \vec{b} n μεταβλητών εκφράζεται ως:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2}$$

Σε αυτή τη μέθοδο μικρότερο αποτέλεσμα μέτρησης μεταφράζεται σε μεγαλύτερη ομοιότητα μεταξύ των δύο συνόλων δεδομένων. Μηδενική απόσταση λοιπόν σημαίνει ότι τα στοιχεία των δύο συνόλων είναι ακριβώς τα ίδια.

Για τα ordinal δεδομένα όμως, δεν είναι δυνατή η σύγκριση ερωτηματολογίων μέσω εφαρμογής ανάλογων μεθόδων μέτρησης απόστασης. Τα ordinal δεδομένα δεν είναι εφικτό να κωδικοποιηθούν με τρόπο ώστε να χρησιμοποιηθούν μέθοδοι των αριθμητικών δεδομένων για τη μέτρηση απόστασης αφού δεν έχει υπόσταση η έννοια της αριθμητικής κλίμακας μεταξύ τους. Δεν υφίσταται δηλαδή η έννοια τρεις φορές πιο ευχαριστημένος ή 30% της βαθμολόγησης κάποιου αντικειμένου ως μέτριο. Οι μόνες πράξεις που είναι εφικτές σε αυτή την περίπτωση είναι ισότητας - ανισότητας και μεγαλύτερο ή μικρότερο από.

4.2 Αρχική Προσέγγιση Υπολογισμού Απόστασης

Το πρώτο ζήτημα που προκύπτει στην υλοποίηση της μεθόδου hot deck σε ordinal δεδομένα, είναι η αναζήτηση αξιόπιστου τρόπου μέτρησης της απόστασης δύο περιπτώσεων. Για το σκοπό αυτό έχουν αναπτυχθεί διάφορες προσεγγίσεις χωρίς να υπάρχει μια κοινώς

αποδεκτή από όλη την επιστημονική κοινότητα. Ακόμα είναι δυνατόν ο κάθε ερευνητής να εφαρμόζει μια δικιά του προσέγγιση μέτρησης απόστασης προσαρμοσμένη στα δεδομένα που έχει να επεξεργαστεί.

Μια απλή, αυθαίρετη προσέγγιση έκφρασης της απόστασης σε ordinal δεδομένα θα μπορούσε να γίνει μέσω αναζήτησης του πλήθους των κοινών στοιχείων. Το αποτέλεσμα της πράξης:

$$d_{ik} = 1 - \frac{\text{κοινά στοιχεία}}{n}$$

όπου d_{ik} η απόσταση των περιπτώσεων i και k , θα μπορούσε να εκφράσει μια έννοια απόστασης μεταξύ των δύο συγκρινόμενων συνόλων ordinal δεδομένων. Η κωδικοποίηση εκτέλεσης των πράξεων μεταξύ δύο περιπτώσεων του πίνακα $X(n, m)$ σε αυτή την προσέγγιση δίνεται μέσω των εκφράσεων “όμοια” $(x_{ij} = x_{kj}) \rightarrow 1$ και “διαφορετικά” $(x_{ij} \neq x_{kj}) \rightarrow 0$. Δηλαδή δύο πανομοιότυπες περιπτώσεις θα έχουν απόσταση 0 αφού το σύνολο των κοινών στοιχείων θα είναι n . Αντίστοιχα δύο περιπτώσεις με αποκλειστικά διαφορετικά στοιχεία θα έχουν απόσταση 1 αφού το σύνολο των κοινών στοιχείων θα είναι 0.

Το πρόβλημα που προκύπτει με αυτή την κωδικοποίηση είναι ότι οι συγκρίσεις όλων των διαφορετικών στοιχείων x εξομοιώνονται ως προς το δυαδικό αποτέλεσμα “όχι”, δηλαδή δεν υφίσταται ταξινόμηση της διαφορετικότητας των x ως προς την ordinal κλίμακα. Όλες οι συγκρίσεις διαφορετικών στοιχείων $(x_{ij} \neq x_{kj})$, συνεισφέρουν το ίδιο στον υπολογισμό της απόστασης (0), ανεξάρτητα από τη θέση τους στην κλίμακα ταξινόμησης. Με λίγα λόγια, σε μια τόσο απλοϊκή προσέγγιση δεν λαμβάνεται υπόψη η εγγύτητα μεταξύ των απαντήσεων, η οποία αν και δεν είναι δυνατόν να μετρηθεί στις ordinal κλίμακες δεν μπορεί να αγνοηθεί.

Για παράδειγμα, σε μια ordinal κλίμακα του τύπου:

Πολύ λίγο – Λίγο – Μέτρια – Πολύ – Πάρα πολύ

είναι προφανές ότι η απάντηση “Λίγο” είναι εγγύτερη στην απάντηση “Μέτρια” παρά στην απάντηση “Πάρα πολύ”. Σύμφωνα με την αρχική προσέγγιση υπολογισμού απόστασης, η σύγκριση στοιχείων μεταξύ “Λίγο” – “Μέτρια” και “Λίγο” – “Πάρα πολύ” θα δώσει αποτέλεσμα “διαφορετικά”, δηλαδή 0 και στις δύο περιπτώσεις και θα έχουν την ίδια βαρύτητα. Αυτό προφανώς δεν αντικατοπτρίζει την πραγματικότητα αφού οι απαντήσεις “Λίγο” και “Μέτρια” έχουν μεγαλύτερη εγγύτητα από τις απαντήσεις “Λίγο” και “Πάρα πολύ” και θα έπρεπε να συνεισφέρουν διαφορετικά (μεγαλύτερη τιμή) στον αριθμητή του λόγου $\frac{\text{κοινά στοιχεία}}{n}$.

Σύμφωνα με αυτή την προσέγγιση υπολογισμού στην έρευνα του παραδείγματος 4.1, η απόσταση μεταξύ του πρώτου με το δεύτερο ερωτηματολόγιο θα είναι $d_{12} = 1 - \frac{0}{5} = 1$, ενώ με το τρίτο $d_{13} = 1 - \frac{1}{4} = 0,75$. Άρα η υλοποίηση της μεθόδου hot deck θα προτιμήσει για απόδοση την τιμή “Πολύ Λίγο” του τρίτου ερωτηματολογίου, πράγμα όμως που έρχεται σε αντίθεση με την κοινή λογική, αφού τα άτομα 1 και 2 έχουν περισσότερη εγγύτητα στις απαντήσεις τους από όσο τα 1 και 3. Η εκτίμηση του πρώτου ατόμου είναι σε γενικές γραμμές αρκετά θετική για το προϊόν της έρευνας άρα, εφόσον αποκλειστεί η ύπαρξη MNAR μηχανισμού έλλειψης, θα περίμενε κανείς να αποδοθεί η τιμή “Πολύ” του δεύτερου ερωτηματολογίου όπου έχει μια πιο θετική εκτίμηση από το τρίτο όπου οι περισσότερες απαντήσεις είναι “Πολύ λίγο”.

4.3 Απόσταση Ordinal Δεδομένων (Μέθοδος Walesiak)

Παρατηρείται λοιπόν ότι κατά την υλοποίηση της μεθόδου hot deck σε ordinal δεδομένα, απαιτείται μια προσεκτική και αξιόπιστη προσέγγιση υπολογισμού αποστάσεων, καθώς η εγγύτητα των απαντήσεων άρα και η ομοιότητα των ερωτηματολογίων είναι κρίσιμος παράγοντας για τη σωστή συμπλήρωση των κενών.

Για την επίλυση του ζητήματος έχουν αναπτυχθεί διάφορες μέθοδοι όπου η μέτρηση γίνεται μέσω συγκρίσεων των ordinal

στοιχείων, τα οποία πρώτα έχουν κωδικοποιηθεί και βαθμονομηθεί με αριθμητικές τιμές, ώστε να είναι εύκολη η εκτέλεση πράξεων σε Η/Υ. Η σύγκριση αριθμητικών δεδομένων προσφέρει τη δυνατότητα όχι μόνο ισότητας – ανισότητας αλλά και ταξινόμησης με βάση την τιμή.

Για τους σκοπούς ανάπτυξης αλγόριθμου hot deck με χρήση σε ordinal δεδομένα στην παρούσα εργασία, χρησιμοποιήθηκε η μέθοδος Walesiak (1993). Σε αυτή τη μέθοδο η απόσταση προκύπτει από τον παρακάτω τύπο:

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^m a_{ikj} b_{kij} + \sum_{j=1}^m \sum_{l=1}^n a_{ilj} b_{klj}}{\sum_{j=1}^m \sum_{l=1, l \neq i, k}^n a_{ilj} b_{klj}},$$

$$2 \left[\left(\sum_{j=1}^m a_{ikj}^2 + \sum_{j=1}^m \sum_{l=1, l \neq i, k}^n a_{ilj}^2 \right) \left(\sum_{j=1}^m b_{kij}^2 + \sum_{j=1}^m \sum_{l=1, l \neq i, k}^n b_{klj}^2 \right) \right]^{\frac{1}{2}},$$

Όπου:

$$a_{ipj} = \begin{cases} 1, & \text{if } x_{ij} > x_{pj} \\ 0, & \text{if } x_{ij} = x_{pj} \\ -1, & \text{if } x_{ij} < x_{pj} \end{cases} \quad \text{για } p = k, l$$

$$b_{krj} = \begin{cases} 1, & \text{if } x_{kj} > x_{rj} \\ 0, & \text{if } x_{kj} = x_{rj} \\ -1, & \text{if } x_{kj} < x_{rj} \end{cases} \quad \text{για } r = i, l$$

και

$i, k, l = 1, \dots, n$ αριθμός περιπτώσεων
 $j = 1, \dots, m$ αριθμός ordinal μεταβλητών
 x_{ij}, x_{kj}, x_{lj} τιμή της j μεταβλητής στην i, k, l περίπτωση

Οι εκφράσεις:

$$\sum_{j=1}^m a_{ikj}^2 + \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq i,k}}^n a_{ilj}^2$$

$$\sum_{j=1}^m b_{kij}^2 + \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq i,k}}^n b_{klj}^2$$

αντιπροσωπεύουν όλες τις συγκρίσεις “μεγαλύτερο από” και “μικρότερο από” για τις περιπτώσεις i και k , σε σχέση με το σύνολο των δεδομένων της έρευνας.

Ο δείκτης απόστασης d_{ik} μπορεί να υπολογιστεί σε δεδομένα εκφρασμένα εξ ολοκλήρου με κλίμακα τύπου ordinal και παίρνει τιμές από 0 έως 1, όπου η τιμή 0 εκφράζει αποκλειστικά σχέσεις ισότητας μεταξύ των μεταβλητών δύο περιπτώσεων, ενώ η τιμή 1 εκφράζει αποκλειστικά σχέσεις ανισότητας. Για τον δείκτη απόστασης ισχύουν τα: $d_{ik} \geq 0$, $d_{ii} = 0$, $d_{ik} = d_{ki}$. Σημειώνεται ότι για τον υπολογισμό του δείκτη απόστασης είναι απαραίτητη η ύπαρξη τουλάχιστον ενός ζευγαριού ανόμοιων περιπτώσεων στον πίνακα $X(n, m)$ ώστε να αποφευχθεί μηδενισμός του παρονομαστή.

Η μέθοδος σε μια παραλλαγή της, δίνει την δυνατότητα του υπολογισμού του δείκτη απόστασης όταν στις μεταβλητές της έρευνας υπάρχουν διαφορετικά βάρη, πολλαπλασιάζοντας κάθε άθροισμα του τύπου με το συντελεστή βάρους w_j .

Ο τετραγωνικός πίνακας $D(n, n)$ που αποτελείται από τα αποτελέσματα d_{ik} στο εξής ονομάζεται πίνακας αποστάσεων. Προφανώς ο πίνακας αυτός είναι συμμετρικός αφού $d_{ik} = d_{ki}$ άρα $D[i, k] = D[k, i]$.

Ο αλγόριθμος Walesiak , σε κάθε βήμα συγκρίνει τις τιμές των περιπτώσεων i, k για τη μεταβλητή j μεταξύ τους και τις τιμές κάθε περίπτωσης i, k για την ίδια μεταβλητή με όλες τις n περιπτώσεις της έρευνας. Οι συγκρίσεις των ζευγών των στοιχείων $a_{ikj}, a_{ilj}, b_{kij}, b_{klj}$, παίρνουν τιμές 1, 0 ή -1 ανάλογα με το εάν προκύπτει αποτέλεσμα “μεγαλύτερο από”, “ίσο” ή “μικρότερο από” αντίστοιχα. Οι πράξεις των αθροισμάτων των παραπάνω συγκρίσεων για όλες τις μεταβλητές j σε

όλες τις n περιπτώσεις δίνουν τελικά την απόσταση των δύο ordinal διανυσμάτων.

Το πλεονέκτημα αυτής της μεθόδου είναι ότι ο υπολογισμός της απόστασης προκύπτει μέσω συγκρίσεων όχι μόνο των ζευγών i, k που υπολογίζονται, αλλά και μέσω της σχετικής θέσης τους με όλα τα ερωτηματολόγια της έρευνας.

Έτσι η πρώτη απλή μορφή του αλγόριθμου hot deck της παραγράφου 4.1 μετασχηματίζεται σε:

Εισαγωγή πίνακα $X(n,m)$

Υπολόγισε πίνακα Walesiak $D(n,n)$.

Εάν x_{ij} = "κενό"

Βρες το ερωτηματολόγιο k για το οποίο ισχύει $\min D(i,k)$ και $x_{kj} \neq$ "κενό".

$x_{ij} \rightarrow x_{kj}$

Ο αλγόριθμος σε έναν πίνακα δεδομένων όπως του παραδείγματος 4.1 υλοποιείται ως εξής:

Πίνακας 4.2: Παράδειγμα έρευνας 3x5 με ένα κενό

	<i>Απάντηση1</i>	<i>Απάντηση2</i>	<i>Απάντηση3</i>	<i>Απάντηση4</i>	<i>Απάντηση5</i>
<i>Άτομο 1</i>	<i>Πάρα Πολύ</i>	<i>Πάρα Πολύ</i>	-	<i>Μέτρια</i>	<i>Πάρα Πολύ</i>
<i>Άτομο 2</i>	<i>Πολύ</i>	<i>Μέτρια</i>	<i>Πολύ</i>	<i>Πολύ</i>	<i>Πολύ</i>
<i>Άτομο 3</i>	<i>Πολύ Λίγο</i>	<i>Πολύ Λίγο</i>	<i>Πολύ Λίγο</i>	<i>Μέτρια</i>	<i>Πολύ Λίγο</i>

Αρχικά κατασκευάζεται ο πίνακας αποστάσεων $D(3,3)$ σύμφωνα με τον αλγόριθμο Walesiak. Στους υπολογισμούς των αποστάσεων d_{12} και d_{13} δεν λαμβάνονται υπόψη τα δεδομένα της μεταβλητής 3 που περιέχει το ελλιπές στοιχείο όπως θα εξηγηθεί αργότερα.

Πίνακας 4.3: Πίνακας αποστάσεων 3x3

	Άτομο 1	Άτομο 2	Άτομο 3
Άτομο 1	0	0,59	0,8
Άτομο 2	0,59	0	0,56
Άτομο 3	0,8	0,56	0

Στη συνέχεια ο αλγόριθμος θα εντοπίσει το ελλιπές στοιχείο x_{13} και θα αναζητήσει την ελάχιστη απόσταση μεταξύ των d_{12} και d_{13} , στην περίπτωση του παραδείγματος 4.2 η $d_{12} = 0,59$. Εφόσον το αντίστοιχο στοιχείο δεν είναι κενό θα γίνει αντικατάσταση του x_{13} με το x_{23} και θα παραχθεί ο πλήρης πίνακας δεδομένων της έρευνας.

Πίνακας 4.4: Συμπληρωμένος πίνακας έρευνας 3x5

	Απάντηση1	Απάντηση2	Απάντηση3	Απάντηση4	Απάντηση5
Άτομο 1	Πάρα Πολύ	Πάρα Πολύ	Πολύ	Μέτρια	Πάρα Πολύ
Άτομο 2	Πολύ	Μέτρια	Πολύ	Πολύ	Πολύ
Άτομο 3	Πολύ Λίγο	Πολύ Λίγο	Πολύ Λίγο	Μέτρια	Πολύ Λίγο

4.4 Ερώτηση Ολικής Ικανοποίησης

Η υλοποίηση του ανωτέρω αλγορίθμου υπολογίζει την απόσταση των δυο ερωτηματολογίων λαμβάνοντας υπόψη ότι όλες οι μεταβλητές έχουν την ίδια βαρύτητα και μετέχουν το ίδιο στην απόφαση εάν δυο ερωτηματολόγια είναι όμοια. Στις περισσότερες έρευνες ικανοποίησης πελατών όμως, υπάρχει μια μεταβλητή που έχει διαφορετική βαρύτητα και ονομάζεται ερώτηση ολικής ικανοποίησης. Σε αυτήν ο πελάτης

καλείται να απαντήσει, μονολεκτικά στην ουσία, για την συνολική εικόνα που αποκόμισε από τις υπηρεσίες – προϊόντα που του παρασχέθηκαν.

Η ερώτηση ολικής ικανοποίησης βοηθάει τον ερωτώμενο να δώσει την συνολική εικόνα πριν μπει στις επιμέρους ερωτήσεις όπου μπορεί για πλήθος λόγων να τον απομακρύνουν από την πραγματικότητα. Μπορεί π.χ. κάποια ερώτηση να είναι δυσνόητη για αυτόν, να έχει σύγχυση για το αντικείμενο της ερώτησης, να μην του δόθηκε η ευκαιρία να αξιολογήσει το αντικείμενο, να κουράστηκε από το πλήθος των ερωτήσεων κ.λπ., με αποτέλεσμα να βαθμολογήσει διαφορετικά από εάν υπήρχαν διαφορετικές συνθήκες έρευνας.

Η εν λόγω ερώτηση δεν φέρει την ίδια βαρύτητα με τις υπόλοιπες ερωτήσεις εφόσον δηλώνει συνήθως και την πρόθεση επαναγοράς του προϊόντος από τον πελάτη. Όταν ο πελάτης συνολικά είναι πολύ ικανοποιημένος από την κάλυψη των προσδοκιών του, αναμένεται και οι επιμέρους τομείς να έχουν καλή βαθμολογία, δηλαδή τουλάχιστον οι περισσότερες ερωτήσεις να κυμαίνονται από “καλό” έως “πολύ καλό” αποτέλεσμα, κανόνας που δεν επιβεβαιώνεται πάντα όμως. Για το λόγο αυτό, οι σύγχρονες έρευνες ικανοποίησης στην πλειοψηφία τους εμπεριέχουν στο ερωτηματολόγιο την ερώτηση ολικής ικανοποίησης.

Η απόδοση βαρύτητας στον υπολογισμό αποστάσεων περιπτώσεων που για την ερώτηση ολικής ικανοποίησης μπορεί να γίνει με τον τύπο:

$$\text{Απόσταση ερωτηματολόγιων} = \left(\frac{w * \text{Απόσταση ολικής ικανοποίησης} + (1 - w) * \text{Απόσταση υπολοίπων ερωτήσεων}}{1} \right)$$

όπου w ένας συντελεστής βαρύτητας που επιλέγει ο χρήστης, με τιμές από 0 έως 1. Η τιμή $w = 0$ δηλώνει ότι δεν υπάρχει ή δεν ενδιαφέρει καθόλου η ερώτηση ολικής ικανοποίησης, καθώς θα ισχύει:

$$\text{Απόσταση ερωτηματολόγιων} = \text{Απόσταση υπολοίπων ερωτήσεων}$$

ενώ η τιμή $w = 1$ δηλώνει ότι η απόσταση ερωτηματολογίων υπολογίζεται μόνο μέσω της ερώτησης ολικής ικανοποίησης καθώς θα ισχύει:

$$\text{Απόσταση ερωτηματολόγιων} = \text{Απόσταση ολικής ικανοποίησης}$$

Ορίζονται λοιπόν, $D1(n,n)$ ο πίνακας αποστάσεων που παράγεται με χρήση της μεθόδου Walesiak μόνο για την ερώτηση ολικής ικανοποίησης και $D2(n,n)$ ο πίνακας αποστάσεων που παράγεται για τις υπόλοιπες ερωτήσεις. Ο πίνακας που εκπροσωπεί την συνολική απόσταση των ερωτηματολογίων ορίζεται ως $Dtotal(n,n)$ και ισχύει

$$Dtotal(n,n) = w * D1(n,n) + (1 - w) * D2(n,n)$$

Ο αλγόριθμος της παραγράφου 4.3 πλέον γίνεται:

Εισαγωγή πίνακα $X(n,m)$

Ορισμός τιμής w από το χρήστη

Υπολόγισε πίνακα Walesiak ολικής ικανοποίησης $D_1(n,n)$

Υπολόγισε πίνακα Walesiak υπόλοιπων μεταβλητών $D_2(n,n)$

$$D_{total}(n,n) = w * D_1(n,n) + (1-w) * D_2(n,n)$$

Εάν x_{ij} = "κενό"

Βρες το ερωτηματολόγιο k για το οποίο ισχύει $\min D_{total}(i,k)$

και $x_{kj} \neq$ "κενό".

$$x_{ij} \rightarrow x_{kj}$$

Στην παρούσα εργασία η έννοια της ολικής ικανοποίησης εισήχθη με την απόδοση βαρύτητας 50% στην εν λόγω ερώτηση. Έτσι, ο υπολογισμός της συνολικής απόστασης δύο περιπτώσεων προκύπτει τον τύπο:

$$Dtotal(n,n) = 0,5 * D1(n,n) + 0,5 * D2(n,n)$$

Ως συνολική απόσταση των περιπτώσεων δηλαδή, λογίζεται ο μέσος όρος της απόστασης που προκύπτει από την ερώτηση ολικής ικανοποίησης και της απόστασης που προκύπτει από τις υπόλοιπες ερωτήσεις.

4.5 Κωδικοποίηση Δεδομένων

Η υλοποίηση του παραπάνω αλγόριθμου σε πρόγραμμα εκτελέσιμο από Η/Υ, απαιτεί την εκτέλεση πράξεων ανισότητας των ordinal δεδομένων για την εύρεση των αποστάσεων Walesiak. Για να είναι αυτό δυνατό, προφανώς ο χρήστης πρέπει να κωδικοποιήσει τα ordinal δεδομένα σε αριθμητικά αφού ο Υ/Η έχει τη δυνατότητα εκτέλεσης μόνο αριθμητικών πράξεων. Μια απλή κωδικοποίηση είναι η έκφραση των αποτελεσμάτων σε αριθμητική κλίμακα ακεραίων από το 1 έως m όπου m το πλήθος των μεταβλητών. Έτσι μια ordinal κλίμακα πέντε επιπέδων:

Πολύ λίγο – Λίγο – Μέτρια – Πολύ – Πάρα πολύ

κωδικοποιείται σε:

$1 - 2 - 3 - 4 - 5.$

Τονίζεται ότι οι αριθμοί αυτοί δεν αντιπροσωπεύουν αριθμητικές τιμές εφόσον αφορούν δεδομένα τύπου ordinal.

Οι κενές απαντήσεις μπορούν να κωδικοποιηθούν με οποιοδήποτε διαφορετικό αριθμό, στην παρούσα εργασία ορίστηκε η έκφραση τους με τον αριθμό 999. Έτσι, ο πίνακας των παραδειγμάτων 4.1 και 4.2 κωδικοποιείται ως:

Πίνακας 4.5: Κωδικοποίηση δεδομένων έρευνας 3x5 με ένα κενό

	Απάντηση1	Απάντηση2	Απάντηση3	Απάντηση4	Απάντηση5
Άτομο1	5	5	999	3	5
Άτομο2	4	3	4	4	4
Άτομο3	1	1	1	3	1

Εφαρμόζοντας την εν λόγω κωδικοποίηση σε πλήρεις πίνακες δεδομένων $X(n, m)$, γίνεται εφικτή η χρήση του αλγόριθμου Walesiak σε ordinal δεδομένα και η παραγωγή του πίνακα αποστάσεων $D_{total}(n, n)$. Όταν όμως εισάγεται η έννοια των ελλιπών δεδομένων, ουσιαστικά αλλοιώνεται η αρχική κλίμακα δεδομένων αφού εισάγεται ο νέος αριθμός 999 για να καλύψει την ανάγκη έκφρασης των κενών. Έτσι, κατά τη χρήση του αλγόριθμου Walesiak, τα κενά θα λειτουργούν ως “φανταστικές” απαντήσεις με πολύ υψηλή βαθμολογία (999), χωρίς ουσιαστικά να ανταποκρίνονται σε κάποια τιμή της αρχικής ordinal κλίμακας.

Άρα, αν κατά τον υπολογισμό του πίνακα $D(n, n)$ ληφθούν υπόψη τα 999, θα επηρεαστούν οι υπολογιζόμενες αποστάσεις αφού για παράδειγμα ένα ζεύγος απαντήσεων $x_{ij} = 999$, $x_{kj} = 1$ θα ερμηνευτεί ως “πιο μακρινό” από ένα ζεύγος $x_{ij} = 999$, $x_{pj} = 5$ το οποίο θα ερμηνευτεί ως “πιο κοντινό”.

Με λίγα λόγια η μέθοδος hot deck, βάσει αυτής της κωδικοποίησης, θα μεροληπτεί συνεχώς, επιλέγοντας για imputation τιμές υψηλότερες από τις αναμενόμενες. Γενικότερα, όποιος αριθμός εάν επιλεχθεί ως κωδικοποίηση των ελλιπών στοιχείων, εάν τα κενά συμπεριλαμβάνονται στους υπολογισμούς των αποστάσεων, το αποτέλεσμα θα είναι το imputation των νέων τιμών να “τραβιέται” προς τον αριθμό αυτό.

Για αποφυγή του προβλήματος αυτού, στον αλγόριθμο που αναπτύχθηκε για την εργασία ο υπολογισμός των αποστάσεων d_{ik} έγινε χωρίς να ληφθούν υπόψη οι στήλες (μεταβλητές) στις οποίες υπάρχει ελλιπής τιμή. Έτσι, κατά τη χρήση του αλγόριθμου Walesiak αρχικά

ελέγχονται τα δύο ερωτηματολόγια για ύπαρξη κενού σε κάθε μεταβλητή ($x_{ij} = 999$ or $x_{kj} = 999$ για $j = 1, 2, \dots, m$) και σε θετικό αποτέλεσμα η μεταβλητή αυτή αφαιρείται από τον υπολογισμό. Όσο περισσότερες μεταβλητές j εμφανίζουν θετικό αποτέλεσμα στον έλεγχο ($x_{ij} = 999$ or $x_{kj} = 999$), τόσο λιγότερες στήλες απομένουν για τον υπολογισμό της απόστασης.

Παρακάτω παρατίθεται παράδειγμα για τον τρόπο υπολογισμού των αποστάσεων σύμφωνα με τη μεθοδολογία που έχει αναλυθεί έως τώρα:

Πίνακας 4.6: Παράδειγμα έρευνας 6x5 με τέσσερα κενά

	<i>Απάντηση1</i>	<i>Απάντηση2</i>	<i>Απάντηση3</i>	<i>Απάντηση4</i>	<i>Απάντηση5</i>
<i>Άτομο1</i>	5	5	999	5	5
<i>Άτομο2</i>	5	999	4	999	5
<i>Άτομο3</i>	999	2	1	4	3
<i>Άτομο4</i>	1	3	2	3	5
<i>Άτομο5</i>	3	4	2	3	4
<i>Άτομο6</i>	3	4	3	3	4

Σύμφωνα με τον αλγόριθμο υπολογισμού αποστάσεων που έχει αναπτυχθεί, αρχικά υπολογίζονται οι αποστάσεις μεταξύ όλων των πιθανών ζευγαριών ερωτηματολογίων μέσω του συμμετρικού, τετραγωνικού πίνακα 6x6 που ακολουθεί:

Πίνακας 4.7: Πίνακας αποστάσεων 6x6

	Άτομο1	Άτομο2	Άτομο3	Άτομο4	Άτομο5	Άτομο6
Άτομο1	0	0	0,8	0,59	0,41	0,41
Άτομο2	0	0	0,92	0,51	0,44	0,35
Άτομο3	0,8	0,92	0	0,7	0,68	0,72
Άτομο4	0,59	0,51	0,7	0	0,24	0,27
Άτομο5	0,41	0,44	0,68	0,24	0	0,04
Άτομο6	0,41	0,35	0,72	0,27	0,04	0

Από ανάλυση των αποτελεσμάτων παρατηρούνται τα κάτωθι:

α. Η απόσταση των ερωτηματολογίων 1 και 2 υπολογίστηκε $d_{12} = 0$ το οποίο είναι λογικό, εφόσον όπως έχει οριστεί προηγουμένως, οι μεταβλητές 2, 3 και 4 εξαιρούνται από τους υπολογισμούς της απόστασης, ενώ στις μεταβλητές 1 και 5 έχουν δοθεί ταυτόσημες απαντήσεις ($x_{11} = x_{21} = 5$ και $x_{15} = x_{25} = 5$).

β. Από τα υπόλοιπα ερωτηματολόγια τα 5 και 6 έχουν τη μικρότερη απόσταση $d_{56} = 0,04$ δηλαδή “μοιάζουν” αρκετά, το οποίο διαπιστώνεται αμέσως αφού διαφοροποιούνται μόνο στην μεταβλητή 3 για ένα βαθμό της κλίμακας απαντήσεων.

Έτσι, ο αλγόριθμος hot deck όπως έχει καθοριστεί μέχρι τώρα θα αντικαταστήσει τα 999 με βάση την ελάχιστη απόσταση Walesiak και θα παραχθεί ο παρακάτω συμπληρωμένος πίνακας:

Πίνακας 4.8: Πίνακας αποτελεσμάτων έρευνας 6x5

	Απάντηση1	Απάντηση2	Απάντηση3	Απάντηση4	Απάντηση5
Άτομο1	5	5	4	5	5
Άτομο2	5	5	4	5	5
Άτομο3	3	2	1	4	3
Άτομο4	1	3	2	3	5
Άτομο5	3	4	2	3	4
Άτομο6	3	4	3	3	4

Από τον πίνακα αποτελεσμάτων προκύπτουν τα παρακάτω συμπεράσματα:

- Το $x_{13} = 999$ του ατόμου 1 αντικαταστάθηκε από την τιμή $x_{23} = 4$ του ατόμου 2 λόγω μικρότερης απόστασης $d_{12} = 0$.
- Τα $x_{22} = 999$, $x_{24} = 999$ του ατόμου 2 αντικαταστάθηκαν από τις τιμές $x_{12} = 5$ και $x_{14} = 5$ αντίστοιχα του ατόμου 1 λόγω μικρότερης απόστασης $d_{12} = 0$.
- Το $x_{31} = 999$ του ατόμου 3 αντικαταστάθηκε από την τιμή $x_{51} = 3$ του ατόμου 5 λόγω μικρότερης απόστασης $d_{35} = 0,68$.

Το imputation του στοιχείου x_{31} από το στοιχείο x_{51} κρίνεται λογικό και σαφές αφού στον υπολογισμό της απόστασης Walesiak λήφθηκαν υπόψη όλες οι μεταβλητές, εκτός της μεταβλητής 1 όπου υπήρχε το ελλιπές στοιχείο. Για το imputation των περιπτώσεων 1 και 2 μεταξύ τους όμως, προκύπτει το ερώτημα κατά πόσον είναι αξιόπιστη η μέτρηση της απόστασης d_{12} με χρήση μόνο των μεταβλητών 1 και 5 (αφού στις 2, 3, 4 υπάρχει στοιχείο με τιμή 999). Μήπως δηλαδή απαιτείται πριν το imputation η χρήση μιας μεθόδου διαγραφής η οποία θα περιορίζει το εύρος των πιθανών περιπτώσεων από τα οποία θα είναι δυνατή η λήψη τιμής?

4.6 Εισαγωγή Συντελεστή Καταλογισμού

Ανάλογα με τα δεδομένα της έρευνας ο ερευνητής, ιδίως στην υλοποίηση μεθόδων hot deck όπου δεν υπάρχει συγκεκριμένη και απαρέγκλιτη θεωρία, πρέπει να προσαρμόζει τη μεθοδολογία που ακολουθεί, τόσο για την εύρεση της ομοιότητας δυο περιπτώσεων όσο και για την ίδια την αντικατάσταση. Για αυτό το λόγο και μετά την τροποποίηση του αλγόριθμου Walesiak για τον υπολογισμό των αποστάσεων, κρίθηκε αναγκαίος ο έλεγχος των πιθανών δοτών ώστε να αποφευχθούν τα προβλήματα αναξιόπιστων αντικαταστάσεων.

Ένα παράδειγμα που φανερώνει το πρόβλημα αξιόπιστης επιλογής δότη είναι το παρακάτω:

Πίνακας 4.9: Παράδειγμα έρευνας 3x5 με τρία κενά

	<i>Απάντηση1</i>	<i>Απάντηση2</i>	<i>Απάντηση3</i>	<i>Απάντηση4</i>	<i>Απάντηση5</i>
<i>Άτομο1</i>	5	5	999	5	5
<i>Άτομο2</i>	4	4	4	4	4
<i>Άτομο3</i>	5	999	2	999	5

Σε αυτό το παράδειγμα ο πίνακας αποστάσεων $D(3,3)$ θα είναι:

Πίνακας 4.10: Πίνακας αποστάσεων 3x3

	<i>Άτομο1</i>	<i>Άτομο2</i>	<i>Άτομο3</i>
<i>Άτομο1</i>	0	0,72	0
<i>Άτομο2</i>	0,72	0	0,74
<i>Άτομο3</i>	0	0,74	0

Δηλαδή η απόσταση d_{13} θα υπολογιστεί 0 αφού οι μεταβλητές 2,3,4 δε θα ληφθούν υπόψη στους υπολογισμούς και στις μεταβλητές 1 και 5 υπάρχουν κοινά στοιχεία. Άρα κατά τη συμπλήρωση των 999 θα προτιμηθεί η αντικατάσταση των στοιχείων μεταξύ των περιπτώσεων 1

και 3 οι οποίες θα θεωρούνται ταυτόσημες από τον αλγόριθμο! Το αποτέλεσμα του αλγόριθμου hot deck σε αυτή την περίπτωση θα είναι ο παρακάτω πίνακας:

Πίνακας 4.11: Πίνακας αποτελεσμάτων έρευνας 3x5 με τρία κενά

	Απάντηση1	Απάντηση2	Απάντηση3	Απάντηση4	Απάντηση5
Άτομο1	5	5	2	5	5
Άτομο2	4	4	4	4	4
Άτομο3	5	5	2	5	5

Γίνεται εύκολα αντιληπτό όμως ότι ποιοτικά αυτή η απόδοση τιμής είναι αμφίβολη, αφού το δείγμα δεδομένων που λαμβάνεται για τον υπολογισμό της απόστασης d_{13} είναι πολύ μικρό (μόλις δύο μεταβλητές) και δεν είναι ασφαλές να εξαχθούν συμπεράσματα για την ομοιότητα των ερωτηματολογίων. Όσο ο αλγόριθμος τεχνητά αφαιρεί στήλες από τον υπολογισμό της απόστασης κάθε ζεύγους περιπτώσεων, τόσο μειώνεται η αξιοπιστία του αποτελέσματος.

Κρίνεται λοιπόν αυτονόητο ότι ο ερευνητής πρέπει να έχει έλεγχο στην επιλογή του δότη για imputation, ώστε να περιορίζονται τα φαινόμενα ανεπιθύμητης αντικατάστασης. Για το λόγο αυτό απαιτείται η εφαρμογή ενός δείκτη όπου θα δείχνει το πόσο αξιόπιστος είναι ο υπολογισμός της απόστασης, λαμβάνοντας υπόψη ότι όσο λιγότερα 999 υπάρχουν στα στοιχεία μεταξύ δύο περιπτώσεων, τόσο πιο αξιόπιστος είναι και ο υπολογισμός της απόστασης. Το βέλτιστο στην εκτέλεση της μεθόδου hot deck για τη συμπλήρωση ενός κενού, είναι ο δότης να μην έχει κανένα κενό στις υπόλοιπες μεταβλητές ώστε να μετριέται ορθά η απόσταση.

Χρησιμοποιήθηκε λοιπόν ο δείκτης “**Συντελεστής Καταλογισμού**” για να γίνει δυνατός ο παραπάνω έλεγχος. Ο συντελεστής Καταλογισμού a_{ik} ενός ζεύγους περιπτώσεων i, k φανερώνει ουσιαστικά το πλήθος των μεταβλητών j που είναι δυνατόν να

χρησιμοποιηθούν στον υπολογισμό της απόστασης d_{ik} . Ο τύπος που χρησιμοποιήθηκε για τον υπολογισμό του εν λόγω δείκτη είναι:

$$a_{ik} = \frac{\text{πλήθος μεταβλητών } j \text{ χωρίς } 999 \text{ σε } i \text{ ή } k}{\text{πλήθος μεταβλητών } j}$$

Για κάθε ζεύγος περιπτώσεων i, k εκτελείται έλεγχος σε όλες τις μεταβλητές j για ύπαρξη $x_{ij} = 999$ ή $x_{kj} = 999$ και δημιουργείται τετραγωνικός και συμμετρικός πίνακας $A(n, n)$ με τα αποτελέσματα του παραπάνω τύπου. Οι τιμές των στοιχείων του πίνακα κυμαίνονται από 0 έως 1, όπου 1 σημαίνει ότι δεν υπάρχει καμία κενή απάντηση μεταξύ των δύο ερωτηματολογίων, ενώ 0 ότι δεν υπάρχει καμία μεταβλητή χωρίς κενό μεταξύ τους.

Με χρήση του παραπάνω δείκτη, ο πίνακας των αποστάσεων “συνοδεύεται” ουσιαστικά από τον πίνακα των συντελεστών Καταλογισμού και δίνεται η δυνατότητα στον ερευνητή να περιορίσει στο εύρος που θέλει την επιλογή περιπτώσεων-δοτών για απόδοση τιμής. Αυτό επιτυγχάνεται με την επιλογή κατάλληλου δότη, όχι μόνο μέσω της ελάχιστης απόστασης δύο περιπτώσεων, αλλά μέσω της ελάχιστης απόστασης αποκλειστικά μεταξύ των περιπτώσεων που έχουν τιμή συντελεστή Καταλογισμού μεγαλύτερη από κάποια προκαθορισμένη τιμή.

Για την υλοποίηση των ανωτέρω στην παρούσα εργασία, επιλέχθηκε ο χρήστης κατά την εισαγωγή των δεδομένων να ορίζει την ελάχιστη επιτρεπτή τιμή του συντελεστή Καταλογισμού, με βάση την οποία θα γίνεται απόδοση τιμής στα κενά. Έτσι λοιπόν ο χρήστης εισάγει την ελάχιστη επιτρεπτή τιμή c του συντελεστή Καταλογισμού και ο αλγόριθμος πριν εκτελέσει imputation ελέγχει τις υποψήφιες επιλογές για:

$$A(i, k) \geq c$$

Επειδή ο έλεγχος αυτός εκτελείται στο τελικό στάδιο του αλγόριθμου για περιπτώσεις όπου προφανώς έχουν τουλάχιστον ένα

999 σε κάποια μεταβλητή, ο συντελεστής a_{ik} παίρνει τιμές από 0 έως $\frac{(m-1)}{m}$, όπου m είναι το σύνολο του πλήθους των μεταβλητών j . Για λόγους διευκόλυνσης του χρήστη, ο έλεγχος ελάχιστης τιμής του συντελεστή καταλογισμού μετατράπηκε σε

$$A(i, k) + \frac{1}{m} \geq c$$

ώστε το εύρος τιμών της ελάχιστης επιτρεπτής τιμής c να είναι από $\frac{1}{m}$ έως 1 και όχι 0 έως $\frac{(m-1)}{m}$. Η τιμή $c = 1$ φανερώνει την απαίτηση του χρήστη για εκτέλεση imputation από δότες μόνο πλήρως συμπληρωμένα ερωτηματολόγια, όπου θα ισχύει η ισότητα

$$\frac{(m-1)}{m} + \frac{1}{m} = c = 1$$

Σε αυτή την περίπτωση η τιμή συντελεστή καταλογισμού a_{ik} είναι $\frac{(m-1)}{m}$ δηλαδή μεταξύ των δύο περιπτώσεων υπάρχει 999 μόνο στη μεταβλητή για την οποία απαιτείται απόδοση τιμής και πουθενά αλλού.

Ο αλγόριθμος hot deck λοιπόν μετασχηματίζεται σε:

Εισαγωγή πίνακα $X(n, m)$

Ορισμός τιμής w από το χρήστη

Ορισμός τιμής c από το χρήστη

Υπολόγισε πίνακα Συντελεστή Καταλογισμού $A(n, n)$.

Υπολόγισε πίνακα Walesiak ολικής ικανοποίησης $D_1(n, n)$.

Υπολόγισε πίνακα Walesiak υπόλοιπων μεταβλητών $D_2(n, n)$.

*$D_{total}(n, n) = w * D_1(n, n) + (1-w) * D_2(n, n)$*

Εάν $x_{ij} = 999$

Βρες το ερωτηματολόγιο k για το οποίο $\min D_{total}(i, k)$,

για το οποίο ισχύει: $x_{kj} \neq 999$ και $A(i, k) + 1/m \geq c$

$x_{ij} \rightarrow x_{kj}$

Υιοθετώντας αυτό τον έλεγχο κατασκευάζεται ο παρακάτω πίνακας συντελεστών Καταλογισμού για το παράδειγμα του πίνακα 4.9:

Πίνακας 4.12: Πίνακας συντελεστών Καταλογισμού

	Άτομο1	Άτομο2	Άτομο3
Άτομο1	0,8	0,8	0,4
Άτομο2	0,8	1	0,6
Άτομο3	0,4	0,6	0,6

Παρατηρείται ότι η μέγιστη τιμή σε κάθε γραμμή ή στήλη του πίνακα δεν είναι απαραίτητα το 1, αφού κάθε μεταβλητή με 999 σε μια περίπτωση αφαιρεί από το αποτέλεσμα a_{ik} το ποσό $1/m$ δηλαδή $1/5$ στην περίπτωση του παραδείγματος. Το μόνο στοιχείο του πίνακα με τιμή 1 φανερώνει την μη ύπαρξη κενού στο ζεύγος του ατόμου 2 με τον εαυτό του, αφού και στα δύο άλλα άτομα υπάρχουν 999.

Έτσι πλέον, με επιλογή ενδεικτικής τιμής ελάχιστου συντελεστή Καταλογισμού $c = 0,8$ από το χρήστη, η ανισότητα

$$A(i, k) + \frac{1}{m} \geq c$$

ισχύει μόνο για τιμές

$$A(i, k) \geq 0,8 - \frac{1}{5} = 0,6$$

Η τιμή 0,8 επιτρέπει δηλαδή, το δανεισμό δεδομένων για συμπλήρωση των κενών μόνο από τα ζεύγη περιπτώσεων που έχουν συντελεστή Καταλογισμού 0,6 ή μεγαλύτερο. Άρα κατά την επιλογή κατάλληλης περίπτωσης για απόδοση τιμής στα άτομα 1 και 3 που παρουσιάζουν κενά, θα απορριφθεί το imputation μεταξύ τους (με τον τρόπο που έγινε στην παράγραφο 4.5) λόγω συντελεστή καταλογισμού 0,4 και τα κενά θα αντικατασταθούν από τις τιμές του ατόμου 2. Ο νέος

και πλέον πιο αξιόπιστος συμπληρωμένος πίνακας αποτελεσμάτων θα έχει την παρακάτω μορφή:

Πίνακας 4.13: Νέος πίνακας αποτελεσμάτων έρευνας 3x5 με τρία κενά

	Απάντηση1	Απάντηση2	Απάντηση3	Απάντηση4	Απάντηση5
Άτομο1	5	5	4	5	5
Άτομο2	4	4	4	4	4
Άτομο3	5	4	2	4	5

Ο χρήστης έχει δηλαδή πλέον έλεγχο του εύρους του περιορισμού των υποψήφιων δοτών, μέσω επιλογής της κατάλληλης τιμής c , ανάλογα με το δείγμα της έρευνας με βάση τον τύπο:

$$A(i, k) + \frac{1}{m} \geq c$$

Εάν δηλαδή γνωρίζει εκ των προτέρων ότι δεν θέλει να γίνει imputation μεταξύ περιπτώσεων που έχουν t μεταβλητές με κενό, μπορεί να λύσει αντίστροφα την εξίσωση και να υπολογίσει το ελάχιστο c ως:

$$c = \frac{(m - t)}{m} + \frac{1}{m} = \frac{m - t + 1}{m}$$

Και εάν θέλει να εκτελείται imputation μόνο από δότες πλήρων συμπληρωμένων περιπτώσεων, θα επιλέξει $t = 1$ (εφόσον πάντα υπάρχει ένα 999 για το οποίο θα γίνει imputation) και άρα:

$$c = \frac{(m - 1 + 1)}{m} = 1$$

Μια εύκολη λύση θα ήταν ο χρήστης να επέλεγε ως βέλτιστη τιμή $c = 1$ συνεχώς. Αυτό όμως εκτός του ότι περιορίζει το εύρος των πιθανών δοτών, ιδιαίτερα σε έρευνες με πίνακες στοιχείων όπου

υπάρχουν πολλά 999, καθιστά αδύνατη την αντικατάσταση κενού σε περιπτώσεις με κενά σε δύο ή περισσότερες μεταβλητές. Δηλαδή εάν $c = 1$ και υπάρχει έστω και μια περίπτωση με δύο ελλιπή στοιχεία, ο έλεγχος

$$\frac{(m-2)}{m} + \frac{1}{m} = \frac{(m-1)}{m} \geq 1$$

θα είναι συνεχώς ψευδής, αφού $m > 0$, και δεν θα βρεθεί κατάλληλη περίπτωση για απόδοση τιμής.

Με βάση το τελευταίο, προκύπτει άλλη μια χρήση του συντελεστή Καταλογισμού ως δικλείδα ασφαλείας για περιπτώσεις με πολλά κενά, τις οποίες δεν θέλει ο χρήστης να συμπεριλάβει στην έρευνα αφού η χρήση τους στη στατιστική επεξεργασία μπορεί να είναι επιβλαβής.

Συνήθως, πριν την επεξεργασία οποιασδήποτε βάσης δεδομένων με ελλιπή στοιχεία προηγείται μια μέθοδος διαγραφής τύπου Listwise, όπου αφαιρούνται τα ανεπιθύμητα ερωτηματολόγια με πολλά κενά, βάσει κριτηρίων που θέτει ο ερευνητής. Το ποσό των επιτρεπτών κενών δύναται να ελεγχθεί με τον συντελεστή Καταλογισμού κατά το imputation όπου για να ολοκληρωθεί, πρέπει να τηρείται η ανισότητα. Π.χ. εάν ο χρήστης ορίσει ελάχιστη τιμή συντελεστή Καταλογισμού $c = 0.7$ σε έρευνα 10 μεταβλητών, η ανισότητα θα γίνει:

$$\frac{(10-t)}{10} + \frac{1}{10} \geq 0,7 \rightarrow t \leq 4$$

δηλαδή δεν θα επιτραπεί απόδοση τιμής σε οποιοδήποτε ερωτηματολόγιο έχει περισσότερα από 4 κενά και σε αυτή την περίπτωση θα ειδοποιηθεί ο χρήστης για σφάλμα στα δεδομένα.

4.7 Λειτουργία Αλγορίθμου Hot Deck

Συνοψίζοντας, ο αλγόριθμος Hot Deck που αναπτύχθηκε για χρήση σε ordinal δεδομένα λειτουργεί σύμφωνα με την παρακάτω δομή:

- Αρχικά κωδικοποιούνται από το χρήστη τα ordinal δεδομένα της έρευνας με βάση μια αριθμητική κλίμακα της επιλογής του όπου ως κενό χρησιμοποιείται ο αριθμός 999.
- Εισάγονται από το χρήστη τα στοιχεία της έρευνας δηλαδή ο πίνακας $X(n, m)$, ο συντελεστής βαρύτητας ολικής ικανοποίησης w και η ελάχιστη τιμή του συντελεστή Καταλογισμού c .
- Στη συνέχεια συμπληρώνονται από τον αλγόριθμο οι πίνακες συντελεστή Καταλογισμού $A(n, n)$ και αποστάσεων Walesiak $D_{total}(n, n)$ για όλα τα ζεύγη των ερωτηματολογίων.
- Ελέγχεται ο πίνακας $X(n, m)$ για τιμή 999 (κενό) και για κάθε κενό που ανακαλύπτεται, γίνεται αναζήτηση του ερωτηματολογίου με την ελάχιστη απόσταση που καλύπτει ταυτόχρονα το κριτήριο του συντελεστή Καταλογισμού και γίνεται αντικατάσταση με την τιμή της μεταβλητής αυτής. Προφανώς έχουν απορριφθεί από την αναζήτηση όλα τα ερωτηματολόγια που έχουν κενό στην ίδια μεταβλητή.

Εάν ένα ερωτηματολόγιο έχει πολλά κενά και η ελάχιστη τιμή του συντελεστή Καταλογισμού που έχει εισάγει ο χρήστης είναι υψηλή, ειδοποιείται είτε για μείωση της ελάχιστης τιμής είτε για έλεγχο και επεξεργασία των δεδομένων.

Σε αυτό το σημείο κρίνεται σκόπιμο να σχολιαστεί η λειτουργία του αλγόριθμου καθώς συμπληρώνονται τα κενά και η διαδικασία συνεχίζεται προς το τέλος. Σε έναν πίνακα δεδομένων $X(n, m)$ όπου υπάρχουν κενά σε διάσπαρτες περιπτώσεις, ο αλγόριθμος αντικαθιστά το πρώτο κενό που ανακαλύπτεται με μια τιμή από άλλη περίπτωση, με βάση την μεθοδολογία που αναλύθηκε. Στα επόμενα κενά που θα

ανακαλυφθούν, το πρώτο κενό πλέον δεν θα υπάρχει αφού θα έχει αντικατασταθεί από μια πραγματική τιμή της έρευνας και θα συμπεριφέρεται ως γνήσια τιμή που προϋπήρχε. Έτσι υπάρχει η πιθανότητα να χρησιμοποιηθεί και αυτή η νέα τιμή ως δότης σε αντικατάσταση επόμενου κενού. Βέβαια, σύμφωνα με τη δομή του αλγόριθμου οι αποστάσεις Walesiak και οι τιμές του συντελεστή Καταλογισμού των ερωτηματολογίων, υπολογίζονται μόνο στην αρχή και ο έλεγχος για imputation θα γίνει με βάση αυτές τις τιμές που υπολογίστηκαν με τα αρχικά κενά.

Άρα προκύπτει το ερώτημα του εάν είναι θεμιτό να γίνεται χρήση αυτών των αρχικών αντικαταστάσεων στη συνέχεια και εάν μπορεί να παρακαμφθεί το ζήτημα αυτό. Τρεις εκδοχές είναι πιθανές: είτε κάθε φορά που γίνεται αντικατάσταση κενού να υπολογίζονται εκ νέου πίνακες αποστάσεων Walesiak και συντελεστή Καταλογισμού ώστε να συμπεριλαμβάνεται η αλλαγή στους ελέγχους, είτε με κάποιο τρόπο να “σημανθούν” αυτές οι τιμές ώστε να μην χρησιμοποιηθούν στη συνέχεια, είτε να χρησιμοποιούνται ως γνήσιες τιμές του αρχικού πίνακα χωρίς υπολογισμό των δύο πινάκων.

Στην πρώτη εκδοχή, σε κάθε επιτυχημένη εκτέλεση imputation, απαιτείται ο υπολογισμός των δύο πινάκων (Walesiak και συντελεστή Καταλογισμού) από την αρχή, θεωρώντας ότι οι νέες τιμές προϋπήρχαν στον αρχικό πίνακα δεδομένων. Αυτή η διαδικασία απαιτεί μεγάλη υπολογιστική ισχύ και ειδικά σε μεγάλες έρευνες είναι πολύ χρονοβόρα χωρίς να είναι δεδομένο ότι έχει θετικό αποτέλεσμα, αφού με αυτόν τον τρόπο ισχυροποιείται η ύπαρξη των νέων τιμών στον αρχικό πίνακα οι οποίες όμως πάντοτε παραμένουν υποθετικές.

Η δεύτερη εκδοχή, ουσιαστικά αναγκάζει τον αλγόριθμο να παραβλέπει τις νέες τιμές που έχουν τοποθετηθεί στον πίνακα και να συμπεριφέρεται σαν να μην έχει γίνει αντικατάσταση των κενών. Άρα κατά την αναζήτηση κάποιας πιθανής περίπτωσης για αντικατάσταση ενός κενού, οι περιπτώσεις που έχουν τιμή προερχόμενη από imputation σε αυτή τη μεταβλητή θα εξαιρεθούν.

Τέλος η τρίτη εκδοχή, αντιλαμβάνεται τις νεοτοποθετημένες τιμές ως τιμές του αρχικού πίνακα δεδομένων και δίνεται η πιθανότητα να χρησιμοποιηθούν ως δότες για imputation, χωρίς όμως να υπολογίζονται οι δύο πίνακες από την αρχή.

Στην παρούσα εργασία επιλέχθηκε η τρίτη εκδοχή λειτουργίας του αλγόριθμου επειδή σε μεγάλες έρευνες με λογικό αριθμό κενών απαντήσεων, αφενός η πιθανότητα δανεισμού των νέων τιμών είναι πολύ μικρή, εφόσον δεν μεταβάλλονται οι τιμές του πίνακα συντελεστή Καταλογισμού (παραμένουν δηλαδή μικρές για αυτές τις περιπτώσεις) και αφετέρου σε MCAR έρευνες οι αποδιδόμενες τιμές θα προσεγγίζουν έτσι και αλλιώς καλά τα στατιστικά δεδομένα της έρευνας και δεν θα επηρεάσουν σημαντικά το αποτέλεσμα. Παρόλα αυτά η χρήση των άλλων δύο εκδοχών του αλγόριθμου και η σύγκριση των αποτελεσμάτων μεταξύ τους, δύναται να αναλυθούν στο πλαίσιο άλλης εργασίας στο μέλλον.

4.8 Προγραμματισμός σε Java

Για την ολοκλήρωση λειτουργίας του αλγόριθμου αποφασίστηκε ο προγραμματισμός του σε γλώσσα προγραμματισμού Java v1.7.0_45 σε περιβάλλον NEATBEANS. Επιλέχθηκε η γλώσσα Java ως μια σύγχρονη γλώσσα η οποία προσφέρει ευκολία ανάπτυξης γραφικού περιβάλλοντος και πλήθος επιλογών και βοηθημάτων για τον προγραμματιστή.

Σκοπός του προγράμματος είναι η ορθή λειτουργία του αλγορίθμου και η συμπλήρωση των κενών της έρευνας σύμφωνα με την ανάλυση που έχει προηγηθεί. Κύριος άξονας του περιβάλλοντος που αναπτύχθηκε, ήταν η απλότητα και η ευκολία χρήσης από οποιονδήποτε χρήστη έχει τις βασικές γνώσεις της μεθόδου hot deck. Το γραφικό περιβάλλον κάνει την εύκολη αναγνωσιμότητα, την απλότητα και την ευκολία χρήσης, ώστε να μην απαιτείται προσπάθεια κατανόησης λειτουργίας του από ένα μέσο χρήστη. Αρκεί ο χρήστης να γνωρίζει τη μορφή εισαγωγής των δεδομένων, τα μεγέθη της έρευνας

και την λειτουργία του συντελεστή Καταλογισμού από τον αλγόριθμο ώστε να βγάλει τα επιθυμητά αποτελέσματα από τη μέθοδο hot deck σε ordinal δεδομένα.

Χρησιμοποιήθηκαν κενά πεδία για την εισαγωγή των απαραίτητων δεδομένων και κουμπιά για την εκτέλεση των εντολών και την αναζήτηση των αρχείων. Γενικά χρησιμοποιήθηκε μια απλή και ευρέως αναγνωρίσιμη μορφή προγράμματος ώστε ο χρήστης να μην αντιμετωπίσει καμιά δυσκολία κατά την εφαρμογή του.

Η εισαγωγή δεδομένων στο πρόγραμμα γίνεται με την εισαγωγή από το χρήστη των διαστάσεων του πίνακα που έχει γραμμές το πλήθος των περιπτώσεων της έρευνας και στήλες τις μεταβλητές της έρευνας, του συντελεστή βαρύτητας ολικής ικανοποίησης w , της ελάχιστης τιμής του συντελεστή Καταλογισμού και του αρχείου που περιλαμβάνει τα δεδομένα του πίνακα. Οι διαστάσεις του πίνακα είναι προφανώς θετικοί ακέραιοι, ενώ ο συντελεστής βαρύτητας ολικής ικανοποίησης και η ελάχιστη τιμή του συντελεστή Καταλογισμού είναι θετικοί δεκαδικοί αριθμοί με πεδίο τιμών από 0 έως 1.

Η εισαγωγή των διαστάσεων του πίνακα είναι απαραίτητη ώστε αφενός το πρόγραμμα να είναι σε θέση να διαβάσει το σύνολο των δεδομένων της έρευνας και αφετέρου για να δίνεται η δυνατότητα στο χρήστη να εκτελεί δοκιμές, μειώνοντας ή αυξάνοντας τις περιπτώσεις και τις μεταβλητές που χρησιμοποιούνται για imputation κάθε φορά. Η συγκεκριμένη δυνατότητα μπορεί να φανεί πολύ χρήσιμη σε ταξινομημένους πίνακες σύμφωνα με το πλήθος των ελλিপών δεδομένων κάθε περίπτωσης, αφού ο ερευνητής μπορεί να προσαρμόζει την ελάχιστη τιμή του συντελεστή Καταλογισμού με διαφορετικό δείγμα έρευνας κάθε φορά χωρίς να μετατρέπει τον αρχικό πίνακα.

Η μορφή αρχείου που διαβάζεται από το πρόγραμμα είναι τύπου κειμένου κατάληξης .txt το οποίο παράγεται εύκολα από αρχείο excel, όπου είναι και το συνηθέστερο για καταχώρηση δεδομένων έρευνας, επιλέγοντας Αποθήκευση ως – Κείμενο (οριοθετημένο με Tab). Διαφορετικά, ο πίνακας μπορεί να γραφτεί απευθείας σε αρχείο .txt

όπου τα στοιχεία κάθε μεταβλητής είναι χωρισμένα με tab μεταξύ τους και οι διαφορετικές περιπτώσεις με αλλαγή γραμμής, δηλαδή οι γραμμές εκφράζουν τις περιπτώσεις, ενώ οι στήλες τις μεταβλητές.

Στο εν λόγω αρχείο έχει οριστεί η πρώτη στήλη ως η στήλη της ερώτησης ολικής ικανοποίησης, οπότε ο χρήστης μεταφέρει τα δεδομένα της ολικής ικανοποίησης στην στήλη αυτή. Εάν σε έρευνα δεν υπάρχει ερώτηση ολικής ικανοποίησης, ο χρήστης χρησιμοποιεί συντελεστή βαρύτητας ολικής ικανοποίησης $w = 0$.

4.9 Λειτουργία προγράμματος

Το πρόγραμμα λειτουργεί μέσω ενός εκτελέσιμου αρχείου .jar το οποίο μεταφέρεται με αντιγραφή – επικόλληση και δεν απαιτείται εγκατάσταση του. Με διπλό κλικ από το ποντίκι ή enter από το πληκτρολόγιο ανοίγει η φόρμα εισαγωγής δεδομένων στην οποία εμφανίζονται τα απαιτούμενα πεδία εισαγωγής και κουμπιά εκτέλεσης εντολών. Συγκεκριμένα, εμφανίζονται τα πεδία εισαγωγής:

“Number of Questions”

“Number of questionnaires”

“Imputation Coefficient Limit”

“Total Satisfaction Weight Coefficient”

όπου εισάγονται τα μεγέθη της έρευνας πλήθος Ερωτηματολογίων, πλήθος Κριτηρίων, ελάχιστη τιμή συντελεστή Καταλογισμού και ο συντελεστής βαρύτητας ολικής ικανοποίησης αντίστοιχα.

The screenshot shows a software window titled "Hot Deck Imputation Program". Below the title is a subtitle: "For customer satisfaction surveys with ordinal data". The interface contains four input fields with labels: "Number of Questions:", "Number of Questionnaires:", "Imputation Coefficient Limit:", and "Total Satisfaction Weight Coefficient:". Below these fields are two buttons: "Browse File" and "Clear". A large, disabled button labeled "Imputation" is positioned below the input fields. At the bottom left, it says "by Idomeneas Kamilakis", and at the bottom right, there is an "Exit" button. The window has standard Windows-style window controls (minimize, maximize, close) in the top right corner.

Σχήμα 4.1: Αρχική οθόνη προγράμματος

Όπως έχει καθοριστεί στον αλγόριθμο, τα μεγέθη πλήθος Ερωτηματολογίων και πλήθος Κριτηρίων είναι θετικοί ακέραιοι αριθμοί μεγαλύτεροι του 0, ενώ η ελάχιστη τιμή συντελεστή Καταλογισμού και ο συντελεστής βαρύτητας ολικής ικανοποίησης έχουν οριστεί δεκαδικοί float με τιμές από 0 έως 1.

Hot Deck Imputation Program
For customer satisfaction surveys with ordinal data

Number of Questions: 20

Number of Questionnaires: 300

Imputation Coefficient Limit: 0.8

Total Satisfaction Weight Coefficient: 0.5

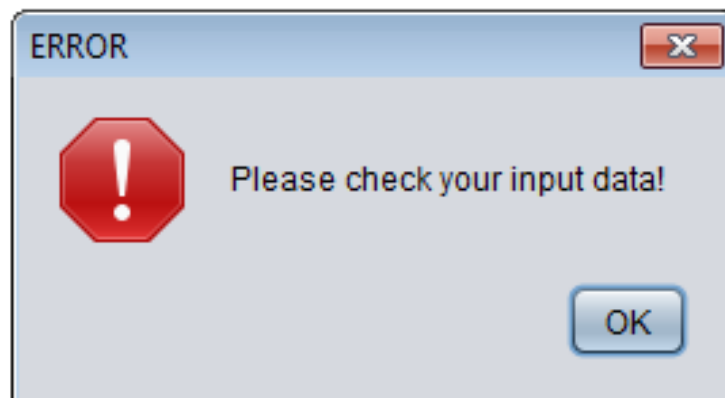
Browse File Clear

Imputation

by Idomeneas Kamilakis Exit

Σχήμα 4.2: Συμπληρωμένη φόρμα προγράμματος

Σε περίπτωση εισαγωγής διαφορετικού τύπου ή τιμής δεδομένων από το χρήστη, κατά το imputation εμφανίζεται παράθυρο σφάλματος, ειδοποιώντας το χρήστη για έλεγχο των δεδομένων που εισήγαγε.



Σχήμα 4.3: Μήνυμα σφάλματος σε δεδομένα έρευνας

Παρακάτω εμφανίζονται τα κουμπιά εκτέλεσης εντολών:

“Browse file”

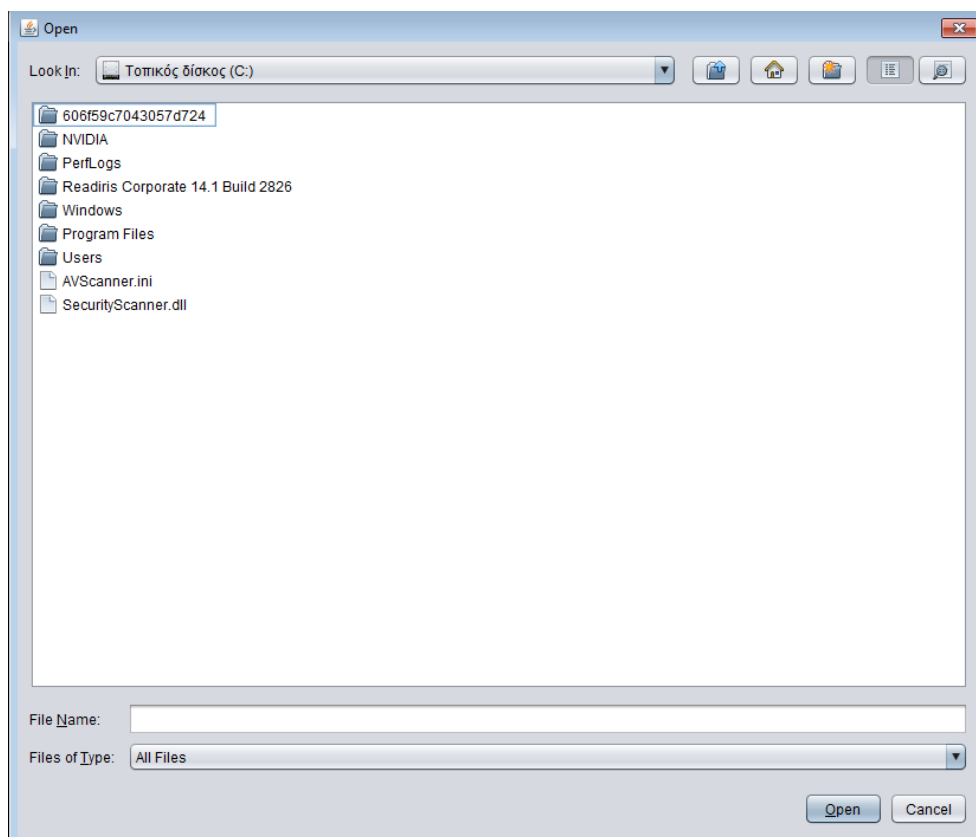
“Clear”

“Imputation”

“Exit”

Με τα οποία εκτελούνται οι εντολές αναζήτησης αρχείου, καθαρισμού δεδομένων και εκτέλεσης hot deck imputation αντίστοιχα.

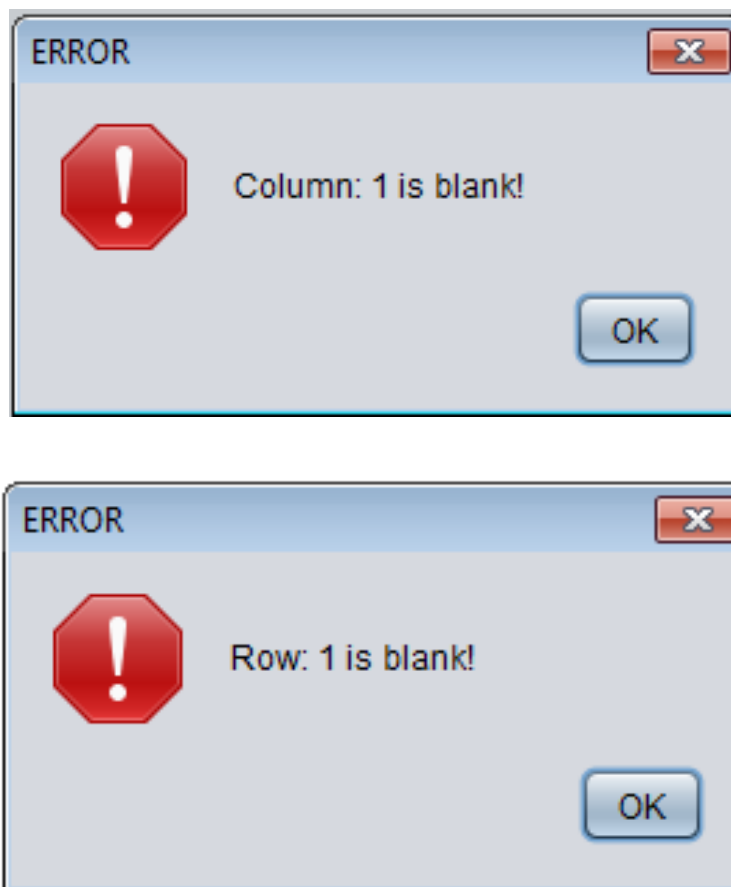
Με το κουμπί αναζήτησης αρχείου “Browse file”, ανοίγει νέο παράθυρο αναζήτησης περιβάλλοντος windows και γίνεται περιήγηση στα αρχεία του Η/Υ. Όταν ο χρήστης βρει το επιθυμητό αρχείο εισαγωγής δεδομένων τύπου .txt το επιλέγει και πατάει το κουμπί “open”.



Σχήμα 4.4: Παράθυρο αναζήτησης αρχείου

Το πρόγραμμα με αυτή την εντολή ανοίγει το αρχείο και διαβάζει τα δεδομένα τύπου integer, καταχωρώντας και αποθηκεύοντας τα στον πίνακα διαστάσεων “πλήθος Ερωτηματολογίων ” x “πλήθος Κριτηρίων”,

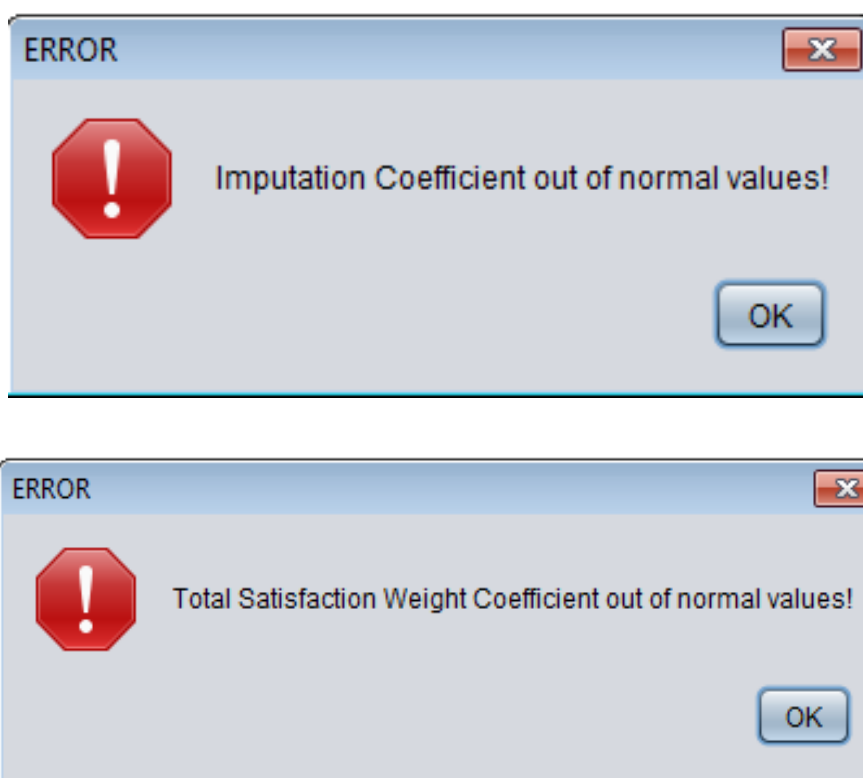
$X(m,n)$ που έχει εισάγει προηγουμένως ο χρήστης. Εάν σε αυτό το σημείο εντοπιστεί κάποιο σφάλμα όπως κατεστραμμένο αρχείο, αδύνατο άνοιγμα αρχείου, μη συμβατά δεδομένα κ.λπ. εμφανίζεται παράθυρο σφάλματος, ειδοποιώντας το χρήστη για έλεγχο των δεδομένων που εισήγαγε. Επίσης, εάν ο χρήστης έχει εισάγει λανθασμένα μεγαλύτερες διαστάσεις πίνακα ή στον πίνακα εντοπιστούν στήλη ή γραμμή μόνο με μηδενικά στοιχεία εμφανίζονται αντίστοιχα μηνύματα σφάλματος.



Σχήμα 4.5: Μήνυμα σφάλματος για κενή στήλη ή γραμμή

Με το κουμπί καθαρισμού δεδομένων “Clear”, διαγράφονται τα δεδομένα που έχει εισάγει ο χρήστης στα πεδία εισαγωγής, εάν θέλει να κάνει κάποια διόρθωση ή να εισάγει νέα. Επίσης εάν έχει εκτελεστεί ήδη το πρόγραμμα προηγουμένως, επαναφέρει τις παραμέτρους που έχουν χρησιμοποιηθεί από τον αλγόριθμο στην αρχική τους τιμή ώστε να μην δημιουργηθεί δυσλειτουργία σε νέα εκτέλεση του.

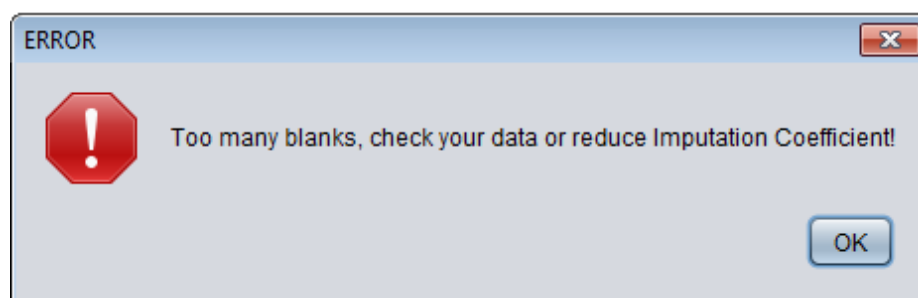
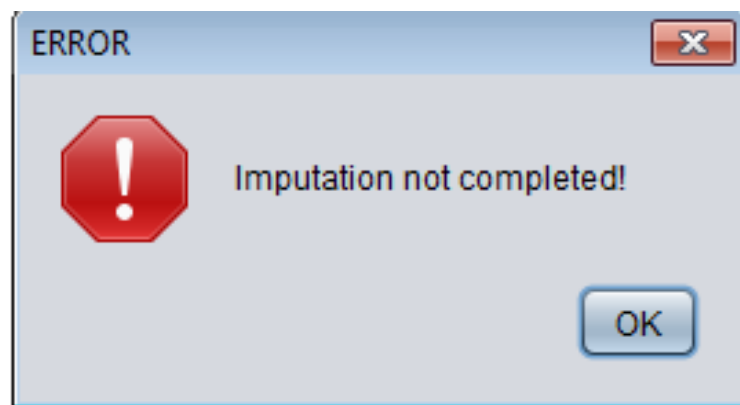
Με το κουμπί *imputation*, το οποίο ενεργοποιείται αφού διαβαστεί επιτυχώς το αρχείο *.txt*, διαβάζονται η ελάχιστη τιμή του συντελεστή Καταλογισμού και ο συντελεστής βαρύτητας ολικής ικανοποίησης που έχει εισάγει ο χρήστης, γίνονται έλεγχοι καταλληλότητας των δεδομένων και εισάγονται στον αλγόριθμο *hot deck*. Εάν βρεθεί μη επιτρεπτή τιμή σε κάποιο συντελεστή ειδοποιείται ο χρήστης ανάλογα με μήνυμα σφάλματος και δεν εκτελείται *imputation*.



Σχήμα 4.6: Μήνυμα σφάλματος για τιμές συντελεστών εκτός ορίων

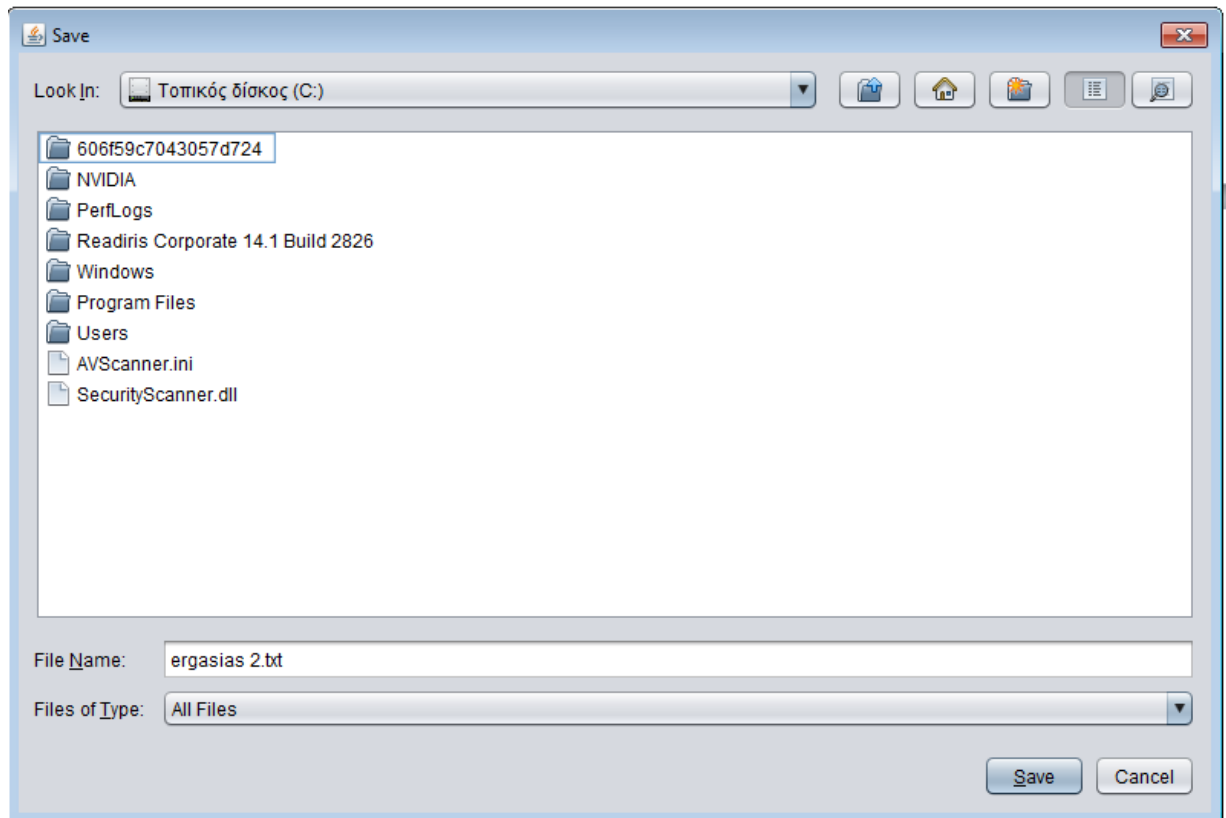
Εντός του αλγορίθμου κατασκευάζονται οι πίνακες συντελεστή Καταλογισμού και αποστάσεων *Walesiak* μεταξύ των περιπτώσεων και για κάθε ελλιπές στοιχείο αναζητείται το βέλτιστο ερωτηματολόγιο για *imputation* βάσει των κριτηρίων ελάχιστης απόστασης και επιτρεπτής τιμής συντελεστή Καταλογισμού. Εάν καλύπτονται οι παραπάνω συνθήκες γίνεται αντικατάσταση τιμής και ο αλγόριθμος αναζητά το επόμενο 999.

Διαφορετικά, εάν δεν καλύπτεται η συνθήκη της ελάχιστης τιμής του συντελεστή Καταλογισμού έστω και σε ένα ερωτηματολόγιο είτε λόγω πολλών ελλιπών στοιχείων, είτε λόγω πολύ υψηλής ελάχιστης τιμής του συντελεστή c , ο αλγόριθμος δεν μπορεί να ολοκληρώσει το imputation και εμφανίζεται μήνυμα μη ολοκλήρωσης. Σε αυτή την περίπτωση, ο χρήστης προτρέπεται να ελέγξει τα δεδομένα του για ύπαρξη ανεπιθύμητων ερωτηματολογίων με πολλά 999 ή να μειώσει την ελάχιστη τιμή του συντελεστή Καταλογισμού.



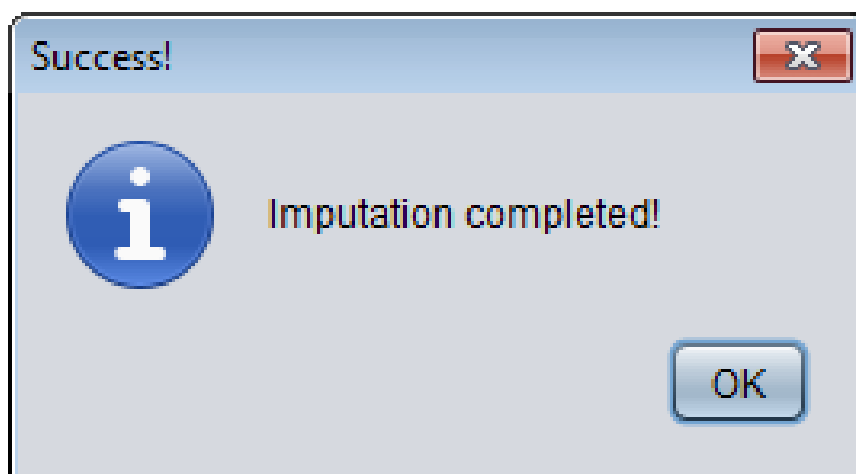
Σχήμα 4.7: Μηνύματα σφάλματος για μη εκτέλεση imputation λόγω πολλών ελλιπών στοιχείων

Όταν πλέον ο αλγόριθμος φτάσει στο τέλος και έχουν συμπληρωθεί όλα τα ελλιπή στοιχεία με κάποια τιμή διαφορετική του 999, εμφανίζεται παράθυρο επιλογών αποθήκευσης του αρχείου όπου είναι αποθηκευμένος ο νέος πίνακας δεδομένων στον Η/Υ του χρήστη. Το αρχείο είναι τύπου excel και τα δεδομένα του πίνακα είναι αποθηκευμένα ακριβώς με την αρχική μορφή που εισήχθησαν.



Σχήμα 4.8: Παράθυρο αποθήκευσης αρχείου

Αφού επιλεγθεί το όνομα αρχείου και η τοποθεσία αποθήκευσης, εμφανίζεται μήνυμα επιτυχούς εκτέλεσης και το πρόγραμμα επανέρχεται στην αρχική φόρμα εισαγωγής δεδομένων. Σε αυτό το σημείο ο χρήστης προτρέπεται πριν την επανεκτέλεση του προγράμματος να καθαρίζει τις τιμές των πεδίων με το κουμπί clear, ώστε να μην υπάρχει ενδεχόμενο αποθήκευσης παλαιών τιμών σε νέο τρέξιμο του προγράμματος.



Σχήμα 4.9: Μήνυμα επιτυχούς εκτέλεσης imputation

Προφανώς με το κουμπί exit τερματίζεται η λειτουργία του προγράμματος και διαγράφονται τα δεδομένα που έχουν ενδεχομένως τοποθετηθεί στα πεδία.

5. Έλεγχος και εφαρμογή αλγόριθμου Hot Deck

5.1 Έλεγχος και επαλήθευση αλγορίθμου

Το τελικό στάδιο για την ολοκλήρωση του προγράμματος είναι κατ' αρχήν η τεκμηρίωση ορθής λειτουργίας και σωστής απόδοσης των αποτελεσμάτων από τον αλγόριθμο και κατά δεύτερον η σωστή λειτουργία της φόρμας εκτέλεσης του προγράμματος. Ο καλύτερος τρόπος για να επιτευχθεί αυτό, είναι η εφαρμογή του σε κάποιο ήδη υπάρχων πίνακα δεδομένων στον οποίο δημιουργούνται τεχνητά ελλείψεις. Δηλαδή, απαιτείται η εφαρμογή του προγράμματος σε πλήρη έρευνα από την οποία έχουν αφαιρεθεί σκόπιμα και τυχαία δεδομένα, προσομοιάζοντας μια έρευνα με κενά στις απαντήσεις των ερωτηθέντων.

Η αξιολόγηση των αποτελεσμάτων γίνεται μέσω σύγκρισης της πραγματικής (αναμενόμενης) με την αποδιδόμενη από το πρόγραμμα τιμή. Για την ευκολότερη αξιολόγηση, τα αποτελέσματα παρουσιάζονται σε συγκεντρωτικούς πίνακες αποτελεσμάτων που ονομάζονται *confusion matrices* ή *confusion tables*. Στους εν λόγω πίνακες οι γραμμές αναφέρονται στο πλήθος των πραγματικών τιμών που αφαιρέθηκαν από τον αρχικό πίνακα δεδομένων, ενώ οι στήλες στο πλήθος των τιμών που αποδόθηκαν από το πρόγραμμα για κάθε διαφορετική πραγματική τιμή. Η κύρια διαγώνιος του πίνακα αποτελείται από τις επιτυχημένες (ακριβής) προβλέψεις του προγράμματος, ενώ οποιοδήποτε στοιχείο εκτός αυτής καταδεικνύει διαφορετική απόδοση τιμής.

Μέσω των *confusion tables* ο ερευνητής μπορεί να διακρίνει εύκολα πόσες από τις ελλειπείς τιμές της έρευνας αποδόθηκαν σωστά από το πρόγραμμα και επιπλέον να εξάγει την κατανομή των λανθασμένων αποτελεσμάτων. Είναι σε θέση δηλαδή εκτός από τον έλεγχο ορθής λειτουργίας του προγράμματος, ανάλογα με τις προδιαγραφές που έχει ο ίδιος θέσει, να ελέγξει πιθανές τάσεις που εμφανίζονται στα αποτελέσματα και να τις αξιολογήσει.

Για πιο αξιόπιστα συμπεράσματα η διαδικασία επαναλαμβάνεται αρκετές φορές ώστε να εξαλειφτούν τυχόν τυχαίες ατέλειες εισαγωγής κενών και πρόβλεψης τιμών από τον αλγόριθμο. Όσες περισσότερες φορές εφαρμοστεί το πρόγραμμα τόσο πιο αξιόπιστα μπορεί να κατασκευαστεί το confusion table. Η τεχνική αφαίρεσης δεδομένων από την αρχική έρευνα επαναλαμβάνεται όσες φορές αποφασίσει ο ερευνητής και δημιουργούνται ισάριθμοι πίνακες με ελλιπή δεδομένα. Το πρόγραμμα εφαρμόζεται σε κάθε ένα από αυτούς τους πίνακες και αποδίδονται οι αντίστοιχοι συμπληρωμένοι πίνακες σε μορφή αρχείου excel.

Στη συνέχεια κατασκευάζονται ισάριθμα confusion tables για κάθε εικονική έρευνα και τελικά τα αποτελέσματα συγκεντρώνονται σε ένα confusion table με τις ποσοστιαίες κατανομές των απαντήσεων. Με αυτό τον τρόπο δίνεται η δυνατότητα στον ερευνητή να γνωρίζει άμεσα την αποδοτικότητα του αλγόριθμου, δηλαδή την επιτυχία του να δίνει σωστές προβλέψεις για τη συμπλήρωση των κενών και με ποιο ποσοστό αποκλίνει ανά λανθασμένο αποτέλεσμα σε κάθε απάντηση.

Σύμφωνα με την παραπάνω μεθοδολογία, επιλέχθηκε έρευνα πραγματικών δεδομένων στην οποία κατασκευάστηκαν τεχνητές ελλείψεις, οι οποίες στη συνέχεια συμπληρώθηκαν μέσω εφαρμογής του προγράμματος hot deck που αναπτύχθηκε στην παρούσα εργασία. Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από έρευνα ικανοποίησης τουριστών. Η έρευνα αποτελείται από οκτώ ερωτήσεις η οποία έγινε σε 3893 ερωτηθέντες με κλίμακα απαντήσεων από 1 έως 5 και αποτελείται από πλήρη στοιχεία όπου η πρώτη ερώτηση αποτελεί την ερώτηση ολικής ικανοποίησης.

Στον πίνακα εισαγωγής δεδομένων 3893x8 τοποθετήθηκαν 200 κενά με τυχαίο τρόπο (συνάρτηση random() του excel) προσομοιάζοντας 200 κενές απαντήσεις στα ερωτηματολόγια της έρευνας. Το νούμερο αυτό επιλέχθηκε ως ένα λογικό ποσοστό ελλιπών δεδομένων σε μια πραγματική και αρκετά καλά σχεδιασμένη έρευνα.

Κατασκευάστηκε το αρχείο εισαγωγής δεδομένων .txt και εισήχθη στο πρόγραμμα επιλέγοντας ελάχιστη τιμή συντελεστή

Καταλογισμού 1, δηλαδή συμπλήρωση των ελλিপών δεδομένων μόνο από ερωτηματολόγια χωρίς κενά και συντελεστή βαρύτητας ολικής ικανοποίησης 0,5. Στη συνέχεια, έτρεξε το πρόγραμμα και δημιουργήθηκε νέο αρχείο του πίνακα δεδομένων με συμπληρωμένες τις ελλειπείς τιμές. Από τη σύγκριση των αποτελεσμάτων με τον αρχικό συμπληρώθηκε το confusion table, όπου παριστάνονται οι πραγματικές και οι αποδιδόμενες τιμές.

Η παραπάνω διαδικασία επαναλήφθηκε 10 φορές και τα αποτελέσματα συγκεντρώθηκαν σε 10 διαφορετικά confusion tables. Τέλος κατασκευάστηκε ένα συνολικό confusion table που συγκεντρώνει τα δέκα διαφορετικά αποτελέσματα το οποίο παρουσιάζεται παρακάτω:

Πίνακας 5.1: Confusion table αποτελεσμάτων με απόλυτες τιμές

		<i>Imputed values</i>				
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>
<i>Original values</i>	<i>1</i>	2	2	11	8	2
	<i>2</i>	2	5	24	20	4
	<i>3</i>	4	12	98	153	35
	<i>4</i>	8	19	184	615	194
	<i>5</i>	3	3	27	215	350
2000						

Σε γενικές γραμμές μπορεί να ειπωθεί ότι ο αλγόριθμος απέδωσε πολύ καλά, αφού σε 2000 αντικαταστάσεις τεχνητά ελλিপών δεδομένων πέτυχε την ακριβή αρχική τιμή 1070 φορές, ενώ άλλες 786 φορές πέτυχε την παραπλήσια τιμή, δηλαδή ± 1 της κλίμακας απαντήσεων.

Εξάλλου όπως έχει αναφερθεί στη θεωρία της μεθόδου hot deck, σκοπός δεν είναι η αντικατάσταση με μια πραγματική τιμή που θεωρητικά θα είχε δοθεί από τον ερωτηθέντα, αλλά η τοποθέτηση τιμών που να μην αλλοιώνουν τα στατιστικά μεγέθη της έρευνας κάνοντας δυνατή την επεξεργασία των δεδομένων.

Από τον παραπάνω πίνακα κατασκευάστηκε νέο confusion table που συγκεντρώνει τα ποσοστά των αποδιδόμενων τιμών για κάθε αρχική τιμή επί του συνολικού πλήθους των δεδομένων και φαίνεται παρακάτω.

Πίνακας 5.2: Confusion table αποτελεσμάτων με ποσοστά % των αποδιδόμενων τιμών επί του συνολικού πλήθους

		<i>Imputed values</i>				
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>
<i>Original values</i>	<i>1</i>	0,1	0,1	0,55	0,4	0,1
	<i>2</i>	0,1	0,25	1,2	1	0,2
	<i>3</i>	0,2	0,6	4,9	7,65	1,75
	<i>4</i>	0,4	0,95	9,2	30,75	9,7
	<i>5</i>	0,15	0,15	1,35	10,75	17,5

Παρατηρείται λοιπόν στο confusion table των ποσοστών ότι οι τιμές της κύριας διαγωνίου, όπου εκφράζεται η ακριβής επιτυχία πρόβλεψης, έχουν άθροισμα 53,5%, ενώ οι δύο παραπλήσιες διαγώνιες 39,3%. Αθροιστικά ο αλγόριθμος προέβλεψε σε ποσοστό 92,8% την ακριβή ή την παραπλήσια τιμή.

Από τους δύο πίνακες επίσης, παρατηρείται ότι ο αλγόριθμος απέδωσε πολύ καλά τις προσδοκώμενες τιμές στις περιπτώσεις των απαντήσεων 4 και 5, όπου τα μεγαλύτερα ποσοστά βρίσκονται στην κύρια διαγώνιο, αλλά λιγότερο καλά στις μικρότερες τιμές όπου οι μέγιστες τιμές βρίσκονται εκτός κύριας διαγωνίου.

Μπορεί να δοθεί εξήγηση για το φαινόμενο αυτό εάν εξεταστεί η κατανομή των τιμών των δεδομένων στην έρευνα και κατ' επέκταση των ελλιπών στοιχείων. Παρατηρείται εύκολα ότι στα αρχικά δεδομένα υπάρχει μεγάλη ανισότητα στην κατανομή του πλήθους των τιμών. Οι τιμές 4 και 5, όπου ο αλγόριθμος είχε τα μεγαλύτερα ποσοστά επιτυχίας, αποτελούν το 50% και 29% στο σύνολο των 31144 απαντήσεων που έχουν δοθεί. Εφόσον λοιπόν τα κενά στον πίνακα

εισήχθησαν με τυχαίο τρόπο, η κατανομή των κενών ακολουθεί σε γενικές γραμμές την κατανομή των τιμών των μεταβλητών. Από τα 2000 ελλιπή στοιχεία που εισήχθησαν, το 51% και το 30% αντιστοιχούσε στις αρχικές τιμές 4 και 5. Στον παρακάτω πίνακα φαίνεται το μέγεθος της άνισης κατανομής των απαντήσεων στον αρχικό πίνακα δεδομένων και η επίσης άνιση κατανομή των τυχαίων 999 που εισήχθησαν στον πίνακα για τις 10 φορές που εκτελέστηκε η διαδικασία.

Πίνακας 5.3: Κατανομή τιμών στο αρχικό δείγμα και κατανομή κενών

Τιμή	Πλήθος τιμών σε αρχικό δείγμα	% πλήθους τιμών σε αρχικό δείγμα	Πλήθος ελλιπών δεδομένων ανά τιμή	% πλήθους ελλιπών δεδομένων ανά τιμή
1	375	1%	25	1%
2	1052	3%	55	3%
3	5058	16%	302	15%
4	15653	50%	1020	51%
5	9006	29%	598	30%

Προκύπτει δηλαδή, ότι ο αλγόριθμος αποδίδει πολύ καλά στις τιμές που συγκεντρώνεται το μεγαλύτερο πλήθος απαντήσεων, ενώ λιγότερο καλά στις χαμηλές τιμές όπου το πλήθος είναι πολύ μικρό. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι ο αλγόριθμος σε τέτοια κατανομή τιμών είναι πολύ πιθανό να βρει κάποια υψηλή τιμή για αντικατάσταση ενός κενού. Δηλαδή, όταν η πραγματική ελλιπής τιμή είναι χαμηλή, ο αλγόριθμος έχει περισσότερες πιθανότητες να βρει “όμοια” περίπτωση με υψηλότερη τιμή στη συγκεκριμένη μεταβλητή. Αντίθετα είναι δύσκολο να βρει “όμοια” περίπτωση με την ίδια ή χαμηλότερη βαθμολογία στην μεταβλητή αυτή αφού ουσιαστικά οι περιπτώσεις αυτές “σπανίζουν”.

Άρα είναι προφανές ότι για κάθε ελλιπή τιμή που τοποθετήθηκε στις θέσεις των αρχικών τιμών 1 και 2, ο αλγόριθμος βρήκε εύκολα

περιπτώσεις-δότες με τιμές 3 ή 4 για αντικατάσταση αφού συμμετέχουν σε πολύ υψηλό ποσοστό στο σύνολο των τιμών της έρευνας.

Συμπερασματικά παρατηρείται ότι ο αλγόριθμος απέδωσε σε μεγάλο βαθμό ακριβώς τις ίδιες τιμές με τις αρχικές και σε επίσης σε μεγάλο βαθμό με τις παραπλήσιες τους. Εάν δηλαδή θεωρήσουμε ως αποδεκτή την αντικατάσταση της παραπλήσιας τιμής ο αλγόριθμος έχει επιτυχία σε ποσοστό 92,8%. Ουσιαστικές αποκλίσεις παρατηρήθηκαν μόνο στις τιμές 1 και 2 όπου “τραβήχτηκαν” προς το κέντρο αποδίδοντας σε μεγάλο ποσοστό την τιμή 3, αλλά εφόσον αποτελούν μόλις το 1% και 3% του συνόλου της έρευνας, κρίνεται ότι δεν επηρεάζουν σημαντικά τα στατιστικά δεδομένα του συνόλου. Εξάλλου, η λειτουργία του αλγόριθμου σε διαφορετικά δεδομένα με ομοιόμορφη κατανομή, θα εξαλείψει αυτά τα φαινόμενα και τα αποτελέσματα θα είναι ίδια σε όλο το εύρος των τιμών.

5.2 Εφαρμογή σε έρευνα ικανοποίησης

Μετά και την επιτυχή τεκμηρίωση ορθής λειτουργίας του προγράμματος και του αλγόριθμου Hot Deck, αποφασίστηκε η εκτέλεση του σε πραγματική έρευνα που αφορά την ικανοποίηση των ασθενών για τις υπηρεσίες που παρέχονται από το ΤΕΠ νοσοκομείου στην Αττική.

Η έρευνα αποτελείται από ερωτηματολόγιο είκοσι ερωτήσεων όπου η τελευταία ερώτηση είναι η ερώτηση ολικής ικανοποίησης για τις υπηρεσίες που παρασχέθηκαν. Στον πίνακα δεδομένων που προέκυψε από την έρευνα υπάρχουν αρκετά ελλιπή στοιχεία τα οποία καθιστούν την στατιστική επεξεργασία επισφαλής. Για αυτό το λόγο κρίνεται απαραίτητη η εφαρμογή μιας μεθόδου συμπλήρωσης των κενών, η οποία θα προσφέρει τη δυνατότητα αξιοποίησης περισσότερων ερωτηματολογίων και θα κάνει τα στατιστικά αποτελέσματα πιο αξιόπιστα.

5.2.1 Γενικά στοιχεία έρευνας

Η έρευνα αφορά τη μέτρηση της ικανοποίησης ενός συνόλου ασθενών, όπου το ερωτηματολόγιο που έχει χρησιμοποιηθεί περιλαμβάνει δεκαεννέα ερωτήσεις – κριτήρια, καθώς και την ερώτηση ολικής ικανοποίησης. Οι ασθενείς κλήθηκαν να απαντήσουν στην ερώτηση “ΠΑΡΑΚΑΛΟΥΜΕ, ΑΞΙΟΛΟΓΗΣΤΕ ΤΙΣ ΕΝΤΥΠΩΣΕΙΣ ΣΑΣ ΣΧΕΤΙΚΑ ΜΕ ΤΑ ΠΑΡΑΚΑΤΩ:” στις δεκαεννέα ερωτήσεις με πιθανές απαντήσεις τις :

Πολύ κακές	Μάλλον κακές	Ούτε καλές ούτε κακές	Μάλλον καλές	Πολύ καλές
---------------	-----------------	--------------------------	-----------------	------------

Όπου αναλυτικά οι ερωτήσεις είναι:

- *χρόνος αναμονής*
- *πινακίδες*
- *καθαριότητα*
- *ελεύθερα καθίσματα*
- *θερμοκρασία*
- *ησυχία*
- *καθαριότητα βοηθ. Χώρων*
- *χρόνος αναμονής εξέτασης*
- *χρόνος έκδοσης αποτελεσμάτων*
- *συμπεριφορά προσωπικού*
- *σεβασμός προσωπικότητας*
- *ποιότητα ιατρικής φροντίδας*
- *συμπεριφορά γιατρών*
- *ενημέρωση από γιατρούς*
- *συμπεριφορά γιατρών σε εξέταση*
- *ποιότητα νοσηλευτικής φροντίδας*
- *συμπεριφορά νοσηλευτών*
- *συμπεριφορά διοικητικού προσωπικού*
- *ταχύτητα διεκπεραίωσης διοικητικών*

Στον πίνακα δεδομένων η κωδικοποίηση της κλίμακας απαντήσεων για τις 19 ερωτήσεις είναι από το 1 έως το 5, όπου το 1

αντιπροσωπεύει την απάντηση “Πολύ κακές” ενώ το 5 την απάντηση “Πολύ καλές”.

Για την ερώτηση ολικής ικανοποίησης η κλίμακα είναι αριθμητική και οι ασθενείς κλήθηκαν να βαθμολογήσουν από το 0 έως το 10, ενώ τα ελλιπή στοιχεία σε όλες τις ερωτήσεις έχουν κωδικοποιηθεί με την τιμή 999.

Η έρευνα εκτελέστηκε σε 307 ασθενείς, ενώ συνολικά υπάρχουν 843 ελλιπείς τιμές από τα συνολικά 6140 στοιχεία του πίνακα απαντήσεων, δηλαδή υπάρχει έλλειψη σε ποσοστό 13.7%. Το ποσοστό αυτό είναι υψηλό και εύκολα γίνεται αντιληπτό ότι απαιτείται επεξεργασία των δεδομένων πριν την εκτέλεση της μεθόδου hot deck, αφού υπάρχουν αρκετά ερωτηματολόγια με πολλά ελλιπή στοιχεία. Για παράδειγμα στα 21 από τα 307 ερωτηματολόγια υπάρχουν πάνω από 10 ελλιπή στοιχεία. Τα ερωτηματολόγια αυτά πρέπει να εξαιρεθούν από την αντικατάσταση και φυσικά να αφαιρεθούν από την έρευνα αφού δεν προσφέρονται για στατιστική επεξεργασία λόγω υψηλού ρίσκου.

Για αυτό το λόγο επιλέχθηκε ελάχιστη τιμή συντελεστή Καταλογισμού 0.8 το οποίο σημαίνει από την ανισότητα:

$$c \geq \frac{(m - t)}{m} + \frac{1}{m}$$

$$\text{ότι επιτρέπονται } t \geq m + 1 - c * m = 5$$

ελλιπή στοιχεία ανά ερωτηματολόγιο. Από την παραπάνω ανάλυση προκύπτει ότι αφαιρέθηκαν 61 ερωτηματολόγια από τον τελικό πίνακα ο οποίος τελικά έχει διαστάσεις 246x20 και περιέχει 296 ελλιπή στοιχεία δηλαδή έλλειψη 6%.

Τα δεδομένα του τελικού πίνακα στη συνέχεια μορφοποιήθηκαν ώστε να είναι δυνατή η εισαγωγή τους στο πρόγραμμα και η ερώτηση ολικής ικανοποίησης μεταφέρθηκε στην πρώτη στήλη του πίνακα όπως έχει καθοριστεί στον αλγόριθμο hot deck.

Στη συνέχεια τα δεδομένα της έρευνας (246 ερωτηματολόγια, 20 ερωτήσεις, 0.8 ελάχιστη τιμή συντελεστή Καταλογισμού, 0,5 συντελεστής βαρύτητας ολικής ικανοποίησης) μαζί με τον τελικό πίνακα απαντήσεων εισήχθησαν στο πρόγραμμα και έγινε απόδοση τιμής στα ελλιπή στοιχεία με επιτυχία. Έτσι η μέθοδος hot deck απέδωσε πλήρως συμπληρωμένο πίνακα 246x20 προσομοιάζοντας μια πλήρως συμπληρωμένη έρευνα η οποία είναι δυνατόν να μελετηθεί από οποιοδήποτε στατιστικό πακέτο.

5.2.2 Απόδοση τιμών και εφαρμογή μεθόδου MUSA

Στην παρούσα εργασία, για την ανάλυση και την εξαγωγή συμπερασμάτων από τον πλήρη πίνακα δεδομένων που προέκυψε, αποφασίστηκε η εφαρμογή της μεθόδου MUSA (MULTicriteria Satisfaction Analysis). Η μέθοδος MUSA αποτελεί την πολυκριτηριακή αναλυτική-συνθετική προσέγγιση για το πρόβλημα της μέτρησης και της ανάλυσης της ικανοποίησης. Η πρωτότυπη αυτή μεθοδολογία βασίζεται στην πολυκριτήρια ανάλυση αποφάσεων, υιοθετώντας τις βασικές αρχές της αναλυτικής- συνθετικής προσέγγισης και της θεωρίας των συστημάτων αξιών ή χρησιμότητας. Έτσι η MUSA έχει τη δυνατότητα να χειριστεί ποιοτικά δεδομένα, παράγοντας ένα σύνολο ποσοτικών αποτελεσμάτων, που μπορούν να αναλύσουν σε βάθος την ικανοποίηση ενός συνόλου πελατών.

Οι συνήθεις στατιστικές μέθοδοι που χρησιμοποιούνται για την μέτρηση μέτρησης ικανοποίησης πελατών, δεν είναι μέθοδοι που έχουν αναπτυχθεί αποκλειστικά για το σκοπό αυτό και κατά την εκτέλεση τους παρουσιάζονται αρκετά προβλήματα (π.χ. με την ποσοτικοποίηση των δεδομένων, την ύπαρξη πολλαπλής συγγραμμικότητας, τη δυσκολία κατανομής πελατών). Αντίθετα, η μέθοδος MUSA έχει αναπτυχθεί ειδικά για την μέτρηση της ικανοποίησης πελατών μίας επιχείρησης ξεπερνώντας τα παραπάνω προβλήματα.

Ορισμένα από τα βασικά πλεονεκτήματα της μεθόδου MUSA είναι:

- Η διαχείριση των δεδομένων, η οποία γίνεται κωδικοποιώντας το σύνολο των προκαθορισμένων λεκτικών απαντήσεων που μπορεί να δοθούν από τους πελάτες.
- Ο μη αυθαίρετος καθορισμός των παραμέτρων ο οποίος γίνεται βάσει των δεδομένων που προκύπτουν από την έρευνα.
- Η πληρότητα των αποτελεσμάτων αφού προσδιορίζονται τόσο η ολική όσο και η μερική ικανοποίηση των πελατών για κάθε διάσταση ικανοποίησης.
- Η δυνατότητα κατασκευής των διαγραμμάτων δράσης και βελτίωσης υποδεικνύοντας τα σημεία στα οποία η επιχείρηση πρέπει να βελτιωθεί για να αυξήσει την ικανοποίηση των πελατών της καθώς και την προτεραιότητα που πρέπει να δώσει στις ενέργειες βελτίωσης.

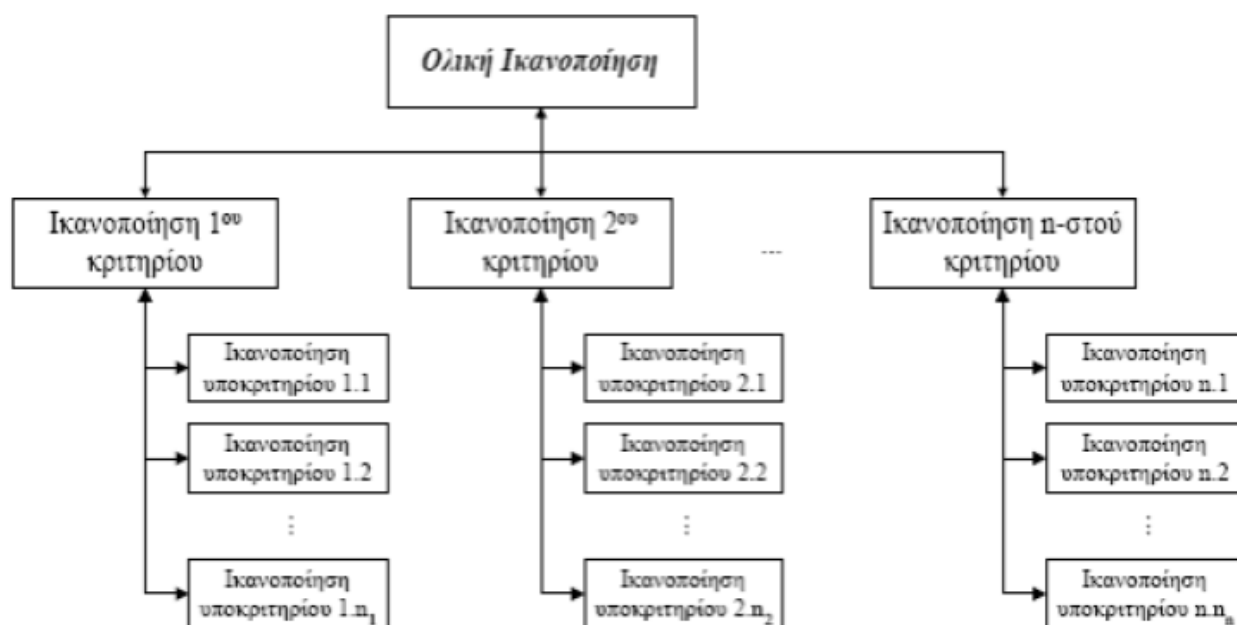
Οι κύριες παραδοχές που αφορούν την ανάπτυξη της μεθόδου MUSA εστιάζονται στα εξής σημεία: (Γρηγορούδης και Σίσκος, 2000):

- Ορθολογικός καταναλωτής: Η συγκεκριμένη υπόθεση σχετίζεται με την ύπαρξη ορθολογικών πελατών και συναντάται στο σύνολο του χώρου της επιστήμης των Αποφάσεων.
- Κριτήρια ικανοποίησης : η μέθοδος MUSA υποθέτει την ύπαρξη ενός συνόλου χαρακτηριστικών του εξεταζόμενου προϊόντος ή υπηρεσίας, σύμφωνα με τα οποία οι πελάτες αντιλαμβάνονται την ικανοποίησή τους. Το σύνολο αυτό των χαρακτηριστικών αποτελεί τα κριτήρια ικανοποίησης των πελατών και οφείλει να πληροί συγκεκριμένες ιδιότητες.
- Προσθετικό μοντέλο σύνθεσης: Τέλος, γίνεται η παραδοχή ύπαρξης ενός προσθετικού μοντέλου σύνθεσης του συνόλου των κριτηρίων ικανοποίησης και ειδικότερα μια προσθετική συνάρτηση αξιών.

Ο βασικός σκοπός της πολυκριτήριας μεθόδου MUSA είναι η σύνθεση των προτιμήσεων ενός συνόλου πελατών σε μια ποσοτική μαθηματική συνάρτηση αξιών. Πιο συγκεκριμένα, η μέθοδος υποθέτει ότι η συνολική ικανοποίηση ενός μεμονωμένου πελάτη εξαρτάται από ένα σύνολο μεταβλητών, τα οποία αντιπροσωπεύουν τα χαρακτηριστικά του προσφερόμενου προϊόντος ή υπηρεσίας.

Έτσι, η εκτίμηση της ικανοποίησης ενός συνόλου πελατών μπορεί να θεωρηθεί σαν ένα πρόβλημα στο επιστημονικό πεδίο της πολυκριτήριας ανάλυσης, υποθέτοντας ότι η συνολική ικανοποίηση ενός πελάτη εξαρτάται από ένα σύνολο κριτηρίων :

$$X = (X_1, X_2, \dots, X_n)$$



Σχήμα 5.6: Σύνθεση Ολικής Ικανοποίησης

Τα απαιτούμενα δεδομένα της μεθόδου συλλέγονται από ένα απλό, αλλά εξειδικευμένο ερωτηματολόγιο, σύμφωνα με το οποίο ζητείται από κάθε πελάτη να αξιολογήσει τις υπηρεσίες που του προσφέρονται, δηλαδή να εκφράσει τόσο τη συνολική όσο και την επιμέρους ικανοποίηση για κάθε ένα από τα κριτήρια-χαρακτηριστικά του προϊόντος ή της υπηρεσίας αυτής.

Οι προτιμήσεις αυτές των πελατών εκφράζονται σύμφωνα με μια μονότονη προκαθορισμένη ποιοτική κλίμακα. Το μοντέλο της MUSA προσπαθεί να εκτιμήσει τη συνολική και τις επιμέρους συναρτήσεις ικανοποίησης αντίστοιχα, δεδομένων των προτιμήσεων που έχει εκφράσει το σύνολο των πελατών (Grigoroudis and Siskos, 2003).

Η βασική εξίσωση ποιοτικής παλινδρόμησης είναι η εξής:

$$\widetilde{Y}^* = \sum_{i=1}^n b_i \cdot X_i^* - \sigma^+ + \sigma^-$$

όπου οι συναρτήσεις \widetilde{Y}^* και X_i^* εκφράζουν τις προτιμήσεις ενός συνόλου καταναλωτών και σ^+ , σ^- είναι αντίστοιχα το σφάλμα υπερεκτίμησης και υποεκτίμησης. Η παραπάνω εξίσωση ισχύει για κάθε ένα πελάτη που έχει εκφράσει μια συγκεκριμένη άποψη ικανοποίησης και για το λόγο αυτό οι μεταβλητές σφάλματος θα πρέπει να οριστούν για κάθε πελάτη χωριστά.

Βάσει των συλλεχθέντων δεδομένων επιλύεται το αντίστοιχο γραμμικό πρόγραμμα με χρήση της MUSA και εξάγονται τα αποτελέσματα που αφορούν στην ικανοποίηση των πελατών τα οποία παρουσιάζονται τόσο αριθμητικά όσο και σε γραφήματα. Τα κύρια αποτελέσματα της μεθόδου MUSA είναι:

Τα βάρη που αποδίδονται σε κάθε διάσταση ικανοποίησης

Τα βάρη των κριτηρίων ικανοποίησης υποδηλώνουν το σχετικό βαθμό σπουδαιότητας που δίνει το σύνολο των πελατών στις αξίες των διαστάσεων ικανοποίησης που έχουν καθοριστεί. Το γεγονός αυτό υποδηλώνει ότι η απόφαση για να κριθεί κάποιο κριτήριο ως «σημαντικό», σε ένα βαθμό, εξαρτάται και από το πλήθος των κριτηρίων που χρησιμοποιούνται. Δε θα πρέπει να λησμονείται η φυσική ερμηνεία των συντελεστών βαρύτητας, ότι τα βάρη είναι βαθμοί παραχώρησης (trade-offs) μεταξύ των αξιών στα διαφορετικά κριτήρια.

Ο μέσος δείκτης ικανοποίησης για κάθε διάσταση ικανοποίησης αλλά και ο ολικός δείκτης ικανοποίησης

Ο ορισμός των μέσων δεικτών ικανοποίησης συμπληρώνει τα δυνατά αποτελέσματα της μεθοδολογίας MUSA και κρίνεται αναγκαίος διότι:

- παρουσιάζει με απλό και κατανοητό τρόπο την κατάσταση της ικανοποίησης ενός συνόλου πελατών,
- συνδυάζει τα αποτελέσματα της μεθόδου MUSA με την περιγραφική στατιστική ανάλυση της έρευνας ικανοποίησης και
- δίνει τη δυνατότητα υλοποίησης ενός συστήματος αξιολόγησης της επιχείρησης.

Ο μέσος δείκτης απαιτητικότητας για κάθε διάσταση ικανοποίησης αλλά και ο ολικός δείκτης απαιτητικότητας

Οι δείκτες απαιτητικότητας εκφράζουν την μέση απόκλιση των συναρτήσεων ικανοποίησης από μια «κανονική» (γραμμική) συνάρτηση αξιών. Οι μέσοι δείκτες απαιτητικότητας δείχνουν το μέγεθος της προσπάθειας που καταβάλλεται για τη βελτίωση ενός χαρακτηριστικού, δεδομένου ότι όσο πιο απαιτητικοί είναι οι πελάτες, τόσο περισσότερο πρέπει να βελτιωθεί το επίπεδο ικανοποίησης για να εκπληρωθούν οι προσδοκίες τους.

Ο μέσος δείκτης αποτελεσματικότητας για κάθε διάσταση ικανοποίησης

Το αποτέλεσμα των ενεργειών βελτίωσης εξαρτάται τόσο από τη σημαντικότητα του κριτηρίου, όσο και από τη συνεισφορά του στη μη-ικανοποίηση (δυσαρέσκεια) των πελατών. Οι δείκτες αυτοί δείχνουν τα περιθώρια βελτίωσης σε ένα συγκεκριμένο κριτήριο ικανοποίησης, λαμβάνοντας υπόψη και τη σπουδαιότητά του. Επομένως, ο δείκτης αποτελεσματικότητας ενός κριτηρίου είναι το γινόμενο του βάρους με το δείκτη δυσαρέσκειας (συμπλήρωμα της ικανοποίησης) του συγκεκριμένου κριτηρίου.

Οι συναρτήσεις ικανοποίησης

Οι εκτιμώμενες συναρτήσεις ικανοποίησης εκφράζουν την πραγματική αξία που προσδίδει το σύνολο των πελατών σε ένα καθορισμένο ποιοτικό επίπεδο ικανοποίησης. Η μορφή των συναρτήσεων αυτών είναι σε θέση να προσδιορίσει το βαθμό απαιτητικότητας των πελατών (τα αποτελέσματα ισχύουν τόσο για την ολική, όσο και για τις μερικές συναρτήσεις ικανοποίησης). Έτσι, έχουμε τις ακόλουθες ομάδες πελατών: (Γρηγορούδης και Σίσκος, 2000)

- Ουδέτεροι πελάτες: η συνάρτηση ικανοποίησης έχει γραμμική μορφή, γεγονός που σημαίνει ότι οι συγκεκριμένοι πελάτες όσο περισσότερο ικανοποιημένοι δηλώνουν ότι είναι, τόσο μεγαλύτερο είναι το ποσοστό των προσδοκιών τους που εκπληρώνεται.
- Απαιτητικοί πελάτες: η συνάρτηση ικανοποίησης έχει κυρτή μορφή, δεδομένου ότι η ομάδα αυτή των πελατών δεν είναι ικανοποιημένη παρά μόνο αν τους προσφέρεται το βέλτιστο επίπεδο υπηρεσιών.
- Μη απαιτητικοί πελάτες: η συνάρτηση ικανοποίησης έχει κοίλη μορφή, γεγονός που υποδηλώνει ότι οι συγκεκριμένοι πελάτες δηλώνουν ότι είναι ικανοποιημένοι παρόλο που μόνο ένα μικρό ποσοστό των προσδοκιών τους εκπληρώνεται.

Το διάγραμμα δράσης

Συνδυάζοντας τα βάρη των κριτηρίων ικανοποίησης με τους μέσους δείκτες ικανοποίησης είναι δυνατός ο υπολογισμός μιας σειράς διαγραμμάτων δράσης τα οποία μπορούν να προσδιορίσουν ποια είναι τα δυνατά και τα αδύνατα σημεία της ικανοποίησης των πελατών, καθώς και το που πρέπει να στραφούν οι προσπάθειες βελτίωσης.

Κάθε διάγραμμα δράσης χωρίζεται σε τεταρτημόρια ανάλογα με την απόδοση και τη σημαντικότητα των κριτηρίων. Με τον τρόπο αυτό είναι δυνατός ο προσδιορισμός των απαιτούμενων ενεργειών για τη βελτίωση ή τη διατήρηση του επιπέδου ικανοποίησης των πελατών. Τα

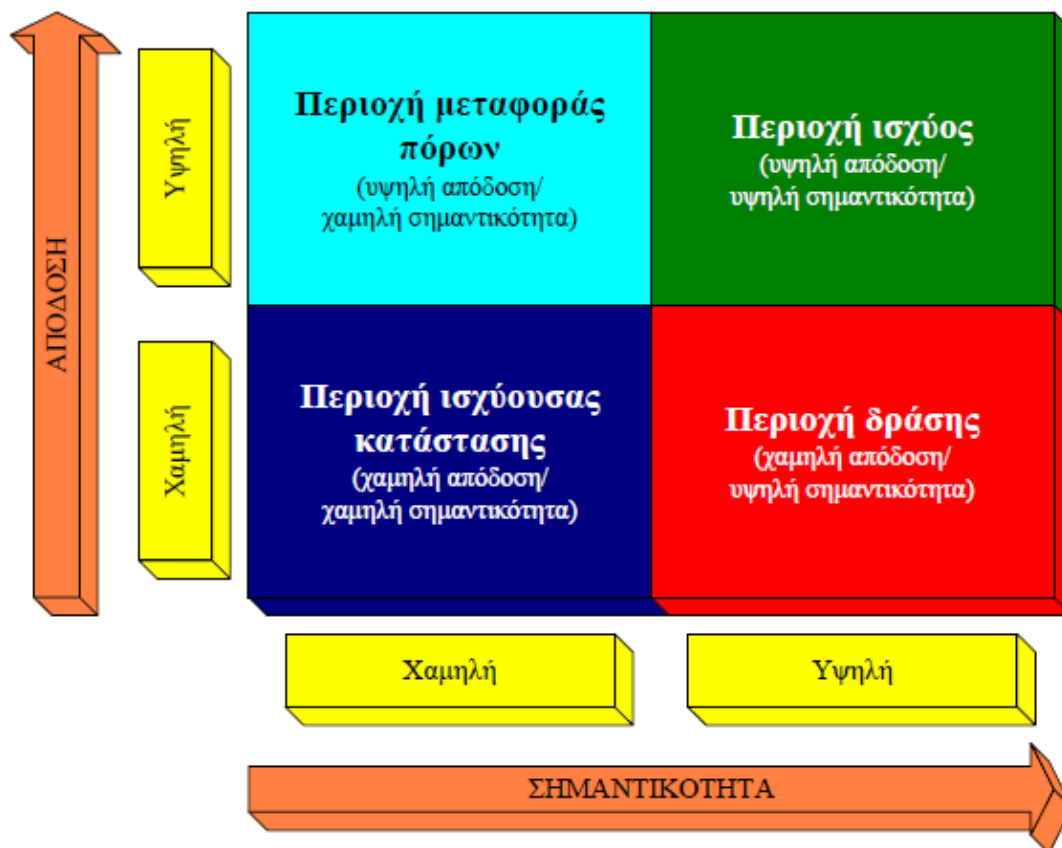
τεταρτημόρια αυτά με βάση την σειρά προτεραιότητας που πρέπει να δοθεί, είναι:

1. Περιοχή δράσης (χαμηλή απόδοση και υψηλή σημαντικότητα). Στο τεταρτημόριο αυτό ανήκουν τα πλέον κρίσιμα χαρακτηριστικά που πρέπει να βελτιωθούν οπωσδήποτε, ώστε να αυξηθεί το επίπεδο ικανοποίησης των τουριστών.

2. Περιοχή ισχύος (υψηλή απόδοση και υψηλή σημαντικότητα). Τα χαρακτηριστικά που ανήκουν σε αυτό το τεταρτημόριο είναι τα δυνατά σημεία και μπορούν να χρησιμοποιηθούν ως το συγκριτικό πλεονέκτημα σε σχέση με άλλα.

3. Περιοχή ισχύουσας κατάστασης (χαμηλή απόδοση και χαμηλή σημαντικότητα). Η περιοχή αυτή παρουσιάζει τα αδύνατα σημεία, δεδομένου ότι οι συγκεκριμένες διαστάσεις ικανοποίησης δεν θεωρούνται σημαντικές από τους πελάτες προς το παρόν, αλλά ενδέχεται να γίνουν σημαντικές στο μέλλον.

4. Περιοχή μεταφοράς πόρων (υψηλή απόδοση και χαμηλή σημαντικότητα). Η περιοχή αυτή θα πρέπει να είναι η τελευταία προτεραιότητα, διότι περιλαμβάνει χαρακτηριστικά τα οποία αφενός δεν είναι σημαντικά για τους πελάτες και αφετέρου η απόδοση είναι υψηλή.



Σχήμα 5.7: Διάγραμμα δράσης

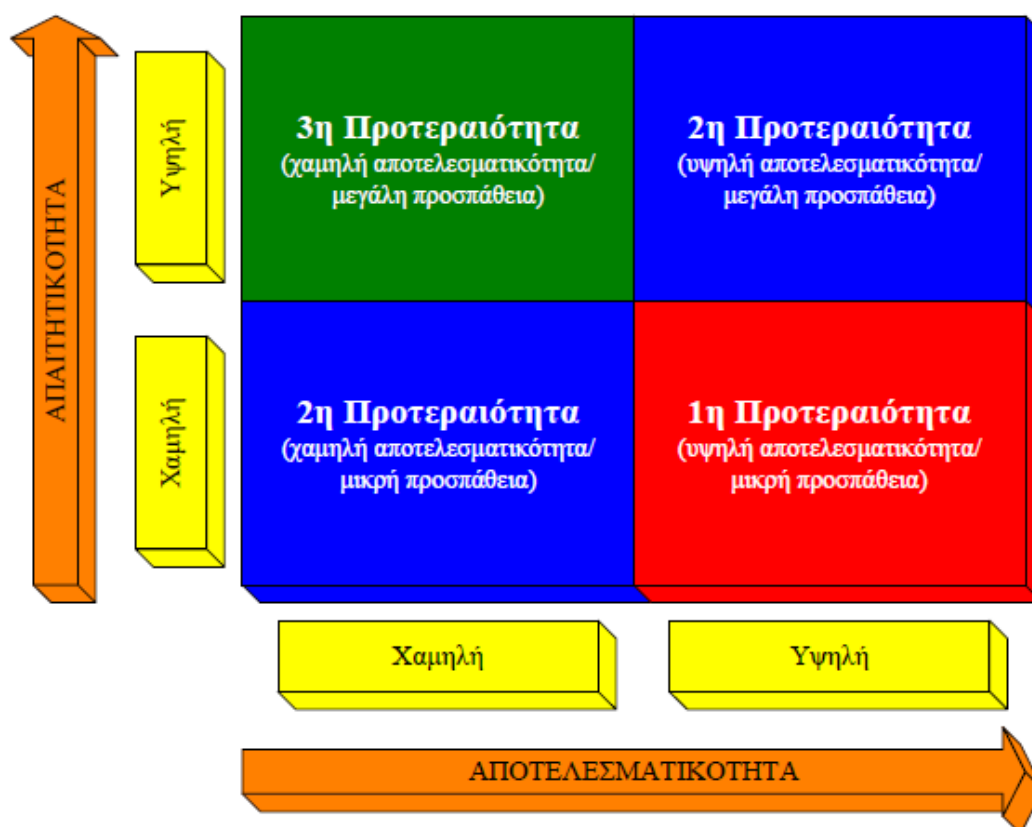
Το διάγραμμα βελτίωσης

Τα διαγράμματα δράσης μπορούν να υποδείξουν ποιες διαστάσεις ικανοποίησης πρέπει να βελτιωθούν, αλλά δεν είναι σε θέση να προσδιορίσουν ποιο είναι το αποτέλεσμα των ενεργειών βελτίωσης, ούτε το μέγεθος της προσπάθειας που χρειάζεται για να επιτευχθεί η προσδοκώμενη βελτίωση. Το πρόβλημα αυτό λύνεται με την κατασκευή διαγραμμάτων βελτίωσης. Τα διαγράμματα αυτά συνδυάζουν τους μέσους δείκτες απαιτητικότητας και αποτελεσματικότητας. Κάθε διάγραμμα βελτίωσης χωρίζεται σε τεταρτημόρια ανάλογα με την απαιτητικότητα και την αποτελεσματικότητα των διαστάσεων ικανοποίησης, με αποτέλεσμα τον προσδιορισμό των προτεραιοτήτων βελτίωσης.

Πρώτη προτεραιότητα αποτελούν οι διαστάσεις ικανοποίησης που έχουν μεγάλη αποτελεσματικότητα ενώ οι πελάτες δεν εμφανίζονται ιδιαίτερα απαιτητικοί.

Τη δεύτερη προτεραιότητα των ενεργειών βελτίωσης αποτελούν τα κριτήρια που είτε παρουσιάζουν μεγάλη αποτελεσματικότητα και μεγάλο βαθμό απαιτητικότητας, είτε εμφανίζουν μικρή αποτελεσματικότητα, ενώ οι πελάτες δεν φαίνονται ιδιαίτερα απαιτητικοί.

Τα χαρακτηριστικά που παρουσιάζουν μικρή αποτελεσματικότητα και μεγάλη απαιτητικότητα αποτελούν την τελευταία προτεραιότητα βελτίωσης.



Σχήμα 5.8: Διάγραμμα βελτίωσης

5.2.3 Αποτελέσματα

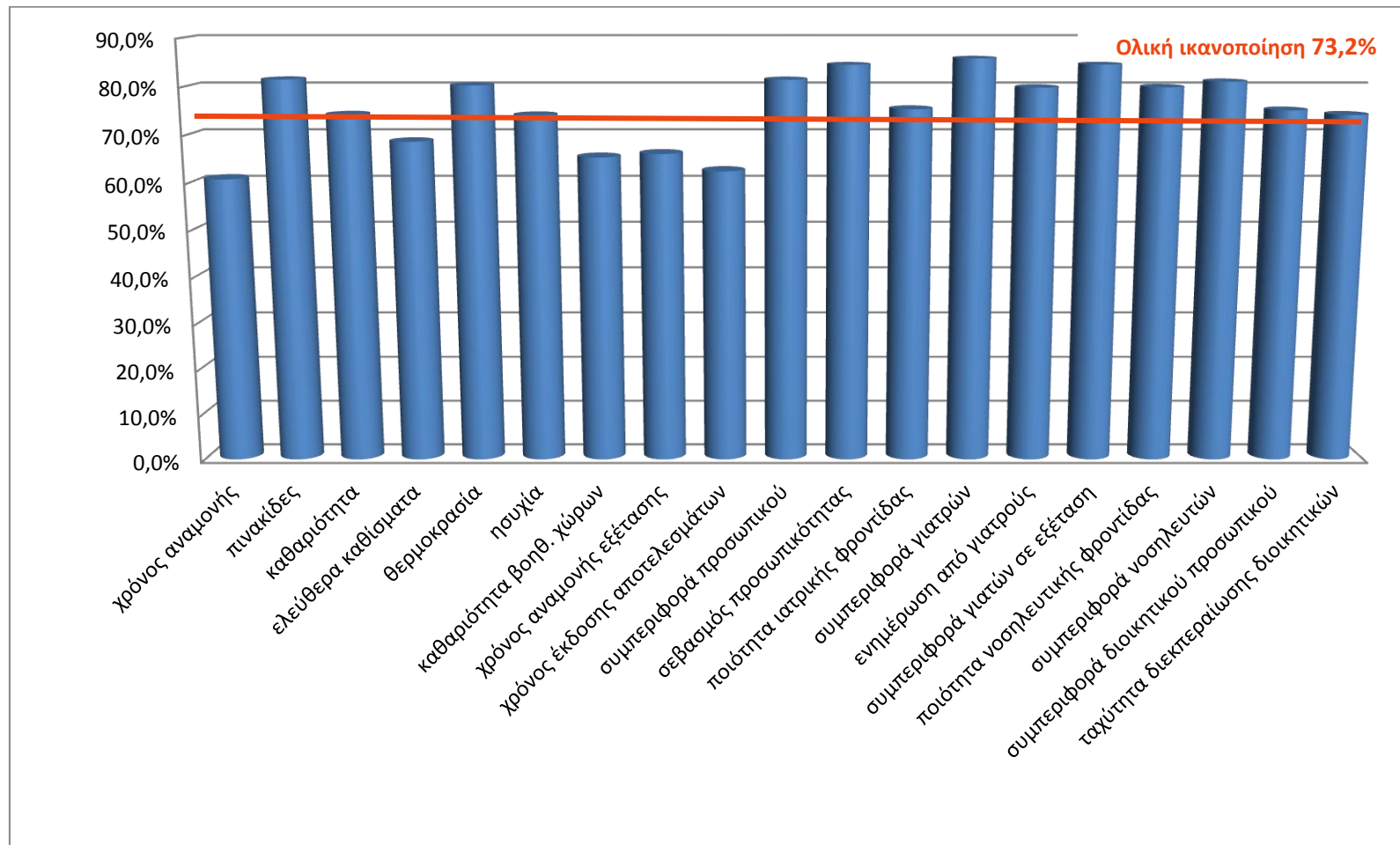
Ο νέος συμπληρωμένος πίνακας που προέκυψε εισήχθη στο πρόγραμμα MUSA και παρακάτω παρουσιάζονται τα βασικά αποτελέσματα ικανοποίησης των ασθενών του νοσοκομείου. Τα βασικά αποτελέσματα της μεθόδου που παρουσιάζονται επικεντρώνονται κυρίως στην εκτίμηση της σημαντικότητας των κριτηρίων ικανοποίησης,

τον υπολογισμό των μέσων δεικτών ικανοποίησης και τον καθορισμό των δυνατών και αδύνατων σημείων μέσω της ανάπτυξης σχετικών διαγραμμάτων δράσης και βελτίωσης.

Σε γενικές γραμμές φαίνεται ότι οι ασθενείς είναι σχετικά ικανοποιημένοι από τις υπηρεσίες που τους παρασχέθηκαν, αφού η μέση τιμή του δείκτη ολικής ικανοποίησης ανέρχεται σε 73,2%.

Όσον αφορά τα επιμέρους κριτήρια, φαίνεται ότι υπάρχουν τομείς που συγκεντρώνουν πολύ υψηλή ικανοποίηση και άλλοι οι οποίοι εμφανίζουν μέτρια ικανοποίηση με σημαντικά περιθώρια βελτίωσης. Συγκεκριμένα:

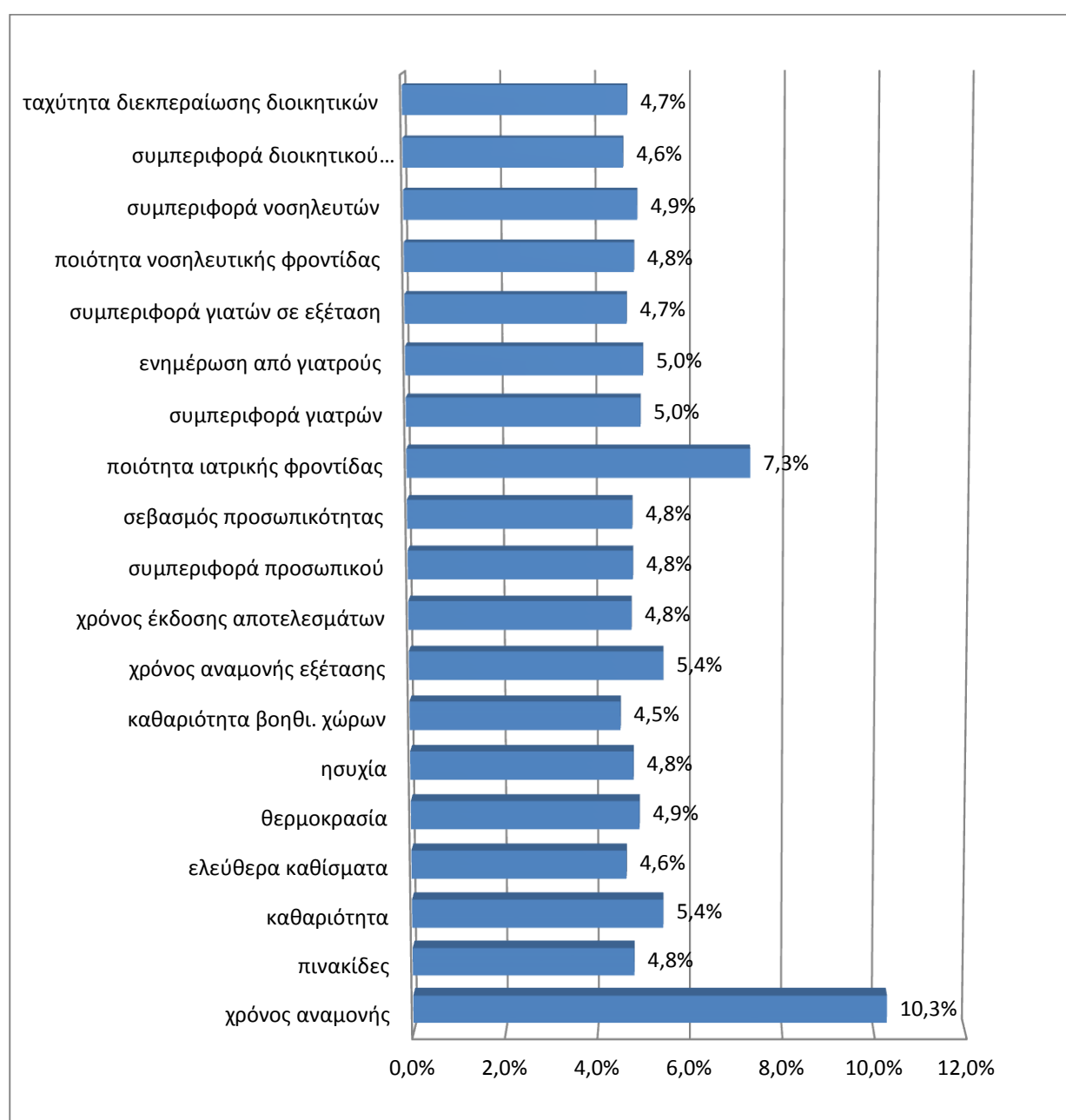
- Οι ασθενείς εμφανίζονται πολύ ικανοποιημένοι από τη συμπεριφορά του προσωπικού που σχετίζεται με τη νοσηλεία τους και το σεβασμό που δείχνουν στην προσωπικότητα τους, με μέσους δείκτες ικανοποίησης πάνω από 80%.
- Αντίθετα, οι ασθενείς χαρακτηρίζουν με μέτρια βαθμολογία τα κριτήρια που σχετίζονται γενικά με χρόνους αναμονής και καθαριότητας βοηθητικών χώρων με μέσους δείκτες ικανοποίησης κοντά στο 60-65%. Συνδυάζοντας τη βαρύτητα που δίνουν οι ασθενείς στους χρόνους αναμονής και τη χαμηλή ικανοποίηση σε αυτά τα κριτήρια, εύκολα διαπιστώνεται ότι απαιτούνται άμεσες ενέργειες για επίλυση των προβλημάτων ώστε να υπάρξει αύξηση των ποσοστών ικανοποίησης.
- Τα υπόλοιπα κριτήρια παρουσιάζουν ένα καλό επίπεδο ικανοποίησης (68-83%), κοντά στην τιμή του μέσου δείκτη ολικής ικανοποίησης.



Σχήμα 5.9: Δείκτες ικανοποίησης ασθενών ανά κριτήριο ικανοποίησης

Στο Παράρτημα IV παρουσιάζονται τα αποτελέσματα ικανοποίησης ανά κριτήριο και για την ερώτηση ολικής ικανοποίησης των ασθενών από τις υπηρεσίες του ΤΕΠ.

Όπως φαίνεται στο σχήμα 5.10, το σύνολο των ασθενών δίνει ιδιαίτερη βαρύτητα στα κριτήρια του χρόνου αναμονής και της ποιότητας ιατρικής φροντίδας, ενώ η σημαντικότητα των υπόλοιπων κριτηρίων είναι περίπου ίση και κυμαίνεται μεταξύ 4,5-5%.



Σχήμα 5.10: Βάρη κριτηρίων

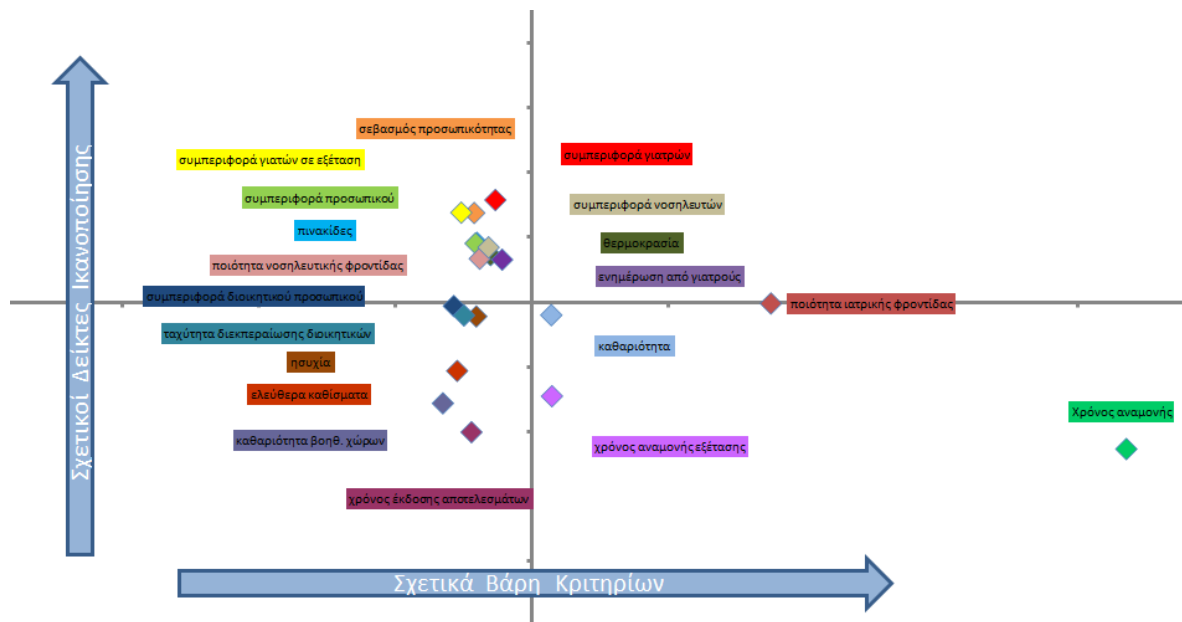
Στο Σχήμα 5.11 παρουσιάζεται το διάγραμμα δράσης για το σύνολο του δείγματος των ασθενών όπου μπορούν να σημειωθούν τα εξής:

- Το κριτήριο του χρόνου αναμονής είναι το σημείο στο οποίο το νοσοκομείο πρέπει να καταβάλει προσπάθειες για αύξηση της απόδοσης, δεδομένου ότι ανήκει στην περιοχή δράσης δηλαδή έχει χαμηλή απόδοση αλλά θεωρείται πολύ σημαντικό για τους ασθενείς. Στην περιοχή δράσης επίσης ανήκει οριακά και το κριτήριο ποιότητα ιατρικής φροντίδας όπου οι ασθενείς έχουν υψηλές απαιτήσεις αλλά ο μέσος δείκτης ικανοποίησης είναι οριακά κάτω από το 0 (-0,004).

- Δεν εντοπίζεται κάποιο κριτήριο στην περιοχή ισχύος (υψηλή απόδοση και υψηλή σημαντικότητα), γεγονός που σημαίνει ότι δεν υπάρχει κάποιο κρίσιμο κριτήριο το οποίο οι ασθενείς να αξιολογούν με υψηλή απόδοση και θα μπορούσε να χαρακτηριστεί ως ισχυρό πλεονέκτημα του νοσοκομείου.

- Στην περιοχή μεταφοράς πόρων (υψηλή απόδοση αλλά χαμηλή σημαντικότητα), ανήκουν τα περισσότερα κριτήρια (συμπεριφορά γιατρών, σεβασμός προσωπικότητας, συμπεριφορά γιατρών σε εξέταση, συμπεριφορά προσωπικού, πινακίδες, συμπεριφορά νοσηλευτών, ποιότητα ιατρικής φροντίδας, θερμοκρασία, ενημέρωση από γιατρούς) στα οποία όμως δεν θα πρέπει να επικεντρωθούν οι προσπάθειες βελτίωσης αφού δεν προσφέρουν σημαντικά στους ασθενείς.

- Τέλος τα κριτήρια που βρίσκονται στην περιοχή ισχύουσας κατάστασης (χρόνος έκδοσης αποτελεσμάτων, καθαριότητα βοηθητικών χώρων, ελεύθερα καθίσματα, ησυχία, ταχύτητα διεκπεραίωσης διοικητικών, συμπεριφορά διοικητικού προσωπικού), αποτελούν τα αδύνατα σημεία του νοσοκομείου, δεδομένου ότι οι έχουν χαμηλή απόδοση αλλά ταυτόχρονα δεν θεωρούνται σημαντικά από τους ασθενείς προς το παρόν.



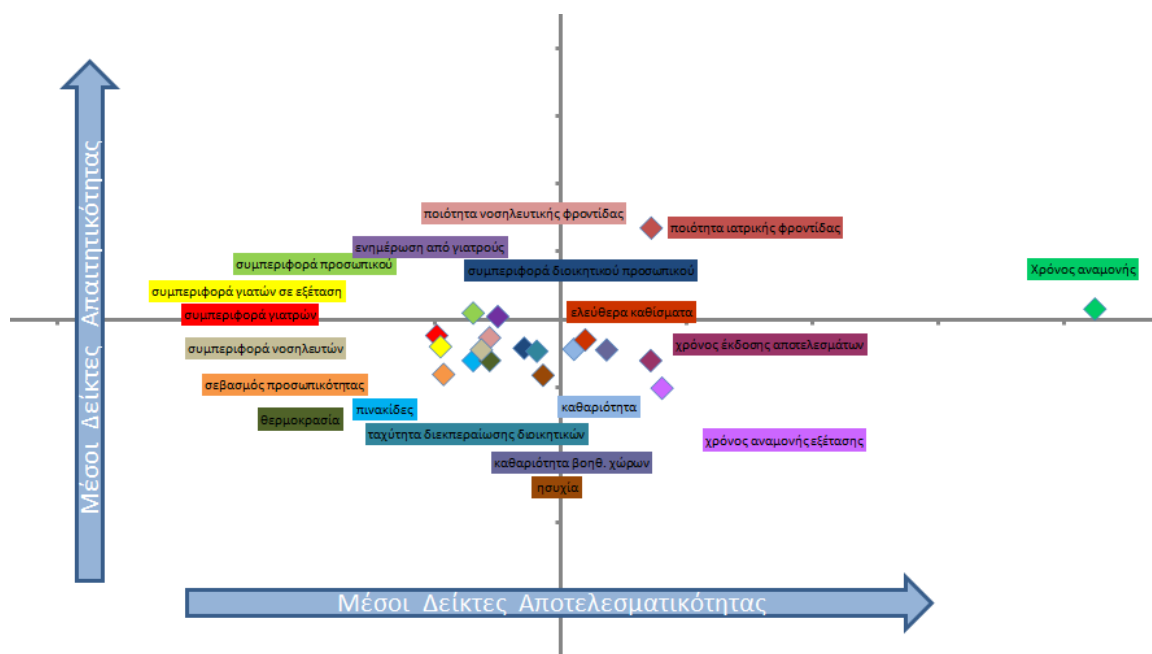
Σχήμα 5.11: Διάγραμμα δράσης

Στο σχήμα 5.12 παρουσιάζεται το διάγραμμα βελτίωσης για το σύνολο του δείγματος των ασθενών όπου μπορούν να σημειωθούν τα εξής:

- Πρώτη προτεραιότητα πρέπει να δοθεί στα κριτήρια χρόνου αναμονής εξέτασης, χρόνου έκδοσης αποτελεσμάτων, καθαριότητας βοηθητικών χώρων, ελεύθερων καθισμάτων και καθαριότητας γενικά. Η βελτίωση της ικανοποίησης των ασθενών στα κριτήρια αυτά αφενός απαιτεί μικρή προσπάθεια γιατί οι ασθενείς δεν είναι ιδιαίτερα απαιτητικοί, αφετέρου θα αυξήσει την ικανοποίηση σε μεγάλο βαθμό λόγω υψηλής αποτελεσματικότητας.
- Στις περιοχές δεύτερης προτεραιότητας βρίσκεται και το μεγαλύτερο πλήθος κριτηρίων. Σε αυτές τις περιοχές βρίσκονται τα κριτήρια που οι ασθενείς αφενός έχουν χαμηλές απαιτήσεις και η βελτίωση τους δεν θα έχει μεγάλο αποτέλεσμα, καθώς και αφετέρου τα κριτήρια που μπορούν να αυξήσουν σε μεγάλο βαθμό την ικανοποίηση των ασθενών αλλά απαιτούν και μεγάλη προσπάθεια λόγω υψηλών προσδοκιών από τους ασθενείς. Στην πρώτη περίπτωση βρίσκονται τα κριτήρια

συμπεριφοράς γιατρών, συμπεριφορά γιατρών σε εξέταση, σεβασμός προσωπικότητας, πινακίδες, συμπεριφορά νοσηλευτών, ποιότητα νοσηλευτικής φροντίδας, θερμοκρασία, συμπεριφορά διοικητικού προσωπικού, ταχύτητα διεκπεραίωσης διοικητικών και ησυχίας. Στην δεύτερη περίπτωση βρίσκονται τα κριτήρια χρόνου αναμονής και ποιότητας ιατρικής φροντίδας.

- Τέλος, τελευταία προτεραιότητα για το νοσοκομείο πρέπει να έχουν τα κριτήρια συμπεριφοράς προσωπικού και ενημέρωσης από γιατρούς, αφού η βελτίωση τους απαιτεί σχετικά υψηλή προσπάθεια, ενώ η αποτελεσματικότητα θα είναι σχετικά χαμηλή.



Σχήμα 5.12: Διάγραμμα βελτίωσης

5.2.4 Συμπεράσματα

Τα αποτελέσματα της μεθόδου MUSA φανερώνουν ότι σε γενικές γραμμές οι ασθενείς είναι σχετικά ικανοποιημένοι από τις υπηρεσίες που τους παρασχέθηκαν με μέση τιμή του δείκτη ολικής ικανοποίησης 73,2%. Όσον αφορά τα επιμέρους κριτήρια, οι ασθενείς εμφανίζονται πολύ ικανοποιημένοι από τη συμπεριφορά του προσωπικού που σχετίζεται με τη νοσηλεία τους και το σεβασμό που δείχνουν στην

προσωπικότητα τους, ενώ αντίθετα χαρακτηρίζουν με μέτρια βαθμολογία τα κριτήρια που σχετίζονται γενικά με χρόνους αναμονής και καθαριότητας βοηθητικών χώρων. Όλα τα υπόλοιπα κριτήρια παρουσιάζουν ένα καλό επίπεδο ικανοποίησης κοντά στην τιμή του μέσου δείκτη ολικής ικανοποίησης.

Το σύνολο των ασθενών δίνει ιδιαίτερη βαρύτητα στα κριτήρια του χρόνου αναμονής και της ποιότητας ιατρικής φροντίδας και έχει αυξημένες απαιτήσεις για την ικανοποίηση του σε αυτούς τους τομείς. Συνδυάζοντας τη βαρύτητα που δίνουν οι ασθενείς σε αυτά τα κριτήρια και τη χαμηλή ικανοποίηση, συνεπάγεται ότι απαιτούνται άμεσες ενέργειες για επίλυση των προβλημάτων ώστε να υπάρξει αύξηση των δεικτών ικανοποίησης.

Από το διάγραμμα δράσης για το σύνολο του δείγματος των ασθενών εύκολα διαπιστώνεται αυτό, αφού τα κριτήρια του χρόνου αναμονής και ποιότητα ιατρικής φροντίδας βρίσκονται στην περιοχή δράσης και απαιτείται προσοχή από τη διοίκηση του νοσοκομείου αφού είναι πολύ σημαντικά για τους ασθενείς. Επίσης από το διάγραμμα δράσης φαίνεται ότι υπάρχουν αρκετά κριτήρια (χρόνος έκδοσης αποτελεσμάτων, καθαριότητα βοηθητικών χώρων, ελεύθερα καθίσματα, ησυχία, ταχύτητα διεκπεραίωσης διοικητικών, συμπεριφορά διοικητικού προσωπικού) στην περιοχή ισχύουσας κατάστασης και αποτελούν τα αδύνατα σημεία του νοσοκομείου με χαμηλή ικανοποίηση, αλλά ενδέχεται να γίνουν σημαντικά στο μέλλον και δεν πρέπει να παραβλέπονται.

Τέλος από το διάγραμμα βελτίωσης παρατηρείται ότι η συνολική ικανοποίηση των ασθενών μπορεί να αυξηθεί με ευκολία εάν η διοίκηση δώσει βαρύτητα στις ενέργειες βελτίωσης των κριτηρίων χρόνου αναμονής εξέτασης, χρόνου έκδοσης αποτελεσμάτων, καθαριότητας βοηθητικών χώρων, ελεύθερων καθισμάτων και καθαριότητας αφού οι ασθενείς δεν είναι ιδιαίτερα απαιτητικοί σε αυτά αλλά τα αποτελέσματα θα έχουν υψηλή απόδοση στην ικανοποίηση.

Επίσης φαίνεται ότι τα κριτήρια χρόνου αναμονής και ποιότητας ιατρικής φροντίδας μπορούν να αυξήσουν σε μεγάλο βαθμό την

ικανοποίηση των ασθενών αλλά απαιτούν και μεγάλη προσπάθεια λόγω υψηλών προσδοκιών από τους ασθενείς, ενώ στα κριτήρια συμπεριφοράς γιατρών, συμπεριφορά γιατρών σε εξέταση, σεβασμός προσωπικότητας, πινακίδες, συμπεριφορά νοσηλευτών, ποιότητα νοσηλευτικής φροντίδας, θερμοκρασία, συμπεριφορά διοικητικού προσωπικού, ταχύτητα διεκπεραίωσης διοικητικών και ησυχίας οι ασθενείς έχουν χαμηλές απαιτήσεις και η βελτίωση τους δεν θα αποδώσει ιδιαίτερα. Τελευταία προτεραιότητα για το νοσοκομείο πρέπει να έχουν τα κριτήρια συμπεριφοράς προσωπικού και ενημέρωσης από γιατρούς, αφού η βελτίωση τους απαιτεί σχετικά υψηλή προσπάθεια αλλά δεν θα αποδώσει και ανάλογα στην ικανοποίηση των ασθενών.

6. Επίλογος

6.1 Σύνοψη

Η έλλειψη δεδομένων σε έρευνες ικανοποίησης πελατών είναι ένα σύνηθες πρόβλημα το οποίο απαντάται γενικότερα στις στατιστικές έρευνες και έχει ονομαστεί με τον όρο missing data. Το πρόβλημα αυτό μειώνει τη δυνατότητα των αναλυτών να εξάγουν ασφαλή συμπεράσματα για το σύνολο του πληθυσμού, γιατί πολλές φορές αναγκάζονται να μειώσουν σημαντικά το δείγμα ώστε να σχηματιστεί μια πλήρης συλλογή δεδομένων.

Η μέθοδος hot deck imputation είναι μια μέθοδος καταλογισμού, δηλαδή απόδοσης τιμών για προβλήματα ελλιπών δεδομένων. Το βασικό πλεονεκτήματα που προσφέρει η μέθοδος είναι ότι η τιμή που τοποθετείται στο κενό είναι πραγματική και με αυτό τον τρόπο μειώνεται το πρόβλημα υποεκτίμησης της διασποράς άλλων μεθόδων. Παρόλα αυτά, επειδή δεν υπάρχει ένας αντικειμενικός τρόπος υπολογισμού της βέλτιστης περίπτωσης από την οποία θα γίνει ο καταλογισμός, απαιτείται μεγάλη προσοχή από τον ερευνητή κατά τον καθορισμό των κριτηρίων για την εύρεση των όμοιων περιπτώσεων.

Ένας απλός τρόπος για την εύρεση όμοιων ερωτηματολογίων στη μέθοδο hot deck, είναι η απόσταση των δεδομένων του συνόλου η οποία σε έρευνες αριθμητικών δεδομένων έχει πλήθος πιθανών τρόπων υπολογισμού. Σε ordinal δεδομένα όμως, που απαντώνται πολύ συχνά στις έρευνες ικανοποίησης πελατών, δεν υπάρχει κάποιος ευρέως αποδεκτός τρόπος υπολογισμού αποστάσεων και ο κάθε ερευνητής μπορεί να εφαρμόσει ένα δικό του κατά την αναζήτηση των όμοιων ερωτηματολογίων.

Στην παρούσα εργασία όπου αναπτύχθηκε αλγόριθμος υλοποίησης της μεθόδου hot deck imputation για έρευνες ικανοποίησης με ordinal δεδομένα, χρησιμοποιήθηκε η μέθοδος υπολογισμού Walesiak (1993). Η εν λόγω μέθοδος προϋποθέτει την ύπαρξη πλήρων συλλογών δεδομένων για τη χρήση της, γεγονός που

από τη φύση του προβλήματος μας αποτρέπει να τη χρησιμοποιήσουμε ως έχει. Τα ελλιπή στοιχεία ως μέρος του προβλήματος δεν είναι δυνατό να αγνοηθούν αφού η αναζήτηση όμοιων περιπτώσεων απαιτείται να γίνει μεταξύ περιπτώσεων με πλήρη και ελλιπή δεδομένα.

Για αυτό το λόγο η μέθοδος Walesiak τροποποιήθηκε ώστε ο υπολογισμός αποστάσεων σε κάθε ζεύγος περιπτώσεων να γίνεται μόνο μεταξύ των μεταβλητών με πλήρη στοιχεία. Αυτός ο τρόπος υπολογισμού όμως δημιουργεί πρόβλημα αξιοπιστίας του αποτελέσματος. Αφαιρώντας διαφορετικό πλήθος δεδομένων από τον υπολογισμό της απόστασης, σημαίνει ότι ο τελικός πίνακας αποστάσεων έχει υπολογιστεί με διαφορετικές συνθήκες κάθε φορά.

Έτσι έγινε επιτακτική η εφαρμογή ενός δείκτη όπου θα δείχνει την αξιοπιστία υπολογισμού της απόστασης σε κάθε ζεύγος περιπτώσεων. Χρησιμοποιήθηκε για αυτό το λόγο ο δείκτης “συντελεστής Καταλογισμού” ο οποίος φανερώνει ουσιαστικά το πλήθος των μεταβλητών που χρησιμοποιήθηκαν στον υπολογισμό της απόστασης.

Με χρήση του παραπάνω δείκτη, ο πίνακας των αποστάσεων “συνοδεύεται” από τον πίνακα των συντελεστών Καταλογισμού και δίνεται η δυνατότητα στον ερευνητή να περιορίσει στο εύρος που θέλει την επιλογή περιπτώσεων-δοτών για imputation. Αυτό επιτυγχάνεται μέσω επιλογής για imputation όχι μόνο μέσω της ελάχιστης απόστασης δύο περιπτώσεων, αλλά μέσω της ελάχιστης απόστασης μεταξύ των περιπτώσεων που έχουν τιμή συντελεστή Καταλογισμού μεγαλύτερη από κάποια προκαθορισμένη τιμή που εισάγει ο χρήστης.

Ο αλγόριθμος ολοκληρώθηκε με προσθήκη της ερώτησης ολικής ικανοποίησης που απαντάται πολύ συχνά στις έρευνες ικανοποίησης πελατών. Αποφασίστηκε λοιπόν, ο υπολογισμός αποστάσεων με τη μέθοδο Walesiak να γίνεται με τρόπο ώστε η ερώτηση ολικής ικανοποίησης να έχει κάποια βαρύτητα, ως ερώτηση που απεικονίζει καλύτερα την ικανοποίηση του πελάτη.

Στη συνέχεια ο αλγόριθμος προγραμματίστηκε σε γλώσσα Java όπου αναπτύχθηκε γραφικό περιβάλλον το οποίο προσφέρει απλότητα και η ευκολία χρήσης από οποιονδήποτε χρήστη έχει τις βασικές

γνώσεις της μεθόδου hot deck. Χρησιμοποιώντας μια απλή και ευρέως αναγνωρίσιμη μορφή προγράμματος το πρόγραμμα περιλαμβάνει κενά πεδία για την εισαγωγή των απαραίτητων δεδομένων και κουμπιά για την εκτέλεση των εντολών και την αναζήτηση των αρχείων.

Στη συνέχεια εκτελέστηκε έλεγχος ορθής λειτουργίας και σωστής απόδοσης των αποτελεσμάτων από τον αλγόριθμο και έλεγχος σωστής λειτουργία της φόρμας εκτέλεσης του προγράμματος. Οι ανωτέρω έλεγχοι έγιναν με εφαρμογή του προγράμματος σε πλήρη πίνακα δεδομένων στον οποίο δημιουργήθηκαν τεχνητά ελλείψεις, προσομοιάζοντας με αυτό τον τρόπο μια έρευνα με κενά στις απαντήσεις ικανοποίησης των πελατών. Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από έρευνα ικανοποίησης τουριστών και η διαδικασία επαναλήφθηκε 10 φορές ώστε τα αποτελέσματα να συγκεντρωθούν σε 10 διαφορετικά confusion tables.

Από τα αποτελέσματα φαίνεται ότι ο αλγόριθμος απέδωσε πολύ καλά στο σύνολο των δεδομένων, με κάποιες αποκλίσεις στον καταλογισμό των χαμηλών τιμών. Η εν λόγω απόκλιση μπορεί να εξηγηθεί από την ανισότητα στην κατανομή του πλήθους των τιμών στα αρχικά δεδομένα και συγκεκριμένα από τα πολύ μικρά συμμετοχής των χαμηλών τιμών.

Αφού πιστοποιήθηκε η ορθή λειτουργία του αλγόριθμου, αποφασίστηκε η εκτέλεση του σε δεδομένα πραγματικής έρευνας που αφορά την ικανοποίηση των ασθενών για τις υπηρεσίες που παρέχονται από το ΤΕΠ νοσοκομείου στην Αττική. Στη συγκεκριμένη έρευνα υπάρχουν πολλά ελλιπή στοιχεία τα οποία καθιστούν την στατιστική επεξεργασία επισφαλή και για αυτό το λόγο απαιτείται η εφαρμογή της μεθόδου hot deck ώστε να συμπληρωθούν όσο το δυνατόν περισσότερα κενά και να αξιοποιηθούν όσο το δυνατόν περισσότερα ερωτηματολόγια. Τα δεδομένα της έρευνας εισήχθησαν στο πρόγραμμα hot deck που αναπτύξαμε, αφού πρώτα αφαιρέθηκαν τα αναξιόπιστα ερωτηματολόγια με μεγάλο αριθμό κενών. Έτσι κατασκευάστηκε πλήρης πίνακας δεδομένων που αποτελείται από 246 ερωτηματολόγια

και τα οποία είναι δυνατόν να αναλυθούν με οποιαδήποτε στατιστική μέθοδο.

Τελικά, για την ανάλυση και την εξαγωγή συμπερασμάτων από τον πλήρη πίνακα δεδομένων που προέκυψε, αποφασίστηκε η εφαρμογή της μεθόδου MUSA (MUlticriteria Satisfaction Analysis) η οποία βασίζεται στην πολυκριτήρια ανάλυση αποφάσεων και έχει τη δυνατότητα να χειριστεί ποιοτικά δεδομένα, παράγοντας ένα σύνολο ποσοτικών αποτελεσμάτων, που μπορούν να αναλύσουν σε βάθος την ικανοποίηση ενός συνόλου πελατών.

Τα αποτελέσματα που συνοψίζουν ολοκληρωμένα την ικανοποίηση των ασθενών και παρουσιάστηκαν στην παρούσα εργασία αφορούν τα βάρη των κριτηρίων, τους δείκτες ικανοποίησης ανά κριτήριο και τα διαγράμματα δράσης και βελτίωσης. Η χρήση των αποτελεσμάτων της μεθόδου MUSA δίνει την δυνατότητα στη διοίκηση να κατανοήσει τα πλεονεκτήματα του νοσοκομείου ώστε να διατηρηθεί η καλή εικόνα που σε γενικές γραμμές έχουν οι ασθενείς αλλά και να θέσει με συγκεκριμένες προτεραιότητες στόχους για βελτίωση των υποτομέων που μειώνουν το επίπεδο ικανοποίησης.

6.2 Μελλοντικές επεκτάσεις

Ο αλγόριθμος Hot Deck για έρευνες ικανοποίησης με ordinal δεδομένα που αναπτύχθηκε στα πλαίσια της εργασίας, αποδείχτηκε ότι λειτουργεί αρκετά ικανοποιητικά. Παρόλα αυτά επειδή κατά την ανάπτυξη του αλγόριθμου και την εφαρμογή της μεθόδου Walesiak σε ελλιπή δεδομένα χρησιμοποιήθηκαν κάποιες παραδοχές, έχει ενδιαφέρον η τροποποίηση τμημάτων του αλγόριθμου και η σύγκριση των αποτελεσμάτων μεταξύ τους.

Αρχικά προτείνεται η εφαρμογή διαφορετικών τρόπων αξιοποίησης από τον αλγόριθμο, των εισαχθέντων τιμών στον αρχικό πίνακα, όπως έχουν αναφερθεί στην παράγραφο 4.7 και η αναζήτηση πιθανώς καλύτερων αποτελεσμάτων με κάποιο από αυτούς. Στην συγκεκριμένη εργασία οι νεοτοποθετημένες τιμές του αρχικού πίνακα

δεδομένων χρησιμοποιήθηκαν ως γνήσιες, δίνοντας έτσι την πιθανότητα να χρησιμοποιηθούν ως δότες για imputation σε μεταγενέστερο κενό. Βέβαια εν μέρη, υπάρχει η προστασία των πινάκων αποστάσεων και συντελεστών καταλογισμού όπου παραμένουν αμετάβλητοι και ουσιαστικά μειώνουν την πιθανότητα μια τοποθετημένη τιμή να λειτουργήσει ως δότης, χωρίς αυτό να είναι απαγορευτικό.

Είναι δυνατή όμως και η εξέταση μιας άλλης εκδοχής, όπου μετά από κάθε επιτυχημένη εκτέλεση imputation, υπολογίζονται εκ νέου οι δύο πίνακες (απόσταση Walesiak και συντελεστή Καταλογισμού) θεωρώντας ότι οι νέες τιμές προϋπήρχαν στον αρχικό πίνακα δεδομένων. Σε αυτή την περίπτωση ο αλγόριθμος εκτελείται ουσιαστικά από το πρώτο του βήμα, τόσες φορές όσα είναι και τα ελλιπή στοιχεία του αρχικού πίνακα. Εδώ υπάρχουν οι κίνδυνοι, αφενός μεγάλου χρόνου ολοκλήρωσης ιδιαίτερα σε μεγάλες έρευνες με πολλά ελλιπή στοιχεία και αφετέρου η συνεχής μεταφορά λανθασμένων τιμών αφού μια αποτυχημένη απόδοση τιμής είναι δυνατό να μεταφερθεί σε επόμενο ελλιπές στοιχείο όταν χρησιμοποιηθεί ως δότης.

Η άλλη εκδοχή του αλγόριθμου που μπορεί να εξεταστεί, ουσιαστικά αγνοεί τις νέες τιμές που έχουν τοποθετηθεί στον πίνακα και συμπεριφέρεται σαν να μην έχει γίνει αντικατάσταση των κενών. Σε αυτή την περίπτωση οι πίνακες απόστασης Walesiak και συντελεστή Καταλογισμού υπολογίζονται μόνο στην αρχή και η θέση κάθε ελλιπούς στοιχείου αποθηκεύεται με ένα δείκτη. Στη συνέχεια κατά την αναζήτηση πιθανών δοτών για imputation, ο δείκτης αυτός χρησιμοποιείται για να αποκλειστούν οι θέσεις όπου υπήρχε αρχικά κενό, ανεξάρτητα από το εάν τη συγκεκριμένη στιγμή έχει αποδοθεί κάποια τιμή σε αυτές.

Σε μελλοντική επέκταση, θα μπορούσε επίσης να εξεταστεί η επιρροή του συντελεστή καταλογισμού στα αποτελέσματα του αλγόριθμου, δηλαδή να γίνει μια ανάλυση ευαισθησίας. Η εν λόγω ανάλυση γίνεται με χρήση πλήρως συμπληρωμένου πίνακα δεδομένων, όπου δημιουργούνται τεχνητά ελλιπή στοιχεία και εξετάζονται τα

αποτελέσματα του αλγόριθμου με χρήση διαφορετικών τιμών συντελεστή καταλογισμού. Προφανώς όσο αυξάνεται η ελάχιστη επιτρεπτή τιμή του συντελεστή καταλογισμού τόσο μειώνεται το πιθανό δείγμα του πίνακα, αφού απορρίπτονται τα ερωτηματολόγια με πολλά ελλιπή στοιχεία. Έτσι, είναι πιθανό μέσω ανάλυσης των αποτελεσμάτων να οριστεί τρόπος αναζήτησης του βέλτιστου μεγέθους δείγματος με την βέλτιστη τιμή συντελεστή καταλογισμού, ώστε να χρησιμοποιούνται όσο το δυνατόν περισσότερα ερωτηματολόγια με τις μικρότερες αποκλίσεις κατά το imputation.

Παράρτημα Ι “ ΒΙΒΛΙΟΓΡΑΦΙΑ ”

- Allison P. D. (2002) Missing Data, Series: Quantitative Applications in social Sciences, Sage Publications Inc. California.
- Churchill G. A. Jr (1988). Marketing research: Methodological foundations, The Dryden Press
- Churchill G. A. Jr and C. Surprenant (1982). An investigation into the determinants of customer satisfaction, Journal of Marketing Research.
- Czarnecki M. T. (1999). Managing by measuring: How to improve your organization's performance through effective benchmarking, AMACOM, New York.
- Day R. L. (1977). Toward a process model of customer satisfaction in: H. K. Hunt (Ed.), Conceptualization and measurement of customer satisfaction and dissatisfaction, Marketing Science Institute, Cambridge. MA.
- Dempster, A. P., Laird, N., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), Journal of the Royal Statistical Society, Ser. B.
- Dutka A. (1995). AMA handbook of customer satisfaction: A guide to research, planning and implementation, NTC Publishing Group, Illinois.
- Engel J. F. and R. D. Blackwell (1982). Consumer behavior, Holt, Rinehart and Winston, New York.
- Gerson R. F. (1993). Measuring customer satisfaction: A guide to managing quality service, Crisp Publications, Menlo Park.
- Graham J. W. & Schafer J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In Statistical Strategies for small- sample research.
- Graham J. W. & Schafer J. L. (2002). Missing Data: Our View of the State of the Art. American Psychological Association.

- Grigoroudis E., Kiriazopoulos P., Siskos G., Spiridatos A., Giannakopoulos D. (2007). Tracking changes of e-customer preferences using multicriteria analysis, *Managing Service Quality*.
- Hill, N. (1996). *Handbook of customer satisfaction measurement* Gower Publishing, Hampshire.
- Howard J. A. and J. Sheth (1969). *The theory of buyer behavior*, John Wiley and Sons, New York.
- Hunt H. K. (1977). Customer satisfaction/ dissatisfaction: Overview and future research directions, in: H. K. Hunt (Ed.), *Conceptualization and measurement of consumer satisfaction and dissatisfaction*, Marketing Science Institute, Cambridge. MA.
- Kim J. O. & Curry J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research*.
- Little R. J. & Rubin D. B. (2001), *Statistical Data Missing*, International Encyclopedia of Social and Behavioral Sciences, Elsevier Science Ltd.
- Nauman E. and K. Giel (1995). *Customer satisfaction measurement and management: Using the voice of the customer*, Thomson Executive Press, Cincinnati.
- Oliver R. L. (1977). Effect of expectation and disconfirmation on post exposure product evaluations: An alternative interpretation, *Journal of Applied Psychology*, (4).
- Oliver R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions, *Journal of Marketing Research*.
- Oliver R. L. (1996). *Satisfaction: A behavioral perspective on the customer*, McGraw-Hill, New York.
- Reilly M. and Pepe M. (1997), The relationship between hot-deck multiple imputation and weighted likelihood, *Statistics in Medicine*.
- Rubin D. B & Little R. J.(1987), *Statistical Analysis with missing data*, New York, John Wiley & sons Ltd.
- Rubin D. B & Little R. J.(2002), *Statistical Analysis with missing data*, second edition, New York, John Wiley & sons Ltd.
- Rubin D.B. (1976), *Inference and missing data*, *Biometrika*.

- Rubin D.B. (1977), Formalizing notions about the effect of nonrespondents in sample surveys, Journal of the American Statistical Association.
- Rubin D.B. (1987), Multiple Imputation for nonresponse in surveys, Wiley series in Probability and Mathematical Statistics
- Rubin D.B. (1996), Multiple imputation after 18+ years, Journal of the American Statistical Association.
- Rubin D.B. and Schenker N. (1986), Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, Journal of the American Statistical Association.
- Rubin, D.B. (1972). A non iterative algorithm for least squares estimation of missing values in any analysis of variance design. Applied Statistics.
- Schafer J. L. (1997). Analysis of incomplete multivariate data. New York: Chapman & Hall.
- Schafer J. L. & Graham J. W., (2002), Missing Data: Our View of the State of the Art, American Psychological Association Inc.
- Schafer J. L., & Olsen M. K. (1998), Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research.
- Tse D. K. and P. C. Wilton (1988). Models of consume satisfaction: An extension, Journal of Marketing Research.
- Walesiak Marek (1999) Distance Measure For Ordinal Data.
- Γρηγορούδης Ε. και Σίσκος Γ. (2000). Ποιότητα Υπηρεσιών και Μέτρηση Ικανοποίησης του Πελάτη, Το Σύστημα MUSA, Εκδόσεις Νέων Τεχνολογιών, Αθήνα.
- Γρηγορούδης Ε., Ι. Μaltse wiltonανδράκης, Ι. Πολίτης και Ι. Σίσκος (1998). Μέτρηση της ικανοποίησης πελατών στις ακτοπλοϊκές συγκοινωνίες, Πρακτικά 12ου Εθνικού Συνεδρίου Ελληνικής Εταιρίας Επιχειρησιακών Ερευνών.
- Σίσκος Ι., (1981). Μεθοδολογία ποιοτικού marketing: Η ανάλυση αγοράς, ΣΠΟΥΔΑΙ, τόμος ΛΔ', Α.Β.Σ.Π., Πειραιάς.

Σίσκος Ι., (1986). Πολυκριτήρια ανάλυση, Εγκυκλοπαίδεια Πληροφορικής & Τεχνολογίας Υπολογιστών, Εκδόσεις Νέων Τεχνολογιών, Αθήνα.

Παράρτημα II

“Confusion Tables τεκμηρίωσης αλγόριθμου Hot Deck”

Confusion Table 1							
imputed						missing data per value	
	1	2	3	4	5		
values	1	0,00	0,25	0,25	0,50	0,00	4
	2	0,00	0,00	0,25	0,75	0,00	4
	3	0,00	0,00	0,30	0,51	0,19	37
	4	0,00	0,01	0,18	0,63	0,18	89
	5	0,00	0,00	0,02	0,32	0,67	66
							200

Confusion Table 2							
imputed						missing data per value	
	1	2	3	4	5		
values	1	0,00	0,00	0,00	1,00	0,00	2
	2	0,00	0,00	0,80	0,20	0,00	5
	3	0,00	0,06	0,36	0,42	0,15	33
	4	0,00	0,04	0,20	0,61	0,15	105
	5	0,00	0,00	0,07	0,44	0,49	55
							200

Confusion Table 3							
imputed						missing data per value	
	1	2	3	4	5		
values	1	0,00	0,00	1,00	0,00	0,00	4
	2	0,00	0,14	0,43	0,29	0,14	7
	3	0,00	0,06	0,36	0,52	0,06	33
	4	0,02	0,04	0,18	0,49	0,26	92
	5	0,00	0,00	0,05	0,34	0,61	64
							200

Confusion Table 4							
imputed						missing data per value	
	1	2	3	4	5		
values	1	0,00	0,00	0,67	0,33	0,00	3
	2	0,20	0,00	0,20	0,60	0,00	5
	3	0,00	0,04	0,21	0,61	0,14	28
	4	0,02	0,03	0,11	0,66	0,19	102
	5	0,03	0,02	0,06	0,50	0,39	62
							200

Confusion Table 5					
values	imputed				
	1	2	3	4	5
	1	n/a	n/a	n/a	n/a
	2	0,11	0,22	0,33	0,00
	3	0,00	0,00	0,16	0,74
	4	0,01	0,02	0,17	0,64
	5	0,00	0,00	0,06	0,37
					0
					9
					19
					109
					63
					200

Confusion Table 6					
values	imputed				
	1	2	3	4	5
	1	0,00	0,00	0,00	1,00
	2	0,00	0,00	0,67	0,33
	3	0,04	0,11	0,33	0,44
	4	0,01	0,00	0,13	0,67
	5	0,00	0,00	0,06	0,27
					0,00
					3
					27
					102
					67
					200

Confusion Table 7					
values	imputed				
	1	2	3	4	5
	1	0,50	0,00	0,50	0,00
	2	0,00	0,33	0,67	0,00
	3	0,00	0,02	0,40	0,52
	4	0,01	0,02	0,18	0,57
	5	0,00	0,00	0,04	0,35
					0,00
					2
					3
					42
					101
					52
					200

Confusion Table 8					
values	imputed				
	1	2	3	4	5
	1	0,00	0,00	0,00	1,00
	2	0,00	0,00	0,40	0,30
	3	0,00	0,00	0,23	0,47
	4	0,00	0,00	0,26	0,52
	5	0,02	0,02	0,06	0,44
					0,00
					1
					10
					30
					97
					62
					200

Confusion Table 9						
values	imputed					missing data per value
	1	2	3	4	5	
	1	0,17	0,17	0,50	0,00	0,17
	2	0,00	0,00	0,33	0,67	0,00
	3	0,04	0,04	0,48	0,39	0,04
	4	0,00	0,02	0,21	0,62	0,15
	5	0,00	0,00	0,02	0,29	0,69
						200

Confusion Table 10						
values	imputed					missing data per value
	1	2	3	4	5	
	1	0,00	0,00	0,00	0,50	0,50
	2	0,00	0,17	0,50	0,33	0,00
	3	0,07	0,07	0,33	0,50	0,03
	4	0,01	0,01	0,19	0,61	0,18
	5	0,00	0,02	0,00	0,29	0,69
						200

Παράρτημα ΙΙΙ
“Ερωτηματολόγιο Έρευνας Νοσοκομείου”

**ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ & ΚΟΙΝΩΝΙΚΗΣ ΑΛΛΗΛΕΓΓΥΗΣ
1^η ΥΓΕΙΟΝΟΜΙΚΗ ΠΕΡΙΦΕΡΕΙΑ ΑΤΤΙΚΗΣ
ΚΩΝΣΤΑΝΤΟΠΟΥΛΕΙΟ Γ.Ν.Ν.Ι.**

**ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ ΙΚΑΝΟΠΟΙΗΣΗΣ ΕΞΩΤΕΡΙΚΩΝ
ΑΣΘΕΝΩΝ**

Παρακαλούμε να μας εκφράσετε την άποψή σας για τις υπηρεσίες του Νοσοκομείου μας, συμπληρώνοντας το παρακάτω ερωτηματολόγιο.

Η συμπλήρωση του ερωτηματολογίου γίνεται ανώνυμα. Το Νοσοκομείο μας έχει λάβει όλα τα απαιτούμενα μέτρα για την προστασία των προσωπικών σας δεδομένων. Οι πληροφορίες που περιλαμβάνονται στο ερωτηματολόγιο θα μας βοηθήσουν να βελτιώσουμε τις υπηρεσίες μας.

ΣΤΟΙΧΕΙΑ ΑΣΘΕΝΟΥΣ:

Φύλο: Άνδρας ☐ Γυναίκα ☐ **Ηλικία:** _____ ετών

Είστε Ασφαλισμένος; 1 ☐ ΝΑΙ ,

(Παρακαλώ σημειώστε το Ταμείο σας)

2 ☐ ΟΧΙ

Ιθαγένεια / Υπηκοότητα (αυτή που αναγράφεται στην ταυτότητα ή το διαβατήριό σας):

1. ☐ Ελληνική

2. ☐ Άλλη από Ελληνική (παρακαλώ σημειώστε : _____)

Επίπεδο εκπαίδευσης: Μέχρι 3^η Γυμνασίου ☐ Τελείωσα Λύκειο ☐

Πανεπιστήμιο ή ΤΕΙ ☐

Τόπος προσέλευσης: Στην εφημερία (στα επείγοντα) ☐

Στα Τακτικά Εξωτερικά Ιατρεία ☐

Στα Τακτικά Ιατρεία της Ολοήμερης Λειτουργίας ☐

Ποιος συμπληρώνει το ερωτηματολόγιο: Ο ίδιος ο ασθενής ☐ Συγγενής ή συνοδός ☐

Ημερομηνία συμπλήρωσης: _____

**ΠΑΡΑΔΩΣΤΕ ΤΟ ΣΥΜΠΛΗΡΩΜΕΝΟ ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ ΣΤΗΝ
ΕΞΟΔΟ:**

- **ΙΣΟΓΕΙΟ: ΤΜΗΜΑ ΚΙΝΗΣΗΣ**
- **2^{ΟΣ} ΟΡΟΦΟΣ: ΓΡΑΜΜΑΤΕΙΑ**

**ΠΑΡΑΚΑΛΟΥΜΕ, ΑΞΙΟΛΟΓΗΣΤΕ ΤΙΣ ΕΝΤΥΠΩΣΕΙΣ ΣΑΣ ΣΧΕΤΙΚΑ ΜΕ ΤΑ
ΠΑΡΑΚΑΤΩ:**

	Πολύ κακές	Μάλλον κακές	Ούτε καλές ούτε κακές	Μάλλον καλές	Πολύ καλές
ΥΠΟΔΟΧΗ - ΠΕΡΙΒΑΛΛΟΝ					
Χρόνος αναμονής					
Ύπαρξη πινακίδων που διευκολύνουν την κίνησή σας στα διάφορα Τμήματα					
Εντυπώσεις από την καθαριότητα και λειτουργικότητα χώρων υποδοχής					
Ύπαρξη ελεύθερων καθισμάτων για τη διάρκεια της αναμονής σας					
Θερμοκρασία χώρων (θέρμανση - ψύξη)					
Ησυχία					
Καθαριότητα βοηθητικών χώρων (W.C., μπάνια κ.ά.)					
ΤΑΧΥΤΗΤΑ ΕΞΥΠΗΡΕΤΗΣΗΣ					
Χρόνος αναμονής για τη διενέργεια εξετάσεων					
Χρόνος έκδοσης αποτελεσμάτων εξετάσεων					
ΚΛΙΝΙΚΕΣ ΚΑΙ ΕΡΓΑΣΤΗΡΙΑΚΕΣ ΕΞΕΤΑΣΕΙΣ					
Συμπεριφορά του προσωπικού κατά τη διάρκεια των εξετάσεων					
Σεβασμός της προσωπικότητας του εξεταζόμενου ασθενούς (παραβάν κ.λπ.)					
ΕΝΤΥΠΩΣΕΙΣ ΑΠΟ ΤΗΝ ΙΑΤΡΙΚΗ ΦΡΟΝΤΙΔΑ					
Ποιότητα ιατρικής φροντίδας (εμπειρία και ικανότητα ιατρών)					
Συμπεριφορά των ιατρών (ευγένεια, φιλικότητα, σεβασμός)					
Πλήρης και κατανοητή ενημέρωση από τους γιατρούς σχετικά με την πορεία της ασθένειας και της θεραπείας σας					
Συμπεριφορά του ιατρικού προσωπικού κατά τη διάρκεια των ιατρικών εξετάσεων					
ΕΝΤΥΠΩΣΕΙΣ ΑΠΟ ΤΗ ΝΟΣΗΛΕΥΤΙΚΗ ΦΡΟΝΤΙΔΑ					
Ποιότητα νοσηλευτικής φροντίδας (εμπειρία και ικανότητα νοσηλευτών)					
Συμπεριφορά νοσηλευτών (ενημέρωση, ευγένεια, φιλικότητα, σεβασμός)					

	Πολύ κακές	Μάλλον κακές	Ούτε καλές ούτε κακές	Μάλλον καλές	Πολύ καλές
ΔΙΟΙΚΗΤΙΚΕΣ ΥΠΗΡΕΣΙΕΣ					
Συμπεριφορά διοικητικού προσωπικού (ενημέρωση, ευγένεια, φιλικότητα, σεβασμός)					
Ταχύτητα διεκπεραίωσης διαδικασιών από το διοικητικό προσωπικό (ταχύτητα εξυπηρέτησης)					

ΣΥΝΟΛΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ Νοσοκομείου:

- Σε κλίμακα από το 0 ως το 10 (όπου το 0 είναι το χειρότερο και το 10 το καλύτερο), με ποιον βαθμό θα αξιολογούσατε τη συνολική σας εμπειρία από το Τμήμα του Νοσοκομείου το οποίο επισκεφθήκατε (Τμήμα Επειγόντων);

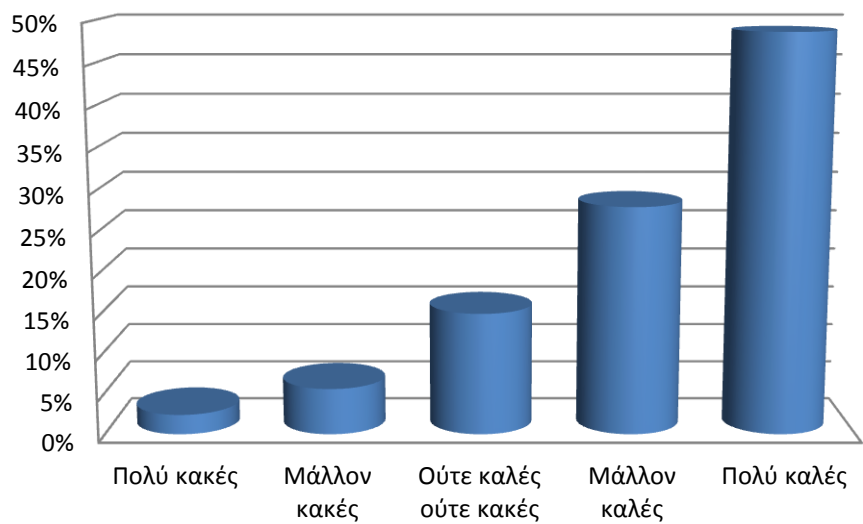
0	1	2	3	4	5	6	7	8	9	10

Παράρτημα IV

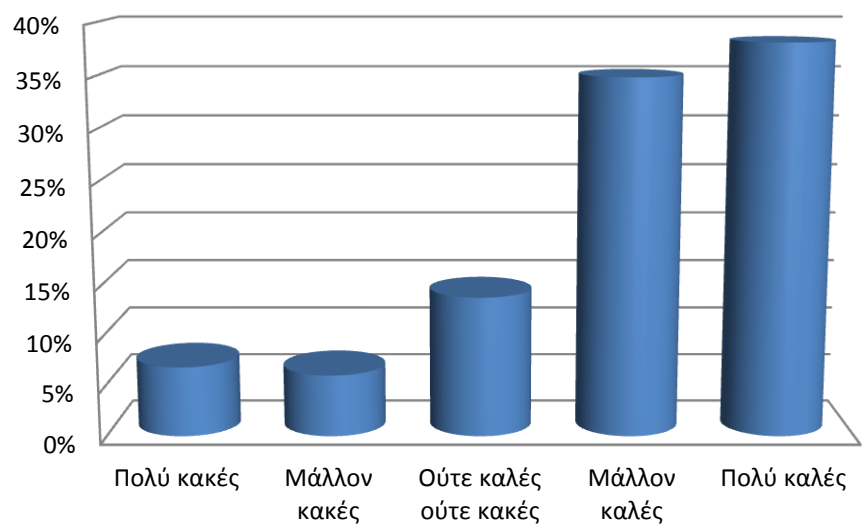
“ Αποτελέσματα ικανοποίησης ασθενών από τις υπηρεσίες του
ΤΕΠ νοσοκομείου ανά κριτήριο”



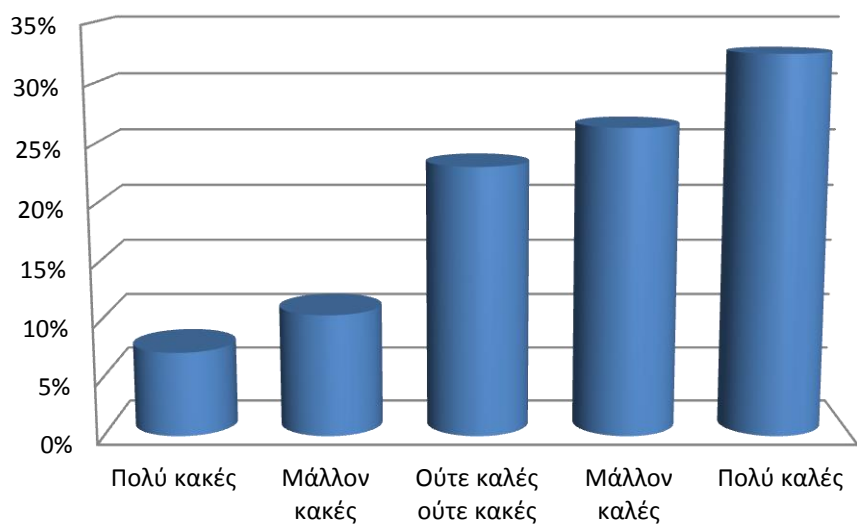
πινακίδες



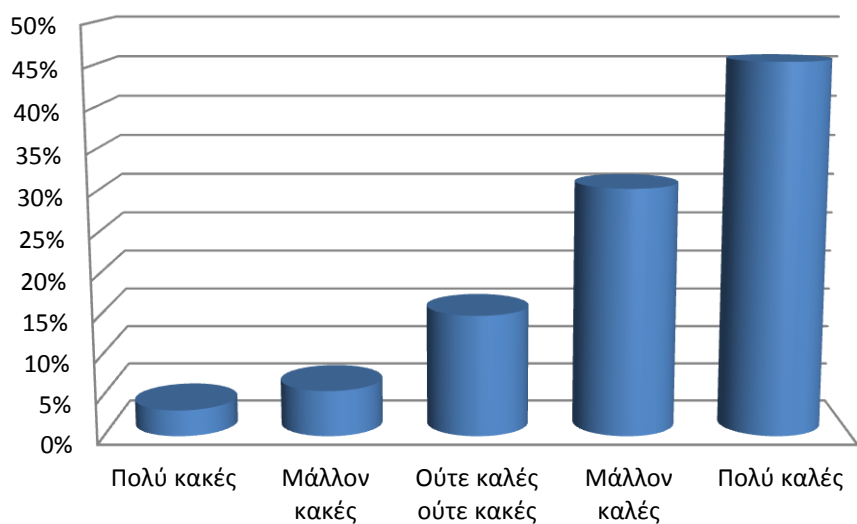
καθαριότητα

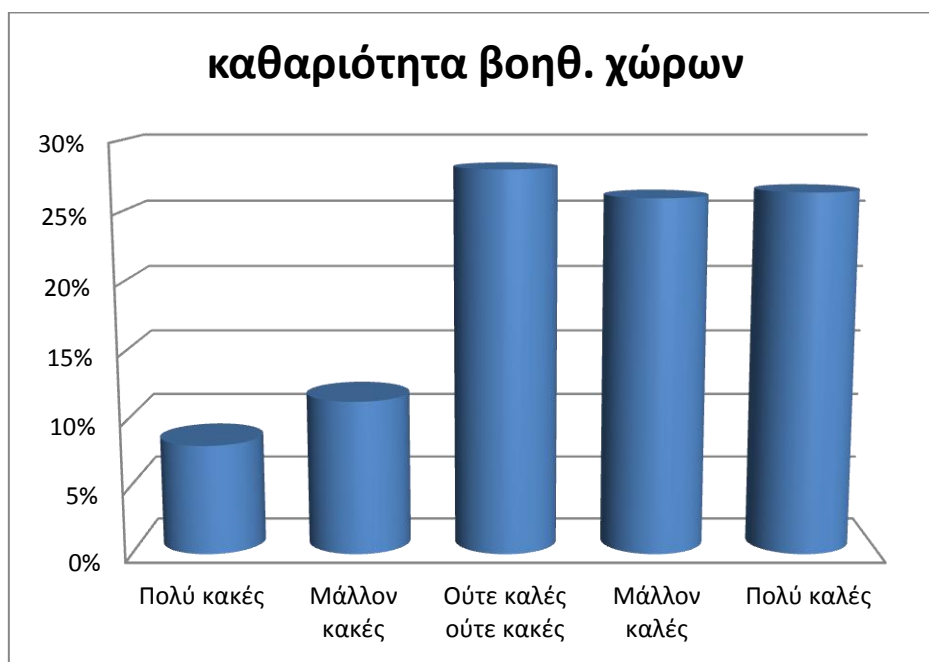
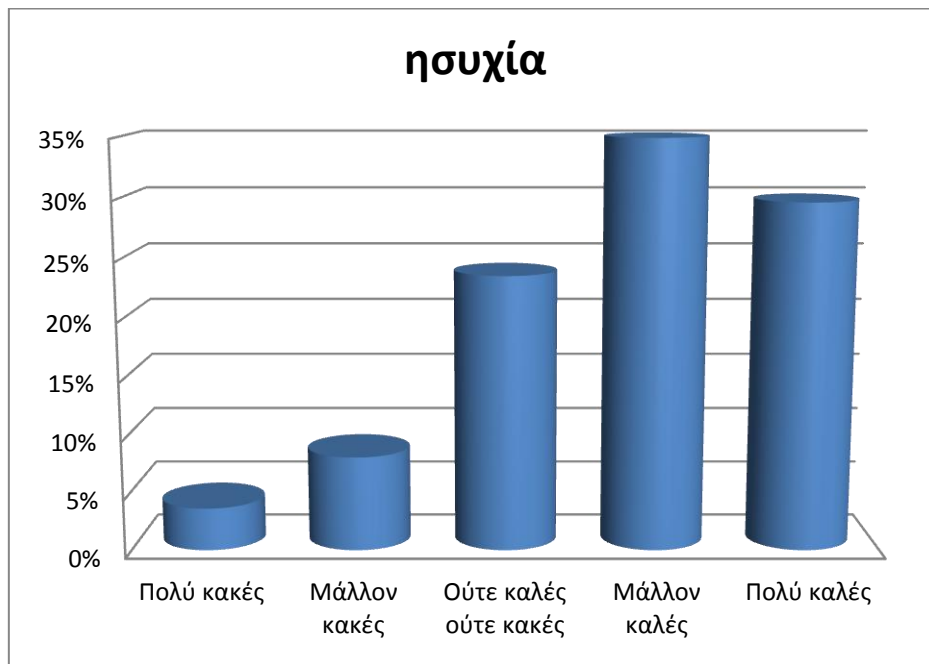


ελεύθερα καθίσματα

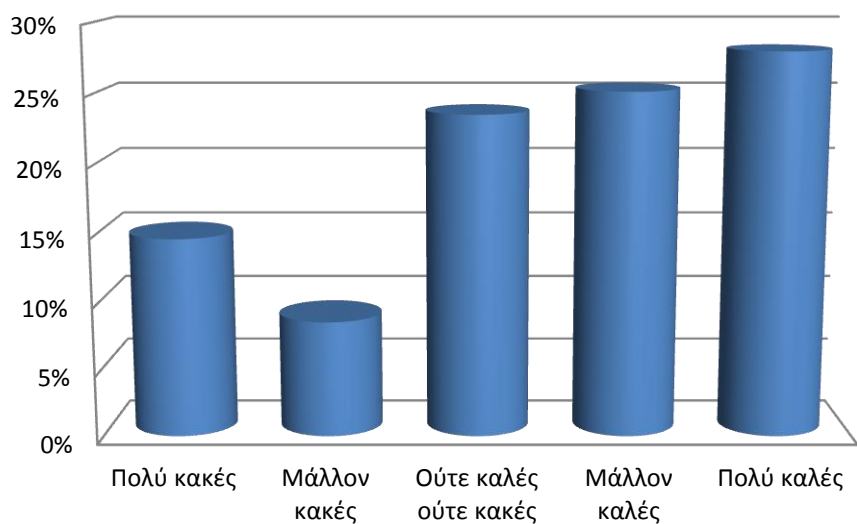


θερμοκρασία

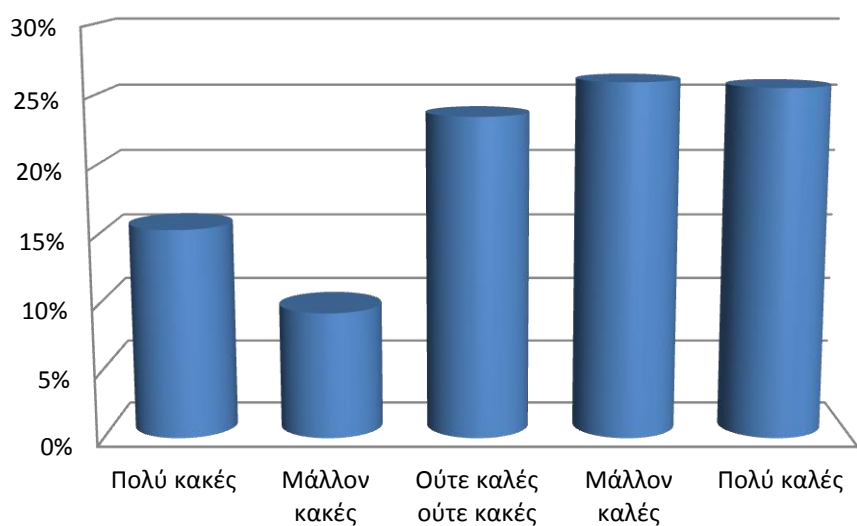




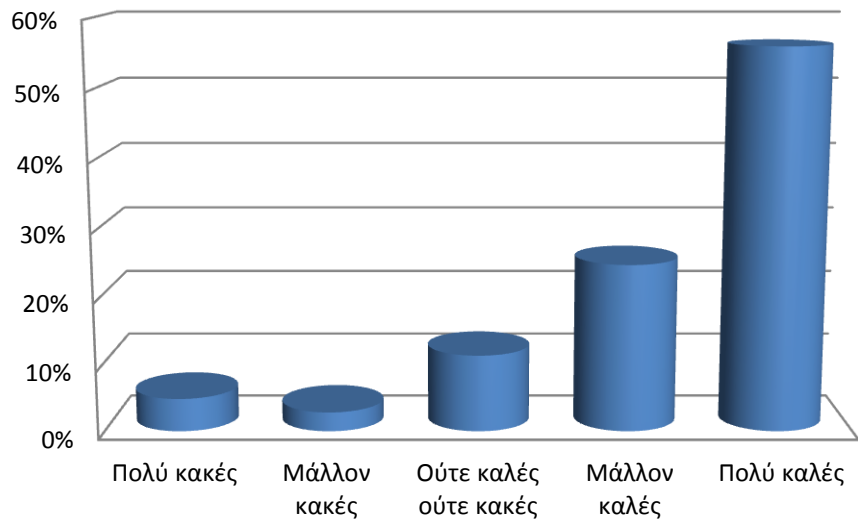
χρόνος αναμονής εξέτασης



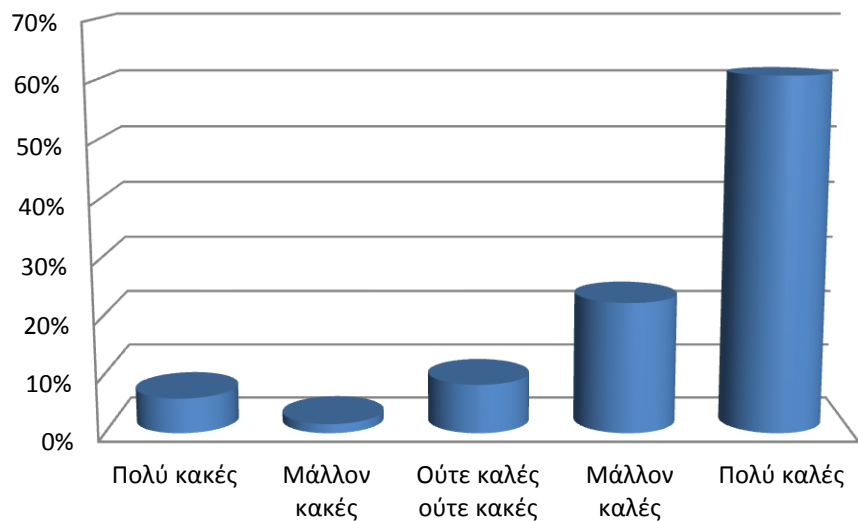
χρόνος έκδοσης αποτελεσμάτων



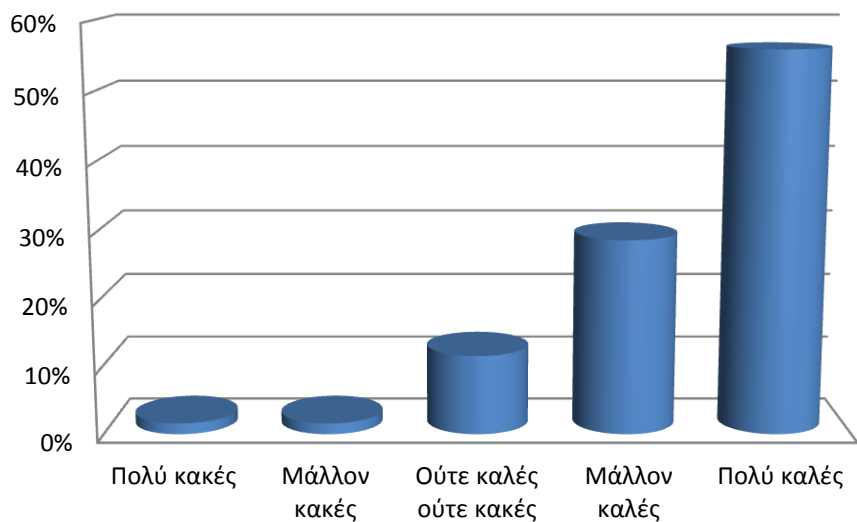
συμπεριφορά προσωπικού



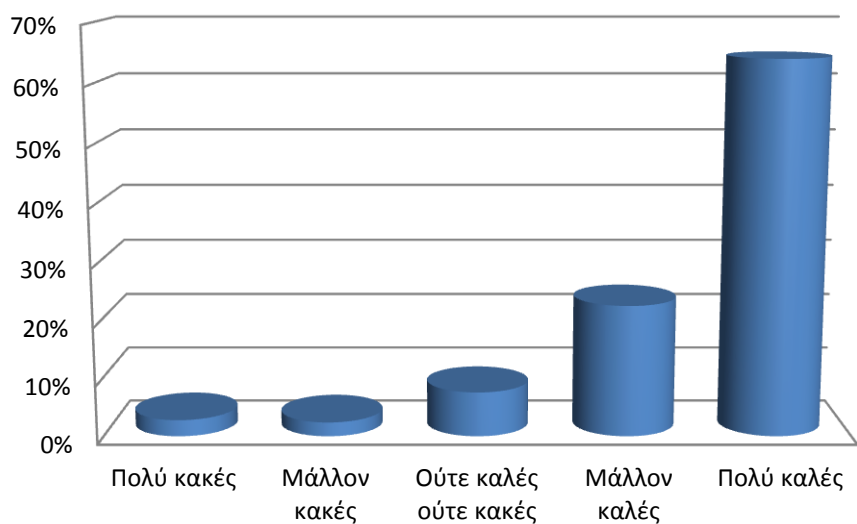
σεβασμός προσωπικότητας



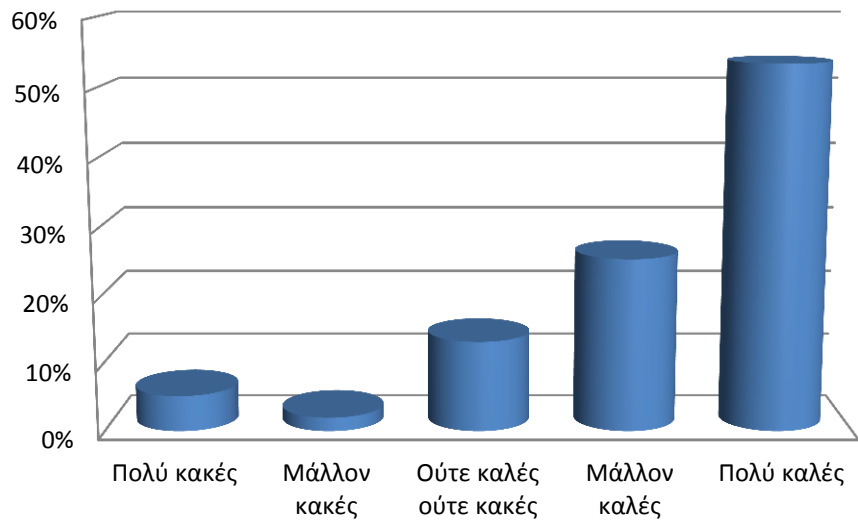
ποιότητα ιατρικής φροντίδας



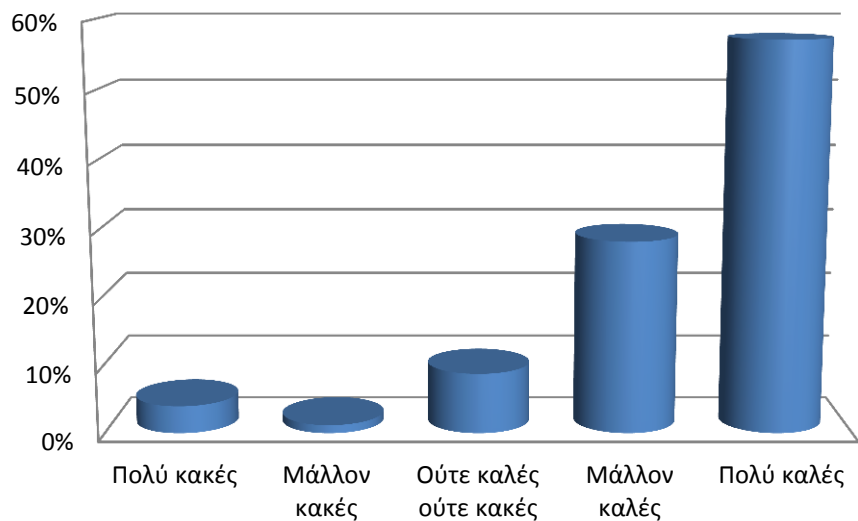
συμπεριφορά γιατρών



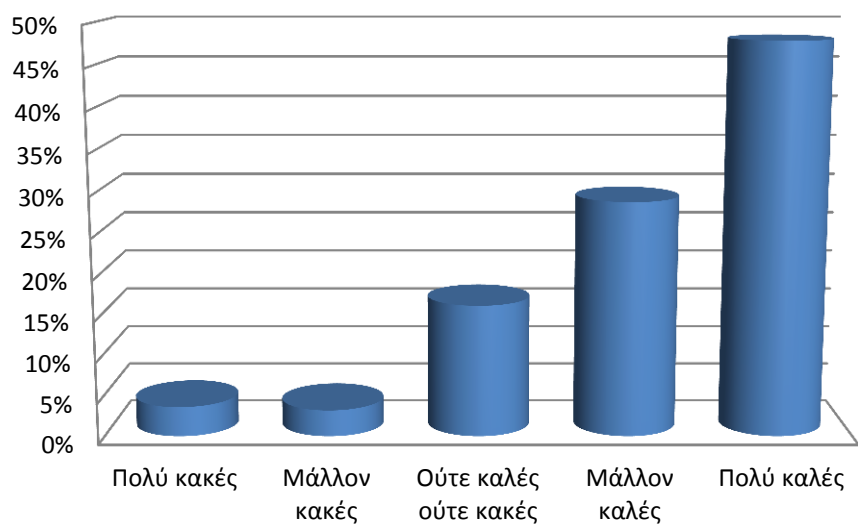
ενημέρωση από γιατρούς



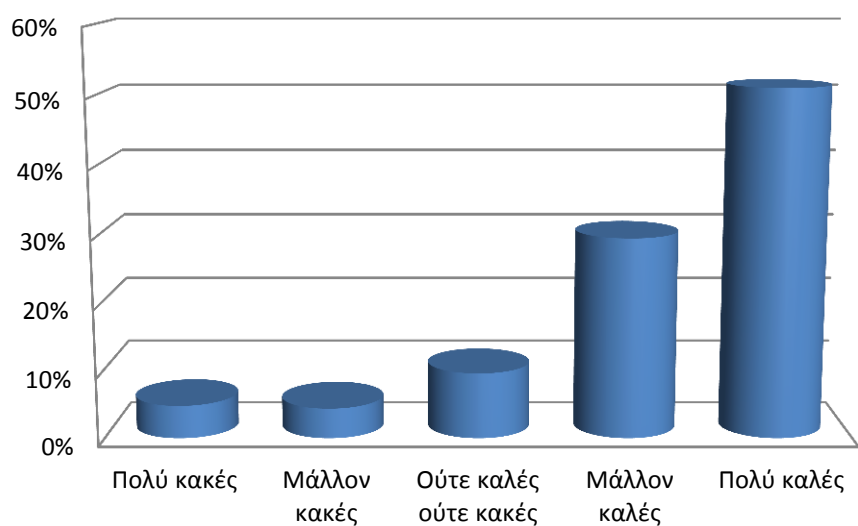
συμπεριφορά γιατών σε εξέταση



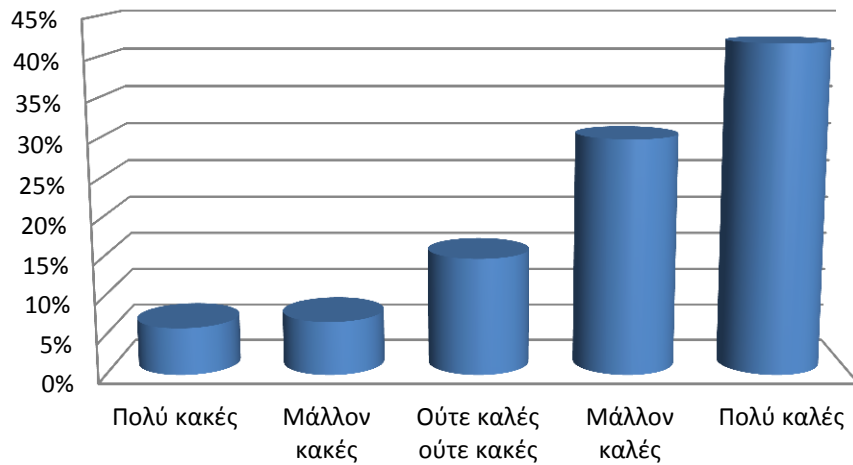
ποιότητα νοσηλευτικής φροντίδας



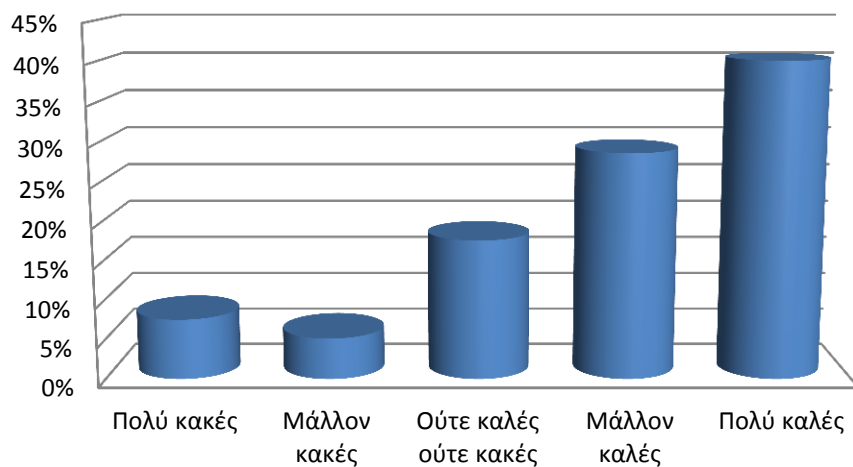
συμπεριφορά νοσηλευτών



συμπεριφορά διοικητικού προσωπικού



ταχύτητα διεκπεραίωσης διοικητικών



Παράρτημα V

“ Προγραμματισμός αλγόριθμου Hot Deck για Ordinal Data σε γλώσσα Java ”

```
package hot.deck;

//import static java.awt.PageAttributes.MediaType.D;
import java.io.BufferedReader;
import java.io.FileReader;
import java.util.StringTokenizer;
import javax.swing.JFileChooser;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileWriter;
import java.io.IOException;
import java.util.logging.Level;
import java.util.logging.Logger;
import javax.swing.JOptionPane;

/**
 *
 * @author idomeneas kamilakis
 */
public class NewJFrame extends javax.swing.JFrame {

    private BufferedReader inputStream = null;
    public int MaxRows;
    public int MaxColumns;
    public double ImputationCoefficient;
    public double WeightCoefficient;
    public int [][] AnswersMatrix;
    private FileReader fr = null;
    private int returnVal;
    private boolean check = true;

    /**
     * Creates new form NewJFrame
     */
    public NewJFrame() {
        initComponents();
        ImputationButton.setEnabled(false);
    }

    /**
     * This method is called from within the constructor to initialize the form.
     * WARNING: Do NOT modify this code. The content of this method is always
     * regenerated by the Form Editor.
     */
    @SuppressWarnings("unchecked")
    // <editor-fold defaultstate="collapsed" desc="Generated Code">
    private void initComponents() {

        jFileChooser1 = new javax.swing.JFileChooser();
        jLabel1 = new javax.swing.JLabel();
        jLabel2 = new javax.swing.JLabel();
        BrowseButton = new javax.swing.JButton();
        ImputationButton = new javax.swing.JButton();
        jLabel3 = new javax.swing.JLabel();
        ImputationCoefficientJText = new javax.swing.JTextField();
    }
}
```



```

jLabel4 = new javax.swing.JLabel();
jLabel5 = new javax.swing.JLabel();
ColumnsJText = new javax.swing.JTextField();
RowsJText = new javax.swing.JTextField();
jLabel6 = new javax.swing.JLabel();
ExitButton = new javax.swing.JButton();
ClearButton = new javax.swing.JButton();
WeightJText = new javax.swing.JTextField();
jLabel7 = new javax.swing.JLabel();

setDefaultCloseOperation(javax.swing.WindowConstants.EXIT_ON_CLOSE);

jLabel1.setFont(new java.awt.Font("Tahoma", 1, 14)); // NOI18N
jLabel1.setText("Hot Deck Imputation Program");

jLabel2.setText("by Idomeneas Kamilakis");

BrowseButton.setText("Browse File");
BrowseButton.addActionListener(new java.awt.event.ActionListener() {
    public void actionPerformed(java.awt.event.ActionEvent evt) {
        BrowseButtonActionPerformed(evt);
    }
});

ImputationButton.setText("Imputation");
ImputationButton.addActionListener(new java.awt.event.ActionListener() {
    public void actionPerformed(java.awt.event.ActionEvent evt) {
        ImputationButtonActionPerformed(evt);
    }
});

jLabel3.setText("Imputation Coefficient Limit:");

ImputationCoefficientJText.addActionListener(new java.awt.event.ActionListener() {
    public void actionPerformed(java.awt.event.ActionEvent evt) {
        ImputationCoefficientJTextActionPerformed(evt);
    }
});

jLabel4.setText("Number of Questions:");

jLabel5.setText("Number of Questionnaires:");

ColumnsJText.addActionListener(new java.awt.event.ActionListener() {
    public void actionPerformed(java.awt.event.ActionEvent evt) {
        ColumnsJTextActionPerformed(evt);
    }
});

RowsJText.addActionListener(new java.awt.event.ActionListener() {
    public void actionPerformed(java.awt.event.ActionEvent evt) {
        RowsJTextActionPerformed(evt);
    }
});

jLabel6.setText("For customer satisfaction surveys with ordinal data");

ExitButton.setText("Exit");
ExitButton.addActionListener(new java.awt.event.ActionListener() {
    public void actionPerformed(java.awt.event.ActionEvent evt) {
        ExitButtonActionPerformed(evt);
    }
});

```

```

ClearButton.setText("Clear");
ClearButton.addActionListener(new java.awt.event.ActionListener() {
    public void actionPerformed(java.awt.event.ActionEvent evt) {
        ClearButtonActionPerformed(evt);
    }
});

WeightJText.addActionListener(new java.awt.event.ActionListener() {
    public void actionPerformed(java.awt.event.ActionEvent evt) {
        WeightJTextActionPerformed(evt);
    }
});

jLabel7.setText("Total Satisfaction Weight Coefficient:");

javax.swing.GroupLayout layout = new javax.swing.GroupLayout(getContentPane());
getContentPane().setLayout(layout);
layout.setHorizontalGroup(
    layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
        .addGroup(layout.createSequentialGroup()
            .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
                .addGroup(layout.createSequentialGroup()
                    .addGap(46, 46, 46)
                    .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
                        .addComponent(BrowseButton, javax.swing.GroupLayout.PREFERRED_SIZE,
117, javax.swing.GroupLayout.PREFERRED_SIZE)
                        .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
                        .addComponent(ClearButton, javax.swing.GroupLayout.PREFERRED_SIZE, 123,
javax.swing.GroupLayout.PREFERRED_SIZE))
                    .addGroup(layout.createSequentialGroup()
                        .addComponent(jLabel2)
                        .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
                        .addComponent(ExitButton))
                    .addComponent(ImputationButton, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)))
                .addGroup(layout.createSequentialGroup()
                    .addGap(0, 82, Short.MAX_VALUE)
                    .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
                        .addComponent(jLabel3, javax.swing.GroupLayout.Alignment.TRAILING)
                        .addComponent(jLabel4, javax.swing.GroupLayout.Alignment.TRAILING)
                        .addComponent(jLabel5, javax.swing.GroupLayout.Alignment.TRAILING)
                        .addComponent(jLabel7, javax.swing.GroupLayout.Alignment.TRAILING)
                        .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)
                        .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
                            .addComponent(ColumnsJText, javax.swing.GroupLayout.DEFAULT_SIZE, 69,
Short.MAX_VALUE)
                            .addComponent(RowsJText)
                            .addComponent(ImputationCoefficientJText,
javax.swing.GroupLayout.DEFAULT_SIZE, 150, Short.MAX_VALUE)
                            .addComponent(WeightJText))))
                    .addGap(29, 29, 29))
            .addGroup(layout.createSequentialGroup()
                .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
                    .addGroup(layout.createSequentialGroup()
                        .addGap(116, 116, 116)
                        .addComponent(jLabel1))
                    .addGroup(layout.createSequentialGroup()
                        .addGap(100, 100, 100)

```

```

        .addComponent(jLabel6)))
        .addContainerGap(javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE))
    );

    layout.linkSize(javax.swing.SwingConstants.HORIZONTAL, new java.awt.Component[]
    { ColumnsJText, ImputationCoefficientJText, RowsJText });

    layout.setVerticalGroup(
        layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
        .addGroup(layout.createSequentialGroup()
            .addGap(18, 18, 18)
            .addComponent(jLabel1)
            .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.UNRELATED)
            .addComponent(jLabel6, javax.swing.GroupLayout.PREFERRED_SIZE, 14,
                javax.swing.GroupLayout.PREFERRED_SIZE)
            .addGap(36, 36, 36)
            .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.BASELINE)
                .addComponent(jLabel4)
                .addComponent(ColumnsJText, javax.swing.GroupLayout.PREFERRED_SIZE,
                    javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE))
            .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.UNRELATED)
            .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.BASELINE)
                .addComponent(RowsJText, javax.swing.GroupLayout.PREFERRED_SIZE, 20,
                    javax.swing.GroupLayout.PREFERRED_SIZE)
                .addComponent(jLabel5))
            .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.UNRELATED)
            .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.BASELINE)
                .addComponent(jLabel3)
                .addComponent(ImputationCoefficientJText,
                    javax.swing.GroupLayout.PREFERRED_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE,
                    javax.swing.GroupLayout.PREFERRED_SIZE))
            .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)
            .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.BASELINE)
                .addComponent(WeightJText, javax.swing.GroupLayout.PREFERRED_SIZE,
                    javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE)
                .addComponent(jLabel7))
            .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED, 36,
                Short.MAX_VALUE)
            .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.BASELINE)
                .addComponent(BrowseButton)
                .addComponent(ClearButton))
            .addGap(18, 18, 18)
            .addComponent(ImputationButton)
            .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.UNRELATED)
            .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
                .addComponent(jLabel2, javax.swing.GroupLayout.PREFERRED_SIZE, 32,
                    javax.swing.GroupLayout.PREFERRED_SIZE)
                .addComponent(ExitButton))
            .addGap(12, 12, 12))
        );

    layout.linkSize(javax.swing.SwingConstants.VERTICAL, new java.awt.Component[]
    { ColumnsJText, ImputationCoefficientJText, RowsJText });

    pack();
} // </editor-fold>

private void ExitButtonActionPerformed(java.awt.event.ActionEvent evt) {
    System.exit(0);
}

private void ClearButtonActionPerformed(java.awt.event.ActionEvent evt) {
    ImputationCoefficientJText.setText("");
    WeightJText.setText("");
}

```

```

ColumnsJText.setText("");
RowsJText.setText("");
check=true;
}

private void ImputationButtonActionPerformed(java.awt.event.ActionEvent evt) {
try {
    ImputationCoefficient = Double.parseDouble(ImputationCoefficientJText.getText());
    if ((ImputationCoefficient < 0) || (ImputationCoefficient > 1)){
        JOptionPane.showMessageDialog(null, "Imputation Coefficient out of normal values!",
"ERROR", JOptionPane.ERROR_MESSAGE);
        check=false;
    }
    WeightCoefficient = Double.parseDouble(WeightJText.getText());
    if ((WeightCoefficient < 0) || (WeightCoefficient > 1)){
        JOptionPane.showMessageDialog(null, "Total Satisfaction Weight Coefficient out of normal
values!", "ERROR", JOptionPane.ERROR_MESSAGE);
        check=false;
    }
    hotDeck();
    if(check==true){
        FileWriter fw;
        returnVal = JFileChooser1.showSaveDialog(this);
        if (returnVal == JFileChooser.APPROVE_OPTION){
            fw = new FileWriter(JFileChooser1.getSelectedFile() + ".xls");    // file with xls format
            for (int row = 0; row < MaxRows; row++){
                for (int column = 0; column < MaxColumns; column++){
                    fw.write(AnswersMatrix[row][column] + "\t");
                }
                fw.write(System.getProperty( "line.separator" )); // next line
            }
            fw.flush();
            fw.close();
            JOptionPane.showMessageDialog(null, "Imputation completed!", "Success!",
JOptionPane.INFORMATION_MESSAGE);
        }
        else { JOptionPane.showMessageDialog(null, "Imputation not completed!", "ERROR",
JOptionPane.ERROR_MESSAGE);}
    }
    catch (NumberFormatException ex) {
        JOptionPane.showMessageDialog(null, "Please check your input data!", "ERROR",
JOptionPane.ERROR_MESSAGE);
    } catch (FileNotFoundException ex) {
        Logger.getLogger(NewJFrame.class.getName()).log(Level.SEVERE, null, ex);
    } catch (IOException ex) {
        Logger.getLogger(NewJFrame.class.getName()).log(Level.SEVERE, null, ex);
    }
}

private void BrowseButtonActionPerformed(java.awt.event.ActionEvent evt) {
try{
    MaxRows = Integer.parseInt(RowsJText.getText());
    MaxColumns = Integer.parseInt(ColumnsJText.getText());
    AnswersMatrix = new int [MaxRows][MaxColumns];
    returnVal = JFileChooser1.showOpenDialog(this);
    if (returnVal == JFileChooser.APPROVE_OPTION){
        ImputationButton.setEnabled(true);
        File measurementsFile = JFileChooser1.getSelectedFile();
        String line;
        fr = new FileReader(measurementsFile.getAbsolutePath());
        inputStream = new BufferedReader(new FileReader(measurementsFile));
        line = inputStream.readLine();
        for (int row = 0; row < MaxRows; row++){
            StringTokenizer measurementsFinder = new StringTokenizer(line);
            for (int column = 0; column < MaxColumns; column++){

```

```

        AnswersMatrix[row][column] = Integer.parseInt(measurementsFinder.nextToken());
    }
    line = inputStream.readLine(); // next line
    }}
    else{
        JOptionPane.showMessageDialog(null, "Please select input file!", "ERROR",
JOptionPane.ERROR_MESSAGE);
    }
    catch (NumberFormatException ex) {
        JOptionPane.showMessageDialog(null, "Please check your input data!", "ERROR",
JOptionPane.ERROR_MESSAGE);
    }
    catch (FileNotFoundException ex) {
        Logger.getLogger(NewJFrame.class.getName()).log(Level.SEVERE, null, ex);
    } catch (IOException ex) {
        Logger.getLogger(NewJFrame.class.getName()).log(Level.SEVERE, null, ex);
    } finally {
        try {
            if (inputStream != null){
                inputStream.close();
            }
            if (fr != null){
                fr.close();
            }
        } catch (IOException ex) {
            Logger.getLogger(NewJFrame.class.getName()).log(Level.SEVERE, null, ex);
        }
    }
}

private void ImputationCoefficientJTextActionPerformed(java.awt.event.ActionEvent evt) {
    // TODO add your handling code here:
}

private void ColumnsJTextActionPerformed(java.awt.event.ActionEvent evt) {
    // TODO add your handling code here:
}

private void RowsJTextActionPerformed(java.awt.event.ActionEvent evt) {
    // TODO add your handling code here:
}

private void WeightJTextActionPerformed(java.awt.event.ActionEvent evt) {
    // TODO add your handling code here:
}

/**
 * @param args the command line arguments
 */
public static void main(String args[]) {
    /* Set the Nimbus look and feel */
    //<editor-fold defaultstate="collapsed" desc=" Look and feel setting code (optional) ">
    /* If Nimbus (introduced in Java SE 6) is not available, stay with the default look and feel.
     * For details see http://download.oracle.com/javase/tutorial/uiswing/lookandfeel/plaf.html
     */
    try {
        for
            (javax.swing.UIManager.LookAndFeelInfo info
                : javax.swing.UIManager.getInstalledLookAndFeels()) {
            if ("Nimbus".equals(info.getName())) {
                javax.swing.UIManager.setLookAndFeel(info.getClassName());
                break;
            }
        }
    } catch (ClassNotFoundException | InstantiationException | IllegalAccessException |
javax.swing.UnsupportedLookAndFeelException ex) {

```

```

java.util.logging.Logger.getLogger(NewJFrame.class.getName()).log(java.util.logging.Level.SEVERE,
null, ex);
}
//</editor-fold>

/* Create and display the form */
java.awt.EventQueue.invokeLater(new Runnable() {
    @Override
    public void run() {
        new NewJFrame().setVisible(true);
    }
});
}

// Variables declaration - do not modify
private javax.swing.JButton BrowseButton;
private javax.swing.JButton ClearButton;
private javax.swing.JTextField ColumnsJText;
private javax.swing.JButton ExitButton;
private javax.swing.JButton ImputationButton;
private javax.swing.JTextField ImputationCoefficientJText;
private javax.swing.JTextField RowsJText;
private javax.swing.JTextField WeightJText;
private javax.swing.JFileChooser jFileChooser1;
private javax.swing.JLabel jLabel1;
private javax.swing.JLabel jLabel2;
private javax.swing.JLabel jLabel3;
private javax.swing.JLabel jLabel4;
private javax.swing.JLabel jLabel5;
private javax.swing.JLabel jLabel6;
private javax.swing.JLabel jLabel7;
// End of variables declaration

public void hotDeck (){
    double Distance;
    double C [][] = new double [MaxRows][MaxRows]; //ImputationCoefficient matrix for i,k
    double questionnairesDistance [][] = new double [MaxRows][MaxRows];
    double totalSatisfactionDistance [][] = new double [MaxRows][MaxRows];
    double totalDistance [][] = new double [MaxRows][MaxRows];

    //Filling ImputationCoefficient matrix C[i][k]
    for (int i = 0; i < MaxRows; i++){
        for (int k = i; k < MaxRows; k++){
            double c=0;
            for (int b = 0; b < MaxColumns; b++){
                if ((AnswersMatrix[i][b] != 999) && (AnswersMatrix[k][b] != 999)){
                    c=c+1; //counter for non blank columns for questionnaires i,k
                }
            }
            C[i][k] = c/MaxColumns;
            C[k][i] = C[i][k];
        }
    }

    //WALESIK algorithm for total satisfaction
    for (int i = 0; i < MaxRows; i++) {
        for (int k = i; k < MaxRows; k++){
            double s1=0; double s2=0; double s3=0; double s4=0; double s5=0; double s6=0;
            double a1, a2, b1, b2;
            if ((AnswersMatrix[i][0] != 999) && (AnswersMatrix[k][0] != 999)){ //distance calculated
                only for non blank total satisfaction

                if (AnswersMatrix[i][0]> AnswersMatrix[k][0]){a1=1;}
                else if (AnswersMatrix[i][0]< AnswersMatrix[k][0]){a1=-1;}
            }
        }
    }

```

```

        else {a1=0;}
        if (AnswersMatrix[k][0]> AnswersMatrix[i][0]){b1=1;}
        else if (AnswersMatrix[k][0]< AnswersMatrix[i][0]){b1=-1;}
        else {b1=0;}
        s1=(s1)+((a1)*(b1));
        s3=(s3)+((a1)*(a1));
        s4=(s4)+((b1)*(b1));

        for (int l = 0; l < MaxRows; l++){
            if ((l!=k) && (l!=i)){
                if (AnswersMatrix[i][0]> AnswersMatrix[l][0]){a2=1;}
                else if (AnswersMatrix[i][0]< AnswersMatrix[l][0]){a2=-1;}
                else {a2=0;}
                if (AnswersMatrix[k][0]> AnswersMatrix[l][0]){b2=1;}
                else if (AnswersMatrix[k][0]< AnswersMatrix[l][0]){b2=-1;}
                else {b2=0;}
                s2=(s2)+((a2)*(b2));
                s5=(s5)+((a2)*(a2));
                s6=(s6)+((b2)*(b2));
            }
        }

        totalSatisfactionDistance[i][k]= 0.5 -((s1+s2)/(2* Math.sqrt((s3+s5)*(s4+s6))));
    }
    else {totalSatisfactionDistance[i][k]=1;}
    totalSatisfactionDistance[k][i] = totalSatisfactionDistance[i][k];
}

//WALESIK algorithm for remaining questions
for (int i = 0; i < MaxRows; i++) {
    for (int k = i; k < MaxRows; k++){
        double s1=0; double s2=0; double s3=0; double s4=0; double s5=0; double s6=0;
        double a1, a2, b1, b2;
        for (int j = 1; j < MaxColumns; j++){
            if ((AnswersMatrix[i][j] != 999) && (AnswersMatrix[k][j] != 999)){ //distance calculated
                only for non blank columns

                if (AnswersMatrix[i][j]> AnswersMatrix[k][j]){a1=1;}
                else if (AnswersMatrix[i][j]< AnswersMatrix[k][j]){a1=-1;}
                else {a1=0;}
                if (AnswersMatrix[k][j]> AnswersMatrix[i][j]){b1=1;}
                else if (AnswersMatrix[k][j]< AnswersMatrix[i][j]){b1=-1;}
                else {b1=0;}
                s1=(s1)+((a1)*(b1));
                s3=(s3)+((a1)*(a1));
                s4=(s4)+((b1)*(b1));
            }
        }
        for (int l = 0; l < MaxRows; l++){
            if ((l!=k) && (l!=i)){
                for (int j = 1; j < MaxColumns; j++){
                    if ((AnswersMatrix[i][j] != 999) && (AnswersMatrix[k][j] != 999)){ //distance
                        calculated only for non blank columns
                        if (AnswersMatrix[i][j]> AnswersMatrix[l][j]){a2=1;}
                        else if (AnswersMatrix[i][j]< AnswersMatrix[l][j]){a2=-1;}
                        else {a2=0;}
                        if (AnswersMatrix[k][j]> AnswersMatrix[l][j]){b2=1;}
                        else if (AnswersMatrix[k][j]< AnswersMatrix[l][j]){b2=-1;}
                        else {b2=0;}
                        s2=(s2)+((a2)*(b2));
                        s5=(s5)+((a2)*(a2));
                        s6=(s6)+((b2)*(b2));
                    }
                }
            }
        }

        questionnairesDistance[i][k] = 0.5 -((s1+s2)/(2* Math.sqrt((s3+s5)*(s4+s6))));
        questionnairesDistance[k][i] = questionnairesDistance[i][k];
    }
}

```

```

        totalDistance[i][k] = ((1-WeightCoefficient) * questionnairesDistance[i][k] + WeightCoefficient
* totalSatisfactionDistance[i][k]);
        totalDistance[k][i] = totalDistance[i][k];
    }
}

//Column only blank check
if( check != false) {
    for (int b = 0; b < MaxColumns; b++){
        int blankColumnCounter=0;
        for (int a = 0; a < MaxRows; a++) {
            if (check == false){break;}
            else {
                if(AnswersMatrix[a][b]!=999)
                    {blankColumnCounter=blankColumnCounter+1;}
            }
        }
        if (blankColumnCounter == 0) {
            check = false;
            JOptionPane.showMessageDialog(null, "Column: " + (b+1) + " is blank!", "ERROR",
JOptionPane.ERROR_MESSAGE);
            break;
        }
    }
}

//Row only blank check
if( check != false) {
    for (int a = 0; a < MaxRows; a++) {
        int blankRowCounter=0;
        for (int b = 0; b < MaxColumns; b++){
            if (check == false){break;}
            else {
                if(AnswersMatrix[a][b]!=999)
                    {blankRowCounter=blankRowCounter+1;}
            }
        }
        if (blankRowCounter == 0) {
            check = false;
            JOptionPane.showMessageDialog(null, "Row: " + (a+1) + " is blank!", "ERROR",
JOptionPane.ERROR_MESSAGE);
            break;
        }
    }
}

//imputation
int imputationcheck;
for (int a = 0; a < MaxRows; a++) {
    for (int b = 0; b < MaxColumns; b++) {
        if (check == false){break;}
        else {
            if (AnswersMatrix[a][b] == 999){
                Distance= 1; //blank check //Distance initialisation
                imputationcheck=0;
                for (int i = 0; i < MaxRows; i++){
                    if (check == false){break;}
                    else {
                        if (i != a){
                            //imputation from different questionnaire
                            if (AnswersMatrix[i][b] != 999){
                                //imputation from
questionnaires with no blank in b question
                                if (totalDistance[a][i] < Distance){
                                    //minimum Distance
                                    if (((C[a][i] * MaxColumns) >= (ImputationCoefficient * MaxColumns -
1.0)){
                                        // normal type is: C[a][i] >= (ImputationCoefficient - 1/MaxColumns) or (c +
1)/MaxColumns >= ImputationCoefficient, where c is the number of non blank questions between 2
questionnaires
                                        Distance = totalDistance[a][i];
                                        AnswersMatrix[a][b] = AnswersMatrix[i][b]; //imputation
                                        imputationcheck=1;

```



```

        }
    }}}
    if (i == MaxRows -1){          //check for successful imputation
        if (imputationcheck == 0) {
            check = false;
            JOptionPane.showMessageDialog(null, "Too many blanks, check your data
or reduce Imputation Coefficient!", "ERROR", JOptionPane.ERROR_MESSAGE);
            }}}}}}}
    }}

```