Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΣΤΑΤΙΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΟΡΘΟΤΗΤΑΣ ΚΑΙ ΔΙΟΡΘΩΣΗΣ ΕΙΣΕΡΧΟΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ

Σωτηρία Β. Γάκη

Διπλωματική Εργασία

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

Πειραιάς
Νοέμβριος 2004

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

Τα μέλη της Επιτροπής ήταν:

- Πολίτης Κ. (Επιβλέπων)
- Κούτρας Μ.
- Ηλιόπουλος Γ.

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

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

STATISTICAL DATA EDITING

By

Sotiria B. Gaki

MSc Dissertation

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

Piraeus, Greece
November 2004

Στην οικογένειά μου
και στη φίλη μου, Αγγελική

Ευχαριστίες

Θα ήθελα προσωπικά να ευχαριστήσω τον κ. Κ. Πολίτη, επιβλέποντα της διπλωματικής εργασίας, για την καθοδήγησή του, τον κ. Γ. Πετράκο για την άψογη συνεργασία και τις επικοδομητικές συζητήσεις μας και τον Μνιέστρη Νίκο για τη πολύτιμη βοήθειά του.

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία επικεντρώνεται στο ρόλο του ελέγχου ορθότητας και διόρθωσης στατιστικών δεδομένων, γνωστό στο χώρο της Επίσημης Στατιστικής ως *statistical data editing*. Συγκεκριμένα, αναλύονται σε θεωρητικό πλαίσιο και με τη βοήθεια παραδειγμάτων, μέθοδοι που αναπτύχθηκαν για το σκοπό αυτό και η εξέλιξή τους μέσα στο χρόνο. Κύριο μεθοδολογικό εργαλείο στην εφαρμογή του ελέγχου, αποτελούν οι κανόνες ορθότητας που αποκαλύπτουν τις ασυνέπειες και τα λάθη στα δεδομένα. Η διόρθωση των λαθών αυτών καταλήγει σε «καθαρά» δεδομένα τα οποία στη συνέχεια χρησιμοποιούνται για αναλυτικούς σκοπούς. Σκοπός της διπλωματικής εργασίας είναι η γενική επισκόπηση των μεθοδολογιών και των αυτοματοποιημένων συστημάτων ελέγχου και διόρθωσης κατηγορικών, αριθμητικών και μεικτών δεδομένων που χρησιμοποιούν διεθνείς οργανισμοί και Στατιστικές Υπηρεσίες.

ABSTRACT

The present dissertation focuses on the role of editing, known as statistical data editing in the Official Statistics community, which includes inspecting for errors and imputing statistical data of any type. Specifically, we elaborate in a theoretical framework methods targeted on editing, illustrated by helpful examples. Additionally, we present improvements on these methods and other approaches and methodological tools. Edit rules are the main methodological tool, used in application of editing, which can reveal inconsistencies and errors in data. Imputing such errors leads to 'clean' data which could then be used for analytic purposes. The main aim of the dissertation is to present a general review of the methods and automated editing systems, applied on qualitative, quantitative and mixed data, used by international organisations and national statistics agencies.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Ο **στατιστικός έλεγχος ορθότητας των δεδομένων** (statistical data editing) αποτελεί ιδιαίτερα σημαντική διαδικασία για την ποιότητα των δεδομένων και εφαρμόζεται σε διάφορες φάσεις του κύκλου ζωής της πληροφορίας. Η εφαρμογή του στατιστικού ελέγχου ορθότητας έγκειται στην επιθεώρηση των στατιστικών δεδομένων, με τη βοήθεια κανόνων ορθότητας (edit rules), οι οποίοι εντοπίζουν ελλειπείς, λανθασμένες ή πιθανόν λανθασμένες ή/και αντιφατικές τιμές μεταβλητών και τελικά τις διορθώνουν. (Fellegi-Holt 1976). Το αποτέλεσμα της εφαρμογής του ελέγχου ορθότητας είναι η παραγωγή «καθαρών» δεδομένων απαλλαγμένων από λάθη και ασυνέπειες, ενώ η χρήση των τελευταίων σε μεθόδους περιγραφικής και επαγωγικής στατιστικής ανάλυσης αποφέρει αξιόπιστα στατιστικά συμπεράσματα.

Μπορούμε να διακρίνουμε δύο είδη ελέγχου ορθότητας :

- Ο **μίκρο-έλεγχος** (micro editing) ο οποίος αναλαμβάνει να εξετάσει κάθε μια εγγραφή ξεχωριστά, ανεξάρτητα από τις υπόλοιπες. Για παράδειγμα σε μια δημογραφική έρευνα η μεταβλητή ηλικία δεν πρέπει να παίρνει τιμές εκτός του διαστήματος 0-120 ή όταν η ηλικία (ενός ατόμου) είναι μικρότερη των 16 δεν μπορεί η οικογενειακή κατάσταση να λαμβάνει τιμή παντρεμένος.
- Ο **μάκρο-έλεγχος** (macro-editing) εξετάζει τις τιμές μιας εγγραφής σε σχέση με τις υπόλοιπες εγγραφές. Η εξέταση των δεδομένων για έκτροπες παρατηρήσεις είναι ένα χαρακτηριστικό παράδειγμα αυτού του είδους ελέγχων. Επίσης η σύγκριση των τιμών μιας εγγραφής με άλλες πηγές δεδομένων αποτελεί μια ακόμη εφαρμογή του μάκρο-ελέγχου.

Μια διαφορετική εφαρμογή των μίκρο και μάκρο ελέγχων αποτελεί ο επιλεκτικός έλεγχος (selective editing), ως ενδιάμεσο βήμα ανάμεσα στον έλεγχο ορθότητας και τη διόρθωση. Τα τελευταία δέκα περίπου χρόνια έχουν αναπτυχθεί διάφορες μέθοδοι επιλεκτικού ελέγχου. Όλες όμως βασίζονται στην κεντρική ιδέα της διάκρισης εκείνων των εγγραφών, που θεωρούνται πιο σημαντικές και πρέπει να υποστούν διόρθωση από τους ειδικούς αναλυτές, ενώ οι υπόλοιπες εγγραφές κατευθύνονται σε αυτοματοποιημένα συστήματα διόρθωσης.

Λέγοντας σημαντικές εγγραφές, εννοούμε αυτές που μας ενδιαφέρουν ιδιαίτερα όπως για παράδειγμα οι εγγραφές που προέρχονται από μεγάλες εταιρείες σε μια έρευνα για τη βιομηχανία ή εγγραφές που επηρεάζουν περισσότερο το αποτέλεσμα σύμφωνα με μια μέθοδο εκτίμησης. Μια παραλλαγή του επιλεκτικού ελέγχου αποτελεί **ο έλεγχος ορθότητας σημαντικότητας** (significance editing) κατά τον οποίο εκτιμάται η επιρροή μιας εγγραφής στο αποτέλεσμα πριν την επεξεργασία όλων των εγγραφών. (Granquist and Kovar 1997).

Σύμφωνα με τον ορισμό που δόθηκε προηγούμενα, το τελικό στάδιο του στατιστικού ελέγχου ορθότητας είναι η διόρθωση των δεδομένων. Συγκεκριμένα, **διόρθωση** (imputation) είναι η διαδικασία που εφαρμόζεται για να καθορίσει τις τιμές που πρέπει να αντικαταστήσουν τις ελλειπείς και λανθασμένες τιμές μεταβλητών ή και ασυνεπείς συνδυασμούς τιμών μεταβλητών, εξασφαλίζοντας την συνέπεια των εγγραφών και την αξιοπιστία των εκτιμητών. Προφανώς, πολλές διορθώσεις στα δεδομένα θα μπορούσαν να αποφευχθούν πριν την έναρξη του ελέγχου, με την επικοινωνία των ερευνητών με τον ερωτώμενο ή μέσω προσεκτικής επισκόπησης των ερωτηματολογίων και διόρθωση από ειδικούς αναλυτές. Όμως ολοκληρωτική διόρθωση είναι δύσκολο να επιτευχθεί με τις παραπάνω ενέργειες, για το λόγο αυτό έχουν δημιουργηθεί μεθοδολογίες αυτόματης διόρθωσης που βασίζονται σε κανόνες ορθότητας (Fellegi-Holt 1976), σε ιστορικά δεδομένα, σε μέσους, πηλίκα ή μοντέλα παλινδρόμησης, τέλος σε εγγραφές που δεν περιέχουν λάθη (Bankier 1999). Η μέθοδος διόρθωσης που επιλέγουμε, εξαρτάται από τα αποτελέσματα της εφαρμογής της, αν δηλαδή καταφέρνει να περιορίσει την μεροληψία και αν τελικά οι εγγραφές έχουν διορθωθεί πλήρως.

Εφόσον αναφέραμε τα είδη ελέγχου και προσδιορίσαμε την έννοια της διόρθωσης, αξίζει να σημειώσουμε ότι στην ουσία ο στατιστικός έλεγχος ορθότητας θέτει κανόνες στα δεδομένα μιας έρευνας. Οι κανόνες μπορεί να θέτουν περιορισμούς σε συγκεκριμένες μεταβλητές ή σε συνδυασμούς μεταβλητών, να θέτουν το άθροισμα των τιμών κάποιων μεταβλητών ίσο με μια άλλη μεταβλητή ή να απαιτούν οι τιμές μιας μεταβλητής να ανήκουν σε κάποιο διάστημα τιμών. Οι κανόνες της μορφής αυτής ονομάζονται **αιτιοκρατικοί ή ντετερμινιστικοί** (deterministic edits) η παραβίαση των οποίων σημαίνει ότι η εγγραφή περιέχει σίγουρα λάθος (Biemer and Lyberg 2003). Στα επόμενα κεφάλαια αναπτύσσονται μεθοδολογίες που χρησιμοποιούν ντετερμινιστικούς κανόνες για τον έλεγχο και τη διόρθωση των δεδομένων. Η δημιουργία των κανόνων αυτών γίνεται από ειδικούς αναλυτές γνώστες της έρευνας και εφαρμόζονται στα δεδομένα για τον έλεγχο και τη διόρθωσή τους μέσω

ειδικών αυτοματοποιημένων συστημάτων ελέγχου, η χρήση των οποίων είναι απαραίτητη σε μεγάλου μεγέθους έρευνες.

Οι κανόνες ορθότητας διακρίνονται ανάλογα με το λάθος που μπορεί να ανιχνεύσουν. Για το λόγο ορισμένοι κανόνες χαρακτηρίζονται κρίσιμοι (fatal or critical edits) και είναι αυτοί που ανιχνεύουν τιμές μεταβλητών που πρέπει οπωσδήποτε να διορθωθούν αν δεν τους ικανοποιούν. Λάθη στα δεδομένα που μπορεί να κριθούν κρίσιμα είναι οι λανθασμένες, ελλειπείς τιμές, μη λογικές (εκτός ενός διαστήματος) τιμές ή σχέσεις των τιμών των μεταβλητών. Από την άλλη πλευρά, υπάρχουν οι **αμφίβολοι** (query) κανόνες που ανιχνεύουν «ύποπτες» τιμές μεταβλητών. Λέγοντας ύποπτες τιμές εννοούμε αυτές που πιθανόν να είναι λάθος και χρειάζονται περαιτέρω μελέτη για την επιβεβαίωση του χαρακτηρισμού τους. Ένα παράδειγμα αμφίβολης τιμής μιας μεταβλητής είναι μια λάθος καταχώρηση και μπορεί να διευκρινιστεί με άμεση επικοινωνία με τον ερωτώμενο. Εφόσον δεν είναι σίγουρο αν υπάρχει λάθος σε κάποια εγγραφή, οι αμφίβολοι κανόνες πρέπει να εφαρμόζονται αν επηρεάζουν τους εκτιμητές της έρευνας σε σημαντικό βαθμό. Αυτοί οι κανόνες ονομάζονται και **στοχαστικοί** (stochastic edits) αφού δεν είναι σίγουρο αν ανιχνεύουν πραγματικό λάθος στα δεδομένα. (Biemer and Lyberg 2003).

Αναφερόμενοι στα δεδομένα μιας έρευνας, πρέπει να διακρίνουμε τα είδη των δεδομένων και την επεξεργασία τους από διάφορες μεθόδους στατιστικού ελέγχου ορθότητας. Είναι γνωστό ότι κάθε μεταβλητή για την οποία συλλέγουμε δεδομένα μπορεί να είναι είτε διακριτή, με τιμές σε ένα πεπερασμένο σύνολο ακεραίων αριθμών ή σε ένα σύνολο γραμμάτων (ονοματική), είτε αριθμητική με τιμές για παράδειγμα στο σύνολο των πραγματικών αριθμών. Οι Fellegi-Holt (1976) (κεφάλαιο 2) εισήγαγαν μια πρωτοποριακή μεθοδολογία για τον έλεγχο ορθότητας δεδομένων, που αφορά κυρίως κατηγορικά δεδομένα αν και το θεωρητικό πλαίσιο της μεθόδου τους μπορεί να καλύψει και αριθμητικά δεδομένα. Σύμφωνα με τους Fellegi-Holt (1976) ορίζονται κανόνες ορθότητας από τους ειδικούς αναλυτές, με συγκεκριμένη μορφή από τους οποίους παράγονται νέοι για τον έλεγχο και τη διόρθωση των δεδομένων. Οι Fellegi-Holt (1976) κατάφεραν να δημιουργήσουν ένα αυτοματοποιημένο σύστημα ελέγχου που βρίσκει εφαρμογή κυρίως σε κατηγορικά δεδομένα.

Βέβαια έχουν δημιουργηθεί και αυτοματοποιημένα συστήματα που βασίζονται στην θεωρία των Fellegi-Holt (1976), αλλά επεξεργάζονται αποκλειστικά αριθμητικά δεδομένα όπως το SPEER (Structured Programs for Economic Editing and Referrals) που χρησιμοποιείται από την Αμερικάνικη Υπηρεσία Απογραφών (Kovar and Winkler 1996).

Τέλος, αναπτύχθηκαν μέθοδοι αποκλειστικά για τον έλεγχο ορθότητας αριθμητικών δεδομένων ανεξάρτητα από την μέθοδο των Fellegi-Holt όπως του Quere (2000) (κεφάλαιο 4). Προφανώς τα δεδομένα μιας έρευνας περιλαμβάνουν κατηγορικές και αριθμητικές μεταβλητές. Για τα δεδομένα που περιέχουν και τα δυο είδη μεταβλητών, τα οποία ονομάζονται μεικτά, έχουν επίσης αναπτυχθεί μεθοδολογίες ελέγχου όπως των Quere and De Waal (2000), De Waal (2003). Συγκεκριμένα, για τον έλεγχο αριθμητικών και μεικτών δεδομένων Quere (2000), Quere and De Waal (2000) έχει χρησιμοποιηθεί η θεωρία γραφημάτων κατά την οποία δημιουργούνται «δέντρα» ανάλογα με τις επιλογές μας στους χειρισμούς των μεταβλητών. Αναλυτικότερη περιγραφή θα γίνει στο κεφάλαιο 4. Επίσης σχηματισμός δέντρων χρησιμοποιείται όπως θα δούμε στο κεφάλαιο 3 για την παραγωγή νέων κανόνων που χρησιμοποιούνται στην εύρεση των μεταβλητών που χρειάζονται διόρθωση. Τέλος, δενδρική αναπαράσταση χρησιμοποιείται στο σχεδιασμό της έρευνας για το συνδυασμό των μεταβλητών που έχουν λογική σχέση ώστε να εξαχθούν οι κανόνες ορθότητας. (Petrakos 2004)

1.1 Στάδια μιας έρευνας και η εφαρμογή του ελέγχου ορθότητας

Ανεξάρτητα από την επιλογή της μεθόδου ελέγχου ορθότητας που θα χρησιμοποιήσουμε, πρέπει να τονίσουμε ότι ο έλεγχος των δεδομένων δεν αποτελεί μόνο την κύρια διαδικασία η οποία εφαρμόζεται μετά την συλλογή των δεδομένων αλλά υπεισέρχεται σε όλα τα στάδια της έρευνας.

Αρχικά πρέπει να αναφέρουμε ότι υπάρχουν τέσσερις διαφορετικοί τρόποι συλλογής δεδομένων που διακρίνονται από τον τρόπο επικοινωνίας του ερευνητή με τον ερωτώμενο: η έρευνα που διεξάγεται μέσω προσωπικής συνέντευξης του ερευνητή με τον ερωτώμενο (face-to-face), η τηλεφωνική έρευνα (telephone survey) κατά την οποία ο ερωτώμενος απαντά στις ερωτήσεις μέσω τηλεφώνου, η έρευνα που διεξάγεται με αποστολή των ερωτηματολογίων μέσω ταχυδρομείου (ή ηλεκτρονικού ταχυδρομείου: e-mail) όπου ο ερωτώμενος καλείται να απαντήσει το ερωτηματολόγιο σύμφωνα με σαφείς οδηγίες και να το επιστρέψει και τέλος οι web εφαρμογές (e-surveys). Βέβαια και στις τέσσερις παραπάνω περιπτώσεις το μέσο συμπλήρωσης του ερωτηματολογίου μπορεί να είναι είτε ένας ηλεκτρονικός υπολογιστής (CAI, computer-assisted interviewing) είτε η παραδοσιακή χειρωνακτική μέθοδος (PAPI, paper-and-pencil interviewing). Κρίνοντας τους αντικειμενικούς στόχους της έρευνας, το κόστος, το ποσό των δεδομένων που κρίνεται ικανοποιητικό για την έρευνα και την ποιότητα

αυτών αποφασίζεται η μέθοδος που θα ακολουθηθεί λαμβάνοντας υπόψη τα πλεονεκτημάτα και των μειονεκτημάτων της κάθε μιας. (Biemer and Lyberg 2003)

Η επεξεργασία των δεδομένων μιας έρευνας περικλείει μια σειρά λειτουργιών όπως τη συλλογή, την κωδικοποίηση, την καταχώρηση, τον έλεγχο, τη διόρθωση και την εξαγωγή συμπερασμάτων. Λειτουργίες όπως η καταχώρηση των δεδομένων, ο έλεγχος, η κωδικοποίηση μπορεί να γίνουν αυτόματα μέσω συστημάτων με στόχο τη μείωση του κόστους και του χρόνου. Η ανάπτυξη της τεχνολογίας βοήθησε στη βελτίωση των παραπάνω λειτουργιών με τη μείωση της ανάγκης για χειρωνακτική εργασία.

Ο στατιστικός έλεγχος ορθότητας έχει εφαρμογές σχεδόν σε κάθε στάδιο της επεξεργασίας των δεδομένων, από τον εμπειρικό έλεγχο που θα κάνει το άτομο που παίρνει τη συνέντευξη, τους ελέγχους κατά την καταχώρηση των δεδομένων, μέχρι τους ελέγχους ορθότητας για την εύρεση των λαθών και τη διόρθωσή τους. Η επεξεργασία αυτή έχει στόχο τη μετατροπή των δεδομένων σε καθαρά και διορθωμένα πεδία εγγραφών ώστε να χρησιμοποιηθούν σε στατιστική ανάλυση και παρουσίαση των αποτελεσμάτων. Οποιοσδήποτε και αν είναι ο τρόπος συλλογής των δεδομένων οι έλεγχοι σε κάθε στάδιο της επεξεργασίας τους είναι οι ακόλουθοι :

- ***Κατά τη διάρκεια της συνέντευξης*** ορισμένοι έλεγχοι μπορούν να γίνουν εμπειρικά από τον ερευνητή για την έγκυρες απαντήσεις, τη σχέση των μεταβλητών, την αποφυγή αναπάντητων πεδίων έτσι ώστε να περιοριστούν όσο το δυνατόν περισσότερο τα ασήμαντα λάθη που θα απαιτούσαν πολύπλοκη διαδικασία για τον εντοπισμό τους. Στην περίπτωση που τα ερωτηματολόγια γίνονται μέσω ηλεκτρονικού ταχυδρομείου σαφείς οδηγίες προς τους ερωτώμενους μπορούν να αποτρέψουν τέτοιου είδους λάθη. Τέλος στην περίπτωση που το ερωτηματολόγιο συμπληρώνεται από τον ερευνητή κατά τη διάρκεια της συνέντευξης μέσω ηλεκτρονικού υπολογιστή, ένα πρόσθετο πρόγραμμα ελέγχων εξετάζει την εγκυρότητα των απαντήσεων και προειδοποιεί τον ερευνητή για ελλειπείς τιμές.
- ***Πριν την καταχώρηση*** των δεδομένων γίνεται έλεγχος στα ερωτηματολόγια που συλλέχθηκαν στον οργανισμό (εταιρεία). Ο έλεγχος αυτός έχει να κάνει με διάκριση των ερωτηματολογίων σε δεκτά ή απορριπτέα ανάλογα με το αν περιέχουν πολλά κενά πεδία ή προφανή λάθη. Στην περίπτωση που ορισμένα ερωτηματολόγια κρίνονται απορριπτέα, μια λύση είναι να επικοινωνήσει μια ακόμη φορά ο ερευνητής με

τον ερωτώμενο για να επιβεβαιώσει ή να διορθώσει ορισμένες απαντήσεις. Αν αυτό δεν είναι εφικτό τα ερωτηματολόγια δεν λαμβάνονται υπόψη σε επόμενη διαδικασία.

- ***Κατά την καταχώρηση*** τα δεδομένα μετατρέπονται σε μορφή αναγνώσιμη από τον υπολογιστή. Η καταχώρηση μπορεί να γίνει είτε με πληκτρολόγηση ή με τη βοήθεια «έξυπνων» συστημάτων αναγνώρισης χαρακτήρων ή ακόμη και αναγνώρισης φωνής. Τελευταία έχουν δημιουργηθεί συστήματα που περιλαμβάνουν ηλεκτρονική εναλλαγή των δεδομένων και αποστολή μέσω Internet. Ο έλεγχος στο στάδιο αυτό, μπορεί να γίνει είτε σε επίπεδο μεταβλητής είτε σε επίπεδο εγγραφής. Η διαδικασία της καταχώρησης διακόπτεται ώσπου να ληφθούν κάποια μέτρα (αν δηλαδή η τιμή μπορεί να γίνει δεκτή ή να σημειωθεί (αμφίβολη τιμή) για μελέτη σε επόμενο στάδιο).
- ***Έπειτα από την καταχώρηση*** ο έλεγχος περιλαμβάνει συγκεκριμένες μεθοδολογίες που θα αναπτυχθούν αναλυτικά στα επόμενα κεφάλαια όπως μίκρο και μάκρο ελέγχους και μεθόδους διόρθωσης (ώστε να λάβουμε καθαρά δεδομένα) που γίνονται αυτόματα με περιορισμένη παρέμβαση των ειδικών αναλυτών. Τα δεδομένα μπορεί να ελέγχονται σε σχέση με προηγούμενες περιστάσεις της έρευνας ή και σε σχέση με ανάλογες έρευνες που πραγματοποιήθηκαν σε άλλη χώρα. Ο έλεγχος δηλαδή εκτός από την κύρια εφαρμογή του σε όλα τα δεδομένα μπορεί να έχει χρονική και γεωγραφική διάσταση.

Η πλήρης απόδοση των αποτελεσμάτων του ελέγχου μπορεί να συνοδεύεται από δείκτες ποιότητας, από παρουσίαση σε μορφή διαγραμμάτων ή πινάκων των λαθών καθε εγγραφής ώστε να διαπιστώνεται η αρχική και τελική ποιότητα των δεδομένων.

1.2. Στόχοι του στατιστικού ελέγχου ορθότητας

Συμπεραίνουμε ότι ο στατιστικός έλεγχος ορθότητας αποτελεί σημαντικό κομμάτι στην επεξεργασία των δεδομένων και εμπλέκεται σε κάθε στάδιο αυτής, ανιχνεύοντας λάθη και διορθώνοντάς τα. Το απόφθεγμα «Κάνε το σωστά από την πρώτη στιγμή» είναι βασικό στο να περιοριστούν λάθη και να αποφευχθεί επιπλέον κόπος και χρόνος για να διορθωθούν σφάλματα που θα αποτρέπονταν με γνώση, εμπειρία και επαγγελματισμό στα αρχικά στάδια της διαδικασίας. Στο σημείο αυτό, έχοντας ορίσει την έννοια του στατιστικού ελέγχου ορθότητας και τους κανόνες, που είναι τα βασικά εργαλεία για την εφαρμογή του, πρέπει να τονίσουμε τους στόχους του ελέγχου ορθότητας από τους οποίους προκύπτει και η σημαντικότητα της εφαρμογής του. Συγκεκριμένα ένας έλεγχος ορθότητας πρέπει να:

- ***Προβάλει πληροφορίες για την ποιότητα των δεδομένων:*** Από το αρχικό στάδιο της επεξεργασίας των δεδομένων, δηλαδή από τη συλλογή αυτών ο έλεγχος μπορεί να βελτιώσει σε σημαντικό βαθμό την εγκυρότητα και τη συνέπειά τους και να διορθώσει λάθη. Μετά την καταχώρηση των δεδομένων, όπου εκτελείται ο έλεγχος ορθότητας μέσα από μεθοδολογίες και με τη βοήθεια προγραμμάτων, εξασφαλίζεται η απόδοση «καθαρών» δεδομένων (ανάλογα με τη μέθοδο ελέγχου) ώστε να μπορούν να χρησιμοποιηθούν για την εξαγωγή αξιόπιστων στατιστικών συμπερασμάτων. Ο έλεγχος ορθότητας μπορεί να δώσει πληροφορίες για τη γνώση και τον επαγγελματισμό του ερευνητή αλλά και της ίδιας της διαδικασίας που χρησιμοποιήθηκε.
- ***Προετοιμάζει το «έδαφος» για μελλοντικές περιστάσεις της έρευνας:*** Επιθεωρώντας τα ερωτηματολόγια, την καταχώρηση των δεδομένων και τη στατιστική ανάλυση ο έλεγχος ορθότητας μπορεί να ανακαλύψει τις αιτίες ύπαρξης των λαθών. Με τον τρόπο αυτό μπορεί να προτείνει βελτιώσεις και να εξαλείψει λάθη που κρίνονται ουσιώδη για τη βελτίωση της έρευνας σε επόμενες περιστάσεις. Για παράδειγμα, στην περίπτωση περιοδικών ερευνών, η εύρεση των αιτίων που προκαλούν τα λάθη στην αρχική διεξαγωγή της έρευνας, είναι ουσιαστική για τη μείωση του χρονικού και οικονομικού κόστους στις επόμενες.
- ***«Τακτοποιεί» τα δεδομένα :*** Ο κυριότερος στόχος του ελέγχου ορθότητας είναι η «τακτοποίηση» των δεδομένων σε κάθε στάδιο της επεξεργασίας τους. Με τον όρο αυτό εννοούμε ότι προσπαθεί να διορθώσει όσο το δυνατόν περισσότερα λάθη υπάρχουν στα δεδομένα για να αγγίξουν το ιδανικό επίπεδο ποιότητας. (Biemer and Lyberg 2003)

Μιλώντας για ποιότητα των δεδομένων και των στατιστικών αποτελεσμάτων πρέπει να τονίσουμε ότι η ποιότητα αποτελεί μια πολυδιάστατη έννοια, οι διαστάσεις της οποίας είναι η σχετικότητα (relevance), η ακρίβεια εκτιμητών (accuracy), η επικαιρότητα (timeliness), η προσβασιμότητα (accessibility), η συγκρισιμότητα (comparability), η συνεκτικότητα (coherence) και τα μετά-δεδομένα (documentation) όπως έχουν ορισθεί από διάφορους στατιστικούς οργανισμούς όπως η EUROSTAT, Στατιστική Υπηρεσία του Καναδά, της Νέας Ζηλανδίας, της Σουηδίας και πολλές άλλες (Biemer and Lyberg 2003). Όμως πολλοί οργανισμοί και Στατιστικές Υπηρεσίες συνδέουν άμεσα την ποιότητα με την ακρίβεια των στατιστικών αποτελεσμάτων παρά το γεγονός ότι αποτελεί μόνο μια διάστασή της. Εξάλλου

δεν θα είχε αξία να στηριζόμαστε σε ακριβή αποτελέσματα όταν αυτά δεν χαρακτηρίζονται από τη συνέπεια, τη συνάφεια, τη συγκρισιμότητα, την επικαιρότητα και άλλες παραμέτρους που αποτελούν διαστάσεις τη ποιότητας.

Όπως αναφέραμε προηγούμενα, ένας από τους στόχους του στατιστικού ελέγχου ορθότητας είναι η βελτίωση της ποιότητας, όμως η διαδικασία του ελέγχου συνδέεται και με άλλες παραμέτρους όπως το κόστος, που προσδιορίζει κατά ένα βαθμό τη διάσταση του ελέγχου στα πλαίσια της επεξεργασίας των δεδομένων. Ο έλεγχος ορθότητας είναι μια διαδικασία που απαιτεί κόστος χρηματικό και χρονικό. Έχει εκτιμηθεί ότι από τον προϋπολογισμό μιας έρευνας το 20% δαπανάται στο στατιστικό έλεγχο όταν πρόκειται για δημογραφικές έρευνες, ενώ αγγίζει το 40% όταν πρόκειται για έρευνες επιχειρήσεων (Granquist and Kovar 1997). Από την άλλη πλευρά, υπέρμετρος έλεγχος ορθότητας (over-editing) ο οποίος τελικά δεν συνεισφέρει θετικά στα αποτελέσματα μπορεί να καθυστερήσει τη δημοσίευση αποτελεσμάτων, μειώνοντας τη χρονική τους συνάφεια. Είναι κοινή ομολογία ότι η σωστή διάσταση του ελέγχου πρέπει να βασίζεται σε στρατηγικές που λαμβάνουν υπόψη το χρονικό και χρηματικό κόστος. Ένας τρόπος αντιμετώπισης του προβλήματος είναι η εφαρμογή του επιλεκτικού ελέγχου που συντελεί στη μείωση χρόνου και χρημάτων χωρίς να υποβιβάζει το επίπεδο της ακρίβειας των στατιστικών αποτελεσμάτων.

Διαπιστώνεται λοιπόν, ότι ο στατιστικός έλεγχος ορθότητας μπορεί να καταναλώσει μεγάλο ποσοστό του προϋπολογισμού μιας έρευνας. Για να μεγιστοποιήσουμε τα οφέλη του, πρέπει να λάβουμε υπόψη τα στάδια στα οποία εφαρμόζεται, κατά την διεξαγωγή μιας έρευνας ώστε να βελτιωθεί η όλη διαδικασία. Συγκεκριμένα, να βελτιωθούν τα ερωτηματολόγια, η κατάρτιση των ερευνητών καθώς και η κύρια διαδικασία του ελέγχου με τη χρήση αποτελεσματικών μεθοδολογιών και αξιόπιστων αυτοματοποιημένων συστημάτων (Biemer and Lyberg 2003).

Όσον αφορά την πραγματική διάσταση των γεγονότων, αξίζει να σημειώσουμε ότι πριν από μερικές δεκαετίες υπήρχε μια αμφίβολη εικόνα για τις στατιστικές έρευνες και την ποιότητά τους. Σε αυτό συνέβαλε και η τεχνολογία που τα χρόνια αυτά, δεν ήταν ιδιαίτερα αναπτυγμένη και εμπόδιζε τόσο την έγκαιρη δημοσίευση των αποτελεσμάτων όσο και την καλή παρουσίαση αυτών. Βέβαια οι οργανισμοί χωρίς να μπορούν να τα αποτρέψουν, περιλάμβαναν στο προϋπολογισμό της έρευνας και τα έξοδα του ελέγχου ποιότητας και την κατάρτιση στο θέμα της ακρίβειας σε ορισμένες έρευνες. Σήμερα η κατάσταση είναι πολύ διαφορετική. Η ραγδαία ανάπτυξη της τεχνολογίας και η ανάγκη αντιμετώπισης του

ανταγωνισμού επηρέασε τις στατιστικές έρευνες θετικά. Οι οργανισμοί και οι Στατιστικές Υπηρεσίες μπορούν να διεξάγουν έρευνες και να δημοσιεύσουν τα αποτελέσματα σε εύλογο χρονικό διάστημα ώστε να είναι επίκαιρα και συγκρίσιμα, καθώς και να εξασφαλίζουν την ποιότητα αυτών εφόσον ακολουθούν κανόνες που ορίζουν μεγάλοι στατιστικοί οργανισμοί όπως η EUROSTAT η οποία όπως αναφέραμε προηγούμενα, έχει εκδώσει επτά διαστάσεις της ποιότητας που πρέπει να διέπουν τα στατιστικά «προϊόντα» (Biemer and Lyberg 2003). Από την άλλη, ο στατιστικός έλεγχος ορθότητας είναι πλέον γεγονός και εφαρμόζεται από αξιόπιστους οργανισμούς και υπηρεσίες. Η ανάπτυξη της τεχνολογίας, η εύρεση νέων μεθοδολογιών και υπολογιστικών συστημάτων κάνει τον έλεγχο ορθότητας προσιτό στην εφαρμογή του και αποτελεσματικό εργαλείο στη βελτίωση της ποιότητας.

Η δομή της εργασίας είναι η ακόλουθη: Στο δεύτερο κεφάλαιο παρουσιάζεται αναλυτικά η πρωτοποριακή και ολοκληρωμένη μέθοδος ελέγχου και διόρθωσης Fellegi-Holt (1976) και η εφαρμογή του στο αυτοματοποιημένο σύστημα που ανέπτυξαν το 1976. Στο τρίτο κεφάλαιο παρουσιάζονται μέθοδοι βασισμένοι στη θεωρία των Fellegi-Holt οι οποίοι βελτιώνουν τη διαδικασία για τη λύση του προβλήματος εντοπισμού λαθών, καθώς παρουσιάζονται και συστήματα όπως το DISCRETE για κατηγορικά δεδομένα και το SPEER για αριθμητικά.Το τέταρτο κεφάλαιο αποτελείται από μεθόδους και προσεγγίσεις της διαδικασίας ελέγχου και διόρθωσης που αφορούν όλα τα είδη των δεδομένων και ακολουθούν νέες αρχές και κανόνες ανεξάρτητες των Fellegi-Holt. Τέλος το πέμπτο κεφάλαιο δίνει μια γενική εικόνα αναπτύσσοντας γενικά συμπεράσματα και σύγκριση των μεθοδολογιών που αναλύθηκαν σε προηγούμενα κεφάλαια.

ΚΕΦΑΛΑΙΟ 2

Μεθοδολογία Fellegi-Holt

2.1. Εισαγωγή

Ο στατιστικός έλεγχος ορθότητας εισερχόμενων δεδομένων είναι η απαραίτητη διαδικασία στην οποία πρέπει να υπόκεινται τα δεδομένα προκειμένου να εξαλειφθούν λάθη και ασυνέπειες που τα χαρακτηρίζουν, ώστε να μπορούν να χρησιμοποιηθούν για περαιτέρω στατιστική ανάλυση για τη εξαγωγή αξιόπιστων στατιστικών συμπερασμάτων. Κάθε φορά που διεξάγεται μια έρευνα ο όγκος των λαμβανόμενων δεδομένων είναι αρκετά μεγάλος. Αναπόφευκτα τα δεδομένα περιλαμβάνουν ελλειπείς ή λανθασμένες τιμές που μπορεί να οφείλονται είτε στον ερευνητή είτε στον ερωτώμενο ή στην ίδια τη διαδικασία. Η αποθήκευση των δεδομένων γίνεται σε ηλεκτρονικές βάσεις οι οποίες είναι εύκολα προσβάσιμες, αλλά είναι δύσκολο για τους ερευνητές να ελέγξουν τις εισόδους για τα προαναφερθέντα λάθη, αφού μπορεί να παραλείψουν ορισμένους ελέγχους ή μπορεί να μην αναγνωρίσουν αντικρουόμενους ελέγχους. Για αυτό το λόγο αναλαμβάνουν οι ειδικοί που είναι γνώστες της έρευνας να ορίσουν αρχικά ένα σύνολο **κανόνων ορθότητας** (edit rules). Το πρόβλημα όμως δεν βρίσκει λύση λόγω του μεγάλου όγκου δεδομένων που αναλαμβάνουν να εξετάσουν. Έτσι έχουν δημιουργηθεί αυτοματοποιημένα συστήματα ελέγχου ορθότητας δεδομένων με πρωτοπόρο αυτό των Fellegi-Holt (FH, 1976). Η μεθοδολογία FH προάγει ένα τρόπο αυτόματης παραγωγής κανόνων ορθότητας για τον έλεγχο των δεδομένων, εντοπίζει τα λανθασμένα δεδομένα και τελικά τα διορθώνει. Η μέθοδος επικεντρώνεται σε κατηγορικά δεδομένα, αν και πολλά θεωρητικά αποτελέσματα βρίσκουν εφαρμογή και σε ποσοτικά δεδομένα.

2.2 Ορολογία

Για τη διεξαγωγή μιας έρευνας, το επόμενο στάδιο έπειτα από το σχεδιασμό και την επιλογή της δειγματοληπτικής μεθόδου, αποτελεί η συλλογή των δεδομένων και ο έλεγχος ορθότητάς τους. Οι πληροφορίες λαμβάνονται σε μορφή ερωτηματολογίων που το σύνολό τους αποτελεί το σύνολο των δεδομένων. Κάθε ερωτηματολόγιο το οποίο περιλαμβάνει τις απαντήσεις ενός ερωτώμενου αποτελεί μια **εγγραφή** (record). Κάθε εγγραφή αποτελείται από πεπερασμένο αριθμό ερωτήσεων: **μεταβλητών ή πεδίων** (field). Κάθε μεταβλητή έχει

προκαθορισμένο πεδίο ορισμού, πεπερασμένο αν η μεταβλητή είναι κατηγορική ή μη πεπερασμένο (π.χ. ευθεία των πραγματικών αριθμών) αν η μεταβλητή είναι συνεχής, μέσα στο οποίο θα πρέπει να κυμαίνονται οι απαντήσεις.

Ο κύριος σκοπός του στατιστικού ελέγχου ορθότητας είναι πρώτον να εξεταστεί αν κάθε μεταβλητή από κάθε εγγραφή περιέχει μη λανθασμένη τιμή δηλαδή τιμή που να ανήκει στο πεδίο ορισμού της μεταβλητής. Δεύτερον, να εξεταστεί αν οι τιμές συγκεκριμένων συνδυασμών πεδίων είναι συμβατές. Ένας κανόνας ορθότητας εκφράζει την κρίση των ειδικών για ένα συνδυασμό τιμών σε συγκεκριμένες μεταβλητές της έρευνας ο οποίος είναι απίθανο να συμβεί. Οι κανόνες ορθότητας που αφορούν ποιοτικά δεδομένα ονομάζονται **λογικοί κανόνες** (logical edits) ενώ αυτοί που αναφέρονται σε ποσοτικά ονομάζονται **αριθμητικοί κανόνες** ορθότητας (arithmetic edits).

2.2.1 Λογικοί κανόνες ορθότητας

Αρχικά θα αναφερθούμε σε κανόνες ορθότητας που αφορούν κατηγορικές μεταβλητές. Υποθέτουμε ότι σε μια έρευνα η κάθε εγγραφή, περιλαμβάνει *n* μεταβλητές ή πεδία. Έστω $y = (y_1, y_2, \ldots y_n)$ είναι μια εγγραφή που αποτελείται από κατηγορικές μεταβλητές και y_i η τιμή που λάβαμε στο *i* πεδίο. Ορίζουμε με R_i το σύνολο των επιτρεπόμενων τιμών για το πεδίο *i* το οποίο είναι πεπερασμένο σύνολο μεγέθους κ_i. Τότε το καρτεσιανό γινόμενο των πεδίων τιμών των μεταβλητών αποτελεί τον δειγματικό χώρο των εγγραφών $D=\prod_{i=1}^{n} R_i$. Ένας κανόνας ορθότητας είναι ένα υποσύνολο του παραπάνω καρτεσιανού γινομένου και είναι της μορφής

$$E^i = \prod_j E_{ij}$$

η οποία λέγεται **κανονική μορφή** (normal form) του κανόνα, όπου E_{ij} είναι το σύνολο των τιμών που παίρνει το πεδίο *j* στον κανόνα *i* . Αναλύοντας το γινόμενο, ένας κανόνας στην κανονική του μορφή γράφεται $E^i : E_{i1} \times E_{i2} \times \ldots . \times E_{in}$, όπου $j=1,2,\ldots,n$. Κάθε κανόνας E^i ορίζει ένα σύνολο $P(E^i)$ που είναι όπως αναφέραμε το καρτεσιανό γινόμενο των συνόλων των τιμών των πεδίων και αποτελεί υποσύνολο του δειγματικού χώρου δηλαδή $P(E^i) \subset D$.

Ο κανόνας E^i λέγεται ότι **περιέχει** το πεδίο *j* ή το πεδίο *j* **υπάρχει** στον κανόνα E^i αν ισχύει το εξής: $E_{ij} \subset R_i$. Αν στον κανόνα E^i για το πεδίο *j* ισχύει $E_{ij} = R_i$ τότε λέμε ότι ο

κανόνας δεν περιέχει το πεδίο *j*. Επίσης ένας κανόνας E^i λέγεται **μέγιστος** (maximal) όταν δεν υπάρχει άλλος κανόνας E τέτοιος ώστε $P(E^i) \subseteq P(E)$. Διαφορετικά ο E^i ονομάζεται **περιττός** (redundant), γιατί ο E έχει τις ίδιες και περισσότερες τιμές από τον E^i. Συνηθίζεται η έκφραση για τον περιττό κανόνα ότι **κυριαρχείται** (dominated) από τον *E*, ή ότι ο *E* **επικρατεί** ή **υπερισχύει** (dominates) επί του E^i. Τέλος, η εγγραφή *y* της οποίας οι τιμές ανήκουν στο σύνολο $P(E^i)$ που ορίζεται από τον κανόνα E^i, λέγεται ότι **χάνει** (fail) τον κανόνα, δηλαδή η εγγραφή είναι λάθος. Αν οι τιμές της εγγραφής y δεν επαληθεύουν τις τιμές του συνόλου $P(E^i)$ λέμε ότι η εγγραφή **ικανοποιεί** ή **περνάει** (satisfy) τον κανόνα E^i.

Παράδειγμα 1 *(FH 1976)*

Έστω ότι σε μια έρευνα το ερωτηματολόγιο περιλαμβάνει τις παρακάτω τρεις μεταβλητές :

- Ηλικία, με πιθανές τιμές : 0-14 (1), 15+ (2)
- Οικογενειακή κατάσταση (O.K), με πιθανές τιμές : Άγαμος (1), Παντρεμένος(2), Διαζευγμένος(3) , Χήρος /Χήρα(4) , Σε Διάσταση (5)
- Σχέση με τον κύριο του νοικοκυριού, με πιθανές τιμές : Νοικοκύρης (1) , Σύζυγος(2) , Άλλο(3)

Οι τιμές στις παρενθέσεις είναι οι κώδικες για κάθε προβλεπόμενη τιμή της μεταβλητής. Η κάθε εγγραφή του ερωτηματολογίου περιέχει 3 πεδία τα εξής:

Πίνακας 1

Πεδία	Όνομα	Πεδίο τιμών	Μέγεθος
1	Ηλικία	R_1={1,2}	κ_1=2
2	O.K	R_2={1,2,3,4,5}	κ_2=5
3	Σχέση με Νοικοκύρη	R_3={1,2,3}	κ_3=3

Ο δειγματικός χώρος είναι $D = \{1,2\} \times \{1,2,3,4,5\} \times \{1,2,3\}$

Έστω ότι οι ερευνητές έχουν ορίσει τους παρακάτω λογικούς κανόνες ορθότητας :

I. (Ηλικία=0-14)∩(O.K=Κάποτε παντρεμένος/η)

II. (O.K=Οχι τώρα παντρεμένος/η)∩(Σχέση με κύριο νοικ.= Σύζυγος)

Οι παραπάνω κανόνες μετατρέπονται στην κανονική τους μορφή :

$E^1 = \{1\} \times \{2,3,4,5\} \times R_3$

$E^2 = R_1 \times \{1,3,4,5\} \times \{2\}$

Λαμβάνοντας μια εγγραφή από το σύνολο των δεδομένων την *y={1,2,2}* παρατηρούμε ότι χάνει τον πρώτο κανόνα αλλά περνάει τον δεύτερο. Αν στην ίδια έρευνα υπήρχε και ένας κανόνας της μορφής $E^k = \{1,2,3\} \times \{2,3,4,5\} \times R_3$ τότε αυτός θα υπερίσχυε του E^1 , και αν δεν υπάρχει άλλος κανόνας ο E^k θεωρείται μέγιστος.

2.2.2 Αριθμητικοί κανόνες ορθότητας

Οι αριθμητικοί κανόνες είναι γραμμικές εκφράσεις των μεταβλητών που περιέχουν. Έστω μια εγγραφή $x = (x_1, x_2, \ldots, x_n)$ μιας έρευνας με n πεδία και ένας αριθμητικός κανόνας ορθότητας της μορφής :

$$f(x_1, x_2, \ldots, x_n) \geq 0$$

όπου f είναι μια πολυγραμμική συνάρτηση και x_i, $i=1,2,\ldots,n$ είναι τα n πεδία της έρευνας που παίρνουν τιμές στα X_i.

Αν υποθέσουμε ότι το πεδίο x_i έχει συντελεστή μη μηδενικό (δηλαδή το πεδίο x_i υπάρχει στον κανόνα) τότε η παραπάνω σχέση γράφεται : $x_1 \geq L(x_2, x_3, \ldots, x_n)$ όπου η L είναι επίσης μια πολυγραμμική συνάρτηση. Τότε ο αριθμητικός κανόνας στην κανονική του μορφή, γράφεται ως εξής:

$$\{x_1 : x_1 \geq L(x_2, x_3, \ldots, x_n)\} \times \{x_2\} \times \{x_3\} \times \ldots \times \{x_n\}$$

Κάθε εγγραφή για την οποία ισχύει η παραπάνω σχέση σημαίνει ότι χρειάζεται διόρθωση γιατί χάνει τον κανόνα. Αν ισχύει η ισότητα στη παραπάνω σχέση τότε ο κανόνας μπορεί να χωριστεί σε δυο νέους όπου ο ένας από αυτούς πρέπει να πολλαπλασιαστεί με -1 για να έχουν οι ανισότητες την ίδια φορά.

Για παράδειγμα σε μια έρευνα που ασχολείται με γεωργικές καλλιέργειες με μεταβλητές x_1 = καλλιεργήσιμη έκταση, x_2 = ακαλλιέργητη έκταση, x_3 = η στρεμματική έκταση που κατέχει ο γαιοκτήμονας, είναι λογικό να περιμένουμε σε κάθε εγγραφή να ισχύει $x_3=x_1+x_2$. Προκύπτει ο γραμμικός αριθμητικός κανόνας που στην κανονική του μορφή θα γράφεται ως εξής:

$$E = \{x_1\} \times \{x_2\} \times \{x_3 : x_3 \neq x_1 + x_2\}$$

Οι πιο συνηθισμένες μορφές αριθμητικών κανόνων ορθότητας είναι οι γραμμικές όπως προαναφέρθηκε, οι **κανόνες πηλίκων** (ratio edits) και οι **κανόνες ισορροπίας** (balance edits).

Οι **κανόνες πηλίκων** είναι της μορφής

$$L_{ij} < V_i \Big/ V_j < U_{ij}$$

όπου V_i, $i= 1,2,...,N$ είναι οι N μεταβλητές (πεδία) της εγγραφής και τα L_{ij}, U_{ij} είναι το κάτω και πάνω όριο αντίστοιχα του κανόνα τα οποία προσδιορίζονται από τους ειδικούς αναλυτές. Μια εφαρμογή τέτοιου είδους κανόνα θα ήταν η ολική παραγωγή μιας βιομηχανίας προς τις εργατοώρες πρέπει να κυμαίνεται μεταξύ δυο σταθερών ποσοτήτων. (L.Draper and W.Winkler 1997)

Οι **κανόνες ισορροπίας** είναι της μορφής

$$\sum_{i \in S} V_i - V_j = 0$$

όπου V_i, $i= 1,2,...,N$ είναι οι N μεταβλητές (πεδία) της εγγραφής και S είναι υποσύνολο του N , $j \notin S$. Μια εφαρμογή αυτού του κανόνα θα ήταν το άθροισμα των μισθών των ατόμων του νοικοκυριού πρέπει να ισούται με το σύνολο των χρημάτων που διαχειρίζεται το νοικοκυριό που δεν έχει επιπλέον εισοδήματα. (L.Draper, W.Winkler 1997)

2.3 Παραγωγή του πλήρους συνόλου κανόνων

Αρχικά οι ειδικοί αναλυτές μιας έρευνας ορίζουν το σύνολο των **ρητά ορισμένων κανόνων** (explicit edits) που συμβολίζεται συνήθως E^0. Με τους ρητούς κανόνες που έχουν ορίσει οι ειδικοί αναλυτές, έχουμε τη δυνατότητα να εντοπίσουμε ποιες εγγραφές χρειάζονται διόρθωση αλλά όχι ποια συγκεκριμένα πεδία της εγγραφής είναι λανθασμένα. Συνήθως οι ρητά ορισμένοι κανόνες υποννοούν νέους κανόνες που είναι δυσδιάκριτοι αρχικά από τους ερευνητές, αλλά υπάρχουν και δίνουν πληροφορίες για την ορθότητα των εγγραφών. Αυτοί οι κανόνες ονομάζονται **έμμεσοι** (implicit edits). Το σύνολο των ρητά ορισμένων κανόνων και το σύνολο των έμμεσων κανόνων αποτελεί το **πλήρες συνολο κανόνων** (complete set) και ορίζεται με E^c.

Όπως ορίζεται από τους FH (1976) το πλήρες σύνολο των κανόνων αποτελείται από τους ρητά ορισμένους κανόνες και τους κατά βάση νέους έμμεσους κανόνες τους οποίους ορίζουμε στη συνέχεια. Η παραγωγή του πλήρους συνόλου κανόνων αποτελεί την πιο χρονοβόρα διαδικασία της μεθόδου Fellegi-Holt και αποτέλεσε το αντικείμενο μελέτης, με στόχο την συντόμευση και αποφυγή άσκοπων (περιττών) ενεργειών για την ολοκλήρωσή της.

2.3.1 Παραγωγή του πλήρους συνόλου λογικών κανόνων ορθότητας .(Fellegi- Holt 1976)

Σύμφωνα με τη θεωρία των Fellegi – Holt κάθε κανόνας γράφεται στην κανονική του μορφή. Η παραγωγή ενός έμμεσου κανόνα, δεδομένων των ρητά ορισμένων κανόνων γίνεται με βάση τα παρακάτω:

Έστω ότι κάθε εγγραφή αποτελείται από *n* κατηγορικές μεταβλητές και R_j όπου *j=1,2,...,n* είναι τα σύνολα των πιθανών τιμών των μεταβλητών. Έστω επίσης ένα σύνολο *s* λογικών κανόνων που γράφονται στην κανονική τους μορφή δηλαδή $E^i : E_{i1} \times E_{i2} \times .... \times E_{in}$ με *i = 1,2,,...s*.

Λήμμα 1 (FH 1976) Από το παραπάνω σύνολο των *s* κανόνων παράγεται ο ακόλουθος κανόνας $E^* : F_1 \times F_2 \times ... \times F_r \times ... \times F_n$.

- Για ένα συγκεκριμένο πεδίο *r* ισχύει $F_r = \bigcup_{l=1}^{s} E_{il} \neq \text{ø}$
- Και για τα υπόλοιπα πεδία ισχύει $F_i = \bigcap_{l=1}^{s} E_{il} \neq \text{ø}$ όπου $i = 1,2,...,n \ , i \neq r$

Το πεδίο *r* ονομάζεται **γεννήτορας πεδίο** (generating field) και οι *s* λογικοί κανόνες ονομάζονται **συμβάλλοντες κανόνες** (contributing edits).

Αν ισχύει $F_r = R_r$ (ο έμμεσος κανόνας δεν περιέχει το γεννήτορα) ενώ $E_{ir} \subset R_r$ με *i = 1,2,...,s* δηλαδή οι συμβάλλοντες κανόνες περιέχουν το γεννήτορα τότε ο έμμεσος κανόνας ονομάζεται «**κατά βάση νέος**» κανόνας (essentially new edit rule). Με τον παραπάνω τρόπο μπορούμε να παράγουμε όλους τους έμμεσους κανόνες χρησιμοποιώντας ως συμβάλλοντες κανόνες όλους τους συνδυασμούς των κανόνων ανα δυο, ανα τρεις, μέχρι να συμβούν όλοι οι συνδυασμοί κανόνων και ορίζοντας ως γεννήτορα πεδίο διαφορετικό κάθε φορά.

Συνέχεια στο παράδειγμα 1(FH 1976)

Χρησιμοποιώντας το παράδειγμα 1 της προηγούμενης παραγράφου *2.2.1* είδαμε ότι οι ρητά ορισμένοι κανόνες είναι :

$E^1 = \{1\} \times \{2,3,4,5\} \times R_3$

$E^2 = R_1 \times \{1,3,4,5\} \times \{2\}$

Αν θεωρήσουμε ως γεννήτορα το πεδίο Οικογενειακή κατάσταση τότε ο έμμεσος κανόνας παράγεται ως εξής:

Για τον γεννήτορα πεδίο

$F_2 = \{2,3,4,5\} \cup \{1,3,4,5\} = R_2$

Για τα υπόλοιπα πεδία

$F_1 = \{1\} \cap R_1 = \{1\}$

$F_3 = R_3 \cap \{2\} = \{2\}$

Τελικά ο έμμεσος κανόνας είναι

$E^* = \{1\} \times R_2 \times \{2\}$ = (Ηλικία = 0-14)∩(O.K = Κάθε τιμή)∩(Σχέση με νοικ. = Σύζυγος)

Παρατηρώ ότι ο γεννήτορας πεδίο (Οικογενειακή κατάσταση) δεν περιέχεται στο έμμεσο κανόνα ενώ περιέχεται στους συμβάλλοντες κανόνες. Αυτό σημαίνει ότι ο κανόνας E^* είναι «κατά βάση νέος» κανόνας. Στο παράδειγμά μας μπορούμε να παράγουμε ακόμη δύο έμμεσους κανόνες, παίρνοντας ως γεννήτορα πεδίο τις μεταβλητές Ηλικία και τη Σχέση με το νοικοκύρη αντίστοιχα στην παραγωγή του κάθε ένα. Όμως οι νέοι έμμεσοι κανόνες που παράγονται με γεννήτορες πεδία την Ηλικία και τη Σχέση με το νοικοκύρη αντίστοιχα δεν είναι κατά βάση νέοι. Τελικά οι κανόνες E^1, E^2 και E^* αποτελούν το πλήρες σύνολο κανόνων όπως αυτό ορίσθηκε από τους Fellegi-Holt.

Βέβαια η διαδικασία της παραγωγής των έμμεσων κανόνων μπορεί να γίνει αρκετά δύσκολη και χρονοβόρα είτε όταν το σύνολο των ρητών κανόνων είναι αρκετά μεγάλο είτε όταν έχουμε να επεξεργαστούμε δεδομένα με αρκετές μεταβλητές.

2.3.2 Παραγωγή του πλήρους συνόλου αριθμητικών κανόνων ορθότητας. (FH, 1976)

Η παραγωγή των έμμεσων αριθμητικών κανόνων πραγματοποιείται με διαφορετικό τρόπο από ότι η παραγωγή των έμμεσων λογικών και καταλήγει σε ένα σύνολο κανόνων οι οποίοι είναι γραμμικοί συνδυασμοί των πεδίων που περιέχουν. Έστω ο αριθμητικός κανόνας $E_r : f(x_1, x_2, ..., x_n) \geq 0$ ο οποίος χάνεται για κάθε εγγραφή που επαληθεύει την ανισότητα. Ο κανόνας προσδιορίζεται από *n+1* συντελεστές και έναν δείκτη που δηλώνει τη φορά της ανισότητας.

Κανόνας	Σταθερά	Πεδίο 1	Πεδίο 2	.....	Πεδίο n	Δείκτης
E_r	a_0^r	a_1^r	a_2^r	……..	a_n^r	d^r

Για κάθε πεδίο που δεν περιέχεται στον κανόνα ο συντελεστής του είναι 0.

Όσο για το δείκτη δ^r ισχύει το εξής :

$$d^r = \begin{cases} 1, & \text{αν } f > 0 \\ 0, & \text{αν } f \le 0 \end{cases}$$

Η παραγωγή του έμμεσου αριθμητικού κανόνα γίνεται με βάση το επόμενο θεώρημα (FH, 1976)

Θεώρημα 1: Δυο αριθμητικοί κανόνες E^r και E^s παράγουν τον κανόνα E^t χρησιμοποιώντας ως γεννήτορα το πεδίο i αν και μόνο αν οι συντελεστές του πεδίου i στους κανόνες E^r και E^s είναι μη μηδενικοί και αντίθετου προσήμου. Οι συντελεστές του κανόνα που παράγεται και ο δείκτης δίνονται αντίστοιχα από τις σχέσεις:

$a_k^t = a_k^s a_i^r - a_k^r a_i^s$ όπου $\kappa = 0,1,2...,n,$ και $d^t = d^r d^s$ με $a_i^r > 0$ και $a_i^s < 0$

Η έννοια των «κατά βάση νέων» κανόνων ορθότητας ισχύει και στην περίπτωση των αριθμητικών κανόνων. Με την διαδικασία που προηγήθηκε μπορούμε να παράγουμε έμμεσους κανόνες συνδυάζοντας τους ρητά ορισμένους κανόνες και ορίζοντας ως γεννήτορα διαφορετικό πεδίο κάθε φορά, έτσι ώστε να αποκτήσουμε το πλήρες σύνολο αριθμητικών κανόνων.

2.3.3 Ασυνεπές σύνολο κανόνων

Με την παραγωγή του πλήρους συνόλου των κανόνων ορθότητας μπορεί να αποκαλυφθεί οποιαδήποτε ασυνέπεια υπάρχει στους ρητά ορισμένους κανόνες. Προφανώς ασυνέπεια δεν εντοπίζεται στη δομή κάθε κανόνα ορθότητας ξεχωριστά, αλλά στο συνδυασμό τους. Ένα ασυνεπές σύνολο κανόνων περιέχει αντικουόμενους κανόνες με συνέπεια καμιά εγγραφή να μην το ικανοποιεί. Διαπιστώνουμε ότι ένα σύνολο κανόνων είναι **ασυνεπές** (inconsistent), αν παράγεται κανόνας ο οποίος περιέχει ένα μόνο πεδίο. Αν ισχύει αυτό, σημαίνει ότι το συγκεκριμένο πεδίο παίρνει τιμές που οδηγούν σε λάθος εγγραφές και στην ουσία δεν θα έπρεπε να ανήκουν στο πεδίο τιμών της συγκεκριμένης μεταβλητής. Η ασυνέπεια ενός συνόλου κανόνων βεβαιώνεται όταν παραχθεί ένας έμμεσος κανόνας της μορφής :

$$E^* = R_1 \times ... R_{i-1} \times F_i \times R_{i+1} ... \times R_n \quad \text{όπου } F_i \subset R_i$$

Ο έλεγχος της συνέπειας των κανόνων γίνεται κατά την διαδικασία παραγωγής των, πριν δηλαδή αρχίσει ο έλεγχος κάθε εγγραφής αν τους ικανοποιεί. Σε ένα συνεπές σύνολο κανόνων υπάρχει τουλάχιστον μια εγγραφή που ικανοποιεί όλους τους κανόνες.

2.3.4 Διευκολύνσεις στην διαδικασία παραγωγής του πλήρους συνόλου κανόνων

Προφανώς, η παραγωγή του πλήρους συνόλου των κανόνων είναι μια χρονοβόρα διαδικασία στην οποία αποδίδεται το τεράστιο υπολογιστικό κόστος που παρατηρείται κατά την εφαρμογή της μεθόδου Fellegi–Holt (1976) σε πραγματικά δεδομένα. Το πλήθος των συνδυασμών των κανόνων για την παραγωγή έμμεσων κανόνων σε συνδυασμό με την επιλογή διαφορετικών γεννητόρων πεδίων θα ήταν αρκετά μεγάλο ακόμη και στην περίπτωση μικρού μεγέθους ερευνών. Οι Fellegi-Holt (1976) απλοποίησαν την διαδικασία προτείνοντας τις ακόλουθες προτάσεις που μειώνουν κατά πολύ το πλήθος των συνδυασμών που πρέπει να γίνουν για να παράγουμε όλους τους κατά βάση νέους κανόνες.

- Κάθε σύνολο κανόνων μπορεί να παράγει έναν «κατά βάση νέο» κανόνα μόνο όταν έχει τουλάχιστον ένα κοινό πεδίο που εμπεριέχεται σε κάθε ένα από αυτούς.
- Συνδυάζοντας ένα έμμεσο κανόνα με το σύνολο των κανόνων που τον παρήγαγε, ο νέος κανόνας που θα παραχθεί δεν είναι «κατά βάση νέος» κανόνας.
- Αν ένα σύνολο κανόνων με συγκεκριμένο πεδίο ως γεννήτορα παρήγαγε ένα «κατά βάση νέο» κανόνα, τότε ένα υπερσύνολο των προηγούμενων κανόνων με τον ίδιο γεννήτορα δεν παράγει «κατά βάση νέο» κανόνα.

Σύμφωνα με τις παραπάνω προτάσεις μπορούν να παραχθούν όλοι οι κατά βάση νέοι κανόνες και παράλληλα να εξασφαλιστεί η συνέπεια του συνόλου των κανόνων.

2.4 Συνοπτική περιγραφή της μεθόδου Fellegi – Holt (1976)

Η πρωτοποριακή μεθοδολογία των Fellegi- Holt όπως παρουσιάστηκε το 1976, αποτελεί μια πλήρη διαδικασία που πρέπει να εφαρμόζεται σε κάθε έρευνα πριν τη στατιστική επεξεργασία των δεδομένων για την εξαγωγή αξιόπιστων στατιστικών αποτελεσμάτων. Η μεθοδολογία χωρίς να λαμβάνει υπόψη το υπολογιστικό κόστος ακολουθεί τα τρία επόμενα βήματα:

I. Οι ρητά ορισμένοι κανόνες ορίζονται από τους ειδικούς που γνωρίζουν την έρευνα, τα δεδομένα της οποίας πρέπει να ελεγχθούν. Δεδομένου λοιπόν, του συνόλου των ρητά ορισμένων κανόνων παράγονται οι έμμεσοι κανόνες σύμφωνα με τη διαδικασία που περιγράφηκε προηγούμενα. Εφόσον αποκτήσουμε το πλήρες σύνολο των κανόνων αυτό ελέγχεται για το αν είναι συνεπές. Στη περίπτωση που έχουμε ένα συνεπές σύνολο κανόνων αυτό χρησιμοποιείται για τον έλεγχο κάθε εγγραφής στο αν ικανοποιεί όλους τους κανόνες ή όχι.

II. Έπειτα από τον παραπάνω έλεγχο εντοπίζονται οι κανόνες που χάνονται από την κάθε εγγραφή ξεχωριστά. Η εύρεση του μικρότερου συνόλου πεδίων που ευθύνονται για τις χαμένες εγγραφές αποτελεί το δεύτερο βήμα της μεθόδου που ονομάζεται **πρόβλημα εντοπισμού λαθών** (error localization).

III. Τέλος έχοντας στη διάθεσή μας τη λύση του παραπάνω προβλήματος δηλαδή το μικρότερο αριθμό πεδίων που ευθύνονται για τις λάθος εγγραφές απομένει να τα **διορθώσουμε** (impute) έτσι ώστε οι διορθωμένες εγγραφές να ικανοποιούν όλους τους κανόνες.

Όσον αφορά την διόρθωση των δεδομένων η μέθοδος FH (1976) στηρίζεται σε τρεις αρχές που πρέπει να τηρούνται :

I. Τα δεδομένα σε κάθε εγγραφή πρέπει να διορθώνονται έτσι ώστε οι διορθωμένες εγγραφές να ικανοποιούν όλους τους κανόνες αλλάζοντας όσο το δυνατόν μικρότερο αριθμό πεδίων.

II. Δεν είναι ανάγκη να παράγουμε νέους κανόνες για την διόρθωση των πεδίων, αντίθετα αυτοί προκύπτουν αυτόματα από τους ήδη υπάρχοντες.

III. Κατά την διόρθωση των εγγραφών, πρέπει να διατηρείται αν όχι η περιθώρια, η από κοινού κατανομή συχνοτήτων των μεταβλητών του συνόλου των σωστών εγγραφών.

2.5 Πρόβλημα εντοπισμού λαθών

Όπως προαναφέρθηκε η γνώση μόνο του συνόλου των ρητά ορισμένων κανόνων είναι ικανή να αποκαλύψει ποιες εγγραφές δεν είναι σωστές. Το πλήρες σύνολο κανόνων είναι απαραίτητο για την εύρεση των συγκεκριμένων πεδίων που χρειάζονται διόρθωση, σύμφωνα με τον επόμενο συλλογισμό:

Έστω ότι μια εγγραφή χάνει ένα αριθμό κανόνων από το πλήρες σύνολο των κανόνων. Επίσης έστω ότι ένα συγκεκριμένο πεδίο περιέχεται σε όλους τους χαμένους κανόνες. Τότε θα υπάρχει μια τιμή για το πεδίο αυτό με την οποία η εγγραφή θα ικανοποιούσε όλους τους κανόνες. Αν δεν υπάρχει τέτοια τιμή, σημαίνει ότι για όλες τις τιμές του πεδίου αυτού, κάποιοι κανόνες δεν ικανοποιούνται (ενώ οι τιμές των άλλων πεδίων παραμένουν ίδιες). Κατά συνέπεια δεν ευθύνεται το πεδίο *i* για το ότι η εγγραφή χάνει τους κανόνες, αλλά κάποιο άλλο πεδίο. Τότε θα έπρεπε να υπάρχει τουλάχιστον ένας άλλος κανόνας που «χάνεται» και δεν περιέχει το πεδίο *i*. Αυτός ο ισχυρισμός έρχεται σε αντίθεση με την αρχική

υπόθεση ότι το συγκεκριμένο πεδίο ανήκει σε όλους τους χαμένους κανόνες. Συνήθως δεν βρίσκουμε μόνο την τιμή του πεδίου *i* που θα διορθώσει όλες τις εγγραφές αλλά ένα συνδυασμό τιμών πεδίων που θα διορθ΄χσουν την λανθασμένη εγγραφή.

Υπενθυμίζουμε ότι οι Fellegi-Holt (1976) επισήμαναν ότι το πλήρες σύνολο κανόνων αποτελείται από τους ρητά ορισμένους κανόνες και τους κατά βάση νέους κανόνες. Για το πρόβλημα εντοπισμού λαθών πρέπει να ορίσουμε τα εξής :

Ορίζουμε Ω το πλήρες σύνολο των κανόνων, $\Omega_k \subseteq \Omega$ το σύνολο που αποτελείται από κανόνες της μορφής $E_r : \prod_{i=1}^{k} E_{ri}$ δηλαδή κανόνες που περιέχουν μόνο τα *1,2,...,κ* πεδία της έρευνας. Για τα υπόλοιπα *j=k+1,k+2,...,n* πεδία ισχύει $E_{rj} = R_j$. Ο συνολικός αριθμός των πεδίων της έρευνας είναι *n.*

Θεώρημα 2 (Fellegi-Holt 1976) : Αν y_i^0 με $i = 1,2,...,k-1$ είναι αντίστοιχα πιθανές τιμές για τα πρώτα *κ-1* πεδία και αυτές οι τιμές των πεδίων ικανοποιούν όλους τους κανόνες στο σύνολο Ω_{k-1}, τότε υπάρχει μια τιμή y_k^0 για το πεδίο *κ* έτσι ώστε οι τιμές y_i^0 με $i = 1,2,...,k$ ικανοποιούν όλους τους κανόνες του συνόλου Ω_k.

Βασισμένοι στο παράδειγμα 1 έχουμε βρει το πλήρες σύνολο κανόνων Ω

$E^1 = \{1\} \times \{2,3,4,5\} \times R_3$

$E^2 = R_1 \times \{1,3,4,5\} \times \{2\}$

$E^* = \{1\} \times R_2 \times \{2\}$

Το σύνολο Ω_2 αποτελείται από τους κανόνες που περιέχουν τα δυο πρώτα πεδία Ηλικία και Ο.Κ, δηλαδή $\Omega_2 = E^1$. Αν λοιπόν μια εγγραφή έχει τιμές για τα πεδία Ηλικία και Ο.Κ. που να ικανοποιούν τον E^1 τότε υπάρχει μια τιμή για το πεδίο Σ.Ν. που ικανοποιεί όλους τους κανόνες. Μια εγγαρφή με τιμές στα πεδία Ηλικία=1 (0-14) και Ο.Κ.= 1(Άγαμος) ικανοποιεί τον πρώτο κανόνα και για την τιμή στο πεδίο Σ.Ν.=3(Άλλο) ικανοποιεί όλους τους κανόνες.
Βασισμένοι στο Θεώρημα 2 προκύπτει το επόμενο πορίσματα.

Πόρισμα 1 (Fellegi-Holt 1976) : Έστω ότι ένα ερωτηματολόγιο αποτελείται από *N* πεδία έχοντας τιμές y_i $(i = 1,2,..N)$. Έστω επίσης ένα σύνολο $S \subseteq N$ πεδίων που έχει την ιδιότητα

μια τουλάχιστον από τις τιμές $y_i \ (i \in S)$ εμφανίζονται σε κάθε κανόνα που χάνεται από μια συγκεκριμένη εγγραφή. Τότε υπάρχουν τιμές $y_i^0 \ (i \in S)$ τέτοιες ώστε οι διορθωμένη εγγραφή που αποτελείται από τις τιμές $y_i \ (i \notin S)$ μαζί με τις $y_i^0 \ (i \in S)$ ικανοποιούν όλους τους κανόνες.

Σύμφωνα με το πόρισμα 1, μπορούμε να επιλέξουμε κάθε σύνολο πεδίων που έχει την ιδιότητα ένα τουλάχιστον από αυτά να υπάρχει σε κάθε κανόνα που δεν περνάει από μια εγγραφή. Τότε υπάρχουν τιμές για αυτό το σύνολο που μαζί με τις υπόλοιπες τιμές (εκτός αυτού του συνόλου) της εγγραφής θα οδηγήσουν στην ικανοποίηση κάθε κανόνα. Τέλος αν καταφέρουμε να επιλέξουμε το σύνολο S με το μικρότερο αριθμό πεδίων θα καταλήξουμε στη βέλτιστη λύση του προβλήματος του εντοπισμού λαθών.

2.6 Διόρθωση δεδομένων κατά Fellegi-Holt (1976)

Το τελικό στάδιο της διαδικασίας των Fellegi-Holt είναι να διορθώσουμε τα πεδία που εντοπίσαμε στο προηγούμενο βήμα γιατί αυτά είναι η αιτία που οι εγγραφές να χάνουν τους κανόνες. Η μέθοδος που πρότειναν οι Fellegi-Holt (1976) για την διόρθωση των πεδίων δεν απαιτεί την παραγωγή νέων κανόνων. Επίσης κατά την διόρθωση των πεδίων παραμένει ίδια η κατανομή των εγγραφών και είναι αυτή που ακολουθούν τα δεδομένα που ικανοποιούν όλους τους κανόνες. Υπάρχουν δυο μέθοδοι διόρθωσης τις οποίες παρουσιάζουμε στη συνέχεια.

2.6.1 Ακολουθιακή διόρθωση (sequential imputation)

Έστω ότι οι εγγραφές της έρευνας περιέχουν *N* πεδία και 1,2,...,*K* είναι τα πεδία που πρέπει να διορθωθούν. Η διόρθωση θα αρχίσει με το *K*-στο πεδίο και θα συνεχιστεί σταδιακά στα υπόλοιπα πεδία, *K-1, K-2,..,.1*. Θεωρούμε όλους τους κανόνες (έστω M) που εμπεριέχουν το πεδίο *K* και όχι τα υπόλοιπα *K-1, K-2 , ...1*.

$$E_r : \prod_{i=K}^{N} E_{ri} \quad \text{με } r = 1,2,...,M$$

Δεδομένης μιας τρέχουσας εγγραφής διαγράφουμε από το παραπάνω σύνολο των κανόνων αυτούς που ικανοποιούνται εκ των προτέρων, από τις τιμές των πεδίων K+1, K+2,...N, εφόσον αυτοί οι κανόνες θα ικανοποιούνται ανεξάρτητα από την τιμή που θα διορθωθεί στο πεδίο *K*. Συνεπώς, κρατάμε μόνο τους κανόνες M' που θα ικανοποιούνται ή

όχι, ανάλογα με την τιμή στο πεδίο K. Τελικά όλοι οι κανόνες ικανοποιούνται διορθώνοντας την τιμή του πεδίου K με τις τιμές που ανήκουν στη παρακάτω σχέση :

$$y_k^* \in \prod_{r=1}^{M'} \overline{E_{rk}} \tag{2.6.1}$$

και $\overline{E_{rk}}$ είναι το συμπληρωματικό του E_{rk} . Για να διατηρηθεί η κατανομή συχνοτήτων βρίσκουμε τις σωστές εγγραφές που έχουν τιμές στα πεδία που δεν χρειάζονται διόρθωση ίδιες με αυτές της τρέχουσας εγγραφής και τιμές στο πεδίο που χρειάζεται διόρθωση τιμές από το σύνολο τιμών (2.6.1). Θα μπορούσαμε τότε να επιλέξουμε τιμή για διόρθωση στο πεδίο αυτή τιμή που εμφανίζεται πιο συχνά στις σωστές εγγραφές. Επιλέγοντας τελικά την τιμή y_k^* για την διόρθωση του πεδίου *K* συνεχίζουμε με σκοπό να διορθώσουμε την τιμή στο πεδίο *K-1*. Βρίσκουμε λοιπόν το σύνολο των κανόνων που εμπεριέχει το πεδίο *K-1* και όχι τα *K-2, ...,1* και εφαρμόζουμε ξανά την παραπάνω διαδικασία.

2.6.2 Από κοινού διόρθωση (joint imputation)

Έστω ότι τα πεδία που χρειάζονται διόρθωση είναι τα *K* πρώτα. Θεωρώ όλους τους κανόνες που περιέχουν τα πεδία αυτά. Καταλήγω στο νέο σύνολο κανόνων *M″* που όπως και στην Ακολουθιακή διόρθωση, προέκυψε από την διαγραφή των κανόνων που ικανοποιούνται από τις τιμές των πεδίων *K+1*,K+2,…,N δεδομένης μιας εγγραφής. Ας θεωρήσουμε τα σύνολα

$$E_i^* = \prod_{r=1}^{M''''} E_{ri} \qquad i = K+1,...,N$$

Βρίσκουμε μια εγγραφή (από το σύνολο των «καθαρών» δεδομένων) με τιμές στα πεδία *K+1,...N* που ανήκουν στα παραπάνω σύνολα. Τότε δανείζομαι από αυτή την εγγραφή τις τιμές που περιέχει στα πεδία *1,2,...,K* για να διορθώσω την υπό μελέτη εγγραφή. Με τη μέθοδο αυτή δεν είναι απαραίτητο να «υπολογίσω» τις τιμές των πεδίων που θα χρησιμοποιήσουμε για διόρθωση έτσι ώστε να ικανοποιούνται οι κανόνες αλλά αυτές προκύπτουν αυτόματα από τις εγγραφές που είναι σωστές.

Συγκριτικά, ενώ με την ακολουθιακή διόρθωση πρέπει να εφαρμοστεί η μέθοδος *K* φορές όσα και τα πεδία που απαιτούν διόρθωση, με την από κοινού διόρθωση η διαδικασία εκτελείται μια φορά. Με την από κοινού μέθοδο διόρθωσης τα σύνολα που ορίζουμε αποτελούν ένα υποσύνολο του πληθυσμού στον οποίο ανήκει η εγγραφή που ψάχνουμε, και ίσως είναι χρονοβόρο να ψάχνουμε ένα υποσύνολο του πληθυσμού που τηρεί κάποιες

προϋποθέσεις από το να ψάχνουμε μια εγγραφή με συγκεκριμένη τιμή στο πεδίο που διορθώνουμε όπως συμβαίνει στην ακολουθιακή μέθοδο. Η επιλογή της μεθόδου κρίνεται ανάλογα με την πολυπλοκότητα των δεδομένων και των κανόνων, το μέγεθος των πεδίων τιμών των μεταβλητών και τη συχνότητα που χάνονται οι κανόνες.

2.7 Αυτοματοποιημένο σύστημα ελέγχου ορθότητας (FH , 1976)

Παράλληλα με την μεθοδολογία που πρότειναν οι Fellegi – Holt (1976) για τον έλεγχο ορθότητας δεδομένων δημιουργήσαν το αυτοματοποιημένο σύστημα για την εφαρμογή της. Το σύστημα αποτελείται από ευανάγνωστους πίνακες στους οποίους καταχωρούνται οι κανόνες και η τρέχουσα εγγραφή με τη χρήση κωδικών 0 και 1 σε μορφή διανυσμάτων. Τα βήματα που ακολουθεί το αυτοματοποιημένο σύστημα παρομοίως με τη θεωρητική βάση του προβλήματος είναι:

- Δημιουργία ενός πίνακα όπου στήλες είναι τα πεδία της έρευνας και γραμμές οι κανόνες ορθότητας, με στοιχεία 0 και 1.
- Παραγωγή του πλήρους συνόλου κανόνων.
- Εύρεση των εγγραφών που αποτυγχάνουν τους κανόνες.
- Εύρεση του μικρότερου συνόλου πεδίων που πρέπει να διορθωθούν (error localization).
- Διόρθωση των πεδίων του προηγούμενου βήματος (Imputation).

Η αναπαράσταση της μεθόδου στο αυτοματοποιημένο σύστημα θα γίνει με τη βοήθεια ενός παραδείγματος που περιέχει κατηγορικές μεταβλητές μόνο.

***Παράδειγμα 2 Fellegi-Holt* (1976)**

- Φύλο : Άνδρας (Α) , Γυναίκα (Γ)
- Ηλικία : 0-14 , 15-16 , 17+
- Οικογενειακή κατάσταση (**Ο.Κ**): Άγαμος/η (Α), Παντρεμένος/η(Π), Διαζευγμένος/η(Δ) , Χήρος /Χήρα(Χ) , Σε Διάσταση (Σ)
- Σχέση με τον κύριο του νοικοκυριού (**Σ.Ν**): Η Σύζυγος(Η), Ο Σύζυγος (Ο), Κόρη/Γιος (ΚΓ), Άλλο(Α)
- Εκπαίδευση : Καμία (Κ) , Δημοτικού (Δ), Πρωτοβάθμια (Π) , Δευτεροβάθμια (Δε)

Οι ρητοί κανόνες που έχουν οριστεί από τους ειδικούς είναι :

E_1 : (Φύλο = Α) και (**Σ.Ν** = Η)

E_2 : (Ηλικία = 0-14) και (**Ο.Κ**= Κάποτε Παντρεμένος/η)

=(Ηλικία = 0-14) και (**Ο.Κ** = Π ή Δ ή Χ ή Σ)

E_3 : (**Ο.Κ** = Όχι τώρα Παντρεμένος/η) και (**Σ.Ν**. = Σύζυγος)

=(**Ο.Κ**= Α ή Δ ή Χ ή Σ) και (**Σ.Ν**. = Ο ή Η)

E_4 : (Ηλικία = 0-14) και (**Σ.Ν** = Ο ή Η)

E_5 : (Ηλικία = 0-16) και (Εκπαίδευση = Δε)

2.7.1 Αναπαράσταση κανόνων και πεδίων σε πίνακα

Πίνακας 2

	Φύλο		Ηλικία			Οικογενειακή Κατάσταση					Σχέση με Νοικοκύρη				Εκπαίδευση			
Κανόνες	Α	Γ	0-14	15-16	17+	Α	Π	Δ	Χ	Σ	Η	Ο	ΚΓ	Α	Κ	Δ	Π	Δε
E_1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1
E_2	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
E_3	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1
E_4	1	1	1	0	0	1	1	1	1	1	1	1	0	0	1	1	1	1
E_5	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1
Νέα εγγραφή	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0

Κάθε κανόνας που δεν περιέχει κάποιο πεδίο σημαίνει ότι το πεδίο αυτό παίρνει κάθε τιμή του πεδίου τιμών του.

Η τελευταία γραμμή του πίνακα αποτελεί μια τρέχουσα εγγραφή. Η αναπαράσταση των λογικών κανόνων σε πίνακα γίνεται ως εξής :

- Κοιτάω αρχικά ποια πεδία περιέχει ο κανόνας. Για τα πεδία που περιέχονται στους κανόνες, βάζω τον κωδικό 1 στα στοιχεία του πεδίου που ορίζει ο κανόνας και 0 διαφορετικά. Δηλαδή για τον πρώτο κανόνα που περιέχει τα πεδία Φύλο και **Σ.Ν**, βάζω τον κωδικό 1 για τις τιμές Άνδρας και Η Σύζυγος και 0 στις υπόλοιπες τιμές των πεδίων αυτών.
- Για τα πεδία που δεν περιέχονται σε ένα κανόνα, καταχωρώ στον πίνακα τον κωδικό 1 για όλες τις τιμές των πεδίων αυτών.

Όμοια με τους κανόνες γίνεται και η καταχώρηση των κωδικών για την τρέχουσα εγγραφή.
Οι τρόποι για να εξετάσουμε ποιοι κανόνες ικανοποιούνται και ποιοι όχι από την τρέχουσα εγγραφή είναι οι εξής :

1. Τοποθετώ την τρέχουσα εγγραφή «επάνω» σε κάθε κανόνα. Αν οι κώδικες 1 της εγγραφής συμπίπτουν με αυτούς του κάθε κανόνα τότε η εγγραφή «χάνει» τον αντίστοιχο κανόνα, διαφορετικά τον ικανοποιεί.
2. Αν υποθέσουμε ότι η κάθε γραμμή του πίνακα είναι ένα διάνυσμα με στοιχεία 0 και 1 υπολογίζω το εσωτερικό γινόμενο της τρέχουσας εγγραφής με κάθε κανόνα. Αν το αποτέλεσμα είναι ίσο με το πλήθος των πεδίων της έρευνας τότε η εγγραφή χάνει τον αντίστοιχο κανόνα.
3. Δημιουργώ ένα νέο πίνακα με στήλες τις στήλες του πίνακα 2 όπου η τρέχουσα εγγραφή έχει στοιχεία 1

Πίνακας 3

Κανόνες	Άνδρας	0-14	Παντρεμένος/η	Η Σύζυγος	Δημοτικού	Γινόμενο
E_1	1	1	1	1	1	1
E_2	1	1	1	1	1	1
E_3	1	1	0	1	1	0
E_4	1	1	1	1	1	1
E_5	1	1	1	1	0	0

Υπολογίζουμε το γινόμενο των στοιχείων της κάθε γραμμής και όταν το αποτέλεσμα δώσει 1 σημαίνει ότι η τρέχουσα εγγραφή χάνει τον κανόνα , ενώ αν δώσει 0 ικανοποιεί τον κανόνα.
Με κάθε ένα από τους παραπάνω τρόπους καταλήγουμε στο ίδιο συμπέρασμα ότι η τρέχουσα εγγραφή χάνει τους κανόνες *E_1, E_2, E_4*.

2.7.2 Παραγωγή πλήρους συνόλου κανόνων

Το επόμενο βήμα της διαδικασίας είναι να παράγουμε το πλήρες σύνολο των κανόνων. Έστω ότι χρησιμοποιούμε τους δύο κανόνες *E_1* και *E_3* με γεννήτορα το πεδίο **O.K** για να παράγουμε ένα νέο κανόνα. Στηριζόμαστε στους εξής κανόνες:

1. Για το γεννήτορα καταχωρούμε τον κωδικό 1 όταν κάποιος από τους συμβάλλοντες κανόνες έχει κωδικό 1. Ενώ καταχωρώ τον κωδικό 0 όταν όλοι οι συμβάλλοντες κανόνες έχουν κωδικό 0.

2. Για τα υπόλοιπα πεδία ακολουθώ αντίθετη διαδικασία, δηλαδή καταχωρώ τον κωδικό 1 όταν όλοι οι συμβάλλοντες κανόνες έχουν κωδικό 1 και 0 όταν κάποιος από τους συμβάλλοντες κανόνες έχει κωδικό 0.

Με την παραπάνω διαδικασία έχω το παρακάτω έμμεσο κανόνα E_6.

Πίνακας 4

	Φύλο		Ηλικία			Ο.Κ (γεννήτορας)					Σ.Ν.				Εκπαίδευση			
E_1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1
E_3	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1
E_6	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1

Ο έμμεσος κανόνας είναι «κατά βάση νέος» κανόνας όταν ισχύουν τα παρακάτω:

1. Κανένα από τα πεδία δεν έχει όλα τα στοιχεία του 0
2. Ο γεννήτορας πρέπει να περιέχει σε όλα τα πεδία του τιμές 1
3. Ο νέος κανόνας δεν πρέπει να περιέχεται σε κανέναν άλλο κανόνα, δηλαδή δεν πρέπει να υπάρχει άλλος κανόνας που να έχει τις τιμές 1 σε περισσότερες «θέσεις» από τις ήδη υπάρχουσες στον νέο κανόνα.

Ο κανόνας E_6 που προέκυψε στο παράδειγμά μας ικανοποιεί τα δυο πρώτα κριτήρια αλλά όχι το τρίτο. Στην ουσία είναι ίδιος με τον κανόνα E_1 και για το λόγο αυτό δεν τον προσθέτουμε στη λίστα κανόνων που θα αποτελέσει το πλήρες σύνολο κανόνων.

Συνδυάζοντας ανά δυο, τρεις μέχρι να εξαντληθούν όλοι οι συνδυασμοί και ορίζοντας ως γεννήτορα άλλο πεδίο κάθε φορά μπορούμε να παράγουμε όλους τους έμμεσους κανόνες. Στη συνέχεια ελέγχοντας ποιοι από αυτούς είναι «κατά βάση νέοι» τους κρατάμε για το πλήρες σύνολο κανόνων. Η διαδικασία βέβαια είναι χρονοβόρα και για το λόγο αυτό η μεθοδολογία Fellegi–Holt προτείνει διευκολύνσεις που μειώνουν το μέγεθος της διαδικασίας (ενότητα *2.4.4*)

2.7.3 Εύρεση ελάχιστου συνόλου πεδίων που χρειάζονται διόρθωση

Εφόσον το πλήρες σύνολο των κανόνων έχει παραχθεί το επόμενο στάδιο είναι να βρούμε το ελάχιστο σύνολο πεδίων που πρέπει να διορθώσουμε.

Δημιουργώ ένα νέο πίνακα με γραμμές τους κανόνες που χάνονται από την τρέχουσα εγγραφή και στήλες όλα τα πεδία της έρευνας. Τα στοιχεία του πίνακα 5 είναι κώδικες 1 όταν το πεδίο περιέχεται στο αντίστοιχο κανόνα και 0 διαφορετικά. Με τον τρόπο αυτό

δημιουργείται ο παρακάτω πίνακας. Στην ενότητα *2.7.1* είχαμε διαπιστώσει ότι η τρέχουσα εγγραφή χάνει τους κανόνες E_1, E_2 και E_4 .

Πίνακας 5

Κανόνες	Φύλο	Ηλικία	Οικογ. Κατάστ.	Σχέση με νοικ.	Εκπαίδευση
E_1	1	0	0	1	0
E_2	0	1	1	0	0
E_4	0	1	1	0	0

Στην ουσία, τα πεδία που έχουν τιμή 1 για κάποιο κανόνα είναι πιθανά πεδία για διόρθωση. Επιλέγουμε ένα κανόνα που περιέχει τον μικρότερο αριθμό πεδίων που θέλω να διορθώσω. Αν κάποιοι κανόνες έχουν ίδιο αριθμό πεδίων επιλέγω έναν τυχαία. Όπως και στα παράδειγμα μας όλοι οι κανόνες περιέχουν ίδιο αριθμό πεδίων για αυτό επιλέγω τυχαία τον E_1 που περιέχει τα πεδία Φύλο και **Σ.Ν.**

Θα δημιουργήσω δυο νέους πίνακες. Ο κάθε ένας θα αποτελείται από τις στήλες με όλα τα πεδία εκτός αυτού που πρότεινα για διόρθωση και με κανόνες όσους δεν περιέχουν το αντίστοιχο πεδίο. Αν ο νέος πίνακας δεν περιέχει κανένα κανόνα τότε έχουμε βρει το ελάχιστο σύνολο των πεδίων για διόρθωση.

Στο πρώτο πίνακα αφαιρούμε τους κανόνες που περιέχουν το πεδίο Φύλο και τη στήλη του πεδίου.

Πίνακας 6 (Φύλο)

Κανόνες	Ηλικία	Ο.Κ.	Σ.Ν.	Εκπαίδευση
E_2	1	1	0	0
E_4	1	1	0	0

Στον δεύτερο πίνακα παρόμοια αφαιρώ τους κανόνες που περιέχουν το πεδίο **Σ.Ν** και τη στήλη του πεδίου.

Πίνακας 7 (**Σ.Ν**.)

Κανόνες	Φύλο	Ηλικία	Ο. Κ.	Εκπαίδευση
E_2	0	1	1	0
E_4	0	1	1	0

Για το κάθε πίνακα ξανακάνω την παραπάνω διαδικασία:

Στον πίνακα 6 επιλέγω τον κανόνα E_2 τυχαία (όλοι οι κανόνες περιέχουν τον ίδιο αριθμό πεδίων). Ο κανόνας αυτός περιέχει τα πεδία Ηλικία και **Ο.Κ**. Δημιουργώντας τους δυο νέους πίνακες διαπιστώνουμε ότι είναι κενοί. Το ίδιο ισχύει αν εφαρμόσω τη διαδικασία στον πίνακα 7. Έτσι οι λύσεις που βρήκα στο πρόβλημα εντοπισμού λαθών είναι :

{Φύλο, Ηλικία}, {Φύλο, Ο.Κ} , {Σ.Ν , Ηλικία} , {Σ.Ν , Ο.Κ}

2.7.4 Διόρθωση των πεδίων

Τέλος το αυτοματοποιημένο σύστημα έχει τη δυνατότητα διόρθωσης των εγγραφών και στηρίζεται στη θεωρητική βάση των μεθόδων που περιγράφηκαν στην παράγραφο *2.6.*

Αυτόματη διόρθωση σύμφωνα με την ακολουθιακή μέθοδο:

Έστω ότι τα πεδία που θα διορθώσουμε είναι Φύλο και η Ηλικία που βρέθηκαν στο προηγούμενο βήμα. Όπως προαναφέραμε η τρέχουσα εγγραφή χάνει τους κανόνες E_1, E_2 και E_4. Αρχικά θα διορθώσουμε την Ηλικία. Από τους κανόνες E_1, E_2, E_4 κρατώ αυτούς που περιέχουν την Ηλικία και όχι το πεδίο Φύλο. Προκύπτει ο επόμενος πίνακας 8 με γραμμές τους δυο κανόνες E_2, E_4 και στήλες όλες τις τιμές του πεδίου Ηλικία θέτοντας στον πίνακα 8 την τιμή 1 όταν μια τιμή του συγκεκριμένου πεδίου υπάρχει στον αντίστοιχο κανόνα.

Πίνακας 8

	Ηλικία		
Κανόνες	0-14	15-16	17+
E_2	1	0	0
E_4	1	0	0

Στη συνέχεια δημιουργώ ένα νέο πίνακα που σε κάθε στήλη έχει τιμή 1 όταν υπάρχει ο κωδικός 1 σε ένα τουλάχιστον κελί της αντίστοιχης στήλης και 0 διαφορετικά. Προκύπτει ο νέος πίνακας :

Πίνακας 9

0-14	15-16	17+
1	0	0

Η τιμή που θα διορθώσει την εγγραφή στο πεδίο Ηλικία είναι αυτή που στον πίνακα 9 παίρνει τιμή 0 και θα είναι αυτή που θα ικανοποιεί τους κανόνες E_2 και E_4. Ψάχνουμε στα «καθαρά» δεδομένα μέχρι να βρούμε μια εγγραφή που να έχει τιμή στο πεδίο Ηλικία 15-16 ή 17+ και με ανάλογα κριτήρια επιλέγουμε μια τιμή από αυτές για να διορθώσουμε το πεδίο Ηλικία.

Στη συνέχεια θα διορθώσουμε το πεδίο Φύλο. Από τους κανόνες E_1, E_2 και E_4 αυτός που περιέχει το πεδίο Φύλο είναι ο E_1. Η μόνη τιμή του Φύλου που ικανοποιεί το E_1 είναι Γ (Γυναίκα) και με αυτή διορθώνουμε το πεδίο Φύλο.

Αυτόματη διόρθωση σύμφωνα με την από κοινού μέθοδο:

Αρχίζουμε την διαδικασία μόνο με τους κανόνες E_1, E_2 και E_4 γιατί οι υπόλοιποι ικανοποιούνται από την τρέχουσα εγγραφή. Τα πεδία που χρειάζονται διόρθωση είναι το Φύλο και η Ηλικία. Στον πίνακα 2, επιλέγω για τα πεδία που δεν χρειάζονται διόρθωση και για κάθε έναν από τους παραπάνω κανόνες τις τιμές που αντιστοιχούν στους κωδικούς 1. Ο πίνακας 2 μας δίνει τις εξής τιμές :

- **Ο. Κ**: Παντρεμένος/η , Χωρισμένος/η, Σε Διάσταση ,Χήρος/α
- **Σ.Ν**: Η σύζυγος
- Εκπαίδευση : Κάθε τιμή.

Ψάχνουμε λοιπόν στα «καθαρά» δεδομένα μέχρι να βρούμε εγγραφή που στα παραπάνω πεδία έχει ένα οποιοδήποτε συνδυασμό των προτεινόμενων τιμών. Δανειζόμαστε τις τιμές των υπολοίπων πεδίων αυτής της εγγραφή για να διορθώσουμε τα πεδία Φύλο και Ηλικία της υπό μελέτης εγγραφής.

2.8 Αξιολόγηση συστήματος Fellegi-Holt

Η αναλυτική περιγραφή της μεθοδολογίας Fellegi-Holt (1976) που δόθηκε προηγουμένως, αποτελεί τη βάση των μεθοδολογιών που αναπτύχθηκαν μεταγενέστερα, καθώς πρωτοποριακή παραμένει η απόδοση του θεωρητικού πλαισίου της μεθόδου σε αυτοματοποιημένη μορφή. Λαμβάνοντας υπόψιν την υπολογιστική ισχύ την περίοδο που αναπτύχθηκε η μεθοδολογία, παρουσιάζουμε ορισμένα χαρακτηριστικά της, αλλά και κάποιους περιορισμούς στην εφαρμογή της.

Όσον αφορά τη μεθοδολογία, μπορεί να χαρακτηριστεί από τα εξής:

- Η διαδικασία παραγωγής του συνόλου των κανόνων τελειώνει, πριν ξεκινήσει ο έλεγχος των δεδομένων. Αυτό επιτρέπει στο σύστημα να διορθώσει οποιεσδήποτε ασυνέπειες εμφανιστούν και να προσαρμόσει τους κανόνες κατάλληλα.
- Οι διαδικασίες που εκτελούνται από το σύστημα είναι σαφώς δομημένες και διαχωρισμένες, ώστε να επιτρέπουν οποιαδήποτε παρέμβαση και μετατροπή απαιτείται.
- Μόνο οι ρητοί κανόνες πρέπει αρχικά να προσδιοριστούν. Από αυτούς στη συνέχεια θα παραχθεί το πλήρες σύνολο κανόνων, ενώ δεν είναι ανάγκη να υπολογίσουμε κανένα κανόνα για τη διαδικασία της διόρθωσης αφού αυτοί παράγονται αυτόματα.

- Η διαδικασία επιτυγχάνει να διατηρήσει τα δεδομένα όσο το δυνατόν περισσότερο όμοια στις αρχικές τους τιμές ελαχιστοποιώντας τον αριθμό των πεδίων που θα διορθώσει ώστε τα τελικά δεδομένα να ικανοποιούν όλους τους κανόνες.
- Με το πρώτο πέρασμα των δεδομένων μέσα από το σύστημα λαμβάνουμε «καθαρά» δεδομένα.

Όσον αφορά το αυτοματοποιημένο σύστημα επισημαίνουμε τα παρακάτω:

- Υπάρχει σαφές πλεονέκτημα της αναπαράστασης των κανόνων σε πίνακες έναντι των κλασικών τύπων κανόνων If-Then-Else. Οι τελευταίοι τύποι κανόνων είναι δύσκολο να μετατραπούν σε κωδικοποιημένη μορφή. Επίσης, οποιαδήποτε μικρή αλλαγή στη έρευνα θα απαιτούσε να αναπροσαρμοστεί ο αλγόριθμος γεγονός που είναι ιδιαίτερα χρονοβόρο. Αντίθετα η χρήση των κανονικών μορφών των κανόνων και η μετατροπή τους σε κώδικες με μορφή πινάκων, είναι απλή και μπορεί να χρησιμοποιηθεί και από άλλες παρόμοιες έρευνες γιατί είναι εύκολη η προσαρμογή τους στα νέα δεδομένα.
- Το πρόγραμμα δεν απαιτεί άριστη γνώση προγραμματισμού από το χρήστη. Αντίθετα ένα σαφώς ορισμένο σύνολο ρητών κανόνων και η σωστή κωδικοποίησή τους μπορεί να ελέγξει δεδομένα χωρίς την παρουσία ειδικών αναλυτών και προγραμματιστών.
- Το υπολογιστικό κομμάτι της παραγωγής του πλήρους συνόλου κανόνων είναι ιδιαίτερα αργό και απαιτεί μεγάλες δυνατότητες από τον υπολογιστή που το εφαρμόζει.
- Το πρόγραμμα δεν έχει εφαρμοστεί σε συνεχείς μεταβλητές. Αν και είναι εύκολο να αναπαρασταθούν οι αριθμητικοί κανόνες σε κανονική μορφή είναι δύσκολο να εφαρμοστεί η μεθοδολογία. Για συνεχείς μεταβλητές ή και μίξη συνεχών και κατηγορικών αναπτύχθηκαν νέες μέθοδοι ελέγχου δεδομένων Quere (Statistics Netherlands 2000) και Quere and De Waal (Statistics Netherlands 2000) αντίστοιχα.

Παρατηρήσεις:

Καταλήγοντας σε ένα σύνολο πεδίων για διόρθωση δεν σημαίνει ότι βρήκαμε τη μοναδική λύση. Για την εύρεση της βέλτιστης λύσης θα μπορούσαμε να θέσουμε βάρη για κάθε πεδίο. Υψηλή τιμή αυτού, σημαίνει ότι η τιμή στο πεδίο είναι πιθανότατα σωστή, δηλαδή αντιμετωπίζουμε με αξιοπιστία τις τιμές των πεδίων που έχουν μεγαλύτερο βάρος. Αντίθετα μικρή τιμή βάρους σημαίνει μεγαλύτερη πιθανότητα η απάντηση (τιμή στο αντίστοιχο πεδίο)

να είναι λάθος. Η βέλτιστη λύση στο πρόβλημα εύρεσης του συνόλου των πεδίων για διόρθωση θα ήταν αυτή με το μικρότερο άθροισμα βαρών. (Fellegi-Holt 1976)

ΚΕΦΑΛΑΙΟ 3

Βελτίωση της μεθόδου Fellegi-Holt στο στάδιο παραγωγής των έμμεσων κανόνων

Αν και η πρωτοποριακή μέθοδος Fellegi-Holt (1976) είναι σαφής, ολοκληρωμένη ως προς το θεωρητικό πλαίσιο, χωρίς να απαιτεί ιδιαίτερες ικανότητες προγραμματισμού όσον αφορά την εφαρμογή της, το τεράστιο υπολογιστικό κόστος και η αδυναμία αντιμετώπισης μεγάλου μεγέθους ερευνών αποτέλεσε την κύρια αιτία για περαιτέρω μελέτη και εκτεταμένη έρευνα για την βελτίωσή της. Στο κεφάλαιο αυτό, θα παρουσιάσουμε μια σειρά βελτιώσεων που υπέστη η μέθοδος Fellegi-Holt (1976) στο στάδιο της παραγωγής του πλήρους συνόλου κανόνων που αρκεί για να λυθεί το πρόβλημα εντοπισμού λαθών. Κοινός στόχος όλων των προσεγγίσεων είναι η μείωση του υπολογιστικού κόστους σε αυτό το στάδιο της διαδικασίας που είναι και το πιο χρονοβόρο. Επίσης παραθέτουμε δυο αυτοματοποιημένα προγράμματα που εφαρμόζουν τις μεθόδους που περιγράφουμε όπως το DISCRETE (Winkler and Petkunas 1996) και το SPEER (Structured Programs for Economic Editing and Referrals) (Kovar and Winkler 1996) που χρησιμοποιούνται από την Αμερικανική Υπηρεσία Απογραφών και επεξεργάζονται κατηγορικά και αριθμητικά δεδομένα αντίστοιχα.

3.1 Πρόβλημα κάλυψης συνόλου όπως προκύπτει από τη μέθοδο FH (1976)

Σύμφωνα με τη μέθοδο Fellegi-Holt (1976) η βέλτιστη λύση στο πρόβλημα εντοπισμού λαθών είναι ο ελάχιστος αριθμών σταθμισμένων πεδίων που χρειάζονται διόρθωση. Το πρόβλημα εντοπισμού λαθών αποτελεί ένα **πρόβλημα κάλυψης συνόλου** (set covering problem) το οποίο διαφορετικά καλείται **μοντέλο ελάχιστου αριθμού σταθμισμένων πεδίων για διόρθωση (Minimum Weighted Fields to Impute MWFI)** και περιγράφεται από τα εξής:

Έστω ότι $\overline{E}_E$ είναι το πλήρες σύνολο κανόνων και έστω μια εγγραφή για την οποία ισχύει $y^0 \in P(\overline{E_E})$, τότε το πρόβλημα εντοπισμού των λαθών μεταφράζεται στις παρακάτω σχέσεις:

$$\text{Ελαχιστοποιώ } \sum_{j\in J} c_j x_j \tag{3.1.1}$$

$$\text{υπό την προϋπόθεση ότι } \sum_{j\in J} a_{ij} x_j \geq 1 \tag{3.1.2}$$

$$\text{όπου } a_{ij} = \begin{cases} 1, & an\ o\ \ E^i\ periecei\ \ to\ \ pedio\ \ j \\ 0, & a\ l\ l\ iwV \end{cases}$$

και x_j =0, 1, j ανήκει στο J ={*1,2,…,n*} το σύνολο των πεδίων της εγγραφής. Η σταθερά c_j είναι ένα μέτρο εμπιστοσύνης για το πεδίο *j*. Αν η τιμή της σταθεράς είναι μεγάλη, δηλαδή δίνει μεγάλο βάρος στο συγκεκριμένο πεδίο, σημαίνει ότι είναι πιο πιθανό η τιμή του πεδίου να είναι σωστή. Αντίθετα μικρή τιμή της σταθεράς σημαίνει ότι η τιμή του συγκεκριμένο πεδίο είναι πιθανότατα λάθος. Επίσης αν c_j = {1,1...,1} το μοντέλο ονομάζεται μοντέλο ελάχιστου αριθμού πεδίων για διόρθωση (MFI). Η σχέση (3.1.1) ελαχιστοποιεί το σταθμισμένο άθροισμα όλων των πεδίων της έρευνας γιατί το μοντέλο επιδιώκει την εύρεση του ελάχιστου αριθμού σταθμισμένων πεδίων για διόρθωση. Η σχέση (3.1.2) δηλώνει ότι αν μια εγγραφή χάνει κάποιο κανόνα τότε ένα τουλάχιστον πεδίο του κανόνα αυτού, πρέπει να διορθωθεί. (Garfinkel Kunnathur and Liepins 1986)

Ένα παρόμοιο πρόβλημα κάλυψης συνόλου χρησιμοποιείται όπως θα δούμε στη συνέχεια της ενότητας, για την εύρεση κανόνων που θα χρησιμοποιηθούν στην παραγωγή κατά βάση νέων κανόνων.

3.2 Μέθοδος R.S.Garfinkel, A.S. Kunnathur, G.E.Liepins (GKL 1986)

Βασισμένοι στη μεθοδολογία των Fellegi-Holt (1976) οι R.S.Garfinkel, A.S. Kunnathur, G.E.Liepins (GKL 1986) εισήγαγαν δυο νέους αλγορίθμους για τον έλεγχο και τη διόρθωση εγγραφών. Οι νέοι αλγόριθμοι αντιμετωπίζουν τις παραπάνω διαδικασίες με λιγότερο υπολογιστικό κόστος, τόσο στην παραγωγή του πλήρους συνόλου κανόνων όσο και στην εύρεση των λαθών. Η μέθοδος GKL (1986) αφορά αποκλειστικά κατηγορικά δεδομένα.

Η ορολογία που χρησιμοποιείται στην περιγραφή των αλγορίθμων είναι ίδια με αυτή που αναλύσαμε στη μεθοδολογία Fellegi-Holt (1976) (ενότητα *2.2*), καθώς επίσης ισχύει το *Λήμμα 1* της παραγωγής έμμεσων κανόνων και οι χαρακτηρισμοί των κανόνων (έμμεσοι, κατά βάση νέοι, περιττοί, μέγιστοι).

Αξίζει να υπενθυμίσουμε αρχικά, τα τρία βασικά βήματα της μεθόδου Fellegi-Holt (1976):

1. Παραγωγή του πλήρους συνόλου των κανόνων, διαδικασία η οποία μπορεί να μειωθεί όσον αφορά το χρόνο υπολογισμού με τις προτάσεις που αναφέρονται στην παράγραφο *2.4.4* του προηγούμενου κεφαλαίου.

2. Έλεγχος κάθε εγγραφής για το αν ικανοποιεί τους κανόνες και εφαρμογή του προβλήματος κάλυψης συνόλου σε κάθε μια για να βρεθεί ο ελάχιστος αριθμός πεδίων που χρειάζονται διόρθωση.
3. Διόρθωση των πεδίων που βρέθηκαν στο βήμα 2.

3.2.1 Η έννοια του επαρκούς συνόλου κανόνων (GKL, 1986)

Ενώ λοιπόν με τη μέθοδο των Fellegi-Holt η λύση του μοντέλου προσδιορίζονταν αν γνωρίζαμε το πλήρες σύνολο των κανόνων ορθότητας, οι GKL (1986) επιτυγχάνουν την αντιμετώπιση του προβλήματος αναγνωρίζοντας ένα επαρκές σύνολο κανόνων, υποσύνολο του πλήρους συνόλου αλλά ισοδύναμο με αυτό.

Λέγοντας ότι τα δύο σύνολα κανόνων $\overline{E'}, \overline{E}$ είναι **ισοδύναμα** (equivalent) εννοούμε ότι $P(\overline{E'}) = P(\overline{E})$ (όπου $P(\overline{E}) = \mathbf{U}_{E^k \in \overline{E}} P(E^k)$) καθώς και η λύση του μοντέλου ελαχίστων σταθμισμένων πεδίων για διόρθωση είναι ίδια για οποιαδήποτε επιλογή ανάμεσα στα δύο σύνολα. Χρησιμοποιείται η παύλα για να δηλώσει σύνολα κανόνων.

Θεώρημα 3 **(GKL 1986)**: Έστω $\overline{E_*}(i)$ είναι το σύνολο των κανόνων που παράγονται από ένα υποσύνολο των ρητά ορισμένων κανόνων με γεννήτορα πεδίο το i. Επίσης έστω $\overline{E_*}(i, j)$ είναι το σύνολο των κανόνων που παράγονται, παίρνοντας αρχικά το i πεδίο ως γεννήτορα και στη συνέχεια το j πεδίο, και ως συμβάλλοντες κανόνες ένα υποσύνολο των ρητά ορισμένων κανόνων μαζί με το $\overline{E_*}(i)$. Τότε $N_{ij} = N_{ji}$ όπου N_{ij} είναι το μη-περιττό (π.χ. μέγιστο) υποσύνολο του $\overline{E_*}(i, j)$.

Σύμφωνα με το Θεώρημα 3 μειώνονται οι υπολογισμοί που απαιτούνται για την παραγωγή νέων κανόνων εφόσον οι μεταθέσεις των πεδίων δεν παίζουν ρόλο όπως θα περιγράψουμε στη συνέχεια, και διακρίνονται για κάθε σύνολο συμβαλλόντων κανόνων εκείνα τα υποσύνολα που πιθανόν θα παράγουν μέγιστους κατά βάση νέους κανόνες. Το παραπάνω αποτελεί και το πλεονέκτημα της μεθόδου GKL (1986) έναντι της μεθόδου FH (1976)

Βασισμένοι στο Θεώρημα 3 προκύπτει ότι για οποιαδήποτε μετάθεση δ και γ των πεδίων ισχύει $N_d(1,2,..,k) = N_g(1,2,...,k)$ και τελικά συμπεραίνεται ότι το σύνολο των κανόνων $N_d(1,2,..,n)$ είναι **επαρκές** (sufficient) για το πλήρες σύνολο των κανόνων για οποιαδήποτε μετάθεση των πεδίων (GKL 1986). Δηλαδή αν γνωρίζουμε το σύνολο των κανόνων που

ορίζεται από τη μετάθεση των πεδίων *δ*, δεν είναι ανάγκη να παράγουμε και τα υπόλοιπα σύνολα που δημιουργούνται από τις αντίστοιχες μεταθέσεις των πεδίων.

3.2.2 Διαδικασία παραγωγής του επαρκούς συνόλου κανόνων

Το πλήρες σύνολο κανόνων σύμφωνα με τους Fellegi-Holt (1976) αποτελείται από το σύνολο των ρητά ορισμένων κανόνων και το σύνολο των κατά βάση νέων κανόνων. Για την παραγωγή του πλήρους συνόλου κανόνων σύμφωνα με τη μέθοδο GKL (1986) χρειαζόμαστε μόνο τους μέγιστους κατά βάση νέους κανόνες. Αν λοιπόν $\overline{E}_C$ είναι το πλήρες σύνολο κανόνων, υπάρχει το επαρκές σύνολο κανόνων $\overline{E}_S$ (δηλαδή το σύνολο $N_d(1,2,..,n)$) ισοδύναμο με το $\overline{E}_C$ και υποσύνολο αυτού που αποτελείται από τους μέγιστους έμμεσους κανόνες.

Ο αλγόριθμος παραγωγής του επαρκούς συνόλου κανόνων θα περιγραφεί με τη βοήθεια ενός παραδείγματος.

Παράδειγμα 3 (GKL 1986)

Πίνακας 10

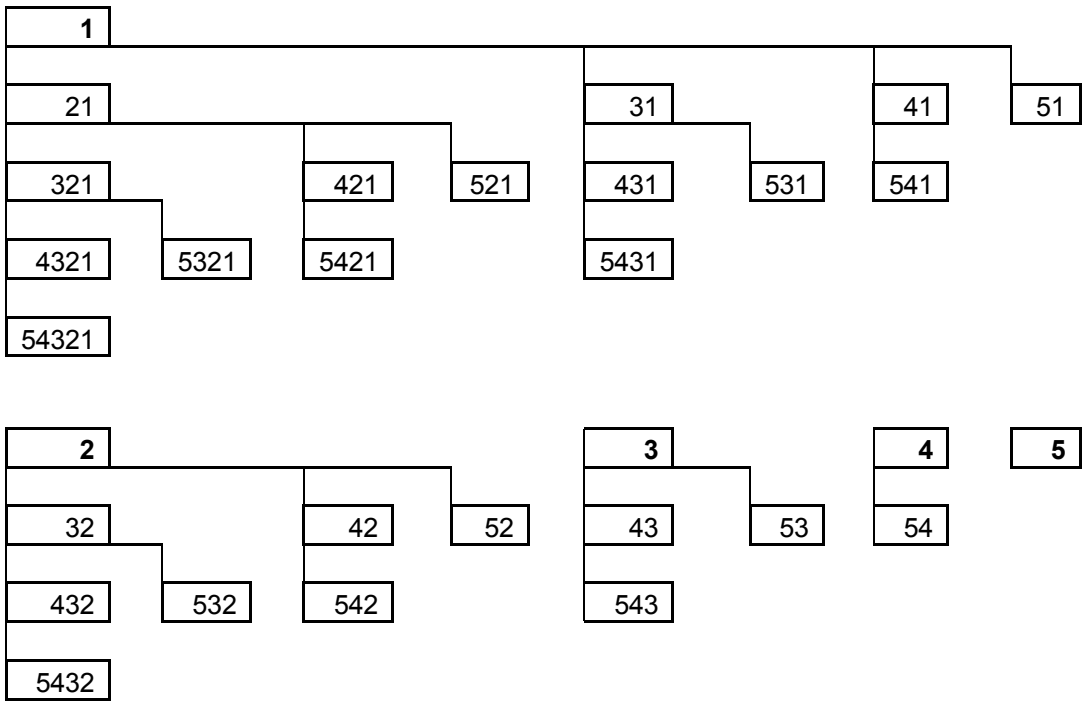
	$R_1=\{0,1\}$	$R_2=\{0,1,2\}$	$R_3=\{0,1\}$	$R_4=\{0,1,2,3\}$	$R_5=\{0,1,2\}$	$R_6=\{0,1,2,3\}$
E^1	R_1	0,1	0	R_4	0,1	R_6
E^2	1	R_2	1	0,1	R_5	2,3
E^3	0	1,2	R_3	1,2,3	R_5	R_6
E^4	R_1	0,2	R_3	R_4	R_5	0,1
E^5	1	R_2	R_3	0	1,2	R_6

Έστω ότι έχουμε έρευνα που κάθε εγγραφή της αποτελείται από έξι πεδία και για την οποία είναι γνωστοί πέντε ρητά ορισμένοι κανόνες.

Αλγόριθμος 1 (GKL 1986)

Βήμα 1 : Δημιουργώ το «δάσος» κωδικών πεδίων (Field Code Forest FCF) για τα έξι πεδία της έρευνας:

Δάσος κωδικών πεδίων (Field Code Forest)



Σχήμα 1

Το παραπάνω «δάσος» πεδίων δημιουργήθηκε με την προϋπόθεση της ύπαρξης 5 μόνο πεδίων, αν και στο παράδειγμά μας τα πεδία είναι 6 αλλά το «δάσος» ήταν ιδιαίτερα μεγάλο. Από το σχήμα διαπιστώνουμε ότι το δάσος αποτελείται από 5 δέντρα πεδίων (τονίσαμε ότι σχηματίζουμε το δάσος για 5 πεδία). Ξεκινώντας από το πρώτο πεδίο διακλαδιζόμαστε ακολουθώντας την αύξουσα αρίθμηση της μορφής του σχήματος μέχρι να δημιουργήσουμε όλους τους δεσμούς λαμβάνοντας υπόψη ότι δεν παίζει ρόλο η μετάθεση όπως τόνισαν οι GKL (1986). Οι αριθμοί στους **δεσμούς** (nodes) αναφέρονται στους συνδυασμούς των πεδίων. Ο πρώτος αριθμός (από αριστερά) σε κάθε δεσμό είναι ο γεννήτορας πεδίο που θα χρησιμοποιηθεί για να παράγει έμμεσους κανόνες με συμβάλλοντες κανόνες, τους κανόνες που έχουν παραχθεί μέχρι τη στιγμή της άφιξης στο συγκεκριμένο δεσμό.

Βήμα 2 : Η διαδικασία παραγωγής κανόνων αρχίζει από τον πρώτο δεσμό (γεννήτορας πεδίο 1). Σε αυτό το δεσμό το σύνολο των κανόνων που χρησιμοποιώ, είναι το σύνολο των ρητά ορισμένων κανόνων $\overline{E^0}$. Δημιουργώ τον **πίνακα κωδικών πεδίων** (Field Code Matrix, FCM) του οποίου οι γραμμές είναι όλες οι πιθανές τιμές του πεδίου και στήλες οι κανόνες που περιέχουν το συγκεκριμένο πεδίο του δεσμού που βρισκόμαστε (δηλαδή το πεδίο 1). Τα στοιχεία του πίνακα γενικά είναι οι τιμές :

f_{st}^{i} = 1 στην περίπτωση που ο κανόνας $E^t \in \overline{E_t}$ περιέχει το *s* στοιχείο του πεδίου *i*

= 0 αλλού

όπου $\overline{E_t}$ είναι το σύνολο των κανόνων που έχει παραχθεί ως τη στιγμή της άφιξης στο συγκεκριμένο δεσμό. Στην προκειμένη περίπτωση αυτό είναι το σύνολο $\overline{E^0}$. Ο γεννήτορας πεδίο περιέχεται μόνο στους κανόνες *E2, E3, E5*.

Πίνακας κωδικών πεδίων (*Πίνακας 11*)

R_1	E_2	E_3	E_5
0	0	1	0
1	1	0	1

Δηλαδή το στοιχείο 0 του γεννήτορα πεδίου 1 υπάρχει μόνο στον κανόνα *E3* , και το στοιχείο 1 υπάρχει στους κανόνες *E2* , *E5*.

Βήμα 3 : Χρησιμοποιώντας τα στοιχεία του παραπάνω πίνακα βρίσκω όλες τις **βασικές καλύψεις** (prime covers) λύνοντας το πρόβλημα κάλυψης συνόλου με περιορισμούς :

$$\sum_{E^t \in \overline{E_t}} f_{st}^i x_t \geq 1 \quad \text{με } s = 0,1,\ldots,|R_i|-1$$

$x_t = 0,1$ (1 αν χρησιμοποιούμε τον κανόνα, 0 διαφορετικά) , $E^t \in \overline{E_t}$

Οι λύσεις του παραπάνω προβλήματος ορίζουν ένα σύνολο κανόνων που πιθανόν να παράγουν μέγιστους κατά βάση νέους κανόνες.

Για το πεδίο 1 οι παραπάνω περιορισμοί είναι:

$0 \times x_2 + 1 \times x_3 + 0 \times x_5 \geq 1 \Rightarrow x_3 \geq 1$

$1 \times x_2 + 0 \times x_3 + 1 \times x_5 \geq 1 \Rightarrow x_2 + x_5 \geq 1$

Τότε οι βασικές καλύψεις είναι (χ_2, χ_3, χ_5) =(1,1,0) , (0,1,1)

Παρατήρηση: Το σύνολο (1,1,0) σημαίνει ότι θα χρησιμοποιήσω στο επόμενο βήμα τους κανόνες *E2* και *E3* ως συμβάλλοντες κανόνες με γεννήτορα πεδίο το 1. Οι κανόνες *E2* και *E3* αποτελούν το υποσύνολο των κανόνων *E2, E3, E5* (που αναφέραμε στην παράγραφο *3.2.1*) που πιθανόν να παράγουν ένα μέγιστο κατά βάση νέο κανόνα. Επίσης οι κανόνες *E2* και *E3* ονομάζονται βασική κάλυψη γιατί περιέχουν μαζί όλες τις τιμές του γεννήτορα πεδίου $\left(E_{21} \bigcup E_{31} = R_1\right)$ και για τους οποίους δεν υπάρχουν υποσύνολά τους με την ίδια ιδιότητα. Τα ίδια ισχύουν και για τους κανόνες *E3, E5*.

Βήμα 4 : Για κάθε βασική κάλυψη που βρήκα στο Βήμα 3 παράγω έμμεσους κανόνες.

Από τους κανόνες E_2 και E_3 στο δεσμό 1 παράγεται (Λήμμα 1, ενότητα *2.4.1*) ο κατά βάση νέος κανόνας:

$E_6 = R_1 \times \{1,2\} \times \{1\} \times \{1\} \times R_5 \times \{2,3\}$

Από τους κανόνες E_3 και E_5 στο δεσμό 1 (Λήμμα 1 ενότητα *2.4.1*) δεν παράγεται έμμεσος κανόνας γιατί:

$E_7 = R_1 \times \{1,2\} \times R_3 \times \otimes \times \{1,2\} \times R_6$

Παρατήρηση: Στην παραγωγή των έμμεσων κανόνων πρέπει να τηρήσουμε τα εξής

- Στην περίπτωση παραγωγής ενός κατά βάση νέου κανόνα που επικρατεί ενός ήδη υπάρχοντα κανόνα τον αντικαθιστά.
- Αν σε κάποιο δεσμό δεν παραχθεί κανένας μη-περιττός κανόνας σταματάμε την διαδικασία του αλγορίθμου χωρίς να «επισκεπτόμαστε» τους επόμενους δεσμούς.

Βήμα 5 : Τους κατά βάση νέους κανόνες που παράχθηκαν στο Βήμα 4 τους προσθέτουμε στο σύνολο των κανόνων με τους οποίους άρχισε η διαδικασία στο συγκεκριμένο δεσμό. Το νέο σύνολο κανόνων χρησιμοποιείται στην παραγωγή κανόνων στον επόμενο δεσμό.

Στο παράδειγμά μας, η διαδικασία άρχισε με το σύνολο $\overline{E^0}$ (των ρητά ορισμένων κανόνων) στο οποίο προσθέτουμε τον E_6. Το νέο σύνολο $\overline{E_t} = \overline{E^0} \cup E_6$ αποτελεί το εισερχόμενο σύνολο κανόνων στο δεσμό 21 και ο αλγόριθμος επαναλαμβάνεται αρχίζοντας από το βήμα 2.

Με τον ίδιο τρόπο, ο αλγόριθμος συνεχίζεται σε όλους τους δεσμούς του «δάσους» και για όλα τα πεδία μέχρι να παραχθεί το επαρκές σύνολο κανόνων. Αφού καταλήξουμε στο επαρκές σύνολο των κανόνων το χρησιμοποιούμε για να λύσουμε το πρόβλημα εντοπισμού των πεδίων που χρειάζονται διόρθωση από το μοντέλο ελάχιστων σταθμισμένων πεδίων για διόρθωση, για κάθε τρέχουσα εγγραφή.

3.2.3 Αλγόριθμος 2 (Cutting Plane Algorithm, GKL, 1986)

Ο αλγόριθμος 1 ελαχιστοποιεί τον αριθμό των κανόνων που παράγονται έτσι ώστε το πρόβλημα κάλυψης συνόλου, που λύνει το πρόβλημα του εντοπισμού λαθών να εφαρμόζεται σε μικρότερο αριθμό κανόνων. Όμως σε περιπτώσεις που οι έρευνες περιέχουν πολλούς κανόνες και πεδία, το επαρκές σύνολο κανόνων μπορεί να είναι επίσης μεγάλο και το πρόβλημα κάλυψης συνόλου δύσκολο να επιλυθεί. Οι GKL (1986) δημιούργησαν έναν ακόμη

αλγόριθμο που παράγει μόνο εκείνους του κανόνες που σχετίζονται με την τρέχουσα εγγραφή και μπορεί να αντιμετωπίσει ανάλογες περιπτώσεις.

Ο αλγόριθμος 2 (GKL, 1986) θα παρουσιαστεί με βάση το παράδειγμα 3 της παραγράφου *3.2.2* : Έστω η τρέχουσα εγγραφή $y^0 = (1,0,0,0,1,0)$ και c =(5,3,1,4,2,4) οι σταθερές εμπιστοσύνης για κάθε πεδίο. Η τρέχουσα εγγραφή χάνει τους κανόνες *E1, E4, E5,* γράφουμε τότε $\overline{E_F} = \{E_1, E_4, E_5\}$.

Βήμα 1: Βρίσκω λύση στο μοντέλο ελάχιστων σταθμισμένων πεδίων για διόρθωση MWFI την $x^* = (0,1,0,0,1,0)$

$$\text{Ελαχιστοποιώ } 5x_1 + 3x_2 + 1x_3 + 4x_4 + 2x_5 + 4x_6$$

$$\text{υπό τους περιορισμούς } \begin{aligned} x_2 + x_3 + x_5 &\geq 1 \\ x_2 + x_6 &\geq 1 \\ x_1 + x_4 + x_5 &\geq 1 \end{aligned}$$

όπου $x_j = \begin{cases} 1 \\ 0 \end{cases} \quad j = 1,2,...,6$.

Η λύση των ανισοτήτων είναι $x_2 = x_5 = 1$. Δηλαδή αυτά είναι τα πεδία που πρέπει να διορθωθούν.

Βήμα 2: Ορίζω το σύνολο $J^* = (2,5)$ με στοιχεία τα πεδία που χρειάζονται διόρθωση. Λαμβάνω υπόψιν τις $|R_2| * |R_5| = 9$ εγγραφές που στα πεδία 2,5 έχουν όλες τις δυνατές τιμές και στα υπόλοιπα πεδία τις υπάρχουσες τιμές τις τρέχουσας εγγραφής.

Πίνακας 12

y_1	(100000)	Χάνει τους *E4, E1*
y_2	(100010)	Χάνει τους *E1, E5*
y_3	(100020)	Χάνει τους *E4, E5*
y_4	(110000)	Χάνει τους *E1*
y_5	(110010)	Χάνει τους *E1, E5*
y_6	(110020)	Χάνει τους *E5*
y_7	(120000)	Χάνει τους *E4*
y_8	(120010)	Χάνει τους *E4, E5*
y_9	(120020)	Χάνει τους *E4, E5*

Κάθε μια από τις 9 εγγραφές δεν ικανοποιεί κανένα κανόνα από τους E_1, E_4, E_5. Αν κάποια από τις εγγραφές ικανοποιούσε ένα τουλάχιστον κανόνα θα είχαμε τη λύση στο πρόβλημα MWFI. Αλλιώς πάω στο επόμενο βήμα.

Βήμα 3: Βρίσκω όλες τις βασικές καλύψεις v^0 που ικανοποιούν τα εξής:

$$Qn \geq 1 \qquad (3.2.3)$$

$$v_i = 0,\ 1$$

για $i = 1,2,\ldots,5$, όπου $Q=(q_{ij})$ είναι ο πίνακας με γραμμές τις εγγραφές του βήματος 2 και στήλες τους κανόνες. Όπου v είναι το διάνυσμα στήλη ή $v=(v_1, v_2, v_3, v_4, v_5)^T$, όπου κάθε στοιχείο του διανύσματος ανταποκρίνεται σε ένα κανόνα.

$q_{ij} = 1$ αν ο E_i κανόνας δεν ικανοποιείται από την j εγγραφή του βήματος 2

$= 0$ διαφορετικά.

$$Q = \begin{bmatrix} \mathbf{10010} \\ \mathbf{10001} \\ \mathbf{00011} \\ \mathbf{10000} \\ \mathbf{10001} \\ \mathbf{00001} \\ \mathbf{00010} \\ \mathbf{00011} \\ \mathbf{00011} \end{bmatrix}$$

τότε η σχέση (3.2.3) καταλήγει στους περιορισμούς

$$\begin{aligned} v_1 + v_4 &\geq 1 \\ v_1 + v_5 &\geq 1 \\ v_4 + v_5 &\geq 1 \\ v_1 &\geq 1 \\ v_1 + v_5 &\geq 1 \\ v_5 &\geq 1 \\ v_4 &\geq 1 \\ v_4 + v_5 &\geq 1 \\ v_4 + v_5 &\geq 1 \end{aligned}$$

των οποίων η λύση είναι $v_1 = v_4 = v_5 = 1$

Προφανώς, θα μπορούσαμε να διαπιστώσουμε τη λύση άμεσα γιατί οι εγγραφές y_4, y_7, y_6 χάνουν μόνο έναν κανόνα η κάθε μια, τους E_1, E_4, E_5 αντίστοιχα.

Ορίζω το σύνολο $I^0 = \{i / v_i^0 = 1\}$. Το σύνολο αυτό στο παράδειγμά μας είναι $I^0 = \{1,4,5\}$

Βήμα 4: Παράγω το νέο κανόνα E^* ως εξής

- $F_j^* = \mathbf{I}_{i \in I^0} F_j^i$ όταν $j \notin J^*$
- $F_j^* = R_j$ όταν $j \in J^*$

Ο νέος κανόνας που παράγεται είναι $E^* = \{1\} \times R_2 \times \{0\} \times \{0\} \times R_5 \times \{0,1\}$ ο οποίος είναι «κατά βάση νέος» κανόνας. Ο αλγόριθμος αρχίζει από το πρώτο βήμα με εισερχόμενο

σύνολο κανόνων το $\overline{E_F} = \overline{E_F} \cup \{E^*\}$ δηλαδή προσθέτουμε στους περιορισμούς του προβλήματος κάλυψης συνόλου τον εξής

$$x_2 + x_3 + x_4 + x_6 \geq 1$$

που προήλθε από τον κανόνα E^*

Συνεχίζοντας τα βήματα αναπροσαρμοσμένα στα νέα αποτελέσματα καταλήγουμε στην εξής λύση $x^* = (0,1,1,0,1,0)$ όπου στο τρίτο βήμα η εγγραφή y= (1,1,1,0,0,0) ικανοποιεί όλους τους κανόνες.

3.2.4 Αξιολόγηση των αλγορίθμων και συγκρίσεις.

Με σκοπό την βελτίωση της μεθόδου Fellegi-Holt (1976) και τη μείωση του υπολογιστικού κόστους -που οι τελευταίοι δεν είχαν υπολογίσει- οι Garfinkel, Kunnathur, Liepins (1986) δημιούργησαν δυο νέους αλγορίθμους, ικανούς να ελέγξουν και να διορθώσουν κατηγορικά δεδομένα σε σαφώς λιγότερο χρόνο. Συγκρίνοντας την προσπάθεια των GKL (1986) με την μέθοδο των Fellegi-Holt (1976) μπορούμε να διαπιστώσουμε τα εξής:

- ο αλγόριθμος 1 των GKL (1986) παράγει ένα επαρκές σύνολο κανόνων, υποσύνολο τους πλήρους συνόλου και ισοδύναμο με αυτό. Στην παραγωγή λοιπόν, του επαρκούς συνόλου χρησιμοποιούνται υποσύνολα των συμβαλλομένων κανόνων (βασικές καλύψεις) που πιθανόν να παράγουν μέγιστα κατά βάση νέους κανόνες. Σε αντίθεση με τη μέθοδο Fellegi-Holt (1976), δοκιμάζονται όλοι οι πιθανοί συνδυασμοί των ρητών και έμμεσων κανόνων για την παραγωγή του πλήρους συνόλου. Προφανώς η μέθοδος GKL με την εφαρμογή του πρώτου αλγορίθμου θα δώσει συντομότερα λύσεις από ότι η μέθοδος Fellegi-Holt.
- Ο αλγόριθμος 1 των GKL (1986) δεν παράγει επιπλέον κανόνες από οποιαδήποτε μετάθεση των πεδίων. Δηλαδή εφόσον έχουμε παράγει τους κανόνες στο δεσμό 321 δεν χρειάζεται να παράγουμε κανόνες από τους δεσμούς 312 αφού η μετάθεση δεν παίζει ρόλο. Αυτό δεν ισχύει στην μέθοδο Fellegi-Holt (1976) που επιτρέπει όλους τους συνδυασμούς των πεδίων.

Όσον αφορά την σύγκριση μεταξύ των δύο αλγορίθμων των GKL επισημαίνουμε τα εξής:

- Ο πρώτος αλγόριθμος καταλήγει στην εύρεση των πεδίων της κάθε εγγραφής που χρειάζεται διόρθωση, ενώ ο δεύτερος καταλήγει σε εγγραφή ήδη διορθωμένη, από τη εφαρμογή του αλγορίθμου.
- Ο δεύτερος αλγόριθμος βγάζει συμπεράσματα πιο γρήγορα από τον πρώτο σε περιπτώσεις μεγάλου πλήθους πεδίων και κανόνων. Το παράδειγμα που παρουσιάστηκε παραπάνω εφαρμόστηκε και στους δύο αλγορίθμους από τους οποίους ο πρώτος χρειάστηκε να παράγει 13 «κατά βάση νέους» κανόνες (GKL 1986) για να εξάγει το αποτέλεσμα ενώ ο δεύτερος μόλις 2. Στην περίπτωση όμως που ο χρόνος παραγωγής του επαρκούς συνόλου στον πρώτο αλγόριθμο είναι μικρός, δηλαδή το επαρκές σύνολο που παράγεται αποτελείται από λίγους κανόνες τότε είναι προτιμότερο να χρησιμοποιήσουμε τον πρώτο αλγόριθμο.

Παρατηρήσεις

1. Στον δεύτερο αλγόριθμο θα μπορούσαμε να επέμβουμε με σκοπό να μειώσουμε το χρόνο μέχρι την εξαγωγή του αποτελέσματος. Στο πρώτο βήμα λοιπόν, έχουμε τη δυνατότητα να ορίσουμε ένα όριο κόστους (τιμή του διανύσματος c), έτσι ώστε οι εγγραφές (λύσεις του προβλήματος κάλυψης συνόλου) που θα το ξεπερνούσαν θα διαγράφονταν. Με τον τρόπο αυτό θα μειώνονταν και ο αριθμός παραγωγής εγγραφών στο δεύτερο βήμα.
2. Στην περίπτωση που ο δεύτερος αλγόριθμος έληγε στο δεύτερο βήμα θα παρήγαγε τη βέλτιστη λύση. Αν όχι, θα μπορούσε να παράγει περισσότερες από μια βέλτιστες λύσεις και για την επιλογή της καλύτερης θα χρειαζόμασταν στατιστικά κριτήρια.

3.3 Εισαγωγή του αλγορίθμου EG από τον Winkler (1997)

Αναλύοντας εκτεταμένα την μεθοδολογία των Fellegi-Holt (1976) στο κεφάλαιο 2 διαπιστώσαμε ότι το Θεώρημα 2 προσδιορίζει ακριβώς την περιοχή όπου βρίσκει λύση το πρόβλημα του εντοπισμού λαθών χωρίς όμως να λαμβάνεται υπόψη το υπολογιστικό κόστος, γεγονός που κατάφεραν να «μετριάσουν» το 1986 οι Garfinkel, Kunnathur και Liepins το 1986 για κατηγορικά δεδομένα. Επειδή όμως το πρόβλημα του εντοπισμού λαθών ήταν ακόμη δύσκολο να επιλυθεί, ο Winkler (1997) παρουσίασε ένα νέο αλγόριθμο EG για την

παραγωγή των έμμεσων κανόνων. Ο αλγόριθμος EG καταφέρνει να παράγει όλους τους μέγιστα έμμεσους κανόνες - γεγονός που δεν πέτυχε ο αλγόριθμος 1 των GKL(1986) - που επαρκούν για να λυθεί το πρόβλημα του εντοπισμού λαθών.

Η ονοματολογία παραμένει ίδια όπως αναπτύχθηκε προηγούμενα στη μέθοδο των Fellegi-Holt (1976), οι κανόνες για τα κατηγορικά δεδομένα γράφονται πάντα στη κανονική τους μορφή και ισχύει το *Λήμμα 1* παραγωγής έμμεσων κανόνων από ένα σύνολο συμβαλλόντων κανόνων. Το πλήρες σύνολο των κανόνων όπως έχει οριστεί από τους GKL (1986) αποτελείται από το σύνολο των ρητά ορισμένων κανόνων και το σύνολο των μέγιστα κατά βάση νέων κανόνων. Ενώ όμως οι GKL (1986) χρησιμοποίησαν την παραδοχή ότι η μετάθεση των γεννητόρων πεδίων σε ένα δεσμό (node) δεν παίζει ρόλο στην παραγωγή κανόνων, αυτό δεν υιοθετήθηκε από τον Winkler(1997) για τον οποίο παίζει ρόλο η σειρά των πεδίων, για παράδειγμα οι κανόνες που θα παραχθούν από το δεσμό (ijk) είναι διαφορετικοί από αυτούς που θα παραχθούν από το δεσμό (ikj) .Οι κανόνες που παράγονται από τον δεσμό (i) λέγονται πρώτου επιπέδου έμμεσοι κανόνες και οι δεσμοί ονομάζονται πρώτου επιπέδου.

Για να αποδείξει ο Winkler (1997) ότι η μετάθεση των πεδίων παίζει ρόλο στην παραγωγή των έμμεσων κανόνων στηρίχθηκε στο παράδειγμα 3 (ενότητα *3.2.2*). Αρχικά άλλαξε τη σειρά των πεδίων του παραδείγματος ως εξής :

1->3 , 2->4 , 3->5 , 4->6 , 5->1 , 6->2

Η αλλαγή αυτή δεν επηρεάζει την κύρια διαδικασία παραγωγής κανόνων. Ενώ λοιπόν ο αλγόριθμος 1 GKL (1986) παρήγαγε 13 έμμεσους κανόνες, αυτοί δεν είναι όλοι οι μέγιστοι κανόνες που επαρκούν για το πλήρες σύνολο όπως ισχυρίστηκαν. Από την άλλη πλευρά, ο Winker (1997) με τον αλγόριθμο EG παράγει όλους τους μέγιστους κανόνες. Συγκεκριμένα, ο αλγόριθμος του Winkler (1997) παρήγαγε ένα κανόνα από το δεσμό 125 ο οποίος είναι μέγιστος. Στους 13 κανόνες που προήλθαν από τον αλγόριθμο 1 GKL (1986) δεν υπάρχει ο κανόνας αυτός. Υπενθυμίζουμε ότι για τους GKL η μετάθεση των πεδίων δεν παίζει ρόλο, που σημαίνει ότι οι δεσμοί 125 και 152 παράγουν τους ίδιους κανόνες. Επίσης, όταν ο αλγόριθμος 1 των GKL (1986) έχει «επισκεφτεί» το δεσμό 125, σημαίνει ότι έχει «επισκεφτεί» και το δεσμό 12. Όμως στους GKL (1986) ο αλγόριθμος δεν επισκέφτηκε το δεσμό 12 (γιατί είχε ήδη παράγει κανόνες από τον 21). Για τον λόγο αυτό δεν υπάρχει στο πλήρες σύνολο κανόνων ο μέγιστος κανόνας που προκύπτει από τον δεσμό 125.

Ο EG αλγόριθμος του Winkler (1997) βασίζεται στα επόμενα Λήμματα που προήλθαν από τη Θεωρία των FH (1976) :

Λήμμα 2 (Winkler 1997): Για την παραγωγή του πλήρους συνόλου των κανόνων E^c, το σύνολο των ρητά ορισμένων κανόνων E^o πρέπει να αντικατασταθεί από το σύνολο E^{om}, στο οποίο κάθε κανόνας είναι μέγιστος και επικρατεί επί τουλάχιστον ενός κανόνα στο σύνολο E^o.

Το σύνολο E^{om} είναι ισοδύναμο με το σύνολο E^o αφού παράγει το ίδιο σύνολο μέγιστων κανόνων με το πλεονέκτημα ότι παράγει μικρότερο αριθμό περιττών κανόνων που σημαίνει μικρότερο υπολογιστικό κόστος. Συνεπώς αν η διαδικασία παραγωγής κανόνων αρχίσει με το σύνολο E^{om} θα λάβουμε το πλήρες σύνολο κανόνων E^c .

Το σύνολο E^{om} μπορεί να παραχθεί ως εξής: αρχικά χρησιμοποιώ τους ρητά ορισμένους κανόνες για να παράγουν πρώτου επιπέδου κανόνες, στη συνέχεια αντικαθιστώ κάθε ρητά ορισμένο κανόνα με πρώτου επιπέδου έμμεσους κανόνες που επικρατούν επί αυτού και επαναλαμβάνω την προηγούμενη διαδικασία μέχρι να μην υπάρχουν άλλοι περιττοί κανόνες.

Λήμμα 3 (Winkler 1997): Έστω το σύνολο E^g των κανόνων που παράγουν τον κανόνα E^i με γεννήτορα πεδίο το j. Επίσης έστω ένα υποσύνολο του E^g το E^{g^*} που παράγει τον κανόνα E^{i^*} με το ίδιο γεννήτορα πεδίο. Τότε ο κανόνας E^{i^*} «επικρατεί» επί του E^i που σημαίνει $P(E^i) \subseteq P(E^{i^*})$ δηλαδή ο κανόνας E^i είναι περιττός

Αλγόριθμος EG Winkler (1997)

1. αντικαθιστώ το σύνολο των ρητά ορισμένων κανόνων E^o με το ισοδύναμο σύνολο των μέγιστων κανόνων E^{om},
2. μεταθέτω τους δεσμούς του δέντρου σε όλους τους δυνατούς συνδυασμούς,
3. σε κάθε δεσμό και για κάθε έμμεσο κανόνα επιλέγω τους κανόνες αυτούς που θα περάσουν στον επόμενο δεσμό για την παραγωγή νέων,
4. μέσα σε κάθε δεσμό και για κάθε νέο έμμεσο κανόνα του τρέχοντα δεσμού παράγω μέγιστους κανόνες.

Οι **διαφορές του αλγορίθμου EG με το αλγόριθμο 1 των GKL(1986)** είναι :

Οι δεσμοί μετατίθενται σε όλους τους δυνατούς συνδυασμούς ενώ στον GKL μόνο κατά τη σειρά i<j<k.

Στο τρίτο βήμα επιλεγώ τους κανόνες που θα περάσουν στο επόμενο δεσμό για την παραγωγή νέων, ενώ στον GKL δεν υπάρχει ανάλογος περιορισμός με αποτέλεσμα το κόστος υπολογισμού να αυξάνεται εκθετικά. (Winkler 1997)

Αυτό που παραμένει ίδιο και στους δυο αλγορίθμους είναι η διαδικασία παραγωγής νέων κανόνων σε κάθε αρχικό δεσμό αφού πρέπει να γίνουν όλοι οι συνδυασμοί κανόνων και γεννητόρων πεδίων για την παραγωγή έμμεσων κανόνων.

3.4 Βελτίωση του Αλγορίθμου EG και δημιουργία νέου EGE (Winkler 1998)

Ο Winkler (1998) επιμένοντας στην μείωση του υπολογιστικού κόστους και στην επεξεργασία μεγαλύτερου μεγέθους δεδομένων, βελτίωσε τον προηγούμενο αλγόριθμο με την δημιουργία του EGE αλγόριθμου. Ο νέος αλγόριθμος μπορεί να παράγει όλους τους μέγιστα έμμεσους κανόνες, όταν αυτοί δεν ακολουθούν συγκεκριμένα πρότυπα (skin patterns), σε ταχύτερους ρυθμούς. Το πλεονέκτημα του αλγορίθμου είναι ότι μπορεί να υπολογίσει τον χρόνο ολοκλήρωσης της παραγωγής των νέων κανόνων.

Ο Winkler χρησιμοποιώντας το επόμενο λήμμα, και λαμβάνοντας υπόψη τα Λήμματα 2 και 3 δημιούργησε τον EGE αλγόριθμο.

Λήμμα 4 (Winkler 1998): Κάθε μέγιστος κανόνας σε ένα ενδιάμεσο δεσμό παράγεται από ένα μέγιστο κανόνα του προηγούμενου δεσμού μαζί με ένα υποσύνολο ρητά ορισμένων κανόνων που χρησιμοποιήθηκαν στην παραγωγή ενός έμμεσου κανόνα στο αρχικό δεσμό.

Ο EGE αλγόριθμος σύμφωνα με τον Winkler (1998) καταφέρνει να παράγει όλους τους μέγιστους κανόνες με λιγότερο υπολογιστικό κόστος αφού χρησιμοποιεί τον μέγιστο κανόνα του προηγούμενου δεσμού με ένα υποσύνολο ρητά ορισμένων κανόνων που πέτυχε τη παραγωγή έμμεσου κανόνα σε αρχικό δεσμό. Ο αλγόριθμος EG (Winkler 1997) χρησιμοποιούσε τον έμμεσο κανόνα του προηγούμενου δεσμού με όλους τους ρητά ορισμένους κανόνες. Τέλος οι GKL (1986) χρησιμοποιούσαν υποσύνολα των έμμεσων κανόνων με όλους τους ρητά ορισμένους κανόνες. Ενώ όλα τα παραπάνω συμβαίνουν σε ενδιάμεσους δεσμούς (π.χ. ij ή ijk και τους υπόλοιπους απογόνους τους), στον αρχικό δεσμό

(i) το υπολογιστικό κόστος και για τους τρεις αλγορίθμους είναι το ίδιο αφού πρέπει να γίνουν όλοι οι δυνατοί συνδυασμοί κανόνων και πεδίων. Συνεπώς το συνολικό υπολογιστικό κόστος οφείλεται περισσότερο στην παραγωγή των κανόνων στους αρχικούς δεσμούς. (Winkler 1998)

3.5 Αυτοματοποιημένο σύστημα DISCRETE

Το αυτοματοποιημένο σύστημα ελέγχου DISCRETE το οποίο χρησιμοποιήθηκε από τη στατιστική υπηρεσία της Αμερικής είναι βασισμένο στη μεθοδολογία των Fellegi-Holt (1976). Ο αρχικός κώδικας δημιουργήθηκε από τον W.Winkler το 1995 και προγραμματίστηκε για κανόνες που ελέγχουν εντοπίζουν και διορθώνουν λάθη σε κατηγορικά δεδομένα. Το λογισμικό είναι γραμμένο σε FORTRAN και μπορεί να προσαρμοστεί και σε περιβάλλον UNIX και DOS.

Το λογισμικό χωρίζεται σε δύο επιμέρους προγράμματα:

1. *gened.for:* είναι υπεύθυνο για την παραγωγή των έμμεσων κανόνων οι οποίοι είναι απαραίτητοι για τη λύση του προβλήματος του εντοπισμού των λαθών. Στο πρόγραμμα εισάγεται το σύνολο των ρητά ορισμένων κανόνων και μέσω του αλγορίθμου παραγωγής κανόνων όπως περιγράφεται στο άρθρο του Winkler το 1995 εξάγει το πλήρες σύνολο κανόνων που αποτελείται από τους ρητά ορισμένους και το σύνολο των μέγιστων κανόνων. Τέλος το πρόγραμμα ελέγχει τη λογικά συνέπεια του πλήρους συνόλου.
2. *edit.for:* καθορίζει τον ελάχιστο αριθμό των πεδίων - που πρέπει να αλλάξουν -από τις εγγραφές που χάνουν τους κανόνες και στη συνέχεια πραγματοποιεί την διόρθωση αυτών. Τα εισερχόμενα αρχεία είναι το αρχείο των κανόνων που παράχθηκε από το προηγούμενο πρόγραμμα και το αρχείο των εγγραφών που θα υποβληθεί σε έλεγχο. Το μοντέλο διόρθωσης απεικονίζει τις τιμές του συνόλου των πεδίων που χρειάζονται διόρθωση έτσι ώστε οι νέες εγγραφές να περνούν όλους τους κανόνες. Σε ορισμένες εφαρμογές του συστήματος DISCRETE, ο αλγόριθμος που χρησιμοποιείται για τη διόρθωση των εγγραφών, περιλαμβάνει if-then-else κανόνες που είναι ορισμένοι από τους ειδικούς αναλυτές. Αυτοί οι κανόνες που ορίζονται για τη διόρθωση των εγγραφών είναι συγκεκριμένοι για κάθε έρευνα. Υπάρχει όμως και η περίπτωση να παράγονται αυτόματα όπως στη μέθοδο Fellegi-Holt (1976) εκτός και αν δεν είναι

αποδεκτοί από τους αναλυτές. Το εξερχόμενο αρχείο περιλαμβάνει συνοπτικά στατιστικά στοιχεία, το αρχείο των εγγραφών που υποβλήθηκε σε έλεγχο και λεπτομέρειες για την κάθε εγγραφή που διορθώθηκε. Winkler και Petkunas (1996)

Το αρνητικό είναι ότι επειδή το πρόγραμμα είναι γραμμένο σε FORTRAN είναι δύσκολα μετατρέψιμο για την εφαρμογή του σε άλλες έρευνες.

Σε μια έρευνα που αφορούσε την εργασιακή εμπειρία νέων γυναικών το σύστημα κατάφερε να επεξεργαστεί ένα μεγάλο όγκο δεδομένων που περιείχαν πολύπλοκα πρότυπα κανόνων. Χρησιμοποιώντας παλαιότερα συστήματα τα δεδομένα δεν κατάφεραν να περάσουν από έλεγχο εξαιτίας των πεπλεγμένων συνδυασμών τους.

3.6 Μείωση του προβλήματος κάλυψης συνόλου με τον αλγόριθμο Chen (1998)

Ο Chen (1998) συνειδητοποίησε την ανάγκη μείωσης του μεγέθους του προβλήματος κάλυψης συνόλου, ώστε η παραγωγή του πλήρους συνόλου των κανόνων να μην είναι τόσο χρονοβόρα διαδικασία, δημιουργώντας ένα νέο αλγόριθμο ο οποίος πλεονεκτεί σε σχέση με τους προγενέστερους, στο χρόνο υπολογισμού παραγωγής έμμεσων κανόνων. Ο αλγόριθμος του Chen καταφέρνει να παράγει όλες τις βασικές καλύψεις κανόνων που πιθανόν να παράγουν κατά βάση νέους κανόνες και αποφεύγει την παραγωγή περιττών καλύψεων.

3.6.1 Είδη προβλημάτων κάλυψης συνόλου

Υπενθυμίζουμε στο σημείο αυτό ότι η λύση του προβλήματος εντοπισμού λαθών λαμβάνεται από τη λύση ενός προβλήματος κάλυψης συνόλων, το λεγόμενο μοντέλο ελάχιστου αριθμού σταθμισμένων πεδίων για διόρθωση. Η λύση του μοντέλου αυτού μας δίνει τον ελάχιστο αριθμό πεδίων κάθε εγγραφής που πρέπει να διορθωθούν για να περάσει όλους τους κανόνες.

Έστω $\overline{E} = \{E^1, E^2, ..., E^m\}$ είναι το σύνολο των κανόνων που χάνονται από την εγγραφή y η οποία αποτελείται από n πεδία . Το πρόβλημα κάλυψης συνόλου εκφράζεται ως εξής:

$$\text{Ελαχιστοποιώ} \sum_{j=1}^{n} c_j x_j$$

$$\text{υπό την προϋπόθεση} \sum_{j=1}^{n} a_{ij} x_j \geq 1$$

$$\text{όπου} \quad a_{ij} = \begin{cases} 1, \; an \; o \;\; E^i \; periecei \;\; to \;\; pedio \;\; j \\ 0, \;\; a \, l \, l \, iwV \end{cases}$$

x_j = 1,0 αν το πεδίο χρειάζεται διόρθωση ή όχι αντίστοιχα και c_j είναι ένα μέτρο εμπιστοσύνης για το πεδίο j. Το σύνολο των κανόνων $\overline{E}$ λαμβάνεται από το πλήρες σύνολο των κανόνων και είναι αυτό που ορίστηκε αρχικά από τους FH (1976) δηλαδή το σύνολο των ρητά ορισμένων κανόνων μαζί με το σύνολο των «κατά βάση νέων» κανόνων. Βεβαίως ισχύει το Λήμμα 1 (παράγραφος *2.4.1*) για την παραγωγή των έμμεσων κανόνων και ο ορισμός των «κατά βάση νέων» κανόνων.

Εκτός όμως από το πρόβλημα κάλυψης συνόλου που χρησιμοποιείται για το πρόβλημα του εντοπισμού λαθών, υπάρχει και το πρόβλημα κάλυψης συνόλου που χρησιμοποιείται για την παραγωγή του πλήρους συνόλου κανόνων. Συγκεκριμένα το πρόβλημα κάλυψης συνόλου για το συγκεκριμένο πεδίο (γεννήτορα) i είναι:

Έστω $\{E^r \mid r \in S\}$ είναι το σύνολο των s κανόνων που περιέχουν το πεδίο i και n_i το πλήθος των στοιχείων του πεδίου i.

$$\text{Ελαχιστοποιώ} \sum_{r \in S} x_r$$

$$\text{υπό τη προϋπόθεση} \sum_{r \in S} g_{rj}^i x_r \geq 1 \qquad j = 1,2,...,n_i \tag{3.6.1}$$

$$\text{όπου} \; g_{rj}^{\;i} = \begin{cases} 1, an \; E^r \; periecei \;\; to \;\; stoiceio \;\; j \; sto \, pedio \;\; i \\ 0 \, , a \, l \, l \, iwV \end{cases}$$

είναι τα στοιχεία ενός πίνακα G με γραμμές τους κανόνες E^r και στήλες τα n_i στοιχεία του πεδίου i. Κάθε x_r αντιστοιχεί σε ένα κανόνα και παίρνει τιμές 1 αν ο κανόνας είναι στην κάλυψη και 0 διαφορετικά σχηματίζοντας ένα διάνυσμα. Η λύση x λέγεται «βασική κάλυψη» της σχέσης (3.6.1) αν για $\mathrm{K} = \{r \mid x_r = 1\} \subset S$ ισχύει $\bigcup_{k \in \mathrm{K}} E_i^k = R_i$. Η βασική κάλυψη είναι ένα μη περιττό σύνολο κανόνων των οποίων τα σύνολα E_i^k καλύπτουν όλες τις τιμές του πεδίου i που είναι ο γεννήτορας πεδίο, που θα παράγει ένα κατά βάση νέο κανόνα με συμβάλλοντες κανόνες αυτούς της βασικής κάλυψης. Τέλος η σχέση (3.6.1) σημαίνει ότι κάθε στοιχείο του πεδίου i εμφανίζεται σε τουλάχιστον ένα κανόνα.

3.6.2 Εισαγωγή στον αλγόριθμο Chen (1998).

Το πρώτο βήμα είναι να δημιουργήσουμε το πίνακα $G = (g_{rj}^{i})_{s\times n_i}$. Προφανώς κάθε γραμμή του πίνακα είναι ένας κανόνας και κάθε στήλη αναφέρεται σε ένα στοιχείο του γεννήτορα πεδίου. Ο χαρακτηρισμός μοναδιαία στήλη δίνεται για να ορίσει τη στήλη η οποία περιέχει ένα στοιχείο 1 και τα υπόλοιπα 0. Αρχικά παρουσιάζουμε ένα απλό παράδειγμα για να κατανοήσουμε ορισμένα βήματα και να εισάγουμε θεωρήματα που χρησιμοποιεί ο αλγόριθμος. (Chen 1998)

Παράδειγμα 4: Έστω ότι υπάρχουν τρεις κανόνες και τρία πεδία $R_1 = \{1,2,3\}$, $R_2 = \{1,2\}$, $R_3 = \{0,1,2,3\}$

Πίνακας 13

	R_1	R_2	R_3
E_1	1	2	0,1,2
E_2	2	1	1,2,3
E_3	3	1	3

Κατασκευάζουμε τον πίνακα G με γεννήτορα πεδίο το 1

$$G = \begin{bmatrix} 100 \\ 010 \\ 001 \end{bmatrix}$$

Αφαιρούμε όλες τις μοναδιαίες στήλες και τις αντίστοιχες γραμμές όπου βρίσκεται το στοιχείο 1 της μοναδιαίας στήλης. Αν ο αριθμός των μοναδιαίων στηλών που διαγράψαμε είναι ίσος με τον αριθμό των στοιχείων του γεννήτορα πεδίου τότε βρήκαμε την βασική κάλυψη. Στο παράδειγμά μας ισχύει η παραπάνω ισότητα, άρα η λύση του προβλήματος κάλυψης συνόλου όταν ο γεννήτορας πεδίο είναι το 1 βρέθηκε και είναι οι κανόνες E_1, E_2, E_3. Αυτοί οι κανόνες θα χρησιμοποιηθούν για την παραγωγή ενός κατά βάση νέου κανόνα με γεννήτορα πεδίο το 1.

Στη συνέχεια προχωράμε στην δημιουργία του πίνακα G με γεννήτορα πεδίο το 2.

$$G = \begin{bmatrix} 01 \\ 10 \\ 10 \end{bmatrix}$$

Θεώρημα 4 (Chen 1998) : Αν ισχύει $g_{p.} = g_{q.}$ (δηλαδή δύο γραμμές στον πίνακα G είναι ίδιες) και έστω η $g_{r_1}.g_{r_2}.\mathbf{U}\{g_{p.}\}$ είναι μια βασική κάλυψη στο μειωμένο πρόβλημα κάλυψης συνόλου (έχει απομείνει η μια από τις ίδιες γραμμές) τότε και οι δύο $g_{r_1}.g_{r_2}.\mathbf{U}\{g_{p.}\}$, $g_{r_1}.g_{r_2}.\mathbf{U}\{g_{q.}\}$ είναι βασικές καλύψεις στο μειωμένο πρόβλημα κάλυψης συνόλου.

Αυτό που σημειώνει το Θεώρημα 4 είναι ότι αν δυο οι περισσότεροι κανόνες εμφανίζουν ακριβώς τα ίδια στοιχεία του γεννήτορα πεδίου, δεν είναι ανάγκη να τους συμπεριλάβουμε όλους στους υπολογισμούς, αλλά μόνο έναν από αυτούς, αυτό θα γίνει πιο σαφές στην περιγραφή του αλγορίθμου.

Εφαρμόζουμε το Θεώρημα 4 στο παράδειγμα που αναλύουμε :

Στο παραπάνω πίνακα G με γεννήτορα πεδίο το 2 αφαιρώ την τρίτη γραμμή η οποία είναι ίδια με τη δεύτερη και καταλήγω στο νέο μειωμένο πίνακα :

$$G_1 = \begin{bmatrix} 01 \\ 10 \end{bmatrix}$$

Ο παραπάνω πίνακας δημιουργεί το δάσος κάλυψης κανόνων (Edit Cover Forest ECF) το οποίο δημιουργείται από δέντρα το πλήθος των οποίων, είναι ίσο με τον αριθμό των κανόνων του μειωμένου (2 =ίσο με τις γραμμές του πίνακα G_1 στην προκειμένη περίπτωση). Το κάθε δέντρο αρχίζει με ένα κανόνα και διακλαδίζεται συνδυάζοντας τους κανόνες κατά αύξουσα σειρά χωρίς να παίζει ρόλο η μετάθεση των κανόνων. Δηλαδή στο δέντρο που δημιουργείται από τον κανόνα 2 δεν υπάρχει επόμενο κλαδί 12 όπως φαίνεται στο επόμενο σχήμα. Κάθε συνδυασμός κανόνων αποτελεί ένα δεσμό.

Δάσος Κάλυψης Κανόνων

Σχήμα 2

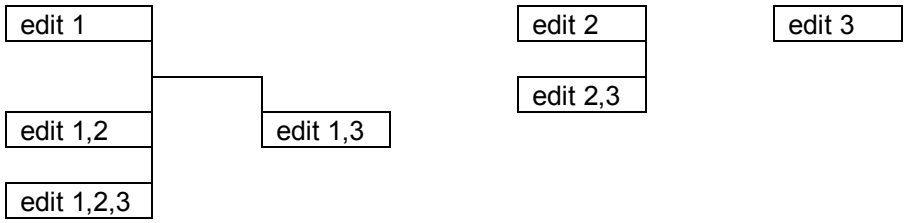
Το μειωμένο πρόβλημα κάλυψης συνόλου το οποίο προέρχεται από το μειωμένο πίνακα G_1 είναι :

$$\text{ελαχιστοποιώ } x_1 + x_2$$

$$\text{υπό τις συνθήκες } \begin{matrix} x_2 \geq 1 \\ x_1 \geq 1 \end{matrix}$$

Άρα η λύση του προβλήματος κάλυψης συνόλου είναι E_1, E_2.

Παρατήρηση: Αν δεν είχα χρησιμοποιήσει το Θεώρημα 4 το δάσος κάλυψης κανόνων θα ήταν



Σχήμα 3

και το πρόβλημα κάλυψης συνόλου θα ήταν

$$\text{ελαχιστοποιώ } x_1 + x_2 + x_3$$

$$\text{υπό τις συνθήκες } \begin{matrix} x_2 + x_3 \geq 1 \\ x_1 \geq 1 \end{matrix}$$

για το οποίο η λύση είναι επίσης E_1, E_2 και η E_1, E_3 Το πλεονέκτημα που δίνει το Θεώρημα 4 είναι το γεγονός ότι μειώνει τις διαστάσεις του δέντρου κάλυψης κανόνων κατά συνέπεια το πρόβλημα κάλυψης συνόλου περισσότερο από 50%. Προφανώς, η χρησιμότητά του αναγνωρίζεται σε πραγματικά δεδομένα όπου ο αριθμός των κανόνων είναι ιδιαίτερα μεγάλος.

Το παράδειγμα συνεχίζεται δημιουργώντας τον πίνακα G με γεννήτορα πεδίο το 3.

$$G = \begin{bmatrix} 1110 \\ 0111 \\ 0001 \end{bmatrix}$$

Λήμμα 5 (Chen 1998): Αν ισχύει $g_{.v} \geq g_{.u}$ για δύο στήλες u, v του πίνακα G, τότε κάθε κάλυψη της στήλης u καλύπτει τη στήλη v, άρα μπορώ να αφαιρέσω τη στήλη v από τον πίνακα G. Δηλαδή αν ένας κανόνας περιέχεται σε κάλυψη μέσω της στήλης u δεν είναι ανάγκη να υπάρχει και μέσω της στήλης v.

Το πρόβλημα κάλυψης συνόλου χωρίς τη χρήση του Λήμματος 5 είναι :

$$\text{ελαχιστοποιώ } x_1 + x_2 + x_3$$

$$\text{υπό τις συνθήκες } \begin{matrix} x_1 \geq 1 \\ x_1 + x_2 \geq 1 \\ x_1 + x_2 \geq 1 \\ x_2 + x_3 \geq 1 \end{matrix}$$

οπότε οι λύσεις στο πρόβλημα είναι E_1, E_2 και E_1, E_3.

Με τη χρήση του Λήμματος 5 ο νέος μειωμένος πίνακας G_1 και το πρόβλημα κάλυψης συνόλου είναι:

$$G_1 = \begin{bmatrix} 10 \\ 01 \\ 01 \end{bmatrix}$$

$$\text{ελαχιστοποιώ } x_1 + x_2 + x_3$$

$$\text{υπό τις συνθήκες } \begin{matrix} x_1 \geq 1 \\ x_2 + x_3 \geq 1 \end{matrix}$$

το οποίο δίνει τις ίδιες λύσεις E_1, E_2 και E_1, E_3 με λιγότερους υπολογισμούς.

Στόχος του αλγορίθμου (Chen 1998) είναι να βρίσκει καλύψεις όσο γίνεται πιο κοντά στη κορυφή του δέντρου δηλαδή στους αρχικούς δεσμούς. Στην περίπτωση αυτή, ο αλγόριθμος δεν «επισκέπτεται» τους απόγονους του δεσμού στον οποίο αντιστοιχεί η κάλυψη γιατί οι απόγονοι δεν θα αποτελούν βασική κάλυψη. Για παράδειγμα αν ο δεσμός 12 βρεθεί ότι είναι κάλυψη του μειωμένου προβλήματος κάλυψης κανόνων, ο αλγόριθμος δεν θα συνεχίσει την αναζήτηση λύσης στους δεσμούς 123 124 και τους υπόλοιπους γιατί αυτές δεν θα είναι βασικές καλύψεις, αλλά θα προχωρήσει στους δεσμούς 13, 14 (για ένα πρόβλημα με τέσσερις κανόνες). Σε κάθε δεσμό ελέγχεται αν η κάλυψη έχει περιττές γραμμές. Αν οι περιττές γραμμές είναι λιγότερες από 2 τότε η κάλυψή μας είναι βασική. Διαφορετικά αφαιρεί τις περιττές γραμμές μια μια, για να φτιάξει υποκαλύψεις. Για παράδειγμα αν ο δεσμός 1234 είναι μια κάλυψη και ο 23 αποτελεί υποκάλυψη, ο αλγόριθμος δεν θα επισκεφτεί τους δεσμούς 23 και 234. Ο Chen (1998) εισάγει στον αλγόριθμο ακόμη ένα βήμα, που διατάσσει τις γραμμές του πίνακα σε φθίνουσα σειρά των μονάδων που περιέχουν γιατί με αυτό τον τρόπο και το οποίο θα περιγραφεί στη συνέχεια.

3.6.3 Αλγόριθμος Chen (1998)

Ο αλγόριθμος (Chen 1998) θα παρουσιαστεί με τη βοήθεια ενός παραδείγματος. Ανάλογη περιγραφή παρουσιάζεται στη διπλωματική εργασία Mniestris N.(2004)

Παράδειγμα 5 : Έστω ο πίνακας G που δημιουργείται από 8 κανόνες με γεννήτορα πεδίο το 1 που περιέχει 5 στοιχεία (στήλες).

$$G = \begin{bmatrix} \mathbf{11011} \\ \mathbf{01100} \\ \mathbf{10011} \\ \mathbf{11011} \\ \mathbf{10001} \\ \mathbf{11101} \\ \mathbf{01010} \\ \mathbf{10001} \end{bmatrix}$$

Βήμα 1 **:** Διαγράφω από τον πίνακα G τις ίδιες γραμμές αν υπάρχουν και παραμένει μόνο η μια από τις ίδιες. Στον πίνακα G οι γραμμές (κανόνες)1 και 4 είναι ίδιες καθώς και οι 5 και 8. Διαγράφοντας τη μία από τις δυο ίδιες παίρνω τον μειωμένο πίνακα H που έχει 6 γραμμές.

$$H = \begin{bmatrix} \mathbf{11011} \\ \mathbf{01100} \\ \mathbf{10011} \\ \mathbf{10011} \\ \mathbf{11101} \\ \mathbf{01010} \end{bmatrix}$$

Βήμα 2 **:** Διαγράφω τις ίδιες ή μεγαλύτερες στήλες του πίνακα H σύμφωνα με το Λήμμα 5. Στον πίνακα H οι στήλες 1 και 5 είναι ίδιες για αυτό αφαιρώ μια από αυτές και παίρνω τον πίνακα B που αποτελείται από 6 γραμμές και 4 στήλες.

$$B = \begin{bmatrix} \mathbf{1101} \\ \mathbf{0110} \\ \mathbf{1001} \\ \mathbf{1000} \\ \mathbf{1110} \\ \mathbf{0101} \end{bmatrix}$$

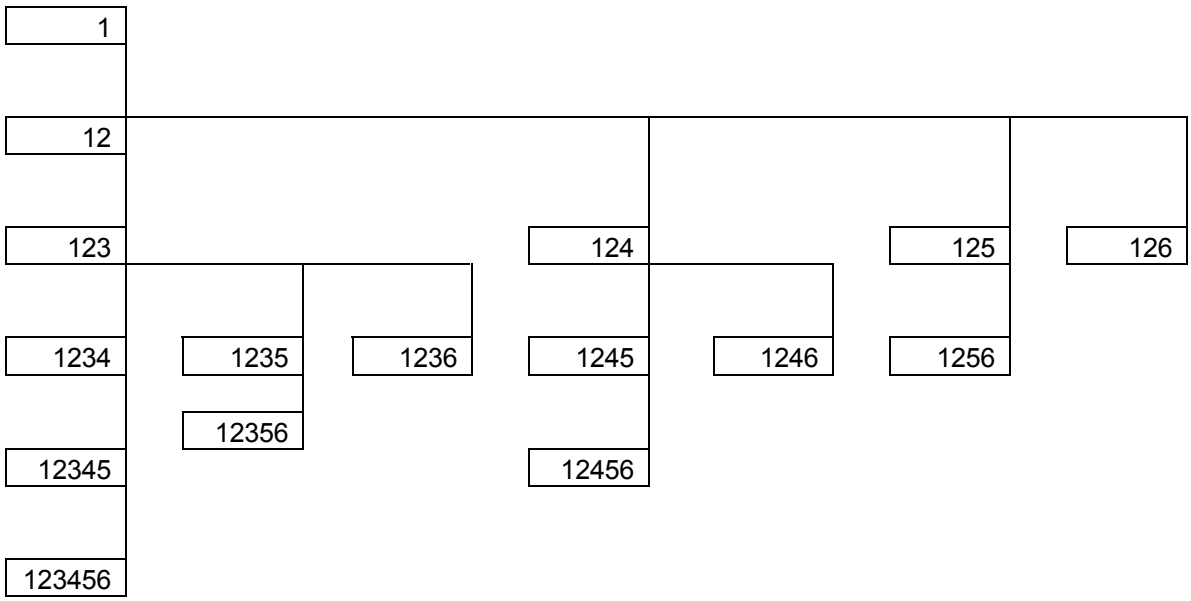
Βήμα 3 **:** Επειδή ο πίνακας B δεν περιέχει μοναδιαίες στήλες παραμένει όπως έχει και συνεχίζω στο επόμενο βήμα. Σε περίπτωση που υπήρχαν μοναδιαίες στήλες θα διαγράφονταν με τις αντίστοιχες γραμμές που υπήρχε η μονάδα.

Βήμα 4 : Διατάσσω τις γραμμές του πίνακα B σε φθίνουσα σειρά των μονάδων που περιέχουν και παίρνω τον πίνακα B_1

$$B_1 = \begin{bmatrix} \mathbf{1101} \\ \mathbf{1110} \\ \mathbf{0110} \\ \mathbf{1001} \\ \mathbf{0101} \\ \mathbf{1000} \end{bmatrix}$$

Βήμα 5: Σχηματίζω το «δάσος» κάλυψης κανόνων (Edit Cover Forest) για τους κανόνες του πίνακα B_1.

Δάσος κάλυψης κανόνων (Edit Cover Forest)



Σχήμα 4

Το παραπάνω δάσος κάλυψης κανόνων δεν είναι ολοκληρωμένο γιατί δεν περιέχει τις διακλαδώσεις 13, 14,..,16 γιατί ήταν ιδιαίτερα μεγάλο.

Το πρόβλημα κάλυψης συνόλου δεδομένου του μειωμένου πίνακα B_1 είναι:

$$\text{ελαχιστοποιώ } x_1 + x_2 + x_3 + x_4 + x_5 + x_6$$

$$\text{υπό τις συνθήκες } \begin{aligned} x_1 + x_2 + x_4 + x_6 &\geq 1 \\ x_1 + x_2 + x_3 + x_5 &\geq 1 \\ x_2 + x_3 &\geq 1 \\ x_1 + x_4 + x_5 &\geq 1 \end{aligned}$$

Βήμα 6 **:** Ελέγχω κάθε δεσμό αν είναι κάλυψη του παραπάνω προβλήματος κάλυψης κανόνων. Για παράδειγμα ο δεσμός 12 δηλαδή οι κανόνες E_1, E_2 είναι μια κάλυψη. Για το λόγο αυτό δεν προχωράω στους δεσμούς που διαδέχονται τον 12 γιατί δεν θα βρω βασικές καλύψεις. Συνεχίζω όμως για την εύρεση κάλυψης στους κόμβους 13, 14, 15, 16 και τους διαδόχους τους αν είναι απαραίτητο.

Βήμα 7 **:** Για κάθε κάλυψη που βρήκα σε ολόκληρο το δέντρο ελέγχω αν κάποια περιέχει περιττή γραμμή. Στο παράδειγμα μας η κάλυψη E_1, E_2 που δεν περιέχει περιττές γραμμές και για το λόγο αυτό αποτελεί μια βασική κάλυψη. Επίσης η κάλυψη E_1, E_3 είναι βασική.

Βήμα 8 **:** Για κάθε βασική κάλυψη βρίσκω σε ποιες γραμμές αντιστοιχεί στον πίνακα *B* (που ήταν ο πίνακας πριν γίνει η διάταξη των κανόνων σε φθίνουσα σειρά). Για παράδειγμα η βασική κάλυψη E_1, E_2 του πίνακα B_1 αντιστοιχεί στις γραμμές 1 και 5 του πίνακα *B*. Δηλαδή η λύση του προβλήματος για τον πίνακα *B* είναι η E_1, E_5

Βήμα 9 **:** Συνδυάζω τις καλύψεις του προηγούμενου βήματος με τις γραμμές που αφαιρέθηκαν στο βήμα 3. Επειδή στο παράδειγμά μας δεν υπήρχαν μοναδιαίες στήλες οι καλύψεις παραμένουν οι ίδιες.

Βήμα 10 **:** Για κάθε βασική κάλυψη (E_1, E_5) βρίσκω τις γραμμές στις οποίες αντιστοιχεί, στον πίνακα *H*. Η αρχική κάλυψη E_1, E_5 αντιστοιχεί στις γραμμές 1 και 5 του πίνακα *H*.

Βήμα 11 **:** Τελικά επιστρέφω στον πίνακα *G* εντοπίζοντας τις γραμμές που αντιστοιχούν οι βασικές καλύψεις του προηγούμενου βήματος. Αν μια γραμμή έχει αντίγραφο τότε δημιουργώ νέες βασικές καλύψεις συνδυάζοντας τις ίδιες με τις υπόλοιπες. Δηλαδή η αρχική κάλυψη του προηγούμενου βήματος E_1, E_5 αντιστοιχεί στις γραμμές 1 και 6 του πίνακα *G* δηλαδή στην βασική κάλυψη E_1, E_6. Όμως η γραμμή 1 είναι ίδια με τη 4 για αυτό έχω δύο νέες βασικές καλύψεις τις E_1, E_6 και E_4, E_6.

Ο παραπάνω αλγόριθμος βρίσκει μόνο τις βασικές καλύψεις και όχι τις περιττές. Το παραπάνω παράδειγμα θεωρείται αρκετά απλό. Εξάλλου έχει παρουσιαστεί ένα μικρό μέρος των υπολογισμών που θα έκανε ο αλγόριθμος μέχρι την εύρεση όλων των βασικών καλύψεων. Για να αναγνωρίσουμε το μέγεθος του προβλήματος αρκεί να αναλογιστούμε ότι τα ίδια βήματα πρέπει να επαναληφθούν θεωρώντας κάθε πεδίο ως γεννήτορα και αναπτύσσοντας ολόκληρο το δέντρο (στο παράδειγμά μας το δέντρο κάλυψης κανόνων ήταν

ένα μικρό μέρος του πλήρους). Εφόσον βρεθούν όλες οι βασικές καλύψεις όλων των γεννητόρων πεδίων αυτές παράγουν κατά βάση νέους κανόνες οι οποίοι μαζί με το σύνολο των ρητά ορισμένων κανόνων αποτελούν το πλήρες σύνολο κανόνων. Το πλήρες σύνολο εισάγεται στο μοντέλο ελάχιστων σταθμισμένων πεδίων για διόρθωση και παίρνουμε τη λύση του προβλήματος εντοπισμού λαθών. Λαμβάνοντας υπόψη τα δεδομένα μιας πραγματικής έρευνας και τον αριθμό των πεδίων και κανόνων που σχετίζονται με αυτή κατανοούμε ότι η υπολογιστική ικανότητα ενός συστήματος πρέπει να είναι τόσο καλή ώστε να μπορεί να αντιμετωπίσει τέτοιου είδους προβλήματα και να εξάγει συμπεράσματα σε λογικό χρονικό πλαίσιο.

3.7 Εύρεση ελάχιστου αριθμού πεδίων για διόρθωση όταν δεν έχει ολοκληρωθεί η παραγωγή όλων των έμμεσων κανόνων. (Winkler and Chen 2002)

Οι αλγόριθμοι των Chen(1998) Winkler(1997) που παρουσιάσαμε, αφορούν την διαδικασία παραγωγής του πλήρους συνόλου κανόνων και στη συνέχεια την εύρεση του ελάχιστου αριθμού πεδίων για διόρθωση χρησιμοποιώντας τις λύσεις του προβλήματος κάλυψης κανόνων. Οι Winkler και Chen (2002) δημιούργησαν ένα αλγόριθμο για την διόρθωση εγγραφών πριν ολοκληρωθεί η παραγωγή του πλήρους συνόλου κανόνων. Οι Winkler and Chen (2002) επέκτειναν τη μέθοδο FH (1976) και βασίστηκαν στον αλγόριθμο 2 (cutting plane algorithm GKL 1986) ο οποίος παράγει νέους έμμεσους κανόνες που χάνονται από την συγκεκριμένη εγγραφή κατά τη διαδικασία εντοπισμού των λαθών και επεκτείνει το σύνολο των πεδίων που πρέπει να διορθωθούν σε αυτή. Ο βελτιωμένος αλγόριθμος των Winkler και Chen (2002) στηρίζεται στο εξής θεώρημα:

Θεώρημα 5 (Winkler and Chen 2002) : Έστω $\overline{E}$ είναι ένα μη-πλήρες σύνολο κανόνων, R ένα σύνολο εγγραφών που χάνουν κάποιους κάνονες του $\overline{E}$, και $r \in R$ μια εγγραφή που χάνει έναν έμμεσο κανόνα που δεν ανήκει στο σύνολο $\overline{E}$. Έστω $\overline{E}_F(r)$ το σύνολο των κανόνων που χάνει η εγγραφή r που ανήκουν στο $\overline{E}$ και $F = \{f_1,...,f_n\}$ είναι το σύνολο των πεδίων που αποτελούν βασική κάλυψη του $\overline{E}_F(r)$. Τότε μπορούν να βρεθούν σύνολα πεδίων $F_r = \{f_{n+1},...,f_q\}$ τέτοια ώστε η ένωση $F \bigcup F_r$ να αποτελεί μια λύση του προβλήματος εντοπισμού λαθών για την εγγραφή r. Επίσης μπορούν να βρεθούν οι έμμεσοι κανόνες που χάνονται από την εγγραφή r και δεν ανήκουν στο σύνολο $\overline{E}$.

3.8 Αυτοματοποιημένο σύστημα SPEER

Στο κεφάλαιο αυτό έχουμε παρουσιάσει μια σειρά αλγορίθμων, οι οποίοι αξιοποιώντας τη βασική μεθοδολογία των FH (1976) και επεκτείνοντάς τη, επιτυγχάνουν τον έλεγχο και τη διόρθωση μόνο κατηγορικών δεδομένων με μειωμένο υπολογιστικό κόστος. Η αντιμετώπιση προβλημάτων που αφορούν αποκλειστικά αριθμητικά δεδομένα επιτυγχάνεται με το αυτοματοποιημένο σύστημα SPEER (Structured Programs for Economic Editing and Referrals) (Kovar and Winkler 1996) του οποίου η μεθοδολογία είναι επίσης βασισμένη στη μέθοδο FH (1976). Το σύστημα SPEER χρησιμοποιείται από την Αμερικάνικη Υπηρεσία Απογραφών για τον έλεγχο και τη διόρθωση αριθμητικών δεδομένων που ελέγχονται από κανόνες πηλίκα και κανόνες ισορροπίας (βεβαιώνοντας ότι τα πεδία αθροίζουν σε σύνολα). Το αρχικό πρόγραμμα είχε δημιουργηθεί από τον Brian Greenberg το 1984 το οποίο περιείχε μόνο κανόνες πηλίκα. Το νέο αυτοματοποιημένο σύστημα SPEER έχει δημιουργηθεί από τον W.Winkler το 1996 σε γλώσσα προγραμματισμού FORTRAN και είναι εύκολα προσαρμόσιμο σε νέα δεδομένα.

Οι κανόνες που αφορούν αριθμητικά δεδομένα και χρησιμοποιεί το σύστημα είναι:

Οι **κανόνες πηλίκων** είναι της μορφής

$$L_{ij} < V_i \big/ V_j < U_{ij}$$

όπου V_i $i= 1,2,...,N$ είναι οι N μεταβλητές (πεδία) της εγγραφής και τα L_{ij}, U_{ij} είναι το κάτω και πάνω όριο αντίστοιχα του κανόνα τα οποία προσδιορίζονται από τους ειδικούς αναλυτές. Αν το πηλίκο δυο μεταβλητών υπερβαίνει το πάνω όριο ή είναι μικρότερο του κάτω ορίου τότε ο κανόνας χάνεται.

Οι **κανόνες ισορροπίας** της μορφής

$$LHS = \sum_{i \in S} V_i = V_j = RHS$$

όπου V_i $i= 1,2,...,N$ είναι οι N μεταβλητές (πεδία) της εγγραφής και S είναι υποσύνολο του N, $j \notin S$. Το αριστερό μέλος της ισότητας αποτελεί το άθροισμα των όρων και το δεξί το σύνολο. Αν το άθροισμα των τιμών των μεταβλητών δεν είναι ίσο με το σύνολο του δεξιού μέλους της ισότητας τότε ο κανόνας χάνεται. Το σύστημα SPEER αναφέρεται μόνο στις περιπτώσεις που κάθε μεταβλητή περιορίζεται από μόνο ένα κανόνα ισορροπίας, το οποίο καλείται **απλό επίπεδο ισορροπίας** (single-level balancing).

Η παραγωγή των έμμεσων κανόνων γίνεται με την αντικατάσταση των κανόνων πηλίκων από συγκεκριμένες μεταβλητές σε αντίστοιχους όρους των κανόνων ισορροπίας. Οι έμμεσοι

κανόνες ονομάζονται **συμπερασματικοί κανόνες** (induced edits). Ο συμπερασματικός κανόνας που παράγεται από την αντικατάσταση ενός όρου του κανόνα ισορροπίας ονομάζεται **απλός συμπερασματικός κανόνας** (simple induced edit). Για παράδειγμα : Αν $V_1 + V_2 = V_3$ είναι ένας κανόνας ισορροπίας και $V_1 / V_j \le U_{1j}$ ένας κανόνας πηλίκο, τότε ένας απλός συμπερασματικός κανόνας που παράγεται είναι ο εξής

$$U_{1j} V_j + V_2 \ge V_3$$

Αρχικά το πρόγραμμα βρίσκει ποιοι από τους ρητά ορισμένους κανόνες χάνονται. Αν υπάρχουν κανόνες που δεν ικανοποιούνται, στη συνέχεια παράγει τους αντίστοιχους συμπερασματικούς κανόνες. Παρατηρούμε ότι οι κανόνες δεν παράγονται εκ των προτέρων αλλά εφόσον έχει αρχίσει η διαδικασία ελέγχου. Από τους κανόνες πηλίκα, ισορροπίας και συμπερασματικούς που δεν ικανοποιούνται, καθορίζονται τα πεδία και οι ισότητες που χρησιμοποιούνται στο πρόβλημα εντοπισμού λαθών για την εύρεση του μικρότερου αριθμού πεδίων για διόρθωση. Όλοι οι συμπερασματικοί κανόνες χρειάζονται για το πρόβλημα εντοπισμού των λαθών, ενώ μόνο οι απλοί συμπερασματικοί κανόνες χρειάζονται για την εύρεση των διαστημάτων, οι τιμές των οποίων διορθώνουν τα πεδία έτσι ώστε να περνούν όλους τους κανόνες. (L.Draper and W.Winkel 1997)

Ο αλγόριθμος που χρησιμοποιεί το σύστημα SPEER περιγράφεται ως εξής:

1. αν κάποιος κανόνας πηλίκο ή ισορροπίας χάνεται παράγω συμπερασματικούς κανόνες και ελέγχω ποιοι από αυτούς χάνονται
2. χρησιμοποιώ του κανόνες πηλίκα ισορροπίας και συμπερασματικούς που χάνονται, σε έναν αλγόριθμο που καθορίζει τα πεδία που χρειάζονται διόρθωση
3. για κάθε πεδίο που πρέπει να διορθωθεί, ελέγχω αν η τιμή του μπορεί να υπολογιστεί από κάποια ισότητα. Αν ναι, τότε διόρθωσα το εν λόγω πεδίο. Αν όχι, τότε χρησιμοποιώ τους κανόνες πηλίκα και τους απλούς συμπερασματικούς κανόνες για να βρω ένα διάστημα οι τιμές του οποίου θα διορθώσουν το εν λόγω πεδίο.
4. Αν οι τιμές του διαστήματος δεν διορθώνουν σωστά το πεδίο, ώστε να περνά τους κανόνες, τότε επιλέγω μια τιμή που είναι λίγο πιο μεγάλη από το κάτω όριο του διαστήματος, όταν η αρχική τιμή του πεδίου είναι μικρότερη από το κάτω όριο του κανόνα πηλίκου. Αντίστοιχα επιλέγω τιμή μικρότερη από το άνω όριο του

διαστήματος, για το πεδίο που η αρχική του τιμή είναι μεγαλύτερη από το πάνω όριο του κανόνα πηλίκου.

Το διαφορετικό στο σύστημα SPEER είναι ότι σε ορισμένες περιπτώσεις ο αλγόριθμος ανακαλείται περισσότερες από μια φορές για να πάρουμε τελικά διορθωμένες εγγραφές που δεν χάνουν κανένα κανόνα. Στην περίπτωση δηλαδή που παράγουμε μόνο τους απλούς συμπερασματικούς κανόνες ο αλγόριθμος πρέπει να εφαρμοστεί περισσότερες από μια φορές για να διορθωθούν όλα τα πεδία. Αν θα θέλαμε η διόρθωση των πεδίων να γίνει με μια μόνο εφαρμογή του αλγορίθμου θα έπρεπε να παράγουμε τέταρτου ή ακόμη και πέμπτου επιπέδου συμπερασματικούς κανόνες.

Για να λάβουμε κάποια εμπειρικά αποτελέσματα της εφαρμογής του αλγορίθμου παραθέτουμε τα αποτελέσματα από την Ετήσια Έρευνα Βιομηχάνων του 1995 που αποτελούνταν από 9765 εγγραφές με 17 πεδία η κάθε μια. Η πρώτη εφαρμογή των δεδομένων στο πρόγραμμα που βρίσκει μόνο τους πρώτου επιπέδου συμπερασματικούς κανόνες, κατέληξε σε 5343 εγγραφές που εξακολουθούσαν να χάνουν κάποιους κανόνες, η δεύτερη εφαρμογή έδωσε 721 εγγραφές, ενώ η τρίτη φορά έδωσε μόνο μια «χαμένη» εγγραφή. Όσο για την ταχύτητα του προγράμματος θα μπορούσαμε να σημειώσουμε ότι είναι ιδιαίτερα γρήγορο αφού μπορεί να επεξεργαστεί 1000 εγγραφές σε πολύ λιγότερο από 4 δευτερόλεπτα (L.Draper and W.Winkler 1997)

ΚΕΦΑΛΑΙΟ 4

Μέθοδοι ελέγχου και διόρθωσης ανεξάρτητοι της μεθόδου Fellegi-Holt (1976)

Στο δεύτερο και τρίτο κεφάλαιο παρουσιάσαμε την μεθοδολογία Fellegi-Holt (1976) και ορισμένους αλγόριθμους, οι οποίοι βασισμένοι στις αρχές και τα θεωρήματα της βασικής μεθοδολογίας των τελευταίων, βελτίωσαν το στάδιο εντοπισμού των πεδίων που χρειάζονται διόρθωση, ώστε να μειωθεί το υπολογιστικό κόστος και να μπορούν να αντεπεξέλθουν σε μεγάλου μεγέθους έρευνες που αφορούν κατά πλειοψηφία κατηγορικά δεδομένα. Ανεξάρτητα από τη μέθοδο των Fellegi-Holt (1976) έχουν δημιουργηθεί νέες μέθοδοι, οι οποίες αν και κλίνουν σε βασικές αρχές των Fellegi-Holt χρησιμοποιούν αλγορίθμους όπως branch and bound (Quere 2000, Quere and De Waal 2000) ή τον αλγόριθμο Chernicova (De Waal 2003, Kovar and Winkler 1996) παρουσιάζοντας νέες προσεγγίσεις ελέγχου και διόρθωσης δεδομένων (κατηγορικών, αριθμητικών και μεικτών), οι οποίες αποτέλεσαν το θεωρητικό πλαίσιο αυτοματοποιημένων συστημάτων που παρουσιάζονται στη συνέχεια. Επίσης έχουν δημιουργηθεί μέθοδοι διόρθωσης όπως είναι η NIM η οποία έχει εφαρμοστεί σε αυτοματοποιημένα συστήματα όπως το CANCEIS, το GEIS και άλλα, που χρησιμοποιούνται από στατιστικές υπηρεσίες. Επίσης στο κεφάλαιο αυτό, θα περιγράψουμε μια νέα προσέγγιση στον έλεγχο ορθότητας δεδομένων τη μέθοδο ελέγχου βασισμένη σε πεδία μεταβλητών, καθώς και ένα μεθοδολογικό εργαλείο που είναι χρήσιμο στο σχεδιασμό της διάταξης των μεταβλητών το αφηρημένο μοντέλο δεδομένων. Τέλος, αναφέρουμε ένα αυτοματοποιημένο πρόγραμμα ελέγχου το GENEDI του χρησιμοποιεί κανόνες ορθότητας της μορφής if-then-else.

4.1 Μέθοδος διόρθωσης MIM

Το τελικό στάδιο της μεθόδου Fellegi-Holt (1976) όπως αναφέραμε στο πρώτο κεφάλαιο αποτελεί η διόρθωση των πεδίων. Αυτοματοποιημένα συστήματα βασισμένα στις βασικές αρχές της μεθόδου διόρθωσης Fellegi-Holt (1976), είναι τα: CANCEIS και GEIS τα οποία χρησιμοποιούνται από τη στατιστική υπηρεσία του Καναδά και τα DISCRETE και SPEER που χρησιμοποιούνται από τη στατιστική υπηρεσία της Αμερικής. Το 1996 στην απογραφή του Καναδά χρησιμοποιήθηκε μια νέα μέθοδος η NIM (New Imputation Methodology), κατά

κάποιο τρόπο διαφορετική από αυτή των FH (1976) που ανέλαβε τον «έλεγχο και διόρθωση» (Edit and Imputation) 11 εκατομμυρίων νοικοκυριών με 5000 κανόνες. Οι δημογραφικές μεταβλητές ήταν ηλικία, φύλο, οικογενειακή κατάσταση, περιουσιακή κατάσταση, και σχέση των μελών του νοικοκυριού, οι οποίες υποβλήθηκαν επιτυχώς σε επεξεργασία σε περίοδο ενός μήνα. Η νέα μέθοδος πέτυχε τη ταυτόχρονη διόρθωση κατηγορικών και ποσοτικών μεταβλητών .

4.1.1 Στόχος της μεθόδου NIM

Χαρακτηριστικό της νέας μεθόδου NIM είναι ότι αποφεύγει τη παραγωγή έμμεσων κανόνων και τη λύση του προβλήματος εντοπισμού λαθών, είναι δηλαδή μια μέθοδος διόρθωσης η οποία χρησιμοποιεί κανόνες, για να διακρίνει τις εγγραφές που χρειάζονται διόρθωση και εφαρμόζεται μετά τη συλλογή και τον έλεγχο των δεδομένων (Claude Poirier 1999). Αντικείμενο της μεθόδου NIM είναι η έρευνα ανάμεσα στις σωστές εγγραφές για δωρητές και στην συνέχεια ο καθορισμός του ελάχιστου αριθμού πεδίων που χρειάζονται διόρθωση. Το γεγονός αυτό βέβαια, αποτελεί και τη διαφορά της μεθόδου με αυτή των Fellegi-Holt(1976) κατά την οποία οι παραπάνω διαδικασίες γίνονται ανάστροφα δίνοντας το προβάδισμα στη μέθοδο NIM να μπορεί να αντιμετωπίσει μεγαλύτερου μεγέθους «ελέγχου και διόρθωσης» προβλήματα. Η εφαρμογή της μεθόδου NIM προϋποθέτει την ύπαρξη αρκετών δωρητών στο σύνολο των δεδομένων.

4.1.2 Ορολογία

Οι παρακάτω ορισμοί είναι βασικοί για την εισαγωγή στην μέθοδο NIM ενώ οι υπόλοιποι δίνονται κατά την ανάπτυξη αυτής:

Δωρητής (*donor*) : είναι μια εγγραφή η οποία ικανοποιεί όλους τους κανόνες και μπορεί να «δανείσει» τις τιμές ορισμένων ή όλων των πεδίων της για να διορθωθούν πεδία εγγραφών που δεν περνούν τους κανόνες. Ο δωρητής και η εγγραφή που πρόκειται να διορθώσει, έχουν κοινές τιμές σε ένα σύνολο πεδίων, δηλαδή οι τιμές τους **ταιριάζουν** σε ένα σύνολο πεδίων (matching fields). Για μια εγγραφή μπορεί να υπάρχουν περισσότεροι από ένας δωρητές.

Πράξη διόρθωσης *(imputation action)* : Έστω μια εγγραφή που δεν ικανοποιεί ορισμένους κανόνες. Υπάρχουν πολλοί συνδυασμοί πεδίων της εγγραφής που αν αλλάξουν θα δώσουν τη βέλτιστη διόρθωση ώστε η εγγραφή να περνάει όλους τους κανόνες. Η διόρθωση του συνόλου των πεδίων μιας εγγραφής ονομάζεται πράξη διόρθωσης.

4.1.3 Αρχές και συνοπτική περιγραφή της μεθόδου NIM

Η μέθοδος NIM εφαρμόζεται προϋποθέτοντας τα παρακάτω:

- Οι διορθωμένες εγγραφές πρέπει να μοιάζουν όσο το δυνατόν περισσότερο στην αρχική τους μορφή, ώστε να επιτύχουμε την μικρότερη διόρθωση (σε αριθμό πεδίων)
- Οι διορθώσεις στις οποίες υποβάλλεται μια εγγραφή πρέπει να προέρχονται μόνο από ένα δωρητή ώστε να είναι **αληθοφανής** (plausible) ο συνδυασμός των τιμών των πεδίων
- Αν υπάρχουν περισσότεροι από ένας υποψήφιοι δωρητές για μια εγγραφή τότε πρέπει να έχουν την ίδια πιθανότητα επιλογής ώστε να μην αυξάνεται λανθασμένα ο αριθμός συγκεκριμένων ομάδων του πληθυσμού.

Οι παραπάνω κανόνες πληρούνται κατά την εφαρμογή της μεθόδου NIM, το οποίο διαπιστώνεται από την συνοπτική περιγραφή της μεθόδου. Αρχικά πρέπει να έχουμε στη διάθεσή μας ένα σύνολο δωρητών που ταιριάζουν με τις χαμένες εγγραφές σε όσο το δυνατόν περισσότερα πεδία. Συγκεκριμένα οι δωρητές και οι υπό διόρθωση εγγραφές, πρέπει να ταιριάζουν σε όσο το δυνατόν περισσότερες κατηγορικές μεταβλητές ενώ να διαφέρουν όσο το δυνατόν λιγότερο στις ποσοτικές μεταβλητές. Εγγραφές με τα παραπάνω χαρακτηριστικά λέγονται **κοντινότεροι γείτονες** (nearest neighbors). Όπως προαναφέρθηκε, οι τιμές των μεταβλητών που θα διορθωθούν είναι αυτές που δεν ταιριάζουν ανάμεσα στο ζευγάρι του κοντινότερου γείτονα και της χαμένης εγγραφής. Στη συνέχεια, για κάθε κοντινότερο γείτονα βρίσκονται τα μικρότερα υποσύνολα μεταβλητών που δεν ταιριάζουν, τα οποία αν διορθώνονταν, η χαμένη εγγραφή θα περνούσε όλους τους κανόνες. Όπως προαναφέραμε, η διόρθωση του συνόλου αυτού των πεδίων ονομάζεται πράξη διόρθωσης. Κάθε πράξη διόρθωσης η οποία οδηγεί την εγγραφή στο να περνά τους κανόνες ονομάζεται **εφικτή** (feasible). Η εφικτή πράξη διόρθωσης που διορθώνει σχεδόν το μικρότερο αριθμό μεταβλητών (με κριτήριο την απόσταση που ορίζεται στη συνέχεια) επιλέγεται τυχαία και ονομάζεται «**πράξη διόρθωσης σχεδόν ελάχιστης αλλαγής**» (near minimum change imputation action NMCIA). Συνεπώς η διορθωμένη εγγραφή θα είναι όσο το δυνατόν περισσότερο όμοια με την αρχική της μορφή, αφού θα μοιάζει στον δωρητή κατά το μέγιστο δυνατό.

Στην ουσία οι πράξεις διόρθωσης σχεδόν ελάχιστης αλλαγής μπορούν να προσδιοριστούν θεωρώντας κάθε κοντινότερο γείτονα ως δωρητή για την χαμένη εγγραφή ως ακολούθως :

Για κάθε ζευγάρι χαμένης εγγραφής/κοντινότερου γείτονα θα διατηρούμε μόνο τους κανόνες που χάνουν οι πιθανές πράξεις διόρθωσης. Με τον τρόπο αυτό χρησιμοποιούμε λιγότερους κανόνες για να εκτιμήσουμε αν η πράξη διόρθωσης είναι εφικτή ή όχι.

Επίσης διακρίνουμε ποιες από τις μεταβλητές είναι πιο πιθανό να χρειάζονται διόρθωση. Αρχικά χρειάζονται διόρθωση οι μεταβλητές που δεν έχουν τιμή (κενές), έπειτα αυτές που έχουν λανθασμένες τιμές δηλαδή τιμές εκτός του πεδίου τιμών τους, στη συνέχεια μεταβλητές που εμπεριέχονται στους κανόνες που δεν περνά η εγγραφή και τέλος οι υπόλοιπες.

Κατά τη μεθοδολογία NIM, ψάχνοντας για πράξεις διόρθωσης, θα εκτιμήσουμε αν είναι εφικτές ή όχι, μόνο αυτές που είναι κοντά στη βέλτιστη διόρθωση και αυτές που είναι «κατά βάση νέες». Λέγοντας «κατά βάση νέες» πράξεις διόρθωσης εννοούμε ότι κανένα υποσύνολο αυτών αν διορθώσει μια εγγραφή δεν θα την κάνει να περάσει τους κανόνες. Αν συνέβαινε αυτό δηλαδή αν κάποιο υποσύνολο μιας πράξης διόρθωσης κατάφερνε να διορθώσει μια εγγραφή έτσι ώστε να περνά τους κανόνες θα ακυρώνονταν η αρχή της μικρότερης αλλαγής στη χαμένη εγγραφή.

4.1.4 Παράδειγμα (M. Bankier 1999).

Το παράδειγμα αναφέρεται σε εγγραφές νοικοκυριών των 6 προσώπων και περιέχει μεταβλητές όπως ηλικία οικογενειακή κατάσταση, σχέση μεταξύ των προσώπων του νοικοκυριού. Οι κανόνες προσδιορίζονται χρησιμοποιώντας τους «**λογικούς πίνακες απόφασης**» (Decision Logic Tables) οι οποίοι είναι πίνακες με στοιχεία της πρώτης στήλης, προτάσεις της μορφής Σχέση με το νοικοκ.(3)=Μητέρα (Το $3^{ο}$ πρόσωπο είναι η μητέρα του $1^{ου}$) και ακολούθως υπάρχουν στήλες με στοιχεία Ναι , Όχι και παύλες παριστάνοντας η κάθε μια ένα κανόνα. Όταν για μια πρόταση υπάρχει το στοιχείο Ν ή Ο σε έναν κανόνα σημαίνει ότι η πρόταση περιέχεται στον κανόνα. Ένας κανόνας χάνεται ανάλογα με τα στοιχεία Ν και Ο που περιέχει η συγκεκριμένη στήλη. Τέλος η παύλα που αντιστοιχεί μια πρόταση σε ένα κανόνα σημαίνει ότι δεν περιέχεται σε αυτόν (Bankier, Luc, et al 1995).

Έστω η παρακάτω εγγραφή που περιέχει τα στοιχεία 3 προσώπων του νοικοκυριού.

Πίνακας 14

Σχέση με Νοικ.	Οικ.κατάσταση	Ηλικία
Πρόσωπο1	Παντρεμένη	38
Σύζυγος	Παντρεμένη	35
Μητέρα	----------	41

Για να διορθωθεί η παραπάνω εγγραφή δανειστήκαμε τιμές πεδίων από ένα δωρητή και η οικογενειακή κατάσταση της μητέρας έγινε Χήρα και η ηλικία της 59.

Πίνακας 15

Σχέση με Νοικ.	Οικ.κατάσταση	Ηλικία
Πρόσωπο1	Παντρεμένη	38
Σύζυγος	Παντρεμένη	35
Μητέρα	Χήρα	59

Υπάρχουν κανόνες που συγκρίνουν στοιχεία δύο προσώπων και λέγονται **κανόνες μεταξύ των προσώπων** (between person edit rule) για παράδειγμα Ηλικία(3)-Ηλικία(1)<15 (Η διαφορά της ηλικίας του 3^{ου} με το 1^{ο} πρόσωπο είναι μικρότερη του 15) και κανόνες που αναφέρονται σε ένα πρόσωπο, για παράδειγμα Σχέση με το νοικοκ.(3)=Μητέρα και λέγονται **ατομικοί κανόνες** (within person edit rule). Ο κανόνας που περνάει μια εγγραφή λέγεται **κανόνας εγκυρότητας** (validity rule) αντίθετα λέγεται **κάνονες αντίφασης** (conflict rule)

Λογικός Πίνακας Απόφασης (***Πίνακας 16***)

Σχέση με το νοικοκ.(3)=Μητέρα	N	N	-	-
Ηλικία(3) – Ηλικία(1) <15	N	-	-	-
Ηλικία (3) < 30	-	N	-	-
Σχέση με το νοικοκ.(3)= Γιαγιά	-	-	N	N
Ηλικία(3) – Ηλικία(1) <30	-	-	N	-
Ηλικία (3)<45	-	-	-	N

Στην παραπάνω εγγραφή (***πίνακας 14***) όντως το τρίτο πρόσωπο είναι η μητέρα του πρώτου και η διαφορά στην ηλικία τους είναι μικρότερη των 15 χρόνων για αυτό η εγγραφή χάνει τον κανόνα που λέει ότι η διαφορά στην ηλικία του τρίτου με το πρώτο πρόσωπο πρέπει να είναι μεγαλύτερη των 15 χρόνων.

Κάνοντας μια έρευνα ανάμεσα στους δωρητές βρίσκουμε αυτόν του πίνακα 17

Πίνακας 17

Σχέση με Νοικ.	Οικ.κατάσταση	Ηλικία
Πρόσωπο1	Παντρεμένη	36
Σύζυγος	Παντρεμένη	37
Πεθερά	Χήρα	59

Παρατηρούμε ότι άλλαξαν 5 στοιχεία από την εγγραφή του πίνακα 13 τα υπογραμμισμένα. Η παραπάνω πράξη διόρθωσης καταλήγει σε εγγραφή που δεν περνάει όλους τους κανόνες (το τρίτο πρόσωπο δεν είναι η μητέρα του πρώτου) και δεν είναι μια εφικτή πράξη διόρθωσης.

Υπάρχουν τρόποι για να μειωθεί το μέγεθος των λογικών πινάκων απόφασης και να είναι ευανάγνωστος.

- Αν καμιά πράξη διόρθωσης δεν έχανε έναν κανόνα αυτός θα μπορούσε να παραβλεφθεί.
- Αν μια πρόταση είναι πάντα σωστή για όλες τις πράξεις διόρθωσης και κάθε κανόνας που αντιστοιχεί σε αυτή την πρόταση έχει στον πίνακα Όχι μπορεί να παραβλεθφεί μαζί με την πρόταση
- Αντίστοιχα αν μια πρόταση είναι πάντα λάθος και ο αντίστοιχος κανόνας της έχει Ναι στο πίνακα μπορεί να παραβλεφθεί μαζί με την πρόταση. Για παράδειγμα το πρόσωπο 3 στον πίνακα 14 έχει ηλικία 41 και στον πίνακα 17 έχει 59 . Η πρόταση Ηλικία (3) < 30 δεν ικανοποιείται ποτέ, όμως ο πίνακας 3 έχει Ναι στο δεύτερο κανόνα άρα μπορεί να παραβλεφθεί μαζί με την πρόταση.

Απλοποιημένος Λογικός Πίνακας Απόφασης (*Πίνακας 18*)

Σχέση με το νοικοκ.(3)=Μητέρα	N	-
Ηλικία(3) – Ηλικία(1) <15	N	-
Σχέση με το νοικοκ.(3)=Πεθερά	-	N
Ηλικία(3) – Ηλικία(2) <15	-	N

Τελικά απέμειναν τέσσερις μεταβλητές Ηλικία(1), Ηλικία(3), Σχέση με το νοικοκ.(3), Ηλικία(2) που περιέχονται στους κανόνες. Γενικά αν *n* είναι ο αριθμός των πεδίων που πρέπει να διορθωθούν, δημιουργούνται 2^n-1 πράξεις διόρθωσης (Bankier, Luc, et al 1995). Οι πράξεις διόρθωσης αποτελούν διανύσματα με στοιχεία όσες οι μεταβλητές που εμπλέκονται στους κανόνες του απλοποιημένου λογικού πίνακα απόφασης. Τα στοιχεία αυτά είναι οι τιμές 1 αν η μεταβλητή θα διορθωθεί και 0 διαφορετικά. Στο παράδειγμά μας οι τέσσερις παραπάνω μεταβλητές εμπλέκονται στους κανόνες του απλοποιημένου πίνακα και δημιουργούν 15 πράξεις διόρθωσης όπως οι (0,0,1,0), (1,1,0,0) και όλοι οι υπόλοιποι συνδυασμοί των τιμών 0 και 1. Από αυτές ελέγχω ποιες περνάνε τους κανόνες του απλοποιημένου λογικού πίνακα απόφασης. Αν βρεθεί ένα διάνυσμα που περνάει τους κανόνες θα διορθώσω τη τιμή κάθε μεταβλητής που αντιστοιχεί σε τιμή 1 στο διάνυσμα, με την αντίστοιχη τιμή του δωρητή. Αν βρεθούν περισσότερες από μια πράξης διόρθωσης να περνάνε τους κανόνες, επιλέγουμε να εφαρμόσουμε αυτή που οδηγεί σε πιο αληθοφανή ή σε πράξη διόρθωσης ελάχιστης αλλαγής. (Bankier 1999, Bankier et al 1997)

Πρέπει να επισημάνουμε επίσης ορισμένες διευκολύνσεις στην εύρεση της εφικτής πράξης διόρθωσης. Αν μια πράξη διόρθωσης και κάθε παράγωγό της δεν είναι σχεδόν ελάχιστης αλλαγής τότε πρέπει να παραβλεφθεί πριν ελεγχθεί από τους κανόνες. Επίσης αν μια πράξη διόρθωσης χάνει ένα κανόνα και όλες οι υπόλοιπες μεταβλητές σε αυτό το κανόνα δεν χρειάζονται διόρθωση τότε αυτή πρέπει να παραβλεφθεί γιατί οποιαδήποτε διόρθωση δεν θα οδηγήσει σε εγγραφή που περνά τους κανόνες. Αν λοιπόν αρχικά διαγράφονται κανόνες και προτάσεις και κατά συνέπεια μεταβλητές και στη συνέχεια αρχίζει η διαδικασία εύρεσης και εκτίμησης σε ένα υποσύνολο πράξεων διόρθωσης εξασφαλίζεται η αποδοτικότητα της μεθόδου σε μεγάλα προβλήματα. (Bankier 1999)

4.1.5 Ορισμός της απόστασης δυο εγγραφών

Η διαδικασία της εύρεσης πράξεων διόρθωσης επαναλαμβάνεται έχοντας στη διάθεσή μας έναν αριθμό κοντινότερων γειτόνων (δωρητών). Έστω ότι μια εγγραφή χάνει κάποιους

κανόνες τότε το σύστημα προσπαθεί να βρει για τη χαμένη εγγραφή έναν δωρητή που προέρχεται από το σύνολο των εγγραφών που περνούν όλους τους κανόνες και μοιάζουν αρκετά σε αυτή που χρειάζεται διόρθωση. Ορίζεται η απόσταση μεταξύ της χαμένης εγγραφής f και μιας σωστής p ως εξής:

$$D_{fp} = D(f, p) = \sum_j w_j D_j(f, p)$$

όπου w_j είναι το βάρος που ορίζεται για κάθε μεταβλητή j (C.Poirier 1999)

Έστω D_{fa} είναι η απόσταση της πράξης διόρθωσης με την χαμένη εγγραφή (απόσταση που μετρά πόσες μεταβλητές χρειάζονται διόρθωση)

Έστω D_{ap} είναι η απόσταση της πράξης διόρθωσης με τον κοντινότερο γείτονα (δωρητή)

Για τις πράξεις διόρθωσης, ορίζεται τελικά η απόσταση

$$D_{fpa} = aD_{fa} + (1-a)D_{ap}$$

όπου α είναι μια παράμετρος με τιμές στο διάστημα (0.5 , 1] δίνοντας ανάλογη βαρύτητα στις δύο αποστάσεις. Η πράξη διόρθωσης για την τιμή της απόστασης

$$D_{fpa} : \min D_{fpa}$$

αποτελεί την πράξη διόρθωσης ελάχιστης αλλαγής. Ενώ αν για μια πράξη διόρθωσης η απόσταση έχει τιμή

$$D_{fpa} \le g \ \min D_{fpa}$$

όπου $g \ge 1$, τότε η πράξη διόρθωσης είναι σχεδόν ελάχιστης αλλαγής. Ο πιο επιεικής χαρακτηρισμός της πράξης διόρθωσης γίνεται για πρακτικούς λόγους στις περιπτώσεις που έχουμε αριθμητικές μεταβλητές, για τις οποίες οι αντίστοιχες πράξεις διόρθωσης δεν είναι τόσο καλές όσο η ελάχιστης αλλαγής (Bankier, Luc, et al 1995).

Κατά την απογραφή του Καναδά του 1996 ορίστηκε α=0.9 δίνοντας μεγαλύτερη βαρύτητα στην εύρεση πράξεων διόρθωσης που μεταβάλλουν την χαμένη εγγραφή όσο το δυνατόν λιγότερο. Οι πράξεις με τη μικρότερη τιμή στην παραπάνω ποσότητα ήταν 5, από τις οποίες επιλέχτηκε τυχαία αυτή που θα δανείσει τις τιμές της στην χαμένη εγγραφή. (M. Bankier 1999).

Τέλος για να επιλέξουμε ανάμεσα στις πράξης διόρθωσης ελάχιστης αλλαγής αυτή που θα χρησιμοποιήσουμε για να διορθώσουμε την εγγραφή ο Bankier (2000) διατύπωσε ένα μέτρο επιλογής που χρησιμοποιεί το πρόγραμμα CANCEIS:

$$M_{fpa} = \left(\min D_{fpa} \Big/ D_{fpa} \right)^t$$

Αν $t = 0$ τότε όλες οι πράξεις διόρθωσης σχεδόν ελάχιστης αλλαγής έχει την ίδια πιθανότητα επιλογής. Αν $t \to \infty$ τότε όλες οι πράξης διόρθωσης ελάχιστης αλλαγής έχουν ίδια πιθανότητα επιλογής και όλες οι υπόλοιπες μηδενική πιθανότητα επιλογής.

Ανακεφαλαιώνοντας, η μέθοδος NIM ταιριάζει κάθε εγγραφή που δεν περνάει τους κανόνες με ένα σύνολο εγγραφών που τους περνά. Βρίσκει τις εγγραφές που έχουν τη μικρότερη απόσταση από τις χαμένες εγγραφές με το κριτήριο της απόστασης. Τα δυο σύνολα εγγραφών διαφέρουν σε ορισμένες μεταβλητές. Διορθώνει τις χαμένες εγγραφές από τους πιθανούς δότες και εξετάζει αν αυτές περνούν τους κανόνες. Επιλέγει τυχαία μια από αυτές για να πραγματοποιήσει τη διόρθωση (W.Winkler and Chen 2002)

4.1.6 Συγκρίσεις της μεθόδου NIM με άλλα αυτοματοποιημένα συστήματα

Στις προηγούμενες απογραφές είχε χρησιμοποιηθεί το σύστημα CANEDIT (Bankier et al 1994) (Στατιστική Υπηρεσία Καναδά) το οποίο είναι βασισμένο στη μέθοδο των Fellegi-Holt. Οι μέθοδοι CANEDIT και NIM συγκρίθηκαν σε 12000 χαμένες εγγραφές. Το 98% αυτών των εγγραφών είχαν τον ίδιο αριθμό μεταβλητών που διορθώθηκαν και στις δυο μεθόδους. Η μικρή διαφορά προέκυψε λόγω των πιο αυστηρών κανόνων της μεθόδου NIM στην καταγραφή της ηλικίας των ατόμων όπου δηλώνονταν η ακριβής χρονολογία ενώ στην μέθοδο CANEDIT καταγράφονταν η δεκαετία. Αυτό κατάληξε στο να διορθώνει μια επιπλέον μεταβλητή η μέθοδος NIM. (Bankier et al 1996)

Σε ορισμένες περιπτώσεις η μέθοδος NIM επιλέγει να πραγματοποιήσει αληθοφανείς διορθώσεις «θυσιάζοντας» τον ελάχιστο αριθμό των διορθώσεων σε αντίθεση με τη μέθοδο Fellegi-Holt που πάντα πραγματοποιείται ο μικρότερος αριθμός αλλαγών.

Χάρη στην αποδοτικότητα του αλγορίθμου το υπολογιστικό κόστος αυξάνεται γραμμικά με την αύξηση των κανόνων σε αντίθεση με την μέθοδο των Fellegi-Holt όπου η αντίστοιχη αύξηση έχει εκθετικούς ρυθμούς. (Bankier 1999)

Τέλος η μέθοδος NIM μπορεί εύκολα να εφαρμοστεί και σε ποσοτικά δεδομένα εκτός από ποιοτικά. Αν και στην πράξη δεν έχουν επεξεργαστεί περισσότερες από μια ποσοτικές μεταβλητές μαζί με ποιοτικές (C.Poirier 1999). Αντίθετα η μέθοδος Fellegi-Holt βρίσκει εφαρμογή κατά πλειοψηφία σε ποιοτικά δεδομένα.

Η NIM είναι η πρώτη μέθοδος που χρησιμοποιεί την έννοια της απόστασης για τη διόρθωση με τη βοήθεια δωρητών. (Bankier 1999)

Στην απογραφή του 1996 είχε χρησιμοποιηθεί και το σύστημα SPIDER για να ελέγξει και να διορθώσει τις μη δημογραφικές μεταβλητές της έρευνας χωρίς όμως επιτυχία. Η αδυναμία του συστήματος να αναλάβει τη διαδικασία αποδείχτηκε από την ανάγκη να χωρίζει τους 2435 κανόνες για νοικοκυριά έξι προσώπων, σε έξι μέρη και να τα επεξεργάζεται ξεχωριστά. Έτσι το 1997 δημιουργήθηκε το αυτοματοποιημένο σύστημα NIM που δεν περιλαμβάνει το SPIDER αλλά αναλαμβάνει τον έλεγχο και τη διόρθωση των εγγραφών με τη βοήθεια των λογικών πινάκων απόφασης, προγραματισμένο σε γλώσσα C. Σε αυτό το σύστημα χρησιμοποιήθηκε και ο όρος των «**βασικών**» (essential) μεταβλητών. Έπειτα από τη διόρθωση των κενών και των λανθασμένων τιμών των μεταβλητών, κάθε κανόνας που χάνεται, αναλύεται με σκοπό την εύρεση μιας μόνο μεταβλητής που αν διορθωθεί οδηγεί σε εγγραφή που περνά τον κανόνα. Κατά συνέπεια η διόρθωση κενών, των λανθασμένων και στη συνέχεια των βασικών μεταβλητών είναι αρκετή για να οδηγήσει σε εγγραφή που περνά τους κανόνες. Αντίθετα με τη μέθοδο του 1996, διατηρούνται όλοι οι κανόνες και όχι μόνο οι κανόνες που μια οι περισσότερες πράξεις διόρθωσης χάνουν. (Bankier 1999)

CANCEIS (Canadian Census Edit and Imputation System)

Από τη στατιστική υπηρεσία του Καναδά έχει αναπτυχθεί ένα ακόμη αυτοματοποιημένο σύστημα το CANCEIS το οποίο αποτελεί μια πιο γενικευμένη εφαρμογή του NIM. Έχει προγραμματιστεί σε γλώσσα C και με μικρές μόνο αλλαγές είναι εύχρηστο σε πολλά συστήματα. Το σύστημα δημιουργήθηκε για να επεξεργαστεί δεδομένα της απογραφής του 2001 στον Καναδά που περιέχουν και άλλες μεταβλητές εκτός από δημογραφικές. Η βασική διαφορά με το σύστημα NIM είναι ότι μεταχειρίζεται όλες τις μεταβλητές ακόμη και τις ποιοτικές ως ποσοτικές και όλες μαζί απεικονίζονται σε λογικούς πίνακες απόφασης, σε μια πιο γενική μορφή προτάσεων και κανόνων από τη NIM. (M.Bankier et al 2000)

4.2 Μέθοδος ελέγχου και διόρθωσης αριθμητικών δεδομένων

Έως τώρα παρουσιάσαμε μια σειρά αλγορίθμων που βρίσκουν εφαρμογή σε αυτοματοποιημένα συστήματα ελέγχου που επεξεργάζονται ως επί το πλείστον κατηγορικά δεδομένα. Η τεχνική των Fellegi-Holt (1976) που δημιουργήθηκε με σκοπό τον έλεγχο και την διόρθωση κατηγορικών δεδομένων, περιλαμβάνει σε θεωρητικό πλαίσιο και αριθμητικά δεδομένα χωρίς ικανοποιητικά αποτελέσματα στην εφαρμογή του. Το έναυσμα της μελέτης των αριθμητικών δεδομένων αποτέλεσε ο αλγόριθμος του N.V. Chernikova το 1964 για την

εύρεση λύσεων από συστήματα γραμμικών ανισοτήτων τον οποίο θα περιγράψουμε σε επόμενη ενότητα. Ο αλγόριθμος Chernikova δεν δημιουργήθηκε με στόχο την ανάπτυξη ενός αυτοματοποιημένου συστήματος αν και στη συνέχεια υιοθετήθηκε για το λόγο αυτό. Ο Quere et al (2000) αναπτύσσει έναν αλγόριθμο ελέγχου και διόρθωσης που αφορά αποκλειστικά αριθμητικά δεδομένα. Ο συγκεκριμένος αλγόριθμος διαφέρει από τη μέθοδο Fellegi-Holt (1976) στο πρόβλημα εντοπισμού λαθών όπως θα δούμε στη συνέχεια.

Αρχικά διακρίνει τους τύπους των εγγραφών μιας έρευνας σε τέσσερις κατηγορίες :

- τις σημαντικές οι οποίες πιθανόν να περιέχουν σημαντικά λάθη και δεν περνούν από αυτόματο έλεγχο αλλά να αναλύονται από τους ειδικούς,
- τις εγγραφές που είναι σωστές, δηλαδή εγγραφές που περνούν όλους τους κανόνες και δεν χρειάζονται διόρθωση,
- τις εγγραφές με λίγα λάθη και περνούν από αυτόματη διόρθωση
- και αυτές που περιέχουν πολλά λάθη για τις οποίες δεν υπάρχει νόημα να γίνουν όλοι οι συνδυασμοί των πεδίων για να διορθωθούν γιατί πιθανόν να μην είναι άξιες να ληφθούν υπόψη. Σε ορισμένα συστήματα ορίζεται ένα άνω όριο πεδίων (cardinality: γ) που πρέπει να διορθωθούν έτσι ώστε αν κάποια μεταβλητή το ξεπερνά να γίνεται η διόρθωση από τους ειδικούς.

4.2.1 Αλγόριθμος Quere

Ο αλγόριθμος Quere (2000) θα αναπτυχθεί παράλληλα με ένα παράδειγμα. Η βασική ιδέα του αλγορίθμου είναι η δημιουργία ενός δέντρου όπου τα κλαδιά θα εκφράζουν την αντικατάσταση ή την απλοποίηση των πεδίων από τους κανόνες ενώ οι δεσμοί είναι τα νέα σύνολα των κανόνων που δημιουργήθηκαν σύμφωνα με τις παραπάνω λειτουργίες. Πιο συγκεκριμένα, έστω ότι επιλέγω ένα πεδίο από αυτά που ανήκουν στους κανόνες που χάνονται. Είναι γνωστό ότι για κάθε κανόνα που χάνεται πρέπει να διορθωθεί τουλάχιστον ένα πεδίο. Για το πεδίο αυτό έχω δυο επιλογές. Είτε να θεωρήσω ότι αυτό έχει σωστή τιμή και να **αντικαταστήσω** (fix) την τιμή του από την τρέχουσα εγγραφή σε όλους τους κανόνες που περιέχεται, είτε να θεωρήσω ότι είναι λάθος και να το **απαλείψω** (eliminate) από όλους τους κανόνες με τη μέθοδο Fourier-Motzkin. Με τον τρόπο αυτό δημιουργείται το **«δίτιμο» δέντρο** (binary tree).

Αντίθετα με τη μέθοδο Fellegi-Holt όπου αρχικά αναπτύσσεται το πλήρες σύνολο κανόνων, στη συνέχεια εντοπίζονται τα πεδία που χρειάζονται διόρθωση και τελικά

διορθώνονται, όπως θα διαπιστώσουμε από το παρακάτω παράδειγμα, δεν είναι ανάγκη να παράγουμε το πλήρες σύνολο κανόνων εκ των προτέρων ούτε απαιτείται να εξάγουμε κανόνες για την διόρθωση των πεδίων. Η αρχή όμως της ελάχιστης αλλαγής που εισήγαγαν οι Fellegi-Holt αποτελεί προϋπόθεση για την εύρεση της βέλτιστης λύσης.

Έστω μια υποτυπώδης έρευνα που αφορά το ετήσιο εισόδημα εργαζομένων. Τα πεδία της έρευνας είναι τα εξής

- x : το καθαρό ετήσιο εισόδημα ενός ατόμου
- y : το μηνιαίο εισόδημα του/της συζύγου του ερωτώμενου
- z : το ολικό ετήσιο εισόδημα του ζευγαριού.

Οι ρητά ορισμένοι κανόνες είναι οι επόμενοι:

$$S_1 = \begin{cases} E_1 & -1250 \times \mathrm{x} \le -15000 \\ E_2 & -12 \times \mathrm{y} \le -15000 \\ E_3 & 875 \times \mathrm{x} - 12 \times \mathrm{y} \le 0 \\ E_4 & -1250 \times \mathrm{x} + 8.4 \times \mathrm{y} \le 0 \\ E_5 & 1250 \times \mathrm{x} + 12 \times \mathrm{y} - 1 \times \mathrm{z} = 0 \end{cases}$$

και έστω μια εγγραφή (30, 0, 37500). Η τρέχουσα εγγραφή αποτυγχάνει τους κανόνες E_2 και E_3. Τα πεδία που περιέχονται στους κανόνες που χάνονται είναι x, y. Έστω ότι επιλέγω να αντικαταστήσω την τιμή του x σε όλους τους κανόνες που περιέχεται θεωρώντας ότι δεν είναι το πεδίο που πρέπει να διορθωθεί. Καταλήγω στο επόμενο σύνολο κανόνων:

$$S_2 = \begin{cases} E_2 & -12 \times \mathrm{y} \le -15000 \\ E_6 & -12 \times \mathrm{y} \le -26250 \\ E_7 & 8.4 \times \mathrm{y} \le 37500 \\ E_8 & 12 \times \mathrm{y} - 1 \times \mathrm{z} = -37500 \end{cases}$$

Με την αντικατάσταση του $x = 30$ ο κανόνας E_1 ικανοποιείται για αυτό δεν το συμπεριέλαβα στο σύνολο S_2. Επίσης παρατηρώ στο σύνολο S_2 ότι ο κανόνας E_2 είναι περιττός λόγω του E_6 για αυτό και τον διαγράφω. Στο νέο σύνολο S_2 βρίσκω πάλι ποιοι κανόνες χάνονται από την τρέχουσα εγγραφή. Χάνεται ο E_6 που περιέχει μόνο το πεδίο y. Επιλέγουμε να αντικαταστήσουμε την τιμή του πεδίου y από την τρέχουσα εγγραφή σε όλους τους κανόνες. Αντικαθιστώντας την τιμή $y = 0$ οι κανόνες E_7, E_8 ικανοποιούνται αλλά ο κανόνας E_6 δεν ισχύει. Άρα η επιλογή μας να αντικαταστήσουμε τις τιμές και των δυο πεδίων x και y δεν είναι η σωστή και σταματά η διαδικασία σε αυτό το «κλαδί» του δέντρου.

Επιστρέφοντας λοιπόν στο σύνολο S_2 επιλέγω να απαλείψω το πεδίο y θεωρώντας ότι είναι το πεδίο που πρέπει να διορθωθεί. (Λύνω την ανισότητα που περιγράφει ο κανόνας E_8 ως προς y και την αντικαθιστώ στις ανισότητες των κανόνων E_6 και E_7). Το νέο σύνολο κανόνων είναι:

$$S_4 = \begin{cases} E_9 & -1 \times z \le -63750 \\ E_{10} & 8.4 \times z \le 765000 \end{cases}$$

Η τρέχουσα εγγραφή χάνει τον πρώτο κανόνα. Επιλέγοντας να αντικαταστήσω την τιμές του z παίρνω το νέο σύνολο κανόνων S_5 από το οποίο δεν ισχύει η μια ανισότητα άρα σταματά η διαδικασία. Ενώ αν επιλέξω να απαλείψω το πεδίο z καταλήγω στο σύνολο S_6 που δεν περιέχει κανόνες που σημαίνει ότι είναι η σωστή επιλογή.

Τελικά βρήκα μια λύση για το πρόβλημα του εντοπισμού των πεδίων που χρειάζονται διόρθωση. Αυτή η λύση είναι τα πεδία y, z και δεν είναι η μοναδική. Αναλύοντας το δέντρο σε όλες τις πιθανές διακλαδώσεις μπορώ να βρω ακόμη μια λύση την x, y.

4.2.2 Βέλτιστη λύση του αλγορίθμου

Στην αναζήτηση λύσης στο πρόβλημα εύρεσης των πεδίων που χρειάζονται διόρθωση θα μπορούσαμε να διακόψουμε τον αλγόριθμο τη στιγμή που εμφανίζει την πρώτη λύση την οποία αποδεχόμαστε. Αν όμως παράγουμε ολόκληρο το δέντρο είναι πιθανό να προκύψουν περισσότερες από μια λύσεις. Κάθε κλαδί ενός πλήρους δέντρου που προέκυψε από την αντικατάσταση και την απαλοιφή κάποιων μεταβλητών αποτελεί μια συγκεκριμένη μορφή της τρέχουσας εγγραφής και ένα συγκεκριμένο σύνολο κανόνων. Αν τελικά η εγγραφή ικανοποιεί τους κανόνες αποτελεί μια λύση στο πρόβλημα εντοπισμού των λαθών. Η λύση με το μικρότερο αριθμό αλλαγών ή η λύση με το μικρότερο σταθμισμένο άθροισμα των μεταβλητών που απαλείφθηκαν αποτελεί τη βέλτιστη λύση.

Παρατηρήσεις :

- Αν στην περίπτωση που επιλέξουμε να αντικαταστήσουμε την τιμή ενός πεδίου και οδηγηθούμε σε μη συνεπές σύνολο κανόνων, ενώ το αρχικό σύνολο ήταν συνεπές αυτό σημαίνει ότι μια προηγούμενη επιλογή μας ήταν λάθος. Δηλαδή αντί να απαλείψουμε ένα πεδίο το αντικαταστήσαμε με την τιμή του. Στο σημείο αυτό

σταματάμε τη διαδικασία σε αυτό το κλαδί του δέντρου και κάνουμε άλλους συνδυασμούς από τις επιλογές που έχουμε.

- Στην περίπτωση που υπάρχουν ελλειπείς τιμές είναι αυτές που πρέπει πρώτα να διορθωθούν, απαλείφοντάς τες είτε αντικαθιστώντας τες με αυθαίρετες τιμές.
- Έχουμε τη δυνατότητα να εισάγουμε σταθερές εμπιστοσύνης για κάθε πεδίο. Μεγάλη τιμή της σταθεράς σε ένα πεδίο σημαίνει ότι αυτό το πεδίο είναι πιθανότερο να έχει σωστή τιμή. Έτσι επιλέγουμε να διορθώσουμε τα πεδία με την μικρότερη τιμή της σταθεράς που φέρουν. Οι σταθερές καθορίζονται από τους ειδικούς της κάθε έρευνας.

4.2.3 Διόρθωση των πεδίων

Αυτό το στάδιο είναι ανεξάρτητο από την εύρεση των πεδίων που πρέπει να αλλάξουν σε κάθε εγγραφή. Υπάρχουν αρκετοί διαθέσιμοι τρόποι διόρθωσης αν και δεν χρησιμοποιούνται όλοι. (Quere 2000)

1. *Διόρθωση με τη μέθοδο παλινδρόμησης*: Βρίσκω τους εκτιμητές όλων των πεδίων και στη συνέχεια βρίσκω μέσω παλινδρόμησης τον εκτιμητή του πεδίου που χρειάζεται διόρθωση. Η μέθοδος αυτή δεν είναι σίγουρο ότι θα διορθώσει την εγγραφή έτσι ώστε να περνά όλους τους κανόνες.
2. *Διόρθωση με τη βοήθεια του μέσου*: Αντικαθιστώ την τιμή του πεδίου που χρειάζεται διόρθωση με τη μέση τιμή που προκύπτει από τις εγγραφές που περνούν όλους τους κανόνες. Επίσης η μέθοδος μπορεί να μη φέρει το επιθυμητό αποτέλεσμα.
3. *Διόρθωση «κοντινότερου γείτονα»:* Έτσι ότι έχουμε μια εγγραφή που χάνει ορισμένους κανόνες και έχουν εντοπιστεί τα πεδία που πρέπει να μεταβληθούν. Από τις έγραφες που περνούν όλους τους κανόνες βρίσκω αυτές που είναι πιο «κοντά» στην συγκεκριμένη εγγραφή. Λέγοντας πιο «κοντά» εννοούμε ότι ταιριάζει σε όσο το δυνατόν περισσότερα πεδία. Χρησιμοποιούμε ορισμένα κριτήρια για να επιλέξουμε την ιδανική εγγραφή και δανειζόμαστε τις τιμές των πεδίων που θα αντικαταστήσουμε στην «χαμένη» εγγραφή. Με τη μέθοδο αυτή η εγγραφή μπορεί να περνάει όλους τους κανόνες.

4.2.4 Εφαρμογή του αλγορίθμου σε αυτοματοποιημένο σύστημα

Για να συγκριθεί ο αλγόριθμος με ένα αυτοματοποιημένο σύστημα χρησιμοποιήθηκε το CherryPi το οποίο είναι προγραμματισμένο σε γλώσσα Pascal και ήταν διαθέσιμο στην

στατιστική υπηρεσία της Ολλανδίας. Η εφαρμογή του αλγορίθμου Quere (2000) ονομάστηκε Leo και δοκιμάστηκε σε αριθμητικά δεδομένα με 10 ρητά ορισμένους κανόνες. Ο αριθμός των κανόνων είναι μικρός γιατί σε διαφορετική περίπτωση θα αποκάλυπτε την αδυναμία του αλγορίθμου να αντιμετωπίσει μεγάλου μεγέθους προβλήματα. Η σύγκριση έγινε σε τρεις έρευνες (δεν έχουμε στοιχεία ποιες έρευνες αφορούσε η σύγκριση), θέτοντας άνω όριο πεδίων που πρέπει να διορθωθούν (cardinality). Το αποτέλεσμα της σύγκρισης ήταν το γεγονός ότι και στις 3 έρευνες ο αλγόριθμος Leo ήταν ταχύτερος σε σχέση με τον CherryPi. (Quere 2000)

4.3 Μέθοδος ελέγχου μεικτών δεδομένων

Επειδή στις περισσότερες έρευνες εμπλέκονται κατηγορικά και αριθμητικά δεδομένα ένας αποδοτικός αλγόριθμος που επεξεργάζεται ταυτόχρονα τέτοιου είδους δεδομένα είναι απαραίτητος. Από ότι γνωρίζουμε η αρχική προσέγγιση μεικτών δεδομένων έγινε από τον G.Sande το 1978. Το 2000 ο G.Sande δημιούργησε ένα σύστημα που επεξεργάζεται μεικτά δεδομένα. Η στατιστική υπηρεσία της Ολλανδίας εισήγαγε μια τεχνική εντοπισμού λαθών που βρίσκει εφαρμογή σε μεικτά δεδομένα, επιδιώκοντας τη διόρθωση όσο το δυνατόν λιγότερων πεδίων που εμπλέκονται σε κανόνες – που περιέχουν κατηγορικές και αριθμητικές μεταβλητές- που δεν καταφέρνει να περάσει η κάθε εγγραφή και η οποία παρουσιάζεται στη συνέχεια. (Quere and De Waal 2000).

4.3.1 Είδη κανόνων

Εφόσον το πρόβλημα επεξεργάζεται αριθμητικές και κατηγορικές μεταβλητές, υπάρχουν τρία είδη κανόνων:

Κατηγορικοί*:* της μορφής

$$(C_1 = 1,2) \times (C_2 = 2) \times (C_3 = 1,3)$$

όπου $C_{1,} C_2, C_3$ είναι τρεις κατηγορικές μεταβλητές της εγγραφής. Οποιαδήποτε εγγραφή ικανοποιεί την παραπάνω σχέση «χάνει» τον κανόνα δηλαδή είναι λάθος αφού η παραπάνω σχέση εκφράζει τους συνδυασμούς των τιμών των πεδίων που είναι απίθανο να συμβούν. Στη συνέχεια οι κατηγορικοί κανόνες θα γράφονται στη μορφή

$$(C_1 = 1,2) \times (C_2 = 2) \times (C_3 = 1,3) => ø$$

(το αριθμητικό μέρος θα είναι το κενό σύνολο)

Αριθμητικοί *:* της μορφής

$$N_1 + 12 \times N_2 \leq 170$$

όπου N_1, N_2 είναι δυο αριθμητικές μεταβλητές για τις οποίες πρέπει να ικανοποιείται η παραπάνω σχέση. Επίσης ο αριθμητικός κανόνας θα γράφεται στο εξής

$$\forall(kathgorikh\ \ metablhth) => N_1 + 12 \times N_2 \leq 170$$

(δηλαδή το κατηγορικό μέρος είναι το σύνολο όλων των πιθανών τιμών των κατηγορικών μεταβλητών)

Μεικτοί *:*της μορφής

$$(C_1 = 1) \times (C_2 = 2,3) \Rightarrow 3 \times N_1 + N_2 \leq 100$$

η οποία μεταφράζεται ως έξης: **AN** το κατηγορικό μέρος επαληθεύεται από την εγγραφή **TOTE** οι αριθμητικές μεταβλητές πρέπει να ικανοποιούν το αριθμητικό μέρος (στην περίπτωση αυτή η εγγραφή περνάει τον κανόνα).

4.3.2 Αλγόριθμος ελέγχου μεικτών δεδομένων

Η ιδέα του αλγορίθμου παραμένει ίδια όπως και στα αριθμητικά δεδομένα με την δημιουργία ενός δίτιμου δέντρου που προκύπτει από τις επιλογές που έχουμε είτε να αντικαταστήσουμε την τιμή μιας μεταβλητής είτε να την απαλείψουμε.

1. Ελέγχω αν η εγγραφή περνά όλους τους κανόνες. Αν αυτό ισχύει έχουμε βρει τη λύση στο πρόβλημα εντοπισμού των πεδίων που χρειάζονται διόρθωση και αυτά είναι τα πεδία που έως αυτή τη στιγμή του δέντρου απαλείψαμε. Έτσι δεν προχωράμε σε επόμενο «κλαδί» του δέντρου.
2. Αν η εγγραφή δεν περνάει ορισμένους κανόνες, επιλέγουμε μια μεταβλητή που περιέχεται σε τουλάχιστον ένα από αυτούς τους κανόνες.
3. Παράγουμε ένα νέο σύνολο κανόνων που προέκυψε από το προηγούμενο σύνολο με απαλοιφή της συγκεκριμένης μεταβλητής. Αν η μεταβλητή που επέλεξα είναι αριθμητική παράγω έμμεσους κανόνες χρησιμοποιώντας μια παραλλαγή της μεθόδου Fourier-Motzkin. Αν είναι κατηγορική παράγω κανόνες με τη τεχνική απαλοιφής των Fellegi-Holt. Η διαδικασία συνεχίζεται με το νέο σύνολο κανόνων.
4. Παράγουμε ένα νέο σύνολο κανόνων που προκύπτει από το προηγούμενο αλλά αυτή τη φορά αντικαθιστώντας την μεταβλητή που επιλέξαμε στο προηγούμενο βήμα με την τιμή της στην τρέχουσα εγγραφή. Συνεχίζεται η διαδικασία με το νέο σύνολο κανόνων.

5. Ο αλγόριθμος σταματά όταν βρεθούν όλες οι λύσεις δηλαδή έχουν πραγματοποιηθεί όλοι οι συνδυασμοί των πεδίων (που απαλείψαμε) ώστε αν διορθωθούν θα οδηγήσουν σε σωστή εγγραφή.

Προφανώς ο παραπάνω αλγόριθμος δεν αναπτύσσει την αντικατάσταση ή την απαλοιφή των κατηγορικών και αριθμητικών μεταβλητών γιατί είναι μακροσκελής διαδικασία αλλά είναι διαθέσιμη στο Quere and De Waal (2000).

Στην πραγματικότητα ο αλγόριθμος επαναλαμβάνεται για κάθε εγγραφή που δεν περνά τους κανόνες. Εντοπίζοντας αρχικά τους κανόνες που χάνονται, για να μειώσουμε το πρόβλημα θα μπορούσαμε να ασχοληθούμε αρχικά με τις αριθμητικές μεταβλητές (αντικαθιστώντας και απαλείφοντας) και στη συνέχεια με τις κατηγορικές. Επιλέγουμε αυτό το τρόπο δημιουργίας του δέντρου γιατί η απαλοιφή κατηγορικών μεταβλητών από μεικτούς κανόνες είναι ιδιαίτερα πολύπλοκη διαδικασία. Επίσης αν στα δεδομένα μας υπάρχουν ελλειπείς τιμές αυτές αναγκαστικά πρέπει να απαλειφθούν εφόσον δεν υπάρχει τιμή για να τις αντικαταστήσουμε στο δίτιμο δέντρο.

4.3.3 Εφαρμογή του αλγορίθμου σε αυτοματοποιημένο σύστημα

Εφαρμογή του αλγορίθμου αποτέλεσε το Leo Simultaneous το οποίο αποτελεί «απόγονο» του Leo (για αριθμητικά δεδομένα). Ο αλγόριθμος επεξεργάζεται με επιτυχία μεικτά αλλά και κατηγορικά δεδομένα. Επίσης μπορεί να αντεπεξέλθει σε ελλειπείς τιμές μεταβλητών απαλείφοντας αρχικά τα πεδία με ελλειπείς τιμές και στην τελική λύση του προβλήματος συμπεριλαμβάνονται και τα συγκεκριμένα πεδία. Επίσης έχουμε τη δυνατότητα χρήσης σταθερών εμπιστοσύνης για κάθε μεταβλητή ακόμη και άνω όριο αριθμού πεδίων που διακόπτει τον αλγόριθμο.

4.4 Λύση του προβλήματος εύρεσης λαθών μέσω παραγωγής κορυφών για μεικτά δεδομένα (De Waal 2003)

Μια διαφορετική προσέγγιση, από αυτή των Fellegi-Holt (1976), περιγράφεται στο άρθρο του De Waal (2003) για τη λύση του προβλήματος του εντοπισμού λαθών, ο οποίος χρησιμοποιεί τον αλγόριθμο Chernicova που δημιουργήθηκε το 1964. Ο αλγόριθμος Chernicova που στόχο έχει την εύρεση λύσεων από ένα ομογενές σύστημα γραμμικών ανισοτήτων, έχει χρησιμοποιηθεί σε πολλά αυτοματοποιημένα συστήματα ελέγχου και

διόρθωσης αριθμητικών δεδομένων όπως το GEIS (Generalized Edit and Imputation System) από τη Στατιστική Υπηρεσία του Καναδά (Kovar and Winkler 1996), το CherryPi από τη Στατιστική Υπηρεσία της Ολλανδίας, το AGGIES (Argicultural Generalized Imputation and Edit System) από την Αμερικάνικη Εθνική Γεωργική Στατιστική Υπηρεσία.

Επειδή ο αλγόριθμος Chernicova ήταν ιδιαίτερα αργός υποβλήθηκε σε αρκετές βελτιώσεις από τον Rubin το 1975 και έπειτα το 1977, καθώς και από τον Sande το 1978. Σύμφωνα με τις παρεμβάσεις τους στον αλγόριθμο Chernicova το σύνολο των αριθμητικών κανόνων αποτελεί ένα κυρτό πολύγωνο που χωρίζει τον δειγματικό χώρο των εγγραφών σε δύο μέρη. Οι εγγραφές που περνούν τους κανόνες αντιστοιχούν σε ένα σημείο μέσα στο πολύγωνο και οι υπόλοιπες σε σημείο έξω από αυτό. Σκοπός της προσέγγισης τους είναι να διορθώσουν τη χαμένη εγγραφή έτσι ώστε να αντιστοιχεί σε ένα σημείο στα όρια του πολυγώνου και όσο πιο κοντά στην αρχική του τιμή. Ο αλγόριθμός Chernicova υπέστη επιπλέον βελτιώσεις από τους Schiopu-Kratina and Kovar το 1989, έπειτα από τους Fillino και Schiopu-Kratina το 1993 αυξάνοντας 60 φορές την ταχύτητα του αλγορίθμου. Από τους παραπάνω μόνο ο G.Sande το 1978 χρησιμοποίησε τον αλγόριθμο Chernicova για μεικτά δεδομένα. Ο De Waal (2003) αναλύει πως η παραγωγή κορυφών μπορεί να χρησιμοποιηθεί στη λύση του προβλήματος εντοπισμού λαθών σε μεικτά δεδομένα.

4.4.1 Συνοπτική περιγραφή του αλγορίθμου Chernikova

Όπως προαναφέραμε, αρχικά ο αλγόριθμος δημιουργήθηκε για αριθμητικά δεδομένα, για τον οποίο το σύνολο των κανόνων μπορεί να εκφραστεί ως εξής:

$$\begin{aligned} AV &\le b \\ V &\ge 0 \end{aligned} \tag{4.4.1}$$

(Kovar and Winkler 1996)

όπου A είναι ο πίνακας των συντελεστών των γραμμικών ανισοτήτων που αντιστοιχούν στους κανόνες, b είναι το διάνυσμα στήλη των σταθερών και V είναι το διάνυσμα στήλη που αντιστοιχεί σε μια εγγραφή με στοιχεία τα πεδία της εγγραφής. Η εγγραφή V που περνά όλους τους κανόνες ικανοποιεί το σύστημα (4.4.1), διαφορετικά δεν το ικανοποιεί. Το πρόβλημα έγκειται στη μείωση της πολλαπλότητας (ο αριθμός των μη-μηδενικών συντελεστών του πίνακα A) για δύο σωστές εγγραφές y, z.

$$A_1\begin{pmatrix} y \\ z \end{pmatrix} \le b_1 \qquad \text{όπου} \quad A_1 = \begin{pmatrix} A & -A \\ -I & I \end{pmatrix} \quad \text{και} \quad b_1 = \begin{pmatrix} b - AV \\ V \end{pmatrix}$$
$$\begin{pmatrix} y \\ z \end{pmatrix} \ge 0$$

(Kovar and Winkler 1996)

Η χρησιμοποίηση του αλγορίθμου Chernicova μπορεί να δώσει τη βέλτιστη λύση του προβλήματος εύρεσης λαθών, με τη εύρεση του μικρότερου αριθμού πεδίων (κατηγορικών και αριθμητικών) που χρειάζονται διόρθωση συμπεριλαμβανομένων και των πεδίων με ελλειπείς τιμές όπως περιγράφεται στο άρθρο του De Waal (2003).

4.5 *Η προσέγγιση INSPECTOR και ο σκοπός της*

Η στατιστική υπηρεσία κάθε χώρας μέλους της Ευρωπαϊκής Ένωσης διεξάγει έρευνες χρησιμοποιώντας σχεδόν ίδια ερωτηματολόγια και στη συνέχεια τα προωθεί στην EUROSTAT για να εξάγει στατιστικά αποτελέσματα. Όμως κάθε στατιστική υπηρεσία χρησιμοποιεί διαφορετικούς κανόνες ορθότητας και διαφορετικές μεθόδους για την εφαρμογή των ελέγχων με άμεση συνέπεια την ανομοιογένεια των δεδομένων που λαμβάνει η EUROSTAT. Η δημιουργία λοιπόν, ενός «κεντρικού» συστήματος ελέγχου δεδομένων, θα επιτύγχανε την διεξαγωγή ερευνών σε διεθνές επίπεδο, την ομοιογένεια της ποιότητας των δεδομένων, την μείωση του κόστους και του χρόνου που απαιτείται για την εξαγωγή κοινών συμπερασμάτων. Για το σκοπό αυτό έχει δημιουργηθεί το INSPECTOR το οποίο προτείνει μια νέα προσέγγιση στην εφαρμογή του **ελέγχου δεδομένων** (data validation).

Η προσέγγιση Inspector είναι μια ερευνητική μελέτη η οποία ολοκληρώθηκε το 2003 με τη συνεργασία δυο πανεπιστημιακών ιδρυμάτων, δυο στατιστικών υπηρεσιών και δυο ιδιωτικών στατιστικών εταιρειών από την Ελλάδα, Αυστρία, Ιταλία και Πορτογαλία. Ο κύριος σκοπός του Inspector είναι να εκφράσει μια νέα προσέγγιση στον έλεγχο ορθότητας στατιστικών δεδομένων κατά τον οποίο πραγματοποιείται ο έλεγχος των στατιστικών δεδομένων για ελλειπείς τιμές, λάθη και λογικές ασυνέχειες. Με άλλα λόγια ο έλεγχος ορθότητας στατιστικών δεδομένων αποτελεί το πρώτο βήμα σ' αυτό που είναι γνωστό στο περιβάλλον της Επίσημης Στατιστικής (Official Statistics) ως έλεγχος ορθότητας (data editing) και δεν καλύπτει τον ακριβή εντοπισμό των λαθών και την διόρθωση αυτών. (Farmakis et al 2004)

Αυτή η προσέγγιση χειρίζεται τους κανόνες ορθότητας ως ιδιότητες των μεταβλητών που εμπλέκονται σ' αυτούς. Πιο συγκεκριμένα χειρίζεται τους κανόνες ορθότητας ως προσδιορισμούς των πεδίων τιμών των μεταβλητών.

Ο έλεγχος δεδομένων ξεκινάει εφόσον έχει σχεδιαστεί η έρευνα. Οι ειδικοί που είναι γνώστες του αντικειμένου της έρευνας και δεδομένου του ερωτηματολογίου προσπαθούν να ανακαλύψουν τις ιδιότητες των μεταβλητών και τις σχέσεις μεταξύ των. Με αυτό τον τρόπο δημιουργούν ένα σύνολο ρητά ορισμένων κανόνων για την έρευνα. Στη συνέχεια το σύνολο των κανόνων «δοκιμάζεται» σε πιλοτικά δεδομένα (που μοιάζουν με αυτά της έρευνας) για να ελεγχθεί κατά πόσο είναι επιεικείς ή αυστηροί ή την ύπαρξη αντικρουόμενων κανόνων. Έπειτα από κατάλληλες τροποποιήσεις, αν αυτό κρίνεται απαραίτητο, το σύνολο των κανόνων θα εφαρμοστεί στα πραγματικά δεδομένα της έρευνας.

4.5.1. Ορισμός των κανόνων ορθότητας

Το καθοριστικό στοιχείο της προσέγγισης Inspector είναι ο **έλεγχος βασισμένος σε πεδία τιμών** (domain-based validation). Η προσέγγιση Inspector θεωρεί τον κανόνα ορθότητας ως μια δήλωση του πεδίου τιμών μιας μεταβλητής ή ενός συνδυασμού μεταβλητών η οποία πρέπει να ικανοποιείται. Για παράδειγμα η μεταβλητή εισόδημα είναι αριθμός με θετικές τιμές. Η οικονομική δραστηριότητα είναι διακριτή μεταβλητή που παίρνει τιμές από μια λίστα κωδικών.

Με άλλα λόγια οι κανόνες ορθότητας ορίζονται ως πεδία τιμών των μεταβλητών ή θέτουν περιορισμούς στα Καρτεσιανά γινόμενα τιμών ξεχωριστών μεταβλητών. Οι κανόνες ορθότητας πρέπει να δηλώνονται και να καταχωρούνται ως νέες ομάδες μεταβλητών με τα πεδία τους. Με αυτό τον τρόπο ελέγχοντας ένα σύνολο δεδομένων εννοούμε απλά τον έλεγχο κάθε μεταβλητής για το αν παίρνει επιτρεπτή τιμή ή όχι.

Το πλεονέκτημα της προσέγγισης Inspector είναι ότι οι κανόνες ορθότητας είναι γενικοί και δεν χρειάζεται να ξαναγράφονται για κάθε νέα εφαρμογή. Απλά οι χρήστες πρέπει να δηλώνουν τις μεταβλητές και τα πεδία τιμών τους για κάθε διαφορετικό σύνολο δεδομένων. Οι κανόνες συνήθως μεταφέρονται στον υπολογιστή με κωδικούς που ακολουθούν την έκφραση If-then-else.

4.5.2. *Ιεράρχηση των στοιχείων που περιλαμβάνουν τα ερωτηματολόγια μιας έρευνας*

Πρωταρχικό σύνολο δεδομένων (primary dataset) είναι τα δεδομένα που συλλέχθηκαν από τις απαντήσεις των ερωτηματολογίων μιας έρευνας. Υποθέτοντας μια εικονική έρευνα έχουμε διαφορετικούς τύπους εγγραφών, ένας τύπος για την κατοικία, ένας άλλος για την οικογένεια και ένας ακόμη για το κάθε άτομο ξεχωριστά. Οι εγγραφές του συνόλου των δεδομένων μπορεί να είναι διαφορετικών τύπων εκφράζοντας η καθεμία διαφορετική στατιστική οντότητα. Οι διαφορετικού τύπου εγγραφές μπορεί να σχετίζονται μεταξύ τους. Όμως υπάρχει μια ιεραρχία στις στατιστικές οντότητες που αντανακλάται στην ιεραρχία των τύπων των εγγραφών. Τα μετα-δεδομένα που θεωρούνται ένας άλλος τύπος εγγραφής είναι τα πρώτα στην ιεραρχία αφού αναφέρονται σε ολόκληρη την έρευνα.

Ομάδες από εγγραφές σχηματίζουν μια **παρατήρηση** (observation). Για παράδειγμα κάθε κατοικία αποτελείται από τα δεδομένα του τύπου εγγραφής της κατοικίας, μια ή περισσότερες εγγραφές της οικογένειας και μιας ή περισσότερων εγγραφών για το κάθε άτομο ξεχωριστά.

Κάθε κανόνας ορθότητας αναφέρεται σε μια στατιστική οντότητα. Ο τύπος εγγραφής που ανταποκρίνεται σ'αυτόν λέγεται **rule's base record type**. Ο κανόνας μπορεί να εμπλέκει μεταβλητές από άλλους τύπους εγγραφών. Για παράδειγμα σε μια έρευνα στο τύπο εγγραφής Οικογένεια υπάρχει η μεταβλητή Ολικό Εισόδημα και στον τύπο εγγραφής Άτομο υπάρχει η μεταβλητή Εισόδημα (κάθε ατόμου). Ο κανόνας που θέτει το συνολικό εισόδημα του σπιτιού να είναι το άθροισμα των εισοδημάτων του κάθε ατόμου αναφέρεται στο τύπο εγγραφής Οικογένεια ο τύπος εγγραφής που ανταποκρίνεται σε αυτό τον κανόνα είναι Οικογένεια αλλά περιλαμβάνει και τον τύπο εγγραφής Άτομο.

4.5.3 *Διαχωρισμός των μεταβλητών*

Κάθε κανόνας αναλύεται στον ορισμό της μεταβλητής και του πεδίου της. Η μεταβλητή μπορεί να είναι μονοδιάστατη και ονομάζεται **ατομική** (atomic variables) ή διάνυσμα μονοδιάστατων μεταβλητών και ονομάζεται **σύνθετη μεταβλητή** (composite variable).

Μερικές ατομικές μεταβλητές υπάρχουν στα δεδομένα χωρίς να οφείλονται στην δημιουργία κανόνων, ονομάζονται **πραγματικές μεταβλητές** (actual variables) και είναι αυτές για τις οποίες συλλέγονται δεδομένα στην έρευνα όπως το εισόδημα. Οι υπόλοιπες ατομικές μεταβλητές ορίζονται χάρη στους κανόνες ελέγχου. Ένα παράδειγμα πραγματικής είναι το άθροισμα των εισοδημάτων των ατόμων του σπιτιού και η διαφορά του από το συνολικό εισόδημα. Οι μεταβλητές που προέρχονται από έναν κανόνα ελέγχου ονομάζονται

«**έμμεσες**» μεταβλητές (implicit variables) όπως για παράδειγμα τα διανύσματα μεταβλητών που πάντα ορίζονται εξαιτίας ενός κανόνα ελέγχου.

Οι έμμεσες μεταβλητές μπορεί να είναι διαφορετικών τύπων ανάλογα με τον τρόπο κατασκευής τους και της προέλευσης τους. Μερικές από αυτές είναι οι **δευτερεύουσες** (secondary), **παραγόμενες** (derived) , **υπολογιστικές** (computed) και οι οποίες χωρίζονται σε επιμέρους ανάλογα με τα ιδιαίτερα χαρακτηριστικά τους.

4.5.4 *Έλεγχος δεδομένων*

Όπως προαναφέρθηκε ο έλεγχος των δεδομένων αφορά την επαλήθευση της τιμής κάθε μεταβλητής στο αντίστοιχο πεδίου τιμών της. Το πεδίο μιας ατομικής μεταβλητής μπορεί να είναι είτε μια κλίμακα αν είναι συνεχής, είτε μια ένωση τιμών αν είναι κατηγορική. Σε κάθε περίπτωση το πεδίο, είναι η ένωση από ένα αριθμό **στοιχείων** (domain elements). Για παράδειγμα αν μια μεταβλητή είναι κατηγορική κάθε τιμή της κατηγοριοποίησης αποτελεί ένα στοιχείο του πεδίου της. Για τις συνεχείς μεταβλητές γνωρίζουμε ότι παίρνουν τιμές σε μια συνεχή κλίμακα με την εξαίρεση κάποιου συνόλου τιμών. Τότε ορίζονται δυο στοιχεία πεδίου, η συνεχής κλίμακα και το σύνολο των εξαιρουμένων τιμών. Για τις σύνθετες μεταβλητές το πεδίο τους, είναι ένα υποσύνολο του Καρτεσιανού Γινομένου των πεδίων των μεταβλητών που τις συνθέτουν. Σε σχέση με το πόσο μεγάλο είναι αυτό το υποσύνολο με το πλήρες Καρτεσιανό Γινόμενο ο χρήστης μπορεί να ορίσει δυο στοιχεία πεδίου το ίδιο το πεδίο της σύνθετης μεταβλητής και το συμπληρωματικό του σε σχέση πάντα με το καρτεσιανό γινόμενο.

4.5.5 Η δημιουργία του τελικού συνόλου δεδομένων

Οι εγγραφές του πρωταρχικού συνόλου δεδομένων προσαυξάνονται με την προσθήκη των απαραίτητων έμμεσων που δημιουργήθηκαν εξαιτίας των ελέγχων. Το τελικό σύνολο δεδομένων ονομάζεται **επαυξημένο σύνολο δεδομένων** (enhanced dataset), στο οποίο εφαρμόζεται ο έλεγχος ορθότητας. Το ενισχυμένο σύνολο δεδομένων περιγράφεται σε μια **φόρμα συνόλου δεδομένων** (dataset template) με όλα τα χαρακτηριστικά των μεταβλητών και το οποίο περιλαμβάνει ακόμη ένα νέο τύπο μεταβλητών τις **προϋποθετικές** (precondition) μεταβλητές οι οποίες αν δεν ικανοποιούν ένα κανόνα δεν επιτρέπουν περαιτέρω έλεγχο στην ίδια παρατήρηση. (Farmakis et al 2004)

Ο προσδιορισμός της φόρμας συνόλου δεδομένων μπορεί να είναι μια χρονοβόρα διαδικασία στις περιπτώσεις που οι έρευνες περιλαμβάνουν μεγάλο αριθμό μεταβλητών και κανόνων. Το γεγονός αυτό μπορεί να οδηγήσει σε λάθη ακόμη και από έμπειρους χρήστες.

Τέλος υπάρχει περίπτωση ένας κανόνας να εμπλέκει δεδομένα από άλλες έρευνες όπως για παράδειγμα έρευνες προηγούμενης περιόδου για να συγκρίνει την τιμή μιας μεταβλητής με αυτή της προηγούμενης έρευνας. Τα σύνολα δεδομένων που παρέχουν τιμές των μεταβλητών τους εξαιτίας κανόνων ορθότητας λέγονται **δευτερεύοντα σύνολα δεδομένων** (secondary datasets).

Το σύστημα ελέγχου πρέπει να έχει πληροφορίες από όλες τις πραγματικές και τις δευτερεύουσες μεταβλητές. Αυτές οι πληροφορίες παράγονται από τις «**αφηρημένες πηγές**» (abstract sources) που είναι σύνολα δεδομένων που παρέχουν τις μεταβλητές αυτές. Για τις πηγές αυτές δεν είναι ανάγκη να γνωρίζουμε την κατασκευή τους (τύποι εγγραφών, μεταβλητές), δηλαδή δεν είναι τόσο περιγραφικό όσο μία φόρμα σύνόλου δεδομένων και ονομάζεται **πηγή φόρμας δεδομένων** (source template).

Ο συνδυασμός της φόρμας συνόλου δεδομένων και της παραπάνω «πηγής» αναφέρεται σε μια συγκεκριμένη έρευνα και για συγκεκριμένο σύνολο κανόνων. Αν τα παραπάνω χαρακτηριστικά μείνουν ίδια τότε η ίδια φόρμα σύνόλου δεδομένων μπορεί να χρησιμοποιηθεί και σε επόμενες περιστάσεις της έρευνας.

Έχουν δημιουργηθεί ακόμη δυο μεθοδολογικά εργαλεία που αφορούν κυρίως τον σχεδιασμό και όχι την εφαρμογή του ελέγχου των δεδομένων, τα οποία μπορούν να χρησιμοποιηθούν στην προσέγγιση Inspector. Αυτές είναι το «αφηρημένο μοντέλο δεδομένων» (abstract data model) που οργανώνει καλύτερα τις μεταβλητές και βοηθά στο σχεδιασμό της προσέγγισης και η μέθοδος της «δίτιμης κατάτμησης» (binary segmentation) που βοηθά στον αυτοματοποιημένο προσδιορισμό των κανόνων. (Petrakos et al 2004), (Farmakis et al 2004)

4.6 Αφηρημένο μοντέλο δεδομένων (*Abstract data model*) (Petrakos et al 2004)

Σκοπός της μεθόδου του **αφηρημένου μοντέλου δεδομένων** δεν είναι η δημιουργία ενός ακόμη αλγορίθμου για τον έλεγχο των δεδομένων αλλά η απόδοση της διαδικασίας πιο αποτελεσματικά. Η νέα μέθοδος καταφέρνει να αποκαλύψει τις σχέσεις μεταξύ των μεταβλητών έπειτα από σαφή κατανόηση των χαρακτηριστικών της συγκεκριμένης έρευνας,

δεδομένου της πραγματικής της πηγής. Ο προσδιορισμός των κανόνων σύμφωνα με τη μέθοδο πραγματοποιείται όχι με απλό συνδυασμό μεταβλητών με τη βοήθεια των ειδικών που είναι γνώστες της έρευνας όπως γινόταν μέχρι τώρα. Οι κανόνες αντανακλούν τους πραγματικούς περιορισμούς που υπεισέρχονται στις σχέσεις μεταξύ των μεταβλητών και στις ιδιότητές τους.

Το αφηρημένο μοντέλο δεδομένων παριστάνεται γραφικά ως ένα «ανάποδο» δέντρο. Η κορυφή, «ρίζα» είναι μια **παρατήρηση** (observation) η οποία χωρίζεται στα επιμέρους κλαδιά κάθε ένα από τα οποία είναι μια **στατιστική μονάδα** (statistical unit). Η διακλάδωση συνεχίζεται για κάθε μια στατιστική μονάδα στις σύνθετες μεταβλητές (composite variables) και καταλήγει στα «φύλλα» του δέντρου κάθε ένα από τα οποία είναι μια ξεχωριστή μεταβλητή της έρευνας η **ατομική μεταβλητή** (atomic variable).

Το μοντέλο θεωρείται αφηρημένο (αφηρημένο) με την έννοια ότι δεν στηρίζεται σε συγκεκριμένες τιμές των μεταβλητών. Το μοντέλο κατασκευάζεται κατά το σχεδιασμό της έρευνας και πριν ακόμη γίνει η συλλογή των δεδομένων. Το πλεονέκτημα έγκειται στο γεγονός ότι μπορεί να χρησιμοποιηθεί από τους ειδικούς γνώστες της έρευνας για τον προσδιορισμό των κανόνων, χωρίς να υπάρχει κίνδυνος παράλειψης κάποιων από αυτούς, και εξασφαλίζοντας την αποφυγή αντικρουόμενων κανόνων.

Το αφηρημένο μοντέλο δεδομένων είναι μια ιεραρχική οργάνωση των **στοιχείων** (συστατικών: components ή blocks) της έρευνας τα οποία περιγράφονται στη συνέχεια με τη χρήση μιας εικονικής δημογραφικής έρευνας που συλλέγει δεδομένα για κατοικίες, την οικογένεια και τα υπόλοιπα πρόσωπα του σπιτιού ξεχωριστά.

1) ***Ατομική μεταβλητή*** είναι η κάθε μια μεταβλητή της έρευνας στην οποία καταλήγει το μοντέλο στην αναπαράσταση του δέντρου. Θα μπορούσε για παράδειγμα να είναι ο αριθμός των ατόμων της οικογένειας. Οι ατομικές μεταβλητές μπορεί να είναι οποιουδήποτε τύπου: κατηγορικές συνεχείς ονοματικές
2) ***Σύνθετη μεταβλητή*** είναι μια ομάδα (δυο ή περισσότερων) μεταβλητών της έρευνας οι οποίες έχουν μια λογική σχέση μεταξύ τους. Κάθε σύνθετη μεταβλητή αναφέρεται σε μια στατιστική μονάδα και μπορεί να είναι οποιουδήποτε τύπου όπως και οι ατομικές. ένα παράδειγμα σύνθετης μεταβλητής στη στατιστική μονάδα Άτομο είναι η μόρφωση και το επάγγελμά του.

3) ***Στατιστική μονάδα*** είναι μια φυσική οντότητα για την οποία η έρευνα συλλέγει δεδομένα. Κάθε στατιστική μονάδα αποτελεί μια εγγραφή για την έρευνα. Ένα ερωτηματολόγιο μπορεί να περιέχει περισσότερες από μια στατιστικές μονάδες (εγγραφές) Για παράδειγμα στατιστικές μονάδες είναι η κατοικία , η οικογένεια το κάθε άτομο της οικογένειας.
4) ***Παρατήρηση*** είναι το σύνολο των δεδομένων που συλλέγονται από ένα ερωτηματολόγιο και μπορεί να περιέχει περισσότερες από μια εγγραφές.
5) ***Αρχείο δεδομένων****(data set)* είναι το σύνολο των παρατηρήσεων μιας έρευνας.

Η ιεραρχία των παραπάνω στοιχείων είναι η ακόλουθη : Ατομική μεταβλητή ∈ Σύνθετες μεταβλητές ∈ Στατιστικές μονάδες ∈ Παρατήρηση ∈ Αρχείο δεδομένων.

Μια ακόμη έννοια που πρέπει να εισάγουμε είναι η «**περίπτωση ελέγχου**» (validation case). Με την έννοια αυτή περιγράφουμε αφηρημένα ένα κανόνα χωρίς τον ακριβή προσδιορισμό του αλλά με την απλή αναφορά των μεταβλητών που περιέχονται σε αυτόν. Μια περίπτωση ελέγχου μπορεί να μεταφραστεί σε περισσότερους από έναν κανόνες. Για παράδειγμα έστω ότι σε μια οικονομικοκοινωνική έρευνα μια περίπτωση ελέγχου απαιτεί τον έλεγχο επαγγέλματος και εισοδήματος. Αυτή η περίπτωση μπορεί να αναλυθεί σε περισσότερους από έναν κανόνες, ο κάθε ένας να συνδέει το επάγγελμα με το αντίστοιχο «συμβατό» εισόδημα.

Η κατασκευή λοιπόν του αφηρημένου μοντέλου δεδομένων αποτελεί το πρώτο βήμα στον προσδιορισμό των κανόνων. Το επόμενο βήμα είναι ο προσδιορισμός των περιπτώσεων ελέγχου και η μετατροπή τους σε κανόνες.

Η περιγραφή των κανόνων μπορεί να γίνει πιο συνοπτική με τη μορφή κατηγοριοποίησης ώστε να μπορούν να παρασταθούν σε μορφή κωδικών και να αναγνωρίζονται από τον υπολογιστή. Η πλειοψηφία των κανόνων περιορίζονται σε μια παρατήρηση αλλά υπάρχουν και κανόνες που

- συγκρίνουν παρατηρήσεις με τα υπόλοιπα δεδομένα όπως συμβαίνει κατά τον έλεγχο έκτροπων παρατηρήσεων,
- κανόνες που συγκρίνουν τις παρατηρήσεις με προηγούμενες περιστάσεις της έρευνας
- και κανόνες που συγκρίνουν παρατηρήσεις με άλλες έρευνες.

Η συνοπτική περιγραφή ενός κανόνα αποτελείται από ένα γράμμα Α(ατομική), Σ(στατιστική μονάδα) , Π(παρατήρηση) , Δ(αρχείο δεδομένων) και έναν αριθμό 1(αν ο κανόνας περιορίζεται μέσα σε ένα συστατικό) , 2(αν συγκρίνει διάφορα συστατικά).

Στη συνέχεια παρουσιάζουμε την εφαρμογή της μεθόδου του αφηρημένου μοντέλου δεδομένων για την παραγωγή των κανόνων σε πραγματικά δεδομένα που προέρχονται από μια έρευνα για το κέρδος των επιχειρήσεων (Survey on Turnover) που διεξήχθη από τη Στατιστική Υπηρεσία της Πορτογαλίας. Η έρευνα συλλέγει μηνιαία δεδομένα, από ένα δείγμα 1000 επιχειρήσεων, και οι μεταβλητές της έρευνας είναι οι ακόλουθες :

- V1: Enterprise tax registry identification number.
- V2: Month of survey.
- V3: Year of survey.
- V4: Activity indication (shows whether the enterprise was active during the reference month).
- V5: Total turnover.
- V5.1: Turnover from sales in Portugal.
- V5.2: Turnover from exports to other EU member states.
- V5.3: Turnover from exports to non-EU countries.
- V6: Turnover from sales of goods purchased for resale in the same condition as received.
- V7: Turnover from sales of products manufactured by the enterprise.
- V8: Turnover from the provision of services.
- V9: Number of employees.
- V10: Total wages.
- V11: Wage payments in arrears.
- V12: Total man-hours worked.
- V13: NACE code (classification of the enterprise's activity using the Statistical Classification of Economic Activities in the European Community).

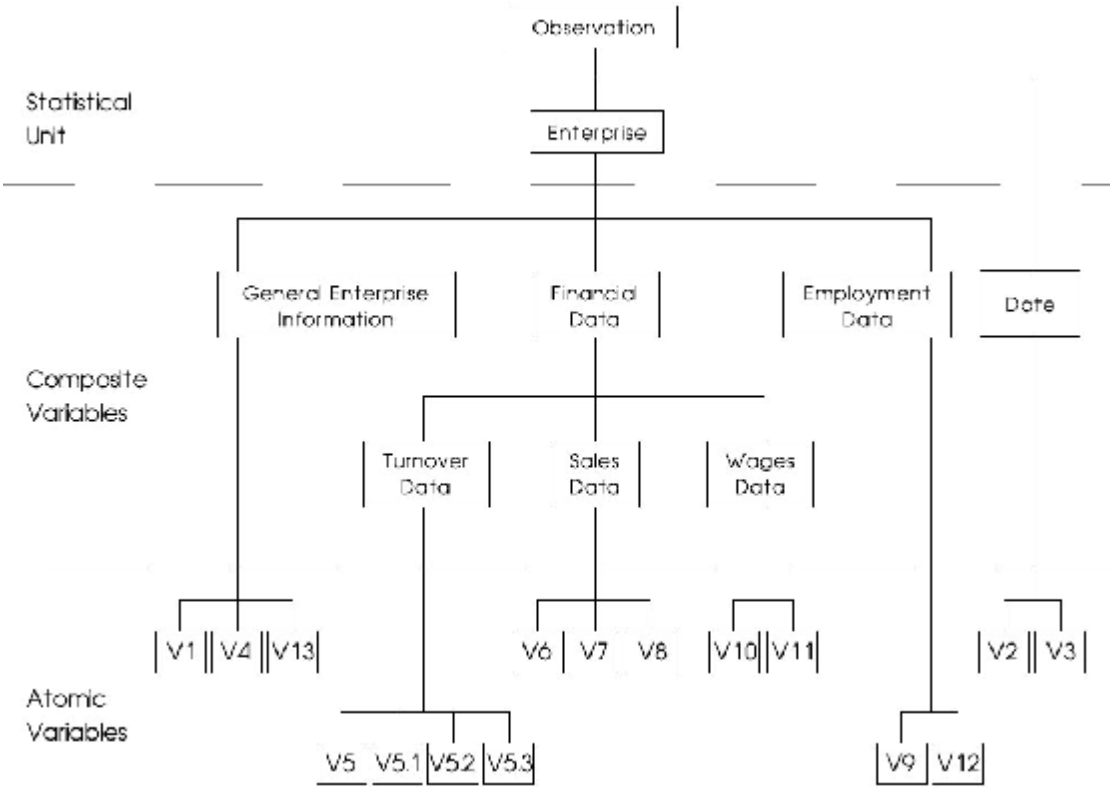
Πηγή : Petrakos et al (2004)

Προσδιορισμός του αφηρημένου μοντέλου δεδομένων :

Αρχείο δεδομένων: αποτελούν οι παρατηρήσεις που λαμβάνονται μηνιαίως από τις επιχειρήσεις.

Παρατήρηση: Τα δεδομένα από μια επιχείρηση

Στατιστική μονάδα: Η επιχείρηση



Πηγή: Petrakos et al (2004)

Από το παραπάνω σχήμα διακρίνω τις εξής **σύνθετες μεταβλητές**:

- Η ημερομηνία (Date): μεταβλητές V2 και V3 που δεν ανήκουν σε καμιά στατιστική μονάδα
- Γενικές πληροφορίες επιχείρησης (General Enterprise Information): αποτελείται από τις ατομικές μεταβλητές V1, V4, V13
- Οικονομικά δεδομένα (Financial Data): που αποτελείται από τις σύνθετες μεταβλητές :
    - Κέρδος (περιέχει τις ατομικές μεταβλητές V5, V5.1, V5.2, V5.3)
    - Πωλήσεις (περιέχει τις ατομικές μεταβλητές V6, V7, V8)
    - Μισθοί (περιέχει τις ατομικές μεταβλητές V10, V11)
- Δεδομένα εργαζομένων : (που περιέχει τις ατομικές μεταβλητές V9, V12)

Ατομικές μεταβλητές : είναι η κάθε μια από τις V1, V2, …, V13 (που αποτελούν τα «φύλλα» του δέντρου)

Εφόσον έχουμε προσδιορίσει το αφηρημένο μοντέλο δεδομένων το επόμενο βήμα είναι η εύρεση των περιπτώσεων ελέγχου και η μετατροπή τους σε κανόνες σύμφωνα με την κατηγοριοποίηση που χαρακτηρίζει κάθε κανόνα με το συνδυασμό γραμμάτων και αριθμών.

Η παραπάνω έρευνα αποτέλεσε εφαρμογή μιας διαδικασίας ελέγχου TREEVAL η οποία προσαρμόζεται στις βασικές αρχές της ολικής διαχείρισης της ποιότητας που είναι γνωστές από τον «κύκλο» του Deming (plan, do, check, act). Η διαδικασία ελέγχου TREEVAL βασίζεται στη μέθοδο δίτιμης κατάτμησης η οποία παράγει τους κανόνες αυτόματα χωρίς την επέμβαση ειδικών αναλυτών. (Petrakos et al 2004)

Τα αποτελέσματα της εφαρμογής της έρευνας στις δύο μεθόδους είναι :

- Το αφηρημένο μοντέλο δεδομένων κατάφερε να εξάγει όλους τους κανόνες ελέγχου από τις περιπτώσεις ελέγχου με επιτυχία, εξασφαλίζοντας την συνέπεια τους. Το σημείο που απαιτεί ιδιαίτερη προσοχή στη μέθοδο αυτή είναι η κατηγοριοποίηση των κανόνων γιατί στηρίζεται σε «λεπτές» διαφορές.
- Η μέθοδος της δίτιμης κατάτμησης μπόρεσε επίσης να παράγει τους κανόνες και μάλιστα τον αριθμό των πιθανόν λαθών στα δεδομένα για κάθε ένα κανόνα. Το αρνητικό στη διαδικασία TREEVAL είναι το γεγονός ότι μπορεί να παράγει κανόνες που περιλαμβάνουν όλες τις μεταβλητές της έρευνας αν εφαρμόζεται στο πλήρες σύνολο δεδομένων και ότι μπορεί να παράγει μεγάλο αριθμό κανόνων ιδίως σε έρευνες με πολλές μεταβλητές.

4.7 Αυτοματοποιημένο σύστημα ελέγχου με χρήση κανόνων If-Then-Else

Ο πλέον εύκολος τρόπος διαμόρφωσης κανόνων για να ελέγξουμε τα δεδομένα μιας έρευνας, είναι η χρήση των κανόνων If-Then-Else που αποτελούν λογική απόρροια των σχέσεων μεταξύ των μεταβλητών. Πολλές Στατιστικές Υπηρεσίες χρησιμοποιούν την παραδοσιακή τεχνική κανόνων για να ελεγχθεί η συνέπεια των μεταβλητών των εγγραφών και να διορθωθούν τυχών λάθη. Αν και οι If-Then-Else κανόνες είναι εύκολο να διατυπωθούν, είναι δύσκολο να παρασταθούν σε μορφή κωδικών αναγνώσιμους από τον υπολογιστή καθώς και οποιαδήποτε αλλαγή στα δεδομένα θα αποτελούσε ιδιαίτερα χρονοβόρα και επίπονη διαδικασία για την αναπροσαρμογή τους. Το δελεαστικό, στο αυτοματοποιημένο σύστημα που εισήγαγαν οι Fellegi-Holt (1976) είναι ότι οι κανόνες

ορθότητας σε αντίθεση με αυτούς της μορφής If-Then-Else παριστάνονται με μορφή πινάκων με κωδικούς 1 και 0 οι οποίοι είναι εύχρηστοι στην μετατροπή και αναπροσαρμογή σε νέα δεδομένα, χωρίς ιδιαίτερη ικανότητα προγραμματισμού και χωρίς την βοήθεια ειδικών αναλυτών, στατιστικών και προγραμματιστών. Ένα παράδειγμα προγράμματος που χρησιμοποιεί κανόνες της μορφής if-then-else είναι το GENEDI το οποίο έχει δημιουργηθεί από τη Eurostat και μπορεί να επεξεργαστεί αριθμητικά και κατηγορικά δεδομένα και βρίσκει εφαρμογή σε διάφορες περιστάσεις ερευνών. Εφαρμογή του προγράμματος αποτέλεσε η έρευνα Structure of Earnings Survey (SES) την οποία θα περιγράψουμε στη συνέχεια για να κατανοήσουμε την διαδικασία του ελέγχου που πραγματοποιεί το πρόγραμμα GENEDI.

4.7.1 Η έρευνα "SES Structure of Earnings Survey" σαν εφαρμογή του προγράμματος GENEDI

Η επόμενη έρευνα "SES Structure of Earnings Survey" διεξήχθη το 2002 υπό την ευθύνη της Eurostat και αφορά το εισόδημα εργαζομένων σε μια επιχείρηση σε συγκεκριμένο μήνα, δηλαδή τον Οκτώβριο του έτους 2002. Το αντικείμενο της μελέτης είναι να προάγει ακριβή δεδομένα που αφορούν εισοδήματα των κρατών μελών της Ευρωπαϊκής Ένωσης και ορισμένων υποψήφιων μελών για τη χάραξη πολιτικής και για ερευνητικούς σκοπούς. Η έρευνα περιλαμβάνει λεπτομερή και συγκρίσιμα στοιχεία για τη σχέση ανάμεσα στο επίπεδο του μισθού σύμφωνα με τα ατομικά χαρακτηριστικά των εργαζομένων και τους εργοδότες.

Τα δεδομένα συγκεντρώθηκαν από ερωτηματολόγια και από στοιχεία προηγούμενων ερευνών. Η διαδικασία της δειγματοληψίας περιλαμβάνει δυο στάδια. Στο πρώτο στάδιο συλλέχθηκαν δεδομένα από ένα απλό τυχαίο στρωματοποιημένο δείγμα επιχειρήσεων. Στο δεύτερο στάδιο πάρθηκε ένα απλό τυχαίο δείγμα από εργαζόμενους από κάθε επιλεγμένη επιχείρηση.

4.7.2 Μεταβλητές της έρευνας

Η Eurostat όρισε της μεταβλητές για τις οποίες απαιτούνται τα δεδομένα οι οποίες χωρίζονται γενικά σε τέσσερις κατηγορίες.

- Πληροφορίες που αφορούν την τοπική μονάδα (επιχείρηση)
- Πληροφορίες που αφορούν τον κάθε εργαζόμενο στο δείγμα
- Πληροφορίες που αφορούν το εισόδημα της πληρωμένες ώρες εργασίας τις ημέρες που δεν εργάστηκε ο εργαζόμενος

- Παράγοντες που ενισχύουν το εισόδημα.

Επίσης ορισμένες μεταβλητές έχουν ορισθεί ως υποχρεωτικές (mandatory) τις οποίες πρέπει να συμπληρώσουν οι ερωτώμενοι ενώ υπάρχουν και οι προαιρετικές (optimal) οι οποίες δεν είναι απαραίτητο να απαντηθούν.

Συγκεκριμένα οι μεταβλητές που απαιτούν πληροφορίες από τις επιχειρήσεις είναι: γεωγραφική περιοχή, μέγεθος της επιχείρησης, οικονομική δραστηριότητα, συμφωνίες πληρωμών, συνολικός αριθμός εργαζομένων, η κύρια αγορά των προϊόντων της, το μέγεθος της ομάδας των επιχειρήσεων στην οποία ανήκει , χώρα στην οποία ανήκει , παράγοντας που ενισχύει τα κέρδη της επιχείρησης.

Ενώ οι μεταβλητές που αφορούν τον κάθε εργαζόμενο είναι οι παρακάτω: φύλο, ηλικία, επάγγελμα, θέση(διαχειριστής ή επόπτης), επίπεδο γνώσης, χρόνος υπηρεσίας στην επιχείρηση, πλήρης ή μερικής απασχόλησης, τύπος συμβολαίου, υπηκοότητα, ποσοστό κάλυψης από την κυβέρνηση, περίοδος διακοπής της εργασίας του στην επιχείρηση, μέσο ωριαίο εισόδημα στον αναφερόμενο μήνα, ολικό εισόδημα τον αναφερόμενο μήνα, εισόδημα από υπερωρίες, επιπλέον εισόδημα (βάρδιες και σαβ/κα) , συνολικό εισόδημα ετησίως (2002), ο αριθμός των εβδομάδων που σχετίζεται με το ετήσιο εισόδημα, ετήσιο bonus (π.χ.δώρα Χριστουγέννων), τακτικό bonus, ετήσιο bonus λόγω παραγωγικότητας, φόροι και μειώσεις στο μισθό, αριθμός πληρωμένων ωρών στον συγκεκριμένο μήνα, αριθμός των υπερωριών στο συγκεκριμένο μήνα, ημέρες απουσίας ετησίως, παράγοντας που αυξάνει το εισόδημα του εργαζομένου.

Βέβαια από τις παραπάνω μεταβλητές διακρίνονται οι αριθμητικές και οι ονοματικές (με τιμές συνδυασμούς γραμμάτων και αριθμών). (Eurostat, European Commission 2003)

4.7.3 Έλεγχοι

Έπειτα από τη συλλογή των δεδομένων η Eurostat αναλαμβάνει να κάνει ορισμένους ελέγχους στα δεδομένα κάθε χώρας, προσαρμογές και άλλους υπολογισμούς. Οι έλεγχοι χωρίζονται σε δυο κατηγορίες. Οι συνολικοί έλεγχοι αφορούν την πληρότητα των υποχρεωτικών και προαιρετικών μεταβλητών (κάθε χώρα αποφασίζει ποιες από τις προαιρετικές μεταβλητές θα συμπληρώσει έτσι ώστε η Eurostat να μπορεί να τις χρησιμοποιήσει). Οι έλεγχοι αληθοφάνειας είναι σχέσεις πιο συγκεκριμένα ανισότητες που πρέπει να ικανοποιούν οι μεταβλητές ώστε να είναι αποδεκτές οι τιμές τους. Για παράδειγμα

η ηλικία ενός εργαζόμενου πρέπει να είναι μεταξύ 14 και 80 ετών. Επίσης αν ο αριθμός των πληρωμένων εβδομάδων είναι μεγαλύτερος του 0 τότε και το ετήσιο εισόδημα είναι μεγαλύτερο του 0.

Επίσης πρέπει να γίνουν ορισμένες προσαρμογές ώστε τα δεδομένα να είναι συγκρίσιμα. Για παράδειγμα αν ένας εργαζόμενος έχει εργαστεί λιγότερες από 52 εβδομάδες πρέπει να μετατραπούν τα υπόλοιπα δεδομένα σε ετήσια βάση.

4.7.4 Αυτοματοποιημένο σύστημα GENEDI

Στη συνέχεια παρουσιάζεται η διαδικασία ελέγχου κανόνων από το πρόγραμμα GENEDI. Το πρόγραμμα μπορεί να εφαρμοστεί σε περιβάλλον Windows NT, Mac OS ή και Unix χωρίς να απαιτεί ιδιαίτερες υπολογιστικές ικανότητες. Αντικείμενό του είναι να αναδιοργανώνει τα εισερχόμενα αρχεία των δεδομένων, να ελέγχει τους κανόνες που πρέπει να ικανοποιούν τα δεδομένα, να τα μετατρέπει σε μια συγκεκριμένη μορφή στατιστικού μηνύματος και στη συνέχεια να τα στέλνει αυτόματα στη Eurostat. (Eurostat GENEDI 2001).

Το πρόγραμμα λαμβάνει δεδομένα σε αρχεία text δύο εναλλακτικών μορφών:

FLR δηλαδή με πεδία ορισμένου μήκους χαρακτήρων το καθένα ή CSV δηλαδή με πεδία όχι ορισμένου μήκους αλλά χωρισμένα με έναν δεδομένο χαρακτήρα (συνήθως , ή ;). Τα αρχεία αυτά μετατρέπονται σε αρχεία GESMES το οποίο είναι ένας συγκεκριμένος τύπος στατιστικού μηνύματος. Στα αρχεία GESMES τα πεδία κάθε εγγραφής εμφανίζονται με ορισμένη σειρά. Αν τα εισερχόμενα αρχεία δεν έχουν τις στήλες (μεταβλητές) στη σωστή σειρά λέγονται "non compliant"(μη συμβατά) ενώ αν τις έχουν στη σωστή σειρά και είναι τύπου CSV λέγονται "compliant" (συμβατά).

Η σειρά διαδικασιών που εκτελούνται με το GENEDI είναι η εξής:

- Αν το εισερχόμενο αρχείο που περιέχει τα δεδομένα είναι FLR ή μη συμβατό CSV τοποθετείται στο φάκελο 0_PreIntray.
- Αν το input αρχείο είναι συμβατό CSV τοποθετείται στο folder 1_Intray.Τα αρχεία πρέπει να μετατραπούν σε συμβατή CSV μορφή για αυτό εφαρμόζουμε την εντολή Create GESMES.
- Στη συνέχεια το συμβατό αρχείο του 1_Intray ελέγχεται για εύρεση λαθών. Στην περίπτωση που δεν περιέχει λάθη μεταφέρεται στο φάκελο 2_Validated. Αν είχε λάθη παραμένει στο 1_Intray και στο 2_Validated παράγεται ένα

αρχείο με το ίδιο όνομα αλλά κατάληξη .log το οποίο περιέχει αναφορά λαθών.

- Το αρχείο από το 2_Validated μετατρέπεται σε GESMES κι η υπόλοιπη διαδικασία έχει να κάνει με συμπίεση και αποστολή στη EUROSTAT.

Με την είσοδο του χρήστη στο πρόγραμμα και με σκοπό την επεξεργασία της παραπάνω έρευνας πρέπει να επιλέξουμε τα εξής:

a. Ορισμός του statistical domain, δηλαδή της θεματικής περιοχής στην οποία αναφέρονται τα δεδομένα της Structure of Earnings Survey (SES).
b. Ορισμός του dataset-Id.. Η έρευνα SES έχει δύο αρχεία δεδομένων το Α που αναφέρεται σε δεδομένα επιχειρήσεων και το Β που αναφέρεται σε δεδομένα εργαζομένων. Τα δύο αυτά αρχεία όπως περιέγραψα προηγουμένως έχουν μετατραπεί σε συμβατή CSV μορφή. Το GENEDI χρειάζεται ένα .txt αρχείο κανόνων για κάθε είδος αρχείου. Ο χρήστης του προγράμματος μπορεί να δημιουργήσει τους κανόνες είτε παραθυρικά από τα εργαλεία που προσφέρει το κεντρικό μενού του προγράμματος είτε γράφοντας σε μορφή εντολών τους κανόνες σε αρχείο text. Οι κανόνες που πρέπει να αφορούν τα δεδομένα της συγκεκριμένης έρευνας έχουν ορισθεί από τη Eurostat.
c. Επιλογή του αρχείου το οποίο θα υποστεί επεξεργασία.

Έπειτα από τα παραπάνω βήματα ξεκινάει η διαδικασία ελέγχου -η οποία είναι ιδιαίτερα σύντομη-για την εύρεση λαθών.

Το πρόγραμμα εμφανίζει λάθος στις εξής περιπτώσεις:

- Όταν μια μεταβλητή είναι υποχρεωτική και δεν περιέχει συγκεκριμένη τιμή.
- Όταν μια μεταβλητή δεν ικανοποιεί έναν συγκεκριμένο κανόνα
- Όταν για μια μεταβλητή έχουν ορισθεί περισσότεροι από ένας κανόνες και δεν ικανοποιεί τουλάχιστον έναν από αυτούς. Αυτό σημαίνει ότι οι κανόνες συνδέονται με AND δηλαδή πρέπει να ισχύουν όλοι οι κανόνες για να είναι σωστή η εγγραφή
- Όταν σε μια εγγραφή έχουν ορισθεί κανόνες για περισσότερες από μια μεταβλητές εμφανίζεται λάθος όταν τουλάχιστον μια μεταβλητή δεν ικανοποιεί έναν κανόνα.

Η διαδικασία ολοκληρώνεται με τη δημιουργία ενός νέου αρχείου όπου περιγράφονται τα λάθη στις εγγραφές που εντοπίστηκαν. Στη συνέχεια μετατρέπεται το αρχείο στη συγκεκριμένη μορφή στατιστικού μηνύματος και αποστέλλεται στη Eurostat. Το πρόγραμμα δεν αναλαμβάνει την διόρθωση λαθών που βρέθηκαν στις εγγραφές.

ΚΕΦΑΛΑΙΟ 5

Γενικά Συμπεράσματα και Συγκρίσεις

Τα προηγούμενα κεφάλαια, περιλαμβάνουν μια αναλυτική περιγραφή των μεθοδολογιών που χρησιμοποιούνται για την εφαρμογή του ελέγχου ορθότητας σε δεδομένα κάθε είδους. Συγκεκριμένα παρουσιάσαμε λεπτομερώς τη πρωτοποριακή και πλήρη μέθοδο των Fellegi-Holt (1976), που αποτέλεσε σταθμό στην ιστορία του ελέγχου ορθότητας και βάση για την ανάπτυξη μεταγενέστερων. Η γνώση και η εμπειρία σε θέματα ελέγχου ορθότητας οδήγησαν στην βελτίωση των βασικών αρχών της μεθόδου Fellegi-Holt (1976), για την απόδοση πιο γρήγορων διαδικασιών που απαιτούν οι σύγχρονες ανάγκες. Παράλληλα, αναπτύχθηκαν και μέθοδοι που αν και υιοθετούν τις βασικές αρχές της θεωρίας Fellegi-Holt (1976) αντιμετωπίζουν τα στάδια της διαδικασίας ελέγχου και διόρθωσης με τελείως διαφορετικό τρόπο δημιουργώντας νέες προσεγγίσεις και πεδία για περαιτέρω έρευνα. Επιπρόσθετα, προσπαθήσαμε να δώσουμε την πραγματική διάσταση του ελέγχου ορθότητας παρουσιάζοντας αυτοματοποιημένα συστήματα που χρησιμοποιούν διεθνείς οργανισμοί και στατιστικές υπηρεσίες σε πραγματικά δεδομένα.

Συγκεκριμένα, στο δεύτερο κεφάλαιο παρουσιάσαμε τη μέθοδο Fellegi-Holt (1976) και αναλύσαμε κάθε στάδιο της διαδικασίας μέχρι τη μετατροπή των δεδομένων σε αξιόπιστες πληροφορίες που χρησιμοποιούνται για την εξαγωγή στατιστικών συμπερασμάτων. Τα βασικά στάδια της μεθόδου είναι η παραγωγή των έμμεσων κανόνων για τη δημιουργία του πλήρους συνόλου κανόνων, η λύση του προβλήματος εντοπισμού των λαθών με την εύρεση του ελάχιστου αριθμού πεδίων για διόρθωση που επιτυγχάνεται με τη λύση ενός προβλήματος κάλυψης συνόλου όταν εφαρμόζεται στο πλήρες σύνολο κανόνων και τελικά η διόρθωση των πεδίων. Οι Fellegi-Holt (1976) κατάφεραν να μετατρέψουν τους if-then-else κανόνες σε κωδικοποιημένη μορφή και να τους αναπαραστήσουν σε πίνακες αναγνώσιμους από τον υπολογιστή και ιδιαίτερα εύχρηστους. Το εμπόδιο στην εφαρμογή της μεθόδου σε πραγματικά δεδομένα ερευνών μεγάλου μεγέθους, είναι το τεράστιο υπολογιστικό κόστος που απαιτεί η διαδικασία γεγονός που δεν έλαβαν υπόψη οι Fellegi-Holt κατά την άκρως μαθηματικοποιημένη απόδοση της θεωρίας τους.

Στη συνέχεια, παρουσιάσαμε νέες βελτιωμένες προσεγγίσεις, που αφορούν κυρίως τη παραγωγή του πλήρους συνόλου κανόνων ικανό να δώσει λύση στο πρόβλημα εντοπισμού

λαθών. Το υπολογιστικό κόστος περιορίζεται με τη μείωση της πιο χρονοβόρας διαδικασίας της παραγωγής έμμεσων κανόνων. Αρχικά οι GKL (1986) θεώρησαν ως πλήρες σύνολο κανόνων, ένα επαρκές σύνολο υποσύνολο του πλήρους και ισοδύναμο με αυτό. Το επαρκές σύνολο αποτελείται από τους μέγιστους κατά βάση νέους κανόνες, ορισμός που δίνει προβάδισμα στη διαδικασία εφόσον κάθε έμμεσος κανόνας μπορεί να αντικατασταθεί από ένα μέγιστο. Επίσης οι GKL (1986) ισχυρίστηκαν ότι οι κανόνες που παράγονται από ένα δεσμό είναι ίδιοι με αυτούς που παράγονται από το δεσμό μεταθέτοντας τα πεδία του. Γεγονός, που στη συνέχεια ο Winkler (1997) απέρριψε ισχυριζόμενος ότι η μετάθεση των πεδίων μέσα σε ένα δεσμό παίζει ρόλο στην παραγωγή των κανόνων. Η χρήση δυο Λημμάτων που προκύπτουν από τη θεωρία των Fellegi-Holt (1976) οδήγησαν στη δημιουργία ενός αλγόριθμου EG που καταφέρνει να εξάγει όλους τους μέγιστα κατά βάση νέους κανόνες πράγμα που δεν συνέβη με τον αλγόριθμο 1 των GKL (1986).

Νέες βελτιώσεις προσδιορίζουν ακριβώς τους κανόνες που πρέπει να χρησιμοποιήσουμε για την παραγωγή μέγιστων κανόνων και καταλήγουν στη δημιουργία του EGE αλγόριθμου από τον Winkler (1998) που επισπεύδει αρκετά τη διαδικασία. Στη συνέχεια ο Chen (1998) ανέπτυξε έναν αλγόριθμο που αποφεύγει περιττές πράξεις που θα επιβράδυναν τη διαδικασία, και βρίσκει μόνο τις βασικές καλύψεις που αποτελούν μη περιττά σύνολα κανόνων που μπορούν να παράγουν κατά βάση νέους κανόνες, ανάγοντας τη διαδικασία της παραγωγής κανόνων σε υπόθεση μερικών δευτερολέπτων (για ένα παράδειγμα με 252 ρητούς κανόνες και 32 πεδία).

Τέλος, στο τέταρτο κεφάλαιο, παρουσιάσαμε νέες προσεγγίσεις σε συγκεκριμένα στάδια της διαδικασίας του ελέγχου ορθότητας. Η μέθοδος NIM που είναι κυρίως μέθοδος διόρθωσης χρησιμοποιεί δωρητές για τη διόρθωση των πεδίων από εγγραφές που χάνουν του κανόνες και εφαρμόζεται σε συστήματα όπως το CANEDIT, GEIS και CANCEIS. Έπειτα, παρουσιάσαμε αλγορίθμους για τη λύση του προβλήματος εντοπισμού λαθών αριθμητικών και μεικτών δεδομένων από τον Quere (2000) και Quere and De Waal (2000) αντίστοιχα. Οι δύο αλγόριθμοι δημιουργούν ένα δίτιμο δέντρο αντικαθιστώντας ή απαλείφοντας την τιμή του πεδίου που επιλέξαμε σχεδόν τυχαία. Με τον τρόπο αυτό, αν οδηγηθούμε σε λάθος επιλογή εγκαταλείπουμε το συγκεκριμένο κλαδί του δέντρου και συνεχίζουμε στην εύρεση σωστής λύσης. Μια ακόμη προσπάθεια αντιμετώπισης μεικτών δεδομένων έγινε με τη βοήθεια της χρήσης του αλγορίθμου Chernicova από τον De Waal (2003).

Τέλος, αναπτύσσεται μια νέα προσέγγιση (η προσέγγιση Inspector) στην έννοια του ελέγχου ορθότητας που αναλαμβάνει μόνο την εύρεση ασυνεπειών στα δεδομένα και δεν ασχολείται με τον εντοπισμό των λαθών και τη διόρθωσή τους. Η νέα προσέγγιση χειρίζεται τους κανόνες ορθότητας ως ιδιότητες των μεταβλητών που εμπλέκονται σ' αυτούς. Πιο συγκεκριμένα χειρίζεται τους κανόνες ορθότητας ως προσδιορισμούς των πεδίων τιμών των μεταβλητών (Farmakis et al 2004). Η προσέγγιση Inspector μπορεί να συνδυαστεί με τη μέθοδο του αφηρημένου μοντέλου δεδομένων (Petrakos et al 2004) που εφαρμόζεται στη διαδικασία του σχεδιασμού της έρευνας κατά την οποία οι κανόνες αντανακλούν τους πραγματικούς περιορισμούς που υπεισέρχονται στις σχέσεις μεταξύ των μεταβλητών και στις ιδιότητές τους. Τέλος παρουσιάζουμε ένα αυτοματοποιημένο σύστημα ελέγχου δεδομένων το GENEDI το οποίο χρησιμοποιεί κανόνες της μορφής if-then-else και εφαρμόζεται σε πολλές έρευνες μια από τις οποίες είναι η έρευνα SES.

Συνοψίζοντας τις μεθόδους και τα γενικά χαρακτηριστικά παραθέτουμε τους επόμενους πίνακες για να λάβουμε μια γενική και παράλληλα συγκριτική εικόνα των όσων παρουσιάσαμε αναλυτικά στα προηγούμενα κεφάλαια:

Στον επόμενο πίνακα παρουσιάζουμε τα χαρακτηριστικά της παραγωγής του πλήρους συνόλου κανόνων όπως αυτά μεταβλήθηκαν από τις βελτιώσεις των GKL, Winkler και Chen.

Χαρακτηριστικά	Fellegi-Holt (1976)	GKL (1986)	GKL cutting plane(1986)	Winkler EG (1997)	Winkler EGE (1998)	Chen (1998)	Winkler-Chen (2001)
Τύπος δεδομένων							
κατηγορικά	+	+	+	+	+	+	+
Ορισμός του πλήρους συνόλου κανόνων: ρητά ορισμένοι κανόνες και							
Κατά βάση νεοι	+					+	
Μέγιστα κατά βάση νέοι		+		+	+		
Το πρόβλημα εντοπισμού των λαθών αρχίζει εφόσον παράχθηκε το σύνολο κανόνων							
Πλήρες σύνολο	+	+		+	+	+	
Ημιτελές			+				+
Ο αλγόριθμος παραγωγής του πλήρους συνόλου κανόνων συνδυάζει τους κανόνες							
έμμεσοι και ρητοί	+						
Υποσύνολα έμμεσων (βασικές καλύψεις) και ρητοί κανόνες		+					
Έμμεσους του προηγούμενου δεσμού και ρητούς				+			
Μέγιστους του προηγούμενου δεσμού και ρητούς					+		
Βασικές καλύψεις του μειωμένου πίνακα B_1						+	
Ο αλγόριθμος παραγωγής κανόνων εξάγει:							
Όλους τους κατά βάση νέους	+					+	
Όλους τους μέγιστους	+	+*		+	+		
Και περιττούς κανόνες	+	+		+	+		

* Ο αλγόριθμος GKL 1986 παράγει μόνο μερικούς από τους μέγιστους κανόνες.

Ο παραπάνω πίνακας συγκρίνει τις μεθόδους (στήλες του πίνακα) που ακολουθούν οι αλγόριθμοι για το πρόβλημα εντοπισμού των λαθών. Προσπαθήσαμε να αποδώσουμε τα χαρακτηριστικά (γραμμές του πίνακα) όπως αυτά προκύπτουν από την διαδικασία των αλγορίθμων. Με τον τρόπο αυτό, διακρίνεται η εφαρμογή όλων των μεθοδολογιών σε κατηγορικά δεδομένα, ο ορισμός του πλήρους συνόλου κανόνων που χρησιμοποιεί η κάθε μέθοδος και πολλά επιμέρους χαρακτηριστικά ώστε να μπορεί ο αναγνώστης να κατανοήσει συνοπτικά τις διαφορές των μεθοδολογιών στην αντιμετώπιση του προβλήματος εντοπισμού των λαθών.

Όσο για την εφαρμογή των μεθόδων που αναφέρονται στον πίνακα, όταν αυτές χρησιμοποιούνται από αυτοματοποιημένα συστήματα μπορούμε να πούμε ότι ο αλγόριθμος Chen (1998) εκτελεί όλες τις δυνατές διαδικασίες για να αποφύγει την παραγωγή περιττών κανόνων (Ενότητα 3.6.3). Αντίθετα, η πιο χρονοβόρα διαδικασία είναι η μέθοδος Fellegi-Holt (1976) αφού για την παραγωγή του πλήρους συνόλου των κανόνων, απαιτεί όλους τους συνδυασμούς των ρητά ορισμένων κανόνων με τους έμμεσους. Βέβαια, αυτό δεν υποβαθμίζει τη μεθοδολογία τους, αρκεί να αναλογιστούμε ότι αποτέλεσε την πιο ολοκληρωμένη και σαφώς ορισμένη μέθοδο στην οποία στηρίχθηκαν οι μεταγενέστερες προσεγγίσεις του στατιστικού ελέγχου ορθότητας.

Ο επόμενος πίνακας συγκρίνει τις μεθόδους ελέγχου και διόρθωσης κατηγορικών και αριθμητικών μεταβλητών ανάλογα με τα χαρακτηριστικά τους.

Χαρακτηριστικά	DISCRETE	SPEER	GEIS	CANEDIT	CANCEIS	NIM
Τύπος δεδομένων						
Αριθμητικά		+	+		+	+
Κατηγορικά	+			+	+	+
Μέθοδος						
Ελέγχου			+		+	
Διόρθωσης			+		+	+
Ελέγχου και διόρθωσης FH	+	+		+		
Παραγωγή έμμεσων κανόνων						
Παράγεται το πλήρες σύνολο κανόνων	+					
Παράγει τους συμπερασματικούς κανόνες που χάνονται		+				
Δεν παράγει έμμεσους κανόνες			+	+	+	+
Εύρεση του <u>ελάχιστου</u> αριθμού πεδίων για διόρθωση						
Επιτυχής	+		+	+	+	+*
Ανεπιτυχής		+				
Η διόρθωση επιτυγχάνεται:						
Αυτόματα από τους έμμεσους κανόνες	+					
Μέσω συντελεστών παλινδρόμησης		+				
Με τη μέθοδο (NIM)			+	+	+	+
Από ιστορικά δεδομένα, μέσους, πηλίκα			+			
Το στάδιο διόρθωσης καταλήγει σε :						
«καθαρά» δεδομένα	+	+**	+ Με τη μέθοδο κοντινότερου γείτονα	+	+	+
«Μη-καθαρά» δεδομένα			+ Με τη χρήση μέσων κ.α.			

* Υπάρχει περίπτωση να επιλέξει για διόρθωση τα πεδία που καταλήγουν σε πιο αληθοφανή πράξη διόρθωσης «θυσιάζοντας» τον ελάχιστο αριθμό πεδίων.
** Το αυτοματοποιημένο σύστημα SPEER όπως περιγράψαμε στην ενότητα 3.8 εξάγει καθαρά δεδομένα μόνο έπειτα από επαναληπτική είσοδο των δεδομένων στο πρόγραμμα όταν αυτό παράγει απλούς συμπερασματικούς κανόνες.

Στον παραπάνω πίνακα εμφανίζονται ορισμένα αυτοματοποιημένα συστήματα ελέγχου και διόρθωσης και η μέθοδος διόρθωσης NIM που χρησιμοποιείται από την Στατιστική Υπηρεσία του Καναδά η οποία διαφέρει από τη μέθοδο διόρθωσης που εισήγαγαν οι Fellegi-Holt (1976). Παρά το γεγονός αυτό, η μέθοδος NIM χρησιμοποιείται από αυτοματοποιημένα συστήματα ελέγχου και διόρθωσης FH όπως το CANEDIT και το GEIS από τη Στατιστική Υπηρεσία του Καναδά και το DISCRETE και SPEER από την Αμερικανική Υπηρεσία Απογραφών. Ο παραπάνω πίνακας περιγράφει τις λειτουργίες των αυτοματοποιημένων συστημάτων δηλαδή αν αυτά χρησιμοποιούνται για έλεγχο ή/και διόρθωση αριθμητικών ή/και κατηγορικών δεδομένων. Επίσης παραθέτονται ορισμένα χαρακτηριστικά, ώστε να διαπιστώσουμε τις λειτουργίες που απαιτούνται για την διόρθωση, τους τρόπους διόρθωσης που ακολουθούν τα αυτοματοποιημένα συστήματα και το αποτέλεσμα αυτών αν δηλαδή παράγουν τελικά «καθαρά» δεδομένα ή όχι.

Καταλήγοντας, πρέπει να αναφέρουμε ένα θέμα που αν και δεν αναπτύχθηκε στην παρούσα διπλωματική εργασία έχει μεγάλο ερευνητικό ενδιαφέρον και είναι ο έλεγχος έκτροπων παρατηρήσεων ως αντικείμενο του μάκρο και επιλεκτικού ελέγχου. Οι έκτροπες παρατηρήσεις που εμφανίζονται στα δεδομένα μας μπορεί να είναι, είτε ιδιαίτερα σημαντικές πληροφορίες που πρέπει να ληφθούν υπόψη γιατί επηρεάζουν τους εκτιμητές, είτε αποτελούν πιθανόν λάθη –σε μονοδιάστατο επίπεδο- και ασυνέπειες –σε πολυδιάστατο επίπεδο- που μπορούν να αποκαλύψουν οι κανόνες ορθότητας μέσα στα δεδομένα. Ο έλεγχος έκτροπων παρατηρήσεων κυμαίνεται από την πιο απλή διαδικασία της δημιουργίας ιστογραμμάτων (σε επίπεδο μιας μεταβλητής), την αναπαράσταση σε διαγράμματα διασποράς (scatterplots) ως τις πιο πολύπλοκες και σύγχρονες διαδικασίες ελέγχου έκτροπων παρατηρήσεων με τα δίκτυα Kohonen. (Morlini I. 1998)

Όπως αναφέραμε και σε προηγούμενο κεφάλαιο, είναι επιτακτική η ανάγκη δημιουργίας ενός ενιαίου συστήματος ελέγχου ορθότητας από διεθνείς οργανισμούς. Το ενιαίο αυτό σύστημα θα μπορεί να το χρησιμοποιεί κάθε εθνική στατιστική υπηρεσία για να ελέγχει τα

δεδομένα της και στη συνέχεια να τα προωθεί στους διεθνείς οργανισμούς για περαιτέρω στατιστική ανάλυση και εξαγωγή συμπερασμάτων σε διεθνές επίπεδο. Με τον τρόπο αυτό εξασφαλίζεται η ομοιογένεια των αποτελεσμάτων και η συγκρισιμότητά τους, η μείωση του κόστους και χρόνου που απαιτείται για την δημοσίευση τους και η επίτευξη ενός υψηλού επιπέδου ποιότητας στα δεδομένα. Δεν πρέπει βέβαια να αγνοούμε ότι, από την αυστηρά μαθηματικοποιημένη μορφή του ελέγχου ορθότητας που εισήγαγαν οι Fellgi-Holt το 1976, εποχή που η υπολογιστική ισχύς ήταν μικρότερη του 1/200 από αυτή που ισχύει σήμερα, η ανάπτυξη νέων μεθοδολογιών και αυτοματοποιημένων συστημάτων και η εξέλιξή τους τα πρόσφατα χρόνια ήταν ραγδαία. Η δημιουργία λοιπόν, ενός «κεντρικού» συστήματος ελέγχου ορθότητας αποτελεί το επόμενο βήμα για την εδραίωση μια νέας άποψης για την ποιότητα, τη συγκρισιμότητα και την αξιοπιστία των στατιστικών δεδομένων και αποτελεσμάτων.

ΠΑΡΑΡΤΗΜΑ

ΛΕΞΙΛΟΓΙΟ

Αιτιοκρατικοί ή ντετερμινιστικοί κανόνες (deterministic edits): είναι κανόνες ορθότητας η παραβίαση των οποίων σημαίνει ότι η εγγραφή περιέχει σίγουρα λάθος.

Ακολουθιακή διόρθωση (sequential imputation): αποτελεί ένα τρόπο διόρθωσης που εισήγαγαν οι FH (1976) κατά την οποία διορθώνουμε σταδιακά τα πεδία

Αληθοφανής (plausible): είναι μια πράξη διόρθωσης της οποίας τα πεδία προέρχονται μόνο από ένα δωρητή. Επίσης μπορεί να μην διορθώνει τον ελάχιστο αριθμό πεδίων αλλά οδηγεί σε πιο αληθοφανή συνδυασμό τιμών.

Αμφίβολοι κανόνες (query): είναι οι κανόνες που ανιχνεύουν «ύποπτες» μεταβλητές (που πιθανόν να είναι λάθος)

Απλό επίπεδο ισορροπίας (single-level balancing): όταν κάθε μεταβλητή περιορίζεται από μόνο ένα κανόνα ισορροπίας

Απλός συμπερασματικός κανόνας (simple induced edit): Ο συμπερασματικός κανόνας που παράγεται από την αντικατάσταση ενός όρου του κανόνα ισορροπίας

Από κοινού διόρθωση (joint imputation): αποτελεί ένα τρόπο διόρθωσης που εισήγαγαν οι FH (1976) κατά την οποία διορθώνουμε ταυτόχρονα όλα τα πεδία

Αριθμητικός κανόνας ορθότητας (arithmetic edit): είναι ο κανόνας που αφορά αριθμητικές μεταβλητές και εκφράζεται γραμμική σχέση αυτών.

Ατομική μεταβλητή (atomic variable): μονοδιάστατη μεταβλητή

Ατομικός κανόνας (within person edit rule): ο κανόνας που αναφέρεται σε στοιχεία ένα προσώπου

Αφηρημένες πηγές (abstract sources): είναι μια πιο γενικά ονομασία της φόρμας δεδομένων που παριστά σύνολα δεδομένων με μερικές ή όλες τις μεταβλητές της φόρμας.

Αφηρημένο μοντέλο δεδομένων (abstract data model): είναι μια δενδρική παρουσίαση των ομάδων των συνδυασμένων μεταβλητών μιας έρευνας χωρίς ακριβή παρουσίαση της σχέσης τους

Βασική κάλυψη (prime cover): είναι η βέλτιστη (μη-περιττή) λύση του προβλήματος κάλυψης συνόλου

Βασική μεταβλητή (essential variables): Αν σε μια εγγραφή έχουν διορθωθεί οι κενές και λανθασμένες τιμές τότε ελέγχω αν υπάρχει μια μεταβλητή που αν διορθωθεί η εγγραφή περνάει τον κανόνα. Η μεταβλητή αυτή ονομάζεται βασική.

Γεννήτορας πεδίο (generating field): είναι το πεδίο που σύμφωνα με το Λήμμα 1 (FH 1976) αποτελεί την ένωση των τιμών των συμβαλλόντων κανόνων στον έμμεσο κανόνα

Δεσμός (node): Ένας δεσμός είναι ο συνδυασμός πεδίων ή κανόνων ανάλογα με το δάσος κάλυψης που σχηματίζουμε. Οι αριθμοί στους δεσμούς αναφέρονται στους συνδυασμούς των πεδίων ή των κανόνων. Στους δεσμούς που συνδυάζουν πεδία ο πρώτος αριθμός είναι ο γεννήτορας πεδίο που θα παράγει έμμεσους κανόνες χρησιμοποιώντας ως συμβάλλοντες κανόνες, τους κανόνες που έχουν παραχθεί μέχρι τη στιγμή της άφιξης στο συγκεκριμένο δεσμό.

Δευτερεύοντα σύνολα (secondary datasets): είναι σύνολα δεδομένων που αποτελούν πηγές για τις δευτερεύουσες μεταβλητές

Δευτερεύουσα μεταβλητή (secondary): είναι μια έμμεση μεταβλητή που παίρνει τιμές από σύνολα δεδομένων όχι όμως του πρωταρχικού

Διόρθωση (imputation): αποτελεί το τελικό στάδιο του στατιστικού ελέγχου ορθότητας κατά το οποίο αλλάζουμε τις ελλειπείς λανθασμένες τιμές ή ασύμβατους συνδυασμούς τιμών των πεδίων που χρειάζονται διόρθωση ώστε να λάβουμε «καθαρές» εγγραφές που ικανοποιούν όλους τους κανόνες

Δίτιμο δέντρο (binary tree): χρησιμοποιείται στη μέθοδο εύρεσης των πεδίων για διόρθωση από αριθμητικά δεδομένα και δημιουργείται από τις επιλογές μας να αντικαταστήσουμε ή να απαλείψουμε ένα πεδίο

Δίτιμη κατάτμηση (binary segmentation): είναι μια στατιστική μέθοδος που χωρίζει ένα σύνολο παρατηρήσεων σε ευδιάκριτες ομάδες ανάλογα με τις τιμές που περνούν με τελικό σκοπό την αυτόματη παραγωγή κανόνων

Δωρητής (donor): είναι μια εγγραφή η οποία ικανοποιεί όλους τους κανόνες και μπορεί να «δανείσει» τις τιμές ορισμένων ή όλων των πεδίων της για να διορθωθούν πεδία εγγραφών που δεν περνούν τους κανόνες.

Εγγραφή (record): αποτελεί το σύνολο των απαντήσεων ενός ερωτώμενου και αποτελείται από πεδία ή μεταβλητές.

Έλεγχος βασισμένος σε πεδία (domain-based validation): εννοούμε απλά τον έλεγχο κάθε μεταβλητής για το αν παίρνει επιτρεπτή τιμή (μέσα στο πεδίο που ορίζει ο κανόνας) ή όχι.

Έλεγχος δεδομένων (data validation): ο έλεγχος των στατιστικών δεδομένων για ελλειπείς τιμές, λάθη και λογικές ασυνέχειες χωρίς να αναλαμβάνει τον εντοπισμό των λαθών και τη διόρθωσή τους

Έλεγχος ορθότητας σημαντικότητας (significance editing): εκτιμάται κατά πόσο μια εγγραφή επηρεάζει το αποτέλεσμα πριν την επεξεργασία όλων των εγγραφών

Έμμεση μεταβλητή (implicit variable): είναι ατομικές μεταβλητές που δεν υπάρχουν στο πρωταρχικό σύνολο δεδομένων και ορίζεται εξαιτίας ενός κανόνα

Έμμεσος κανόνας ορθότητας (implicit edit): είναι ο κανόνας που παράγεται με το Λήμμα 1(FH 1976) από συνδυασμούς άλλων κανόνων.

Επαρκές σύνολο κανόνων (sufficient set of edits): είναι το υποσύνολο του πλήρους συνόλου κανόνων αλλά ισοδύναμο με αυτό. Κατά τη μέθοδο GKL 1986 είναι το σύνολο των μέγιστα κατά βάση νέων κανόνων

Επαυξημένο σύνολο (enhanced dataset): αποτελεί την ένωση του πρωταρχικού συνόλου δεδομένων με όλες τις σύνθετες μεταβλητές που παράχθηκαν λόγω των κανόνων

Επιλεκτικός έλεγχος (selective editing): διακρίνει τις πιο σημαντικές εγγραφές που χρειάζονται διόρθωση και τις παραπέμπει στους ειδικούς αναλυτές ενώ οι υπόλοιπες διορθώνονται αυτόματα.

Εφικτή πράξη διόρθωσης (feasible): χαρακτηρίζεται μια πράξη διόρθωσης που περνά όλους τους κανόνες

Ισοδύναμα σύνολα κανόνων (equivalent set of edits): τα δύο σύνολα κανόνων $\overline{E'}, \overline{E}$ λέγονται ισοδύναμα και γράφουμε $P(\overline{E'}) = P(\overline{E})$ όταν η λύση του μοντέλου ελαχίστων σταθμισμένων πεδίων για διόρθωση είναι ίδια για οποιαδήποτε επιλογή ανάμεσα στα δύο σύνολα.

Κανόνας αντίφασης (conflict rule): όταν μια εγγραφή δεν περνάει τον κανόνα αυτός λέγεται κανόνας αντίφασης στους λογικούς πίνακες απόφασης

Κατά βάση νέος κανόνας (essentially new implicit edit): είναι ο έμμεσος κανόνας στον οποίο δεν περιέχεται ο γεννήτορας πεδίο ενώ ο τελευταίος περιέχεται στους συμβάλλοντες κανόνες.

Κανόνας βασισμένος στον τύπο εγγραφής (rule's base record type): ο κάθε κανόνας αναφέρεται σε ένα συγκεκριμένο τύπο εγγραφής σύμφωνα με τη μέθοδο ελέγχου βασισμένου σε πεδία μεταβλητών

Κανόνας εγκυρότητας (validity rule): Όταν μια εγγραφή περνάει ένα κανόνα αυτός ονομάζεται κανόνας εγκυρότητας στους λογικούς πίνακες απόφασης

Κανόνας ισορροπίας (balance edit): Αθροίζει τις τιμές μεταβλητών με τη τιμή μιας άλλης. Ο κανόνας χάνεται όταν δεν ισχύει η ισότητα.

Κανόνας μεταξύ προσώπων (between person edit rule): ο κανόνας που συγκρίνει στοιχεία δύο προσώπων, που αναπαριστώνται στους λογικούς πίνακες απόφασης

Κανόνας ορθότητας (edit rule): αποτελεί την κρίση των ειδικών για τις τιμές που λαμβάνει ένα πεδίο ή ο συνδυασμός πεδίων και είναι απίθανο να συμβεί.

Κανόνας πηλίκο (ratio edit): περιορίζει το πηλίκο δύο μεταβλητών ανάμεσα στο κάτω και το πάνω όριο. Τα όρια ορίζονται από τους ειδικούς. Ο κανόνας χάνεται όταν το πηλίκο των τιμών των μεταβλητών έχουν τιμή εκτός των ορίων.

Κανονική μορφή (normal form) του κανόνα E^i: είναι το καρτεσιανό γινόμενο της μορφής $E^i = \prod_j E_{ij}$ όπου E_{ij} είναι το σύνολο των τιμών που παίρνει το πεδίο j στον κανόνα i

Κοντινότερος γείτονας (nearest neighbor): είναι οι δωρητές οι οποίοι ταιριάζουν σε όσο το δυνατόν περισσότερες κατηγορικές μεταβλητές ενώ να διαφέρουν όσο το δυνατόν λιγότερο στις ποσοτικές μεταβλητές με τη εγγραφή που χρειάζεται διόρθωση

Κρίσιμοι κανόνες (fatal or critical edits): είναι αυτοί που ανιχνεύουν τιμές μεταβλητών που πρέπει οπωσδήποτε να διορθωθούν αν δεν τους ικανοποιούν

Λογικός κανόνας ορθότητας (logical edit): είναι ο κανόνας ορθότητας που αφορά κατηγορικές μεταβλητές και εκφράζεται ως καρτεσιανό γινόμενο των τιμών των μεταβλητών που είναι απίθανό να εμφανιστούν στα δεδομένα.

Μάκρο-έλεγχος (macro editing): ελέγχει τα δεδομένα μιας εγγραφής σε σχέση με τις υπόλοιπες (έλεγχος έκτροπων παρατηρήσεων, σύγκριση των τιμών μιας εγγραφής με άλλες πηγές)

Μέγιστος έμμεσος κανόνας (maximal implicit edit): E^i λέγεται μέγιστος όταν δεν υπάρχει άλλος κανόνας E τέτοιος ώστε $P(E^i) \subseteq P(E)$. Ο μέγιστος κανόνας δεν είναι υποσύνολο κανενός άλλου

Μεικτά δεδομένα (mixed data): αριθμητικά και κατηγορικά δεδομένα

Μεταβλητή ή πεδίο (field): η τιμή που προσδιορίζει ο ερωτώμενος σε μια απάντηση. Το σύνολο των πεδίων αποτελεί μια εγγραφή.

Μίκρο-έλεγχος (micro editing): ελέγχει για ελλείπουσες, λανθασμένες τιμές και ασυνέπειες σε κάθε εγγραφή ξεχωριστά.

Μοντέλο ελάχιστου αριθμού σταθμισμένων πεδίων για διόρθωση (minimal weighted fields to impute MWFI): είναι η μαθηματικοποιημένη έκφραση του προβλήματος εντοπισμού λαθών. Συγκεκριμένα ελαχιστοποιώ το σταθμισμένο άθροισμα όλων των πεδίων με την προϋπόθεση ότι πρέπει να διορθωθεί τουλάχιστον ένα πεδίο από κάθε εγγραφή.

Παραγόμενη μεταβλητή (derived): είναι μια έμμεση μεταβλητή που παίρνει τιμές από άλλες εγγραφές του ίδιου συνόλου δεδομένων και όχι από την εγγραφή στην οποία ανήκει στο αφηρημένο μοντέλο δεδομένων

Παρατήρηση (observation): Ομάδες από εγγραφές σχηματίζουν μια παρατήρηση σύμφωνα με το μοντέλο βασισμένο σε πεδία μεταβλητών. Για παράδειγμα κάθε κατοικία αποτελείται από τα δεδομένα του τύπου εγγραφής της κατοικίας, μια ή περισσότερες εγγραφές της οικογένειας και μιας ή περισσότερων εγγραφών για το κάθε άτομο ξεχωριστά.

Περίπτωση ελέγχου (validation case): με την έννοια αυτή περιγράφεται αφηρημένα ένας κανόνας χωρίς τον ακριβή προσδιορισμό του αλλά με την απλή αναφορά των μεταβλητών που περιέχονται σε αυτόν. Μια περίπτωση ελέγχου μπορεί να μεταφραστεί σε περισσότερους από έναν κανόνες.

Περιττός κανόνας (redundant edit): ονομάζεται ένας κανόνας όταν περιέχεται σε κάποιον άλλο ως υποσύνολο του καρτεσιανού του γινομένου.

Πηγή φόρμας δεδομένων (source template): είναι η περιγραφή της κατασκευής της αυθαίρετης πηγής. Λέγοντας κατασκευή εννοούμε τον ορισμό των τύπων των εγγραφών εκείνων των μεταβλητών που θα περάσουν στο ενισχυμένο σύνολο δεδομένων σύμφωνα με το μοντέλο βασισμένο σε πεδία μεταβλητών

Πλήρες σύνολο κανόνων (complete set of edits): αποτελεί το σύνολο των κανόνων που είναι επαρκές για να λύσει το πρόβλημα εντοπισμού λαθών

Πραγματική μεταβλητή (actual variable): είναι η ατομική μεταβλητή για την οποία συλλέγονται τα δεδομένα στο πρωταρχικό σύνολο σύμφωνα με το μοντέλο βασισμένο σε πεδία μεταβλητών

Πράξη διόρθωσης (imputation action): Η διόρθωση ενός συνόλου πεδίων μιας εγγραφής ώστε η εγγραφή να περνάει όλους τους κανόνες σύμφωνα με τη μέθοδο NIM

Πράξη διόρθωσης σχεδόν ελάχιστης αλλαγής (near minimum change imputation action NMCIA): Η εφικτή πράξη διόρθωσης που διορθώνει σχεδόν το μικρότερο αριθμό μεταβλητών (με κριτήριο την απόσταση)

Πρόβλημα εντοπισμού λαθών (error localization problem): αποτελεί την εύρεση του ελάχιστου αριθμού πεδίων για διόρθωση

Πρόβλημα κάλυψης συνόλου (set covering problem): ένα παράδειγμα αυτού είναι το MWFI. ένα ανάλογο παράδειγμα είναι η εύρεση του συνόλου των κανόνων για την παραγωγή έμμεσων κατά βάση νέων κανόνων.

Προϋποθετική μεταβλητή (precondition variables): είναι μια μεταβλητή που αν χάνει ένα κανόνα δεν επιτρέπει περαιτέρω έλεγχο στην ίδια παρατήρηση

Πρωταρχικό σύνολο δεδομένων (primary dataset): είναι τα δεδομένα που συλλέχθηκαν από τις απαντήσεις των ερωτηματολογίων μιας έρευνας

Ρητά ορισμένος κανόνας (explicit edit): είναι ο κανόνας που ορίζεται από τους ειδικούς αναλυτές. Το σύνολό τους αποτελεί το σύνολο των ρητά ορισμένων κανόνων που χρησιμοποιείται για την παραγωγή νέων

Στατιστική μονάδα (statistical unit): είναι ένα στοιχείο του πληθυσμού της έρευνας

Στοιχείο πεδίου (domain element): η ένωση όλων των στοιχείων πεδίου συνθέτει το πεδίο ορισμού της μεταβλητής

Συμβάλλοντες κανόνες (contributing edits): είναι οι κανόνες που συνδυάζονται για την παραγωγή του έμμεσου κανόνα

Συμπερασματικός κανόνας (induced edits): είναι ο έμμεσος κανόνας που δημιουργείται με την αντικατάσταση των κανόνων πηλίκων από συγκεκριμένες μεταβλητές, σε αντίστοιχους όρους των κανόνων ισορροπίας

Συνεπές σύνολο (consistent set): λέγεται το σύνολο των κανόνων που δεν περιέχει αντικρουόμενους κανόνες. Επίσης αν υπάρχει τουλάχιστον μια εγγραφή από το σύνολο των δεδομένων που περνάει όλους τους κανόνες το σύνολο θεωρείται συνεπές.

Σύνθετη μεταβλητή (composite variable) : ένα διάνυσμα ατομικών μεταβλητών.

Υπολογιστική μεταβλητή (computed): είναι μια έμμεση μεταβλητή που παράγεται ως συνάρτηση άλλων μεταβλητών

Φόρμα δεδομένων (dataset template): η περιγραφή της κατασκευής του ενισχυμένου συνόλου δεδομένων μαζί με τα πεδία ορισμού των μεταβλητών

ΒΙΒΛΙΟΓΡΑΦΙΑ

Bankier, M., Fillion, J.-M., Luc, M. and Nadeau, C. (1994). Imputing Numeric and Qualitative Variables Simultaneously. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 242-247.

Bankier, M., Lue, M., Nadeau, C. and Newcombe,P. (1995). Additional Details on Imputing Numeric and Qualitative Variables Simultaneously. *Proceedings of the Section on Survey Research Methods*, American StatisticalAssociation, 287-292.

Bankier, M., Luc, M., Nadeau, C. and Newcombe, P.(1996). Imputing Numeric and Qualitative Census Variables Simultaneously. *Proceedings of the Survey Research Methods Section,* American Statistical Association, (1996).

Bankier, M., Houle, A.-M., Luc, M. and Newcombe, P. (1997). 1996 Canadian Census DemographicVariables Imputation. *American Statistical Association, Proceedings of the 1997 Section on Survey Research Methods*, 389-394.

Bankier, M. (2000). 2001 Canadian Census Minimum Change Donor Imputation Methodology. *U.N. Economic Commission for Europe Work Session on Statistical Data Editing*, Cardiff, UK, October 2000

Bankier, M.(1999). Experience with the New Imputation Methodology used in the 1996 Canadian Census with Extensions for Future Censuses. *U.N. Economic Commission for Europe Work Session on Statistical Data Editing*, Rome, Italy, June 1999

Biemer, P. Lyberg, L. (2003). *Introduction to Survey Quality*. New York: Wiley

Chen, B.-C. (1998). Set Covering Algorithms in Edit Generation, American Statistical Association. *Proceedings of the Section on Statistical Computing*, 91-96

De Waal, T. and Van de Pol, F. (1997). A recipe for applying CherryPi to the edit process. *UN Work Session on Statistical Data Editing,* 14-17, Prague, Czech Republic October 1997

De Waal, T. (2003). Solving The Error Localization Problem by Means of Vertex Generation. *Survey Methodology* , Vol. **29**, No. 1, pp 71-79 June (2003)

De Waal, T. and Quere, R. (2003). A Fast and Simple Algorithm for Automatic Editing of Mixed Data. *Journal of Official Statistics*, Vol. **29**, No. 4, 2003, pp. 383-402 (2003)

Draper, L. and Winkler, W.E. (1997), Balancing and Ratio Editing with the new SPEER system, American Statistical Association, *Proceedings of the 1997 Section on Survey Research Methods*, 570-575

European Commission, Eurostat (2003). Structure of Earnings Survey 2002, Eurostat arrengenets for implementing the Council Regulation 530/1999 and the Commission Regulation 1916/2000. (2003)

Eurostat, GENEDI , Generic EDI toolbox. *Parameters Guide.* European Commission (2001)

Farmakis G., Figueiredo J., Mota D., Petrakos G., Santos D., Stavropoylos P., (2004)"An alternative approach for the implementation of data editing: the INSPECTOR project", Q2004, *European Conference on Quality and methodology in Official Statistics,* Mainz, Germany. (2004)

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, **71**, 17-35.

Garfinkel, R. S., Kunnathur, A. S. and Liepins, G. E., (1986). Optimal Imputation of Erroneous Data: Categorical Data, General Edits. *Operations Research*, **34**, 744-751.

Granquist, L. and Kovar, J. G. (1997). Editing of Survey Data: How much is enough? *Survey Measurement and Process Quality,* L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin (eds), New York: Wiley, 415-435

Kovar, J.G., and Winkler, W.E., (1996). Editing Economic Data. American Statistical Association*, Proceedings of the Section on Survey Research Methods*, 81-87

Mniestris, N., (2004). *An overview of automatic data editing*. MSc thesis, Athens University of Economic and Business (2004)

Morlini, I., (1998). *Multivariate outlier detection with Kohonen networks: an useful tool for routine exploration of large data sets.* International Conference on New Techniques and Technologies for statistics. Sorrento, Italy (1998) 345-350

Petrakos G., Conversano C., Farmakis G., Mola F., Siciliano R., Stavropoulos P., (2004). News Ways of Specifying Data Edits. *J. R. Statist. Soc.***167**, Part 2, pp. 249-274 (2004)

Poirier, C. (1999). A Functinal Evaluation of Edit and Imputation Tools. *U.N. Economic Commission for Europe Work Session on Statistical Data Editing*, Rome, Italy, June 1999

Quere, R. (2000). Automatic editing of numerical data. *Technical Report*, Statistics Netherlands.

Quere, R. and De Waal, T. (2000). Error lacalization in mixed data sets. *Technical Report*, Statistics Netherlands.

Winkler, W.E. (1998). Set-Covering and Editing Discrete Data. American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 564-569

Winkler, W.E. (1999). State of Statistical data editing and current research problems. *Working paper No 29 in the UN/ECE Work Session on Statistical data editing*, Rome, Italy, June 1999

Winkler, W. E., and Petkunas, T. (1996). The DISCRETE Edit System, in J. Kovar and L. Granquist, (eds.) Statistical Data Editing, Vol **II**, *U.N. Economic Commission for Europe*, 56-62.

Winkler, W. E., (1997). Editing Discrete Data, *Statistical Research Report Series,* US Bureau of the Census

Winkler, W. E. and Chen, B. C. (2002). Extending the Fellegi-Holt model of statistical data editing, *Statistical Research Report* 2002/02, Statistical Research Division, US Bureau of the Census, Washington DC