ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Χημικών Μηχανικών

Τομέας ΙΙ: Ανάλυσης, Σχεδιασμού και Ανάπτυξης

Διεργασιών και Συστημάτων

Διπλωματική Εργασία:

Ανάπτυξη Μοντέλων Ανάλυσης Κύκλου Ζωής για Βέλτιστο Σχεδιασμό Καινοτόμων Χημικών Προϊόντων

Όνομα: Παντελής Μπαξεβανίδης-Τάρος

Υπεύθυνος Καθηγητής: κ. Αντώνης Κοκόσης

Επιβλέποντες: Δρ. Ευτυχία Μαρκουλάκη, Ερευνήτρια Β', ΕΚΕΦΕ ΔΗΜΟΚΡΙΤΟΣ

Δρ. Σταύρος Παπαδοκωνσταντάκης, Ερευνητής, ETH Zurich

ΑΘΗΝΑ, 2014

Ευχαριστίες

Η παρούσα διπλωματική εργασία αποτελεί μια ενδελεχή έρευνα για τη βελτίωση των μεθοδολογιών Μοριακού Σχεδιασμού(CAMD) με στόχο τη βελτιστοποίηση διεργασιών και μια καινοτόμα δουλειά όσον αφορά στην ανάπτυξη προβλεπτικών μοντέλων Συνεισφοράς Ομάδων για τη μεθοδολογία Ανάλυσης Κύκλου Ζωής (LCA).

Οφείλω ιδιαίτερες ευχαριστίες στον επιβλέποντα καθηγητή κ. Αντώνη Κοκόση, για την καθοδήγηση και την αμέριστη βοήθεια του σε όλη τη διάρκεια του project. Η ευκαιρία που μου έδωσε να εργαστώ και να εκπονήσω μέρος της διπλωματικής στο Πολυτεχνείο της Ζυρίχης (ETH Zurich) ήταν μια μοναδική εμπειρία ζωής και τον ευχαριστώ θερμά για την εμπιστοσύνη του σε εμένα και τη στήριξη του.

Να εκφράσω, επίσης, την απεριόριστη ευγνωμοσύνη μου στην κα. Έφη Μαρκουλάκη, Δρ. Χημικό Μηχανικό και Κύρια Ερευνήτρια στο Εργαστήριο Βιομηχανικής Ασφάλειας και Αξιοπιστίας Συστημάτων του Ε.Κ.Ε.Φ.Ε. Δημόκριτος, πάνω στη δουλειά της οποίας βασίστηκε η παρούσα εργασία και δίχως τη βοήθεια της οποίας η εργασία αυτή θα ήταν αδύνατο να ολοκληρωθεί. Ήταν ξεχωριστή εμπειρία και μεγάλη μου τιμή και χαρά να έχω συνεργαστεί μαζί της και την ευχαριστώ θερμότατα για όλα όσα μου δίδαξε αλλά και για την υπομονή της, τη στήριξη της, την εμπιστοσύνη της και την πίστη της στο πρόσωπο μου όλο αυτό το χρονικό διάστημα.

Τέλος, να εκφράσω θερμότατες ευχαριστίες στον κο. Σταύρο Παπαδοκωνσταντάκη, Δρ. Χημικό Μηχανικό και Senior Scientist στο ETH Zurich, ο οποίος συνείσφερε καθοριστικά στην επιτυχία αυτής της εργασίας. Ήταν επίσης, μεγάλη χαρά να βρεθώ και να συνεργαστώ μαζί του στη Ζυρίχη και τον ευχαριστώ για την αφοσίωση του στη δουλειά μας, αλλά και για τον αγώνα που έδινε μαζί μου σε καθημερινή βάση για να δημιουργηθεί μια επιστημονικά άρτια μελέτη. Ήταν μεγάλη μου τιμή που δούλεψα στο Εργαστήριο Ασφάλειας και Περιβάλλοντος (Safety and Environmental Technology Group) της Σχολής Χημικών Μηχανικών του ETH Zurich υπό τον καθηγητή K. Hungerbuehler και ευχαριστώ πολύ όλους τους ανθρώπους που έκαναν την παραμονή μου εκεί ευχάριστη.

Περιεχόμενα

Λίστα πινάκων

Σελίδα

Λίστα Διαγραμμάτων

Σελίδα

Σελίδα

Σελίδα

ΠΕΡΙΛΗΨΗ

Στην παρούσα διπλωματική εργασία αναπτύσσονται μοντέλα συνεισφοράς ομάδων για τον υπολογισμό τριών περιβαλλοντικών δεικτών και τη μετέπειτα χρήση τους σε διαδικασία Μοριακού Σχεδιασμού υποβοηθούμενου από Ηλεκτρονικό Υπολογιστή (Computer Aided Molecular Design – CAMD). Ο τελικός σκοπός είναι η αναζήτηση και ο σχεδιασμός βέλτιστων χημικών ουσιών οι οποίες να ικανοποιούν και περιβαλλοντικά κριτήρια.

Οι τρεις δείκτες Δυναμικό Παγκόσμιας Υπερθέρμανσης (Global Warming Potential), Σωρευτική Ενεργειακή Απαίτηση (Cumulative Energy Demand) και ο Οικολογικός Δείκτης 99 (Ecoindicator 99). που μελετώνται εδώ εμπίπτουν στη μεθοδολογία της Ανάλυσης Κύκλου Ζωής (Life Cycle Analysis – LCA). Τα μοντέλα LCA αναπτύσσονται από διαθέσιμα δεδομένα στο Τεχνολογικό Ίδρυμα της Ζυρίχης (ETH Zurich), τα οποία είναι είτε πειραματικά είτε υπολογισμένα μέσω του εργαλείου FineChem. Τα μοντέλα αυτά είναι τύπου συνεισφοράς λειτουργικών ομάδων, ώστε να επιτρέπεται η απρόσκοπτη χρήση τους σε εργαλεία CAMD. Για την ανάπτυξή τους γίνεται χρήση στατιστικών μεθόδων Γραμμικής Παλινδρόμησης Πολλών Μεταβλητών (Multiple Linear Regression), Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis - PCA), Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων (Partial Least Squares), Παλινδρόμησης τύπου «Kriging», Συναρτήσεων Ακτινικής Βάσης (Radial Basis Functions - RBF), καθώς και συνδυασμού RBF με PCA.

Από την παραπάνω στατιστική ανάλυση επιλέγονται τα βέλτιστα μοντέλα για τον προσδιορισμό των δεικτών GWP, CED και EI 99, τα οποία και ενσωματώνονται στις βάσεις δεδομένων υπάρχοντος λογισμικού CAMD. Το λογισμικό CAMD βασίζεται στο στοχαστικό αλγόριθμο βελτιστοποίησης της Προσομοιωμένης Ανόπτησης (Simulated Annealing) και το ανέπτυξαν οι Marcoulaki και Kokossis (2000a). Επιλέγεται ως μελέτη περίπτωσης ο σχεδιασμός βέλτιστων διαλυτών για διεργασία εκχύλισης υγρού-υγρού του συστήματος κανονικής βουτανόλης – νερού. Το σύστημα αυτό παρουσιάζει αζεότροπο που καθιστά αδύνατο τον διαχωρισμό του με απλή κλασματική απόσταξη. Η προσθήκη των κριτηρίων LCA επιτρέπει το σχεδιασμό διαλυτών, οι οποίοι, πέραν των συμβατικών ιδιοτήτων όπως η διαλυτότητα ή οι απώλειες διαλύτη κ.ά., ικανοποιούν και περιβαλλοντικές απαιτήσεις. Όπως αναμένεται, οι τελικές λύσεις διαμορφώνονται ανάλογα με τις απαιτήσεις LCA, ενώ η χρήση του στοχαστικού εργαλείου CAMD επιτρέπει την λήψη ενός συνόλου τελικών λύσεων για κάθε ομάδα περιορισμών σχεδιασμού.

Κεφάλαιο 1. Εισαγωγή

Η βελτιστοποίηση διεργασιών αποτελεί βασική ενασχόληση της επιστήμης του Χημικού Μηχανικού. Μία από τις πιο καλώς μελετημένες και ανεπτυγμένες τεχνικές για την επίτευξη βέλτιστων διεργασιών είναι ο Μοριακός Σχεδιασμός με χρήση Ηλεκτρονικού Υπολογιστή. Πρόκειται για μεθοδολογίες που αξιοποιούν ένα ευρύ φάσμα στατιστικών μαθηματικών μοντέλων και πλήθος μοντέλων πρόρρησης ιδιοτήτων με σκοπό την αναζήτηση και σχεδιασμό χημικών μέσων (διαλυτών, ψυκτικών, θερμαντικών κ.α.) συγκεκριμένων προδιαγραφών που καθιστούν μια διεργασία βέλτιστη.

Η σύγχρονη, όμως, τάση και πολιτική για αειφόρο και πράσινη ανάπτυξη επιτάσσουν τη στροφή προς περιβαλλοντικά φιλικές διεργασίες. Το πρόβλημα του Μοριακού Σχεδιασμού ανάγεται, λοιπόν, σε πρόβλημα εύρεσης «πράσινων» χημικών μέσων, τα οποία να επιτυγχάνουν τους βέλτιστους στόχους και παράλληλα, θα υπακούν σε συγκεκριμένους περιβαλλοντικούς περιορισμούς. Το πιο κοινώς εφαρμοσμένο και ανεπτυγμένο πλαίσιο για να εξεταστεί η περιβαλλοντική απόδοση ενός χημικού μέσου είναι η Ανάλυση Κύκλου Ζωής, ένα εργαλείο που επιτρέπει τον ποσοτικό προσδιορισμό της περιβαλλοντικής απόδοσης ενός χημικού, εξετάζοντας ολόκληρο τον Κύκλο Ζωής του. Η μεθοδολογία αυτή περιλαμβάνει πλήθος περιβαλλοντικών δεικτών, οι οποίοι εξετάζουν διάφορες συνιστώσες της περιβαλλοντικής επιβάρυνσης που μπορεί να επιφέρει ένα προϊόν στον Κύκλο Ζωής του.

Γίνεται εύκολα κατανοητό πως η μεθοδολογία του Μοριακού Σχεδιασμού μπορεί να πλέον να κατευθυνθεί, σε συνδυασμό με την Ανάλυση Κύκλου Ζωής, προς λύσεις που να καλύπτουν και τους περιβαλλοντικούς στόχους και τους στόχους της βελτιστοποίησης. Παρά ταύτα, η έλλειψη αξιόπιστων και συστηματικών μοντέλων πρόβλεψης δεικτών Ανάλυσης Κύκλου Ζωής, που να επιτρέπουν την εύκολη και γρήγορη χρήση τους για την εξαγωγή αποτελεσμάτων στις επαναληπτικές διαδικασίες του Μοριακού Σχεδιασμού, καθιστά αδύνατη την υλοποίηση της παραπάνω ιδέας. Είναι, λοιπόν, απαραίτητη η ανάπτυξη μοντέλων πρόβλεψης δεικτών με χρήση πειραματικών δεδομένων.

Η παρούσα εργασία πραγματεύεται την ανάπτυξη μοντέλων υπολογισμού δεικτών Ανάλυσης Κύκλου Ζωής και η ενσωμάτωση τους σε ήδη υπάρχουσα διαδικασία Μοριακού Σχεδιασμού. Στο κεφάλαιο 2, γίνεται εισαγωγή στις Μεθόδους Συνεισφοράς Ομάδων, που αποτελούν το είδος μοντέλων που χρησιμοποιείται στο Μοριακό Σχεδιασμό και το είδος των μοντέλων που αναπτύσσεται και ακολουθεί μια παρουσίαση των ήδη υπαρχόντων μοντέλων. Στο 3ο κεφάλαιο γίνεται παρουσίαση της μεθοδολογίας Ανάλυσης Κύκλου Ζωής και των ήδη υπαρχόντων μοντέλων για τον υπολογισμό των δεικτών. Έπειτα, ακολουθεί, στο 4ο κεφάλαιο, εισαγωγή στις στατιστικές διαδικασίες που χρησιμοποιούνται για την ανάπτυξη των μοντέλων πρόβλεψης και στο 5ο αναλύεται η χρήση κάθε στατιστικής διαδικασίας στην ανάπτυξη μοντέλων κάθε δείκτη. Ασφαλώς, προτείνεται εκείνη η διαδικασία που επιτρέπει την ανάπτυξη των πιο αξιόπιστων συσχετίσεων. Θεωρείται, επίσης, χρήσιμο να παρουσιαστούν και να επεξηγηθούν

αναλυτικά στο 6ο κεφάλαιο όλα τα στάδια που ακολουθούνται στη μεθοδολογία Μοριακού Σχεδιασμού. Τέλος, γίνεται ο συνδυασμός των δύο μεθοδολογιών και επανεξετάζεται μία μελέτη περίπτωσης υπό το πρίσμα της περιβαλλοντικά φιλικής βελτιστοποίησης.

Κεφάλαιο 2. Μέθοδοι Συνεισφοράς Ομάδων

2.1 Εισαγωγή σε μεθόδους Συνεισφοράς Ομάδων

Για τη χρήση του εργαλείου μοριακής βελτιστοποίησης που περιγράφηκε στο Κεφάλαιο 1 απαιτούνται μοντέλα υπολογισμού των θερμοδυναμικών, φυσικών και άλλων ιδιοτήτων των σχεδιαζόμενων μορίων. Τα μοντέλα αυτά ονομάζονται μοντέλα Συνεισφοράς Ομάδων (Group Contribution models)[1] είναι ιδιαίτερα διαδεδομένα στον υπολογισμό των διάφορων ιδιοτήτων μορίων για τα οποία δεν υπάρχουν σχετικά πειραματικά δεδομένα.

Τα μοντέλα GC ξεκίνησαν ως μοντέλα υπολογισμού θερμοδυναμικών και φυσικών ιδιοτήτων (Lydersen, 1955) και σύντομα επεκτάθηκαν ώστε να προβλέπουν πληθώρα άλλων ιδιοτήτων όπως η τοξικότητα (Gao, 1991). Για ένα μόριο του οποίου είναι επιθυμητή η πρόβλεψη των ιδιοτήτων του, αναγνωρίζονται πρώτα ποιες ομάδες αποτελούν τα επιμέρους τμήματα του. Δηλαδή, «αποδομείται» το μόριο στις επιμέρους ομάδες. Κάθε ομάδα έχει ξεχωριστή επίδραση (συνεισφορά) ως σύνολο στη διαμόρφωση της κάθε ιδιότητας, η οποία εκφράζεται από έναν αριθμό που ποσοτικοποιεί την επίδραση που έχει η συγκεκριμένη ομάδα στην τιμή της ιδιότητας του μορίου όπου η ομάδα συμμετέχει. Στο τέλος, συνυπολογίζονται οι συνεισφορές όλων των ομάδων, μέσω μιας εξίσωσης και προκύπτει η τιμή της κάθε ιδιότητας. Επομένως, τα μοντέλα GC μπορούν να προκύπτουν μετά από παλινδρόμηση σε πειραματικά δεδομένα υποθέτοντας τη σχέση:

$$X_{GC} = f(N_i \cdot C_i) \qquad (2.1)$$

όπου X_{GC} η εκτιμώμενη τιμή της ιδιότητας X, N_i οι επαναλήψεις της ομάδας i στο μόριο και C_i η συνεισφορά της ομάδας i στην τιμή της ιδιότητας X και f η συνάρτηση του μοντέλου GC. Συχνά οι μέθοδοι GC υποθέτουν προσθετικότητα, ότι δηλαδή μια ομάδα συνεισφέρει στις ιδιότητες του μορίου ανεξάρτητα με τις άλλες ομάδες που συμμετέχουν. Έτσι καταλήγουμε σε γραμμικά μοντέλα GC. Περισσότερα στοιχεία σχετικά με την ανάπτυξη μεθόδων GC παρέχονται στο Κεφάλαιο 5 όπου περιγράφεται η ανάπτυξη τέτοιων εργαλείων για χαρακτηριστικά σχετικά με τον κύκλο ζωής ενός μορίου.

Τα πλεονεκτήματα των μεθόδων GC είναι ότι είναι απλές στη χρήση και γρήγορες στο υπολογισμό των ιδιοτήτων. Κυρίως όμως, δεν απαιτούν πειραματικά μετρημένες παραμέτρους για το κάθε μόριο. Αυτό μειώνει δραστικά τις απαιτήσεις για διεξαγωγή πειραμάτων για οργανικές ενώσεις που το πλήθος τους αυξάνεται εκθετικά με το πλήθος το ατόμων άνθρακα. Για παράδειγμα, τα αλκάνια με 32 άτομα άνθρακα είναι πάνω από 27 δισεκατομμύρια μόρια και για την πλειοψηφία τους δεν έχουμε πειραματικές μετρήσεις. Με τη χρήση μεθόδων GC αρκούν οι συνεισφορές των τριών ομάδων CH_k, $k \in \{0,1,2,3\}$ για να προσδιορίσουμε τις ιδιότητες τους.

[1] Στο εξής τα μοντέλα και οι μέθοδοι συνεισφοράς ομάδων θα αναφέρονται για λόγους συντομίας ως GC.

Τα μειονεκτήματα των μεθόδων GC είναι ότι δεν ξεχωρίζουν ισομερή μόρια, εφόσον λαμβάνονται υπόψη μόνο οι ομάδες που τα απαρτίζουν, και ότι οι υπολογισμοί τους έχουν πολύ μεγαλύτερα σφάλματα από μοντέλα που βασίζονται σε πειραματικές μετρήσεις για το υπό μελέτη μόριο. Τα σφάλματα αυτά ανέρχονται κατά μέσο όρο σε 40 %. .

Στο Παράρτημα Α παρουσιάζεται μια σύντομη μελέτη σχετικά με την αποδόμηση μορίων σε λειτουργικές ομάδες και τα σφάλματα στους υπολογισμούς κάποιων βασικών ιδιοτήτων.

2.2 Μοντέλα Συνεισφοράς Ομάδων για καθαρές ουσίες

Μια πλειάδα μοντέλων GC έχουν κατά καιρούς αναπτυχθεί για την πρόρρηση θερμοδυναμικών και φυσικών ιδιοτήτων καθαρών ουσιών (βλ. Poling et al., 2001, κεφάλαιο 8, σελίδα 335). Τα μοντέλα που καλούνται στην τελευταία έκδοση του λογισμικού των Marcoulaki και Kokossis (2000) είναι τα ακόλουθα: η κρίσιμη θερμοκρασία, η κρίσιμη πίεση, ο κρίσιμος όγκος, η πρότυπη ενθαλπία εξάτμισης, η πρότυπη ενέργεια Gibbs, η ειδική θερμότητα υγρών, το κανονικό σημείο βρασμού, το κανονικό σημείο τήξης, η τοξικότητα LC 50 και η μεθοδολογία UNIFAC για την πρόβλεψη συντελεστών ενεργότητας για μίγματα ενώσεων.

2.2.1 Πρότυπη ενέργεια Gibbs καθαρής ουσίας (standard Gibbs energy – ΔG_f)

Η ενέργεια Gibbs σε πρότυπες συνθήκες ορίζεται θερμοδυναμικά ως ΔG=ΔS-T·ΔH, όπου με ΔS συμβολίζεται η μεταβολή της εντροπίας μίας ουσίας και με ΔH, η μεταβολή της ενθαλπίας μίας ουσίας, σε θερμοκρασία 273 K και 1 bar. Οι μονάδες στις οποίες μετριέται η ενέργεια Gibbs είναι kJ/mol. Για την πρόρρηση της πρότυπης ενέργειας Gibbs βάσει των λειτουργικών ομάδων μιας ουσίας, οι Constantinou και Gani (1994), προτείνουν την παρακάτω συνάρτηση:

$$\Delta G_f - g_0 = \sum_i N_i \cdot g_{1i} + W \sum_j M_j \cdot g_{2j} \quad (2.2)$$

όπου g_0 η σταθερά του γραμμικού μοντέλου ίση με -14,828 kJ/mol, N_i αριθμός επανάληψης της κάθε ομάδας και g_{1i} η συνεισφορά της κάθε λειτουργικής ομάδας για την ενέργεια Gibbs. Το πρώτο άθροισμα του δεξιού μέλους της εξίσωσης (5.2) αφορά τη διάσπαση των μορίων σε απλές ομάδες ατόμων. Το δεύτερο άθροισμα αναφέρεται σε πιο λεπτομερείς πληροφορίες για τη δομή του μορίου.

Η παράμετρος W παίρνει την τιμή 0 ή 1. Αν W=0, τότε η εκτίμηση της ενέργειας Gibbs καθορίζεται μόνο από το πρώτο άθροισμα. Αντίθετα αν W=1, τότε λαμβάνονται υπόψη και οι δομικές πληροφορίες σχετικές με το πώς συνδέονται οι λειτουργικές ομάδες μεταξύ τους. Για την απεικόνιση τέτοιων

πληροφοριών, οι Constantinou και Gani πρότειναν τη χρήση εκτενέστερων λειτουργικών ομάδων τις οποίες ονόμασαν ομάδες δεύτερης τάξης. Με M_j συμβολίζεται ο αριθμός των επαναλήψεων των ομάδων δεύτερης τάξης και με g_{2j}, συμβολίζεται η συνεισφορά τους. Αυτές οι ομάδες δεύτερης τάξης αποτελούνται από ομάδες πρώτης τάξης ενωμένες σε μεγαλύτερα σύνολα.

Σημειώνεται πως στο λογισμικό των Marcoulaki και Kokossis (2000α) χρησιμοποιείται μόνο η προσέγγιση πρώτης τάξης (δηλαδή W=0) για όλα τα μοντέλα πρόβλεψης ιδιοτήτων και ο αριθμός N_i των επαναλήψεων των ομάδων δίνεται από το μοριακό διάνυσμα της ένωσης.

2.2.2 Κρίσιμη Θερμοκρασία καθαρής ουσίας (critical temperature – T_c)

Η κρίσιμη θερμοκρασία είναι η θερμοκρασία που έχει μια καθαρή ουσία στο κρίσιμο σημείο της. Το κρίσιμο σημείο βρίσκεται εκεί που οι ιδιότητες του κορεσμένου υγρού και του κορεσμένου ατμού είναι ταυτόσημες και πέρα από αυτό δεν υπάρχει ευδιάκριτη διαδικασία αλλαγής φάσης. Η κρίσιμη θερμοκρασία μετριέται σε μονάδες θερμοκρασίας (Kelvin). Για την πρόρρηση της κρίσιμης θερμοκρασίας βάσει των λειτουργικών ομάδων μιας ουσίας, οι Constantinoukαι Gani (1994), προτείνουν την παρακάτω συνάρτηση:

$$exp\left(\frac{T_c}{t_{c0}}\right) = \sum_i N_i \cdot t_{c1i} + W \sum_j M_j \cdot t_{c2j} \qquad (2.3)$$

όπου t_{c0}, η σταθερά του μοντέλου ίση με 181,128 K και t_{cl1}, η συνεισφορά της κάθε λειτουργικής ομάδας για την κρίσιμη θερμοκρασία. Όπως και στο μοντέλο της πρότυπης ενέργειας Gibbs, μπορούν προαιρετικά να χρησιμοποιηθούν οι ομάδες δεύτερης τάξης θέτοντας W=1 και με t_{c2j} να συμβολίζονται οι συνεισφορές των ομάδων δεύτερης τάξης για την πρόβλεψη της κρίσιμης θερμοκρασίας.

2.2.3 Κρίσιμη Πίεση καθαρής ουσίας (critical pressure – P_c)

Η κρίσιμη πίεση είναι η πίεση που έχει μια καθαρή ουσία στο κρίσιμο σημείο της. Μετριέται σε μονάδες πίεσης (bar). Για την πρόρρηση της κρίσιμης πίεσης βάσει των λειτουργικών ομάδων μιας ουσίας, οι Constantinoukαι Gani (1994), προτείνουν την παρακάτω συνάρτηση:

$$(P_c - p_{c1})^{-0,5} - p_{c2} = \sum_i N_i \cdot p_{c1i} + W \sum_j M_j \cdot p_{c2j} \qquad (2.4)$$

όπου p_{c1} και p_{c2} οι σταθερές του μοντέλου ίσες με 1,3705bar και 0,10022 $bar^{-0,5}$, αντιστοίχως και p_{cl1}, η συνεισφορά της κάθε λειτουργικής ομάδας για την κρίσιμη πίεση. Όπως και στο μοντέλο της πρότυπης ενέργειας Gibbs, μπορούν προαιρετικά να χρησιμοποιηθούν οι ομάδες δεύτερης τάξης για W=1 και με p_{c2j} να συμβολίζονται οι συνεισφορές των ομάδων δεύτερης τάξης για την πρόβλεψη της κρίσιμης πίεσης.

2.2.4 Κρίσιμος Όγκος καθαρής ουσίας (critical volume – V_c)

Ο κρίσιμος όγκος είναι ο όγκος που καταλαμβάνει ένα kmol μιας ουσίας στο κρίσιμο σημείο της. Μετριέται με μονάδες όγκου ανά kmol (m^3/kmol). Για την πρόρρηση του κρίσιμου όγκου βάσει των λειτουργικών ομάδων μιας ουσίας, οι Constantinou και Gani (1994), προτείνουν την παρακάτω συνάρτηση:

$$V_c - u_{c0} = \sum_i N_i \cdot u_{c1i} + W \sum_j M_j \cdot u_{c2j} \quad (2.5)$$

όπου u_{c1}, η σταθερά του μοντέλου, ίση με -0,004350 m^3/kmol και u_{cl1}, η συνεισφορά της κάθε λειτουργικής ομάδας για τον κρίσιμο όγκο. Όπως και στο μοντέλο της πρότυπης ενέργειας Gibbs, μπορούν προαιρετικά να χρησιμοποιηθούν οι ομάδες δεύτερης τάξης για W=1 και με u_{c2j} να συμβολίζονται οι συνεισφορές των ομάδων δεύτερης τάξης για την πρόβλεψη του κρίσιμου όγκου.

2.2.5 Κανονικό σημείο βρασμού (normal boiling point – T_b)

Το κανονικό σημείο βρασμού ή κανονική θερμοκρασία βρασμού είναι η θερμοκρασία όπου ένα υγρό βράζει σε πίεση 1 atm. Το κανονικό σημείο βρασμού μετριέται σε μονάδες θερμοκρασίας (Kelvin). Για την πρόρρηση του κανονικού σημείου βρασμού βάσει των λειτουργικών ομάδων μιας ουσίας, οι Constantinou και Gani (1994), προτείνουν την παρακάτω συνάρτηση:

$$exp\left(\frac{T_b}{t_{b0}}\right) = \sum_i N_i \cdot t_{b1i} + W \sum_j M_j \cdot t_{b2j} \quad (2.6)$$

όπου t_{b0}, η σταθερά του μοντέλου ίση με 204,359 K και t_{b1i}, η συνεισφορά της κάθε λειτουργικής ομάδας για το κανονικό σημείο βρασμού. Όπως και στο μοντέλο της πρότυπης ενέργειας Gibbs, μπορούν προαιρετικά να χρησιμοποιηθούν οι ομάδες δεύτερης τάξης θέτοντας W=1 και με t_{b2j} να συμβολίζονται οι συνεισφορές των ομάδων δεύτερης τάξης για την πρόβλεψη του κανονικού σημείου βρασμού.

2.2.6 Πρότυπη ενθαλπία εξάτμισης (standard enthalpy of vaporization– ΔH_v)

Η πρότυπη ενθαλπία εξάτμισης είναι η λανθάνουσα θερμότητα που απαιτείται από ένα mol μιας ουσίας, ώστε να μεταβεί από την υγρή στην ατμώδη φάση σε πρότυπες συνθήκες (273K, 1 bar). Οι μονάδες της πρότυπης ενθαλπίας εξάτμισης είναι kJ/mol. Για την πρόρρηση της πρότυπης ενθαλπίας εξάτμισης βάσει των λειτουργικών ομάδων μιας ουσίας, οι Constantinou και Gani (1994), προτείνουν την παρακάτω συνάρτηση:

$$\Delta H_v - h_{v0} = \sum_i N_i \cdot h_{v1i} + W \sum_j M_j \cdot h_{v2j} \quad (2.7)$$

όπου h_{v0}, σταθερά του μοντέλου ίση με 6,829 kJ/mol και h_{v1i}, η συνεισφορά της κάθε λειτουργικής ομάδας για την πρότυπη ενθαλπία εξάτμισης. Όπως και στο μοντέλο της πρότυπης ενέργειας Gibbs, μπορούν προαιρετικά να χρησιμοποιηθούν οι ομάδες δεύτερης τάξης θέτοντας W=1 και με h_{v2j}, να συμβολίζονται οι συνεισφορές των ομάδων δεύτερης τάξης για την πρόβλεψη της πρότυπη ενθαλπία εξάτμισης.

2.2.7 Θερμοχωρητικότητα υγρού (liquid heat capacity– C_{pl})

Η θερμοχωρητικότητα υγρού (ή ειδική θερμότητα) είναι η θερμότητα που πρέπει να προσλάβει ένα mol ενός υγρού ώστε να μεταβληθεί η θερμοκρασία του κατά ένα βαθμό Kelvin. Οι μονάδες της είναι J/(mol·K). Για την πρόρρηση της θερμοχωρητικότητας βάσει των λειτουργικών ομάδων μιας ουσίας, ο Missenard (1965), προτείνει την παρακάτω συνάρτηση:

$$C_{pl}(T_j) = \sum_i N_i \cdot C_{pli}(T_j) \qquad (2.8)$$

όπου C_{pli}, η συνεισφορά της κάθε λειτουργικής ομάδας για τη θερμοχωρητικότητα υγρού και T_j, η θερμοκρασία του υγρού. Επειδή, η θερμοχωρητικότητα είναι και συνάρτηση της θερμοκρασίας, ο Missenard προτείνει συνεισφορές για ένα εύρος θερμοκρασιών (και συγκεκριμένα 248 K, 273 K, 298 K, 323 K, 348 K, 373 K), όπου μπορεί να προβλεφθεί η θερμοχωρητικότητα του κάθε υγρού.

2.2.8 Τοξικότητα LC 50 (LC 50 toxicity)

Η τοξικότητα LC50 εκφράζει εκείνη την περιεκτικότητα που για ένα συγκεκριμένο οργανισμό (fathead minnow), το 50% του δείγματος καταλήγει σε θάνατο μετά από έκθεση σε αυτή. Οι μονάδες της τοξικότητας LC 50 είναι mol/L. Για την πρόρρηση της τοξικότητας LC 50 βάσει των λειτουργικών ομάδων μιας ουσίας, οι Gao et al. (1991), προτείνουν την παρακάτω συνάρτηση:

$$-log(LC\ 50) = \sum_i N_i \cdot a_i \qquad (2.9)$$

όπου, η συνεισφορά της κάθε λειτουργικής ομάδας για την τοξικότητα LC 50.

2.3 Εισαγωγή στο μοντέλο της UNIFAC

2.3.1 Αρχικό μοντέλο UNIFAC

Η UNIFAC είναι ένα μοντέλο πρόρρησης του συντελεστή ενεργότητας που βασίζεται στη Συνεισφορά Λειτουργικών Ομάδων, όπως περιγράφτηκε στο Κεφάλαιο 2.1 της παρούσας εργασίας.

Αποτελεί μια αποτελεσματική και πολύ κοινή λύση στο πρόβλημα προσδιορισμού της συμπεριφοράς πολυσυστατικών μιγμάτων, κυρίως μιας και οι καταστατικές εξισώσεις και τα μοντέλα περίσσειας ενέργειας Gibbs (G^E models), απαιτούν πειραματικά δεδομένα, τα οποία μπορεί να μην υπάρχουν.

Στο μοντέλο UNIFAC (Fredenslund et al., 1975) ο νεπέριος λογάριθμος του συντελεστή ενεργότητας εκφράζεται ως το άθροισμα δύο συνεισφορών, ενός συνδυαστικού μέρους (combinatorial part) και ενός υπολειμματικού μέρους (residual part):

$$\ln \gamma_i = \ln \gamma_i{}^{comb} + \ln \gamma_i{}^{res} \qquad (2.10)$$

Το συνδυαστικό μέρος ($ln\gamma^{comb}$) λαμβάνει υπόψη τις διαφορές στο μέγεθος και το σχήμα των μορίων (εντροπική συνεισφορά), ενώ το υπολειμματικό μέρος ($ln\gamma^{res}$) τις ενεργειακές αλληλεπιδράσεις μεταξύ όλων των δομικών ομάδων (ενθαλπική συνεισφορά).

Η αρχική UNIFAC χρησιμοποιεί το συνδυαστικό μέρος της UNIQUAC, το αποκαλούμενο Staverman-Guggenheim combinatorial:

$$\ln \gamma_i{}^{comb} = \ln \frac{\varphi_i}{x_i} + 1 - \frac{\varphi_i}{x_i} - \frac{z}{2} \cdot q_i \cdot \left(\ln \frac{\varphi_i}{\theta_i} + 1 - \frac{\varphi_i}{\theta_i} \right) \qquad (2.11)$$

Στην εξίσωση (2.15), το φ_i είναι το κλάσμα όγκου του συστατικού i:

$$\varphi_i = \frac{x_i \cdot r_i}{\sum_j x_j \cdot r_j} \qquad (2.12)$$

όπου r_i είναι η παράμετρος όγκου του καθαρού συστατικού *i*. Επίσης, θ_i είναι το κλάσμα επιφάνειας του συστατικού *i* που ορίζεται από τη σχέση:

$$\theta_i = \frac{x_i \cdot q_i}{\sum_j x_j \cdot q_j} \qquad (2.13)$$

όπου q_i είναι η παράμετρος επιφανείας του καθαρού συστατικού *i*. Τέλος, *z* είναι ο αριθμός σύνταξης που τυπικά παίρνει την τιμή 10.

Οι παράμετροι όγκου και επιφανείας r_i και q_i είναι στην πραγματικότητα ο όγκος και η επιφάνεια van der Waals του συστατικού *i* και υπολογίζονται με τη βοήθεια των αντίστοιχων τιμών για τις ομάδες που δίνονται από τον Bondi (1968). Τιμές για τις παραμέτρους επιφάνειας και όγκου για τις διάφορες ομάδες δίνονται στους πίνακες παραμέτρων του μοντέλου UNIFAC. Στη UNIFAC καθορίζονται δύο διαφορετικά είδη λειτουργικών ομάδων, οι υποομάδες (sub-groups) και οι βασικές ομάδες (main groups). Οι υποομάδες είναι οι δομικές μονάδες στις οποίες χωρίζεται το μόριο, ενώ οι βασικές ομάδες χρησιμοποιούνται για την κατηγοριοποίηση των υποομάδων. Οι υποομάδες μιας βασικής ομάδας έχουν διαφορετικές παραμέτρους όγκου και επιφάνειας. Αντιθέτως, οι παράμετροι αλληλεπίδρασης ορίζονται μεταξύ των βασικών ομάδων. Για παράδειγμα, στη βασική ομάδα «CH_2» περιλαμβάνονται οι υποομάδες CH_3, CH_2, CH και C.

Το υπολειμματικό μέρος υπολογίζεται από τις ακόλουθες εκφράσεις:

$$\ln \gamma_i{}^{res} = \sum_k \nu_k{}^{(i)} \left(\ln \Gamma_k - \ln \Gamma_k{}^{(i)} \right) \qquad (2.14)$$

όπου

$$\ln \Gamma_k = Q_k \cdot \left[1 - \ln(\sum_m \theta_m \cdot \Psi_{mk}) - \sum_m \frac{\theta_m \cdot \Psi_{mk}}{\sum_n \theta_n \cdot \Psi_{nm}}\right] \qquad (2.15)$$

όπου το κλάσμα της επιφάνειας της ομάδας, θ_m, δίνεται από την ακόλουθη εξίσωση:

$$\theta_m = \frac{Q_m \cdot X_m}{\sum_n Q_n \cdot X_n} \qquad (2.16)$$

και το γραμμομοριακό κλάσμα της ομάδας *m*, X_m, από τη σχέση:

$$X_m = \frac{\sum_j \nu_m^{(j)} \cdot x_j}{\sum_j \sum_n \nu_n^{(j)} \cdot x_j} \qquad (2.17)$$

Τέλος, η παράμετρος Ψ_{mn}δίνεται από:

$$\Psi_{mn} = e^{-\frac{\alpha_{mn}}{T}} \quad (2.18)$$

όπου a_{mn}είναι η παράμετρος αλληλεπίδρασης ανάμεσα στις βασικές ομάδες *m* και *n*. Οι παράμετροι αλληλεπίδρασης προκύπτουν από προσαρμογή σε πειραματικά δεδομένα ισορροπίας φάσεων (Fredenslund et al., 1975).

2.3.2 Τροποποιήσεις του αρχικού μοντέλου UNIFAC

Μετά την ανάπτυξη της αρχικής UNIFAC το 1975, έχουν προταθεί στη βιβλιογραφία αρκετές τροποποιήσεις του αρχικού μοντέλου με σκοπό τη βελτίωση του στην πρόρρηση της ισορροπίας φάσεων ατμού-υγρού, υγρού-υγρού, στερεού-υγρού, συντελεστών ενεργότητας άπειρης αραίωσης (γ^{∞}) και ενθαλπιών ανάμειξης. Στη συνέχεια αναφέρονται τα πιο επιτυχημένα μοντέλα τύπου UNIFAC στην πρόρρηση της Ισορροπίας Φάσης Ατμού-Υγρού (Vapor-Liquid Equilibrium – VLE).

Οι Larsen et al. (1987) παρουσίασαν επέκταση του μοντέλου UNIFAC, όπου έκαναν τροποποιήσεις στο συνδυαστικό μέλος, καθώς και στους συντελεστές αλληλεπίδρασης ώστε να έχουν εξάρτηση από τη θερμοκρασία. Η νέα μέθοδος ονομάστηκε UNIFAC [Lyngby] και παρατηρήθηκε πως έχει καλύτερα αποτελέσματα από την αρχική UNIFAC. Για τον υπολογισμό των παραμέτρων αλληλεπίδρασης α_{mn} χρησιμοποιήθηκε μια βάση πειραματικών δεδομένων που περιλάμβανε δεδομένα VLE και ενθαλπίες ανάμειξης. Τα συνδυαστικό μέλος της UNIFAC [Lyngby] δίδεται από:

$$\ln \gamma_i^{comb} = \ln\left(\frac{\varphi_i'}{x_i}\right) + 1 - \left(\frac{\varphi_i'}{x_i}\right) \qquad (2.19)$$

με

$$\varphi_i' = \frac{x_i \cdot r_i^{\frac{2}{3}}}{\sum_j x_j \cdot r_j^{\frac{2}{3}}} \quad (2.20)$$

Επίσης, για τη βελτίωση της θερμοκρασιακής εξάρτησης των συντελεστών ενεργότητας, η παράμετρος Ψ_{mn} στο υπολειμματικό μέρος υπολογίζεται από τη σχέση:

$$\Psi_{mn} = exp\left(-\frac{a_{mn}+b_{mn}\cdot(T-T_o)+c_{mn}\cdot\left(T\cdot\ln\frac{T_o}{T}+T-T_o\right)}{T}\right) \quad (2.21)$$

όπου T_o είναι μια θερμοκρασία αναφοράς ίση με 298.15 K.

Οι Hansen et al. (1991) επέκτειναν το αρχικό μοντέλο UNIFAC και για τον υπολογισμό των παραμέτρων αλληλεπίδρασης α_{mn}, χρησιμοποιήθηκε μια εκτενής βάση πειραματικών δεδομένων VLE, η Dortmund DataBank.

Οι Gmehling, Li και Schiller (1993) παρουσιάζουν τη δική τους επέκταση στο μοντέλο, τη UNIFAC [Dortmund] όπου έχουν κάνει αλλαγές και στο συνδυαστικό και το υπολειμματικό μέρος σε σχέση με την αρχική σχέση και πρότειναν συνδυαστικό μέρος αυτού του μοντέλου να δίνεται από την παρακάτω σχέση:

$$\ln\gamma_i{}^{comb} = \ln\frac{\varphi_i{}'}{x_i} + 1 - \frac{\varphi_i{}'}{x_i} - \frac{z}{2}\cdot q_i\cdot\left(\ln\frac{\varphi_i}{\theta_i} + 1 - \frac{\varphi_i}{\theta_i}\right) \quad (2.22)$$

με

$$\varphi_i{}' = \frac{x_i\cdot r_i^{\frac{3}{4}}}{\sum_j x_j\cdot r_j^{\frac{3}{4}}} \quad (2.23)$$

και

$$\Psi_{mn} = exp\left(-\frac{a_{mn}+b_{mn}\cdot T+c_{mn}\cdot T^2}{T}\right) \quad (2.24)$$

όπου τα θ και φ υπολογίζονται με τους τύπους της αρχικής μορφής της UNIFAC.

Πρέπει να σημειωθεί ότι το μοντέλο UNIFAC [Dortmund], σε αντίθεση με τα μοντέλα των Hansen et al. και UNIFAC [Lyngby] (Larsen et al., 1987), δεν χρησιμοποιεί τις τιμές των παραμέτρων όγκου και επιφάνειας του van der Waals για τις ομάδες όπως αυτές υπολογίζονται από τη μέθοδο του Bondi (1968). Οι παράμετροι αυτές για το μοντέλο UNIFAC [Dortmund] (Gmehling et al. 1993) έχουν προέλθει από ταυτόχρονη προσαρμογή μαζί με τις παραμέτρους αλληλεπίδρασης του υπολειμματικού μέρους σε πειραματικά δεδομένα ισορροπίας φάσεων υγρού-υγρού, στερεού-υγρού, ατμού-υγρού, γ^{∞} και ενθαλπιών ανάμιξης.

Παρότι είναι αρκετά δύσκολη η εξαγωγή ενός γενικού κανόνα για το ποιο από τα μοντέλα UNIFAC δίνει τα καλύτερα αποτελέσματα σε κάθε περίπτωση, προκύπτει ότι τα τρία μοντέλα δίνουν εξίσου καλά αποτελέσματα πρόρρησης VLE. Οι Lohnman και Gmehling (2001) έχουν εκπονήσει εκτενή αξιολόγηση των μοντέλων UNIFAC, όπου φαίνεται το μοντέλο UNIFAC [Dortmund] να έχει ένα ελαφρύ προβάδισμα έναντι των άλλων. Το μοντέλο UNIFAC [Dortmund] παρέχει καλύτερες περιγραφές της σχέσης των παραμέτρων από τη θερμοκρασία και της συμπεριφοράς της διαλυμένης ουσίας στο μίγμα. Αυτά καθιστούν τη μέθοδο εφαρμόσιμη με μεγαλύτερη ακρίβεια σε συστήματα με μόρια που είναι πολύ διαφορετικά ως προς το μέγεθός τους.

Το πλεονέκτημα της UNIFAC είναι ότι μπορούν να αναπαρασταθούν και να υπολογιστούν συντελεστές ενεργότητας σε ικανοποιητικό βαθμό για δυαδικά και πολυσυστατικά μίγματα ενώσεων. Έτσι επιτρέπεται , σε συνδυασμό με τη γ-φ μεθοδολογία, πρόρρηση VLE σε χαμηλές πιέσεις, στα πλαίσια προκαταρτικού σχεδιασμού. Η εναλλακτική λύση ελλείψει πειραματικών δεδομένων θα ήταν ο νόμος ιδανικών μιγμάτων του Rault, που όμως δίνει μεγάλες αποκλίσεις από τα πειραματικά δεδομένα(Tassios,2001). Τέτοιες αποκλίσεις θα είχαν σημαντική επίδραση στην περίπτωση δύσκολων διαχωρισμών, όπως ο σχεδιασμός μιας αποστακτικής στήλης.

Τα μειονεκτήματα της μεθόδου UNIFACμπορούν να συνοψιστούν κυρίως σε αυτά που προαναφέρθηκαν στην Ενότητα 2.1 για τις μεθόδους GC καθαρών ουσιών. Επίσης, η μεθοδολογία *γ-φ*, όταν το *φ* υπολογίζεται από την καταστατική εξίσωση Virial αποκομμένη στο δεύτερο όρο, περιορίζει την εφαρμογή της UNIFAC σε πιέσεις κάτω από περίπου 10-15 atm και δε συμπεριλαμβάνει μη συμπυκνώσιμα ρευστά π.χ. CO_2, O_2, CH_4. Ακόμα, το βέλτιστο θερμοκρασιακό εύρος εφαρμογής της μεθόδου είναι 280-393 K, δηλαδή πολύ χαμηλές θερμοκρασίες και δεν περιλαμβάνει προβλέψεις για μίγματα που περιέχουν ηλεκτρολύτες. Τέλος, δεν παρέχει ικανοποιητικά αποτελέσματα για μίγματα υδρογονανθράκων και νερού. Ακόμα, σε αρκετές περιπτώσεις δεν υπάρχουν διαθέσιμα δεδομένα παραμέτρων αλληλεπίδρασης. Τα δεδομένα, όμως, συνεχώς ανανεώνονται.

2.3.3 Χρήσεις του μοντέλου UNIFAC στη μοριακή βελτιστοποίηση

Η μέθοδος UNIFAC είναι ένα από τα βασικότερα μοντέλα GC που χρησιμοποιούνται στη μοριακή βελτιστοποίηση διαλυτών καθώς επιτρέπει την πρόρρηση του συντελεστή ενεργότητας σε άπειρη αραίωση, γ_{inf}. Το γ_{inf} απαιτείται για τον υπολογισμό απαραίτητων ιδιοτήτων για ένα διαλύτη, όπως η διαλυτική ισχύς (solvent power),η διαλυτότητα (solubility), ο συντελεστής κατανομής (distribution coefficient), η σχετική πτητικότητα (relative volatility) κ.ά.

Είναι σημαντικό να αναφερθεί ότι η ακρίβεια της μεθόδου UNIFAC, καθώς και των υπολοίπων μοντέλων GC, καθίσταται λιγότερη σημαντική στα πλαίσια μιας διαδικασίας βελτιστοποίησης έναντι της προσομοίωσης. Ο λόγος είναι ότι στη βελτιστοποίηση το ζητούμενο δεν είναι η ακρίβεια στον υπολογισμό της ιδιότητας μιας ουσίας, αλλά περισσότερο το να διατηρείται η σχετική τιμή της ιδιότητας για μια ομάδα ουσιών. Δηλαδή, δεν είναι τόσο σημαντική η τιμή π.χ. της διαλυτότητας της ουσίας Α, όσο το ότι η διαλυτότητα της ουσίας Α είναι μεγαλύτερη από τη διαλυτότητα της ουσίας Β.

Κεφάλαιο 3. Εισαγωγή στην Ανάλυση Κύκλου Ζωής

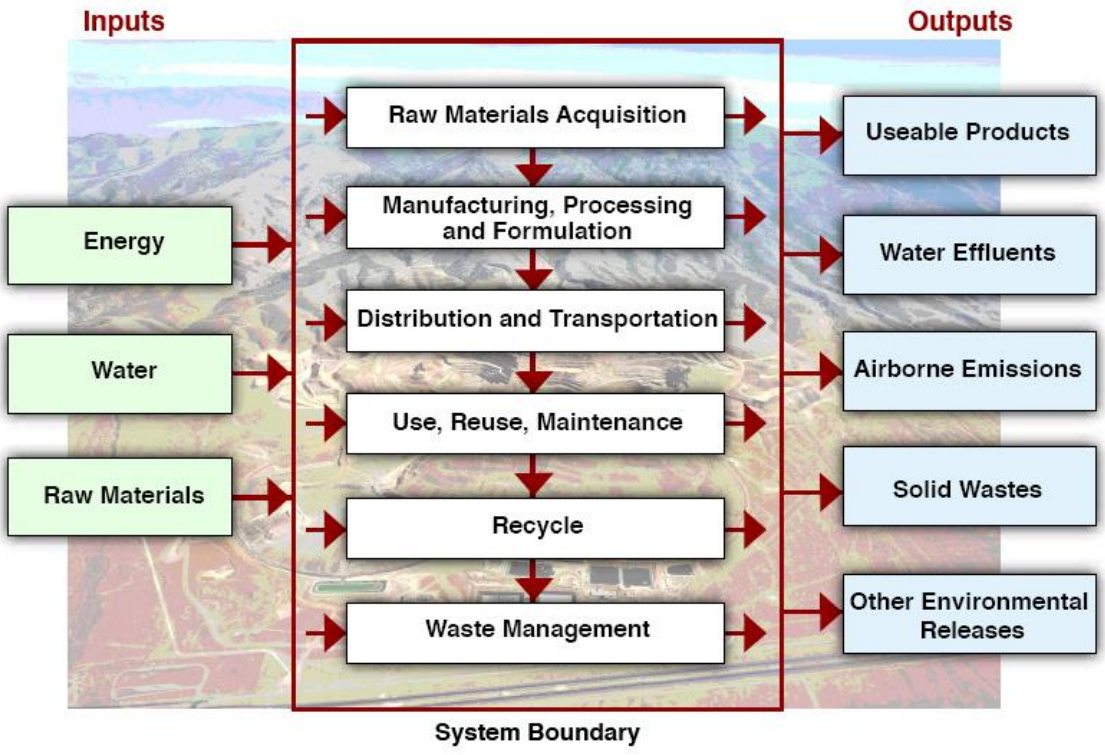
Όσο η περιβαλλοντική συνείδηση αυξάνεται, κάτι που εκφράζεται από την συνεχώς αυστηρότερη πολιτική και νομοθεσίες, τόσο οι επιχειρήσεις και βιομηχανίες τείνουν να αξιολογούν πιο έντονα την επίδραση των δραστηριοτήτων τους στο περιβάλλον. Οι κοινωνίες αποκτούν επίγνωση των ζητημάτων κατασπατάλησης φυσικών πόρων και της περιβαλλοντικής υποβάθμισης. Πολλές βιομηχανίες ανταποκρίνονται στις ανησυχίες αυτές με «πράσινα» προϊόντα και με χρήση «πράσινων» διεργασιών. Η περιβαλλοντική απόδοση των προϊόντων και διεργασιών έχει γίνει βασικό θέμα, που έχει σαν αποτέλεσμα πολλές βιομηχανίες να επιχειρούν να ελαχιστοποιήσουν τις επιπτώσεις τους στο περιβάλλον. Πολλές εταιρίες θεώρησαν χρήσιμη πρακτική να προχωρήσουν πέρα από τις στρατηγικές μείωσης της ρύπανσης και τη διαχείριση των περιβαλλοντικών συστημάτων. Ένα εργαλείο για αυτήν την αντιμετώπιση είναι η Ανάλυση Κύκλου Ζωής (Life Cycle Assessment, στο εξής LCA). Η Ανάλυση Κύκλου Ζωής είναι ένα από τα πιο σύγχρονα και που πιο γνωστά εργαλεία στην αντιμετώπιση ενός από τα πιο γνωστά προβλήματα στην παγκόσμια περιβαλλοντική πολιτική και ανάπτυξη: την αξιολόγηση περιβαλλοντικών επιπτώσεων προϊόντων και χημικών. Να σημειωθεί πως δεν αποτελεί ένα οικονομικό μέτρο για το αν συμφέρει να παραχθεί ένα προϊόν, αλλά περισσότερο για να εντοπιστούν αδύναμα σημεία στον κύκλο ζωής του, όπου υπάρχει μεγάλη περιβαλλοντική επιβάρυνση, αλλά και τρόποι βελτίωσης των υπαρχουσών τεχνολογιών και διεργασιών. Το σχήμα 3.1 δείχνει τα βήματα που εκτελούνται κατά την ανάλυση του κύκλου ζωής ενός προϊόντος.

3.1 Φάσεις εκτέλεσης του LCA

Πιο συγκεκριμένα, η Ανάλυση Κύκλου Ζωής είμαι μια τεχνική με την οποία αξιολογούνται τα περιβαλλοντικά και οικολογικά αποτυπώματα που είναι συνδεδεμένα με ένα προϊόν, μια διεργασία ή μια υπηρεσία. Αυτό γίνεται, καταρχάς συντάσσοντας ένα σύνολο από σχετικές ενεργειακές ροές, εισροές υλικών και περιβαλλοντικών εκπομπών. Δευτερευόντως, αξιολογώντας τα πιθανά περιβαλλοντικά αποτυπώματα που συνδέονται με συγκεκριμένες εισροές (inputs) και εκπομπές (releases) και τέλος, την ερμηνεία των αποτελεσμάτων για να βοηθήσει στη λήψη πιο εμπεριστατωμένων και έγκυρων αποφάσεων (Life Cycle Assessment:Principle and Practice, 2006).

Η διαδικασία εκτέλεσης του LCA είναι συστηματική και κατηγοριοποιημένη σε τέσσερις φάσεις: ορισμός του σκοπού της ανάλυσης (goal definition), ανάλυση του καταλόγου (εισροών και εκροών) (inventory analysis), αξιολόγηση του αποτυπώματος (impact assessment) και ερμηνεία (interpretation). Ο ορισμός του σκοπού περιλαμβάνει τον ορισμό και την περιγραφή του προϊόντος, της διεργασίας ή της

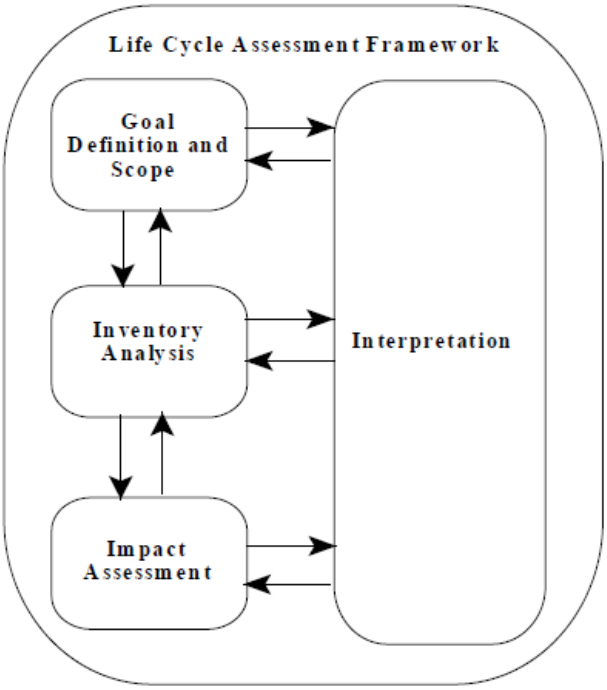
γενικότερης δραστηριότητας. Καθιερώνεται, ακόμα, ένα πλαίσιο εργασίας, μέσα στο οποίο θα γίνει η ανάλυση LCA και αναγνωρίζονται τα όρια του συστήματος και οι περιβαλλοντικές επιπτώσεις που πρέπει να μελετηθούν. Στην ανάλυση του καταλόγου-απογραφής ροών (inventory analysis ή αλλιώς, Life Cycle Inventory, LCI) περιλαμβάνονται την ποιοτική και ποσοτική αναγνώριση της ενέργειας, του νερού και των πρώτων υλών που χρησιμοποιούνται και των περιβαλλοντικών εκπομπών (π.χ. αέριες εκπομπές, απόθεση στερεών απορριμμάτων, απόρριψη υδατικών αποβλήτων). Η αξιολόγηση του αποτυπώματος περιλαμβάνει την εκτίμηση των πιθανών συνεπειών της χρήσης, αυτής, του νερού, της ενέργειας και των υλικών (που έχουν αναγνωρισθεί στο κατάλογο-απογραφή) στην ανθρώπινη υγεία και στο οικολογικό σύστημα. Τέλος, στο κομμάτι της ερμηνείας των αποτελεσμάτων, αξιολογούνται τα ευρήματα από την ανάλυση της απογραφής και του καταλόγου (2ο βήμα) και της αξιολόγησης του αποτυπώματος (3ο βήμα), για να υποστηριχθεί η επιλογή ενός συγκεκριμένου προϊόντος, διεργασίας ή δραστηριότητας με ξεκάθαρη κατανόηση της αβεβαιότητας και των παραδοχών που έχουν γίνει για να παραχθούν τα αποτελέσματα.



Σχήμα 3.1. Γενικό πλαίσιο εκτέλεσης μια ανάλυσης LCA

Είναι σημαντικό να τονιστεί ο κύκλος ζωής δεν ξεκινάει από την παραγωγή ενός προϊόντος, παρά από την ίδια την εξόρυξη και την παραγωγή των πρώτων υλών, του νερού και της ενέργειας που θα

συμμετάσχουν στην παραγωγή του (cradle-to-grave). Συνεπώς, και η ανάλυση που συνδέεται με ένα προϊόν ξεκινάει ήδη από τις πρώτες ύλες. Αυτό το χαρακτηριστικό του LCA (η ενσωμάτωση όλων των σταδίων ξεκινώντας ήδη πριν από την παρασκευή του προϊόντος) είναι που κάνει μοναδική αυτή τη μεθοδολογία. Πρέπει, επίσης, να επισημανθεί πως ο Κύκλος Ζωής ενός προϊόντος διαφέρει και από χώρα σε χώρα. Αυτό θα γίνει κατανοητό και στη συνέχεια, όπου θα αναφερθούν οι δείκτες του LCA. Αυτό συμβαίνει, διότι, το μονοπάτι (ακολουθία διαδικασιών στην παραγωγή και διανομή) που ακολουθείται στον Κύκλο Ζωής ενός προϊόντος είναι διαφορετικό από χώρα σε χώρα και κατά συνέπεια, η περιβαλλοντική επιβάρυνση διαφορετική. Ως ένα χαρακτηριστικό παράδειγμα για να γίνει πιο κατανοητό το παραπάνω, είναι το εξής: όπως αναφέρθηκε, στο LCA καταγράφονται και οι ενεργειακές ροές (μαζί με τις ροές υλικών). Στην ανάλυση του Κύκλου Ζωής, οι ενεργειακές ροές για ένα προϊόν ενσωματώνουν αρκετή πληροφορία και εξαρτώνται άμεσα από το ενεργειακό μίγμα (η κατανομή της χρήσης των διαφόρων πηγών ενέργειας) της εκάστοτε χώρας όπου γίνεται το LCA. Όμως, αφού το ενεργειακό μίγμα από χώρα σε χώρα αλλάζει, έτσι θα αλλάζει και η ενεργειακή απόδοση και κατανάλωση ενός προϊόντος και συνεπώς, και η επιβάρυνση στο περιβάλλον. Πρέπει να γίνει κατανοητό, πως οι διεργασίες παραγωγής μπορούν να παραμένουν ίδιες παντού για το ίδιο προϊόν. Αυτό που αλλάζει από χώρα σε χώρα είναι η αποτελεσματική χρήση της ενέργειας (σε κάποιες χώρες η ενέργεια χρησιμοποιείται πιο αποτελεσματικά), κάτι που ενσωματώνεται και στον Κύκλο Ζωής και συνεπώς, το ίδιο προϊόν μπορεί να φαίνεται πιο «πράσινο» σε μια χώρα από κάποια άλλη.



Σχήμα 3.2. Φάσεις εκτέλεσης LCA

3.2 Τα οφέλη διεξαγωγής LCA

Όπως προαναφέρθηκε, το LCA είναι ένα χρήσιμο εργαλείο για την υποστήριξη αποφάσεων επιλογής προϊόντων και διεργασιών με το ελάχιστο περιβαλλοντικό και οικολογικό αποτύπωμα. Μπορεί κάλλιστα, να χρησιμοποιηθεί σε συνδυασμό με άλλα κριτήρια (π.χ. οικονομικά και απόδοσης) με σκοπό για πιο ολοκληρωμένες επιλογές και αποφάσεις. Τα δεδομένα LCA βοηθούν στην ταυτοποίηση της μεταφοράς των περιβαλλοντικών επιπτώσεων από ένα μέσο σε ένα άλλο (π.χ. εξάλειψη αέριων ρύπων δημιουργώντας ένα υδατικό ρεύμα που περιλαμβάνει τους ρύπους αυτούς από μια μονάδα επεξεργασίας τους). Αν δε διεξαγόταν LCA, δε θα ήταν δυνατόν, να αναγνωριστεί αυτή η μεταφορά και να συμπεριληφθεί στην ανάλυση γιατί είναι έξω από την τυπική μελέτη της επιλογής διεργασίας. Γίνεται κατανοητό πως, το LCA βοηθά τους μελετητές να αποφασίσουν υπό το πλαίσιο ολόκληρου του συστήματος ενός προϊόντος και όχι μόνο, αν επικεντρωθούν σε μια διεργασία μόνο. Παραδείγματος χάριν, στη διαδικασία επιλογής μεταξύ δυο ανταγωνιστικών προϊόντων, μπορεί να φαίνεται πως η μια επιλογή μπορεί να είναι περιβαλλοντικά φιλικότερη, επειδή παράγει λιγότερα στερεά απόβλητα από τη δεύτερη επιλογή. Παρολαυτά, μετά τη διεξαγωγή LCA μπορεί να αποδειχθεί πως η πρώτη επιλογή είναι περισσότερο περιβαλλοντικά δαπανηρή από τη δεύτερη όταν η επιβάρυνση της μετράται και στους τρεις παράγοντες ταυτόχρονα (γη, αέρας, υδροφόρος ορίζοντας). Αυτό βοηθά στην απόφαση για το δεύτερο προϊόν, γιατί συνολικά έχει μικρότερο περιβαλλοντικό αποτύπωμα.

Η ικανότητα να εντοπίζονται μετατοπίσεις στο περιβαλλοντικό αποτύπωμα μπορεί να βοηθήσει τους ερευνητές να εντοπίσουν και να χαρακτηρίσουν όλες οι πιθανές αντισταθμίσεις (trade-off) που συνδέονται με τις διάφορες εναλλακτικές ενός προϊόντος ή μιας διεργασίας. Διεξάγοντας ένα LCA, μπορεί να:

- Αναπτυχθεί μια συστηματική διαδικασία για την αξιολόγηση των περιβαλλοντικών επιπτώσεων που συνδέονται άμεσα με ένα προϊόν.
- Αναλυθούν οι περιβαλλοντικές αντισταθμίσεις που συνδέονται με ένα ή περισσότερα προϊόντα/διεργασίες
- Ποσοτικοποιηθούν περιβαλλοντικές εκπομπές στον αέρα, τον υδροφόρο ορίζοντα και τη γη αναφορικά στα στάδια του κύκλου ζωής και/ή διεργασιών με μεγάλη συνεισφορά στην περιβαλλοντική συμπεριφορά.
- Υποβοηθηθεί η αναγνώριση σημαντικών μετατοπίσεων στα περιβαλλοντικά αποτυπώματα μεταξύ των σταδίων του κύκλου ζωής και των περιβαλλοντικών μέσων (γη, αέρας, νερό)
- Αξιολογηθούν οι επιδράσεις στην ανθρώπινη υγεία και οικολογία από την κατανάλωση πρώτων υλών και περιβαλλοντικών εκπομπών στην τοπική κοινότητα, περιοχή και γενικά, στον κόσμο.
- Συγκριθούν οι επιπτώσεις στην υγεία και την οικολογία μεταξύ δυο ή περισσοτέρων προϊόντων/διεργασιών.

- Αναγνωρισθούν επιπτώσεις σε μια ή περισσότερες συγκεκριμένες περιβαλλοντικές περιοχές ενδιαφέροντος.

3.3 Περιορισμοί στην διεξαγωγή του LCA

Η διεξαγωγή LCA μπορεί να είναι οικονομικά και χρονικά δαπανηρή. Η συλλογή των δεδομένων μπορεί να είναι ένα αρκετά μεγάλο πρόβλημα, κάτι που εξαρτάται και από το πόσο σε βάθος είναι επιθυμητό από το χρήστη, να είναι η ανάλυση. Το κατά πόσο διαθέσιμα είναι τα δεδομένα είναι βασικός παράγοντας της εγκυρότητας των τελικών αποτελεσμάτων. Συνεπώς, είναι απαραίτητο να σταθμιστεί η διαθεσιμότητα των δεδομένων, ο χρόνος διεξαγωγής της ανάλυσης και των οικονομικών πόρων που απαιτούνται σε σχέση με τα οφέλη του LCA. Το LCA δε θα καθορίσει πιο προϊόν είναι πιο αποδοτικό ή πιο οικονομικό. Κατά συνέπεια, η Ανάλυση Κύκλου Ζωής που θα διεξαχθεί θα πρέπει να είναι μέρος μιας ευρύτερης μελέτης η οποία αξιολογεί και άλλους παράγοντες, όπως το κόστος και την απόδοση.

3.4 Ανασκόπηση βιβλιογραφίας

Όπως έχει γίνει αντιληπτό, μεγάλο μέρος της αξιοπιστίας εκτέλεσης μεθοδολογίας LCA βασίζεται στην ύπαρξη αξιόπιστων πειραματικών δεδομένων. Μέχρι στιγμής, η δημιουργία της βάσης δεδομένων Ecoinvent αποτελεί μια σημαντική κατάκτηση για τη διευκόλυνση των υπολογισμών (Frischknecht et al., 2005). Η κατασκευή της Ecoinvent ξεκίνησε το 2000 στην Ελβετία και έκτοτε ανανεώνεται συνεχώς. Για την ανάπτυξη της έλαβαν μέρος αρκετές ελβετικές ομοσπονδιακές υπηρεσίες και εργαστήρια του Ομοσπονδιακού Τεχνικού Ινστιτούτου (ETHZ). Να σημειωθεί πως η Ecoinvent περιλαμβάνει πληροφορίες για πάνω από 2500 προϊόντα και διεργασίες, καθώς και διάφορες μεθοδολογίες που έχουν αναπτυχθεί με τα χρόνια. Επίσης, δόθηκε σημασία στο ζήτημα συγκέντρωσης δεδομένων από διάφορες χώρες εντός και εκτός Ευρώπης. Όσον αφορά στις διεργασίες έχουν γίνει κατάλληλες παραδοχές και έχουν συγκεντρωθεί μέσες τιμές από τις τεχνολογίες που χρησιμοποιούνται και από αυτές που πρόκειται να χρησιμοποιηθούν στο μέλλον. Ένα σημαντικό κομμάτι της πληροφορίας ενσωματώνεται στα εκτενή δεδομένα εκπομπών ρύπων από διάφορες πηγές για πληθώρα χρονικών οριζόντων, αλλά και στα δεδομένα διαχείρισης αποβλήτων. Ένα κομμάτι της Ecoinvent χρησιμοποιείται και στα πλαίσια της παρούσας εργασίας για την ανάπτυξη των μοντέλων.

Όσον αφορά στην πρόβλεψη των δεικτών LCA, οι Wegnet et al (2009) πρότειναν το μοντέλο πρόβλεψης Finechem tool το οποίο θα παρέχει υπολογισμούς των δεικτών LCI χρησιμοποιώντας μόνο πληροφορίες που αφορούν τη δομή του μορίου, ανεξάρτητα από τις διαδικασίες παραγωγής. Να σημειωθεί

πως είναι περιορισμένος ο αριθμός μοντέλων πρόβλεψης LCA δεικτών, τα οποία έχουν αναπτυχθεί κυρίως από εταιρίες και προορίζονται για εσωτερική χρήση χωρίς να είναι διαθέσιμα στο ευρύ κοινό. Επίσης, η έλλειψη πειραματικών δεδομένων και το κόστος χρόνου και κεφαλαίου που απαιτούνται, αποθαρρύνει την προσπάθεια ανάπτυξης μεθόδων. Το μοντέλο χρησιμοποιεί την προσέγγιση των νευρωνικών δικτύων για την εξαγωγή αποτελεσμάτων και ως είσοδο απαιτεί δομικές ποσοτικές πληροφορίες, όπως το μοριακό βάρος, αριθμό ατόμων άνθρακα, αζώτου και αλογόνων, αρωματικούς και αλειφατικούς δακτυλίους και άλλα. (σύνδεσμος για το Finechem tool: http://www.sust-chem.ethz.ch/tools/finechem)

3.5 CAMD και LCA

Όπως γίνεται αντιληπτό, το LCA είναι το πιο κατάλληλο εργαλείο για να συνδυαστεί με μεθοδολογίες CAMD για να ξεχωρίσει μεταξύ των διάφορων επιλογών χημικών μέσων (και προϊόντων γενικότερα) και να περιορίσει την αναζήτηση σε λύσεις περιβαλλοντικά φιλικές και συνεπώς, σε «πράσινα» χημικά μέσα (διαλύτες, ψυκτικά κ.τ.λ.). Αυτό γίνεται με τη βοήθεια δεικτών που περιλαμβάνονται στο LCA και που εκφράζουν , από διάφορες σκοπιές, την περιβαλλοντική απόδοση μιας ουσίας. Με τη σωστή διαμόρφωση του προβλήματος βελτιστοποίησης (στόχων, objective και περιορισμών, constraints) και την ενσωμάτωση αυτών των δεικτών στην αλγοριθμική διαδικασία μπορεί το πρόβλημα που αντιμετωπίζει το CAMD να αναχθεί στο αντίστοιχο πρόβλημα αναζήτησης «πράσινων» προϊόντων/μέσων. Με τη κατάλληλη χρήση, δηλαδή, αυτών των μέτρων (στους περιορισμούς, για παράδειγμα) μπορούν να σχεδιαστούν διαλύτες, που η επιβάρυνση τους να μην ξεπερνά μια συγκεκριμένη τιμή ενός δείκτη. Η αλγοριθμική διαδικασία διεξάγεται κανονικά, μόνο που πρέπει να συμπεριληφθούν και υπολογισμοί LCA, με τον ίδιο τρόπο, όπως και στον υπολογισμό των υπόλοιπων ιδιοτήτων.

Όπως έχει, όμως, ήδη αναφερθεί, η συστηματικότητα της μεθόδου εξασφαλίζεται σε μεγάλο βαθμό από τη συστηματικότητα των μεθόδων πρόβλεψης. Αυτό αφορά σε απλά, στη χρήση και στον υπολογισμό μοντέλα, τα οποία, παρολαυτά, διατηρούν την αξιοπιστία τους. Σε αυτό το σημείο, όμως, το εγχείρημα, σύζευξης των δυο εργαλείων LCA και CAMD φαίνεται αδύνατο, μιας και υπάρχει μεγάλη έλλειψη αξιόπιστων υπολογιστικών εργαλείων, που να διατηρούν και αυστηρά επίπεδα, αξιοπιστίας. Είναι, λοιπόν, επιτακτική η ανάγκη ανάπτυξης τέτοιων αξιόπιστων εργαλείων.

3.6 Μια σύντομη ιστορική αναδρομή στο LCA

Η Ανάλυση Κύκλου Ζωής έκανε το ξεκίνημα της το 1960. Είχε ήδη αρχίσει να υπάρχει ανησυχία για την κατασπατάληση και εύρεση νέων φυσικών πόρων. Στο Παγκόσμιο Συνέδριο Ενέργειας (World Energy Conference) του 1963, σε μια από τις πρώτες δημοσιεύσεις του είδους του, ο Harold Smith, general project

manager της Douglas Point Nuclear Generating Station στον Canada, ανέφερε τον υπολογισμό της συνολικής ενεργειακής απαίτησης. Αργότερα, το 1960, οι Meadows et al. (1972) και Goldsmith et al. (1972) δημοσίευσαν μοντέλα, που προβλέπουν τις επιπτώσεις του συνεχώς αυξανόμενου παγκόσμιου πληθυσμού στη ζήτηση των περιορισμένων φυσικών πόρων και ενεργειακών πηγών. Η προβλέψεις για ραγδαία κατασπατάληση των ορυκτών καυσίμων και κλιματικές αλλαγές αποτέλεσε βάση για περισσότερες μελέτες χρήσης ενέργειας στη βιομηχανία. Οι μελέτες, αυτήν την περίοδο, επικεντρώνονταν στον υπολογισμό κόστους και περιβαλλοντικών επιπτώσεων χρήσης εναλλακτικών πηγών ενέργειας.

Το 1969, ξεκίνησε για λογαριασμό της Coca-Cola μια εσωτερική έρευνα, η οποία έθεσε τα θεμέλια για την καθιέρωση των σύγχρονων μεθόδων LCA για τις ΗΠΑ. Συγκεκριμένα, η έρευνα περιλάμβανε τον καθορισμό κουτιού για το αναψυκτικό το οποίο είχε τη μικρότερη περιβαλλοντική επιβάρυνση και συνείσφερε στο ελάχιστο στην σπατάλη των φυσικών πόρων. Η έρευνα κατάφερε να ποσοτικοποιήσει τη ποσό των πρώτων υλών και καυσίμων που ενσωματώνονταν στο προϊόν και τα περιβαλλοντικά φορτία από τις παραγωγικές διαδικασίες για κάθε κουτί. Άλλες εταιρίες στις ΗΠΑ και στην Ευρώπη εφάρμοσαν την ίδια τεχνική στην αρχή της δεκαετίας του ’70. Να σημειωθεί πως, εκείνη την εποχή, πολλές από τις πληροφορίες που απαιτούνταν για την ανάλυση, προέρχονταν από πηγές ανοικτές για όλο το κοινό, όπως κυβερνητικά έγγραφα ή τεχνικές αναφορές, καθώς βιομηχανικά δεδομένα δεν ήταν διαθέσιμα.

Η διαδικασία ποσοτικοποίησης φυσικών πόρων και περιβαλλοντικών εκπομπών προϊόντων έγινε γνωστή ως Ανάλυση Πόρων και Περιβαλλοντικού Προφίλ (Resource and Environmental Profile Analysis, REPA), στις ΗΠΑ. Στην Ευρώπη ονομάστηκε Ecobalance. Όσο σχηματίζονταν ομάδες κοινού ενδιαφέροντος, οι οποίες αύξησαν τη ζήτηση για ακριβή δεδομένα από τη βιομηχανία και με την έλλειψη πετρελαίου στις αρχές του ’70, διεξήχθησαν περίπου 15 REPA, μεταξύ 1970 και 1975. Μέσα σε αυτήν την περίοδο σχηματίστηκε ένα πρωτόκολλο ή μεθοδολογία έρευνας για τέτοιου είδους μελέτες. Αυτή η μεθοδολογία απαιτούσε έναν αριθμό παραδοχών, οι οποίες πέρασαν από διεξοδικό έλεγχο, μεταξύ άλλων και από εκπροσώπους μεγάλων βιομηχανιών, με σκοπό να σχηματιστεί μια έγκυρη και αξιόπιστη μεθοδολογία.

Από το 1975 έως τις αρχές του ’80, καθώς έγινε η ανάκαμψη της οικονομίας του πετρελαίου και υπονομεύτηκε η χρήση των παραπάνω μελετών, το ενδιαφέρον μετακινήθηκε στα θέματα διαχείρισης βλαβερών οικιακών απορριμμάτων. Παρολαυτά, το LCA συνέχισε να διεξάγεται και να αναπτύσσεται με αργό ρυθμό (περίπου δυο μελέτες το χρόνο). Σε αυτό το χρονικό διάστημα, το ενδιαφέρον στη Ευρώπη αυξήθηκε, λόγω της έναρξης λειτουργίας της Ευρωπαϊκής Διεύθυνσης Περιβάλλοντος (DG X1), που καθιερώθηκε από την Ευρωπαϊκή Επιτροπή. Οι ευρωπαίοι εκπρόσωποι του LCA ξεκίνησαν να εφαρμόζουν τις πρακτικές που εφάρμοζαν και στις ΗΠΑ. Η DG X1 καθιέρωση οδηγία για τις συσκευασίες υγρών τροφών (Liquid Food Container Directive), το 1985, όπου η κάθε βιομηχανία ήταν υποχρεωμένη να καταγράφει την κατανάλωση της ενέργειας και των πρώτων υλών των συσκευασιών και τα στερεά

απόβλητα που προέκυπταν από τη χρήση συσκευασιών. Όταν τα στερεά απορρίμματα ξαναέγιναν ένα παγκόσμιο θέμα το 1988, το LCA επανήλθε ως ένα εργαλείο για την ανάλυση περιβαλλοντικών προβλημάτων. Το LCA βρίσκεται πάλι στο επίκεντρο της ερευνητικής δραστηριότητας και γίνεται προσπάθεια βελτίωσης της μεθοδολογίας από τους επιστήμονες σε όλο τον κόσμο.

Το 1991, προέκυψαν καταγγελίες χρήσης της ανάλυσης LCA, από κατασκευαστές στις ΗΠΑ, για λόγους marketing. Σε αυτήν την περίπτωση, τα ένδικα μέσα των ΗΠΑ κατήγγειλαν τη χρήση του LCA για σκοπούς προώθησης, μέχρι να αναπτυχθούν ομοειδείς μέθοδοι για να γίνονται τέτοιες αναλύσεις. Αυτήν την περίοδο, μαζί και με τις πιέσεις από άλλες περιβαλλοντικές οργανώσεις, για την προτυποποίηση του LCA, οδήγησε στην καθιέρωση προτύπων για το LCA από το Διεθνή Οργανισμό Προτυποποίησης (ISO) και συγκεκριμένα, στη σειρά 14000 (1997-2002).

Το 2002, το Πρόγραμμα Περιβάλλοντος Ηνωμένων Εθνών (UNEP) συνεργάστηκε με την Κοινότητα Περιβαλλοντικής Τοξικολογίας και Χημείας και για μια διεθνή συνεργασία που ονομάστηκε Πρωτοβουλία Κύκλου Ζωής (Life Cycle Initiative). Σκοπός ήταν να μπει σε πρακτική εφαρμογή μεγαλύτερης κλίμακας το LCA και να βελτιωθούν τα εργαλεία του, μέσα από καλύτερα δεδομένα και δείκτες. Καθιερώθηκε, μάλιστα, και ο όρος Life Cycle Management,όπου έχει σκοπό την ευαισθητοποίηση και τη βελτίωση των ικανοτήτων αυτών που χρησιμοποιούν το LCA, σαν εργαλείο λήψης αποφάσεων μέσα από την καλύτερη εκπαίδευση και διανομή της πληροφορίας.

3.7 Δείκτες του LCA

Όπως αναφέρθηκε στην ενότητα 3.1 ο βασικός ρόλος του LCA βασίζεται πάνω στο ότι υπάρχει ένας επαρκής αριθμός δεικτών βάσει του οποίου μπορεί να γίνει η αξιολόγηση ενός προϊόντος ή μιας διεργασίας από διάφορα περιβαλλοντικά πρίσματα. Αυτό είναι λογικό, δεδομένου του πλήθους οικολογικών προβλημάτων που αντιμετωπίζεται και πως για να αποκτηθεί μια γενικότερη άποψη πάνω στην περιβαλλοντική απόδοση ενός προϊόντος ή μιας διαδικασίας, πρέπει να φαίνεται η επίδραση του σε κάθε κατηγορία προβλήματος (π.χ. φαινόμενο του θερμοκηπίου, φαινόμενο ευτροφισμού, τρύπα του όζοντος κ.τ.λ.).

3.7.1 Δυναμικό Παγκόσμιας Υπερθέρμανσης

Ο πρώτος βασικός δείκτης, που χρησιμοποιείται, κυρίως, για να χαρακτηρίσει ενώσεις που μπορούν να αποτελέσουν θερμοκηπικά αέρια είναι το Δυναμικό Παγκόσμιας Υπερθέρμανσης (Global Warming Potential-GWP). Η μονάδα, αυτή, χρησιμοποιείται για τη συγκριτική μελέτη μεταξύ δύο αερίων, ως προς το

πόση θερμότητα μπορεί να παγιδευτεί στην ατμόσφαιρα ανά μονάδα μάζας κάθε αερίου. Με άλλα λόγια, πόσο συνεισφέρει ο κύκλος ζωής μιας ουσίας στο φαινόμενο του θερμοκηπίου. Ο υπολογισμός του GWP βασίζεται στην ικανότητα απορρόφησης θερμότητας του εν λόγω αερίου ως προς την ικανότητα απορρόφησης θερμότητας του CO_2, όπως και στο ρυθμό αποδόμησης του αερίου με το χρόνο. Το GWP χρησιμοποιείται για να χαρακτηρίσει το πόσο συνεισφέρει ένα αέριο στο φαινόμενο του θερμοκηπίου, για ένα χρονικό ορίζοντα. Ο ορίζοντας αυτός είναι, συνήθως, 20, 100 και 500 χρόνια. Όπως είναι λογικό, η τιμή του GWP φθίνει μετά από χρόνια. Αυτό συμβαίνει, επειδή, η ουσία αποσυντίθεται μέσα από φυσικούς μηχανισμούς και μειώνεται η επίδραση της στο φαινόμενο του θερμοκηπίου. Το GWP μετριέται, στην παρούσα περίπτωση, με τον ορισμό που δίνεται από το Ecoinvent Report No.3 (Implementation of Life Cycle Assessment Methods) και με ισοδύναμα CO_2 ανά μονάδα μάζας της μελετώμενης ουσίας (Ecoinvent report No. 3) . Δηλαδή, το GWP μετριέται σε μάζα του CO_2 σε kg, που έχει το ίδιο GWP όσο και 1 kg της μελετώμενης ουσίας. Όταν μια ουσία έχει, για παράδειγμα GWP=10 kg CO_2-eq, σημαίνει πως 1kg της μελετώμενης ουσίας προκαλεί στο φαινόμενο του θερμοκηπίου την ίδια επιβάρυνση όσο και 10kg CO_2. Είναι κατανοητό, ο λόγος που χρησιμοποιείται το CO_2 σαν αέριο αναφοράς, μιας και είναι το πιο γνωστό και καλά μελετημένο θερμοκηπικό αέριο. Τυπικές τιμές για το δείκτη αυτό 1-10.

3.7.2 Σωρευτική Ενεργειακή Απαίτηση

Ο δεύτερος βασικό δείκτης του LCA είναι η Σωρευτική Ενεργειακή Απαίτηση (Cumulative Energy Demand-CED). Το CED για ένα προϊόν εκφράζει την άμεση και έμμεση ενεργειακή χρήση που ενσωματώνεται σε αυτό, δηλαδή όλη την ενέργεια που χρειάζεται για να διανύσει το προϊόν όλο τον κύκλο ζωής του. Συμπεριλαμβάνονται σε αυτό το δείκτη και η ενέργεια που χρησιμοποιείται για την εξόρυξη, παρασκευή και απόθεση των πρώτων υλών (Ecoinvent report No. 3). Αποτελεί έναν αρκετά παλιό και ανεπτυγμένο δείκτη για το LCA. Είναι αρκετά αξιόπιστος για την επιλογή μεταξύ συγκεκριμένων αποφάσεων και απαιτεί αρκετές παραδοχές, ως προς τη μέτρηση της απαιτούμενης ενέργειας. Για παράδειγμα, μπορεί κάποιος μελετητής να επιλέξει μετρώντας την Ανώτατη Θερμογόνο Δύναμη της πρωτογενούς ενέργειας ή την Κατώτατη Θερμογόνο Δύναμη. Το μέτρο αυτό φαίνεται αρκετά χρήσιμο στη σύγκριση των αποτελεσμάτων μιας μελέτης LCA, όπου χρησιμοποιούνται μόνο πρωτογενείς πηγές ενέργειας. Παρολαυτά, έχει αναπτυχθεί αρκετά, ώστε να χειρίζεται πυρηνική και υδροηλεκτρική ενέργεια. Το μειονέκτημα του είναι πως δεν έχει τυποποιηθεί ακόμα ο τρόπος μελέτης και μέτρησης του. Η έννοια και ο τρόπος μέτρησης του CED εκφράζονται πάλι από το Ecoinvent Report No.3, όπου το CED χωρίζεται σε 2 μεγάλες κατηγορίες ενεργειακών πηγών: τις ανανεώσιμες και τις μη ανανεώσιμες. Οι ανανεώσιμες πηγές ενέργειας χωρίζονται σε 5 υποκατηγορίες: βιομάζα, αιολική ενέργεια, υδροηλεκτρική ενέργεια, ηλιακή και γεωθερμική. Η μη ανανεώσιμες χωρίζονται σε 3 υποκατηγορίες: ορυκτά καύσιμα, πυρηνικά και πρωτογενή ξυλεία (primary forest). Η παραδοχή που γίνεται εδώ είναι πως όλοι οι ενεργειακοί φορείς έχουν μια εγγενή

τιμή. Αυτή η εγγενής τιμή καθορίζεται από την ενέργεια που παραλαμβάνεται από το περιβάλλον, η οποία πρέπει να είναι κοινή και για τις 8 υποκατηγορίες. Γι' αυτό και το CED εκφράζεται σε ισοδύναμα MJ, MJ-eq. Ο λόγος που εκφράζεται με αυτή τη μονάδα μέτρησης είναι ώστε να είναι κοινός ο ορισμός της ενέργειας για όλες τις πηγές. Έτσι, λοιπόν η τελική ενέργεια που χρησιμοποιείται για τον Κύκλο Ζωής ενός προϊόντος δεν αποτελείται από την πρωτογενή ενέργεια της ενεργειακής πηγής (της εκάστοτε υποκατηγορίας) μόνο, αλλά και π.χ. από την ενέργεια που έχει δαπανηθεί για τη μεταφορά των πρώτων υλών στο εργοστάσιο κ.τ.λ.. Για παράδειγμα, για κάθε MJ που χρησιμοποιείται για την παραγωγή ενός κιλού μιας ουσίας να δαπανώνται 3MJ πρωτογενούς ενέργειας και 2,5 MJ ενέργεια για μεταφορές. Τότε στην παραγωγή του αυτό το προϊόν έχει 5,5 MJ-eq. Τυπικές τιμές για το δείκτη αυτό είναι από 50-200 MJ-eq.

3.7.3 Οικολογικός Δείκτης 99

Το τρίτο και εξίσου σημαντικό μέτρο του LCA είναι ο Οικολογικός Δείκτης 99 (EcoIndicator 99-EI 99). Ο μέτρο αυτό δεν επικεντρώνεται μόνο σε μια σκοπιά επιβάρυνσης του περιβάλλοντος, παρά εκφράζει το συνολικό περιβαλλοντικό αποτύπωμα μιας ουσίας, συνολικά, σε όλο τον κύκλο ζωής της (Ecoindicator Manual, 2010). Συνεισφέρει σημαντικά στη μελέτη του Κύκλου Ζωής από την άποψη πως ξεπερνάει τα προβλήματα που υπάρχουν όσον αφορά στη δυσκολία συλλογής πληροφοριών και στη μεγάλη δαπάνη χρόνου για την εκπόνηση της μελέτης. Ακόμα, ενώ αρκετοί δείκτες φανερώνουν πληροφορίες για τα διάφορα επιμέρους φορτία (φαινόμενο του θερμοκηπίου, όξινη βροχή κ.τ.λ.), ο Οικολογικός Δείκτης 99, όχι μόνο συνενώνει τις διάφορες περιβαλλοντικές μελέτες σε ένα κατανοητό και φιλικό-προς-το-μελετητή αριθμό, αλλά σταθμίζει και τους διάφορους περιβαλλοντικούς κινδύνους, για να φανεί ποιος από όλους έχει πραγματικά μεγαλύτερη συνεισφορά στην ποιότητα του περιβάλλοντος. Η φιλοσοφία του ο Οικολογικός Δείκτης 99 είχε γίνει ήδη γνωστή από τον ο Οικολογικό Δείκτη 95 (EcoIndicator 95). Ο Οικολογικός Δείκτης 99 επεκτείνει τη λίστα με τους δείκτες που λαμβάνει υπόψη και εισάγει νέες μεθοδολογίες υπολογισμού και συνεπώς, μεγαλύτερη αξιοπιστία. Υπολογίζεται αφού πρώτα καθοριστούν όλα τα στάδια ζωής ενός προϊόντος. Έπειτα, εξετάζονται 5 παράγοντες/στάδια στη ζωή του προϊόντος, οι οποίοι συνεισφέρουν στην επιβάρυνση του περιβάλλοντος: η χρήση υλικών, οι διεργασίες, η μεταφορά του προϊόντος, η ενέργεια, όχι μόνο για τη μεταφορά των πρώτων υλών αλλά και για τα επιμέρους στάδια του Κύκλου Ζωής και τέλος, τις τεχνικές απόθεσης και ανακύκλωσης. Κάθε μια από τις παραπάνω κατηγορίες περιέχει επιμέρους στάδια, για τα οποία υπάρχει ένας συντελεστής στάθμισης περιβαλλοντικής επιβάρυνσης. Βάσει του ποσού χρήσης και του συντελεστή στάθμισης που υπάρχει για τα επιμέρους στάδια προκύπτει ένας αριθμός από βαθμούς (points). Αθροίζοντας τους βαθμούς του κάθε σταδίου, προκύπτει ένα σύνολο από βαθμούς για κάθε μια από τις 5 κατηγορίες. Αθροίζοντας τους βαθμούς κάθε κατηγορίας

προκύπτει ο αριθμός βαθμών, συνολικά, για τον κύκλο ζωής του προϊόντος. Όσο μεγαλύτερος ο αριθμός, τόσο μεγαλύτερη η επιβάρυνση. Τυπικές τιμές για το δείκτη αυτό από 0,1-1.

Κεφάλαιο 4. Στατιστικές Μέθοδοι για ανάπτυξη μοντέλων

Σε αυτό το κεφάλαιο αναφέρονται οι στατιστικές μέθοδοι που θα χρησιμοποιηθούν στο κεφάλαιο 5 για την ανάπτυξη των μοντέλων πρόβλεψης των δεικτών LCA

4.1 Ανάλυση Κυρίων Συνιστωσών

Η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis - PCA) είναι μια στατιστική μέθοδος που έχει ως σκοπό τη μείωση του αριθμού των μεταβλητών που περιγράφουν ένα σύστημα. Δεν παράγει άμεσα ένα μοντέλο σα μέθοδος (όπως η Γραμμική Παλινδρόμηση Πολλών Μεταβλητών), αλλά βοηθά στην περιγραφή του συστήματος με λιγότερες μεταβλητές, οι οποίες μπορούν να συμμετάσχουν σε μια παλινδρόμηση μεταξύ αυτών και των παρατηρήσεων (observations). Έτσι, στα μοντέλα που σχηματίζονται, συσχετίζονται λιγότερες μεταβλητές με τις παρατηρήσεις (πιθανότατα πειραματικά δεδομένα) και οδηγούμαστε σε πιο ακριβή μοντέλα.

Ας γίνει η υπόθεση πως υπάρχει ένα σύνολο μεταβλητών n ,X=[x_1, x_2,…,x_n]. Επίσης, υπάρχουν p παρατηρήσεις, όπου η κάθε μια αναλύεται στις παραπάνω n διαστάσεις και δίνει τιμές στην κάθε μεταβλητή. Αν αυτός ο αριθμός n είναι μικρός (ή αν έστω και ακόμα έχουμε δύο μεταβλητές), τότε είναι αρκετά απλό να παρατηρήσουμε τις σχέσεις μεταξύ αυτών των μεταβλητών. (Πρέπει να γίνει κατανοητό, πως είναι απαραίτητο να μπορούν να γίνουν συσχετίσεις μεταξύ των μεταβλητών x_1,…,x_n ,διαφορετικά δεν είναι εφικτή αυτή η μέθοδος). Αν όμως ο αριθμός αυτός είναι μεγάλος, τότε είναι ανέφικτο να παρατηρηθούν όλες οι συσχετίσεις μεταξύ τους και είναι δύσκολο να παρατηρείται το σύστημα μέσα από πολλές μεταβλητές. Τότε μπορούν να αναπτυχθούν νέες μεταβλητές, λιγότερες σε αριθμό (και ασφαλώς πιο εύκολες στο συσχετισμό τους), οι οποίες ονομάζονται Κύριες Συνιστώσες (Principal Components-PC) και προκύπτουν από το αρχικό σύνολο μεταβλητών και να περιγράφουν επαρκώς το σύστημα. Αυτές οι νέες μεταβλητές θα είναι m στον αριθμό ,όπου m<<n για να έχει νόημα η μέθοδος. Αυτές οι μεταβλητές (PC's) είναι γραμμικοί συνδυασμοί των αρχικών, δηλαδή υπάρχουν m διανύσματα C, όπου:

$$C_1 = [c_{1,1}, c_{2,1}, \dots , c_{n,1}] \quad (4.1)$$

$$C_2 = [c_{1,2}, c_{2,2}, \dots , c_{n,2}] \quad (4.2)$$

$$C_3 = [c_{1,3}, c_{2,3}, \dots , c_{n,3}] \quad (4.3)$$

⋮

$$C_m = [c_{1,m}, c_{2,m}, \dots, c_{n,m}]$$

Και

$$PC_1 = C_1 \cdot X = c_{1,1} \cdot x_1 + c_{2,1} \cdot x_2 + \cdots + c_{n,1} \cdot x_n \text{ (4.4)}$$

$$PC_2 = C_2 \cdot X = c_{1,2} \cdot x_1 + c_{2,2} \cdot x_2 + \cdots + c_{n,2} \cdot x_n \text{(4.5)}$$

$$\vdots$$

$$PC_m = C_m \cdot X = c_{1,m} \cdot x_1 + c_{2,m} \cdot x_2 + \cdots + c_{n,m} \cdot x_n \text{ (4.6)}$$

Να σημειωθεί πως τα PC's ονομάζονται και «λανθάνουσες μεταβλητές» (latent variables), διότι είναι προφανές πως αποτελούν την προβολή των αρχικών μεταβλητών σε έναν «λανθάνοντα» διανυσματικό χώρο.

Είναι πολύ σημαντικό να επισημανθεί τι εννοούμε, όταν λέμε πως το σύστημα πρέπει να περιγράφεται ικανοποιητικά από τα PC's. Όπως γίνεται αντιληπτό, ένα σύστημα περιέχει αρκετή πληροφορία, η οποία εκφράζεται μέσα από την κατανομή των τιμών των μεταβλητών, τη μεταβλητότητα και μέσα από τους γενικότερους στατιστικούς δείκτες που εκφράζουν τη συμπεριφορά του συστήματος. Είναι πολύ λογικό, όταν σε ένα σύστημα μειώνεται ο αριθμός των μεταβλητών που το περιγράφουν, να χάνεται ένα κομμάτι της πληροφορίας και κατά συνέπεια, όταν αργότερα γίνεται η παλινδρόμηση για να συσχετιστούν οι μεταβλητές με τα πειραματικά δεδομένα, να οδηγούμαστε σε σχέσεις λιγότερο αξιόπιστες, διότι ενσωματώνουν λιγότερο αντιπροσωπευτικό κομμάτι της τάσης/συμπεριφοράς (trend) του συστήματος. Όσο, λοιπόν, μικρότερος είναι αυτός ο αριθμός m, με τόσο λιγότερα PC's επιλέγουμε να μοντελοποιήσουμε ένα σύστημα και συνεπώς, μεγαλύτερο κομμάτι «πληροφορίας» χάνεται και σε λιγότερο αξιόπιστα μοντέλα οδηγούμαστε. Το πρόβλημα επιλογής του αριθμού PC's, τα οποία θα αναπαριστούν το σύστημα, είναι ένα ξεκάθαρο πρόβλημα αντισταθμίσεων (trade-off): λιγότερα PC's οδηγούν σε λιγότερο αξιόπιστο σύστημα, αλλά αυτό αντισταθμίζεται από την πιο αξιόπιστη συσχέτιση λόγω των λιγότερων μεταβλητών. Αντίθετα, όσο περισσότερα τα PC's, τόσο μικρότερο κομμάτι «πληροφορίας» χάνεται, οπότε υπάρχει μεν αξιοπιστία εκεί, αλλά δυσκολότερα εξάγεται ένα αξιόπιστο μοντέλο με περισσότερες μεταβλητές, οπότε πιθανόν να αντισταθμίζεται η αξιοπιστία του μοντέλου από μια λιγότερο καλή συσχέτιση. Η «πληροφορία» αυτή εκφράζεται άμεσα από τη διακύμανση που έχει η κάθε μεταβλητή στις παρατηρήσεις που συμμετέχουν στο σύστημα. Είναι, λοιπόν, απαραίτητο, όταν παράγεται το κάθε PC, να παράγεται έτσι το διάνυσμα C του, ώστε να διατηρεί τη μέγιστη διακύμανση (variance) μεταξύ των μεταβλητών, ώστε να διατηρείται η βέλτιστη μεταβλητότητα και η καλύτερη δυνατή κατανομή μεταξύ των μεταβλητών του αρχικού πληθυσμού. Πρέπει, επιπλέον, να σημειωθεί πως όλα τα PC's, μεταξύ τους, ανά δύο είναι ορθογώνια

διανύσματα. Αυτό συμβαίνει, διότι, τα PC's προκύπτουν άμεσα από τα ιδιοδιανύσματα του πίνακα συσχέτισης (covariance matrix) των μεταβλητών.

Ένα ακόμα χαρακτηριστικό είναι πως το πρώτο PC που εξάγεται (PC_1) ενσωματώνει το μεγαλύτερο ποσοστό της αρχικής διακύμανσης των μεταβλητών. Ακολουθούν τα υπόλοιπα PC's με φθίνουσα σειρά ποσοστού κάλυψης της διακύμανσης. Τελικά, από μια PCA, προκύπτουν δυο πίνακες. Ο πρώτος πίνακας περιλαμβάνει τους συντελεστές (coefficients) $c_{i,j}$ που ανταποκρίνεται στην i μεταβλητή και ανήκει στο κάθε διάνυσμα C_j.

Πίνακας 4.1. Πίνακας συντελεστών PCA (Coefficient matrix)

Μεταβλητές/PC's	PC_1	PC_2		PC_n
x_1	$c_{1,1}$	$c_{1,2}$	⋱	$c_{1,n}$
x_2	$c_{2,1}$	$c_{2,2}$		$c_{2,n}$
⋮	⋮	⋮		⋮
x_n	$c_{n,1}$	$c_{n,2}$		$c_{n,n}$

Όπως φαίνεται, παράγονται n PC's (δηλαδή όσες και οι αρχικές μεταβλητές) και ένας πίνακας 1xn, όπου φαίνεται τη διακύμανση που εξασφαλίζει το κάθε PC. Προφανώς και αν προστεθούν τα στοιχεία του πίνακα, θα προκύψει η αρχική διακύμανση του πληθυσμού των μεταβλητών x. Εξαρτάται από το χρήστη, ποιο ποσοστό της αρχικής διακύμανσης θέλει να καλύψει και να επιλέξει ανάλογα τον αριθμό των PC's που θέλει.

Όμως, πριν την παραγωγή των PC's κάθε παρατήρηση που αποτελεί το σύστημα αναλυόταν σε n διαστάσεις: σε κάθε παρατήρηση έπαιρνε η κάθε μεταβλητή μια συγκεκριμένη τιμή. Χαρακτηριζόταν, δηλαδή, η κάθε παρατήρηση από ένα διάνυσμα 1 x n, όπου οι τιμές του διανύσματος ήταν οι τιμές της κάθε μεταβλητής. Τώρα που οι ανεξάρτητες μεταβλητές άλλαξαν και η κάθε παρατήρηση προβάλλεται σε διαφορετικές διαστάσεις πρέπει να προκύψει ένας νέος πίνακας (p x m), όπου η κάθε γραμμή θα είναι οι τιμές των νέων μεταβλητών για την κάθε παρατήρηση. Αυτός είναι ο πίνακας των αποτελεσμάτων (score):

Πίνακας 4.2. Πίνακας αποτελεσμάτων (score)

Παρατηρήσεις/PC's	PC_1	PC_2		PC_n
παρατήρηση$_1$	$a_{1,1}$	$a_{1,2}$	⋱	$a_{1,n}$
παρατήρηση$_2$	$a_{2,1}$	$a_{2,2}$		$a_{2,n}$
⋮	⋮	⋮		⋮
παρατήρηση$_n$	$a_{n,1}$	$a_{n,2}$		$a_{n,n}$

Ακολουθεί η Παλινδρόμηση Κύριων Συνιστωσών, όπου όταν είναι επιθυμητό να αναπτυχθεί ένα γραμμικό μοντέλο, αντί να γίνει άμεσα παλινδρόμηση με τις αρχικές μεταβλητές και τις παρατηρήσεις, γίνεται έμμεσα παλινδρόμηση μεταξύ των παρατηρήσεων και των PC's. Άρα το αρχικό γραμμικό πρόβλημα:

$$y = b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_n \cdot x_n + \alpha \text{ (4.7)}$$

αλλάζει σε:

$$y = b_1{}' \cdot PC_1 + b_2{}' \cdot PC_2 + \cdots + b_m{}' \cdot PC_m + a' \text{ (4.8)}$$

,όπου δεν προσδιορίζονται n+1 παράμετροι, παρά m+1 παράμετροι (το +1 προκύπτει λόγω του σταθεράς του μοντέλου). Για να προκύψει η πρώτη εξίσωση και τα $b_1,\ldots,b_n$ από τα $b_1{}',\ldots,b_n{}'$, ακολουθούνται τα επόμενα βήματα:

1. Αντικαθιστώνται οι εκφράσεις των PC's στη δεύτερη εξίσωση, εφαρμόζεται η επιμεριστική ιδιότητα και εξάγονται κοινοί παράγοντες:

$$y = b_1{}' \cdot \left(c_{1,1} \cdot x_1 + c_{2,1} \cdot x_2 + \cdots + c_{n,1} \cdot x_n\right) + b_2{}' \cdot \left(c_{1,2} \cdot x_1 + c_{2,2} \cdot x_2 + \cdots + c_{n,2} \cdot x_n\right) + \cdots + b_m{}' \cdot \left(c_{1,m} \cdot x_1 + c_{2,m} \cdot x_2 + \cdots + c_{n,m} \cdot x_n\right) + a' \text{ (4.9)}$$

$$y = \left(b_1{}' \cdot c_{1,1} + b_2{}' \cdot c_{1,2} + \cdots + b_m{}' \cdot c_{1,m}\right) \cdot x_1 + \left(b_1{}' \cdot c_{2,1} + b_2{}' \cdot c_{2,2} + \cdots + b_m{}' \cdot c_{2,m}\right) \cdot x_2 + \cdots + \left(b_1{}' \cdot c_{n,1} + b_2{}' \cdot c_{n,2} + \cdots + b_m{}' \cdot c_{n,m}\right) \cdot x_n + a' \text{ (4.10)}$$

2. Προκύπτει ο κάθε συντελεστής του γραμμικού μοντέλου για κάθε μεταβλητή b_i ως εξής:

$$b_i = \left(b_1{}' \cdot c_{\iota,1} + b_2{}' \cdot c_{\iota,2} + \cdots + b_m{}' \cdot c_{\iota,m}\right) \text{ (4.11)}$$

Περισσότερες πληροφορίες για το PCA δίνονται από τον Joliffe (2002).

4.2 Παλινδρόμηση μερικών ελαχίστων τετραγώνων (Partial Least Squares, PLS)

Η μέθοδος αξιοποιεί, επίσης, τη χρήση λανθανουσών μεταβλητών και την παλινδρόμηση τους. Λειτουργεί, σχεδόν ακριβώς, όπως και η PCA/PCR. Μόνο που σε αυτήν την περίπτωση, δεν εξάγονται λανθάνουσες μεταβλητές μόνο για τις μεταβλητές εισόδου στο μοντέλο, αλλά και για τις μεταβλητές εξόδου, δηλαδή για τις μεταβλητές που περιγράφονται για τις παρατηρήσεις. Όπως προαναφέρθηκε θεωρήσαμε, πως στο PCA έγινε η συσχέτιση για την ανάπτυξη γραμμικού μοντέλου μεταξύ των PC's και των y, μεταβλητών εξόδου (το y μπορεί να είναι και ένα διάνυσμα που να περιλαμβάνει k μεταβλητές εξόδου αντί για μία), αντί για απευθείας συσχέτιση μεταξύ x και y. Στην περίπτωση, του PLS, και οι μεταβλητές εξόδου προβάλλονται σε ένα φανταστικό διανυσματικό χώρο, ας υποθέσουμε U, ενώ οι μεταβλητές εισόδου σε έναν άλλο διανυσματικό χώρο, ας υποθέσουμε T. Αυτή τη φορά, η συσχέτιση θα αναπτυχθεί μεταξύ των μεταβλητών του χώρου T και του χώρου U. Με άλλα λόγια, η μέθοδος αυτή «εντοπίζει» πως ο λόγος που είναι επιθυμητή η μείωση του αρχικού χώρου μεταβλητών είναι, επειδή, είναι επιθυμητή η ανάπτυξη ενός μοντέλου και ότι οι νέες «λανθάνουσες μεταβλητές», θα συσχετιστούν με μια έξοδο (output). Αυτό δεν ήταν εφικτό με το PCA: το PCA μειώνει, μεν, το μέγεθος του χώρου των μεταβλητών εισόδου X, αλλά δε «γνωρίζει» ότι αυτές οι νέες μεταβλητές που παράγει (PC's) θα συμμετάσχουν στην ανάπτυξη ενός μοντέλου. Υπό αυτήν την άποψη, η Ανάλυση Κύριων Συνιστωσών και η Παλινδρόμηση Κύριων Συνιστωσών είναι δύο διαφορετικές, ανεξάρτητες μεταξύ τους μεθοδολογίες: η μια φτιάχνει τις συνιστώσες, ενώ η δεύτερη σχηματίζει το μοντέλο.

Το PLS έχοντας, λοιπόν, αυτήν την παραπάνω «γνώση», όταν εξάγει τις συνιστώσες (δε λέγονται πια κύριες συνιστώσες) των μεταβλητών του X, τις εξάγει με τέτοιο τρόπο ώστε να έχουν το βέλτιστο δυνατό συσχετισμό με τις συνιστώσες των μεταβλητών εξόδου του Y. Η μέθοδος κινείται, δηλαδή, προς τέτοιες κατευθύνσεις στο διανυσματικό χώρο για να αναζητήσει τις νέες συνιστώσες του X, ώστε αυτά να έχουν τη μεγαλύτερη δυνατή συσχέτιση με τις συνιστώσες του Y. Μπορούμε, λοιπόν, να πούμε πως δεν υπάρχουν μόνο, οι περιορισμοί για τα συνιστώσες του X, να είναι ορθογώνια και να εξασφαλίσουν τη μέγιστη διακύμανση μεταξύ των μεταβλητών, αλλά υπάρχει και ένας τρίτος περιορισμός: η συσχέτιση μεταξύ των X και Y μεταβλητών να είναι μέγιστη.

Όπως είναι αναμενόμενο, εξάγονται επιπλέον δυο πίνακες: ένας για τους συντελεστές μεταξύ των συνιστωσών και των μεταβλητών εξόδου και ένας για τα αποτελέσματα, των παρατηρήσεων και των νέων τους λανθανουσών μεταβλητών. Το πρόβλημα επιλογής του αριθμού των συνιστωσών παραμένει και εδώ το ίδιο με το PCA. Περισσότερες πληροφορίες για το PLS στο An Introduction to Partial Least Squares Regression του Randall D. Tobias

4.3 Παρεμβολή τύπου «Kriging»

Η παρεμβολή τύπου «Kriging» ακολουθεί μια μη γραμμική μέθοδο για την ανάπτυξη ενός μοντέλου. Το χαρακτηριστικό αυτής της μεθόδου είναι πως δεν ακολουθεί μια διαδικασία παλινδρόμησης για τον υπολογισμό κάποιων γενικών παραμέτρων και τη μετέπειτα χρήση αυτών σε ένα τύπο για τον υπολογισμό μιας νέας τιμής, όπως γίνεται στα γραμμικά μοντέλα. Ας θεωρήσουμε, όπως και στις προηγούμενες περιπτώσεις, ένα σύνολο p παρατηρήσεων, οι οποίες αναλύονται/προβάλλονται σε n διαστάσεις (εκφράζονται με x_n μεταβλητές). Αυτές αποτελούν ένα σύνολο παρατηρήσεων με τις οποίες γίνεται η «εκπαίδευση» του μοντέλου και στο εξής θα ονομάζονται σύνολο εκπαίδευσης. Αφού προκύψει το μοντέλο, οποιαδήποτε νέα μέτρηση που είναι επιθυμητό-γνωρίζοντας την προβολή της στις n διαστάσεις-, να βρεθεί η τιμή της, θεωρείται από το μοντέλο ως παρεμβολή (interpolation) μεταξύ των σημείων του συνόλου εκπαίδευσης και η τιμή της υπολογίζεται ,όπως θα φανεί στην πορεία (Engineering Design via Surrogate Modelling, 2008).

Για να γίνει η διαδικασία εκπαίδευσης, πρέπει να σχεδιαστούν νοητά πρώτα οι n διαστάσεις στις οποίες αναλύονται οι μεταβλητές των παρατηρήσεων που αποτελούν το σύνολο εκπαίδευσης, καθώς και οι νέες μετρήσεις που θέλουμε να προβλέψουμε την τιμή τους. Σα δεύτερο βήμα, φέρουμε πάνω σε αυτό το χώρο των n διαστάσεων, ως σημεία τις παρατηρήσεις του συνόλου εκπαίδευσης. Για παράδειγμα, ας υποθέσουμε, πως οι παρατηρήσεις που αποτελούν το σύνολο εκπαίδευσης περιγράφονται από 5 μεταβλητές (x_1, x_2,…, x_5). Τότε, έχουμε το σχεδιασμό χώρου 5 διαστάσεων. Αν για ένα σημείο του συνόλου εκπαίδευσης, έστω $x^{(1)}$ οι τιμές των μεταβλητών έχουν τιμή, για παράδειγμα: x_1=1.5, x_2=-8.4, x_3=0.7, x_4=4.2 και x_5=100. Τότε, στο χώρο 5 διαστάσεων που έχει σχηματιστεί, φέρουμε το σημείο (1.5, -8.4, 0.7, 4.2, 100). Αν για το δεύτερο σημείο, έστω $x^{(2)}$, του συνόλου εκπαίδευσης, έχουν οι μεταβλητές τιμές, π.χ. x_1=2.7, x_2=-10, x_3=0.1, x_4=5.6, και x_5=209, τότε το δεύτερο σημείο που θα φέρουμε στο χώρο των 5 διαστάσεων, θα είναι το (2.7, -10, 0.1, 5.6, 209) κ.ο.κ.. Όταν όλα τα σημεία σχηματιστούν στο χώρο, τότε σχηματίζουν ένα πλέγμα (grid) και η προβολή του κάθε σημείου αποτελεί ένα κέντρο του χώρου. Αυτό το πλέγμα μαζί με τα κέντρα, αποτελούν το βασικό στοιχείο για τον υπολογισμό μιας νέας τιμής.

Όταν, λοιπόν, προκύψει μια νέα μέτρηση, για την οποία είναι γνωστές οι μεταβλητές της και είναι επιθυμητή η πρόβλεψη της τιμής της, τότε το πρώτο και βασικότερο βήμα της μεθόδου είναι να προβληθεί το καινούργιο σημείο στο χώρο και να υπολογιστεί η «απόσταση» του από τα κέντρα του πλέγματος. Ο ορισμός της «απόστασης» αυτής δε συμπίπτει σε καμία περίπτωση με τη γνωστή απόσταση που ορίζεται μεταξύ δυο σημείων στο χώρο. Συγκεκριμένα, η απόσταση που χρησιμοποιείται δίνεται από:

$$cor\left(Y\left(x^{(i)}\right), Y(x)\right) = exp\left(-\sum_{j=1}^{n} \theta_j \cdot \left|x_j{}^{(i)} - x_j\right|^{p_j}\right) \quad (4.12)$$

Στον παραπάνω τύπο φαίνεται η απόσταση μεταξύ του νέου σημείου (x) , για το οποίο γίνεται η πρόβλεψη και ενός σημείου του συνόλου εκπαίδευσης, i ($x^{(i)}$). Κατά σύμβαση, χρησιμοποιείται ο όρος

correlation μεταξύ των σημείων και όχι distance. Αυτό οφείλεται στη φύση της μεθόδου και στη χρήση των διάφορων εννοιών και τύπων. Ως j απεικονίζεται η κάθε διάσταση και ο όρος p_j είναι σταθερός για όλες τις διαστάσεις και ίσος με 2, επειδή εμπίπτει στον Ευκλείδειο χώρο. Ο όρος θ_j είναι μια σταθερή παράμετρος χαρακτηριστική για κάθε διάσταση και υπολογίζεται. Όπως γίνεται αντιληπτό, για τον υπολογισμό της απόστασης, υπολογίζεται για την κάθε διάσταση, η διαφορά των συντεταγμένων (μεταβλητών) μεταξύ του σημείου της νέας μέτρησης και του σημείου του συνόλου εκπαίδευσης και αφού σταθμιστεί αυτή η διαφορά αθροίζεται με αυτές των άλλων διαστάσεων. Η παραπάνω διαδικασία εκτελείται και για τα p σημεία του συνόλου εκπαίδευσης και σχηματίζεται το διάνυσμα ψ (p x 1), που περιέχει τις αποστάσεις της νέας μέτρησης από κάθε κέντρο.

Ακολουθεί η ίδια διαδικασία υπολογισμού αποστάσεων μεταξύ όλων των κέντρων του πλέγματος και σχηματίζεται ο πίνακας συσχέτισης (correlation matrix):

$$\Psi = \begin{bmatrix} cor\left(Y\left(x^{(1)}\right), Y\left(x^{(1)}\right)\right) & \cdots & cor\left(Y\left(x^{(1)}\right), Y\left(x^{(n)}\right)\right) \\ \vdots & \ddots & \vdots \\ cor\left(Y\left(x^{(n)}\right), Y\left(x^{(1)}\right)\right) & \cdots & cor\left(Y\left(x^{(n)}\right), Y\left(x^{(n)}\right)\right) \end{bmatrix} \quad (4.13)$$

Όπως είναι κατανοητό, ο παραπάνω πίνακας έχει τη διαγώνιο του ίση με 1, γιατί τα στοιχεία που την απαρτίζουν είναι τα $cor\left(Y\left(x^{(i)}\right), Y\left(x^{(i)}\right)\right)$ και η απόσταση ενός σημείου από τον εαυτό του είναι 0, άρα το exp(0)=1. Επιπλέον, ο πίνακας αυτός είναι συμμετρικός γιατί για ένα στοιχείο Ψ(i,j)= $cor\left(Y\left(x^{(i)}\right), Y\left(x^{(j)}\right)\right)$ είναι το Ψ(j,i)= $cor\left(Y\left(x^{(j)}\right), Y\left(x^{(i)}\right)\right)$, όμως η απόσταση μεταξύ σημείων i και j είναι ίδια μεταξύ j και i. Έπειτα ο αντίστροφος του παραπάνω πίνακα (Ψ^{-1}) αποσυντίθεται με τη μέθοδο LU (LU decomposition).

Αν με y συμβολιστεί το διάνυσμα p x 1, που περιέχει τις τιμές των παρατηρήσεων και με μ συμβολιστεί το κέντρο των σημείων που απαρτίζουν το σύνολο εκπαίδευσης, $X = \{x^{(1)}, x^{(2)}, \dots, x^{(p)}\}$ τότε υπολογίζεται το $\hat{\mu}$, που είναι η εκτίμηση της μέγιστης πιθανότητας για το μ:

$$\hat{\mu} = \frac{1^T \Psi^{-1} y}{1^T \Psi^{-1} 1} \quad (4.14)$$

Η νέα εκτίμηση, $\hat{y}$, θα δίνεται από τον τύπο:

$$\hat{y}(x) = \hat{\mu} + \psi^T \Psi^{-1}(y - 1\hat{\mu}) \quad (4.15)$$

Το τελευταίο κομμάτι που είναι χρήσιμο να αναφερθεί για την παρεμβολή τύπου «kriging» είναι η χρήση της παραμέτρου λ. Για να γίνει κατανοητή η χρησιμότητα της παραμέτρου λ, πρέπει να επισημανθούν πρώτα κάποια βασικά χαρακτηριστικά του μοντέλου παρεμβολής τύπου «kriging».

Καταρχάς, όταν χρησιμοποιείται αυτή η διαδικασία, στο μοντέλο που αναπτύσσεται γίνεται εξαρχής υπερπροσαρμογή των δεδομένων ή υπερπαραμετροποίηση (overfitting). Αν θέλουμε να το παραστήσουμε διαγραμματικά θα μπορούσαμε να φανταστούμε την υπερπροσαρμογή ως εξής: έστω πως έχουμε ένα σύνολο δεδομένων που αναπαριστώνται από σημεία στο χώρο των δύο διαστάσεων (σύνολο εκπαίδευσης). Αν είναι επιθυμητή η ανάπτυξη ενός μοντέλου, τότε με την παλινδρόμηση, θα σχηματιζόταν μια καμπύλη (βέλτιστη) του μοντέλου η οποία θα ελαχιστοποιούσε την απόσταση των πειραματικών σημείων από αυτή. Στην υπερπαραμετροποίηση, η καμπύλη δεν ελαχιστοποιεί, απλά, τις αποστάσεις, αλλά περνάει από τα πειραματικά σημεία. Αυτό έχει σαν αποτέλεσμα, να προβλέπονται με απόλυτη ακρίβεια τα σημεία του συνόλου εκπαίδευσης (0% σφάλμα), αλλά να μην έχει απολύτως καμία ακρίβεια σε νέα σημεία που δεν ανήκουν μέσα στο σύνολο εκπαίδευσης. Αυτό είναι φυσικό, μιας και το μοντέλο «αναλώνεται» περισσότερο στο να περάσει την καμπύλη από όλα τα σημεία και δε μπορεί να «αντιληφθεί» πως τα νέα σημεία που θα επακολουθήσουν, μπορεί να ακολουθούν μια άλλη τάση. Αντίθετα, το μοντέλο που δεν κάνει υπερπαραμετροποίηση, «αντιλαμβάνεται» και ακολουθεί τη γενικότερη τάση των σημείων του συνόλου εκπαίδευσης και δέχεται πως δε χρειάζεται να περάσει από όλα τα πειραματικά σημεία, αλλά αποτυπώνει τη γενικότερη μεταβλητότητα του. Αποτέλεσμα αυτού είναι πως υπάρχει σαφώς μεγαλύτερο σφάλμα στην πρόβλεψη τιμών στο σύνολο εκπαίδευσης, αλλά και μειωμένο σφάλμα στο σύνολο δοκιμής. Το γεγονός ότι η παρεμβολή τύπου «kriging», σα μέθοδος ανάπτυξης μοντέλου, κάνει υπερπαραμετροποίηση φαίνεται από τον ορισμό της απόστασης. Έχει ήδη αναφερθεί πως:

$$cor\left(Y\left(x^{(i)}\right), Y(x)\right) = exp\left(-\sum_{j=1}^{n} \theta_j \cdot \left|x_j{}^{(i)} - x_j\right|^{p_j}\right) \quad (4.16)$$

Αν είναι επιθυμητό να προβλεφθεί η τιμή για ένα μόριο του συνόλου εκπαίδευσης, έστω του πρώτου, πρέπει να σχηματιστεί το διάνυσμα ψ, που περιέχει τις αποστάσεις του μορίου αυτού από τα υπόλοιπα:

$$cor\left(Y\left(x^{(i)}\right), Y\left(x^{(1)}\right)\right) = exp\left(-\sum_{j=1}^{n} \theta_j \cdot \left|x_j{}^{(i)} - x_j{}^{(1)}\right|^{p_j}\right) \quad (4.17)$$

Η απόσταση του πρώτου μορίου, όμως, με τον εαυτό του είναι:

$$cor\left(Y\left(x^{(1)}\right), Y(x^1)\right) = exp\left(-\sum_{j=1}^{n} \theta_j \cdot \left|x_j{}^{(1)} - x_j{}^{(1)}\right|^{p_j}\right) = 1 \quad (4.18)$$

Ενώ οι αποστάσεις με όλα τα υπόλοιπα κέντρα θα είναι:

$$cor\left(Y\left(x^{(i)}\right), Y\left(x^{(1)}\right)\right) = \frac{1}{exp\left(\sum_{j=1}^{n} \theta_j \cdot \left|x_j{}^{(i)} - x_j{}^{(1)}\right|^{p_j}\right)} \quad (4.19)$$

,δηλαδή η τιμή αυτή είναι κατά πολύ μικρότερη του ενός. Φαίνεται πως η απόσταση του ενός κέντρου από τον εαυτό του δίνει μια πολύ μεγαλύτερη τιμή στο $cor\left(Y\left(x^{(i)}\right), Y(x)\right)$ από ότι οι υπόλοιπες, άρα επηρεάζει αργότερα και κατά πολύ περισσότερο την νέα εκτίμηση, «ωθώντας» την πολύ περισσότερο προς την πειραματική τιμή του πρώτου κέντρου. Εν τέλει, αυτή η εκτίμηση θα συμπέσει με την πειραματική τιμή.

Δεν είναι, συνεπώς, αναμενόμενο να έχει αυτό το μοντέλο καλές προβλεπτικές δυνατότητες για καινούργια μόρια. Παρολαυτά, είναι δυνατόν παραλείποντας να εισαχθούν κάποια δεδομένα, δηλαδή κάποια σημεία, του συνόλου εκπαίδευσης, να αποφευχθεί η υπερπαραμετροποίηση. Αυτό είναι λογικό, μιας και το μοντέλο δε θα προσπαθήσει καν να περάσει από αυτά τα σημεία, συντελώντας έτσι σε μια πιο «ομαλή» καμπύλη. Όμως, δεν κρίνεται πάντα ως η κατάλληλη λύση, αφού έτσι παραλείπεται ένα κομμάτι του συνόλου εκπαίδευσης και συνεπώς, της συνολικής πληροφορίας που υπάρχει μέσα στο αρχικό σύνολο δεδομένων.

Αντί της παραπάνω αντιμετώπισης, προτείνεται η χρήση της παραμέτρου λ. Συγκεκριμένα, για να αποφύγουμε να αφαιρεθούν δεδομένα, προτείνεται στα υπάρχοντα δεδομένα να προστεθεί ένα είδος «θορύβου». Με τον όρο αυτό, εννοούμε οι πειραματικές τιμές για το κάθε μόριο που συμμετέχει στο σύνολο εκπαίδευσης, να μεταβληθεί κατά ένα πολύ μικρό ποσοστό η τιμή του δείκτη που πρέπει να προβλεφθεί, δηλαδή η πειραματική τιμή του GWP, CED, EI 99. Με αυτό τον τρόπο, το μοντέλο θα «βλέπει» νέες πειραματικές τιμές (που περιλαμβάνουν το «θόρυβο») και δε θα προσπαθήσει να καλύψει τα πραγματικά σημεία του συνόλου εκπαίδευσης. Αν για παράδειγμα ένα σημείο για το CED έχει πειραματική μέτρηση 250, τότε το μοντέλο με τη διαταραχή που έχει επιβληθεί θα εντοπίζει μια διαφορετική τιμή π.χ. 245. Και το μοντέλο θα προσπαθήσει να περάσει την καμπύλη από εκεί. Έτσι, η καμπύλη που θα σχηματιστεί δε θα περνά από τα πραγματικά σημεία του συνόλου εκπαίδευσης και η καμπύλη θα είναι πιο ομαλή. Όπως είναι προφανές, η παράμετρος λ είναι μια παράμετρος βελτιστοποίησης. Στην παραπάνω προγραμματιστική διαδικασία MATLAB, που έφερε αποτελέσματα για την παρεμβολή τύπου «kriging», ακολουθήθηκε μια επαναληπτική διεργασία που εύρισκε κάθε φορά το βέλτιστο λ, που δίνει την καλύτερη συσχέτιση.

4.4 Μέθοδος συναρτήσεων ακτινικής βάσης

Η μέθοδος αυτή (Radial Basis Functions-RBF) αποτελεί ακόμα μια μέθοδο παρεμβολής, που χρησιμοποιεί διαφορετικά κέντρα για τον υπολογισμό των αποστάσεων. Αυτή η μέθοδος λειτουργεί παρόμοια με την παρεμβολή τύπου «kriging». Η διαφορά εδώ, είναι πως το πλέγμα που δημιουργείται διαφέρει από το πλέγμα της παρεμβολής τύπου «kriging»: ενώ στην παρεμβολή τύπου «kriging», τα κέντρα αποτελούνταν από τα ίδια τα στοιχεία του συνόλου εκπαίδευσης (παρατηρήσεις προβεβλημένες στις n διαστάσεις), στα RBF, τα κέντρα του πλέγματος, αποτελούνται από άλλα σημεία, τα οποία επιλέγονται από μια αλγοριθμική διαδικασία, έτσι ώστε να έχουν την ελάχιστη απόσταση από ένα ή περισσότερα σημεία του συνόλου εκπαίδευσης (δηλαδή, από προβολές σημείων που αντιστοιχούν σε παρατηρήσεις). Πιο συγκεκριμένα, τα νέα αυτά κέντρα επιλέγονται έτσι, ώστε να «εκπροσωπούν» ένα η περισσότερα σημεία του συνόλου εκπαίδευσης. Ένα ή περισσότερα σημεία του συνόλου εκπαίδευσης, λοιπόν, «ανήκουν» σε ένα

από τα νέα κέντρα και όλες οι πράξεις που θα περιλάμβαναν τα σημεία αυτά, γίνονται περιλαμβάνοντας το νέο αντιπροσωπευτικό τους κέντρο (Alexandridis et al. 2003).

Για να επιτευχθεί αυτό, πρέπει πρώτα η κάθε διάσταση να χωριστεί σε ίσο αριθμό τμημάτων. Δεν είναι απαραίτητο, ασφαλώς, πως κάθε διάσταση θα χωρίζεται στον ίδιο αριθμό τμημάτων όπως και οι υπόλοιπες n-1. Αυτά τα κομμάτια της κάθε διάστασης θα ονομάζονται υποδιαστήματα και έχουν, όπως γίνεται αντιληπτό, άκρα τα οποία είναι κοινά μεταξύ δυο υποδιαστημάτων που συνορεύουν. Τα μέσα αυτών των υποδιαστημάτων για την κάθε διάσταση, αποτελούν και υποψήφια κέντρα, τα οποία ευρίσκονται με το σχηματισμό των υποδιαστημάτων και θα ενεργοποιηθούν αν ένα σημείο του συνόλου εκπαίδευσης βρεθεί ικανοποιητικά κοντά. Η διαδικασία ξεκινά με την εισαγωγή του πρώτου σημείου του συνόλου εκπαίδευσης στο χώρο των n διαστάσεων. Ευρίσκεται, για την κάθε διάσταση, το κοντινότερο κέντρο υποδιαστήματος και τελικά, συγκροτείται το πρώτο κέντρο του πλέγματος. Έπειτα γίνεται η εισαγωγή του δεύτερου στοιχείου του συνόλου εκπαίδευσης και συνεχίζεται με τα υπόλοιπα. Πρώτα, η αλγοριθμική διαδικασία ελέγχει αν πληρούνται τα κριτήρια απόστασης του κέντρου του πλέγματος από το νέο σημείου του συνόλου εκπαίδευσης και αν είναι εφικτό να ενταχθεί σε κάποιο ήδη ενεργοποιημένο κέντρο. Αν δεν πληρούνται τα κριτήρια απόστασης, τότε εγκαινιάζεται ένα καινούργιο κέντρο, σύμφωνα με τον παραπάνω τρόπο, διαφορετικά εντάσσεται στα ήδη υπάρχοντα. Η διαδικασία επαναλαμβάνεται μέχρι να ενταχθούν στο πλέγμα όλα τα σημεία του συνόλου εκπαίδευσης. Ακολουθεί μια παρόμοια διαδικασία υπολογισμού της νέας τιμής, όπως και στην παρεμβολή τύπου «kriging».

Το όφελος αυτής της πρακτικής είναι σημαντικό: το να μην ανατίθεται ένα κέντρο του πλέγματος σε κάθε στοιχείο, αλλά να αντιπροσωπεύεται ένας αριθμός στοιχείων από ένα κέντρο συντελεί στο μικρότερο αριθμό κέντρων, που μπορεί να οδηγήσει στο πιο εύκολο σχηματισμό συσχετίσεων και συνεπώς, στην πιο αξιόπιστη ανάπτυξη μοντέλων. Παράλληλα, όμως, η περιγραφή ενός συστήματος από λιγότερα κέντρα, μπορεί να συντελεί σε λιγότερο καλή ενσωμάτωση πληροφορίας (όπως και στα PC's) στο μοντέλο. Είναι προφανές, λοιπόν, πως ο αριθμός των υποδιαστημάτων που χωρίζεται μια διάσταση είναι παράμετρος βελτιστοποίησης. Με τη μεταβολή αυτού του αριθμού, μπορεί να προκύπτει πιο αραιό ή πιο πυκνό πλέγμα και συνεπώς, να δημιουργούνται λιγότερα ή και περισσότερα κέντρα.

4.5 Ανάλυση Διακύμανσης (Analysis of Variance, ANOVA)

Στα πλαίσια αυτής της διπλωματικής χρησιμοποιείται ως στατιστικό τεστ η Ανάλυση Διακύμανσης, γι' αυτό και αναφέρονται συνοπτικά η λειτουργία της και τα χαρακτηριστικά της. Η ANOVA είναι μια στατιστική μέθοδος, κατά την οποία η μεταβλητότητα που υπάρχει σε ένα σύνολο διασπάται σε επιμέρους συνιστώσες, σκοπό την κατανόηση την κατανόηση της σημαντικότητας των διαφορετικών πηγών προέλευσής της. Με άλλα λόγια, τα δεδομένα ενός συνόλου ομαδοποιούνται με βάση κάποια κοινά

χαρακτηριστικά τους και μελετάται η στατιστική κατανομή τους, ώστε να επισημανθεί η σημασία του κάθε παράγοντα που συνεισφέρει σε όλο το σύνολο συνολικά. Για να γίνει πιο κατανοητή σα μέθοδος, κρίνεται σκόπιμο να αναφερθεί ένα παράδειγμα. Ας θεωρήσουμε πως έχουμε τις παρακάτω παρατηρήσεις (1 έως 3) χωρισμένες σε δυο διαφορετικές κατηγορίες (ομάδα 1, ομάδα 2):

Πίνακας 4.3. Παράδειγμα συνόλων για Ανάλυση Διακύμανσης

	Ομάδα 1	**Ομάδα 2**
Παρατήρηση 1	2	6
Παρατήρηση 2	3	7
Παρατήρηση 3	1	5
Μέση τιμή group	2	6
Sum of Squares(SS)	2	2
Γενικός μέσος όρος	4	
SS συνόλου παρατηρήσεων	28	

Έχει υπολογιστεί το άθροισμα των τετραγώνων των αποκλίσεων (διαφορών) από τη μέση τιμή για το σύνολο των παρατηρήσεων και για την κάθε ομάδας, καθώς και η μέση τιμή για το σύνολο των παρατηρήσεων και για την κάθε ομάδα. Η ομαδοποίηση σε ομάδες μπορεί να γίνεται βάσει ενός χαρακτηριστικού. Οι διάφορες τιμές (ποιοτικές και ποσοτικές) αυτού του χαρακτηριστικού ονομάζονται επίπεδα (levels) και αυτό το χαρακτηριστικό του οποίου μελετάται, με την ομαδοποίηση, η επίδρασή του ονομάζεται παράγοντας (factor). Παρατηρείται πως υπάρχει ικανή διαφορά μεταξύ των μέσων όρων των επιμέρους ομάδων. Το άθροισμα των τετραγώνων και των δύο ομάδων είναι 2. Αν προστεθούν τα δυο αθροίσματα προκύπτει 4. Όμως το άθροισμα των τετραγώνων συνολικά για το σύνολο δεδομένων (αν δε ληφθεί υπόψη η κατηγοριοποίηση σε ομάδες) είναι 28. Είναι φανερό, πως υπολογίζοντας το άθροισμα τετραγώνων (Sum of Squares-SS) βασιζόμενοι στην εντός-της-ομάδας μεταβλητότητα έχει ως αποτέλεσμα ένα κατά πολύ μικρό αποτέλεσμα με το αν υπολογιστεί βασιζόμενοι σε όλο το σύνολο. Η διαφορά αυτή έγκειται στη διαφορά που έχουν και οι μέσοι όροι. Αποδεικνύεται πως το άθροισμα των τετραγώνων του συνόλου των παρατηρήσεων έχει δύο συνιστώσες, δηλαδή αποτελεί άθροισμα δύο παραμέτρων: του αθροίσματος τετραγώνων απόκλισης από τη μέση τιμή της κάθε ομάδας (SS_{within} ή SS_{error}), που αντικατοπτρίζει τις αποκλίσεις εντός των ομάδων και αποτελεί την ανερμήνευτη διακύμανση, λόγω της τυχαιότητας της δειγματοληψίας των παρατηρήσεων και του αθροίσματος των τετραγώνων απόκλισης από τη μέση τιμή του μέσου των ομάδων ($SS_{between}$ ή SS_{effect}), που αντικατοπτρίζει τις αποκλίσεις μεταξύ των διαφόρων ομάδων και το οποίο εξαρτάται άμεσα από τη διαφορά των μέσων των ομάδων. Με άλλα λόγια, η κατηγοριοποίηση σε ομάδες εξηγεί αυτή τη μεταβλητότητα, επειδή είναι γνωστό πως οφείλεται στη διαφορά μεταξύ των μέσων των ομάδων. Αντίστοιχα, τα δυο παραπάνω αθροίσματα (SS_{within} και $SS_{between}$)

μπορούν να διαιρεθούν με τους αντίστοιχους βαθμούς ελευθερίας και να προκύψουν χρήσιμοι δείκτες που να εκφράσουν τη μεταβλητότητα. Από το SS_{within} προκύπτει ο δείκτης MSE_{within} (mean square effect), που εκφράζει την (ανεξήγητη) μεταβλητότητα εσωτερικά των ομάδων και από το $SS_{between}$, το $MSE_{between}$, που εκφράζει τη μεταβλητότητα μεταξύ των ομάδων. Η σύγκριση, δηλαδή, ο λόγος των παραπάνω δεικτών είναι υψίστης σημασίας.

Στην Ανάλυση Διακύμανσης γίνεται η μηδενική υπόθεση, η οποία εκφράζεται με H_0, πως ο μέσος όρος των επιμέρους ομάδων είναι ο ίδιος. Αν συμβαίνει κάτι τέτοιο, αναμένεται πως οι παραπάνω δείκτες ($MSE_{between}$ και MSE_{within}, οι οποίοι επειδή είναι υπεύθυνοι για τη συνολική διακύμανση λέγονται και εκτιμητές) αναμένεται να είναι σχεδόν ίδιοι. Όταν δεν ισχύει η μηδενικοί υπόθεση, το $MSE_{between}$ θα είναι μεγαλύτερο από το MSE_{within}. Στο συγκεκριμένο παράδειγμα το MSE_{within}=1 και το $MSE_{between}$=24. Άρα ο λόγος των δεικτών είναι 24. Πρέπει, όμως να γίνει κατανοητό πως η ύπαρξη του τυχαίου σφάλματος που οφείλεται στη δειγματοληψία δεν επιτρέπει στα $MSE_{between}$ και MSE_{within} να είναι ίσα ακόμη και στην περίπτωση που η μηδενική υπόθεση είναι αληθινή. Για το λόγο αυτό θα πρέπει να έχουμε ένα μέτρο ανοχής για το πόσο μεγάλη θα πρέπει να είναι η παρατηρούμενη διαφορά προκειμένου να συμπεράνουμε ότι δεν οφείλεται μόνο σε τυχαίο σφάλμα. Απάντηση στο ερώτημα αυτό μας δίνει η κατανομή δειγματοληψίας του λόγου των διακυμάνσεων $MSE_{between}/MSE_{within}$. Προσδιορίζεται, λοιπόν, μια κρίσιμη τιμή F (βάσει του ότι ο λόγος ακολουθεί μια κατανομή F), η οποία εξαρτάται από την κατανομή F και από το επίπεδο σημαντικότητας, η οποία καθορίζει τα διαστήματα αποδοχής και αν ξεπερνιέται από το λόγο των MSE να απορρίπτεται η μηδενική υπόθεση. Στο συγκεκριμένο παράδειγμα, η κρίσιμη τιμή F είναι ίση με 7,71. Άρα προκύπτει πως η μηδενική υπόθεση δεν ισχύει και η μέση τιμή δεν είναι ίδια για όλα τις ομάδες. Αυτό μπορεί να οφείλεται ότι οι ομάδες μπορεί να προέρχονται από διαφορετικές κατανομές. Συνοψίζοντας, μπορούμε να πούμε πως η ANOVA είναι μια στατιστική μέθοδος ανάλυσης της διακύμανσης μεταξύ μέσων τιμών ομάδων. Αυτό γίνεται χωρίζοντας τη συνολική διακύμανση σε δύο συνιστώσες, μία για να περιλάβει το τυχαίο σφάλμα και μια η οποία οφείλεται στις διαφορές των μέσων των ομάδων. Η τελευταία εξετάζεται για να βρεθεί η συνεισφορά της στη συνολική διακύμανση. Αν είναι σημαντική, τότε απορρίπτεται η μηδενική υπόθεση, ότι δηλαδή, οι μέσες τιμές των ομάδων είναι ίδιες και γίνεται αποδεκτή η αντίθετη, πως δηλαδή, οι ομάδες προέρχονται από διαφορετικές κατανομές.

Είναι, όμως, πιθανό να πρέπει να μελετηθεί η επίδραση περισσοτέρων παραγόντων (factors) στις παρατηρήσεις. Δηλαδή, να πρέπει όχι μόνο να γίνει ομαδοποίηση κατά ένα χαρακτηριστικό σε ομάδες, αλλά σε περισσότερα. Σε αυτή την περίπτωση έχουμε ANOVA δύο παραγόντων (two-factor ANOVA), το οποίο αποτελεί προέκταση της προηγούμενης μεθοδολογίας και τα παραπάνω βήματα δεν εκτελούνται πλέον, μόνο, για ένα παράγοντα (για στήλες), αλλά και για ένα επιπρόσθετο παράγοντα (κατά γραμμές). (Περισσότερες πληροφορίες στους παρακάτω δικτυακούς συνδέσμους: http://www.statsoft.com/textbook/anova-manova, *infoman.teikav.edu.gr/e_education/118/files/ANOVA.doc*)

Κεφάλαιο 5. Ανάπτυξη μοντέλων LCA

5.1 Προετοιμασία των δεδομένων

Για την ανάπτυξη των μοντέλων Συνεισφοράς Ομάδων GWP, CED και EI 99, χρησιμοποιήθηκαν πειραματικές και υπολογισμένες τιμές των παραπάνω δεικτών από το ήδη υπάρχον εργαλείο Finechem tool (επειδή δεν ήταν πάντα διαθέσιμες πειραματικές τιμές). Για να αναπτυχθούν μοντέλα GC, πρέπει να δίνεται η είσοδος στο μοντέλο με τη μορφή των περιγραφέων (descriptors) που απαιτούνται από τα GC μοντέλα, δηλαδή τις ομάδες. Χρησιμοποιούνται οι 106 ομάδες που χρησιμοποιούνται στα GC μοντέλα για τον υπολογισμό των ιδιοτήτων στη μεθοδολογία CAMD των Marcoulaki & Kokossis (2000). Κάθε μόριο, λοιπόν, από τις βάσεις δεδομένων που χρησιμοποιούνται για το σχηματισμό του μοντέλου, έχει ως χαρακτηριστικά τις τρεις τιμές των παραπάνω δεικτών και ένα διάνυσμα, 1x106, που το κάθε στοιχείο του είναι ένας ακέραιος αριθμός, ο οποίος εκφράζει τον αριθμό των φορών που εμφανίζεται αυτή η ομάδα στο μόριο. Η ανάπτυξη των μοντέλων γίνεται με απλές παλινδρομήσεις, στις μεθόδους που περιγράφηκαν παραπάνω και με μη γραμμικά μοντέλα στις υπόλοιπες περιπτώσεις.

Χρησιμοποιήθηκαν δυο βάσεις δεδομένων: η πρώτη αποτελείται από 339 οργανικά μόρια, όλων των κατηγοριών (κυκλικά, αλειφατικά και αρωματικά) και από όλες τις χαρακτηριστικές ομάδες και δεν περιλαμβάνει πειραματικά μετρημένες τιμές για τα μόρια, παρά υπολογισμένες τιμές. Ακολούθησε η καταγραφή των μορίων που δε μπορούν να εκφραστούν και να περιγραφούν με τις συγκεκριμένα ομάδες τα οποία χρησιμοποιούνται από τη μεθοδολογία CAMD. Τέτοια μόρια είναι, για παράδειγμα, όσα περιέχουν άτομα φθορίου ή θείου ή όσα έχουν ένα κυάνιο ενωμένο με ένα αρωματικό δακτύλιο. Τα παραπάνω μόρια απορρίφθηκαν από αυτή τη βάση δεδομένων, διότι είναι αδύνατος ο σχηματισμός του παραπάνω διανύσματος αριθμών, το οποίο θα δοθεί ως είσοδος. Επιπλέον, απορρίφθηκαν από τη διαδικασία και μόρια, που δεν ήταν γνωστή ή δοσμένη η δομή τους σε αυτή τη βάση δεδομένων με αποτέλεσμα να είναι, επίσης, αδύνατος ο σχηματισμός του παραπάνω διανύσματος. Στα υπόλοιπα μόρια έγινε κανονικά η αποδόμηση τους σε ομάδες και ο σχηματισμός των διανυσμάτων που περιγράφουν τον αριθμό των επαναλήψεων της κάθε ομάδας στο μόριο. Τελικά, προέκυψαν 101 μόρια, έτοιμα για την είσοδό τους στο μοντέλο.

Χρησιμοποιήθηκε, ακόμα, και ένα τμήμα της βάσης δεδομένων Ecoinvent που περιλαμβάνει 290 μετρήσεις για 190 διαφορετικά μόρια όλων των ειδών (αλειφατικά, αρωματικά και κυκλικά) και όλων των χαρακτηριστικών ομάδων. Ο λόγος που υπάρχουν παραπάνω μετρήσεις για κάποια μόρια είναι επειδή, συμπεριλαμβάνονται και πειραματικές μετρήσεις για τους τρεις δείκτες, όχι μόνο για τα ευρωπαϊκά δεδομένα, αλλά και για δεδομένα που έχουν ληφθεί από χώρες που θεωρούνται ως ξεχωριστές κατηγορίες και συγκεκριμένα, από την Ελβετία, την Ολλανδία, τη Δυτική Ευρώπη, τη Σουηδία, αλλά και χώρες εκτός Ευρώπης, δηλαδή, τη Βραζιλία, την Κίνα και την Αμερική. Τα μόρια που επαναλαμβάνονται πάνω από δυο

φορές στη βάση δεδομένων έχουν, ασφαλώς, διαφορετικές τιμές για τους τρεις δείκτες (όχι όμως με μεγάλες διαφορές).

Από την παραπάνω λίστα απορρίπτονται μόρια, τα οποία, επίσης, δεν είναι δυνατόν, να περιγραφούν από τους 106 ομάδες του CAMD εργαλείου. Τα υπόλοιπα μόρια αποδομούνται στις ομάδες του εργαλείου των Marcoulaki & Kokossis (2000) και προκύπτουν τα διανύσματα που περιγράφουν τον αριθμό των επαναλήψεων των ομάδων. Έπειτα, απορρίπτονται, αν είναι δυνατόν, μόρια τα οποία επαναλαμβάνονται πάνω από δύο φορές. Παρολαυτά, ένα μόριο που υπάρχει πάνω από μια φορές μπορεί να παραμείνει αν υπάρχουν πάνω από ένας, έγκυροι τρόποι για να περιγραφεί ένα μόριο από περισσότερους συνδυασμούς των ομάδων. Για παράδειγμα, η ένωση τριχλωροπροπάνιο, μπορεί να περιγραφεί από τον συνδυασμό ομάδων [-CH_2Cl,CHCl], αλλά και από το συνδυασμό των ομάδων [$C(Cl)_3$,CH_3,CH_2] και [CH_3,$CH(Cl)_2$,CHCl]. Οι παραπάνω συνδυασμοί δίνουν, ασφαλώς, τις ίδιες τιμές για τους τρεις δείκτες και μπορεί να αποτελούν ισοδύναμες διαμορφώσεις για την ίδια ένωση, αλλά για το μοντέλο θα αποτελέσουν διαφορετικές εισόδους και θα συντελέσουν στην καλύτερη και πιο σφαιρική και ολοκληρωμένη διαμόρφωση των τιμών της συνεισφοράς των ομάδων. Γι' αυτό και θα συμπεριληφθούν και οι τρεις στην είσοδο των δεδομένων εισόδου. Όσον αφορά στο παραπάνω παράδειγμα, αν δίναμε στο μοντέλο, ως είσοδο μόνο τον πρώτο συνδυασμό, τότε οι τιμές των τριών δεικτών (output), θα επηρέαζαν τη διαμόρφωση των συνεισφορών μόνο των CH_2Cl και CHCl. Η παραπάνω κίνηση, δεν είναι λάθος, είναι απλά ανεπαρκής. Αφού, λοιπόν, και οι άλλοι συνδυασμοί είναι εξίσου πιθανοί, μπορούν και πρέπει να συνεισφέρουν οι τιμές της εξόδου και στις συνεισφορές των υπολοίπων ομάδων.

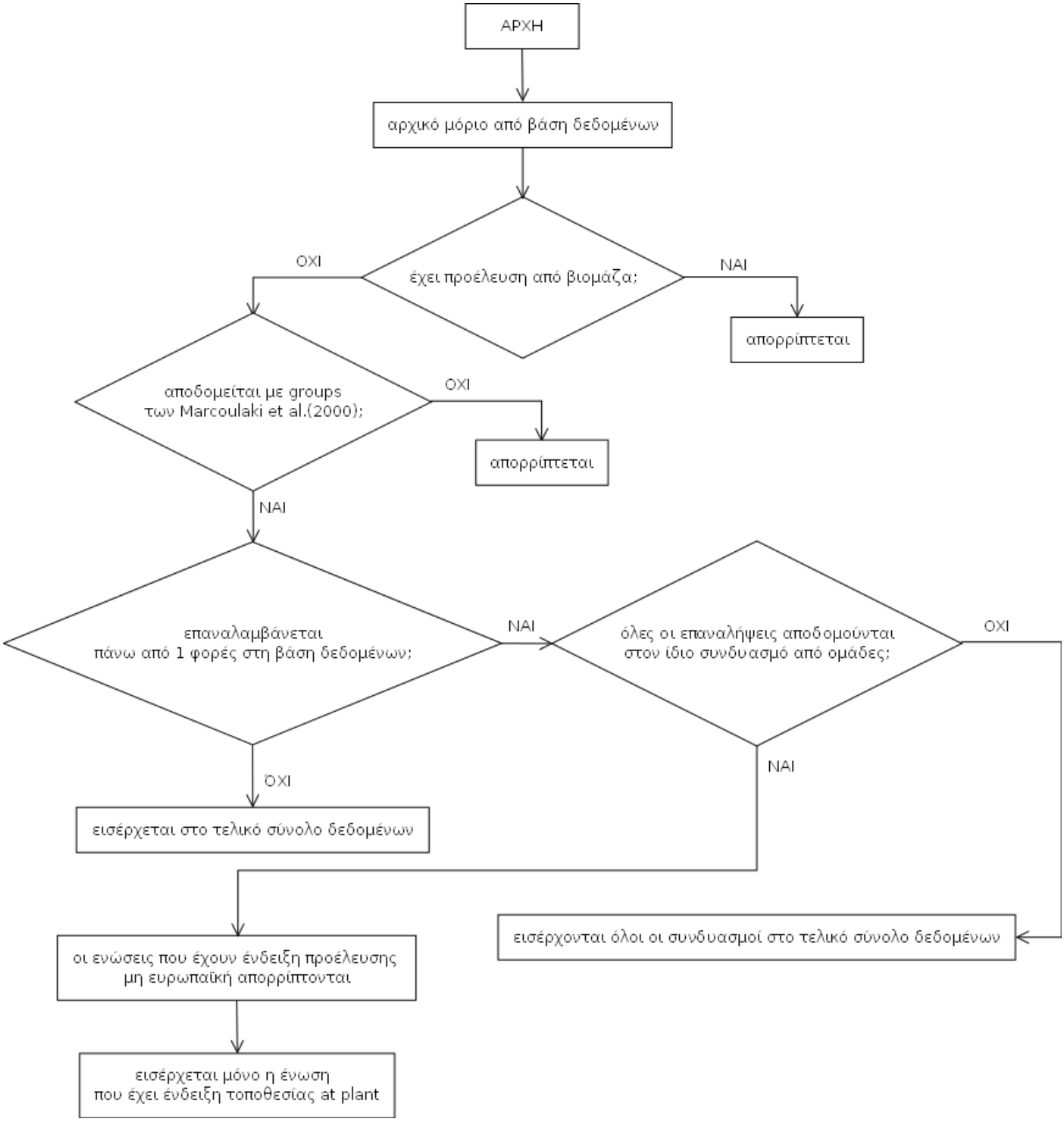
Το πρώτο κριτήριο, λοιπόν, όταν συναντάται ένα μόριο με πολλαπλές εμφανίσεις είναι να ελεγχθεί αν όλες αποδομούνται στον ίδιο συνδυασμό ομάδων. Αυτές που έχουν διαφορετικό συνδυασμό παραμένουν. Μεταξύ, όμως, των εμφανίσεων που έχουν τον ίδιο συνδυασμό, επιλέγεται να παραμείνουν αυτές που οι μετρήσεις τους προέρχονται από ευρωπαϊκά δεδομένα (αυτά τα δεδομένα για τα οποία έχουμε την ένδειξη RER). Για οποιεσδήποτε άλλες εμφανίσεις, του ίδιου μορίου, με τον ίδιο συνδυασμό ομάδων, των οποίων οι ενδείξεις διαφέρουν από την ένδειξη της «Ευρώπης» (έστω και αν έχουν ενδείξεις όπως CH, για χώρες ,όπως η Ελβετία, που υπάγονται στην Ευρώπη) απορρίπτονται. Αυτό συμβαίνει επειδή είναι επιθυμητό να υπάρχει όσο το δυνατόν μεγαλύτερη ομοιομορφία στο τρόπο μέτρησης και στην πηγή προέλευσης των δεδομένων. Κάτι τέτοιο δεν είναι γενικά δυνατόν, διότι για μερικές ενώσεις έχουμε δεδομένα, τα οποία έχουν ,ούτως ή άλλως, προέλευση μόνο εκτός Ευρώπης. Σε μια ιδανική περίπτωση, θα έπρεπε να διαγραφούν, όλες οι ενώσεις που τα δεδομένα τους έχουν προέλευση μη ευρωπαϊκή. Τα δεδομένα, όμως είναι αρκετά λίγα εξ αρχής και συνεπώς, θα ήταν ανεπαρκή για την ανάπτυξη ενός μοντέλου.

Έπειτα, το δεύτερο κριτήριο για τη διαγραφή μιας ουσίας (με την προϋπόθεση πως εμφανίζεται διπλά και έχει τον ίδιο συνδυασμό από ομάδων) είναι το στάδιο του κύκλου ζωής για το οποίο έχει γίνει η

μέτρηση. Υπάρχουν δυο στάδια του κύκλου ζωής για τα οποία γίνεται μια μέτρηση των δεικτών: μέτρηση στην παραγωγική διαδικασία και μέτρηση στην περιφερειακή αποθήκη (regional storehouse). Επιλέγεται μόνον η μέτρηση για την παραγωγική διαδικασία και οι τιμές που αναφέρονται στην αποθήκη απορρίπτονται. Ο λόγος γι' αυτό είναι, όπως και πριν, η ομοιογένεια των δεδομένων. Εν κατακλείδι, περιγράφηκαν οι λόγοι για τους οποίους ένα μόριο μπορεί να εμφανίζεται περισσότερες από μια φορές στη βάση δεδομένων (περισσότεροι συνδυασμοί ομάδων, πολλαπλές μετρήσεις από διαφορετικές πηγές και μέτρηση σε πολλαπλά σημεία του κύκλου ζωής) και περιγράφηκαν οι τρόποι ιεράρχησης για τη διαγραφή των περιττών δεδομένων. Επίσης, απορρίπτονται και όσα μόρια προέρχονται από επεξεργασία βιομάζας.

Τελικά συγκεντρώνονται 113 μόρια από τη βάση δεδομένων Ecoinvent. Επίσης, υπάρχουν αρκετές ενώσεις που είναι κοινές και στις δυο βάσεις δεδομένων. Διαγράφονται, λοιπόν, από την πρώτη βάση δεδομένων (με τα υπολογισμένα δεδομένα) και κρατούνται οι ενώσεις και οι τιμές από την Ecoinvent. Ο λόγος που προτιμάται να κρατηθούν τα δεδομένα της Ecoinvent, είναι επειδή ανταποκρίνονται σε πειραματικά δεδομένα και όχι σε υπολογισμούς (ο τελικός αριθμός μορίων 101 που αναφέρεται στην προηγούμενη βάση δεδομένων προκύπτει μετά την αφαίρεση των κοινών ενώσεων). Να σημειωθεί πως μερικά από τα μόρια της βάσης δεδομένων που χρησιμοποιείται μπορεί να περιγράφονται από μια μόνο ομάδα. Αυτό συμβαίνει, επειδή, μερικές ομάδες αποτελούν αυτόνομα μόρια (όπως η μεθυλαμίνη, CH_3-NH_3) με σθένος=0. Όπως γίνεται κατανοητό, γίνεται εξαρχής αποδεκτό στην κατασκευή των μοντέλων που θα ακολουθήσουν πως θα ενσωματώνουν και το σφάλμα των δεδομένων, που είναι υπολογισμένων. Δηλαδή, στα μοντέλα που θα αναπτυχθούν, το τελικό σφάλμα θα συμπεριλαμβάνει το σφάλμα της μοντελοποίησης και το σφάλμα από τα πειραματικά δεδομένα, που υπάρχει λόγω του ότι χρησιμοποιήθηκαν υπολογισμένα και όχι πειραματικά δεδομένα (ο υπολογισμός έχει γίνει με ήδη υπάρχον εργαλείο που προβλέπει δείκτες LCA και ονομάζεται Finechem,Wernet et al., 2009). Να σημειωθεί πως για να έχει όσο το δυνατόν μεγαλύτερη συσχέτιση το μοντέλο, πρέπει ο λόγος μορίων προς περιγραφείς (groups) να είναι όσο το δυνατόν μεγαλύτερος.

Τα βήματα επιλογής μορίων από τις βάσης δεδομένων συνοψίζονται στο σχήμα 5.1. (όπου τελικό σύνολο δεδομένων είναι το τελικό σύνολο των 214 μορίων).



Σχήμα 5.1. Βήματα επιλογής μορίων προς ανάπτυξη μοντέλων

Αφού καθοριστούν τα μόρια που θα συμμετάσχουν, παρατηρείται πως αρκετές από τις ομάδες που είναι διαθέσιμες για την περιγραφή των διάφορων μορίων παραμένουν αχρησιμοποίητες (η τιμή στο διάνυσμα που δείχνει πόσες φορές εμφανίζονται στο μόριο είναι συστηματικά 0). Αυτές οι ομάδες δε θα χρησιμοποιηθούν από το μοντέλο καθόλου και γι' αυτό διαγράφονται. Το πλήθος των 106 ομάδων μειώνεται σε 57. Να σημειωθεί πως σε αυτές τις ομάδες έχει προστεθεί και το AC-Cl, δηλαδή άτομο χλωρίου συνδεδεμένο με αρωματικό άνθρακα, το οποίο δε συμπεριλαμβανόταν στο αρχικό σύνολο από ομάδες. Είναι, όμως, αρκετά χρήσιμο, μιας και πολλά μόρια του συνόλου που χρησιμοποιούνται το χρησιμοποιούν. Συνδυάζοντας τα μοριακά διανύσματα (1 x 57 πλέον) με τα 214 μόρια και εκφράζοντας τα

σε μορφή πίνακα προκύπτει ο πίνακας της εισόδου (input) στο μοντέλο, δηλαδή 214 x 57 (όπου η κάθε γραμμή αντιστοιχεί στο μοριακό διάνυσμα ενός μορίου και η κάθε στήλη αντιστοιχεί στις επαναλήψεις της κάθε ομάδας). Θα μπορούσαμε να χρησιμοποιήσουμε ως παράδειγμα μοριακού διανύσματος το προπάνιο: λαμβάνοντας υπόψη πως οι ομάδες $–CH_3$ και $>CH_2$ στις οποίες αποδομείται είναι οι δύο πρώτες με αυτή τη σειρά στη λίστα των ομάδων που χρησιμοποιούνται τελικά, το μοριακό διάνυσμα θα είναι:

[2 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

Αφού οι μέθοδοι που είναι επιθυμητό να αναπτυχτούν είναι GC, αναμένεται το αποτέλεσμα του κάθε μοντέλου να είναι ένας πίνακας με 57 συνεισφορές και αφού έχουμε την ανάπτυξη απλών γραμμικών μοντέλων έχουμε και τη σταθερά του γραμμικού μοντέλου (intercept). Άρα αναμένεται ένα διάνυσμα 1x58. Το γραμμικό μοντέλο που θα αναπτυχθεί θα είναι:

$$GWP = contr_{GWP\_1} \cdot N_1 + \cdots + contr_{GWP\_57} \cdot N_{57} + intercept_{GWP} \quad (5.1)$$

$$CED = contr_{CED\_1} \cdot N_1 + \cdots + contr_{CED\_57} \cdot N_{57} + intercept_{CED} \quad (5.2)$$

$$EI = contr_{EI\_1} \cdot N_1 + \cdots + contr_{EI\_57} \cdot N_{57} + intercept_{EI} \quad (5.3)$$

όπου N_i, ο αριθμός επαναλήψεων του κάθε ομάδας και $contr_i$, η συνεισφορά της κάθε ομάδας στην για την εκάστοτε ιδιότητα.

Ο πίνακας της εξόδου (output) είναι, αντίστοιχα, ένας πίνακας 214 x 3. Η κάθε στήλη περιέχει τις υπολογισμένες τιμές των τριών δεικτών για 101 πρώτα μόρια και τις πειραματικές τιμές για τα υπόλοιπα. Ο λόγος, λοιπόν, μορίων προς ομάδων στη συγκεκριμένη περίπτωση είναι 214/57=3.75. Ο λόγος αυτός κρίνεται αρκετά μικρός (θα έπρεπε να τουλάχιστον 10), κάτι που εξηγείται από την έλλειψη πειραματικών δεδομένων, που με τη σειρά του οφείλεται στη δυσκολία και χρονοβόρα διαδικασία συγκέντρωση πειραματικών δεδομένων. Στο παράρτημα Β φαίνονται οι ομάδες που συμμετέχουν στο αρχικό εργαλείο. Με αστερίσκο σημειώνονται οι ομάδες που απέμειναν μετά την αφαίρεση.

Οι ενώσεις που συμμετέχουν στο αρχικό σύνολο με τα πειραματικά (και υπολογισμένα) δεδομένα τους φαίνονται στο παράρτημα Γ (αναφέρονται μόνο οι 113 ενώσεις της βάσης δεδομένων Ecoinvent).

5.2 Διαδικασία πριν την ανάπτυξη των μοντέλων

Αν στη δημιουργία του μοντέλου συμπεριληφθούν και τα 214 μόρια, τότε θα είναι αδύνατον να δοκιμάσουμε τις δυνατότητες του και τις προβλεπτικές του ικανότητες σε νέα μόρια. Είναι αυτονόητο, πως είναι λάθος να δοκιμαστεί η ικανότητα ενός μοντέλου στην πρόβλεψη με τα ίδια μόρια, τα οποία συμμετείχαν στην εκάστοτε παλινδρόμηση (ασφαλώς, όμως, λαμβάνονται υπόψη και αυτές οι προβλέψεις).

Είναι, λοιπόν, αναγκαίο να γίνει η δημιουργία του μοντέλου με ακόμα λιγότερα δεδομένα, να αφαιρεθούν κάποια μόρια, δηλαδή, από την αρχική βάση δεδομένων και το μοντέλο, αφού δημιουργηθεί, να δοκιμαστεί με αυτά. Να σημειωθεί πως αφαιρώντας επιπλέον μόρια από τη βάση δεδομένων για την επικύρωση του μοντέλου, επιδεινώνεται ο παραπάνω λόγος παρατηρήσεων (μορίων)/περιγραφείς, ο οποίος μπορεί να έχει ξεκάθαρη επίδραση πάνω στην αξιοπιστία του μοντέλου. Κρίνεται πρακτικό να αφαιρεθεί το 20% (43 μόρια) των παρατηρήσεων (μορίων) και η όποια παλινδρόμηση για τη δημιουργία του συσχετισμού να γίνει με το υπόλοιπο 80% (171 μόρια). Το 20% των μορίων θα χρησιμοποιηθεί για τη δοκιμασία των προβλεπτικών ικανοτήτων του μοντέλου. Το μεγαλύτερο κομμάτι, δηλαδή, το σύνολο 171 μορίων ονομάζονται και θα αποκαλούνται στο εξής, σύνολο εκπαίδευσης (training set). Το σύνολο των 43 μορίων ονομάζεται και θα αποκαλείται στο εξής σύνολο δοκιμής (testing set). Η διαδικασία όπου χωρίζεται σε σύνολο εκπαίδευσης και δοκιμής το αρχικό σύνολο ονομάζεται διαμέριση ή διχοτόμηση (partition). Το ποσοστό στο οποίο θα διαχωριστεί το αρχικό σύνολο είναι αυθαίρετη επιλογή. Θεωρήθηκε, δηλαδή, πως η διαμέριση 80-20 των μορίων από το αρχικό σύνολο δεδομένων είναι ένας αρκετά καλός καταμερισμός: όχι, δηλαδή, πολύ λίγα μόρια στο σύνολο εκπαίδευσης ώστε να μην είναι αξιόπιστο το μοντέλο, αλλά ούτε και πολύ λίγα μόρια στο σύνολο δοκιμής, έτσι ώστε να μην εξετάζονται όλες οι πτυχές του. Στη είσοδο του στατιστικού εργαλείου για την ανάπτυξη του μοντέλου δε θα δοθεί ένας πίνακας, λοιπόν, 214x57 (input) και 214x3 (output), αλλά 171x57 (input) και 171x3 (output). Να σημειωθεί πως όταν έχουμε περιπτώσεις μορίων που αναπαρίστανται μόνο από μια ομάδα, η σειρά εκείνη του πίνακα εισόδου (214x57) που αντιστοιχεί σε αυτό το μόριο θα έχει όλα τα στοιχεία της ίσα με 0 και ένα μόνο στοιχείο της (αυτό που αντιστοιχεί στην ομάδα) θα είναι ίσο με 1. Όλες οι παραπάνω διαδικασίες και η ανάπτυξη του μοντέλου γίνονται σε περιβάλλον MATLAB.

Αν είναι επιθυμητό να μη γίνει η αφαίρεση ενός πλήθους από μόρια, τότε μπορεί να αφαιρεθεί ένα μόριο μόνο για να δοκιμαστεί από το μοντέλο (leave-one-out validation). Η συγκεκριμένη τεχνική δεν κρίθηκε πρακτικό να ακολουθηθεί στη συγκεκριμένη περίπτωση, καθώς ένα μόνο μόριο ίσως να μη δώσει μια πλήρη εικόνα για τις προβλεπτικές ικανότητες του μοντέλου σε όλες τις κατηγορίες μορίων (αλκοόλες, οξέα, κετόνες, υδρογονάνθρακες κ.τ.λ).

Είναι φανερό, πως ο τρόπος με τον οποίο γίνεται η κατανομή των μορίων στο σύνολο εκπαίδευσης και δοκιμής επηρεάζει και τα αποτελέσματα του μοντέλου. Με άλλα λόγια, ας υποθέσουμε πως κάποια μόρια ανατίθενται στο σύνολο εκπαίδευσης και τα υπόλοιπα στο σύνολο δοκιμής, εξάγονται κάποια αποτελέσματα και μετά επαναλαμβάνεται η παραπάνω διαδικασία, αυτή τη φορά, όμως, ο καταμερισμός γίνεται διαφορετικά και το σύνολο εκπαίδευσης περιέχει ένα διαφορετικό σύνολο μορίων και αντίστοιχα, το σύνολο δοκιμής. Ο συνδυασμός, λοιπόν, έχει αλλάξει. Τα αποτελέσματα που θα ληφθούν θα είναι κι αυτά διαφορετικά. Με την ίδια διαδικασία μπορεί να εξαχθεί ένας τεράστιος αριθμός διαφορετικών συνδυασμών στο χωρισμό των μορίων σε σύνολο δοκιμής και εκπαίδευσης. Δεν υπάρχει, λοιπόν, κάποιος κανόνας ή κάποιο μοτίβο πάνω στο οποίο γίνεται ο καταμερισμός στα δύο σύνολα και ο χωρισμός στηρίζεται καθαρά

στην τύχη, κάτι που λειτουργεί δυσμενώς για το μοντέλο, διότι το κάνει να λειτουργεί υποκειμενικά. Έτσι, είναι αναγκαίο να εξαλειφθεί αυτός ο παράγοντας της τυχαιότητας. Για να γίνει αυτό, δε στηρίζεται η δημιουργία του μοντέλου σε ένα μόνο διαχωρισμό, αλλά σε 1000 διαφορετικούς. Δηλαδή, δημιουργούνται, 1000 τυχαίοι διαμερισμοί (1000 σύνολα εκπαίδευσης και αντίστοιχα 1000 σύνολα δοκιμής) και συνεπώς, δημιουργούνται 1000 διαφορετικά μοντέλα (άρα και 1000 διαφορετικά διανύσματα 1x58). Για την κάθε μια από τις 58 παραμέτρους που προκύπτουν στο τέλος, λαμβάνεται η μέση τιμή που έχει και στους 1000 διαμερισμούς. Με αυτόν τον τρόπο εξαλείφεται η υποκειμενικότητα του μοντέλου και δε συνεισφέρει ένας αλλά χίλιοι διαφορετικοί συνδυασμοί ισάξια στη διαμόρφωση του μοντέλου. Έτσι, αυτό που προκύπτει είναι ένας αντικειμενικός τρόπος υπολογισμού ιδιοτήτων. Ο αριθμός 1000 είναι και αυτός αυθαίρετος: θεωρείται ένα καλό πλήθος συνόλων δεδομένων, όπου αν ξεπεραστεί, θα συντελέσει σε μεγάλους υπολογιστικούς χρόνους (πάνω από 4 ημέρες), κυρίως για πολύπλοκες διαδικασίες, όπως την παρεμβολή τύπου «kriging».

Είναι, επίσης, απαραίτητο, να βρεθούν για κάθε σύνολο δοκιμής από τα χίλια, ποια μόρια θεωρούνται ως προς το αντίστοιχο σύνολο εκπαίδευσης, παρεμβολές (interpolations) ή προεκβολές (extrapolations). Το σε ποια από τις παραπάνω κατηγορίες εμπίπτει ένα μόριο εξαρτάται καθαρά από τις ομάδες στους οποίους αποδομείται. Συγκρίνονται, λοιπόν, οι επαναλήψεις των ομάδων των μορίων του συνόλου δοκιμής με αυτές του αντίστοιχου συνόλου εκπαίδευσης. Ο λόγος για τον οποίο είναι επιθυμητό να καθοριστεί αυτό είναι έτσι ώστε αν κατά τη διαδικασία επικύρωσης του μοντέλου, κάποια μόρια του συνόλου δοκιμής φανεί να έχουν μεγαλύτερο σφάλμα, αυτό να μπορεί να δικαιολογηθεί από το γεγονός πως αποτελούν προεκβολές του συνόλου εκπαίδευσης και όχι γενικά, πως έχει δημιουργηθεί ένα αναξιόπιστο μοντέλο. Για να εκφράσουμε το αν ένα μόριο αποτελεί παρεμβολή ή προεκβολή χρησιμοποιείται η έννοια της απόστασης που έχει από το μέσο του συνόλου του συνόλου εκπαίδευσης.

Η απόσταση που χρησιμοποιείται, στη συγκεκριμένη περίπτωση, ονομάζεται απόσταση mahalanobis (Maesschalck et al., 2000). Η μέθοδος αυτή εντοπίζει την απόσταση ενός σημείου από ένα σύνολο διεσπαρμένων δεδομένων και την εκφράζει ως πολλαπλάσιο της διακύμανσης. Ας θεωρήσουμε πως έχουμε ένα σύνολο δεδομένων, κατανεμημένα γύρω από μια μέση τιμή. Η διακύμανση θα εκφράζει τα «σύνορα» της περιοχής του συνόλου δεδομένων, κάτι που είναι λογικό αφού η διακύμανση εκφράζει την κατανομή του συνόλου γύρω από το μέσο όρο. Υπολογίζεται, έτσι, με στατιστικό τρόπο ένα είδος απόστασης από τη μέση τιμή. Αν, λοιπόν, ένα σημείο απέχει περισσότερες φορές από μια διακύμανση, σημαίνει πως είναι έξω από το σύνορο του συνόλου και αποτελεί ένα εξωτερικό στοιχείο του συνόλου εκπαίδευσης, που στο εξής θα ονομάζεται ακραία παρατήρηση (outlier), ενώ αν η απόσταση είναι υποπολλαπλάσια της διακύμανσης θα βρίσκεται μέσα στο σύνολο και θα αποτελεί παρεμβολή. Κάθε ένα από τα 43 μόρια του κάθε συνόλου δοκιμής θα συνοδεύεται από μια τιμή που θα αντικατοπτρίζει την απόσταση του από το αντίστοιχο σύνολο εκπαίδευσης. Οι τιμές που προέκυψαν ως αποστάσεις δεν ξεπερνούσαν στις περισσότερες περιπτώσεις το 10. Παρολαυτά, υπήρχαν αρκετές περιπτώσεις που η απόσταση του ξεπερνούσε το 10000. Παρατηρώντας

πως δεν υπάρχει μια κατανομή μεταξύ των αποστάσεων της πρώτης και της δεύτερης κατηγορίας (δηλαδή, δεν κατανέμονται ίσα σε όλο το φάσμα οι αποστάσεις των στοιχείων, παρά προκύπτουν ή πολύ μικρές ή πολύ μεγάλες αποστάσεις) φαίνεται ξεκάθαρα πως οι μεγάλες τιμές αντιστοιχούν σε ακραίες παρατηρήσεις. Οι παραπάνω ακραίες παρατηρήσεις σημειώνονται για περαιτέρω ανάλυση μετά τη δημιουργία των μοντέλων. (Βλ. Παράρτημα Z για περισσότερες πληροφορίες)

5.3 Διαδικασία μετά την ανάπτυξη του μοντέλου

Αφού γίνει η συσχέτιση και αναπτυχθούν συσχετισμοί, παράγεται το διάνυσμα 1x58 που περιέχει τις συνεισφορές και τη σταθερά του γραμμικού μοντέλου (intercept). Χρησιμοποιείται το παραπάνω μοντέλο και γίνονται οι προβλέψεις των τριών δεικτών για το κάθε σύνολο εκπαίδευσης και δοκιμής. Έπειτα, υπολογίζεται ένα σύνολο από 15 στατιστικούς δείκτες, οι οποίοι για τον κάθε διαμερισμό, εξετάζουν τις συσχετίσεις μεταξύ των πειραματικών και υπολογισμένων δεδομένων. Αυτοί οι δείκτες αναφέρονται στο Παράρτημα ΣΤ. Προκύπτουν, λοιπόν, για κάθε διαμερισμό δυο σύνολα, που το κάθε ένα περιέχει τους 15 στατιστικούς δείκτες. Το ένα σύνολο περιλαμβάνει τους δείκτες που αναφέρονται στη σχέση των πειραματικών τιμών με τις υπολογισμένες από το μοντέλο τιμές για το σύνολο εκπαίδευσης και το άλλο για το σύνολο δοκιμής. Να σημειωθεί πως κάποιοι δείκτες λαμβάνουν και αρνητικές τιμές όταν αναφέρονται στο σύνολο δοκιμής. Αυτοί είναι η Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων και ο Συντελεστής Προσδιορισμού.

Επειδή, είναι σημαντικό να παρατηρηθεί η σχέση που έχει ο κάθε δείκτης στο σύνολο εκπαίδευσης και αντίστοιχα, στο σύνολο δοκιμής, δηλαδή να παρατηρηθεί κατά πόσο αυτός έχει μεγαλύτερη τιμή στο εκπαίδευσης από το δοκιμής ή πόσο κοντά βρίσκονται οι δυο τιμές, θεωρήθηκε πρακτικό να ληφθεί ο μεταξύ τους λόγος. Έτσι, για την κάθε διαμέριση δημιουργείται και ένα τρίτο σύνολο τιμών: αυτό περιλαμβάνει το λόγο του κάθε δείκτη στο σύνολο δοκιμής προς το λόγο του ίδιου δείκτη στο σύνολο εκπαίδευσης (π.χ. Μέσο Σχετικό Σφάλμα$_{testing}$/Μέσο Σχετικό Σφάλμα$_{training}$ κ.ο.κ.). Όπως έχει αναφερθεί για τις συνεισφορές των ομάδων, ο μέσος όρος των τιμών (και από τους 1000 διαμερισμούς) δίνει μια αντικειμενική εκτίμηση για τη συνεισφορά του κάθε ενός. Έτσι και σε αυτή την περίπτωση, για να ληφθεί μια συνολική εικόνα του κάθε δείκτη για το σύνολο εκπαίδευσης και δοκιμής, λαμβάνεται ο μέσος όρος του κάθε δείκτη και για τα 1000 σύνολα δοκιμής συνολικά. Για παράδειγμα, όσον αφορά σε ένα συγκεκριμένο δείκτη (π.χ. Μέσο Σχετικό Σφάλμα), συγκεντρώνονται 1000 τιμές αυτού του δείκτη που προκύπτουν από τα 1000 διαφορετικά σύνολα εκπαίδευσης (1000 διαφορετικές τιμές Μέσο Σχετικό Σφάλμα) και υπολογίζεται ο μέσος όρος. Το παραπάνω βήμα εκτελείται και για τους 15 δείκτες. Το ίδιο συμβαίνει και για τους 15 λόγους δεικτών μεταξύ του συνόλου εκπαίδευσης και δοκιμής. Τελικά, λοιπόν, προκύπτουν δύο σύνολα για το μοντέλο: το ένα περιέχει τους 15 (μέσους) στατιστικούς δείκτες για το σύνολο εκπαίδευσης, όπου ο κάθε

ένας δείκτης αποτελεί το μέσο όρο, όλων (και των 1000) τιμών που λαμβάνει σε κάθε διαμέριση και ένα σύνολο με τους λόγους της τιμής του κάθε ένα από τους 15 δείκτες στο σύνολο δοκιμής προς την τιμή του στο σύνολο εκπαίδευσης (ο μέσος λόγος του κάθε δείκτη αποτελεί μέση τιμή των 1000 τιμών που λαμβάνει σε όλους τους διαμερισμούς). Οι μέσες τιμές των δεικτών του συνόλου δοκιμής δεν υπολογίζονται άμεσα, όπως στο σύνολο εκπαίδευσης (λαμβάνοντας, δηλαδή, τις μέσες τιμές του κάθε δείκτη και στα 1000 διαμερισμούς), αλλά πολλαπλασιάζοντας το μέσο λόγο του κάθε δείκτη (τιμή δοκιμής/τιμή εκπαίδευσης) με τη μέση τιμή του ίδιου δείκτη στο σύνολο εκπαίδευσης. Έγινε η παραδοχή πως:

$$\overline{\left(\frac{\text{μέση τιμή δείκτη στο σύνολο δοκιμής}}{\text{μέση τιμή δείκτη στο σύνολο εκπαίδευσης}}\right)} \approx \frac{\overline{\text{μέση τιμή δείκτη στο σύνολο δοκιμής}}}{\overline{\text{μέση τιμή δείκτη στο σύνολο εκπαίδευσης}}} \quad (5.4)$$

δηλαδή η μέση τιμή του λόγου (κάθε δείκτη) είναι περίπου ίση με το λόγο των μέσων τιμών.

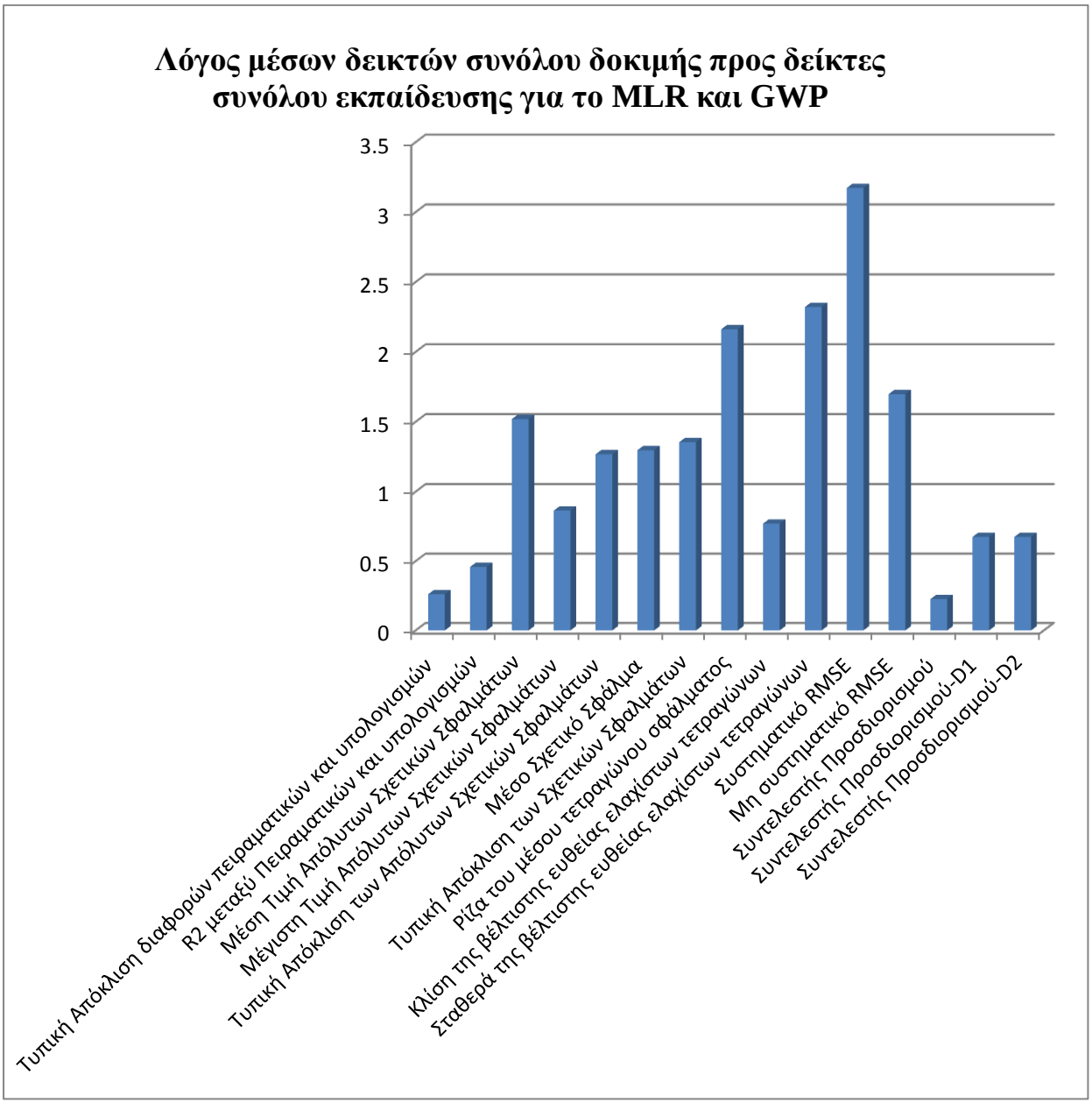
Να σημειωθεί πως, κατά τη διαμόρφωση των λόγων των τιμών των δεικτών, κάποιοι λόγοι δεικτών, που, όπως προαναφέρθηκε, μπορούν να πάρουν αρνητικές τιμές στο σύνολο δοκιμής, θα προκύπτουν αρνητικοί. Αυτό θα έχει σαν αποτέλεσμα σε κάποιες περιπτώσεις να προκύπτουν αρνητικοί μέσοι λόγοι στα τελικά σύνολα (αφού οι αρνητικές τιμές συμψηφίζονται με τις θετικές στην εύρεση του μέσου όρου) και συνεπώς, να μην είναι δυνατόν να αποφανθούμε σε καμία περίπτωση για το αν αυτό το μοντέλο έχει καλές ή όχι δυνατότητες. Επισκιάζονται, δηλαδή, από μεγάλες αρνητικές τιμές, τα οφέλη χρήσης του μοντέλου. Γι' αυτό και γίνεται αποδεκτό, αποκλειστικά για τους δύο αυτούς δείκτες να βρεθούν σε ποιες διαμερίσεις λαμβάνουν αρνητικές τιμές οι λόγοι τους και να τεθούν αυτόματα ίσοι με το 0. Το 0 είναι προφανώς η καλύτερη (μέγιστη) τιμή που μπορούμε να τους δώσουμε ώστε να μην πάρουν αρνητική τιμή. Το να μη συμμετάσχουν αρνητικές τιμές στο μέσο όρο είναι στοιχειώδες, ώστε κατά πρώτον να μη ληφθούν αρνητικές τελικές (μέσες) τιμές και κατά δεύτερον, να αναδειχθούν οι πραγματικές του δυνατότητες, μη επηρεαζόμενες από κάποια «κακά» διαμερίσεις, όπου μπορούν φαινομενικά τουλάχιστον, να υποβιβάσουν την αξιοπιστία του μοντέλου.

Να σημειωθεί πως ο βασικότερος δείκτης που μας πληροφορεί για την ποιότητα του μοντέλου είναι ο Συντελεστής Προσδιορισμού (Coefficient of Determination). Γι' αυτό και κατά τη σύγκριση του μοντέλου αυτός αποτελεί το βασικότερο μέτρο αναφοράς. Οι λόγοι για αυτήν την επιλογή φαίνονται στο Παράρτημα Δ.

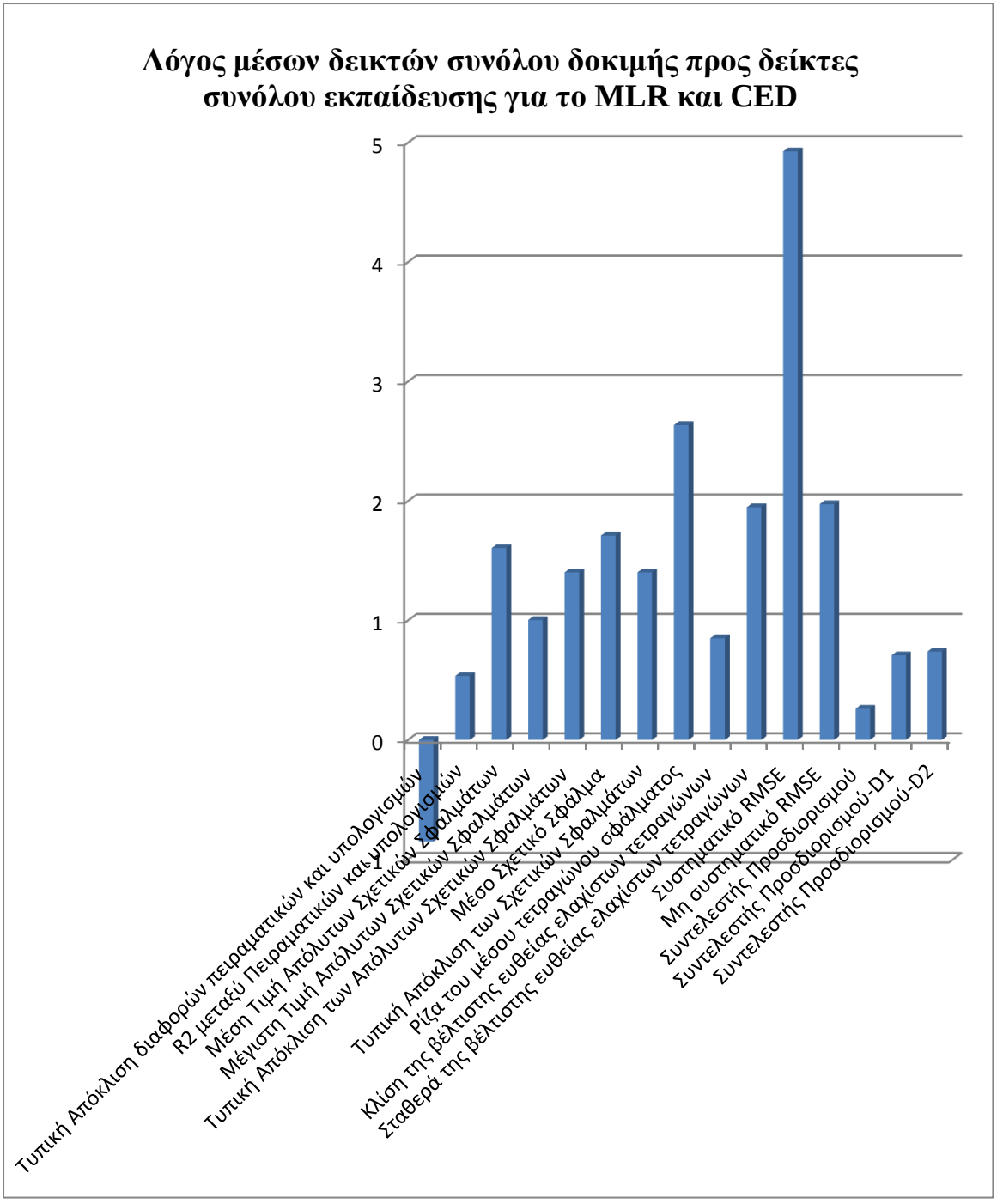
5.4 Γραμμική Παλινδρόμηση Πολλών Μεταβλητών

Το πρώτο είδος συσχετισμού που επιχειρήθηκε είναι η απλή γραμμική παλινδρόμηση πολλών μεταβλητών (Multiple Linear Regression,MLR). Ο πίνακας που δίνεται ως είσοδος είναι ο πίνακας του συνόλου εκπαίδευσης που προκύπτει από το κάθε διαμέριση, 171x57, μόνο που έχει προστεθεί και μια ακόμα στήλη με τον αριθμό ένα (δηλαδή 171x58) για να προκύψει και η παράμετρος του γραμμικού

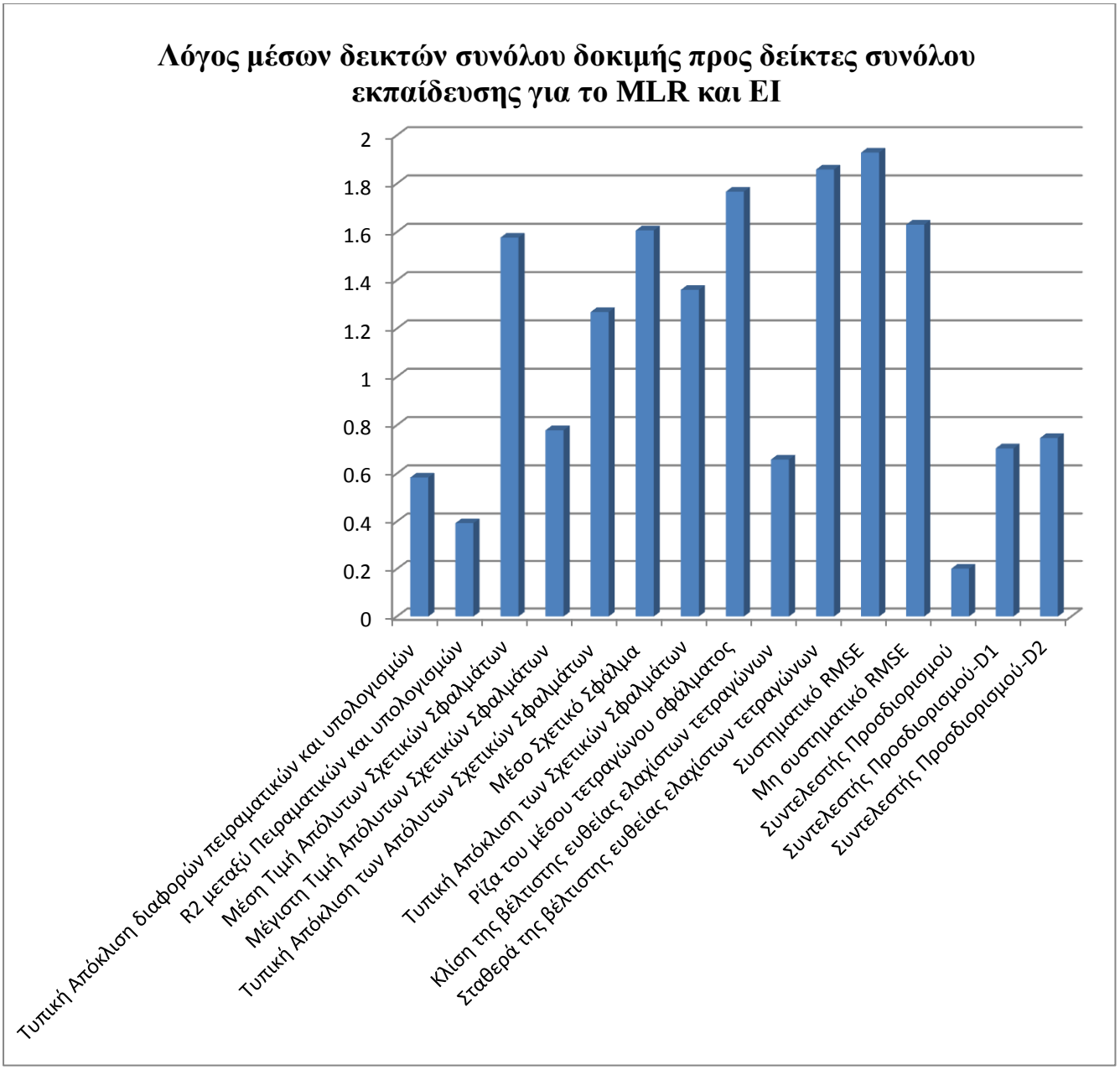
μοντέλου. Επίσης, ο πίνακας 171x3, όπου η κάθε στήλη περιέχει για κάθε μόριο που συμμετέχει στο σύνολο εκπαίδευσης τις μετρημένες τιμές (πειραματικές και υπολογισμένες). Εννοείται πως για να προκύψει το κάθε μοντέλο, χρησιμοποιείται κάθε φορά μια από τις τρεις στήλες του πίνακα. Έπειτα, πραγματοποιούνται για το κάθε σύνολο δοκιμής, υπολογισμοί για Απόσταση Mahalanobis και προκύπτουν οι ακραίες περιπτώσεις. Τέλος, χρησιμοποιείται η συνάρτηση του MATLAB, (…)= regress(training set, experimental/calculated data), για να προκύψουν οι τιμές των παραμέτρων. Υπολογίζεται η τιμή του κάθε δείκτη (GWP,CED, EI 99) από τα μοντέλα που προέκυψαν και προκύπτουν με τη διαδικασία που περιγράφηκε στο προηγούμενο κεφάλαιο. Στα διαγράμματα 5.1-5.3 φαίνεται ο λόγος των μέσων δεικτών του συνόλου δοκιμής προς τους μέσους δείκτες του συνόλου εκπαίδευσης και στο Παράρτημα Ι τα αποτελέσματα των στατιστικών δεικτών.



Διάγραμμα 5.1. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για GWP και MLR



Διάγραμμα 5.2. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για CED και MLR



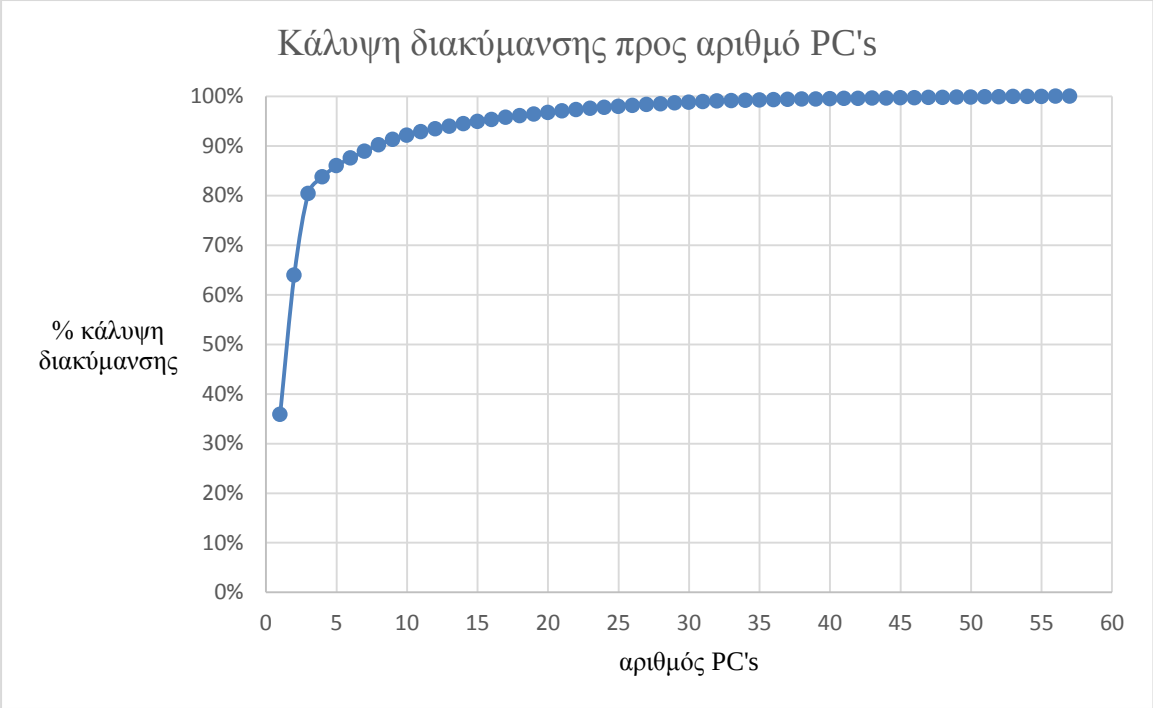
Διάγραμμα 5.3. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για EI 99 και MLR

5.5 Ανάλυση Πρωτευόντων Συνιστωσών

Η ανάπτυξη των μοντέλων συνεχίστηκε με PCA. Αρχικά, πραγματοποιήθηκε ο διαχωρισμός και σχηματίστηκαν χίλιες διαφορετικές διαμερίσεις, όπως και στις προηγούμενες περιπτώσεις. Στη συγκεκριμένη περίπτωση είναι επιθυμητό να μειώσουμε το πλήθος των ανεξάρτητων μεταβλητών (δηλαδή, τον αριθμό εμφάνισης κάθε ενός από τις 57 ομάδες) σε ένα μικρότερο σύνολο. Αναπτύσσοντας ένα μοντέλο με βάση ένα μικρότερο αριθμό μεταβλητών μπορεί, όπως προαναφέρθηκε, να συντελέσει σε ένα πιο αξιόπιστο μοντέλο. Πρέπει να αντικατασταθούν, λοιπόν, κάθε μια από τις 57 μεταβλητές με PC's. Μετά τον χωρισμό σε σύνολα εκπαίδευσης και δοκιμής ακολούθησε όπως και πριν, ο υπολογισμός της απόστασης mahalanobis για τα σύνολα δοκιμής. Για την κάθε διαμέριση υπολογίστηκαν δύο πίνακες: ο ένας είναι ο πίνακας των συντελεστών που συνδέουν τον αριθμό εμφάνισης των ομάδων με τα PC's (πίνακας των coefficients) και ο δεύτερος, είναι ο πίνακας που περιλαμβάνει τα αποτελέσματα (scores) (βλ. εισαγωγικά για PCA για περισσότερες πληροφορίες), που περιλαμβάνει τις τιμές που λαμβάνει κάθε PC, για το κάθε μόριο. Ο πίνακας των συντελεστών περιλαμβάνει τόσα PC's, όσα και τις αρχικές μεταβλητές, δηλαδή είναι πίνακας 57x57. Ο πίνακας των αποτελεσμάτων περιλαμβάνει 171 γραμμές για τα 171 μόρια τα οποία θα συμπεριληφθούν στη δημιουργία του μοντέλου και 57 στήλες για τα αποτελέσματα των 57 πιθανών PC's. Όπως έχει γίνει ξεκάθαρο, δε θα χρησιμοποιηθούν και οι 57 στήλες, διότι τότε δεν επιτυγχάνεται η μείωση του αριθμού των αρχικών μεταβλητών.

Για να αποφασιστεί πόσα PC's είναι πρακτικό να χρησιμοποιηθούν, ώστε να μειωθεί επαρκώς ο αριθμός των μεταβλητών, αλλά να μην χάνεται ένα μεγάλο μέρος της πληροφορίας του αρχικού συνόλου, υπολογίζεται τι ποσοστό της διακύμανσης καλύπτει κάθε PC. Όπως γίνεται αντιληπτό, αφού σε κάθε διαμερισμό συμμετέχουν άλλα μόρια, θα είναι διαφορετικό το αρχικό σύνολο και μαζί η κατανομή των τιμών των μεταβλητών. Συνεπώς, από διαμερισμό σε διαμερισμό είναι διαφορετικά τα PC's (coefficient και scores) και το ποσοστό διακύμανσης, που καλύπτει κάθε ένα. Γι' αυτό και έγινε η μελέτη σε μερικούς τυχαίους διαμερισμούς, για να βρεθεί αν μεταβάλλεται σημαντικά η πληροφορία που «ενσωματώνεται» από κάθε ένα PC. Βρέθηκε πως το ποσοστό αυτό δε μεταβάλλεται σημαντικά. Με άλλα λόγια, είναι δυνατόν με τον ίδιο αριθμό PC's να καλυφθεί η ίδια ή σχεδόν η ίδια διακύμανση του εκάστοτε αρχικού συνόλου (του κάθε διαμερισμού). Επιλέγεται, λοιπόν, η κατανομή της διακύμανσης στα PC's του αρχικού συνόλου του πρώτου διαμερισμού, ως αντιπροσωπευτικό δείγμα. Να σημειωθεί ότι θεωρείται γενικά, πως καλύπτεται επαρκής πληροφορία όταν καλύπτεται τουλάχιστον το 80-90% της αρχικής διακύμανσης με τα PC's.

Έπειτα γίνεται η αθροιστική καμπύλη (δηλαδή, τι ποσοστό καλύπτεται αν χρησιμοποιήσουμε το 1ο PC, το 1ο και το 2ο PC, το 1ο, το 2ο και το 3ο PC κ.ο.κ.). Προφανώς τα 57 PC's μαζί καλύπτουν το 100% της διακύμανσης. Στο διάγραμμα 5.4 φαίνεται η αθροιστική καμπύλη και εξάγονται τα συμπεράσματα:



Διάγραμμα 5.4. Κάλυψη της διακύμανσης από τον αριθμό των PC's

Φαίνεται πως ήδη το πρώτο PC καλύπτει ένα αρκετά ικανό ποσοστό της διακύμανσης, ίσο με 37% περίπου και ότι με 2 PC's έχουμε ήδη περιγράψει το 64% περίπου της συνολικής πληροφορίας. Το υπάρχον σύστημα είναι ένα αρκετά καλό σύστημα, δηλαδή, το οποίο μπορεί να περιγραφεί πολύ ικανοποιητικά από μόλις λίγα PC's (κρίνεται θετική η απότομη κλίση της καμπύλης στην αρχή). Είναι, ακόμα, ξεκάθαρο πως με τα τρία PC's έχει καλυφθεί ήδη το 80% της ελάχιστης (τυπικά) διακύμανσης που πρέπει να υπάρχει για να θεωρείται η συσχέτιση του μοντέλου αξιόπιστη. Για στατιστικούς και ερευνητικούς λόγους αποφασίζεται να μη δημιουργηθεί ένα μοντέλο, μόνο για τα τρία PC, αλλά να δοκιμαστεί η απόδοση της συσχέτισης έως και με 12 διαφορετικά PC's (θα αναπτυχθούν, δηλαδή, 10 διαφορετικά μοντέλα). Όμως με το να αναπτυχθεί πλήθος μοντέλων με τα PC προκύπτει ένα ακόμα όφελος: αν τυχόν υπάρχουν κάποιοι διαμερισμοί, οι οποίοι δεν αντιπροσωπεύονται από την παραπάνω κατανομή της κάλυψης της διακύμανσης (και χρειάζονται περισσότερα PC's για να είναι αξιόπιστα), τότε με την ανάπτυξη μοντέλων με περισσότερα PC, μπορούν αυτοί οι διαμερισμοί να αναδειχθούν και να υποδείξουν ποιος αριθμός PC's έχει το καλύτερο αποτέλεσμα.

Για να δημιουργηθεί οι πίνακες συντελεστών και αποτελεσμάτων χρησιμοποιείται η εντολή MATLAB: [coefficient, score, variance coverage]=princomp(αρχικός πίνακας). Στην παλινδρόμηση που γίνεται κάθε φορά για διαφορετικό αριθμό PC's χρησιμοποιείται και διαφορετικός αριθμός από στήλες των πινάκων συντελεστών και αποτελεσμάτων (score). Για παράδειγμα, όταν αναπτύσσεται ένα μοντέλο χρησιμοποιώντας 5 PC's, χρησιμοποιούνται οι 5 πρώτες στήλες του πίνακα των συντελεστών και αποτελεσμάτων (scores).

Όπως έχει αναφερθεί στο εισαγωγικό κεφάλαιο του PCA, προκύπτουν οι συντελεστές contr των PC's στο γραμμικό μοντέλο, π.χ. για 5 PC's:

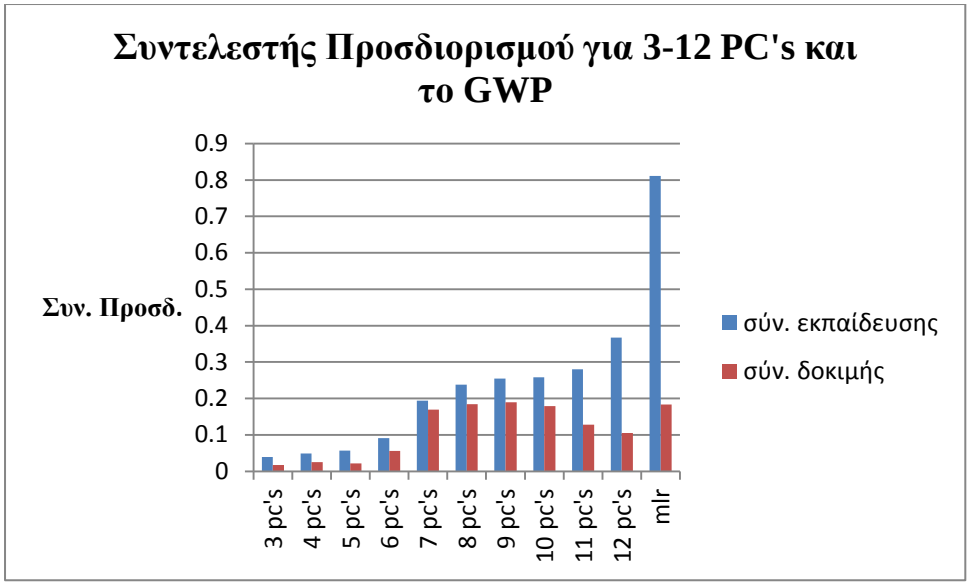
$$GWP = N_1 \cdot contr_{GWP\_PC_1} + N_2 \cdot contr_{GWP\_PC_2} + N_3 \cdot contr_{GWP\_PC_3} + N_4 \cdot contr_{GWP\_PC_4} + N_5 \cdot contr_{GWP\_PC_5} + intercept_{GWP} \quad (5.5)$$

$$CED = N_1 \cdot contr_{CED\_PC_1} + N_2 \cdot contr_{CED\_PC_2} + N_3 \cdot contr_{CED\_PC_3} + N_4 \cdot contr_{CED\_PC_4} + N_5 \cdot contr_{CED\_PC_5} + intercept_{CED} \quad (5.6)$$

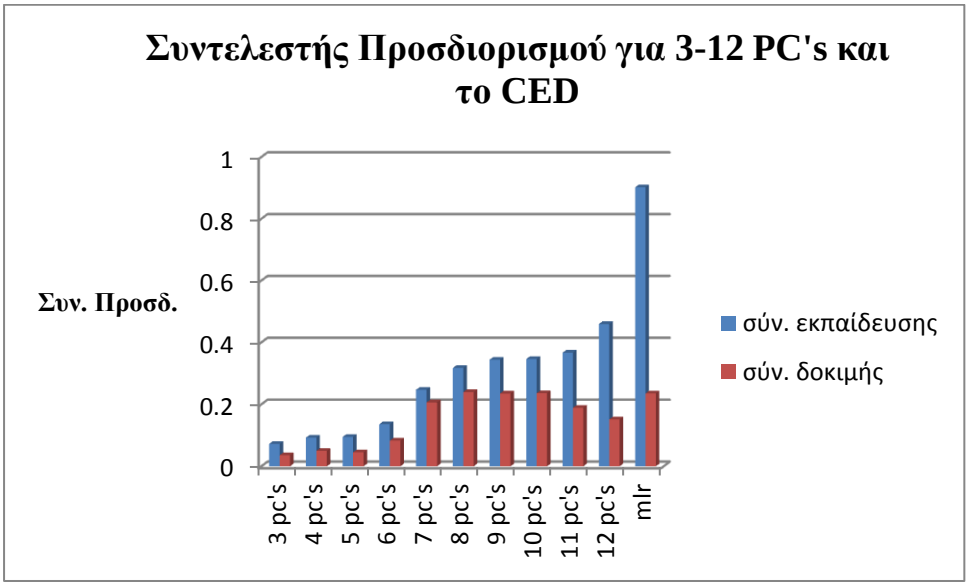
$$EI = N_1 \cdot contr_{EI\_PC_1} + N_2 \cdot contr_{EI\_PC_2} + N_3 \cdot contr_{EI\_PC_3} + N_4 \cdot contr_{EI\_PC_4} + N_5 \cdot contr_{EI\_PC_5} + intercept_{EI} \quad (5.7)$$

,όπου $contr_{PC_i}$, η συνεισφορά του i PC και N_i, το αποτέλεσμα για το κάθε PC για το συγκεκριμένο μόριο (κατ' αναλογία με τον αριθμό των επαναλήψεων). Για να προκύψουν οι συνεισφορές των 57 ομάδων και το μοντέλο από την παραπάνω μορφή των PC's, να μετατραπεί στη μορφή των εξισώσεων 5.1, 5.2, 5.3 είναι απαραίτητο, για την κάθε ομάδα να πολλαπλασιαστεί το διάνυσμα που αποτελείται από τη σειρά του πίνακα των συντελεστών (coefficient), που αντιστοιχεί σε αυτή την ομάδα (στο παραπάνω παράδειγμα, αφού χρησιμοποιούνται 5 PC's, έτσι και ο πίνακας και το διάνυσμα θα έχει 5 στήλες) με το διάνυσμα που αποτελείται από τους συντελεστές των PC's του γραμμικού μοντέλου (στο παράδειγμα μας είναι $[contr_{PC_1}, contr_{PC_2}, contr_{PC_3}, contr_{PC_4}, contr_{PC_5}]$). Η σταθερά του μοντέλου παραμένει ίδια.

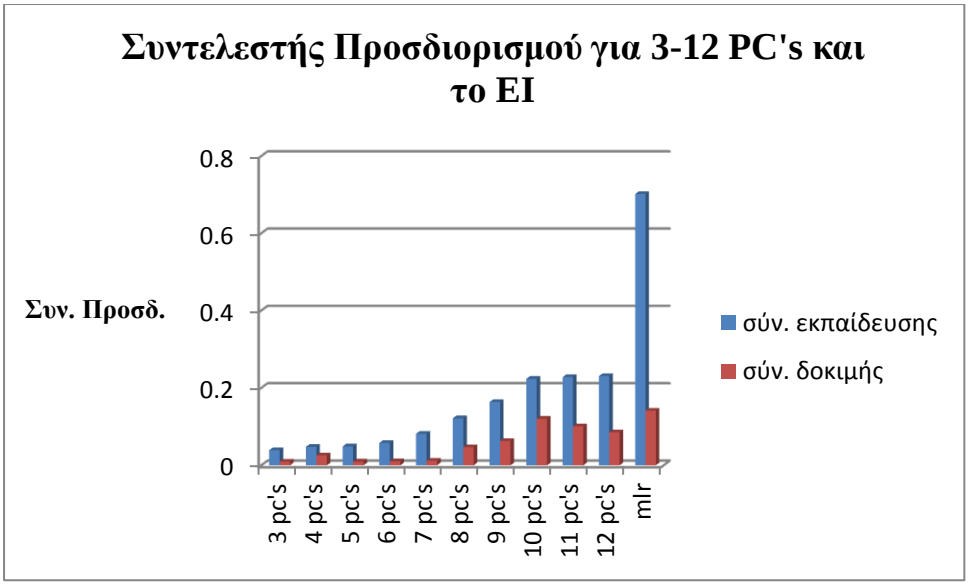
Έπειτα, και αφού εξαχθούν για την κάθε περίπτωση PC, όλοι οι μέσοι στατιστικοί δείκτες (κατά τα προηγούμενα κεφάλαια) για το σύνολο εκπαίδευσης και δοκιμής, συγκρίνονται οι μέσοι Συντελεστές Προσδιορισμού του συνόλου δοκιμής για το κάθε μοντέλο με διαφορετικό αριθμό PC's, ώστε να διαπιστωθεί για ποιο αριθμό PC's επετεύχθη η καλύτερη σχέση:



Διάγραμμα 5.5. Σύγκριση των Συντελεστών Προσδιορισμού για το GWP στα μοντέλα με 3-12 PC's

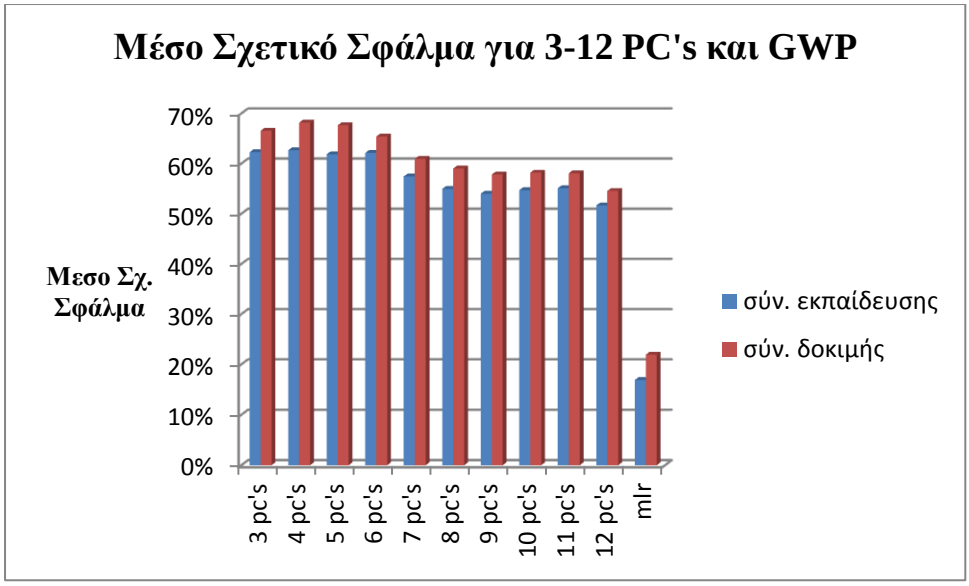


Διάγραμμα 5.6. Σύγκριση των Συντελεστών Προσδιορισμού για το CED στα μοντέλα με 3-12 PC's

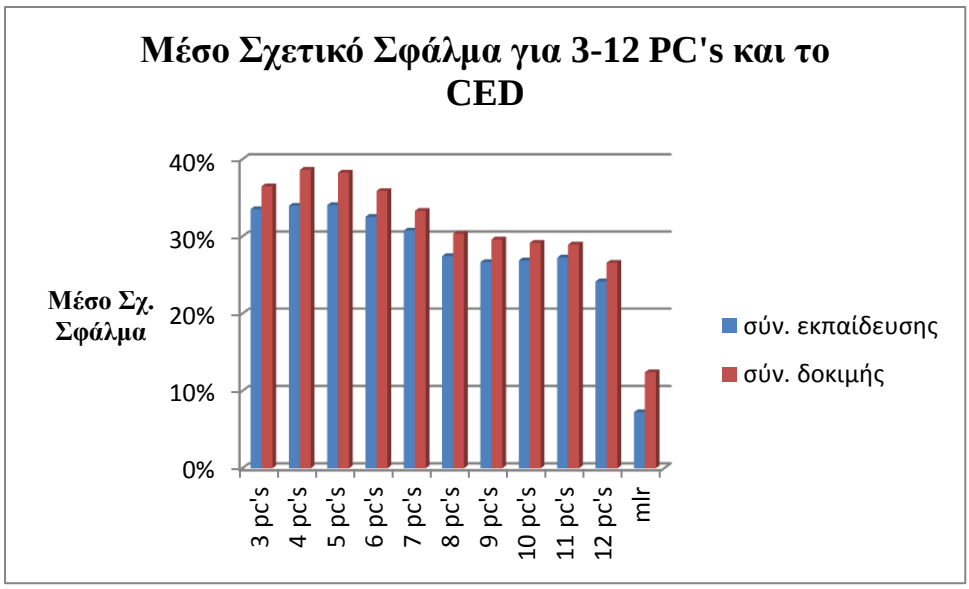


Διάγραμμα 5.7. Σύγκριση των Συντελεστών Προσδιορισμού για το EI 99 στα μοντέλα με 3-12 PC's

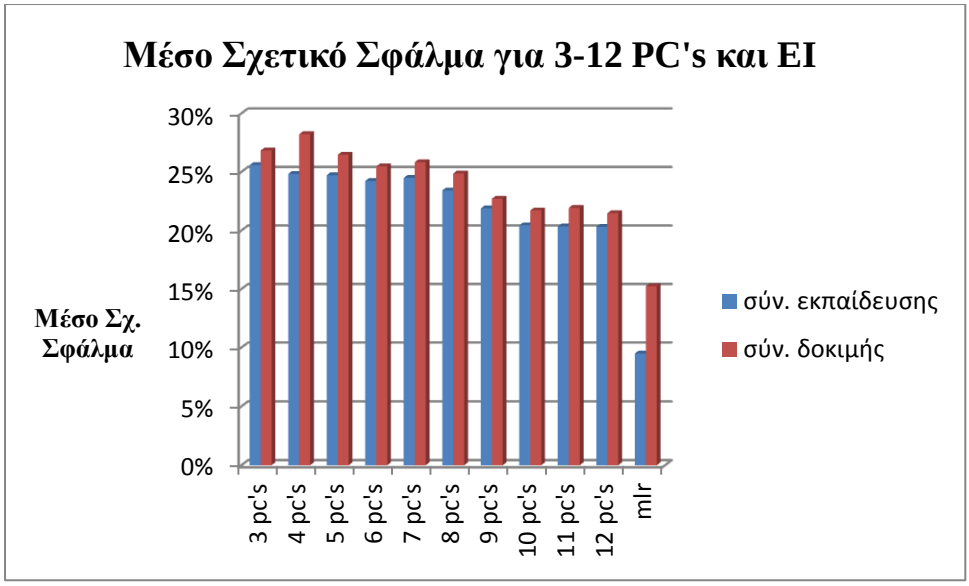
Είναι γενικά φανερό πως το μοντέλο του δείκτη GWP, που έχει τις χειρότερες συσχετίσεις και είναι κατά κάποιο τρόπο το πιο «δύσκολο» για να γίνουν συσχετίσεις γι' αυτό. Αυτό δε φαίνεται, μόνο, από το PCA και το MLR, αλλά θα φανεί και από τα επόμενα μοντέλα. Σαφώς, μπορεί ο χρήστης να χρησιμοποιήσει για το μοντέλο κάθε δείκτη, τον αριθμό PC's που δίνει τα καλύτερα αποτελέσματα. Χάριν απλότητας, όμως, θεωρούμε πως ο ιδανικός αριθμός για τα τρία μοντέλα είναι αυτός στον οποίο, βελτιστοποιείται η συσχέτιση στο μοντέλο του δείκτη GWP. Από τα παραπάνω διαγράμματα φαίνεται ότι το μοντέλο του GWP είναι το καλύτερο για 9 PC's στο σύνολο δοκιμής, μιας και έχει το μεγαλύτερο Συντελεστή Προσδιορισμού. Ο λόγος που επιλέγονται οι δείκτες του συνόλου δοκιμής για να συμμετάσχουν στην επιλογή των PC's αναλύεται στο Κεφάλαιο 4. Αυτή η επιλογή δεν είναι και για τα άλλα μοντέλα τόσο κακή, μιας και τα 9 PC's για το CED έχουν από τις καλύτερες συσχετίσεις με ελαφρύ προβάδισμα τα 10 και 8 PC's, ενώ είναι σχετικά καλό και το μοντέλο των 9 PC's και για το EI 99 με σαφές προβάδισμα του μοντέλου με 10 και 11 PC's. Για συγκριτική μελέτη και του σφάλματος, που είναι επίσης σημαντικός δείκτης, παρατίθενται και συγκριτικά για τα τρία μοντέλα:



Διάγραμμα 5.8. Σύγκριση του Μέσου Σχετικού Σφάλματος για το GWP στα μοντέλα με 3-12 PC's

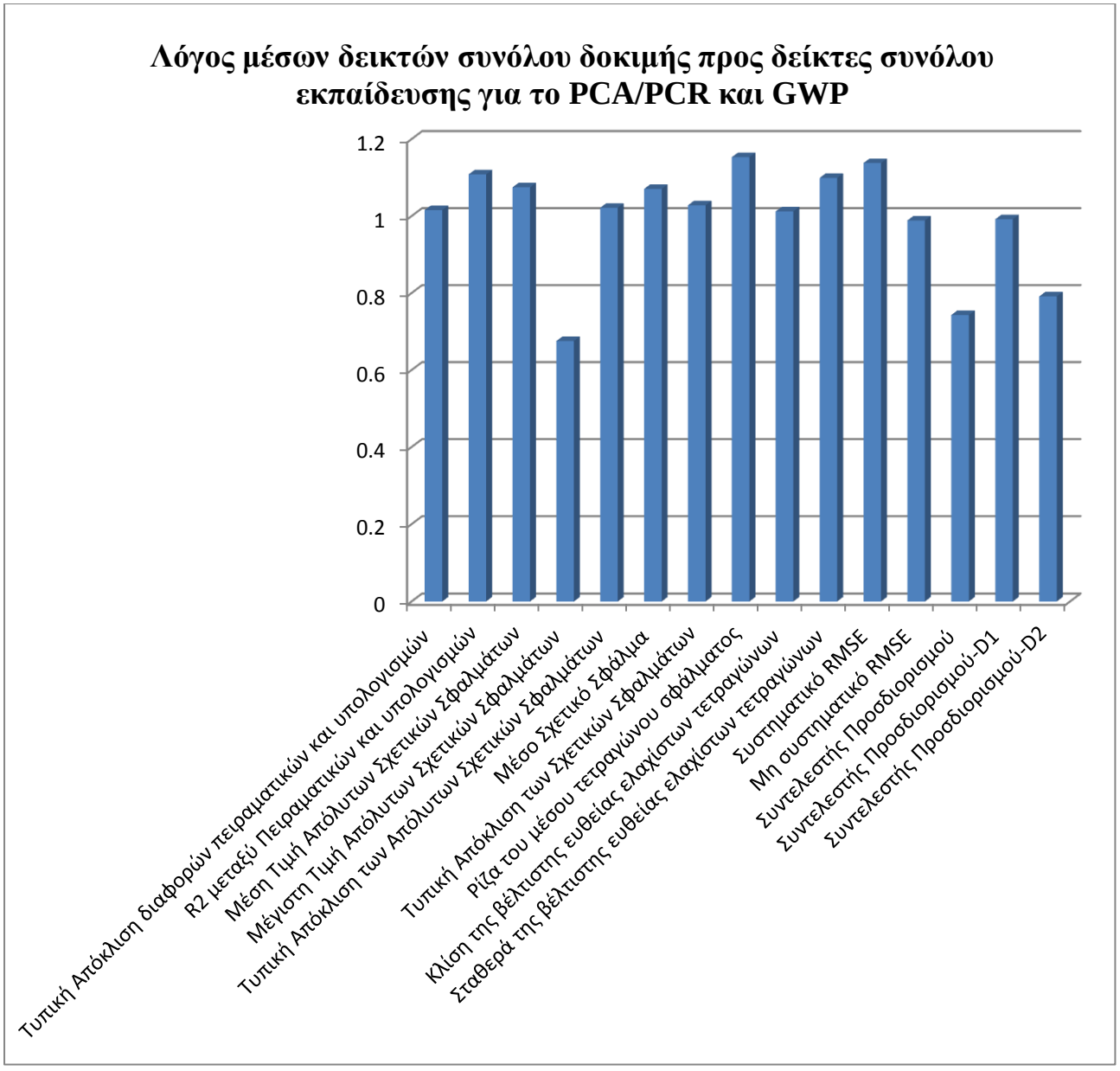


Διάγραμμα 5.9. Σύγκριση του Μέσου Σχετικού Σφάλματος για το CED στα μοντέλα με 3-12 PC's

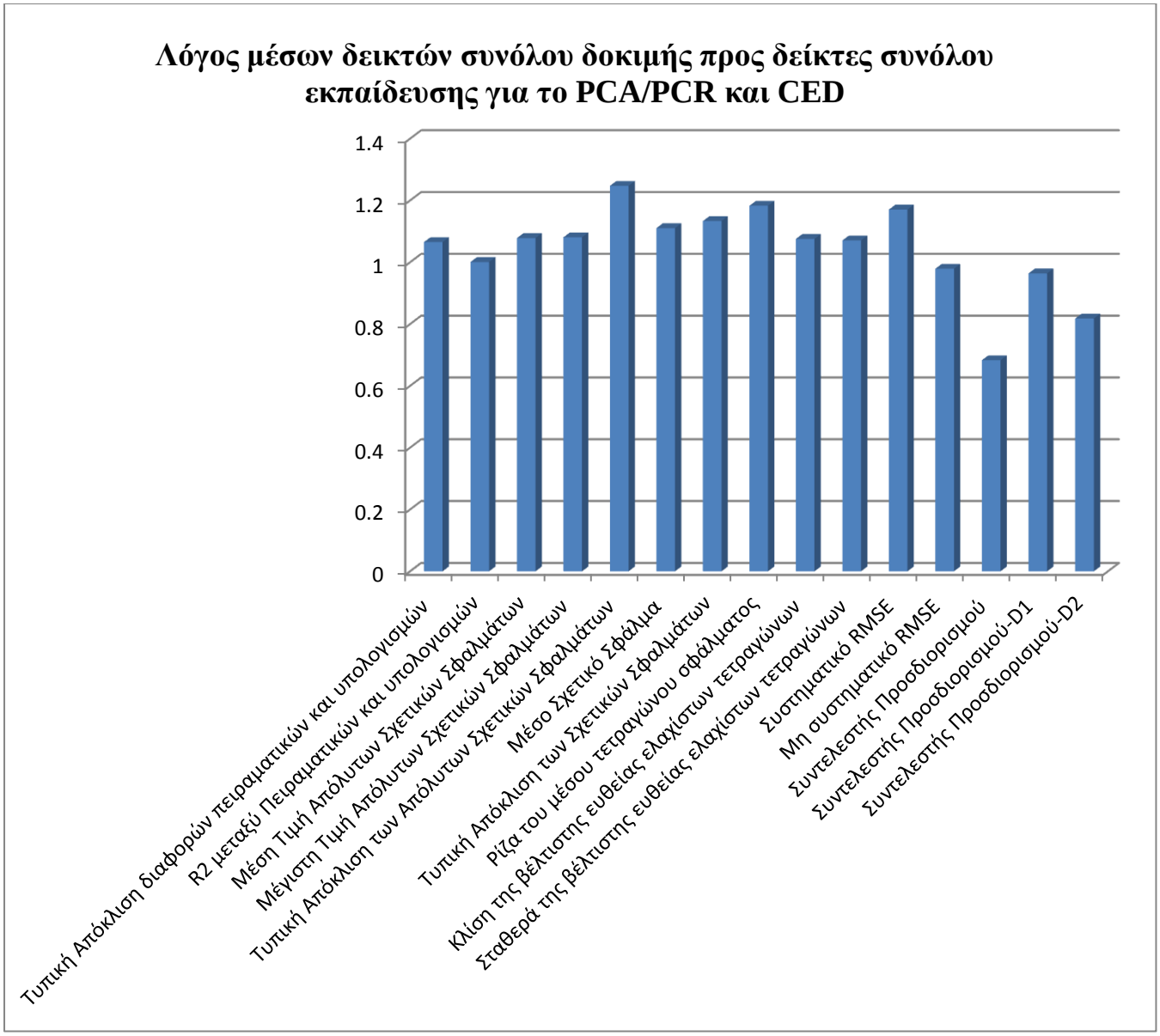


Διάγραμμα 5.10. Σύγκριση του Μέσου Σχετικού Σφάλματος για το EI 99 στα μοντέλα με 3-12 PC's

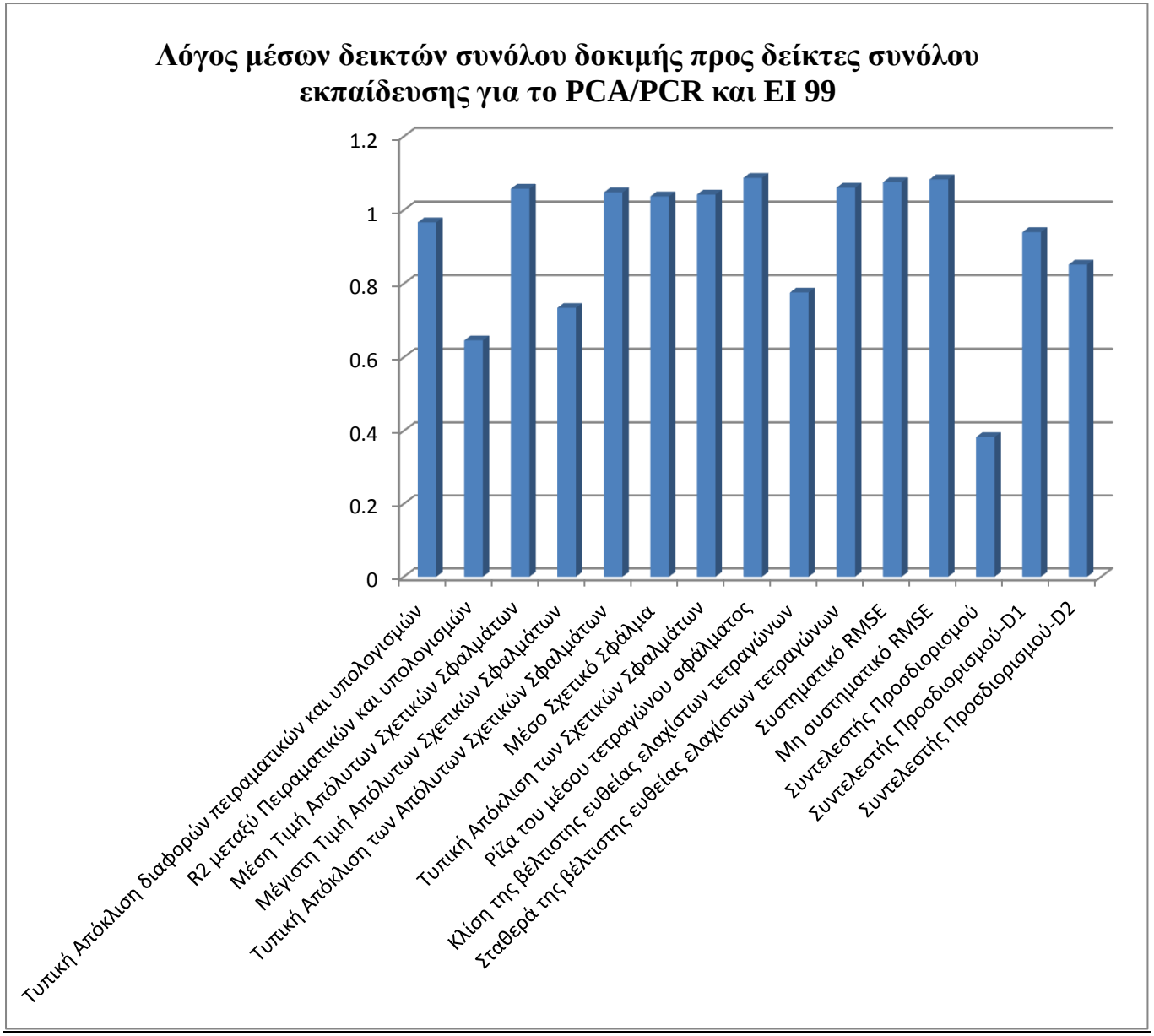
Φαίνεται πως η επιλογή για τα 9 PC's ήταν καλή και για τα σφάλματα, καθώς τα σφάλματα των μοντέλων με 9 PC's είναι από τα χαμηλότερα. Είναι ξεκάθαρο, πως υπάρχει μια τάση μείωσης του μέσου σφάλματος με την αύξηση του αριθμού των PC's. Στο παράρτημα Ι φαίνονται τα αποτελέσματα των στατιστικών δεικτών για το PCA/PCR. Στα διαγράμματα 5.11-5.13 παρατίθενται και οι 15 λόγοι στατιστικών δεικτών για τα μοντέλα των δεικτών GWP, CED και EI 99:



Διάγραμμα 5.11. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για GWP και το PCA/PCR για 9 PC's



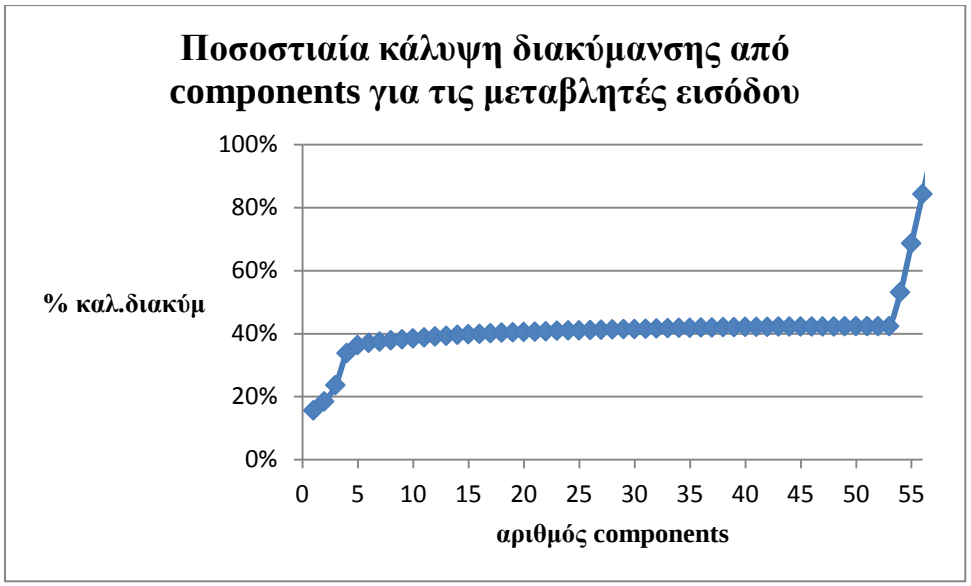
Διάγραμμα 5.12. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για CED και το PCA/PCR για 9 PC's



Διάγραμμα 5.13. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για EI 99 και το PCA/PCR για 9 PC's

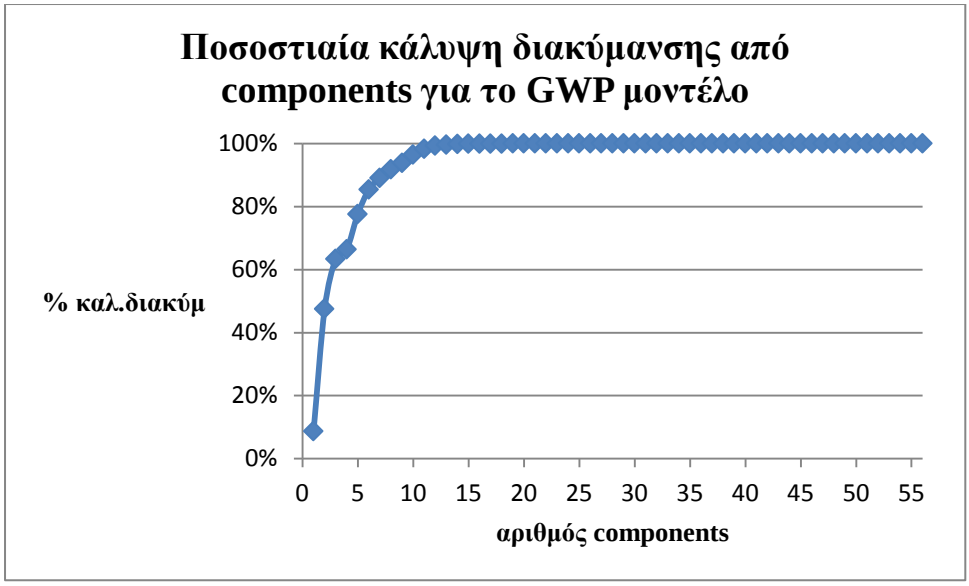
5.6 Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων

Στην περίπτωση αυτή ακολουθούνται, ακριβώς, τα ίδια βήματα με το PCA. Πραγματοποιούνται οι διαμερισμοί, εξάγεται η απόσταση mahalanobis για τα στοιχεία κάθε σύνολο δοκιμής. Αυτή τη φορά, όμως, δεν θα εξαχθούν λανθάνουσες μεταβλητές μόνο για τις μεταβλητές της εισόδου, αλλά και για τους τρεις δείκτες του LCA της εξόδου. Ο μέγιστος αριθμός συνιστωσών είναι 57. Στο διάγραμμα 5.14 παρατίθεται το διάγραμμα που δείχνει το ποσοστό κάλυψης της διακύμανσης από τα συνιστωσών για την είσοδο:

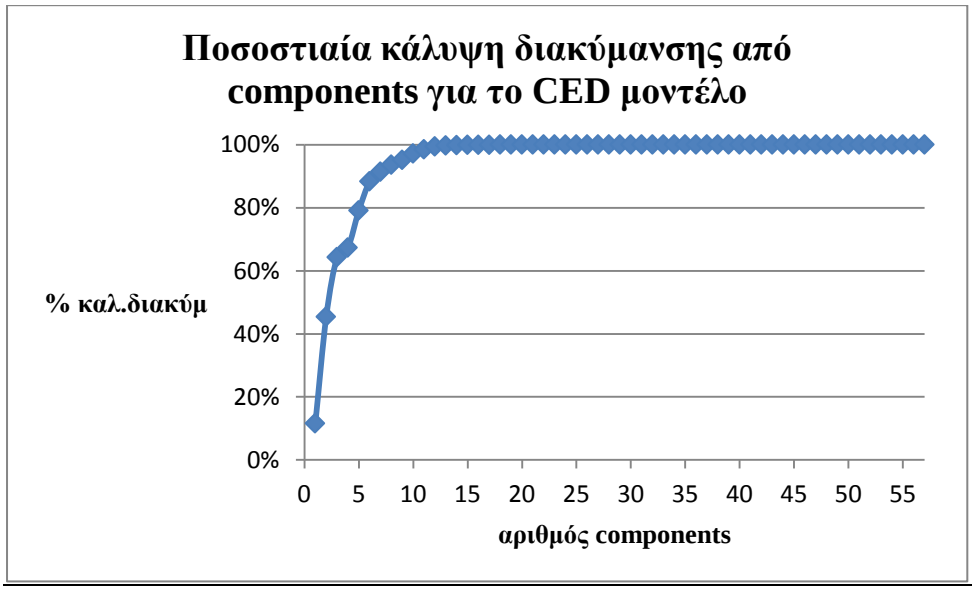


Διάγραμμα 5.14. Ποσοστό κάλυψης διακύμανσης στο πλήθος μεταβλητών εισόδου από τις συνιστώσες

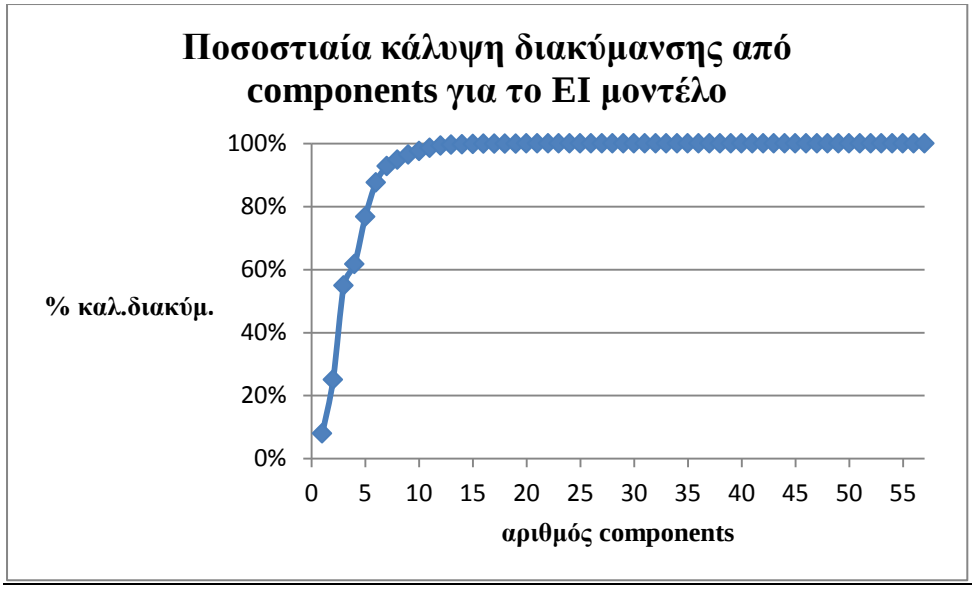
Παρατηρείται πως μετά τις 5 συνιστώσες, το ποσοστό κάλυψης μεταβάλλεται πολύ λίγο μέχρι να φτάσει στις 54, όπου ξεκινάει να καλύψει απότομα την εναπομένουσα διακύμανση. Είναι προφανές πως δεν είναι καθόλου πρακτικό να χρησιμοποιηθούν πάνω από 50 συνιστώσες για να καλυφθεί ποσοστό που έστω προσεγγίζει το 80% της συνολικής διακύμανσης, διότι δε θα έχει μειωθεί καθόλου, πρακτικά, ο αριθμός των ανεξάρτητων μεταβλητών. Επειδή, λοιπόν, είναι σχεδόν το ίδιο να χρησιμοποιηθεί οποιοσδήποτε αριθμός από συνιστώσες μεταξύ των 5 και 50, επιλέγεται να χρησιμοποιηθεί οτιδήποτε μεταξύ 5-10 συνιστωσών. Να σημειωθεί πως, όπως και στα PC's, το παραπάνω διάγραμμα αντιστοιχεί στο πρώτο διαμερισμό και αποτελεί αντιπροσωπευτικό της κάλυψης της διακύμανσης για όλες τις διαμερίσεις. Όμως, όπως έχει γίνει ξεκάθαρο, κατά το σχηματισμό των συνιστωσών επιτυγχάνεται όσο το δυνατόν μεγαλύτερος συσχετισμός μεταξύ των μεταβλητών εισόδου και εξόδου. Πρέπει, λοιπόν, να φανεί και ποιο ποσοστό της διακύμανσης του συνόλου των τιμών των μεταβλητών εξόδου καλύπτεται. Τα διαγράμματα 5.15-5.17 δείχνουν το ποσοστό της διακύμανσης αυτής που καλύπτεται από τα συνιστώσες που σχηματίζονται για κάθε ένα από τα τρία μοντέλα:



Διάγραμμα 5.15. Ποσοστό κάλυψης διακύμανσης στο πλήθος μεταβλητών εξόδου από τις συνιστώσες για το μοντέλο GWP

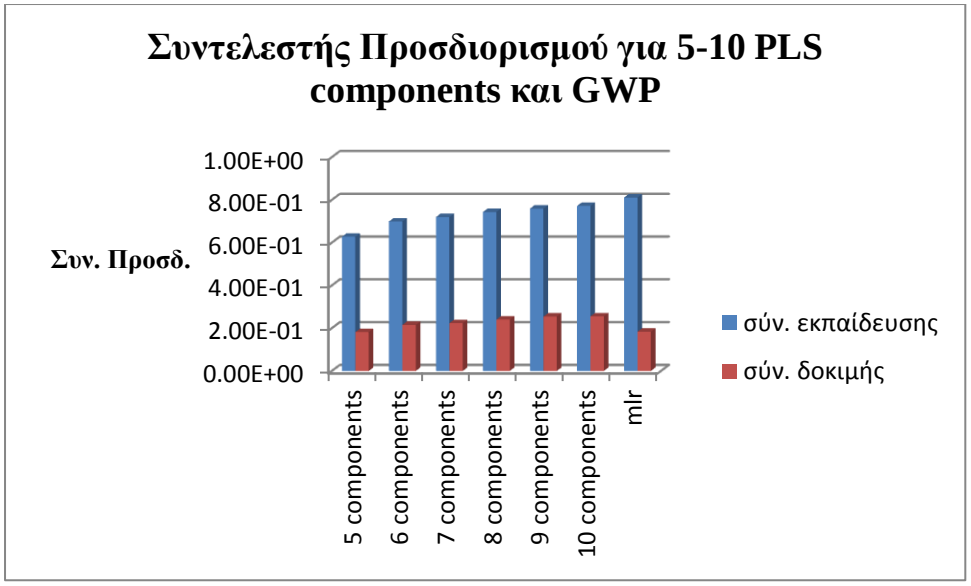


Διάγραμμα 5.16. Ποσοστό κάλυψης διακύμανσης στο πλήθος μεταβλητών εξόδου από τις συνιστώσες για το μοντέλο CED

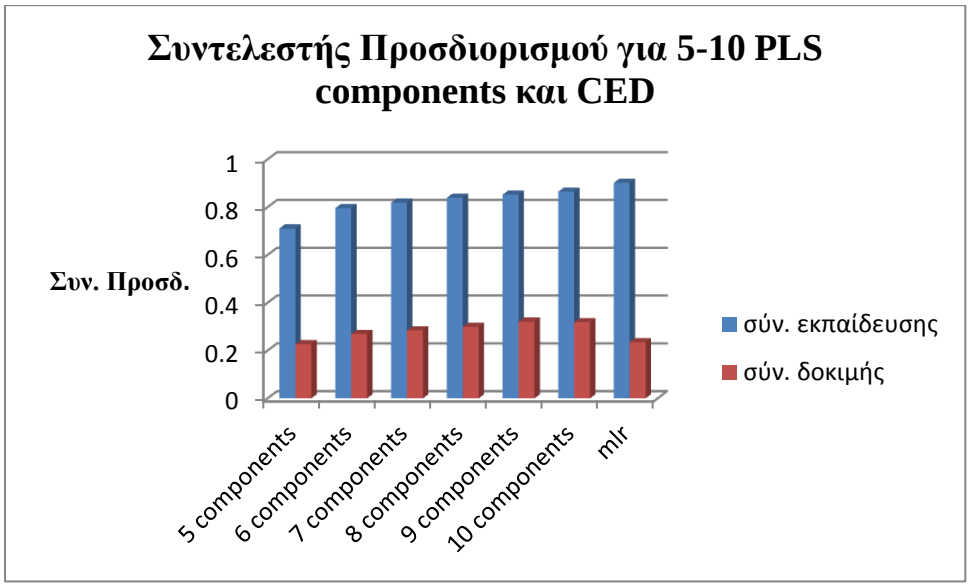


Διάγραμμα 5.17. Ποσοστό κάλυψης διακύμανσης στο πλήθος μεταβλητών εξόδου από τις συνιστώσες για το μοντέλο EI 99

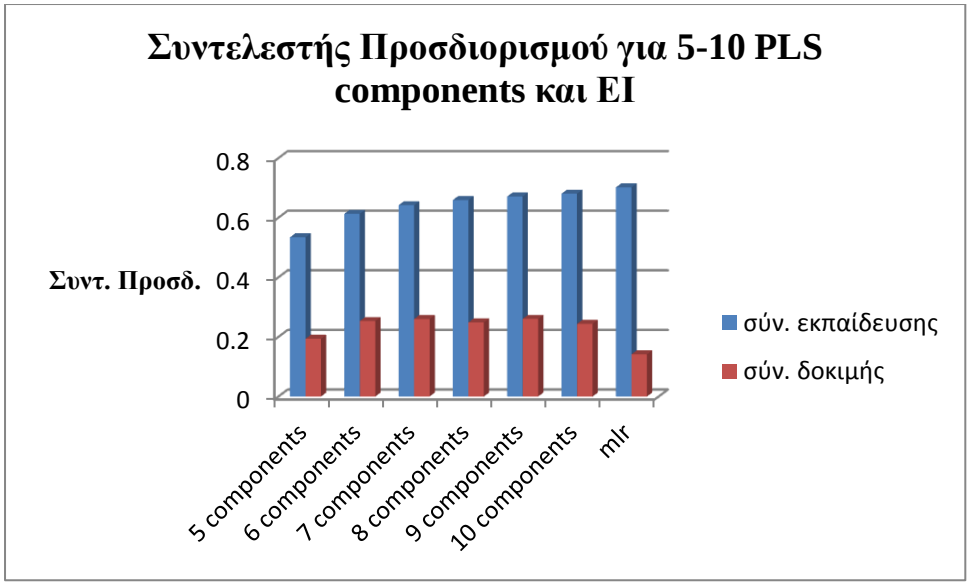
Φαίνεται πως η επιλογή για χρήση 5-10 συνιστωσών είναι αρκετά καλή μιας και με αυτά μπορεί να αναπαρασταθεί επαρκέστατα η πληροφορία και για τους τρεις δείκτες. Με 5 συνιστώσες φαίνεται πως καλύπτεται, για την έξοδο του μοντέλου, περίπου το 80% της πληροφορίας. Ο λόγος που μελετάμε το σχηματισμό μοντέλων με αριθμό συνιστωσών από 5-10 είναι ο ίδιος όπως και στο PCA: ερευνητικά είναι ενδιαφέρον να φανεί πως επιδρά ο αριθμός συνιστωσών στην απόδοση του PLS, αλλά και μπορεί να αναδειχτούν διαμερισμοί, οι οποίοι δεν έχουν τόσο καλά αποτελέσματα στα μοντέλα τους με λίγες συνιστώσες. Η συνάρτηση του MATLAB, η οποία χρησιμοποιείται για τη συγκεκριμένη περίπτωση του PLS, είναι η (…)=plsregress(input matrix, output matrix, number of components). Πρώτα εκτελείται αυτή η εντολή για 57 συνιστώσες, ώστε να υπολογιστεί η κατανομή της κάλυψης της διακύμανσης από τις συνιστώσες. Έπειτα, το τρέξιμο αυτό αγνοείται και ο αριθμός των συνιστωσών μεταβάλλεται κάθε φορά, ώστε να καλύψει το εύρος από 5-10. Να σημειωθεί πως το MATLAB εξάγει απευθείας τις συνεισφορές των 57 ομάδων και ασφαλώς, η σταθερά του μοντέλου και κάνει εσωτερικά όλους τους υπολογισμούς μεταξύ κανονικών και λανθανουσών μεταβλητών. Τελικά προκύπτει για την κάθε διαμέριση, το επιθυμητό διάνυσμα των συνεισφορών και εκτιμήσεις δεικτών για το σύνολο δοκιμής και εκπαίδευσης. Εξάγονται τα στατιστικά και ακολουθείται η παραπάνω διαδικασία για να υπολογιστούν οι μέσοι στατιστικοί δείκτες για το σύνολο δοκιμής και εκπαίδευσης. Τελικά, και εδώ, η σύγκριση για το ποιος αριθμός συνιστωσών είναι ο βέλτιστος θα εξαχθεί βάσει του Συντελεστή Προσδιορισμού. Ακολουθεί το συγκριτικό διάγραμμα 5.18-5.20 για το κάθε μοντέλο:



Διάγραμμα 5.18. Σύγκριση των Συντελεστών Προσδιορισμού για το GWP στα μοντέλα με 5-10 συνιστώσες

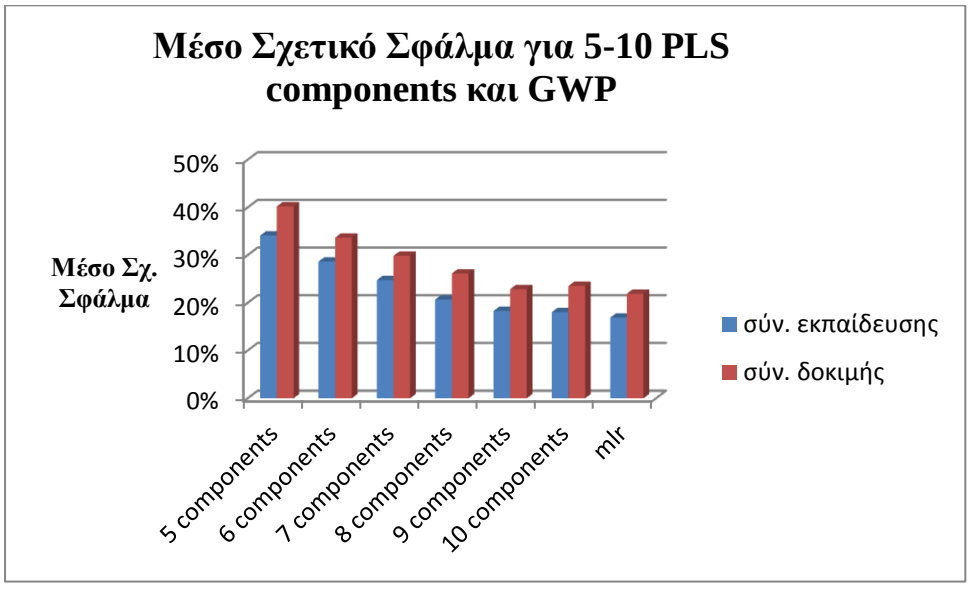


Διάγραμμα 5.19. Σύγκριση των Συντελεστών Προσδιορισμού για το CED στα μοντέλα με 5-10 συνιστώσες

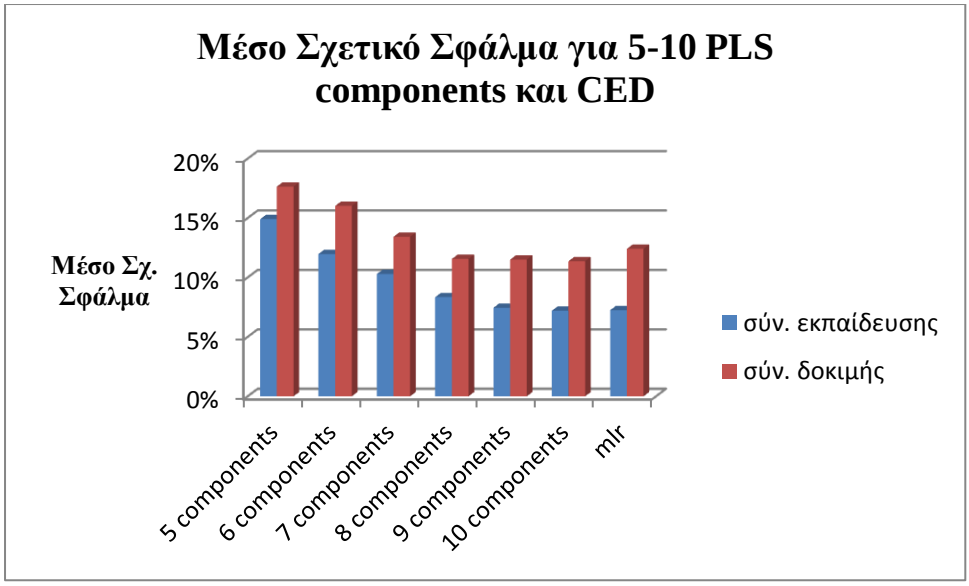


Διάγραμμα 5.20. Σύγκριση των Συντελεστών Προσδιορισμού για το EI 99 στα μοντέλα με 5-10 συνιστώσες

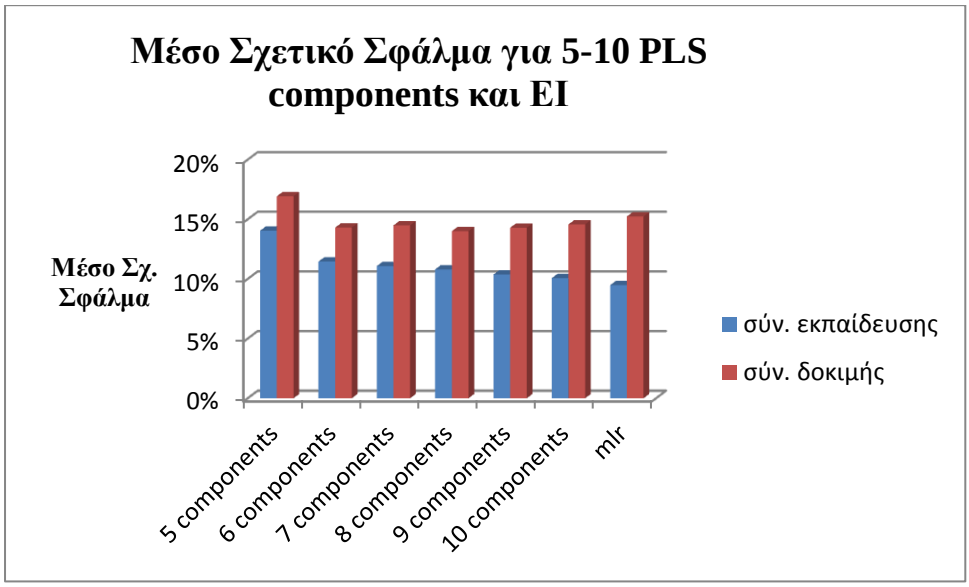
Είναι ξεκάθαρο πως ο βέλτιστος αριθμός συνιστωσών είναι ίσος με 10. Στο GWP και CED έχει το μοντέλο των 10 συνιστωσών το μεγαλύτερο Συντελεστή Προσδιορισμού, ενώ συναγωνίζονται το μοντέλο των 9 συνιστωσών. Στο EI 99 το μοντέλο των 10 συνιστωσών είναι από τα καλύτερα με ελαφρύ προβάδισμα του μοντέλου με 9 και ίσες δυνατότητες με το μοντέλο των 8 συνιστωσών. Όπως έχει προαναφερθεί, ο χρήστης μπορεί να χρησιμοποιήσει σε κάθε περίπτωση, όποιο μοντέλο επιθυμεί. Φαίνονται και τα συγκριτικά διαγράμματα 5.21-5.23 των μέσων σφαλμάτων:



Διάγραμμα 5.21. Σύγκριση του Μέσου Σχετικού Σφάλματος για το GWP στα μοντέλα με 5-10 συνιστώσες

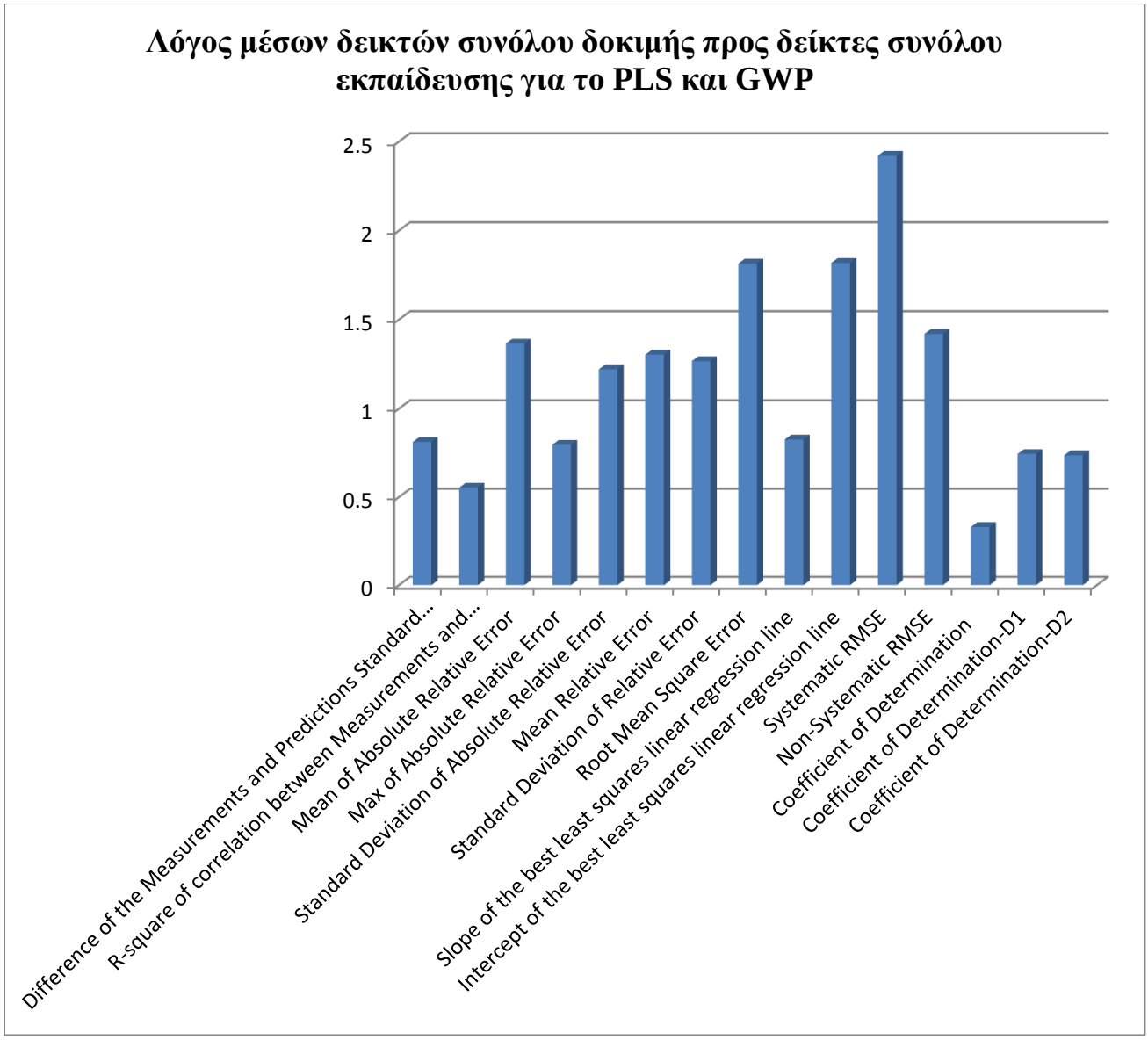


Διάγραμμα 5.22. Σύγκριση του Μέσου Σχετικού Σφάλματος για το CED στα μοντέλα με 5-10 συνιστώσες

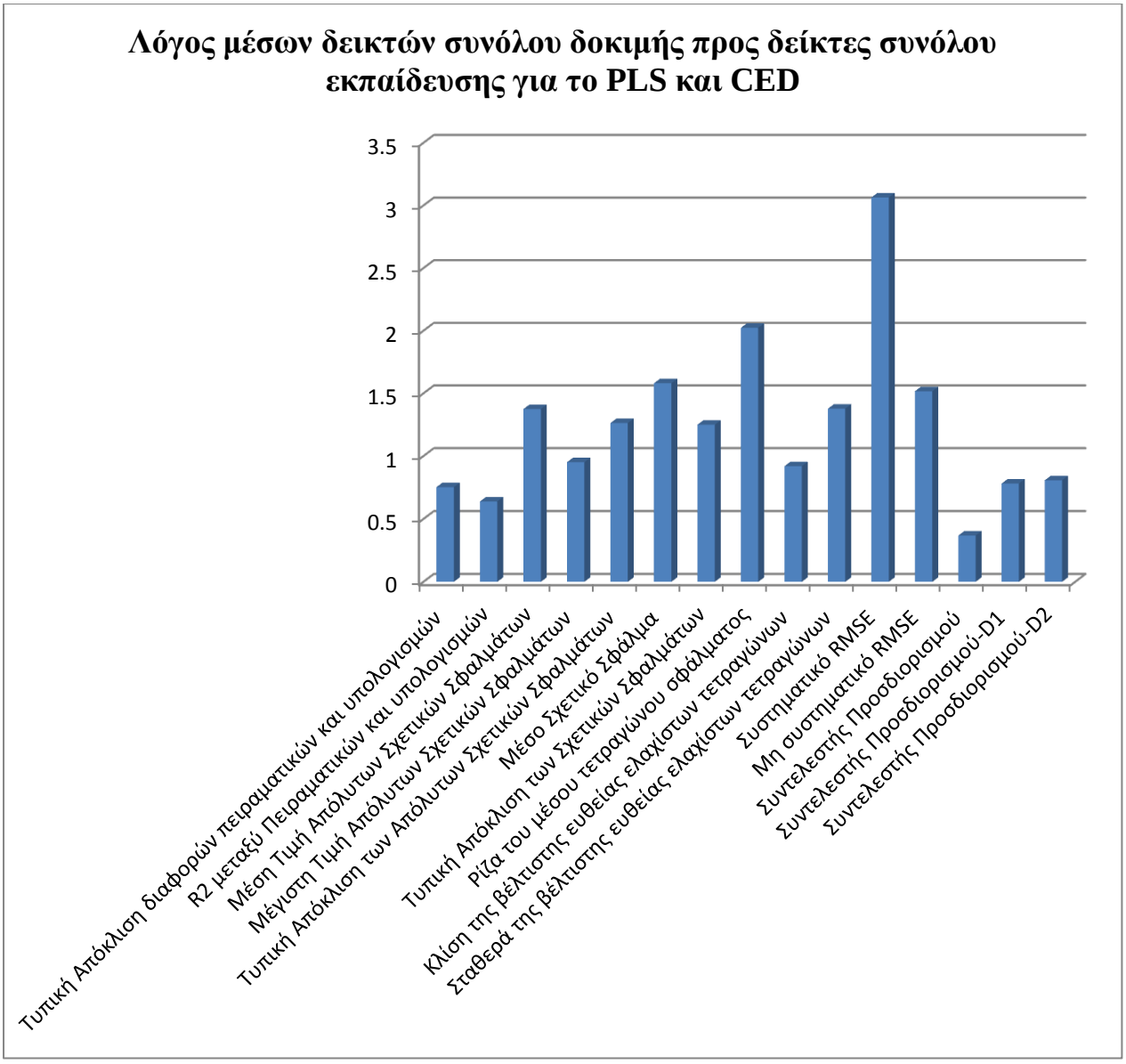


Διάγραμμα 5.23. Σύγκριση του Μέσου Σχετικού Σφάλματος για το EI 99 στα μοντέλα με 5-10 συνιστώσες

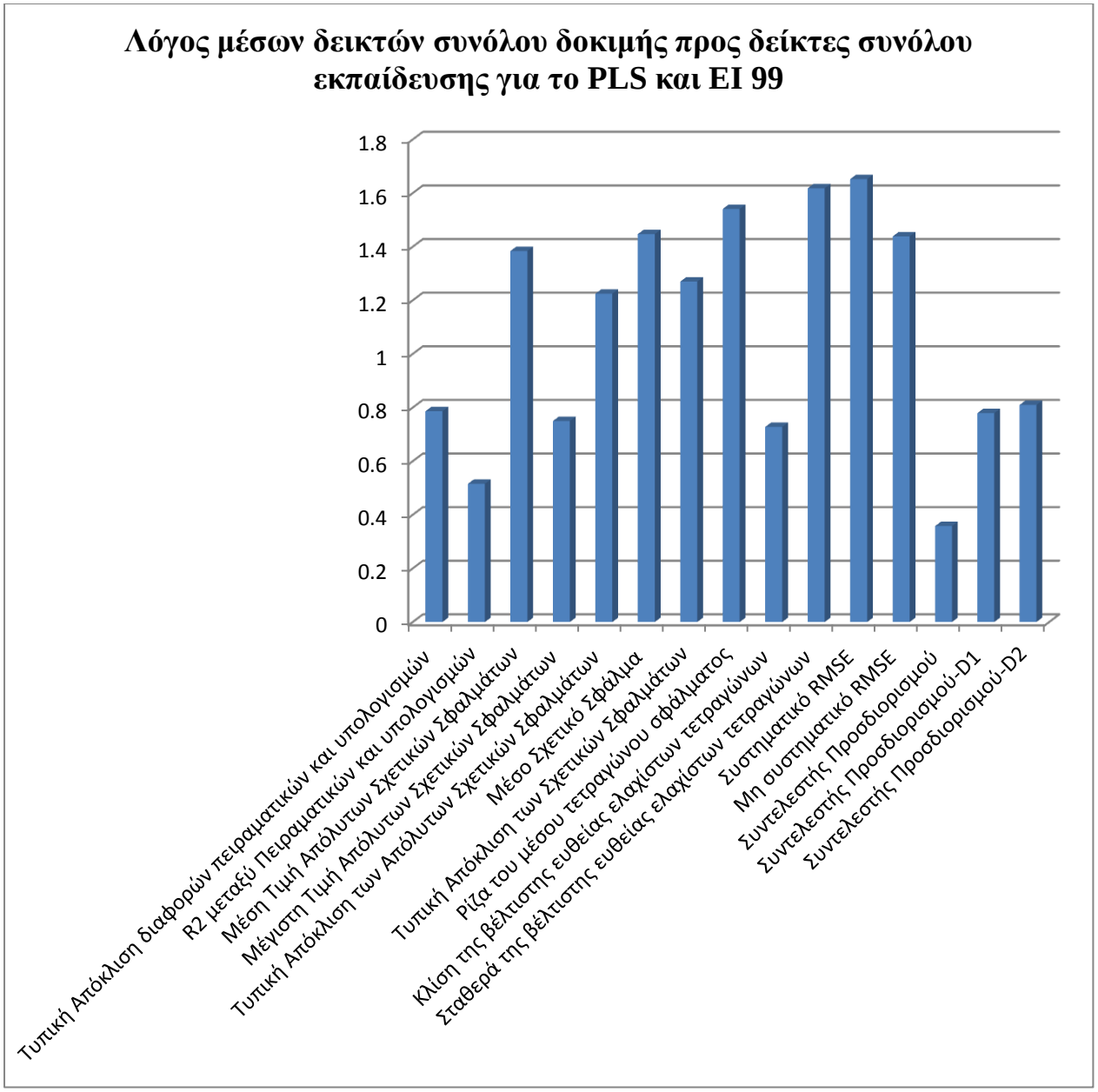
Και εδώ, όπως και στο PCA, παρατηρείται μείωση του μέσου σφάλματος με αύξηση του αριθμού των συνιστωσών. Η επιλογή του αριθμού των 10 συνιστωσών είναι αρκετά καλή, λοιπόν, δεδομένου πως τα μοντέλα αυτά έχουν σφάλματα που είναι από τα χαμηλότερα και για τους τρεις δείκτες. Τα αποτελέσματα της μεθοδολογίας PLS με 10 συνιστώσες βρίσκονται στο παράρτημα Ι. Φαίνονται στα διαγράμματα 5.24-5.26 οι λόγοι των μέσων στατιστικών δεικτών συνόλου δοκιμής προς εκπαίδευσης:



Διάγραμμα 5.24. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για GWP και PLS με 10 συνιστώσες



Διάγραμμα 5.25. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για CED και PLS με 10 συνιστώσες



Διάγραμμα 5.26. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για EI 99 και PLS με 10 συνιστώσες

5.7 Παρεμβολή τύπου «Kriging»

Όπως και στις προηγούμενες μεθόδους, έτσι και εδώ γίνεται, χωρισμός σε 1000 διαμερίσεις και υπολογισμός απόστασης mahalanobis για τα στοιχεία των συνόλων δοκιμής. Εκτελείται η μεθοδολογία παρεμβολής τύπου «kriging» για κάθε ένα από τους 1000 διαφορετικούς διαμερισμούς. Η διαδικασία αυτή είναι η πιο χρονικά δαπανηρή: για μία μόνο διαμέριση για την ανάπτυξη και των τριών μοντέλων για τους τρεις δείκτες (GWP, CED, EI 99) χρειάζεται περίπου 5 λεπτά. Για να δημιουργηθούν για τους 1000 διαμερισμούς, από 3 μοντέλα δεικτών χρειάζονται περίπου 4 ημέρες φυσικού χρόνου. Για το κάθε μοντέλο για ένα διαμερισμό παράγονται ένα διάνυσμα 1x57 με τις τιμές της παραμέτρου θ για την κάθε διάσταση

(ανεξάρτητη μεταβλητή), μια τιμή της παραμέτρου λ, στην οποία θα αναφερθούμε αργότερα και τον πίνακα Ψ (περιλαμβάνει τις αποστάσεις). Να σημειωθεί πως οι τιμές των ανεξάρτητων μεταβλητών, δηλαδή, οι τιμή που λαμβάνει ο αριθμός επαναλήψεων της κάθε ομάδας στο σύνολο εκπαίδευσης και δοκιμής κανονικοποιούνται πριν να συμμετάσχουν στο μοντέλο. Ο λόγος οφείλεται στην εσωτερική λειτουργία του αλγορίθμου που εκτελείται. Συνεπώς, κάθε στήλη του αρχικού πίνακα δεδομένων, 171x57, του κάθε διαμερισμού πρέπει να έχει εύρος από 0 έως 1. Η κανονικοποίηση (normalization) είναι η εφαρμογή ενός γραμμικού μοντέλου της μορφής, $y = ax + b$, σε ένα σύνολο δεδομένων, κατά το οποίο πρέπει να ισχύει πως όταν το x είναι ίσο με τη μικρότερη τιμή του συνόλου αριθμών (x_{min}) που πρέπει να κανονικοποιηθεί, πρέπει να ισχύει πως:

$$0 = a \cdot x_{min} + b \text{ (5.8)}$$

και όταν το x είναι ίσο με τη μεγαλύτερη τιμή του συνόλου, x_{max}, τότε:

$$1 = a \cdot x_{max} + b \text{ (5.9)}$$

Επιλύοντας το σύστημα προκύπτει πως:

$$a = \frac{1}{x_{max} - x_{min}} \text{(5.10)}$$

$$b = -\frac{x_{min}}{x_{max} - x_{min}} \text{ (5.11)}$$

,άρα:

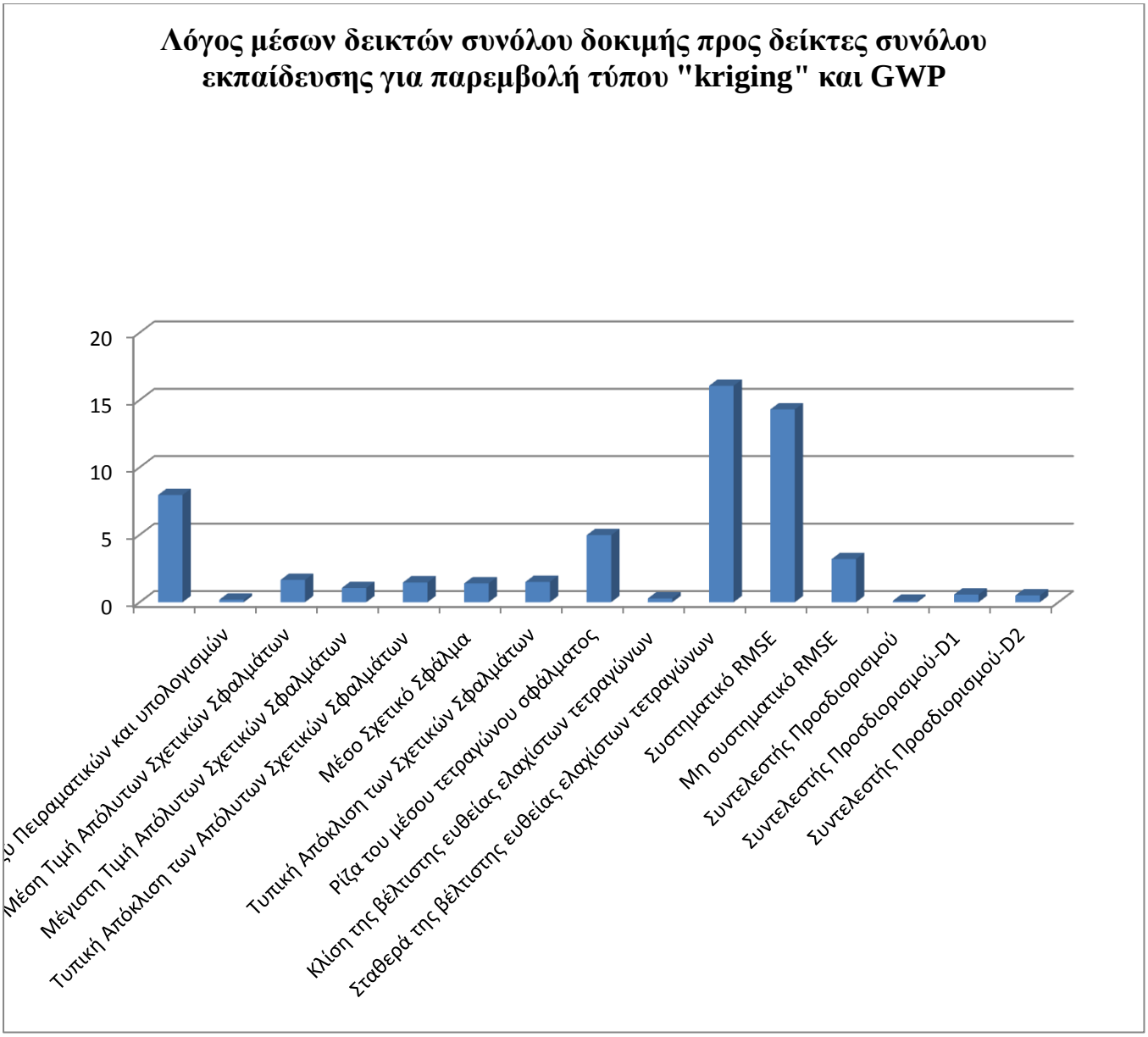
$$y = \frac{1}{x_{max} - x_{min}} \cdot x - \frac{x_{min}}{x_{max} - x_{min}} \text{ (5.12)}$$

Για κάθε στήλη, λοιπόν, του αρχικού πίνακα ευρίσκονται τα x_{min} και x_{max}, και εφαρμόζεται σε αυτά ο τύπος της κανονικοποίησης. Αποτέλεσμα είναι πως οι στήλες του κάθε πίνακα εισόδου θα έχουν τιμές από 0-1. Να σημειωθεί πως χρησιμοποιώντας τους ίδιους συντελεστές ($\frac{1}{x_{max} - x_{min}}$ και $\frac{x_{min}}{x_{max} - x_{min}}$) που εξήχθησαν από το σύνολο εκπαίδευσης, κανονικοποιούνται και οι τιμές των μεταβλητών στο σύνολο δοκιμής. Π.χ. αν η ομάδα CH_2 έχει ελάχιστο αριθμό εμφανίσεων στο σύνολο εκπαίδευσης ίσο με 1 και μέγιστο ίσο με 8, τότε στη κανονικοποίηση η στήλη του CH_2 θα κανονικοποιηθεί με x_{min}=1 και x_{max}=8. Αν η στήλη του CH_2 έχει στο σύνολο δοκιμής x_{min}=0 και x_{max}=5, τότε αυτοί οι αριθμοί δε θα χρησιμοποιηθούν, αλλά για να κανονικοποιηθεί το σύνολο δοκιμής, πάλι τα x_{min}=1 και x_{max}=8 θα χρησιμοποιηθούν. Αυτό συμβαίνει, ασφαλώς, ώστε τα δεδομένα και στο σύνολο δοκιμής και εκπαίδευσης να κανονικοποιούνται με τον ίδιο τρόπο και να υπάρχει κοινή βάση.

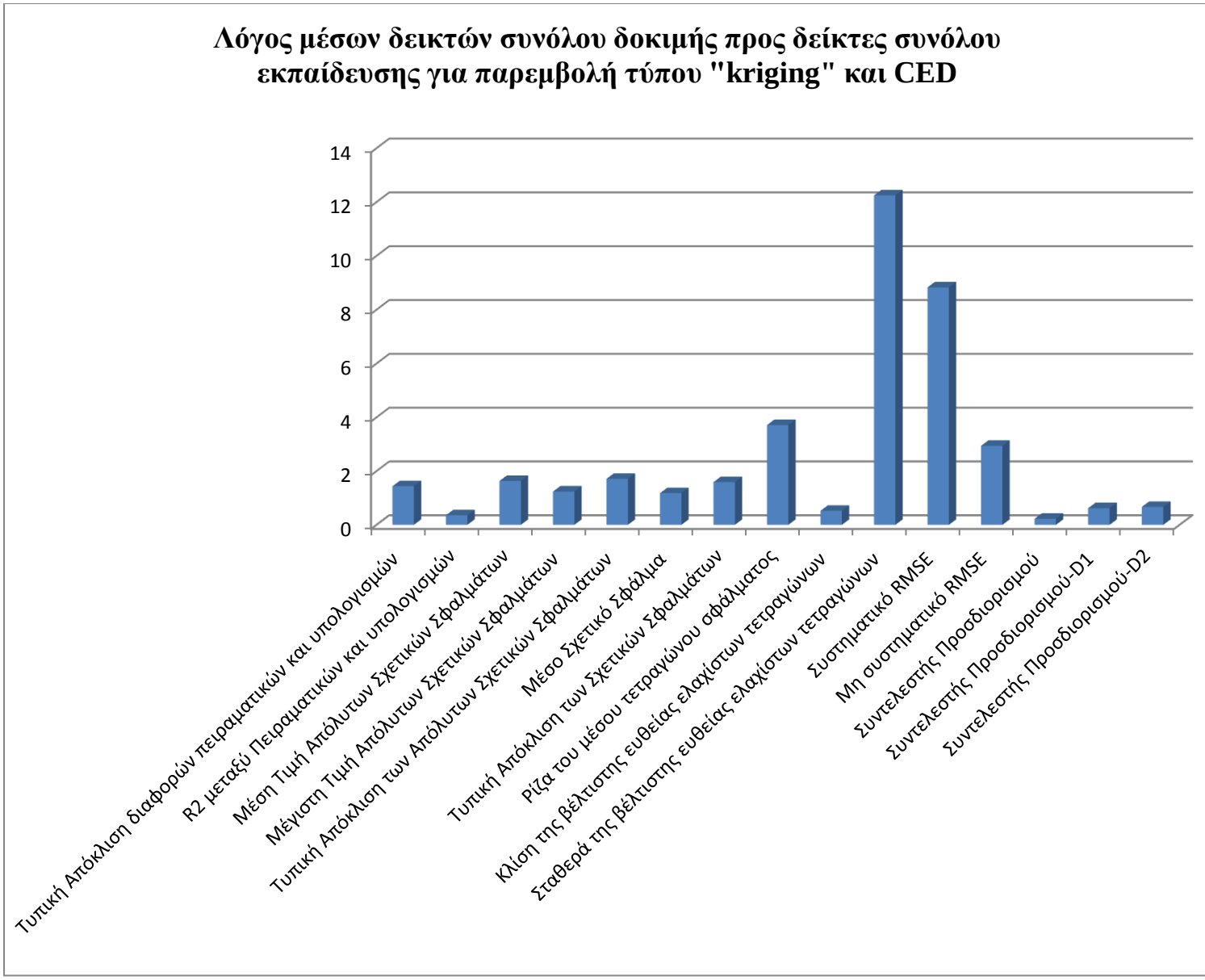
Όπως έχει ήδη αναφερθεί υπάρχουν μόρια τα οποία αντιπροσωπεύονται από μια ομάδα. Αυτές οι ομάδες, δηλαδή αποτελούν από μόνα τους μια ένωση και έχουν σθένος 0. Όπως είναι φυσικό, οι γραμμές

που αντιστοιχούν στα μόρια αυτά, στον αρχικό πίνακα εισόδου (214x57) θα έχουν όλα τα στοιχεία τους ίσα με 0 εκτός από τη στήλη εκείνη που βρίσκεται η ομάδα τους, όπου θα είναι ίσο με μονάδα, όπως και η στήλη των ομάδων αυτών θα έχουν όλα τα στοιχεία ίσα με 0, εκτός από τη γραμμή όπου υπάρχει η ένωση την οποία εκφράζουν. Στους διαμερισμούς εκείνους που κάποια από τις παραπάνω ενώσεις επιλέγεται να βρεθεί στο σύνολο δοκιμής, η στήλη από εκείνη τη μία ομάδα που εκφράζει την ένωση, έχει πλέον όλα τα στοιχεία της ίσα με 0. Αυτό είναι αναμενόμενο, αφού το μόνο μη μηδενικό στοιχείο αυτής της στήλης ανήκει σε μια γραμμή, που το μόριο της βρίσκεται στο σύνολο δοκιμής. Έτσι, στην αυτοματοποιημένη διαδικασία, η οποία εκτελεί τη κανονικοποίηση δημιουργούνται αρκετά προβλήματα, καθώς πλέον το εύρος της στήλης ($x_{max} - x_{min}$) είναι ίσο με 0. Άρα οι συντελεστές της κανονικοποίησης τείνουν στο άπειρο και τελικά, δε μπορεί να κανονικοποιηθεί το σύστημα. Για να αντιμετωπιστεί το πρόβλημα αυτό, ευρίσκονται όλες οι παραπάνω «προβληματικές» ενώσεις και καθορίζεται από τη διαδικασία, αυτές να συμμετέχουν πάντα στο σύνολο εκπαίδευσης, έτσι ώστε το εύρος των στηλών να είναι πάντα διάφορο του μηδενός (ίσο με ένα για τις συγκεκριμένες ομάδες). Αυτό επηρεάζει ασφαλώς, τον τρόπο που γίνονται οι διαμερισμού, αφού τελικά, με το να μένουν μόνιμα κάποιες ενώσεις πάντα στο σύνολο εκπαίδευσης, δεν επιτυγχάνεται ο αρχικός σκοπός τους, δηλαδή να συμμετέχουν όλες οι ενώσεις με διαφορετικούς συνδυασμούς στα επιμέρους σύνολα. Παρολαυτά, δεδομένου πως το αρχικό μας σύνολο δεν είναι αρκετά μεγάλο, η παραπάνω λύση κρίνεται ως κατάλληλη και απαραίτητη. Όπως προαναφέρθηκε μαζί με τις απαραίτητες παραμέτρους δίδεται και από το μοντέλο η παράμετρος λ.

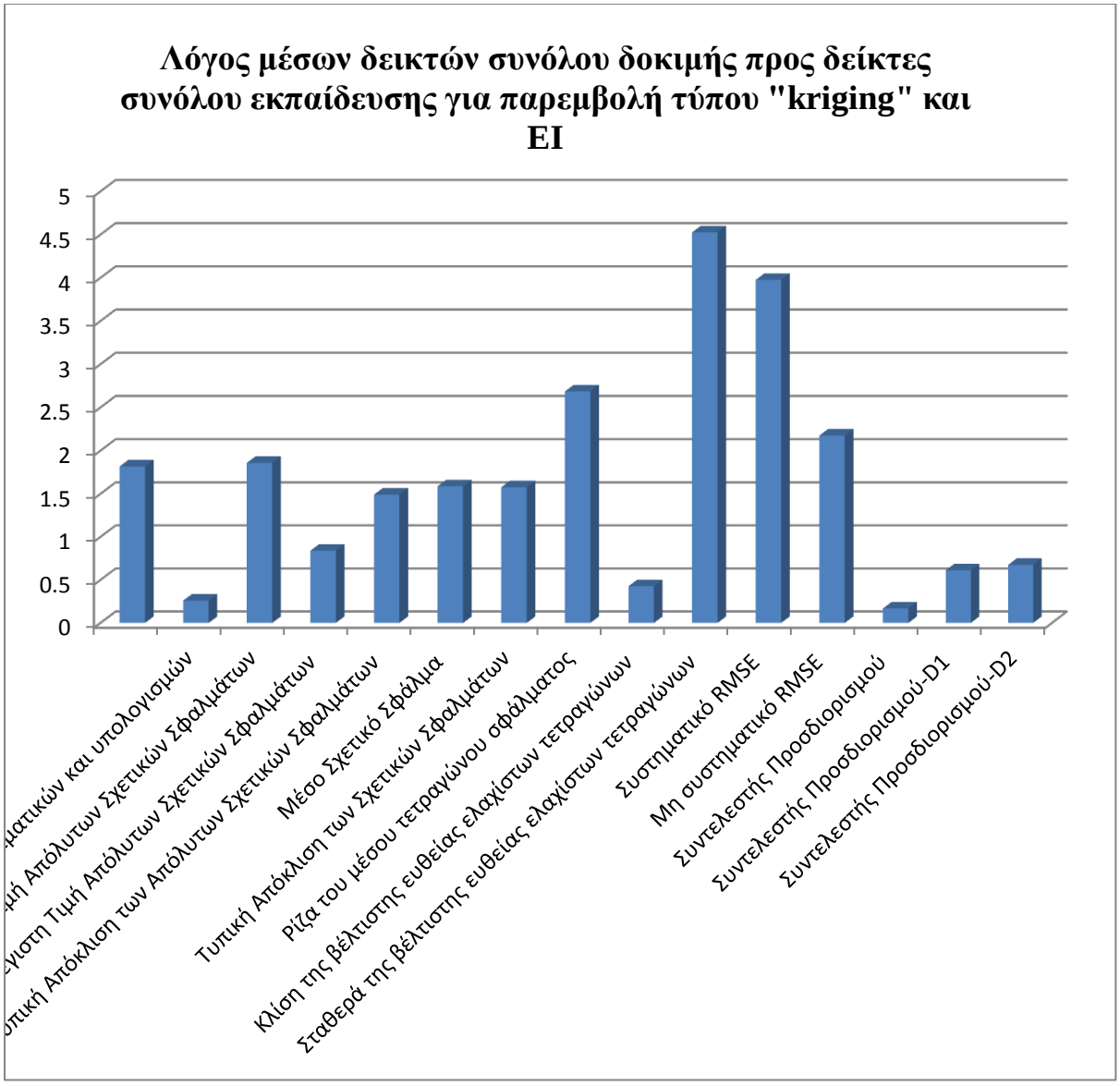
Μετά από τα παραπάνω ακολουθήθηκε η διαδικασία εύρεσης των μέσων στατιστικών δεικτών. Προέκυψαν τα διαγράμματα 5.27-5.29, που απεικονίζουν τους λόγους των μέσων στατιστικών δεικτών, ενώ τα αποτελέσματα βρίσκονται στο παράρτημα Ι:



Διάγραμμα 5.27. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για GWP και παρεμβολή τύπου «kriging»



Διάγραμμα 5.28. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για CED και παρεμβολή τύπου «kriging»

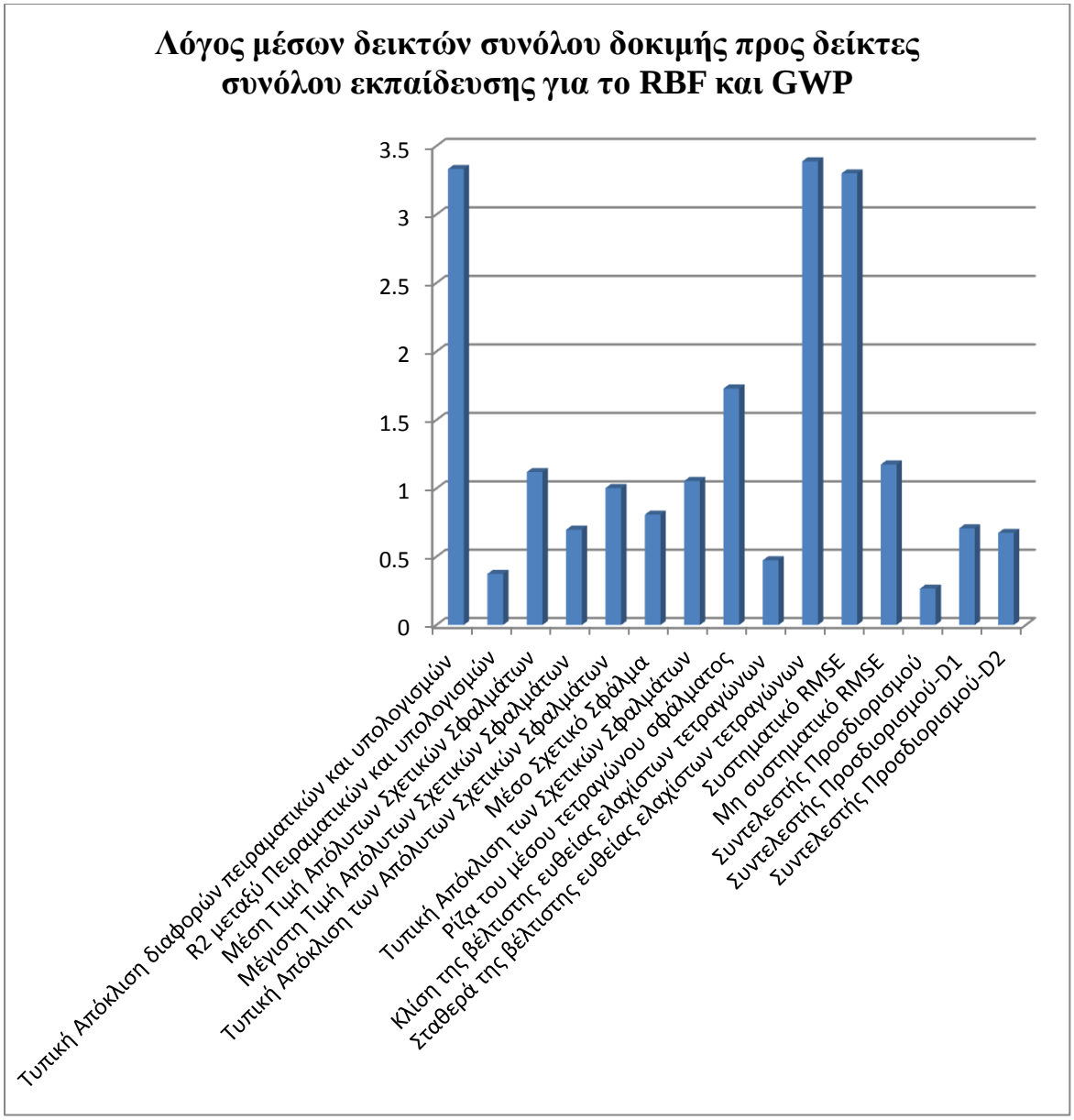


Διάγραμμα 5.29. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για EI 99 και παρεμβολή τύπου «kriging»

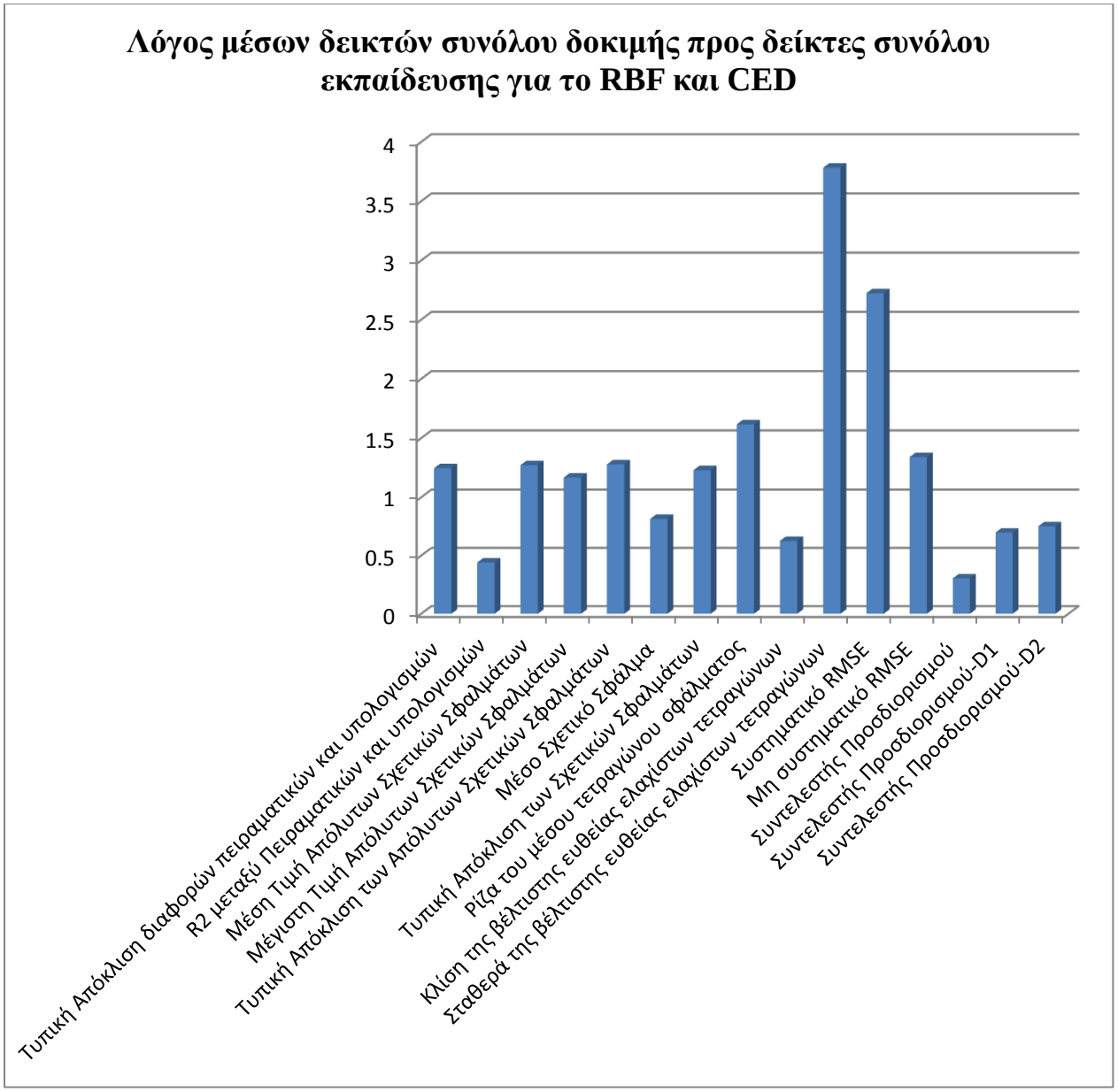
5.8 Μέθοδοι Συναρτήσεων Ακτινικής Βάσης

Όπως και στις προηγούμενες μεθόδους, έτσι και εδώ γίνεται, χωρισμός σε 1000 διαμερίσεις και υπολογισμός απόστασης mahalanobis για τα στοιχεία των συνόλων δοκιμής. Να σημειωθεί πως, επειδή και σε αυτή τη μέθοδο γίνεται κανονικοποίηση των τιμών που λαμβάνει κάθε μεταβλητή στο σύνολο εκπαίδευσης, είναι απαραίτητο για τις ομάδες που χρησιμοποιούνται μια φορά, να εξασφαλιστεί πως τα μόρια τα οποία αντιπροσωπεύονται από αυτές να βρίσκονται πάντα στο σύνολο εκπαίδευσης. Επίσης, σε αυτή τη μέθοδο, όπως έχει αναφερθεί και στο εισαγωγικό κεφάλαιο, ο χρήστης καθορίζει την πυκνότητα του πλέγματος. Πρέπει να καθοριστεί από τον χρήστη, λοιπόν, σε πόσα τμήματα θα χωρίζεται η κάθε

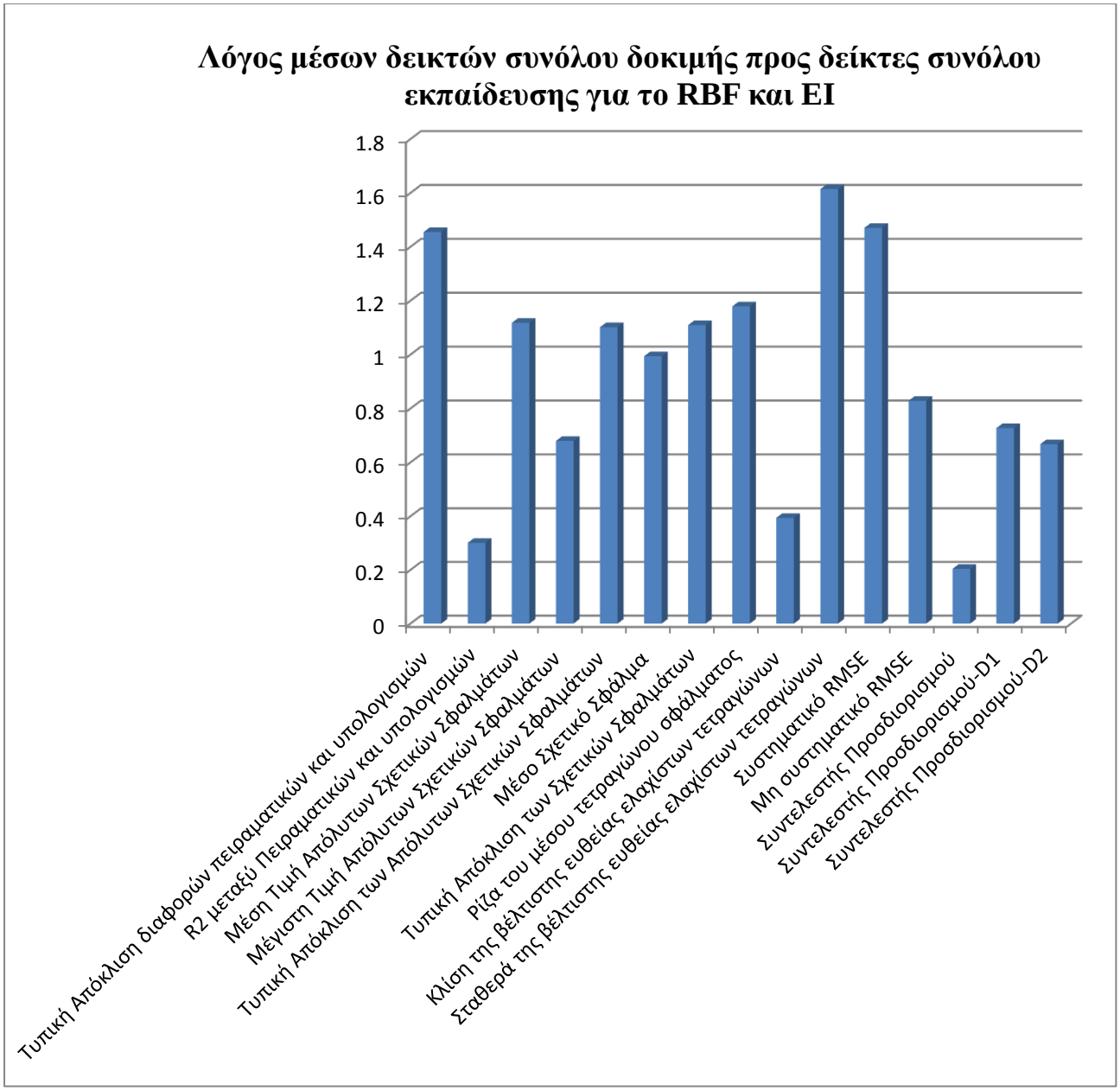
διάσταση. Θεωρείται πως η σωστότερη προσέγγιση σε αυτό το ζήτημα είναι η κάθε μεταβλητή να χωρίζεται σε τέτοιο αριθμό υποδιαστημάτων, ανάλογα με το μέγιστο αριθμό εμφανίσεων σε ένα μόριο (δηλαδή με το μέγιστο αριθμό σε μια στήλη). Η παράμετρος αυτή είναι θέμα ευριστικής αναζήτησης και αποτελεί παράμετρο βελτιστοποίησης. Το πρόβλημα βελτιστοποίησης αυτό δε μελετάται στα πλαίσια αυτής της εργασίας. Στη συγκεκριμένη περίπτωση καταλήγουμε να χωρίζονται οι δύο πρώτες μεταβλητές ,δηλαδή ,οι ομάδες $-CH_3$ και CH_2 σε 12 και 8 τμήματα αντίστοιχα και οι υπόλοιπες μεταβλητές σε 7. Να σημειωθεί πως ο κώδικας και η μεθοδολογία που χρησιμοποιείται είναι των Alexandridis et al. (2003). Αφού, λοιπόν, κανονικοποιηθούν τα δεδομένα, εισάγονται στο μοντέλο, οι κανονικοποιημένες μεταβλητές, μαζί με την έξοδο του μοντέλο (πειραματικά) και ένα διάνυσμα 1x57 ,όπου κάθε στοιχείο του περιλαμβάνει τον αριθμό των υποδιαστημάτων που χωρίζεται η κάθε μεταβλητή. Αυτό το διάνυσμα έχει τα δύο πρώτα του στοιχεία ίσα με 12 και 8 και τα υπόλοιπα ίσα με 7. Για τις εκτιμήσεις των δεικτών για τον κάθε διαμερισμό με τα πειραματικά δεδομένα και για όλους τους διαμερισμούς μαζί, εξάγονται οι λόγοι των μέσων στατιστικών δεικτών, συνολικά για το RBF (διαγράμματα 5.30-5.32), ενώ τα αποτελέσματα βρίσκονται στο παράρτημα I:



Διάγραμμα 5.30. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για GWP και RBF



Διάγραμμα 5.31. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για CED και RBF

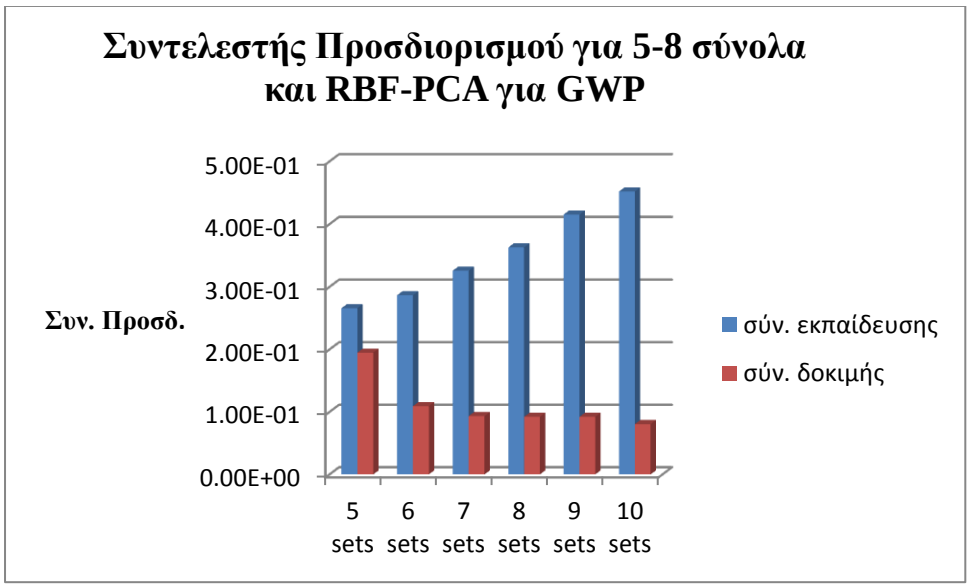


Διάγραμμα 5.32. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για EI 99 και RBF

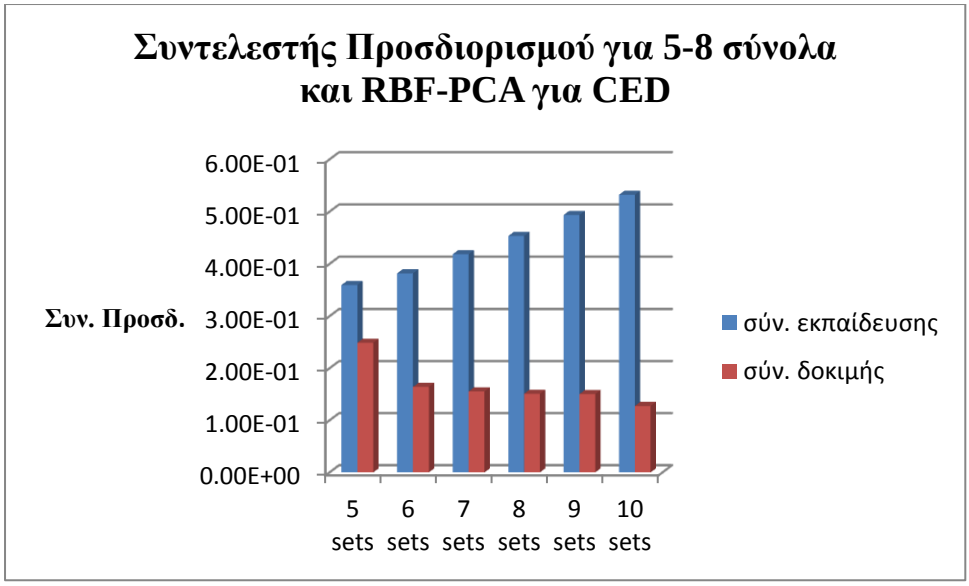
5.9 Μέθοδοι Συναρτήσεων Ακτινικής Βάσης-Ανάλυση Πρωτευόντων Συνιστωσών

Όπως και στις προηγούμενες μεθόδους, έτσι και εδώ γίνεται, χωρισμός σε 1000 διαμερίσεις και υπολογισμός απόστασης mahalanobis για τα στοιχεία των συνόλων δοκιμής. Ακουλουθεί η διαδικασία RBF, αλλά αντί να χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές για τη δημιουργία του πλέγματος (grid) οι ομάδες, συμμετέχουν τα PC's που έχουν βρεθεί από την προηγούμενη μεθοδολογία. Συγκεκριμένα, θα χρησιμοποιηθούν, όπως είναι φυσικό, 9 PC's μιας και αυτά μας δίνουν τα καλύτερα αποτελέσματα. Αυτό

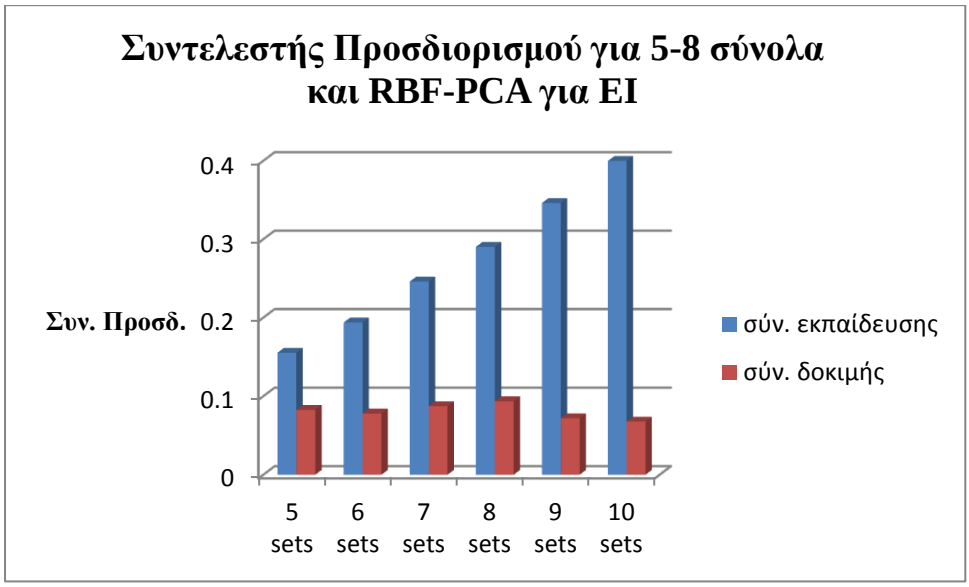
έχει ως αποτέλεσμα, την ακόμα μεγαλύτερη μείωση της πυκνότητας του πλέγματος, αφού χρησιμοποιούνται λιγότερες μεταβλητές και συνεπώς, θα δημιουργηθούν και λιγότερα κέντρα. Παράλληλα, όμως, δε χάνεται κάποιο κομμάτι πληροφορίας, μιας και όπως έχει αποδειχτεί, τα 9 PC's ενσωματώνουν υψηλό ποσοστό της αρχικής διακύμανσης. Πραγματοποιείται, όπως και πριν, ο χωρισμός σε 1000 διαμερισμούς, αυτή τη φορά με τα 9 PC's ως ανεξάρτητες μεταβλητές και υπολογίζεται για το κάθε σύνολο δοκιμής η απόσταση mahalanobis για το κάθε μόριο. Το πρόβλημα που αντιμετωπίστηκε εδώ, ήταν σε πόσα υποδιαστήματα πρέπει να χωριστεί η κάθε μεταβλητή, δηλαδή, το κάθε PC. Επειδή το πρόβλημα αυτό απαιτεί ξεκάθαρα ευριστική λύση, αποφασίζεται να χωρίζονται τα PC's σε ίσα διαστήματα όλα τα PC's. Η διαδικασία ξεκινάει, λοιπόν, με το χωρισμό σε 5-10 διαστήματα. Για τον κάθε διαφορετικό χωρισμό σε υποδιαστήματα, αναπτύσσεται το μοντέλο, όπως και πριν και συνολικά για όλες τις διαμερίσεις, εξάγονται οι μέσοι στατιστικοί δείκτες. Τέλος, συγκρίνεται ο μέσος Συντελεστής Προσδιορισμού και για τα 6 διαφορετικά υποδιαστήματα, για να αποφασιστεί ποιος χωρισμός σε υποδιαστήματα επέτρεψε στα 9 PC's να φτιάξουν το καλύτερο πλέγμα για το RBF. Αφού, αναπτυχθούν οι συσχετίσεις, όπως κάθε φορά, υπολογίζονται οι μέσοι στατιστικοί δείκτες. Η σύγκριση των συντελεστών συσχέτισης παρουσιάζονται στα διαγράμματα 5.33-5.35 για τους τρεις διαφορετικούς δείκτες:



Διάγραμμα 5.33. Σύγκριση του Συντελεστή Προσδιορισμού για τους διάφορους χωρισμούς σε υποδιαστήματα για το GWP

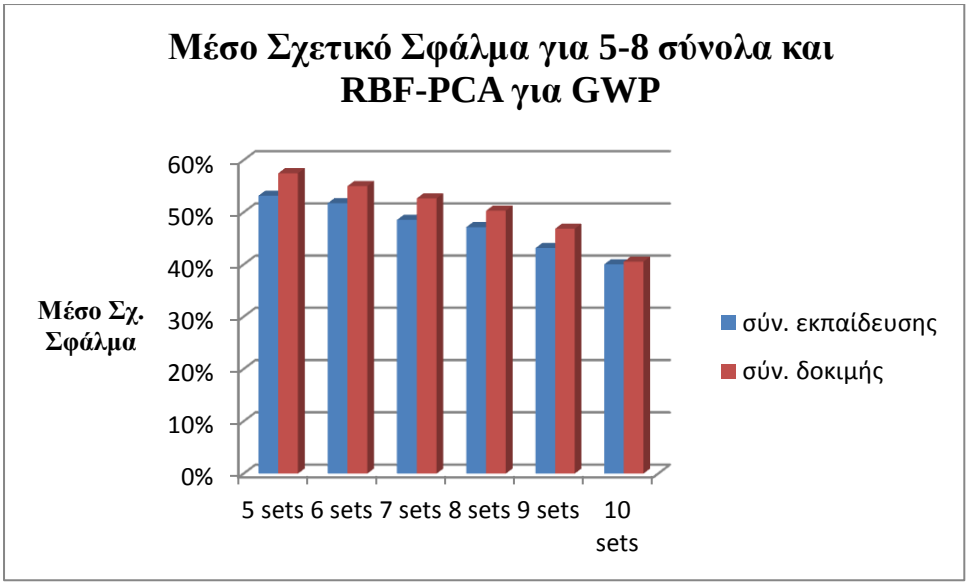


Διάγραμμα 5.34. Σύγκριση του Συντελεστή Προσδιορισμού για τους διάφορους χωρισμούς σε υποδιαστήματα για το CED

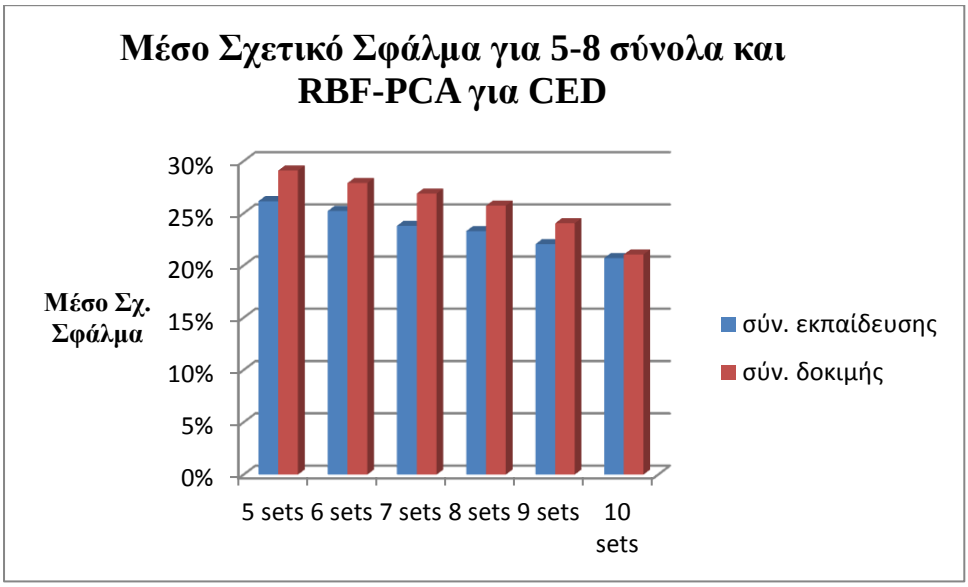


Διάγραμμα 5.35. Σύγκριση του Συντελεστή Προσδιορισμού για τους διάφορους χωρισμούς σε υποδιαστήματα για το EI 99

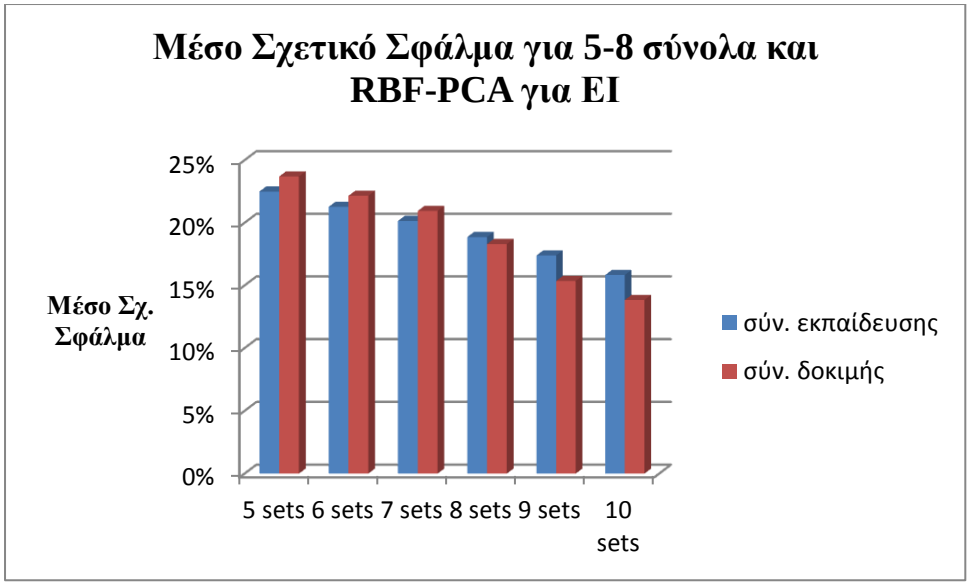
Για το GWP και το CED, φαίνεται πως τη βέλτιστη συσχέτιση την επιτυγχάνει ο διαχωρισμός σε 5 υποδιαστήματα στο κάθε PC, ενώ στο EI 99 έχει ένα ελαφρύ προβάδισμα ο διαχωρισμός στα 8 διαστήματα. Παρλαυτά, τα 5 διαστήματα θεωρούνται ο βέλτιστος διαχωρισμός. Για λόγους αναφοράς φαίνονται και τα συγκριτικά για το μέσο σφάλμα στα διαγράμματα 5.36-5.38:



Διάγραμμα 5.36. Συγκριτικά μέσου σφάλματος για τους διάφορους διαχωρισμούς για το GWP

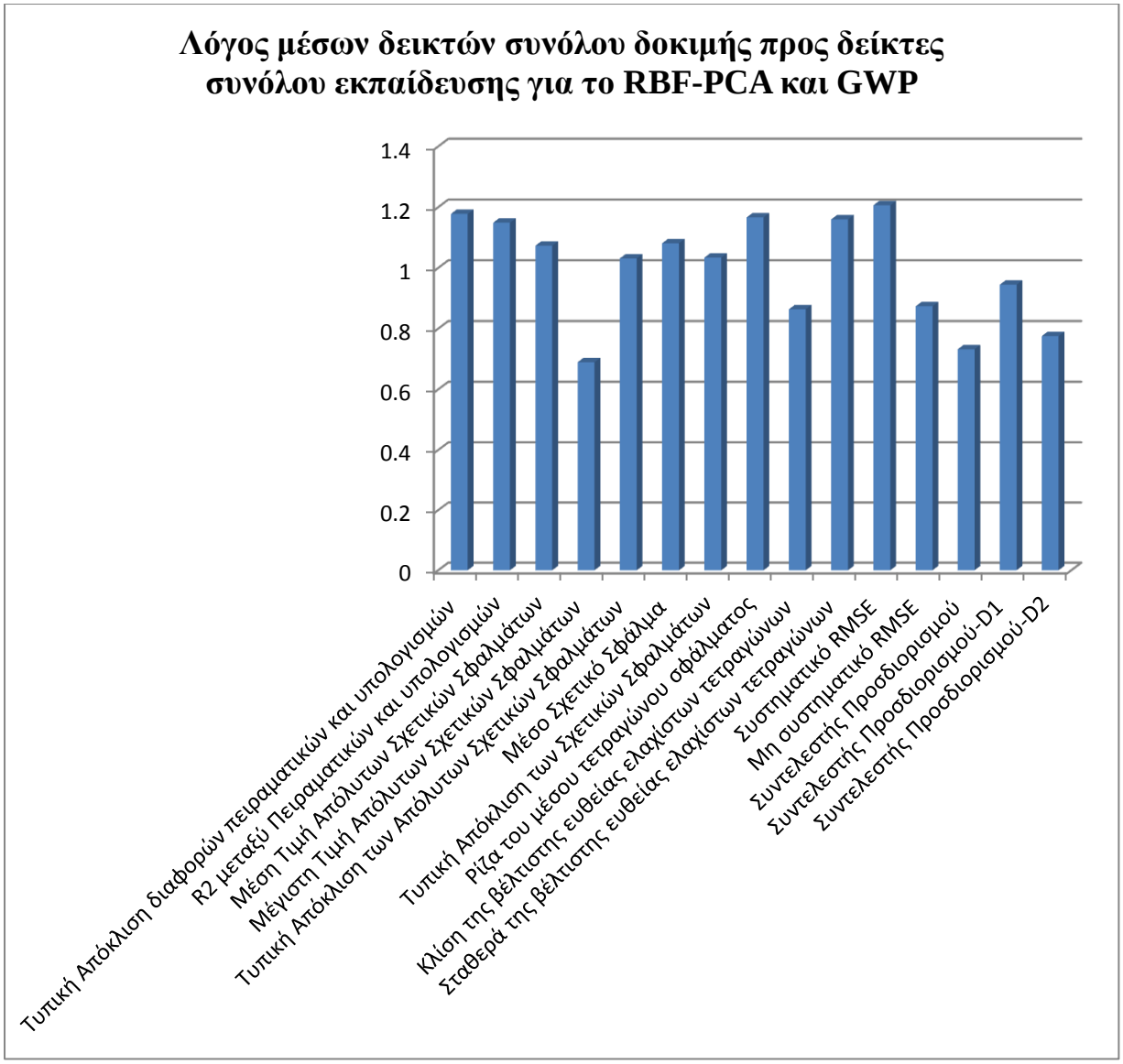


Διάγραμμα 5.37. Συγκριτικά μέσου σφάλματος για τους διάφορους διαχωρισμούς για το CED

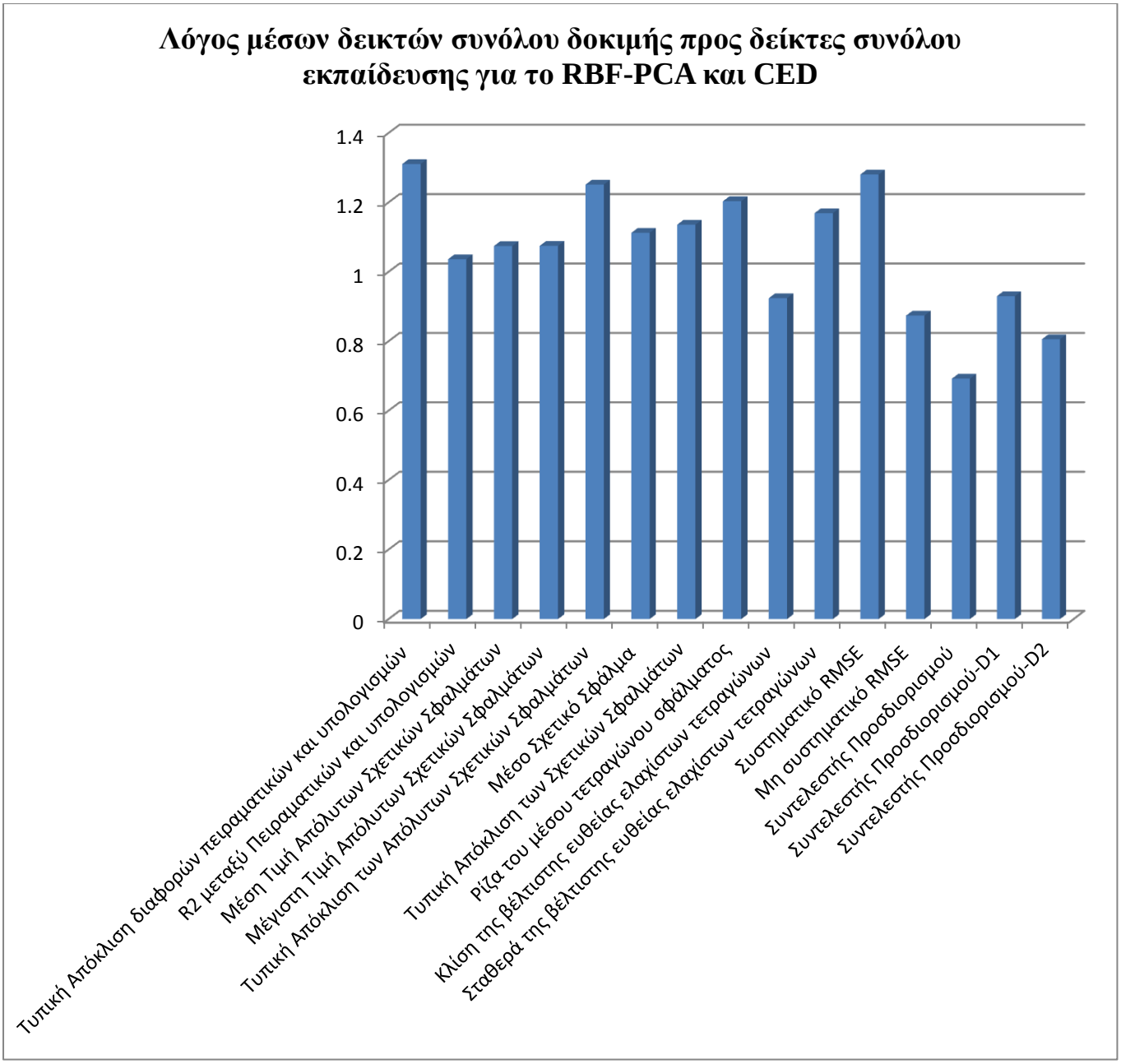


Διάγραμμα 5.38. Συγκριτικά μέσου σφάλματος για τους διάφορους διαχωρισμούς για το EI 99

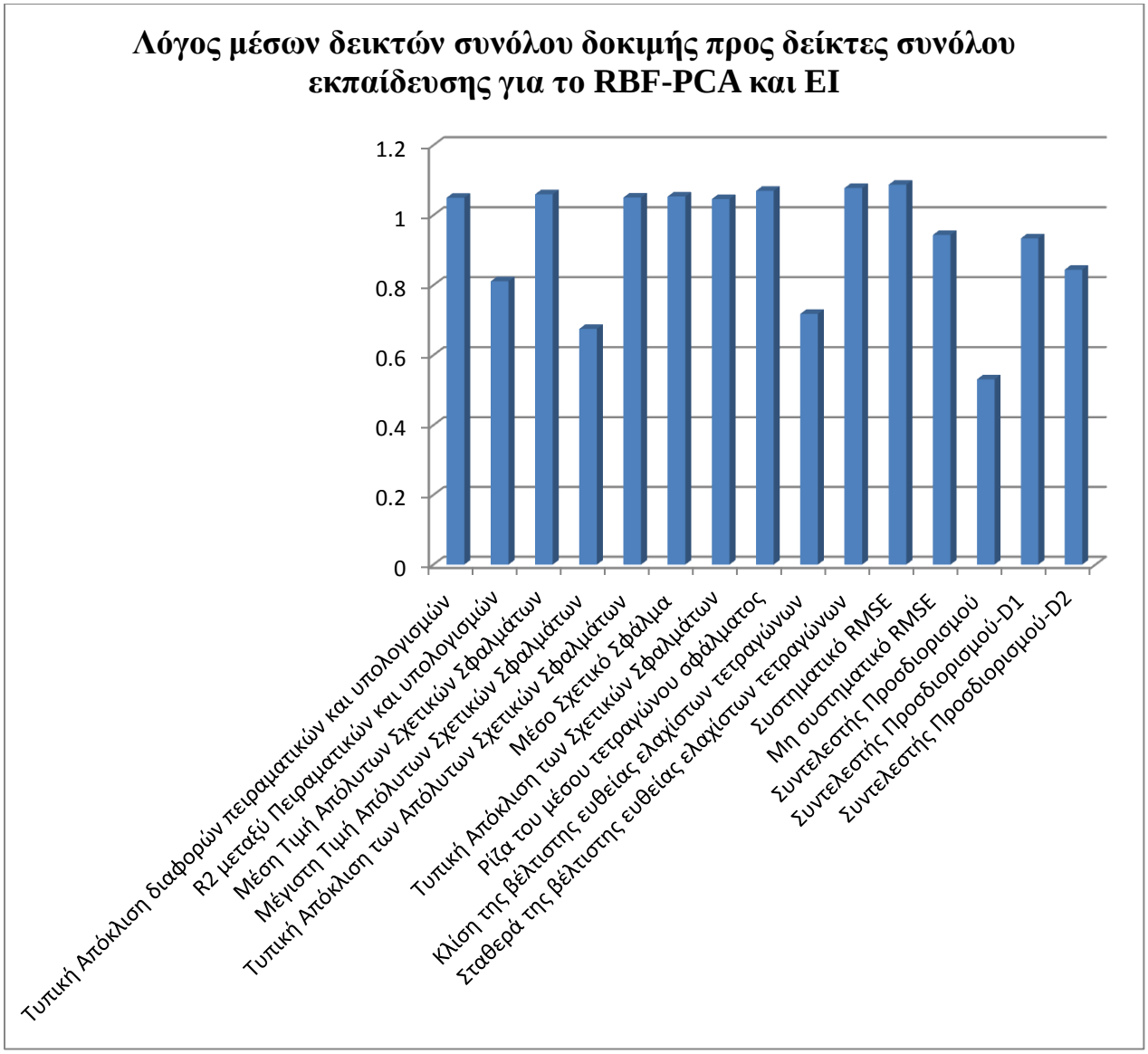
Παρατηρείται πως το σφάλμα ακολουθεί φθίνουσα τάση στην αύξηση του αριθμού των υποδιαστημάτων. Συνεπώς, το σφάλμα των 5 υποδιαστημάτων είναι μεγαλύτερο περίπου κατά 10%. Αυτό δε θεωρείται, όμως, πως υποσκιάζει το όφελος από τον υψηλότερο Συντελεστή Προσδιορισμού. Εν τέλει, το καλύτερο μοντέλο προκύπτει να είναι αυτό με 9 PC's, στο οποίο χωρίζεται το κάθε PC σε 5 υποδιαστήματα. Στο παράρτημα Ι παρουσιάζονται τα αποτελέσματα της μεθόδου RBF-PCA. Στα διαγράμματα 5.39-5.41 εμφανίζονται οι λόγοι των μέσων στατιστικών δεικτών για το RBF-PCA.



Διάγραμμα 5.39. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για GWP και RBF-PCA



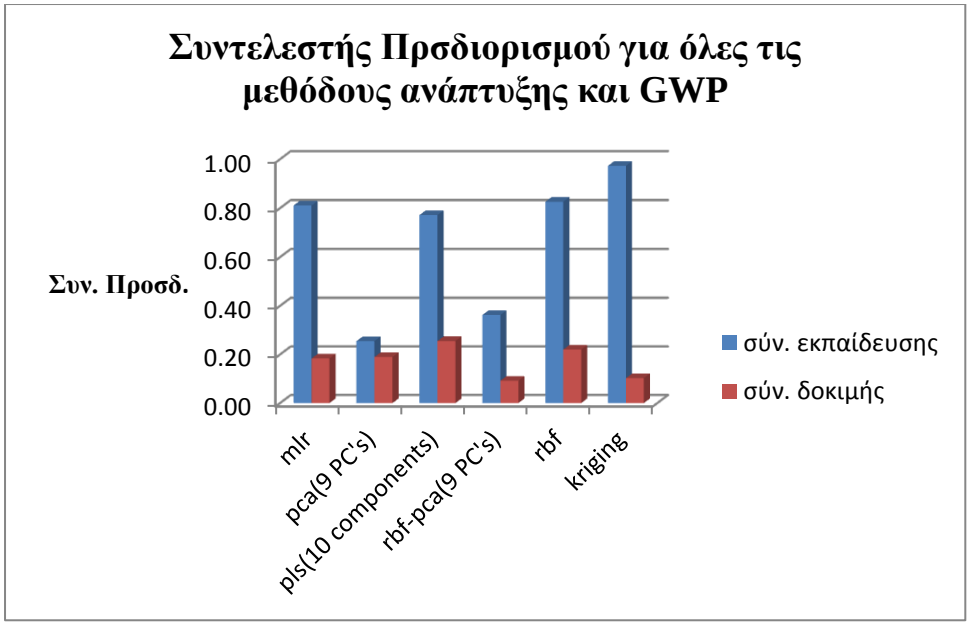
Διάγραμμα 5.40. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για CED και RBF-PCA



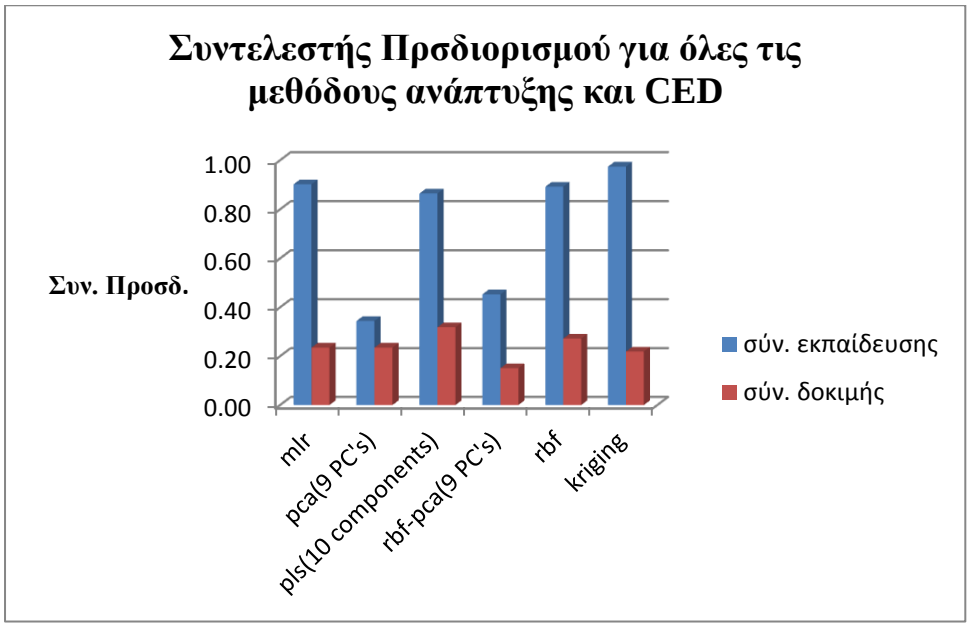
Διάγραμμα 5.41. Λόγοι μέσων στατιστικών δεικτών συνόλου εκπαίδευσης προς δοκιμής για EI 99 και RBF-PCA

5.10 Σύγκριση των μοντέλων

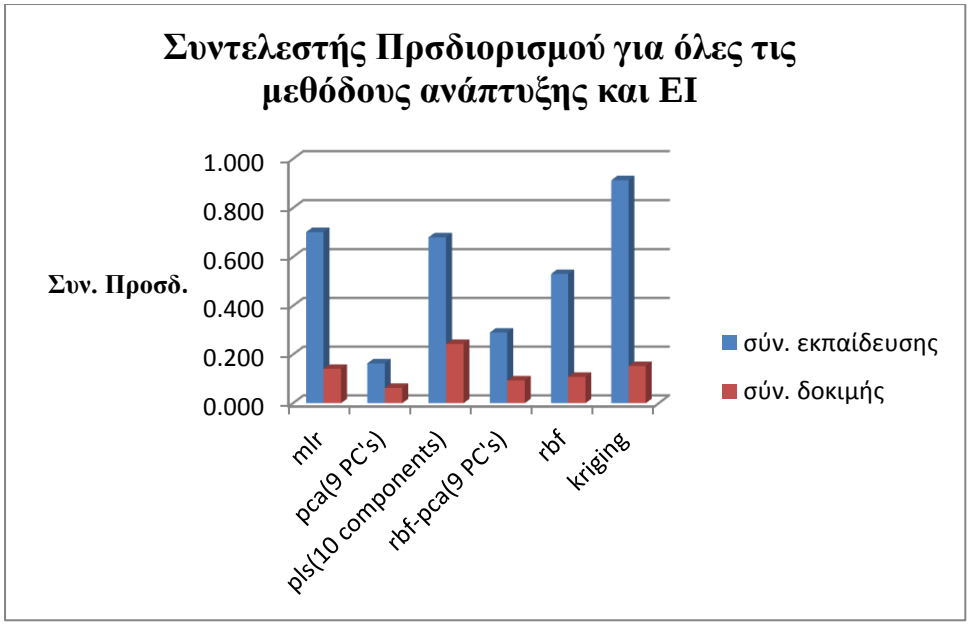
Τα μοντέλα που αναπτύσσονται με την κάθε μεθοδολογία συγκρίνονται μεταξύ τους, ώστε να φανεί ποια μεθοδολογία αναπτύσσει την καλύτερη συσχέτιση. Όπως είναι φυσικό, το μέτρο σύγκρισης για την ποιότητα της συσχέτισης είναι το Συντελεστή Προσδιορισμού. Στα διαγράμματα 5.42-5.44, παρατίθεται το διάγραμμα σύγκρισης των συντελεστών αυτών για το βέλτιστο μοντέλο της κάθε μεθοδολογίας: το μοντέλο MLR, το μοντέλο PCA/PCR με 9 PC's, το μοντέλο PLS με 10 συνιστώσες, το μοντέλο παρεμβολής τύπου «kriging», το μοντέλο RBF και τέλος, το μοντέλο RBF-PCA με 9 PC's και 5 υποδιαστήματα σε κάθε PC.



Διάγραμμα 5.42. Σύγκριση Συντελεστών Προσδιορισμού για όλα τα μοντέλα GWP

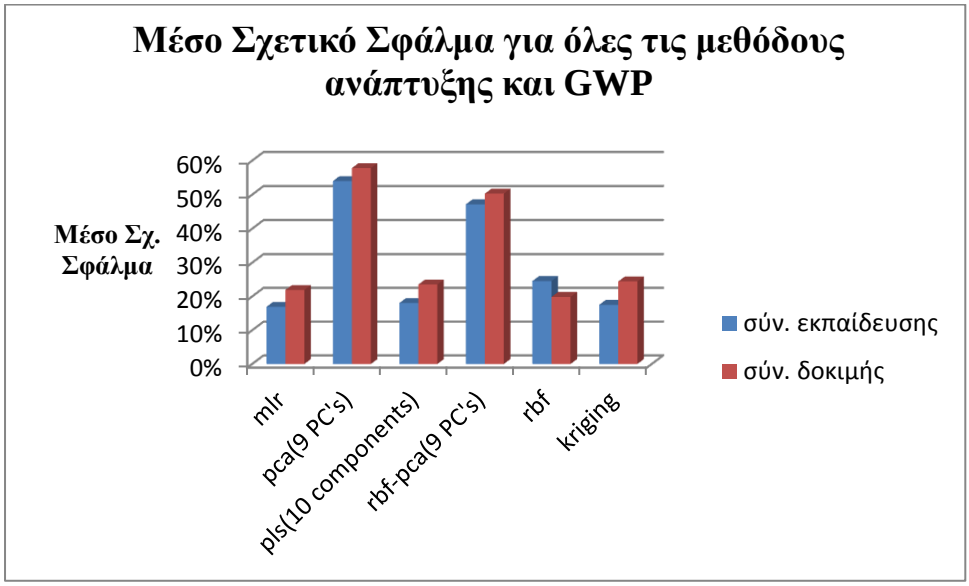


Διάγραμμα 5.43. Σύγκριση Συντελεστών Προσδιορισμού για όλα τα μοντέλα CED

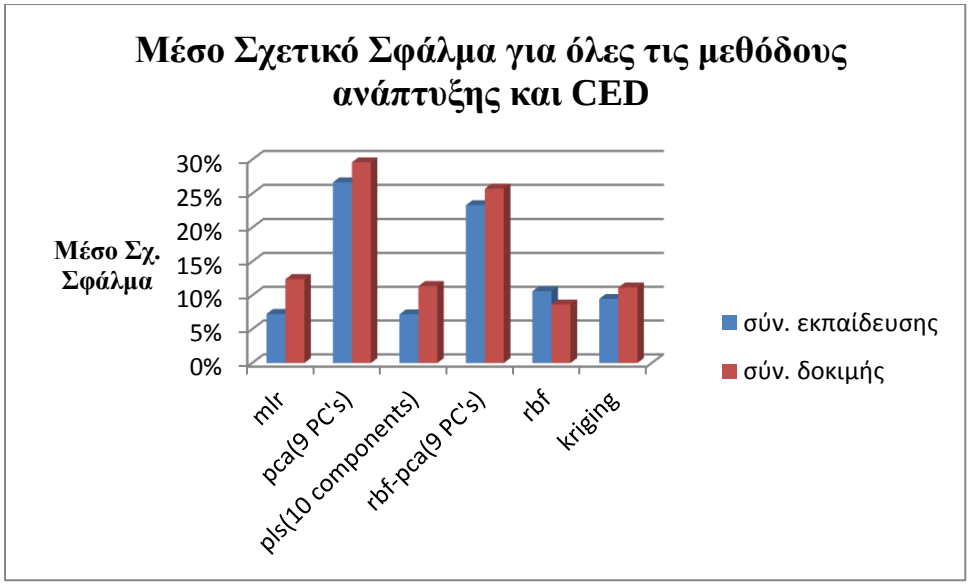


Διάγραμμα 5.44. Σύγκριση Συντελεστών Προσδιορισμού για όλα τα μοντέλα EI 99

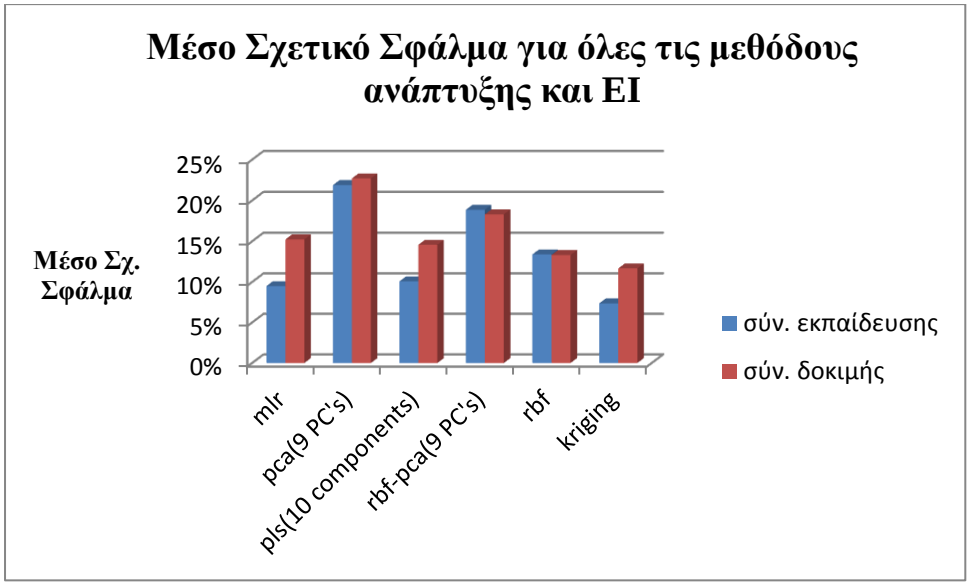
Είναι, πλέον, ξεκάθαρο πως το μοντέλο PLS με 10 συνιστώσες επιτυγχάνει την καλύτερη συσχέτιση για όλα τα μοντέλα, με Συντελεστές Προσδιορισμού για το GWP, CED και EI 99 ίσα με 0.2552, 0.3184 και 0.2436 αντίστοιχα. Στα διαγράμματα 5.45-5.47, παρουσιάζονται τα συγκριτικά των μέσων σφαλμάτων για όλα τα μοντέλα και όλους τους δείκτες:



Διάγραμμα 5.45. Συγκριτικά μέσου σφάλματος μεταξύ των 6 μοντέλων GWP



Διάγραμμα 5.46. Συγκριτικά μέσου σφάλματος μεταξύ των 6 μοντέλων CED



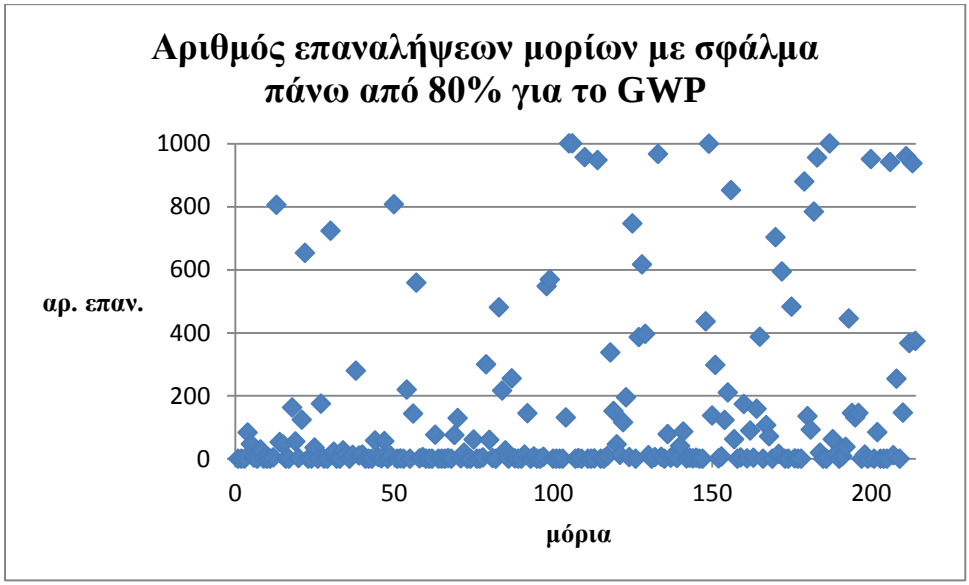
Διάγραμμα 5.47. Συγκριτικά μέσου σφάλματος μεταξύ των 6 μοντέλων EI 99

Έστω και αν δεν είναι το μοντέλο των 10 συνιστωσών δεν επιτυγχάνει την ελαχιστοποίηση του σφάλματος, αποτελεί ένα μοντέλο με τα χαμηλότερα σφάλματα. Αυτό το γεγονός, σε συνδυασμό με τη βέλτιστη ποιότητα συσχέτισης, μας οδηγεί στο να καταλήξουμε πως το μοντέλο PLS με 10 συνιστώσες είναι το βέλτιστο.

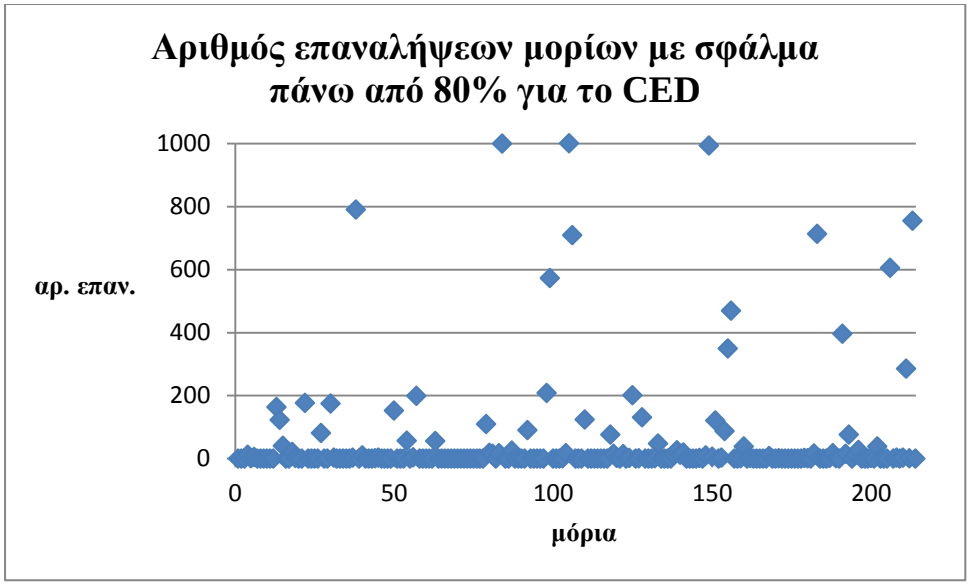
5.11 Διαδικασία Ανάλυσης (Post-Analysis) του μοντέλου

Αφού έχει αποφασιστεί πως το καλύτερο μοντέλο που αναπτύχθηκε ήταν το μοντέλο από το PLS με 10 συνιστώσες, εκτελούνται τα απαραίτητα βήματα για να εκτιμηθεί η αξιοπιστία του μοντέλου και τα αποτελέσματα του, καθώς και οι δυνατότητες για σωστές προβλέψεις μορίων.

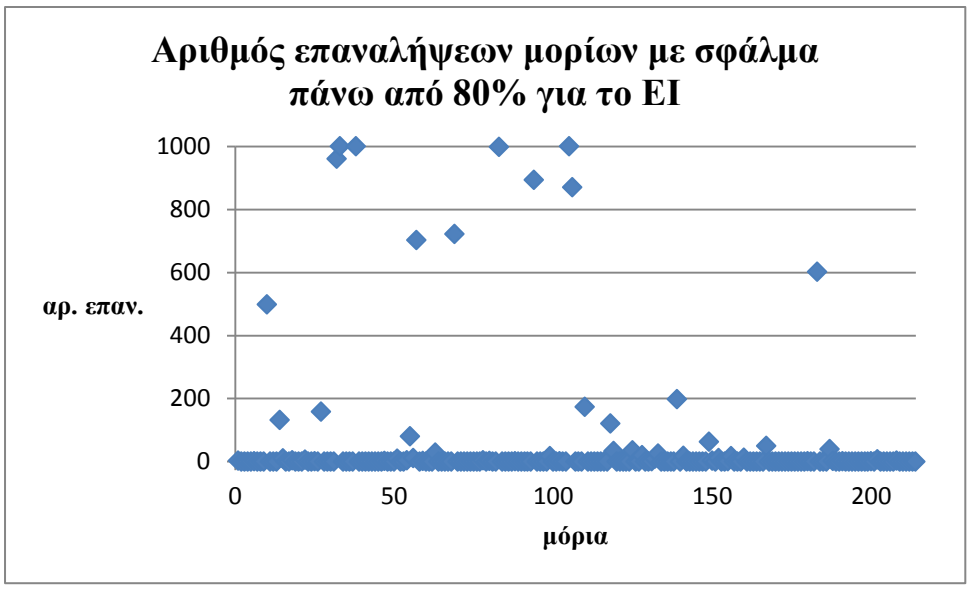
Το πρώτο βήμα που εκτελείται είναι η καταγραφή του αριθμού των φορών, που κάθε μόριο του συνόλου (των 214 μορίων) χρησιμοποιείται, εμφανίζεται να έχει σφάλμα πάνω από 80% στο σύνολο εκπαίδευσης και δοκιμής. Καθένα, δηλαδή, από τα 214 μόρια του αρχικού συνόλου συνοδεύεται πλέον από 2 αριθμούς: ο ένας υποδηλώνει πόσες φορές αυτό το μόριο εμφανίστηκε στα 1000 διαφορετικά σύνολα εκπαίδευσης να έχει σφάλμα πάνω από 80% και ο δεύτερος, πόσες φορές το ίδιο μόριο εμφανίστηκε σε κάθε ένα από τα 1000 σύνολα δοκιμής με σφάλμα πάνω από 80%. Έπειτα, αυτοί οι δυο αριθμοί προστίθενται, ώστε να φανεί συνολικά κάθε μόριο, πόσες φορές εμφανίστηκε να έχει σφάλμα πάνω από 80% (εννοείται πως μεγαλύτερη σημασία έχουν πάντα τα στοιχεία που μας πληροφορούν για το σύνολο δοκιμής). Επειδή, το σφάλμα 80% είναι αρκετά μεγάλο, το άθροισμα αυτό μπορεί να παρέχει πληροφορίες σχετικά με το ποια μόρια είναι γενικώς «προβληματικά», διότι εμφανίζονται συστηματικά να έχουν ένα σημαντικό σφάλμα και να παρατηρηθούν τάσεις σχετικά με χαρακτηριστικές ομάδες μορίων. Οι παραπάνω αριθμοί τοποθετούνται σε ένα διάγραμμα τύπου scatter (στον οριζόντιο άξονα ο αύξων αριθμός του κάθε μορίου στην αρχική λίστα) (διαγράμματα 5.48-5.50):



Διάγραμμα 5.48. Αριθμός φορών εμφάνισης κάθε μορίου στο σύνολο δεδομένων με σφάλμα πάνω από 80% για το GWP



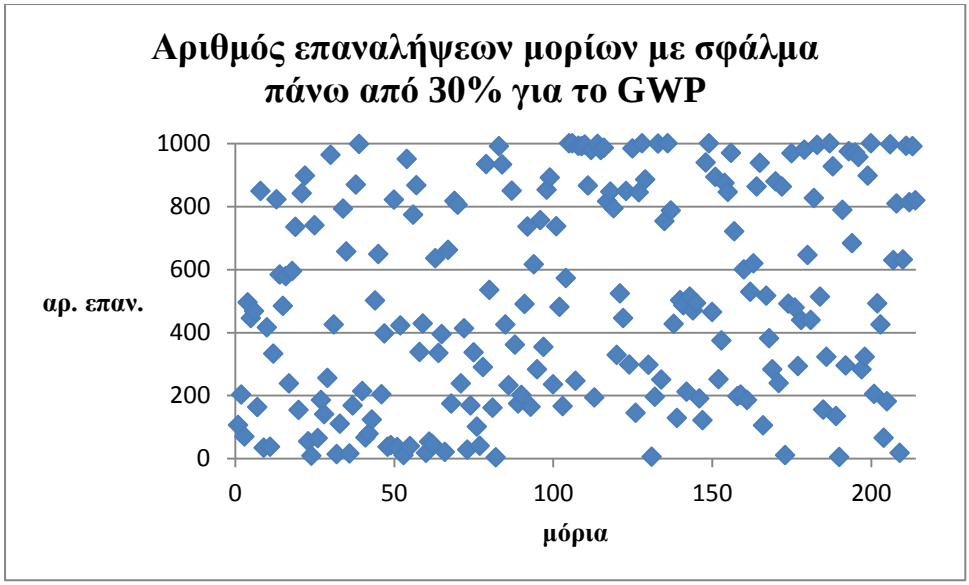
Διάγραμμα 5.49. Αριθμός φορών εμφάνισης κάθε μορίου στο σύνολο δεδομένων με σφάλμα πάνω από 80% για το CED



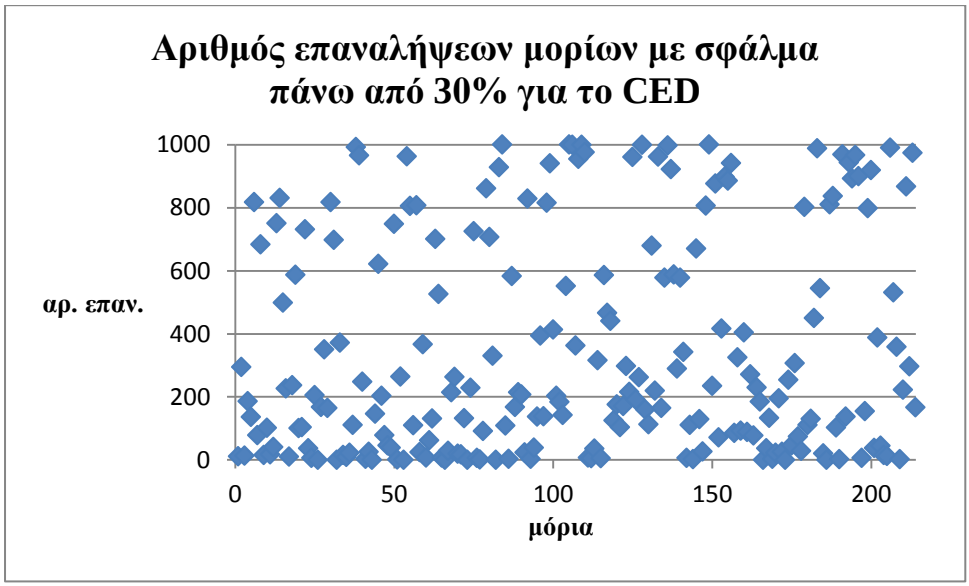
Διάγραμμα 5.50. Αριθμός φορών εμφάνισης κάθε μορίου στο σύνολο δεδομένων με σφάλμα πάνω από 80% για το EI 99

Όπως φαίνεται, τα περισσότερα μόρια παραμένουν σε ποσοστά κάτω από 80% (βρίσκονται επάνω στον οριζόντιο άξονα). Ο μεγαλύτερος αριθμός μορίων που «ξεφεύγει» από τον οριζόντιο άξονα βρίσκεται στις προβλέψεις του GWP, αποδεικνύοντας πως είναι το μοντέλο στο οποίο γίνονται δυσκολότερα αποτελεσματικές συσχετίσεις.

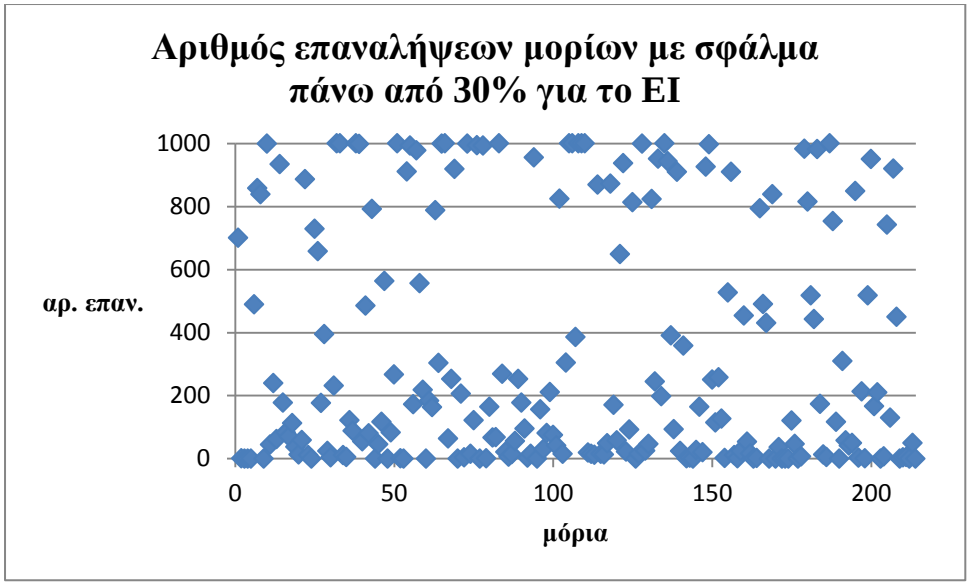
Για συγκριτικούς λόγους και για να φανεί η διασπορά του σφάλματος των μορίων σε διάφορες κατηγορίες σφάλματος επαναλαμβάνεται η παραπάνω διαδικασία καταγραφής των φορών εμφάνισης των μορίων με σφάλμα πάνω από 30% (διαγράμματα 5.51-5.53).



Διάγραμμα 5.51. Αριθμός φορών εμφάνισης κάθε μορίου στο σύνολο δεδομένων με σφάλμα πάνω από 30% για το GWP



Διάγραμμα 5.52. Αριθμός φορών εμφάνισης κάθε μορίου στο σύνολο δεδομένων με σφάλμα πάνω από 30% για το CED



Διάγραμμα 5.53. Αριθμός φορών εμφάνισης κάθε μορίου στο σύνολο δεδομένων με σφάλμα πάνω από 30% για το EI 99

Όπως είναι λογικό, καθώς το κατώφλι σφάλματος συνεχώς μειώνεται, στα διαγράμματα που ακολουθούν, όλο και περισσότερα μόρια «μεταναστεύουν» από τον οριζόντιο άξονα (όπου έχουν μηδενικές επαναλήψεις) και το «νέφος» των σημείων πάνω από τον οριζόντιο άξονα πυκνώνει. Είναι σημαντικό να επισημανθεί στις κατανομές των μορίων, πως στους δείκτες των CED και EI 99, ακόμα και όταν τα όρια των σφαλμάτων έχουν μειωθεί συνεχίζουν να υπάρχουν αρκετά μόρια που δεν έχουν παρουσιάσει σφάλμα πάνω από αυτά τα όρια και συνεχίζουν να βρίσκονται πάνω στον οριζόντιο άξονα. Αυτό επιβεβαιώνει πως τα μοντέλα αυτά προβλέπουν με πολύ ικανοποιητική ακρίβεια τους δείκτες από νέες μοριακές δομές. Εν αντιθέσει, στο μοντέλο στο GWP, τα περισσότερα μόρια εμφανίζονται να έχουν αναλογικά μεγαλύτερα σφάλματα.

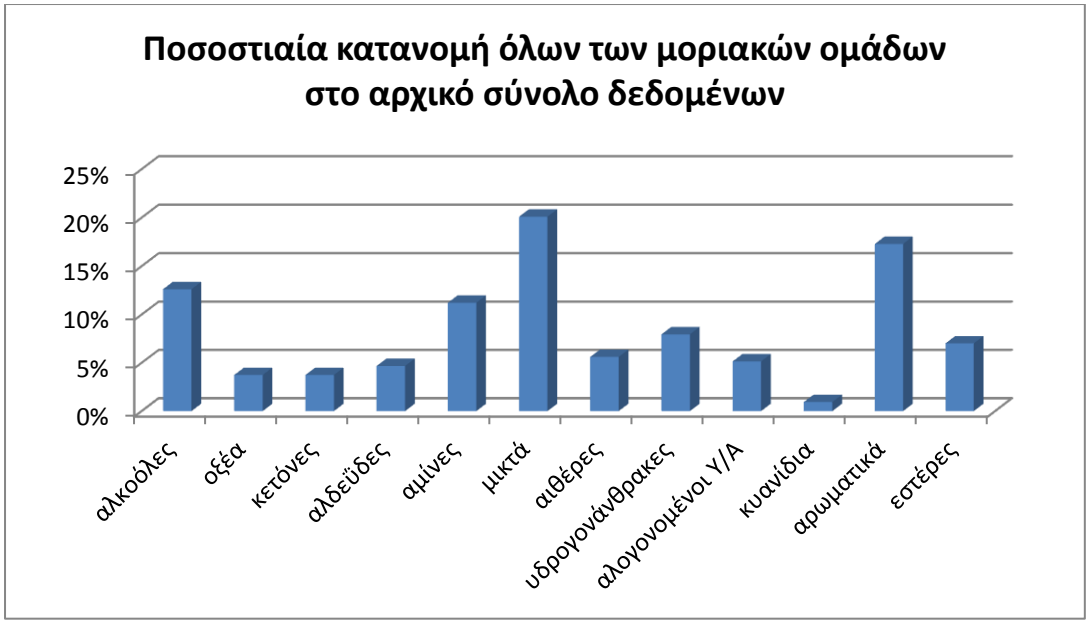
Όπως έχει γίνει αναφερθεί, για το μοντέλο που είναι βέλτιστο (στη συγκεκριμένη περίπτωση το PLS με 10 συνιστώσες), εξάγονται οι μέσες τιμές των παραμέτρων που είναι απαραίτητες για να εξαχθούν οι νέες προβλέψεις. Για κάθε διαμέριση, υπάρχει και από μια τιμή συνεισφοράς (contribution) κάθε ομάδας. Σκοπός είναι η εξαγωγή του τελικού διανύσματος συνεισφορών,1x58 (57 ομάδες και ο σταθερός όρος). Για την κάθε ομάδα συγκεντρώνονται και οι 1000 συνεισφορές από όλους τους διαμερισμούς (δηλαδή, 58 σύνολα των 1000 αριθμών το καθένα). Έπειτα, εξάγεται ο μέσος όρος των 1000 επιμέρους συνεισφορών που λαμβάνει κάθε ομάδα σε όλους τους διαμερισμούς, αλλά και η διακύμανση που έχει το κάθε σύνολο των επιμέρους συνεισφορών της κάθε ομάδας. Δεχόμενοι πως το κάθε σύνολο συνεισφορών για την κάθε ομάδα αποτελεί κανονική κατανομή, υπολογίζεται το τριπλάσιο της διακύμανσης του εκάστοτε συνόλου του, με σκοπό να εξάγουμε την κατά 95% αβεβαιότητα της κάθε μιας παραμέτρου. Στο Παράρτημα Ε παρουσιάζονται οι συνεισφορές και η αβεβαιότητα για τους τρεις δείκτες.

Ένα σημαντικό βήμα για τον έλεγχο της αξιοπιστίας του μοντέλου είναι να καθιερωθεί ένας τρόπος να διαχωρίζονται τα αποτελέσματα του μοντέλου σε χαμηλό σφάλμα, μέσο και υψηλό. Ο λόγος που αυτό κρίνεται απαραίτητο είναι ώστε με την κατηγοριοποίηση των αποτελεσμάτων στις παραπάνω κατηγορίες να

αξιολογούνται οι προβλεπτικές δυνατότητες του μοντέλου κατά περίπτωση. Κρίνεται σκόπιμο να χωριστούν, λοιπόν, τρία διαστήματα χαμηλού, μέσου και υψηλού σφάλματος, όπου τα όρια της κάθε κατηγορίας θα είναι μέχρι 30%, μεταξύ 30% και 60% και πάνω από 60% μέσο απόλυτο σφάλμα στο σύνολο δοκιμής αντίστοιχα. Όσα αποτελέσματα, λοιπόν, έχουν στο σύνολο δοκιμής συνολικά κάτω από 30% μέσο απόλυτο σφάλμα θεωρούνται ενώσεις χαμηλού σφάλματος, αυτά που έχουν από 30%-60% μέσο απόλυτο σφάλμα και αυτές που έχουν πάνω από 60% μέσο απόλυτο σφάλμα θα έχουν υψηλό σφάλμα. Η θέσπιση των ορίων αυτή είναι καθαρά βάσει σύμβασης (εννοείται πως θα μπορούσαν να καθιερωθούν και διαφορετικά όρια). Δεδομένης, όμως, της φύσης των μοντέλων GC, τα οποία μπορεί να έχουν περίπου 20%-30% σφάλμα, τότε το όριο για «καλό» σφάλμα ίσο με 30% είναι αρκετά λογικό. Ακόμα, δεδομένου πως το σφάλμα του μοντέλου «μέσου όρου» είναι ίσο με 50% και δεδομένου πως πρέπει να είμαστε αυστηρότεροι ως προς το σφάλμα, για να επιτευχθεί μια καλύτερη εκτίμηση, τότε η θέσπιση του 60% ως «κακό» σφάλμα είναι εξίσου καλή.

Εν συνεχεία, για κάθε μια από τις 214 ενώσεις που συμμετείχαν στο αρχικό σύνολο δεδομένων συγκεντρώνεται το μέσο απόλυτο σφάλμα της, από όλες τις φορές που αυτή η ένωση συμμετείχε στο σύνολο δοκιμής. Με άλλα λόγια, έχει προαναφερθεί πως ο λόγος που είναι επιθυμητός ένας τόσο μεγάλος αριθμός διαμερίσεων, είναι ώστε στη διαμόρφωση του μοντέλου να συμμετάσχουν όσο το δυνατόν περισσότεροι συνδυασμοί μορίων στο σύνολο εκπαίδευσης και σύνολο δοκιμής. Έτσι, όλα τα μόρια είναι επιθυμητό κάποιες φορές να βρίσκονται στο ένα σύνολο και κάποιες στο άλλο. Για όσες φορές, λοιπόν, ένα μόριο συμμετείχε στο σύνολο δοκιμής, υπολογίστηκε και το σχετικό απόλυτο σφάλμα του από την πειραματική του τιμή. Από όλα αυτά τα απόλυτα σφάλματα, εξάγεται ο μέσος όρος τους. Αυτό είναι χρήσιμο για να εντοπιστεί η απόδοση του μοντέλου σε συγκεκριμένες χαρακτηριστικές ομάδες (αλκοόλες, οξέα, κετόνες κ.τ.λ.). Προκύπτουν, λοιπόν, 214 μέσα απόλυτα σφάλματα για το μοντέλο κάθε δείκτη (GWP, CED,EI 99). Να σημειωθεί πως δεν υπολογίζεται νέα εκτίμηση για την κάθε ένωση με τις νέες μέσες συνεισφορές για να εξαχθεί το σφάλμα του μοντέλου και να φανεί η απόδοση του, αλλά λαμβάνεται ο μέσος όρος των σφαλμάτων, άρα συμψηφίζονται οι μέσες επιμέρους αποδόσεις του μοντέλου για την κάθε ένωση.

Έπειτα, χωρίζονται οι 214 ενώσεις σε χαρακτηριστικές ομάδες. Προκύπτουν οι παρακάτω κατηγορίες: αλκοόλες, καρβοξυλικά οξέα, κετόνες, αλδεΰδες, αμίνες, αιθέρες, υδρογονάνθρακες, αλογονομένοι υδρογονάνθρακες, αρωματικά, εστέρες, μια κατηγορία που έχει ονομαστεί κυανίδια (cyanides ή και νιτρίλια) και περιλαμβάνει ενώσεις με τις ομάδες $–CH_3CN$ και $>CH_2CN$ και τέλος, η κατηγορία μικτών (mixed), δηλαδή μια κατηγορία που περιλαμβάνει μόρια με παραπάνω από μια χαρακτηριστικές ομάδες. Αυτά τα μόρια δεν είναι δυνατόν να τοποθετηθούν σε μια από τις παραπάνω κατηγορίες ανάλογα με μια από τις χαρακτηριστικές ομάδες τους, διότι θα αγνοούνταν η επίδραση των υπολοίπων στην ανάλυση των σφαλμάτων. Στο διάγραμμα 5.54 φαίνεται η ποσοστιαία κατανομή των ενώσεων του αρχικού συνόλου σε κατηγορίες:



Διάγραμμα 5.54. Κατανομή των χαρακτηριστικών ομάδων από το αρχικό σύνολο

Είναι αρκετά ενδιαφέρον να παρατηρηθεί η κατανομή των ενώσεων του αρχικού συνόλου δεδομένων στις παραπάνω κατηγορίες. Αν μια κατηγορία έχει κατά πολύ μεγαλύτερη συμμετοχή στο αρχικό σύνολο είναι λογικό πως θα επηρεάζει τη διαμόρφωση και των τελικών αποτελεσμάτων πολύ περισσότερο (π.χ. τα σφάλματα). Στη συγκεκριμένη περίπτωση, τα αρωματικά θα έχουν τη μεγαλύτερη συμμετοχή σα μια αυτόνομη ομάδα και έπειτα, οι αλκοόλες. Αμέσως, μετά οι αμίνες και μετά οι υδρογονάνθρακες. Αντίθετα, τα κυανιούχα (νιτρίλια), θα έχουν τη μικρότερη επίδραση στη διαμόρφωση αποτελεσμάτων. Να σημειωθεί πως η κατηγορία των μικτών, παρότι συνιστά το μεγαλύτερο ποσοστό της κατανομής δε συνιστά μια αυτόνομη χαρακτηριστική ομάδα και αυτό γιατί αποτελείται από οποιοδήποτε συνδυασμό δύο ή περισσοτέρων χαρακτηριστικών ομάδων που μπορεί να συνυπάρχουν μέσα σε ένα μόριο και όχι μια συγκεκριμένη χαρακτηριστική ομάδα μορίων. Επίσης, η τυχαιότητα που υπάρχει ως προς το ποιες χαρακτηριστικές ομάδες συμμετέχουν μέσα σε αυτή, το καθιστά αδύνατο να εντάξουμε αυτήν την κατηγορία μαζί με τις υπόλοιπες και να λάβουμε υπόψη τη συνεισφορά της στη διαμόρφωση των τελικών αποτελεσμάτων.

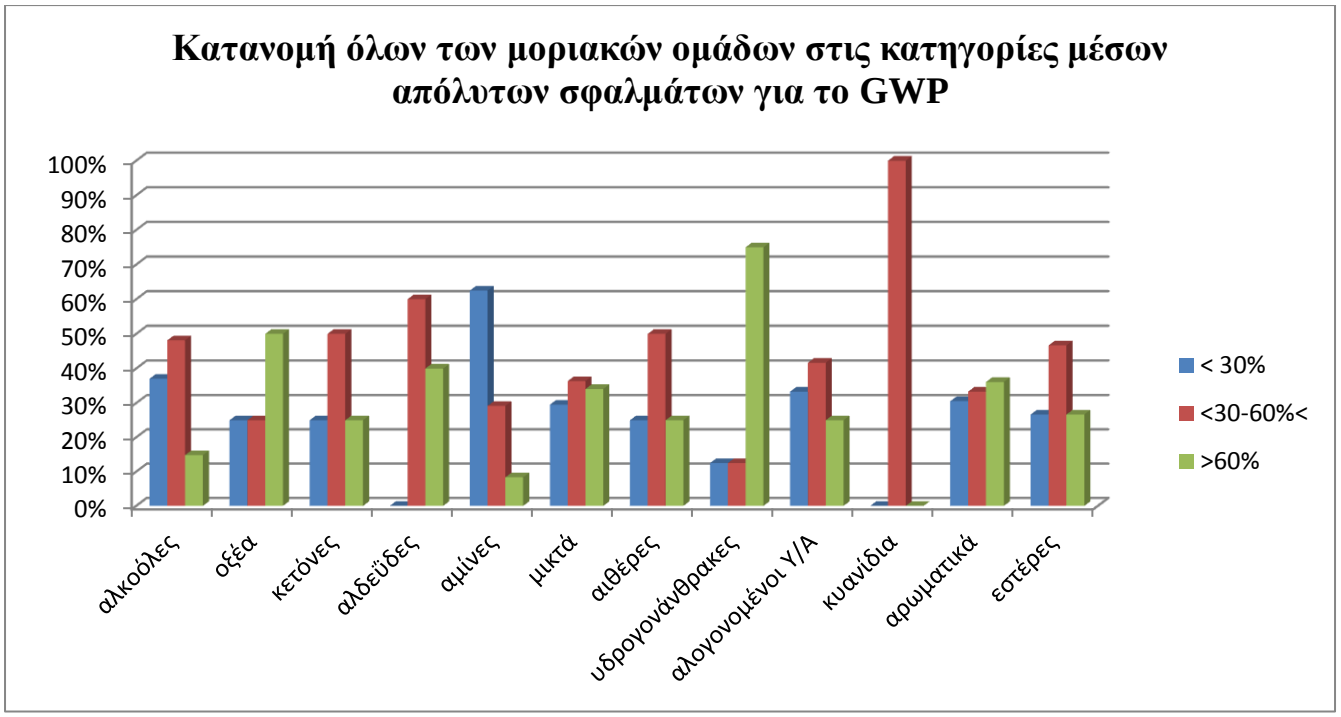
Για την κάθε χαρακτηριστική ομάδα, εξάγεται από τις ενώσεις που την απαρτίζουν, ο μέσος όρος των μέσων σφαλμάτων (της κάθε ένωσης). Με αυτόν τον τρόπο προκύπτει το μέσο απόλυτο σφάλμα του μοντέλου για την κάθε κατηγορία. Είναι μεγίστης σημασίας να εξαχθεί, όχι μόνο ένα σφάλμα που να χαρακτηρίζει γενικά το μοντέλο, αλλά και για κάθε κατηγορία. Κατ' αυτόν τον τρόπο, όταν στο μοντέλο δίνεται μια ένωση από μια συγκεκριμένη κατηγορία, θα αναμένεται και συγκεκριμένο σφάλμα από αυτή την εκτίμηση. Στο διάγραμμα 5.55, δίδεται το μέσο σφάλμα που εμφάνισε η κάθε κατηγορία για τους τρεις δείκτες:

Διάγραμμα 5.55. Μέσα σφάλματα για την κάθε κατηγορία στους τρεις δείκτες

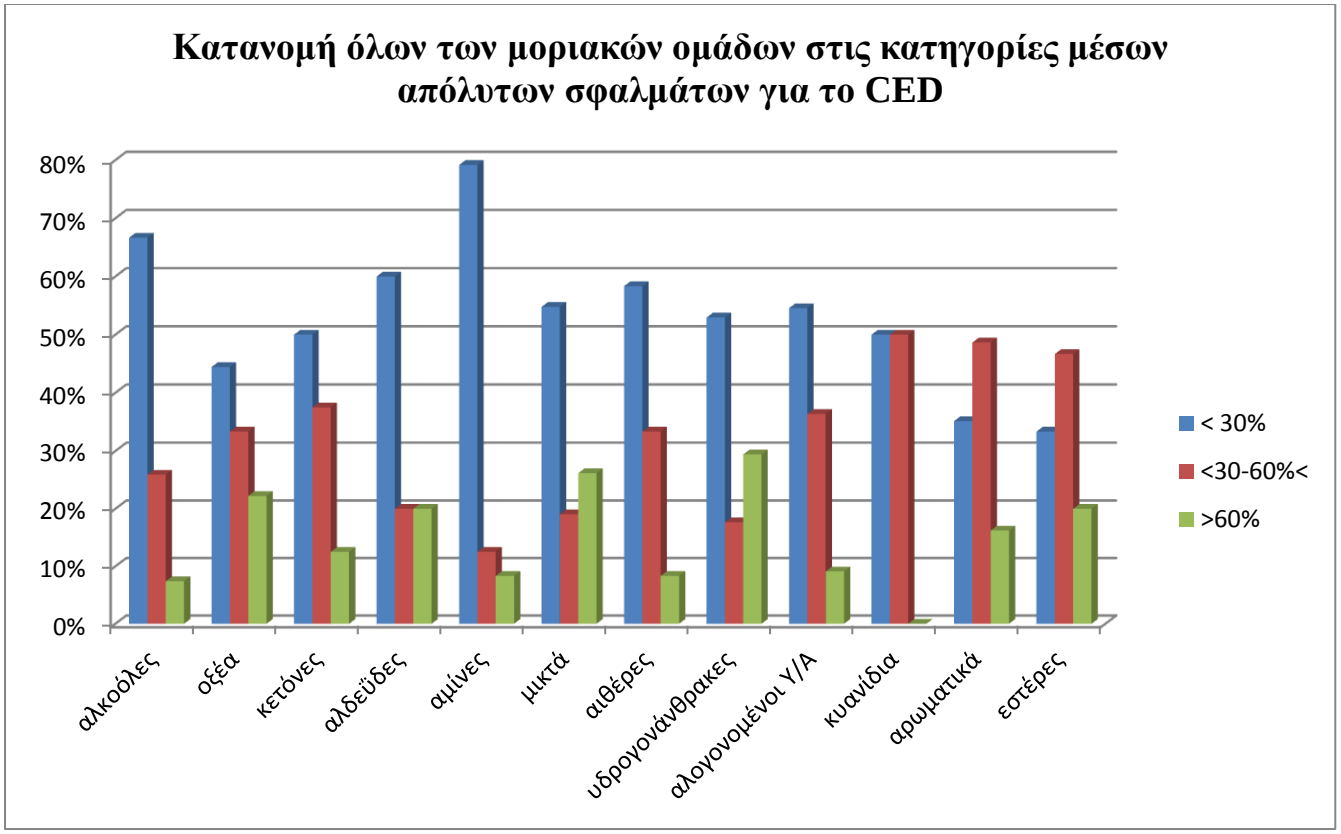
Για το μοντέλο του GWP, που είναι και το χειρότερο, φαίνεται πως οι καλύτερες κατηγορίες που μπορεί να προβλέψει είναι οι αμίνες και οι κετόνες, ενώ η χειρότερη κατηγορία είναι οι απλοί υδρογονάνθρακες. Για το μοντέλο του CED, φαίνεται πως οι καλύτερες κατηγορίες είναι οι αμίνες, οι αλογονομένοι υδρογονάνθρακες και κετόνες , ενώ η χειρότερη κατηγορία παραμένει η κατηγορία των υδρογονανθράκων. Τέλος, για το μοντέλο του EI 99, οι αλδεΰδες και οι εστέρες, ενώ η χειρότερη κατηγορία είναι αυτή των μικτών. Να σημειωθεί πως, ενώ η κατηγορία των κυανιούχων (νιτριλίων) περιλαμβάνει δύο μόνο ενώσεις. Παρόλο που σημειώνονται αρκετά μικρά σφάλματα για όλες τις κατηγορίες δε θεωρείται αρκετά αξιόπιστο να εξαχθούν συμπεράσματα γενικά για αυτήν την κατηγορία, μόνο επικεντρώνοντας την προσοχή μας σε αυτά τα μόρια.

Παραπάνω αναφέρθηκαν τρεις κατηγορίες σφαλμάτων: η κατηγορία χαμηλού σφάλματος (κάτω από 30%), η κατηγορία μέσου σφάλματος (30%-60%) και η κατηγορία υψηλού σφάλματος (πάνω από 60%). Είναι επίσης σημαντικό βάσει των κατηγοριών σφαλμάτων να παρατηρηθεί η κατανομή των μορίων της κάθε χαρακτηριστικής ομάδας στις παραπάνω κατηγορίες, να σημειωθεί δηλαδή, ποιο ποσοστό του συνόλου των μορίων κάθε ομάδας ανήκει στην κατηγορία των χαμηλών σφαλμάτων, ποιο ποσοστό έχει μέσα και ποιο ποσοστό έχει υψηλά σφάλματα. Με αυτό τον τρόπο, θα παρατηρηθεί και η συνεισφορά του κάθε υποσυνόλου μορίων στη διαμόρφωση της τελικής μέσης τιμής του σφάλματος της αντίστοιχης κατηγορίας. Πρέπει να ληφθεί, όμως, υπόψη πως η μέση τιμή κάθε κατηγορίας μπορεί να επηρεάζεται από πολύ λίγες, πολύ υψηλές (ή χαμηλές) τιμές προβλέψεων κάποιων μορίων, οι οποίες δεν είναι αντιπροσωπευτικές του συνόλου της κάθε κατηγορίας. Έτσι, το ίδιο το μέσο σφάλμα επηρεάζεται από αυτά τα μόρια και η γενικότερη τάση των μορίων είναι διαφορετική. Συνεπώς, αν η πλειοψηφία των ενώσεων

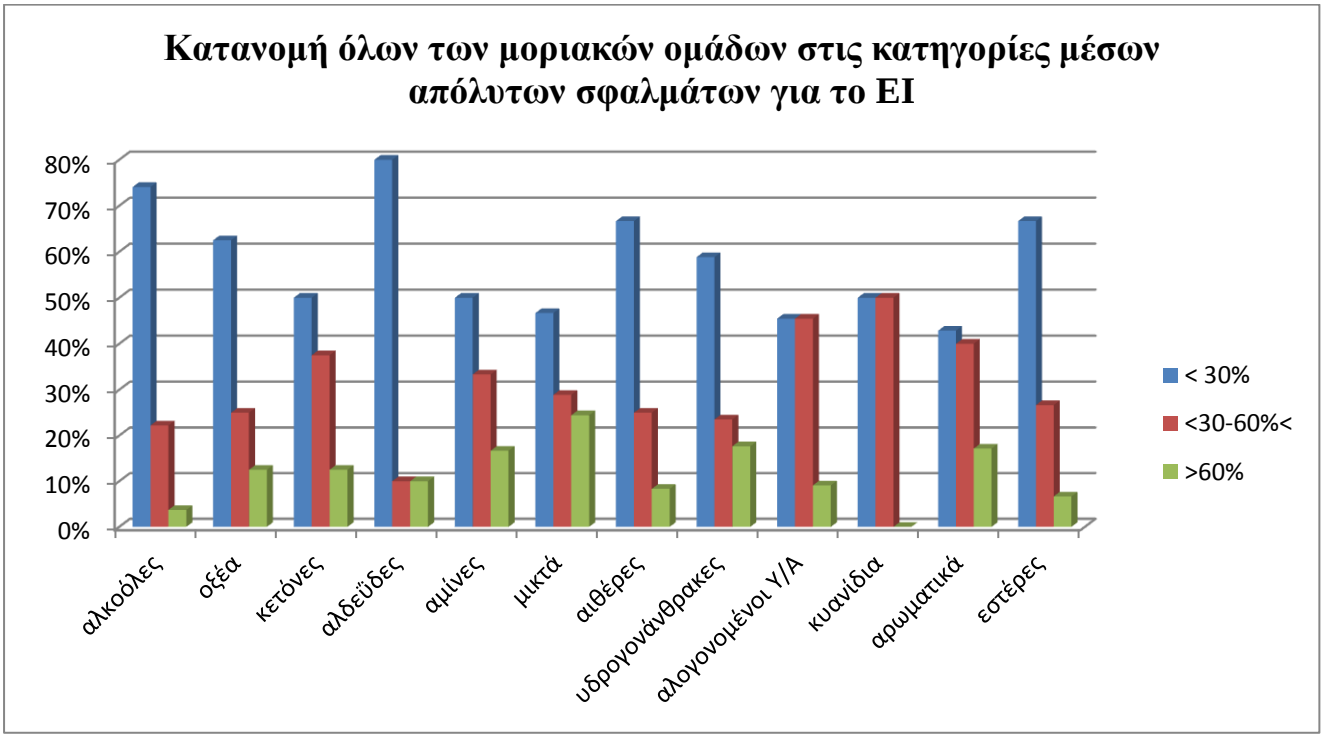
συγκεντρώνεται σε σφάλμα κάτω από 30% για ένα δείκτη και μια συγκεκριμένη ομάδα, ενώ το μέσο σφάλμα της ομάδας συνολικά είναι 60%, τότε η μέση τιμή έχει σαφώς επηρεαστεί ισχυρά από ακραίες τιμές. Εκφράζεται, λοιπόν, κατά μια κάπως γενικότερη έννοια, η πιθανότητα για μια καινούργια, ξένη ένωση της οποίας την τιμή του δείκτη δε γνωρίζουμε και πρέπει να προβλεφθεί, η πιθανότητα που μπορεί το σφάλμα της πρόβλεψης να συμπίπτει (ή να προσεγγίζει) με την τιμή του μέσου σφάλματος της χαρακτηριστικής ομάδας ,όπου ανήκει. Ασφαλώς, το παραπάνω μέτρο μπορεί να επιστήσει μόνο την προσοχή και όχι να εξάγει άμεσα στατιστικά αποτελέσματα, παρά μόνο να επισημάνει τρανταχτά παραδείγματα της παραπάνω υπόθεσης. Να σημειωθεί πως ο λόγος αυτών των ακραίων τιμών είναι και η αιτία, που αργότερα χρησιμοποιούνται οι στατιστικοί δείκτες του εκατοστημορίου (percentile) και μέσου (median). Στα διαγράμματα 5.56-5.58 φαίνεται η κατανομή των στοιχείων της κάθε ομάδας στις διάφορες κατηγορίες σφαλμάτων:



Διάγραμμα 5.56. Κατανομή των μορίων κάθε ομάδας σε τρεις διαφορετικές κατηγορίες σφαλμάτων για το GWP



Διάγραμμα 5.57. Κατανομή των μορίων κάθε ομάδας σε τρεις διαφορετικές κατηγορίες σφαλμάτων για το CED



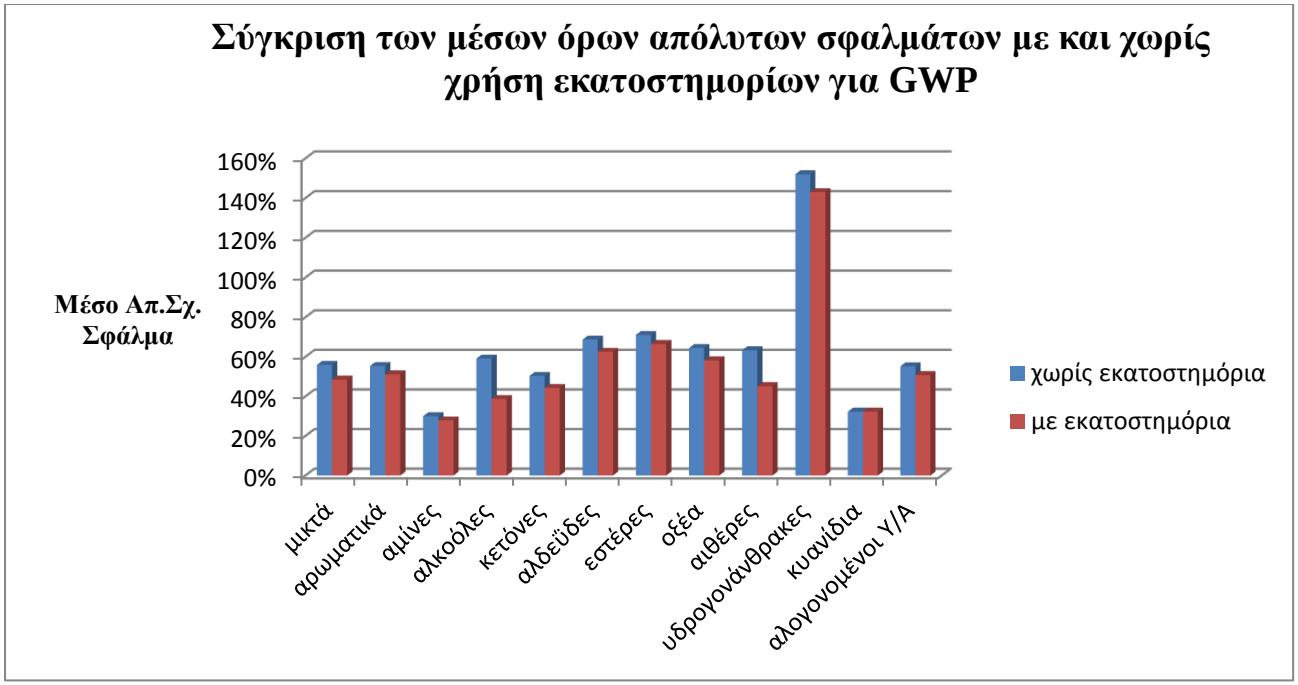
Διάγραμμα 5.58. Κατανομή των μορίων κάθε ομάδας σε τρεις διαφορετικές κατηγορίες σφαλμάτων για το EI 99

Η παραπάνω περίπτωση των ακραίων τιμών επιβεβαιώνεται και από τα διαγράμματα 5.56-5.58. Συγκεκριμένα για την περίπτωση των αλκοολών, παρατηρείται πως στις προβλέψεις για το GWP, ενώ η

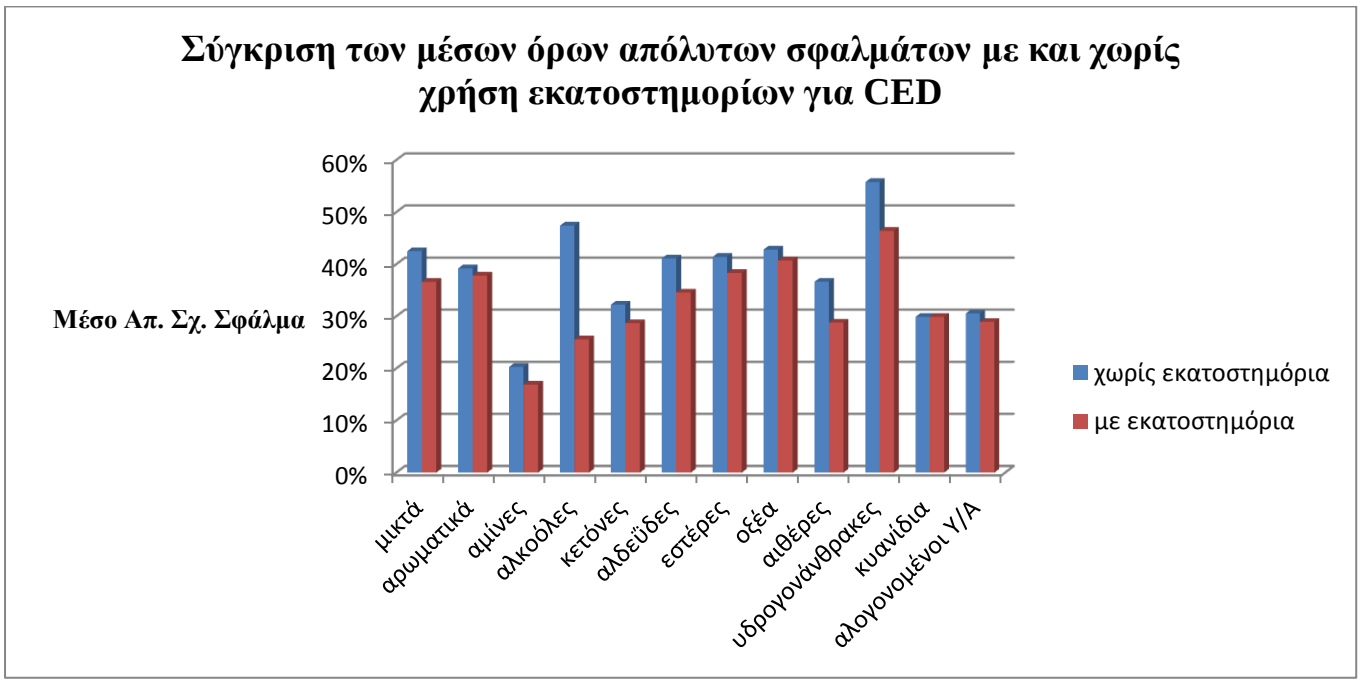
πλειοψηφία των μορίων εντοπίζεται στις κατηγορίες χαμηλού και μέσου σφάλματος και ενώ, μόλις το 13% των ενώσεων παρατηρείται να υπερβαίνουν το κατώφλι του 60% σφάλματος, το συνολικό σφάλμα βρίσκεται στο 58% (αρκετά κοντά στην περιοχή υψηλού σφάλματος). Αυτό μας ενημερώνει για την πιθανή ύπαρξη ακραίων τιμών. Το ίδιο συμβαίνει, ξανά για τις αλκοόλες, στην περίπτωση του EI 99, σε ένα ακόμα πιο προφανές παράδειγμα. Ενώ το 74% των αλκοολών βρίσκεται στην περιοχή σφάλματος κάτω από 30%, και υποθέτοντας πως υπάρχει μια κανονική κατανομή στο σύνολο των δεδομένων, αναμένεται πως το σφάλμα θα εντοπίζεται στη περιοχή του 30% και όχι του 47%.

Ακολουθεί ένας συστηματικός τρόπος για την εύρεση των ακραίων τιμών που υπάρχουν για κάθε κατηγορία. Συγκεκριμένα, για κάθε χαρακτηριστική ομάδα ευρίσκεται ο μέσος της (median), καθώς και το 5ο και 95ο εκατοστημόριο της. Δηλαδή, ευρίσκονται οι τιμές κάτω από τις οποίες υπάγεται το 5% και το 95% των τιμών του σφάλματος της κάθε ομάδας ενώσεων. Μπορεί να βρεθεί με αυτό τον τρόπο το 5% των ενώσεων με το μικρότερο σφάλμα όλης της ομάδας και το 5% των ενώσεων με το μεγαλύτερο σφάλμα. Εξαιρώντας αυτά τα μόρια από τον υπολογισμό του μέσου σφάλματος συντελεί στο να έχει αφαιρεθεί ο μεγαλύτερος αριθμός από τις ακραίες τιμές (αν όχι όλες). Το 5% είναι ένας αυθαίρετος αριθμός για την αφαίρεση των ακραίων τιμών και εξασφαλίζει μεν, την αφαίρεση των ακραίων τιμών, αλλά χωρίς να μειώνει το πλήθος του συνόλου τόσο, ώστε να μην είναι αξιόπιστη και αντιπροσωπευτική η μέση τιμή. Συγκεντρώνονται, λοιπόν, τα σφάλματα όλων των ενώσεων που βρίσκονται μεταξύ του 5% των χαμηλότερων και 5% των υψηλότερων τιμών και εξάγεται από εκεί ο μέσος όρος σφαλμάτων χωρίς ακραίες τιμές για κάθε κατηγορία. Ο μέσος (median), τα δύο εκατοστημόρια και η νέα μέση τιμή σφάλματος δίνονται στο παράρτημα Η.

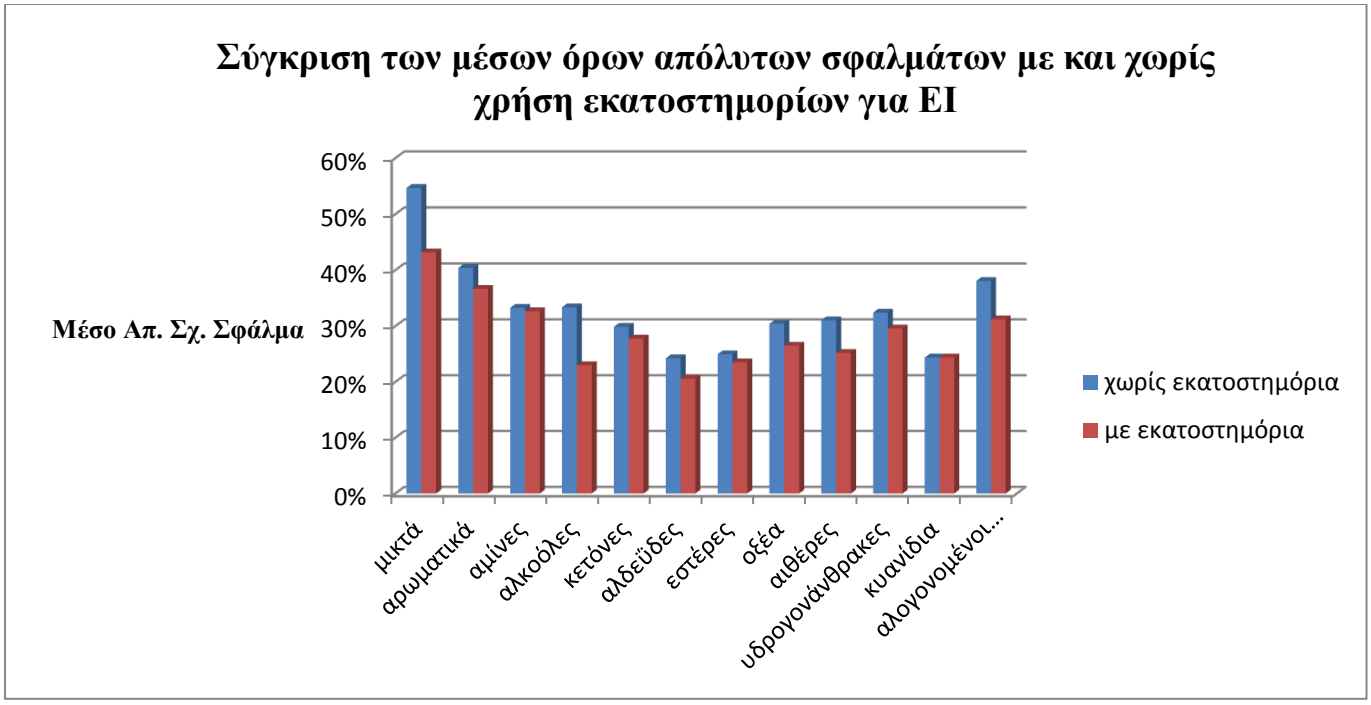
Στα διαγράμματα 5.59-5.61 φαίνεται η σύγκριση των μέσων σφαλμάτων κάθε κατηγορίας χωρίς να ληφθεί υπόψη το εκατοστημόριο με το μέσο όρο, αν αφαιρεθεί το 5% των μεγαλύτερων και το 5% των μικρότερων σφαλμάτων:



Διάγραμμα 5.59. Σύγκριση των μέσων όρων πριν και μετά την αφαίρεση των ενώσεων έξω από το εκατοστημόριο 5% και 95% για το μοντέλο GWP



Διάγραμμα 5.60. Σύγκριση των μέσων όρων πριν και μετά την αφαίρεση των ενώσεων έξω από το εκατοστημόριο 5% και 95% για το μοντέλο CED



Διάγραμμα 5.61. Σύγκριση των μέσων όρων πριν και μετά την αφαίρεση των ενώσεων έξω από το εκατοστημόριο 5% και 95% για το μοντέλο EI 99

Όπως φαίνεται, οι κατηγορίες των μορίων που επισημάνθηκαν νωρίτερα ως ύποπτες για ακραίες τιμές (αλκοόλες στο GWP και EI 99 και μικτά στο EI 99), επιβεβαιώνονται και εδώ να έχουν μεγάλες διαφορές στους μέσους όρους τους. Ακόμα, μεγάλες διαφορές παρατηρούνται και στις αλκοόλες στο δείκτη CED και στους αιθέρες στο δείκτη GWP.

Θεωρείται αρκετά χρήσιμο να επισημανθούν και τα αποτελέσματα των μοντέλων των τριών δεικτών έχοντας εξαιρεθεί οι ακραίες περιπτώσεις. Να αναφερθεί πως οι ακραίες περιπτώσεις είναι οι ενώσεις του συνόλου δοκιμής που ευρέθησαν να έχουν υψηλές τιμές στην απόσταση mahalanobis και συνεπώς, θεωρήθηκαν προεκβολές ως προς τα αντίστοιχα σύνολα εκπαίδευσης. Συγκεκριμένα, για κάθε διαμέριση συγκεντρώνονται οι εκτιμήσεις των μη ακραίων παρατηρήσεων από τα μοντέλα και τα πειραματικά τους δεδομένα και υπολογίζονται οι στατιστικοί δείκτες μεταξύ αυτών των δύο συνόλων (υπολογισμένων και πειραματικών). Έπειτα, υπολογίζονται οι μέσοι δείκτες, ακριβώς, όπως και στις άλλες περιπτώσεις. Αυτό συμβαίνει, διότι οι ακραίες παρατηρήσεις επηρεάζουν την αξιοπιστία του μοντέλου αρνητικά ως προεκβολές και επισημαίνονται ώστε να είναι αναμενόμενη μια υψηλότερη τιμή σφάλματος σε αυτές τις περιπτώσεις. Στους πίνακες 5.1-5.3 φαίνονται τα στατιστικά των μη ακραίων περιπτώσεων, αλλά γίνεται και μια σύγκριση με τα κανονικά στατιστικά του μοντέλου:

Πίνακας 5.1. Σύγκριση στατιστικών για το μοντέλο GWP με και χωρίς ακραίες περιπτώσεις

Στατιστικοί δείκτες	Σύν. Εκπαίδ. (χωρίς ακραία)	Σύν. Δοκιμής (χωρίς ακραία)	Σύν. Εκπαίδ. (με ακραία)	Σύν. Δοκιμής (με ακραία)
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,62	0,25	0,62	0,50
R^2 μεταξύ Πειραματικών και υπολογισμών	0,77	0,42	0,77	0,43
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,46	0,65	0,46	0,62
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	5,02	4,00	5,02	4,00
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,61	0,78	0,61	0,75
Μέσο Σχετικό Σφάλμα	0,18	0,29	0,18	0,24
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,74	0,97	0,74	0,94
Ρίζα του μέσου τετραγώνου σφάλματος	2,42	4,31	2,42	4,40
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,77	0,70	0,77	0,64
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,98	1,72	0,98	1,79
Συστηματικό RMSE	1,15	2,72	1,15	2,78
Μη συστηματικό RMSE	2,12	3,03	2,12	3,02
Συντελεστής Προσδιορισμού	0,77	0,24	0,77	0,26
Συντελεστής Προσδιορισμού-D1	0,68	0,49	0,68	0,51
Συντελεστής Προσδιορισμού-D2	0,93	0,68	0,93	0,69

Πίνακας 5.2. Σύγκριση στατιστικών για το μοντέλο CED με και χωρίς ακραίες περιπτώσεις

Στατιστικοί δείκτες	Σύν. Εκπαίδ. (χωρίς ακραία)	Σύν. Δοκιμής (χωρίς ακραία)	Σύν. Εκπαίδ. (με ακραία)	Σύν. Δοκιμής (με ακραία)
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	7,25	-0,51	7,25	-3,68
R^2 μεταξύ Πειραματικών και υπολογισμών	0,86	0,55	0,86	0,48
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,29	0,63	0,29	0,18
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	5,16	1,98	5,16	10,21
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,46	0,36	0,46	0,17
Μέσο Σχετικό Σφάλμα	0,07	-0,53	0,07	-0,04
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,54	0,51	0,54	0,28
Ρίζα του μέσου τετραγώνου σφάλματος	38,38	90,70	38,38	3481,53
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,86	0,86	0,86	0,74
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	14,88	-36,89	14,88	-548,78
Συστηματικό RMSE	13,93	65,50	13,93	912,58
Μη συστηματικό RMSE	35,72	53,83	35,72	1923,13
Συντελεστής Προσδιορισμού	0,86	0,18	0,86	0,15
Συντελεστής Προσδιορισμού-D1	0,73	0,46	0,73	0,34
Συντελεστής Προσδιορισμού-D2	0,96	0,70	0,96	0,68

Πίνακας 5.3. Σύγκριση στατιστικών για το μοντέλο EI 99 με και χωρίς ακραίες περιπτώσεις

Στατιστικοί δείκτες	Σύν. Εκπαίδ. (χωρίς ακραία)	Σύν. Δοκιμής (χωρίς ακραία)	Σύν. Εκπαίδ. (με ακραία)	Σύν. Δοκιμής (με ακραία)
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,03	0,02	0,03	0,03
R^2 μεταξύ Πειραματικών και υπολογισμών	0,68	0,34	0,68	0,35
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,27	6,59	0,27	0,38
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	3,15	24,81	3,15	2,36
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,36	4,20	0,36	0,44
Μέσο Σχετικό Σφάλμα	0,10	6,66	0,10	0,15
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,44	4,10	0,44	0,56
Ρίζα του μέσου τετραγώνου σφάλματος	0,11	1,78	0,11	0,17
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,68	0,51	0,68	0,50
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,11	1,96	0,11	0,18
Συστηματικό RMSE	0,06	1,78	0,06	0,10
Μη συστηματικό RMSE	0,09	0,13	0,09	0,13
Συντελεστής Προσδιορισμού	0,68	0,00	0,68	0,24
Συντελεστής Προσδιορισμού-D1	0,68	0,07	0,68	0,53
Συντελεστής Προσδιορισμού-D2	0,90	0,14	0,90	0,73

Παρατηρούνται αρκετές διαφορές μεταξύ των στατιστικών στις δύο περιπτώσεις. Είναι σημαντικό να αναφερθεί, μιας και η συσχέτιση είναι από τα σημαντικότερα μεγέθη που μας ενδιαφέρει, πως δε βελτιώνεται πάντα με την εξαίρεση των ακραίων περιπτώσεων.

Το τελευταίο βήμα για την αναλυτική διαδικασία του μοντέλου είναι η ανάλυση σφαλμάτων των ομάδων, δηλαδή να φανεί κατά πόσο μια ομάδα, όταν εμφανίζεται σε μια ένωση συνεισφέρει στο σφάλμα της εκτίμησης από την πειραματική τιμή. Αυτό συνίσταται στο να παρατηρηθεί αν σε κάποιες ενώσεις που έχουν χαμηλό σφάλμα (κάτω από 30%) επαναλαμβάνεται συστηματικά η εμφάνιση κάποιων ομάδων. Σε αυτές τις περιπτώσεις δε μπορεί παρά σε αυτές να οφείλεται το χαμηλό σφάλμα και να είναι «καλές» ομάδες. Κατ' αναλογία, σε ενώσεις που έχουν υψηλό σφάλμα (από 60% και πάνω) μπορούν να παρατηρηθούν αν έχουν κοινά σημεία. Αν υπάρχουν κάποιες ομάδες που επαναλαμβάνονται, επίσης, συστηματικά, τότε αυτές θα χαρακτηρίζονται ως ομάδες υψηλού σφάλματος. Στη συγκεκριμένη ανάλυση και επειδή, το ενδιαφέρον επικεντρώνεται στις ομάδες δε γίνεται διαχωρισμός ως προς τις χαρακτηριστικές ομάδες των μορίων. Τα βήματα που ακολουθούνται για να γίνει η παραπάνω ανάλυση είναι τα ακόλουθα: πρώτα κατηγοριοποιούνται όλες οι ενώσεις (ανεξαρτήτου ομάδας) στις τρεις κατηγορίες σφάλματος και έπειτα καθορίζεται ποιες ομάδες επαναλαμβάνονται συστηματικά στην κάθε κατηγορία για να προκύψουν ομάδες χαμηλού, μέσου και υψηλού σφάλματος.

Πρέπει να οριστεί πρώτα, όμως, η σημασία του όρου «συστηματικά». Με άλλα λόγια, πότε θεωρούμε πως μια ομάδα εμφανίζεται συστηματικά σε μια κατηγορία και πότε όχι. Για να εξαχθεί αυτό το συμπέρασμα, δεν καταγράφεται μόνο ο αριθμός των φορών που εμφανίζεται το κάθε ομάδα, αλλά αυτός ο αριθμός διαιρείται εν συνεχεία και με το συνολικό αριθμό εμφανίσεων της ομάδας αυτής συνολικά, σε όλες τις ενώσεις του αρχικού συνόλου. Κατά αυτόν τον τρόπο, λαμβάνεται για την κάθε μια η ποσοστιαία κατανομή του στις τρεις κατηγορίες σφάλματος. Αν, για παράδειγμα, μια ομάδα εμφανίζεται συνολικά σε όλες τις ενώσεις που χρησιμοποιούνται στο αρχικό σύνολο γενικά, 32 φορές, όπου τις 4 φορές σε ενώσεις χαμηλού σφάλματος, τις 20 σε ενώσεις μέσου σφάλματος και τις υπόλοιπες 10 σε ενώσεις υψηλού σφάλματος, τότε το 4/32=12.5% των εμφανίσεων έχει χαμηλό σφάλμα, το 20/32=62.5% μεσαίο σφάλμα και το 10/32=31.25% υψηλό σφάλμα. Γίνεται η παραδοχή πως αν μια ένωση εμφανίζεται σε μια κατηγορία με ποσοστό από 50% και πάνω των συνολικών φορών που έχει εμφανιστεί στο σύνολο, τότε καθιερώνεται και ως σφάλμα της αντίστοιχης κατηγορίας. Για το παραπάνω παράδειγμα, η ομάδα αυτή έχει εμφανιστεί στην κατηγορία μέσου σφάλματος σε αναλογία μεγαλύτερης του 50%. Άρα για πάνω από τις μισές φορές που εμφανίστηκε «προσέδωσε» στις ενώσεις που συμμετείχε μεσαίο σφάλμα (μεταξύ 30%-60%). Αποτελεί, λοιπόν, ομάδα μέσου σφάλματος. Να σημειωθεί πως οι ομάδες $-CH_3$, $>CH_2$ δε συμμετέχουν στην παραπάνω ανάλυση. Αυτό συμβαίνει διότι αποτελούν ομάδες που επαναλαμβάνονται σχεδόν σε κάθε μόριο, οποιασδήποτε κατηγορίας και είναι τα μόνα, τα οποία έχουν τη μεγαλύτερη συχνότητα (243 φορές το $-CH_3$ και 332 φορές το $>CH_2$). Αφού, λοιπόν, βρίσκονται στα περισσότερα μόρια δεν είναι δυνατόν να βρεθεί αν συνεισφέρουν σε χαμηλό ή υψηλό σφάλμα. Δίνονται οι πίνακες 5.4-5.6 που περιέχουν όλες τις ομάδες που

συμμετείχαν συστηματικά σε μόρια χαμηλού σφάλματος και για τους τρεις δείκτες (δίνεται μαζί και η συμμετοχή τους σε άλλες κατηγορίες σφάλματος):

Πίνακας 5.4. Ομάδες χαμηλού σφάλματος στο GWP και η κατανομή τους στις άλλες κατηγορίες

Ομάδες (GWP)	**30% >**	**30%< & < 60%**	**> 60%**
CH2NH	91,67%	0,00%	8,33%
CH2NH2	88,92%	11,00%	0,00%
C=C	50,00%	50,00%	0,00%
AC-NH2	62,50%	25,00%	12,50%
CH-Cl	50,00%	0,00%	50,00%
AC-CH=CH2	100,00%	0,00%	0,00%
AC-CH	100,00%	0,00%	0,00%
CF3	66,67%	33,00%	0,00%
CH-Cl2	100,00%	0,00%	0,00%
CCl4	100,00%	0,00%	0,00%
CH3NH	100,00%	0,00%	0,00%

Πίνακας 5.5. Ομάδες χαμηλού σφάλματος στο CED και η κατανομή τους στις άλλες κατηγορίες

Ομάδες (CED)	30% >	30%< & < 60%	> 60%
CH	67,57%	21,62%	10,81%
CH2=CH	50,00%	33,33%	16,67%
C=C	50,00%	50,00%	0,00%
OH	54,93%	25,35%	19,72%
CH3CO	66,67%	0,00%	33,33%
CH2CO	63,64%	36,36%	0,00%
HCO	57,14%	14,29%	28,57%
CH3COO	50,00%	41,67%	8,33%
CH2NH2	94,44%	5,56%	0,00%
CH3NH	100,00%	0,00%	0,00%
CH2NH	91,67%	8,33%	0,00%
CH3N	50,00%	25,00%	25,00%
CH2N	72,22%	5,56%	22,22%
CH2CN	100,00%	0,00%	0,00%
CHCl	50,00%	0,00%	50,00%
CHCl2	100,00%	0,00%	0,00%
CHCl3	100,00%	0,00%	0,00%
CCl3	100,00%	0,00%	0,00%
CF3	66,67%	33,33%	0,00%
CH2=O	100,00%	0,00%	0,00%
CH3Cl	100,00%	0,00%	0,00%
AC-CH	100,00%	0,00%	0,00%
AC-CH=CH2	100,00%	0,00%	0,00%
AC-COOH	50,00%	50,00%	0,00%
AC-CF3	50,00%	50,00%	0,00%
AC-NH2	50,00%	50,00%	0,00%
AC-Cl	50,00%	50,00%	0,00%

Πίνακας 5.6. Ομάδες χαμηλού σφάλματος στο EI 99 και η κατανομή τους στις άλλες κατηγορίες

Ομάδες (EI 99)	30% >	30%< & < 60%	> 60%
CH	75,68%	10,81%	13,51%
C	55,00%	25,00%	20,00%
CH2=CH	83,33%	11,11%	5,56%
CH=C	57,14%	28,57%	14,29%
C=C	50,00%	50,00%	0,00%
OH	57,75%	25,35%	16,90%
CH3CO	50,00%	16,67%	33,33%
CH2CO	63,64%	36,36%	0,00%
HCO	92,86%	7,14%	0,00%
CH3COO	66,67%	25,00%	8,33%
HCOO	50,00%	25,00%	25,00%
CH2O	60,71%	21,43%	17,86%
CH3NH2	100,00%	0,00%	0,00%
CH2NH2	66,67%	16,67%	16,67%
CHNH2	50,00%	50,00%	0,00%
CH2NH	58,33%	33,33%	8,33%
CH3N	50,00%	50,00%	0,00%
CH3CN	100,00%	0,00%	0,00%
CHCl	50,00%	50,00%	0,00%
CHCl2	100,00%	0,00%	0,00%
CCl3	100,00%	0,00%	0,00%
CH#C	100,00%	0,00%	0,00%
C#C	50,00%	50,00%	0,00%
CF3	66,67%	33,33%	0,00%
CH3Cl	100,00%	0,00%	0,00%
AC-CH3	52,38%	23,81%	23,81%
AC-CH	100,00%	0,00%	0,00%
AC-CH=CH2	100,00%	0,00%	0,00%
AC-CF3	100,00%	0,00%	0,00%
AC-NO2	75,00%	25,00%	0,00%
AC-NH2	50,00%	37,50%	12,50%
AC-Cl	50,00%	50,00%	0,00%

Δίνονται οι πίνακες 5.7-5.9 που περιέχουν όλες τις ομάδες, που συμμετείχαν συστηματικά σε μόρια χαμηλού σφάλματος και για τους τρεις δείκτες (δίνεται μαζί και η συμμετοχή τους σε άλλες κατηγορίες σφάλματος):

Πίνακας 5.7. Ομάδες μέσου σφάλματος στο GWP και η κατανομή τους στις άλλες κατηγορίες

Ομάδες (GWP)	< 30%	30%< & <60%	60% <
CH2=C	0,00%	100,00%	0,00%
C=C	50,00%	50,00%	0,00%
CH2CO	26,91%	64,00%	9,09%
HC=O	7,29%	57,00%	35,71%
CH3COO	41,67%	50,00%	8,33%
CH3O	0,00%	80,00%	20,00%
CH3-NH2	0,00%	100,00%	0,00%
CH-NH2	0,00%	100,00%	0,00%
CH3N	25,00%	50,00%	25,00%
CH3-CN	0,00%	100,00%	0,00%
CH2-CN	0,00%	100,00%	0,00%
CH2Cl2	0,00%	100,00%	0,00%
CHCl3	0,00%	100,00%	0,00%
C-Cl3	0,00%	100,00%	0,00%
CH#C	0,00%	100,00%	0,00%
C#C	0,00%	100,00%	0,00%
C(Cl)F2	0,00%	100,00%	0,00%
CH3Cl	0,00%	100,00%	0,00%
AC-CH2Cl	0,00%	100,00%	0,00%
AC-CH(Cl)2	0,00%	100,00%	0,00%
AC-C(F)3	0,00%	100,00%	0,00%
AC-NO2	25,00%	75,00%	0,00%
AC-Cl	0,00%	75,00%	25,00%

Πίνακας 5.8. Ομάδες μέσου σφάλματος στο CED και η κατανομή τους στις άλλες κατηγορίες

Ομάδες (CED)	< 30%	30%< & <60%	60% <
CH2=C	0,00%	100,00%	0,00%
CH=C	0,00%	57,14%	42,86%
C=C	50,00%	50,00%	0,00%
CH2COO	0,00%	50,00%	50,00%
HCOO	0,00%	50,00%	50,00%
CH3-NH2	0,00%	100,00%	0,00%
CH-NH2	0,00%	50,00%	50,00%
CH3-CN	0,00%	100,00%	0,00%
CH2-Cl	12,50%	50,00%	37,50%
CH2(CL)2	0,00%	100,00%	0,00%
C-(Cl)4	0,00%	100,00%	0,00%
CH#C	0,00%	100,00%	0,00%
C#C	0,00%	100,00%	0,00%
C(Cl)F2	0,00%	100,00%	0,00%
AC-CH3	28,57%	57,14%	14,29%
AC-COOH	50,00%	50,00%	0,00%
AC-CH2Cl	0,00%	100,00%	0,00%
AC-CH(Cl)2	0,00%	100,00%	0,00%
AC-C(F)3	50,00%	50,00%	0,00%
AC-NO2	25,00%	50,00%	25,00%
AC-NH2	50,00%	50,00%	0,00%
AC-Cl	50,00%	50,00%	0,00%

Πίνακας 5.9. Ομάδες μέσου σφάλματος στο EI 99 και η κατανομή τους στις άλλες κατηγορίες

Ομάδες (EI 99)	< 30%	30%< & <60%	60% <
CH=CH	28,57%	71,43%	0,00%
CH=C	0,00%	100,00%	0,00%
C=C	50,00%	50,00%	0,00%
CH-NH2	50,00%	50,00%	0,00%
CH3N	50,00%	50,00%	0,00%
CH2-CN	25,00%	50,00%	25,00%
CH2-Cl	25,00%	75,00%	0,00%
CH-Cl	50,00%	50,00%	0,00%
CH2(CL)2	0,00%	100,00%	0,00%
CH-Cl3	0,00%	100,00%	0,00%
C#C	50,00%	50,00%	0,00%
C(Cl)F2	0,00%	100,00%	0,00%
AC-CH2	16,67%	50,00%	33,33%
ACOH	12,50%	50,00%	37,50%
AC-CH2Cl	0,00%	100,00%	0,00%
AC-CH(Cl)2	0,00%	100,00%	0,00%
AC-Cl	50,00%	50,00%	0,00%

Τέλος, δίνονται οι πίνακες 5.10-5.12 που περιέχουν όλες τις ομάδες που συμμετείχαν συστηματικά σε μόρια υψηλού σφάλματος και για τους τρεις δείκτες (δίνεται μαζί και η συμμετοχή τους σε άλλες κατηγορίες σφάλματος):

Πίνακας 5.10. Ομάδες υψηλού σφάλματος στο GWP και η κατανομή τους στις άλλες κατηγορίες

Ομάδες (GWP)	>60%	60%< & 30%<	30%>
CH=CH	57,14%	28,57%	14,29%
CH=C	57,14%	42,85%	0,00%
CH2COO	60,00%	40,00%	0,00%
HCOO	75,00%	25,00%	0,00%
CHO	50,00%	40,00%	10,00%
COOH	78,26%	13,04%	8,70%
CH2Cl	62,50%	25,00%	12,50%
CHCl	50,00%	0,00%	50,00%
CH2=O	100,00%	0,00%	0,00%
C5H5N	100,00%	0,00%	0,00%
AC-CH2	83,33%	8,33%	8,33%
AC-C	100,00%	0,00%	0,00%
AC-OH	62,50%	18,75%	18,75%
AC-COOH	100,00%	0,00%	0,00%

Πίνακας 5.11. Ομάδες υψηλού σφάλματος στο CED και η κατανομή τους στις άλλες κατηγορίες

Ομάδες (CED)	>60%	60%< & 30%<	30%>
CH=CH	57,14%	28,57%	14,29%
CH2COO	50,00%	50,00%	0,00%
HCOO	50,00%	50,00%	0,00%
CHO	50,00%	30,00%	20,00%
CHNH2	50,00%	50,00%	0,00%
CHCl	50,00%	0,00%	50,00%
C5H5N	100,00%	0,00%	0,00%
AC-C	80,00%	20,00%	0,00%
AC-OH	50,00%	43,75%	6,25%

Πίνακας 5.12. Ομάδες υψηλού σφάλματος στο EI 99 και η κατανομή τους στις άλλες κατηγορίες

Ομάδες(EI 99)	>60%	60%< & 30%<	30%>
CH2COO	50,00%	10,00%	40,00%
CH3NH	100,00%	0,00%	0,00%
CCl4	100,00%	0,00%	0,00%
CH2=O	100,00%	0,00%	0,00%
C5H5N	100,00%	0,00%	0,00%
AC-C	80,00%	20,00%	0,00%
AC-COOH	100,00%	0,00%	0,00%

Για να επιβεβαιωθεί πως το παραπάνω μέτρο (το ποσοστό που εμφανίζεται ένα μόριο να είναι πάνω από 50% σε μια κατηγορία για να συνδεθεί το κάθε ομάδας με το αντίστοιχο σφάλμα) είναι στατιστικά αποδεκτό, για κάθε ένα από τους παραπάνω πίνακες εκτελείται Ανάλυση Διακύμανσης (ANOVA). Τα αποτελέσματα βρίσκονται στο παράρτημα Θ.

Παρατηρείται πως σε όλες τις περιπτώσεις η τιμή του δείκτη F για τις γραμμές είναι κατά πολύ μικρότερη από την κρίσιμη, ενώ για τις στήλες κατά πολύ μεγαλύτερη. Αυτό μας οδηγεί στο συμπέρασμα ότι η κατανομή στις γραμμές είναι όμοια για όλες τις γραμμές ενός συγκεκριμένου πίνακα. Αυτό είναι απόλυτα λογικό, μιας και για μια συγκεκριμένη ομάδα (που αντιπροσωπεύεται από μια γραμμή), θα κατανέμεται το μεγαλύτερο ποσοστό του σε μια συγκεκριμένη κατηγορία σφάλματος (διαφορετική κάθε φορά), που αντιπροσωπεύεται από μια στήλη. Τα υπόλοιπα ποσοστά σφαλμάτων κατανέμονται στις άλλες κατηγορίες (στήλες) κατά τυχαίο τρόπο (αυτός είναι και ο λόγος που οι στήλες δεν έχουν την ίδια κατανομή). Το παραπάνω συμπέρασμα με τη βοήθεια της ANOVA, δείχνει πως ο τρόπος με τον οποίο απεικονίστηκε η «συστηματική» εμφάνιση των ομάδων είναι ένα μέτρο που χαρακτηρίζει ικανοποιητικότατα την τάση που έχει το σύνολο των αποτελεσμάτων και αποτελεί έναν αξιόπιστο δείκτη.

Κεφάλαιο 6. Μοριακός Σχεδιασμός με Χρήση Ηλεκτρονικού Υπολογιστή

Ο Μοριακός Σχεδιασμός με χρήση Ηλεκτρονικού Υπολογιστή (Computer Aided Molecular Design,CAMD) είναι μια από τις πιο γνωστές προσεγγίσεις στο θέμα τις βελτιστοποίησης και γνώρισε ιδιαίτερη ανάπτυξη με την ανάπτυξη των υπολογιστικών εργαλείων και υπολογιστικών δυνατοτήτων. Με τη χρήση σύγχρονων εργαλείων μας επιτρέπεται η διαλογή (screening) μεταξύ λύσεων που υπόσχονται υψηλές αποδόσεις.

Η μεθοδολογία αυτή απευθύνεται σε κάθε διεργασία που επιτελείται με τη βοήθεια ενός χημικού μέσου. Τέτοιες διεργασίες είναι για παράδειγμα η εκχύλιση, η οποία χρησιμοποιεί διαλύτη, η εκχυλιστική απόσταξη, τα κύκλα παραγωγής ισχύος, τα οποία χρησιμοποιούν θερμαντικά μέσα, τα κύκλα ψύξης, που χρησιμοποιούν ψυκτικά και άλλα. Σε όλες τις παραπάνω περιπτώσεις, η διεργασία πραγματοποιείται με τη βοήθεια μιας χημικής ένωσης που είναι και η βασική παράμετρος βελτιστοποίησης. Αλλάζοντας, δηλαδή, το μόριο που συμμετέχει μεταβάλλονται και οι ιδιότητές του με αποτέλεσμα να βελτιώνεται ή να επιδεινώνεται η απόδοση της διεργασίας. Ως πιο χαρακτηριστικό παράδειγμα να αναφερθεί το παράδειγμα της εκχύλισης: αν αλλάξει ο διαλύτης, μεταβάλλεται και η εκλεκτικότητά του και ο συντελεστής κατανομής των προς διαχωρισμό συνιστωσών του αρχικού μίγματος και συνεπώς, μπορεί να οδηγούμαστε σε καλύτερη κατανομή των συνιστωσών στο εκχύλισμα και στο υπόλειμμα. Δεν πρέπει να ξεχνάμε, πως υπάρχει η απαίτηση ο ίδιος αυτός διαλύτης, να μπορεί να διαχωριστεί εύκολα από το εκχύλισμα, έπειτα, για να είναι αποδοτική η διεργασία. Δε μπορεί, όμως, να αγνοηθούν και άλλες παράμετροι για την εύρεση του βέλτιστου. Πρέπει για παράδειγμα, όταν η μεθοδολογία βελτιστοποίησης επιλέγει το βέλτιστο διαλύτη, να μην επιλέγει μόρια που είναι πολύ τοξικά. Παρατηρείται, λοιπόν, πως όσο μελετάται το πρόβλημα της βελτιστοποίησης μπορεί να προκύψει ένα πλήθος παραμέτρων στις οποίες πρέπει να βρεθεί μια «συμβιβαστική» λύση. Φαίνεται ξεκάθαρα, πως οι θερμοδυναμικές ιδιότητες συνδέονται άμεσα με την οικονομική απόδοση της διεργασίας (εκλεκτικότητα, συντελεστής κατανομής, σχετική πτητικότητα κ.τ.λ.), ενώ οι περιβαλλοντικές ιδιότητες και τοξικότητα μπορούν να προστεθούν ως περιορισμοί στη βελτιστοποίηση.

6.1 Ανασκόπηση βιβλιογραφίας

Από τα πρώτα προβλήματα της επιλογής κατάλληλου χημικού μέσου τίθεται από τους Prausnitz και Kumar (1975) και από τον Tassios (1972) για την επιλογή διαλυτών σε διεργασίες διαχωρισμού και αζεοτροπικών αποστάξεων αντίστοιχα. Παρόλο που το πρόβλημα αυτό επιλύθηκε σε σχετικά ικανοποιητικό βαθμό με ευριστικές μεθόδους, η ανάπτυξη των υπολογιστικών μεθόδων επέτρεψε μια πιο

αυτοματοποιημένη και πιο αποδοτική αναζήτηση μέσων. Οι Gani και Brignole (1983) και οι Stephanopoulos και Townsend (1985) πρότειναν συστηματικές μεθοδολογίες επιλογής μορίων με επιθυμητές ιδιότητες. Προτείνεται και η χρήση του θερμοδυναμικού μοντέλου UNIFAC για την πρόβλεψη της συμπεριφοράς και απόδοσης του μέσου στο διαχωρισμό. Οι Kolbe et al. (1979) και Magnussen et al. (1983) για τη λύση του προβλήματος της αζεοτροπικής απόσταξης και εκχυλιστικής απόσταξης, αντίστοιχα, επέλεξαν τη διαλογή μεταξύ πληθώρας διαλυτών με κριτήριο τις προβλεπόμενες, επίσης, από το μοντέλο της UNIFAC ιδιότητες. Μέχρις αυτό το σημείο, δεν υπήρχαν διαθέσιμες αρκετές συστηματικές μεθοδολογίες, κάτι που βασίστηκε στην έλλειψη αξιόπιστων μοντέλων για την πρόβλεψη της συμπεριφοράς των μη ιδανικών μιγμάτων στις παραπάνω διεργασίες. Έπειτα, ο Gani και Brignole (1986) εισήγαγαν μεθόδους χαρακτηρισμού των ομάδων με σκοπό για να εξασφαλίσουν τη δημιουργία εφικτών μοριακών δομών για διαλύτες.

Αργότερα, οι Cockrem et al. (1989) εισήγαγαν μια μέθοδο για την επιλογή διαλυτών για την εκχύλιση ενώσεων από αραιά υδατικά διαλύματα. Στην επέκταση των δυνατοτήτων αυτών των εργαλείων συνεισφέρουν και οι Joback και Stephanopoulos (1989) εισάγοντας ένα τρόπο μείωσης του μεγέθους του προβλήματος, αν αρκεί μόνο η εύρεση των ιδιοτήτων των καθαρών ουσιών των μορίων. Οι ίδιοι επέκτειναν τη μεθοδολογία μοριακού σχεδιασμού ώστε να περιλάβει σχεδιασμό ψυκτικών και πολυμερών. Σχεδιασμό πολυμερών επιχειρούν και οι Vadyanathan και El-Halwagi (1994). Οι Gani et al. (1991) επεκτείνουν την αρχική μεθοδολογία που εισήχθη από τον Gani (1983) και τη διαχωρίζουν σε τέσσερα διακριτά στάδια κατά τα οποία γίνεται μια αρχική διαλογή χαρακτηριστικών ενώσεων που θα συμμετάσχουν και τέλος, η πρόβλεψη των ιδιοτήτων των υποψήφιων ενώσεων και η επιλογή των βέλτιστων.

Οι Porter et al. (1991) επέκτειναν τη μεθοδολογία CAMD για σχεδιασμό διαλυτών για διεργασίες απορρόφησης αερίων. Αυτό επιτυγχάνεται με την εισαγωγή μεθόδων υπολογισμού διαλυτότητας αέριων συνιστωσών στη μέθοδο. Την ίδια περίοδο, οι Naser και Fournier (1991) εκτελούν μοριακό σχεδιασμό με μια διαδικασία τριών βημάτων που χρησιμοποιεί συνεχή βελτιστοποίηση (continuous optimization), η οποία ξεκινάει από ένα αρχικό μόριο, το οποίο είναι βέλτιστο για μέχρι και επτά κριτήρια βελτιστοποίησης και από αυτό εξάγονται παρόμοιες δομές που συνιστούν πιθανές λύσεις. Οι Pretel et al (1994) πρότειναν μια διαφορετική μεθοδολογία την οποία ονόμασαν MOLDES (Molecular Design of Solvents) και χρησιμοποιούν πειραματικά δεδομένα για την επικύρωση του μοντέλου τους. Εν συνεχεία οι Venkatasubramanian et al. (1994) επιχειρούν να εκτελέσουν CAMD στοχαστικά και συγκεκριμένα με χρήση γενετικών αλγορίθμων.

Αργότερα ο Pistikopoulos και Stefanis (1998) προτείνει μια μη συστηματική διαδικασία, τριών βημάτων, όπου γίνεται εμπειρικά η επιλογή ενός συνόλου διαλυτών και η επιλογή με βάση κάποια λειτουργικά και περιβαλλοντικά κριτήρια. Αργότερα το 1999, οι Pistikopoulos et al. εισάγουν μεθοδολογία για σχεδιασμό διαλυτών για διεργασίες απορρόφησης χωρίς αντίδραση, με γνώμονα το κόστος και την

περιβαλλοντική επιβάρυνση. Τέλος, το 2008 οι Pistikopoulos et al. επεκτείνουν τη μεθοδολογία CAMD και σε σχεδιασμό διαλυτών για αντιδράσεις. Συγκεκριμένα, σχεδιάζονται διαλύτες που σκοπό έχουν να ελαχιστοποιήσουν τη δημιουργία παραπροϊόντων. Είναι σημαντικό να τονιστεί πως ενώ όλες οι μεθοδολογίες έχουν τα δικά τους πλεονεκτήματα, δεν είναι πάντα εφικτή η σύγκριση τους, μιας και πρέπει να συγκριθούν το τελικό αποτέλεσμα (δομή του μορίου), ο ορισμός των ομάδων, οι εφικτές δομές και το σφάλμα των υπολογιστικών μεθόδων.

Το 2001, οι Wang και Achenie θέτουν το ζήτημα της συμβατότητας υλικών και ασφάλειας στο σχεδιασμό χημικών μέσων και χρησιμοποιούν τη μεθοδολογία για εύρεση διαλυτών για εκχυλιστική ζύμωση. Οι Gani, Achenie και Venkatasubramanian (2002) αναφέρουν τις νέες προκλήσεις στο CAMD, όπου ανάμεσα σε άλλα αναφέρουν το σχεδιασμό νέων υλικών και προτείνουν ένα πλαίσιο εργασίας για την αναζήτηση του βέλτιστου μονοπατιού σύνθεσης χημικών ενώσεων μέσα από αντιδράσεις. Επίσης, οι Kim και Diwekar (2002) επιχειρούν να εξαλείψουν τις αδυναμίες που επιφυλάσσει η στοχαστική αναζήτηση, εκτελώντας CAMD με χρήση του αλγορίθμου στοχαστικής ανόπτησης Hammersley (Hammersley stochastic annealing,HSTA) Το 2003 οι Sinha, Achenie και Gani παρουσιάζουν την εφαρμογή των μεθοδολογιών CAMD σε διαλύτες καθαρισμού που χρησιμοποιούνται από την τυπογραφία. Οι Sahinidis, Tawarmalani και Yu (2003) επανεκτελούν το πρόβλημα εύρεσης βέλτιστων ψυκτικών, αυτή τη φορά επιλύοντας το πρόβλημα με μικτό, ακέραιο, μη γραμμικό προγραμματισμό (MINLP) και με αλγόριθμο εντοπίζει το καθολικά βέλτιστο. Ακόμα, οι Giovanoglou et al. (2003) εκτελούν το CAMD με χρήση MIDO (mixed integer dynamic optimization) για έυρεση διαλυτών σε διαλείποντος έργου διεργασίες. Η διαδικασία αυτή εκτελείται σε πολλά βήματα με σκοπό να απορριφθούν έγκαιρα άχρηστες λύσεις και να μειωθεί το υπολογιστικό κόστος.

Το 2004 προτείνεται από τους Karunanithi et al. η χρήση μιγμάτων διαλυτών για διεργασίες διαχωρισμού και αντιμετωπίζουν χωρίζουν το αρχικό πρόβλημα σχεδιασμού μορίων και μιγμάτων σε υποπροβλήματα. Οι ίδιοι συγγραφείς επανέρχονται το 2005 στο ίδιο πρόβλημα και το αντιμετωπίζουν ξανά, με χρήση MINLP (mixed integer non linear programming). Το πρόβλημα χωρίζεται σε δυο στάδια, όπου επιλύεται το πρώτο και ευρίσκεται το βέλτιστο μόριο και μετά, αφού αναγνωριστεί ένα σύνολο από βέλτιστα μόρια, προσδιορίζεται το μίγμα τους. Οι Gani, Jimenez-Gonzalez και Constable (2005) αντιμετωπίζουν το πρόβλημα σχεδιασμού διαλυτών για οργανικές αντιδράσεις. Συγκεκριμένα, χρησιμοποιούν συσχετισμούς μεταξύ θερμοδυναμικών ιδιοτήτων και ρυθμών αντίδρασης για να καθορίσουν μια λίστα από βέλτιστους διαλύτες.

Οι Karunanithi et al. (2006) επέκτειναν την αρχική μεθοδολογία μοριακού σχεδιασμού με σκοπό να σχεδιαστούν διαλύτες για διεργασίες κρυστάλλωσης. Οι Gani et al. (2006) επεκτείνουν το υπάρχον πλαίσιο εργασίας για να συμπεριλάβουν και διαλύτες που προορίζονται για αντιδράσεις και λαμβάνουν υπόψιν περιβαλλοντικούς περιορισμούς και περιορισμούς ασφάλειας και υγιεινής. Οι Song και Song (2006)

αναφέρουν μια μέθοδο CAMD,η οποία χρησιμοποιεί ευριστικούς τρόπους στο σχηματισμό μορίων που δοκιμάζονται ως βέλτιστες λύσεις και χρησιμοποιούν γενετικούς αλγορίθμους και προσομοιωμένη ανόπτηση για τη βελτιστοποίηση των δομών. Οι Kossack et al. (2008) αντιμετωπίζουν ξανά το πρόβλημα της εκχυλιστικής απόσταξης, αυτή τη φορά κάνοντας τη διαλογή των υποψήφιων διαλυτών με κριτήριο την εκλεκτικότητα σε άπειρη αραίωση και τη μέθοδο RBM (rectification body method). Εξετάζονται και λειτουργικές παράμετροι σε στήλες εκχυλιστικής απόσταξης. Να αναφερθεί πως το 2008, οι Modarresi et al. (2008) χωρίς να απευθύνονται σε προβλήματα βελτιστοποίησης, προτείνουν μια σειρά από μοντέλα διαλυτότητας στερεών για την αξιοποίηση τους σε επιλογή βέλτιστων διαλυτών. Οι Gani et al. (2008) επεκτείνουν την αρχική μεθοδολογία CAMD και πλέον απευθύνονται σε προβλήματα που περιλαμβάνουν αντιδράσεις πολλαπλών σταδίων (multi-step reaction) και αντικατάστασης διαλυτών σε αντιδράσεις, σε ήδη υπάρχοντα συστήματα. Να αναφερθεί, ακόμα, και η δουλειά των Karunanithi, Acquah και Achenie (2008), οι οποίοι χρησιμοποίησαν το CAMD για να επιτύχουν μέσα από βελτιστοποίηση του διαλύτη, συγκεκριμένη κρυσταλλική μορφολογία σε φαρμακευτικά προϊόντα. Αυτό επετεύχθη μέσα από τρία στάδια: στο πρώτο καθορίζονται οι ιδιότητες με στόχο την επίτευξη συγκεκριμένης μορφολογίας, στο δεύτερο, σχηματίζεται το πρόβλημα σχεδιασμού ως πρόβλημα μικτού ακέραιου μη γραμμικού προγραμματισμού (MINLP) και τέλος, επιβεβαιώνεται πειραματικά το αποτέλεσμα. Οι Satyanarayana, Abildskov, Gani (2009) επαναφέρουν το πρόβλημα του σχεδιασμού πολυμερών και προτείνουν τη χρήση GC^{+} μοντέλων πρόβλεψης ιδιοτήτων και το συνδυασμό τους με μια βελτιωμένη μέθοδο CAMD. Οι Gani et al. (2012) χρησιμοποιούν το υπάρχον πλαίσιο εργασία του Gani (1991) και επεκτείνουν τη μεθοδολογία για τις διεργασίες διαχωρισμού για να σχεδιάσουν και ιονικά ρευστά. Να σημειωθεί πως οι Gani et al. (2012) χρησιμοποιούν σε άλλη δημοσίευση, το τροποποιημένο θερμοδυναμικό της UNIFAC, UNIFAC-IL για την πρόβλεψη της συμπεριφοράς των ιονικών ρευστών σε αζεοτροπικές αποστάξεις.

Το 2013 προτείνεται από τους Karunanithi και Merkesh ένα νέο πλαίσιο εργασίας για το σχεδιασμό ιονικών ρευστών, το οποίο χρησιμοποιεί ημιεμπειρικά μοντέλα πρόβλεψης ιδιοτήτων. Οι Sahinidis και Samudra (2013) εισάγουν μια καινούργια μεθοδολογία CAMD, η οποία χωρίζεται σε τρία στάδια. Στο πρώτο αναγνωρίζονται οι βέλτιστες ομάδες που μπορούν να απαρτίζουν ένα μόριο, στο δεύτερο ευρίσκεται η βέλτιστη δομή, την οποία μπορεί να σχηματίζουν αυτές τις ομάδες και στο τρίτο στάδιο χρησιμοποιούνται καινοτόμα μοντέλα πρόβλεψης ιδιοτήτων. Οι ίδιοι ερευνητές επεκτείνουν την έρευνα τους στα περιβαλλοντικώς φιλικά ψυκτικά.

6.2 Σύντομη περιγραφή της μεθοδολογίας Marcoulaki και Kokossis

Οι Marcoulaki και Kokossis (2000) πρότειναν μια πρωτοποριακή μεθοδολογία σύνθεσης μορίων CAMD, η οποία χρησιμοποιεί Στοχαστικές Μεθόδους για βελτιστοποίηση και συγκεκριμένα, τη μέθοδο της Προσομοιωμένης Ανόπτησης (Simulated Annealing).

Το πρόβλημα της βελτιστοποίησης που αντιμετωπίζεται μπορεί να περιγραφεί ως εξής. Δίνεται, καταρχάς, ένα σύνολο από ομάδες ατόμων (δομικές μονάδες), τα οποία ενωμένα με διάφορους πιθανούς συνδυασμούς σχηματίζουν ένα σύνολο δομών. Δεύτερον, ένα σύνολο μοντέλων τα οποία δέχονται τις δομικές μονάδες ως είσοδο και παρέχουν, ως έξοδο, θερμοφυσικές και άλλες ιδιότητες. Τρίτον, στόχοι της σύνθεσης μορίων που περιγράφονται από μια αντικειμενική συνάρτηση (objective) και περιορισμών (constraints), τα οποία υπαγορεύονται από τη φύση της διεργασίας. Πρέπει, λοιπόν, να καθοριστεί μια βέλτιστη τιμή για την αντικειμενική συνάρτηση, καθώς και ένα σύνολο από κατάλληλες μοριακές δομές, οι οποίες ικανοποιούν ή έστω, να πλησιάζουν τους στόχους της αντικειμενικής συνάρτησης. Οι περιορισμοί και η αντικειμενική συνάρτηση αποτελούνται από θερμοδυναμικές ιδιότητες και τοξικότητα. Χαρακτηριστικά παραδείγματα που αντιμετωπίζονται με τη χρήση αυτού του εργαλείου είναι ο σχεδιασμός διαλυτών, ψυκτικών και πολυμερών.

Η δημιουργία μορίων και για την πρόβλεψη των ιδιοτήτων βασίζεται στην αναπαράσταση των μορίων ως σύνολο από ομάδες. Στις παραπάνω περιπτώσεις εμπίπτουν αρωματικά και μη αρωματικά μόρια, όπως και κάποιες κυκλικές δομές. Αυτός ο τρόπος αναπαράστασης είναι απολύτως συμβατός με τα μοντέλα πρόβλεψης. Είναι σημαντικό να αναφερθεί, πως το συγκεκριμένο εργαλείο περιλαμβάνει ένα σύνολο από περιορισμούς (feasibility rules) μεταξύ του σθένους των διάφορων ομάδων που εξασφαλίζει πως το σύνολο των ομάδων που προτείνεται από τη μεθοδολογία μπορεί, όντως, να σχηματίσει μια πραγματική μοριακή δομή. Η παραπάνω μεθοδολογία βασίζεται σε μια επαναληπτική διαδικασία, όπου «παράγονται» με συστηματικό τρόπο μόρια σε μορφή διανυσμάτων (όπου στο κάθε διάνυσμα περιλαμβάνονται οι ομάδες και ο αριθμός που αυτά εμφανίζονται). Εν συνεχεία, τιμή της αντικειμενικής συνάρτησης ευρίσκεται και γίνεται ο έλεγχος, αν πληρούνται τα κριτήρια των περιορισμών. Αφού εκτελεστούν τα παραπάνω βήματα για το πρώτο μόριο, τότε γίνονται οι απαραίτητες κινήσεις ή μετατοπίσεις (perturbations) για την τροποποίηση της μοριακής δομής και την «παραγωγή» μιας νέας μοριακής δομής από διαφορετικές ομάδες (εννοείται πως οι κινήσεις γίνονται με σεβασμό στα feasibility rules) και ακολούθως, ο παραπάνω έλεγχος. Η διαδικασία συνεχίζεται με στοχαστική αναζήτηση και επιλογή μορίων (απόρριψη ή αποδοχή των νέων δομών) βάσει των κριτηρίων της Προσομοιωμένης Ανόπτησης, μέχρι να σχηματιστεί ένα σύνολο από βέλτιστες λύσεις. Να σημειωθεί πως το βασικό κριτήριο για την αποδοχή ή απόρριψη των δημιουργημένων δομών και για το ποιο μόριο αποτελεί βέλτιστη λύση είναι η απόδοση που συνδέεται με κάθε μόριο. Η απόδοση υπολογίζεται συναρτήσει της τιμής της αντικειμενικής συνάρτησης, της πολυπλοκότητας και των ποινών που έχει το κάθε μόριο που αντικατοπτρίζουν τη δομή του και το είδος των ομάδων.

Το παραπάνω εργαλείο αποτέλεσε αντικείμενο συνεχούς μελέτης και επεκτάθηκε σε αρκετές εφαρμογές. Οι Marcoulaki και Kokossis (2000) επιδεικνύουν τις δυνατότητες του εργαλείου αυτού για σχεδιασμό διαλυτών για εκχυλίσεις και εκχυλιστικές αποστάξεις για διάφορα συστήματα. Το θερμοδυναμικό μοντέλο της UNIFAC παίζει μείζονα ρόλο, μιας και με αυτό γίνονται όλοι οι υπολογισμοί για την αποδοτικότητα ενός διαλύτη. Επιδεικνύουν, επίσης, πως η διαμόρφωση του προβλήματος (objective και constraints) επιδρούν στο τελικό αποτέλεσμα και στην απόδοση της διεργασίας. Τέλος, παρουσιάζουν πως με την υπάρχουσα μεθοδολογία μπορεί να επιτευχθεί ολοκλήρωση και της διεργασίας και του σχεδιασμού του διαλύτη. Οι Marcoulaki, Kokossis και Batzias (2000) επικεντρώνονται στο σχεδιασμό πράσινων διαλυτών προσαρμόζοντας τη μεθοδολογία να αναζητά μόρια με χαμηλή τοξικότητα. Οι Marcoulaki και Batzias (2003) προσαρμόζουν τη διαδικασία σύνθεσης βέλτιστου διαλύτη και τη συνδυάζουν με βελτιστοποίηση διεργασιών εκχυλιστικής ζύμωσης (extractive fermentation).

Οι Papadopoulos και Linke (2004) δοκιμάζουν να αντιμετωπίσουν, με την υπάρχουσα μεθοδολογία, το πρόβλημα σύνθεσης διαλυτών ως υπερδομή και να δοκιμάσουν άλλες μεθόδους βελτιστοποίησης, όπως οι γενετικοί αλγόριθμοι και τις αποικίες μυρμηγκιών (ant colonies) με σκοπό να παρατηρηθεί η αξιοπιστία και η ταχύτητα της μεθόδου. Αργότερα, την ίδια χρονιά οι ίδιοι ερευνητές αναβαθμίζουν την υπάρχουσα μεθοδολογία ώστε να εμπεριέχει πολλαπλές αντικειμενικές συναρτήσεις (multi-objective) και εισάγουν μια μέθοδο, όπου επιτρέπει να εκτελεστεί ταυτόχρονα η σύνθεση του βέλτιστου μορίου και η βελτιστοποίηση της διεργασίας. Στη μεθοδολογία τους χρησιμοποιούν το μέτωπο Pareto σε συνδυασμό με το πλήθος των αντικειμενικών, ώστε να καθοριστεί ο βέλτιστος διαλύτης. Για να υποβοηθήσουν την ολοκλήρωση της βελτιστοποίησης του διαλύτη με τη βελτιστοποίηση της διεργασίας, χωρίζουν τα μόρια που προκύπτουν σε ομάδες (clusters), όταν τα μόρια που προκύπτουν έχουν παραπλήσιες ιδιότητες. Έτσι, είναι ευκολότερο να επιλεγεί ένα αντιπροσωπευτικό μόριο από κάθε ομάδα (cluster) και να διευκολυνθεί η επιλογή διαλύτη. Η παραπάνω ενοποιημένη μεθοδολογία διαλύτη διεργασίας επεκτείνεται και στο σχεδιασμό διεργασιών που περιλαμβάνουν αντίδραση και διαχωρισμό (reactive-separation) (Papadopoulos, Linke, 2005). Ακόμα, προτείνουν παραπάνω από μια στρατηγικές για το διαχωρισμό των μορίων σε ομάδες (clusters) και μελετάται καλύτερα η σχέση μεταξύ διαλύτη και διεργασίας (Papadopoulos, Linke, 2006).

Οι Papadopoulos, Stijepovic και Linke (2010) χρησιμοποιούν αυτή τη μεθοδολογία για να την εφαρμόσουν στην εύρεση βέλτιστων μέσων για Οργανικά Κύκλα Rankine (ORC). Κάνουν, όμως, και αναφορά όχι μόνο στη χρήση οικονομικών κριτηρίων για την επιλογή, αλλά και σε περιορισμούς ασφάλειας και περιβάλλοντος. Αργότερα, αυτή η μεθοδολογία (Papadopoulos et al. 2013) αναλύεται σε δυο βήματα, όπου αναζητούνται μόρια τα οποία σε μια βέλτιστη αναλογία σχηματίζουν ένα μίγμα από ένα βέλτιστο ρευστό για χρήση σε ORC.

Στις επόμενες υποενότητες ακολουθεί μια εκτενής περιγραφή της μεθοδολογίας CAMD και της μεθόδου Προσομοιωμένης Ανόπτησης.

6.3 Η έννοια της βελτιστοποίησης

Ο όρος βελτιστοποίηση είναι μια έννοια που συναντάται σε κάθε δραστηριότητα, ενέργεια και λειτουργία με σκοπό την επίτευξη του καλύτερου δυνατού αποτελέσματος. Η βελτιστοποίηση βασίζεται στην επιλογή τιμών για διάφορες μεταβλητές, τέτοιων ώστε να μεγιστοποιείται ή να γίνεται ελάχιστη μια επιθυμητή ποσότητα που συνδέεται με τις μεταβλητές αυτές, λαμβάνοντας παράλληλα υπόψη τυχόν περιορισμούς στην επιλογή τιμών των μεταβλητών.

Παραδείγματα εφαρμογής βελτιστοποίησης:

- **Βιομηχανική διεργασία**
  - Σχεδιασμός της διεργασίας με σκοπούς όπως: η αύξηση της απόδοσης, η μείωση του κόστους, η αύξηση της αξιοπιστίας της.

- **Διαδικασία μεταφοράς προϊόντων**
  - Επιζήτηση της μικρότερης δυνατής κατανάλωσης καυσίμου για δεδομένο χρόνο ολοκλήρωσης της μεταφοράς, πλήθος και τύπο οχημάτων μεταφοράς.
  - Ελαχιστοποίηση του χρόνου μεταφοράς για δεδομένο πλήθος και τύπο οχημάτων μεταφοράς.

- **Χημικές ουσίες**
  - Επίτευξη βέλτιστων ιδιοτήτων σε μια χημική ουσία ενώ παράλληλα θα πληροί και κάποιους περιβαλλοντικούς ή/και θερμοδυναμικούς κανόνες.

6.3.1 Μαθηματική βελτιστοποίηση

Στην προσπάθεια επίλυσης πραγματικών προβλημάτων (όπως αυτά που προαναφέρθηκαν στην ενότητα 6.3) αναπτύσσονται μαθηματικά μοντέλα τα οποία αναπαριστούν τα πρακτικά αυτά προβλήματα που θέλουμε να επιλύσουμε. Εκφράζοντας το σύστημα μας με μαθηματικό τρόπο, εστιάζουμε στην βελτίωση της αντικειμενικής συνάρτησης f υπό ένα σύνολο παραμέτρων (δεδομένων), μεταβλητών (ζητούμενων), περιορισμών και φραγμάτων. Οι περιορισμοί του προβλήματος είναι αυτοί που ορίζουν το εφικτό πεδίο λύσεων και η παραβίαση έστω και ενός από τους περιορισμούς καθιστά τη λύση μη εφικτή.

6.3.2 Μαθηματικές μέθοδοι βελτιστοποίησης

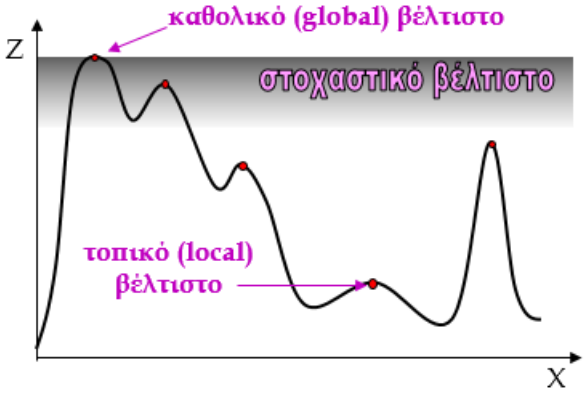
Στις μεθόδους βελτιστοποίησης μπορούμε να διακρίνουμε τις εξής δυο κατηγορίες:

- **Αιτιοκρατικές (deterministic)**

Στις Αιτιοκρατικές μεθόδους βελτιστοποίησης συγκαταλέγονται ειδικοί αλγόριθμοι, όπου ακολουθούν την κλίση της αντικειμενικής μας συνάρτησης και καθοδηγούν την διαδικασία σε τιμές εντός του πεδίου τιμών που διαρκώς βελτιώνουν τα κριτήρια απόδοσης. Σε μεθόδους αυτού του είδους, ένα τοπικό βέλτιστο δεν πιστοποιεί και το καθολικό βέλτιστο του προβλήματος, διότι η τελική λύση εξαρτάται από την αρχική τιμή των μεταβλητών καθώς και τις τοπικές τιμές της κλίσης.

- **Στοχαστικές (stochastic)**

Οι στοχαστικές μέθοδοι βελτιστοποίησης ενδείκνυνται για συνδυαστικά συστήματα μεγάλης κλίμακας που έχουν πολλά ακρότατα, καθώς επίσης και για προβλήματα με περισσότερες της μιας αντικειμενικής συνάρτησης. Η στοχαστική αναζήτηση εφαρμόζει τυχαίες μεταβάσεις στο σύνολο των εναλλακτικών λύσεων, με πιθανότητες μετάβασης που ενισχύουν στατιστικά την προοδευτική προσέγγιση των περιοχών βέλτιστων λύσεων. Το γεγονός ότι δεν ακολουθούν την κλίση της συνάρτησης αλλά επιτρέπουν σε περιοχές τιμών που δεν βελτιώνουν την αντικειμενική συνάρτηση δίνει την δυνατότητα η λύση να συγκλίνει πολύ κοντά στο καθολικό βέλτιστο.



Διάγραμμα 6.1. Αναπαράσταση της επίλυσης προβλήματος βελτιστοποίησης με στοχαστικές μεθόδους

6.3.3 Μέθοδοι στοχαστικής βελτιστοποίησης

Σε αυτήν την κατηγορία ανήκουν οι:

- Γενετικοί αλγόριθμοι (genetic algorithms)

Οι Γενετικοί αλγόριθμοι αναζήτησης μιμούνται τους βιολογικούς μηχανισμούς της φυσικής εξέλιξης που συναντώνται στους ζωντανούς οργανισμούς προκειμένου να επιτευχθεί καλύτερη προσαρμογή στις συνθήκες του περιβάλλοντος. Αυτού του τύπου οι αλγόριθμοι πραγματοποιούν αναζήτηση σε πολλές κατευθύνσεις και έχουν την δυνατότητα να συνδυάζουν και να αναπαράγουν τις σχετικά «καλές» λύσεις ενώ οι σχετικά «κακές» αφαιρούνται. Ο διαχωρισμός και η αξιολόγηση των λύσεων γίνεται με την βοήθεια της αντικειμενικής συνάρτησης.

- Αλγόριθμοι ευφυΐας σμήνους (swarm intelligence)

  Η αναζήτηση λύσεων με τον αλγόριθμο ευφυΐας σμήνους βασίζεται στην ίδια τακτική. Φυσικά παραδείγματα της νοημοσύνης σμήνους περιλαμβάνουν τις αποικίες μυρμηγκιών, τα σμήνη πουλιών κ.λπ. Η αρχή της συλλογικής συμπεριφοράς είναι για παράδειγμα η εύρεση της συντομότερης διαδρομής μεταξύ τροφής-αποικίας, με χρήση φερορομόνης.

- Αναζήτηση με απαγορευμένες μεταβάσεις (tabu search)

  Μετα-ευρετική μέθοδος τοπικής αναζήτησης με προσωρινή «απαγόρευση» της αντιστροφής των πιο πρόσφατων μεταβάσεων. Στην αναζήτηση ταμπού γίνονται αποδεκτές λύσεις που δεν οδηγούν στην βελτίωση. Έτσι είναι δυνατό ο αλγόριθμος να επιστρέψει σε λύσεις που έχει ήδη επισκεφθεί με αποτέλεσμα το πρόβλημα της ανακύκλωσης λύσεων. Για την υπερνίκηση αυτού του προβλήματος εφαρμόζεται η «στρατηγική απαγόρευσης» οπού ελέγχει και ενημερώνει τον κατάλογο περιορισμένης αναζήτησης

- Προσομοιωμένης ανόπτησης (simulated annealing)

  Ο αλγόριθμος προσομοιωμένης ανόπτησης είναι εμπνευσμένος από την μεταλλουργία και περιγράφεται εκτενώς στη συνέχεια του κεφαλαίου 6

6.4 Ανόπτηση

Ο όρος ανόπτηση προέρχεται από τον κλάδο της μεταλλουργίας και περιγράφει την θερμική κατεργασία στην οποία υποβάλλεται ένα μέταλλο ή κράμα. Η ανόπτηση είναι μία διαδικασία που το μεταλλουργικό προϊόν εκτίθεται για παρατεταμένη χρονική περίοδο σε υψηλή θερμοκρασία και στη συνέχεια ψύχεται με χαμηλούς ρυθμούς, μέχρι και την θερμοκρασία περιβάλλοντος. Ο χαμηλός ρυθμός απόψυξης εγγυάται πως το υλικό βρίσκεται σε συνθήκες θερμοδυναμικής ισορροπίας (προσεγγιστικά) κάθε χρονική στιγμή. Εφαρμόζεται σε περιπτώσεις όπου το υλικό έχει υποστεί κάποιο είδος κατεργασίας που του έχει επιφέρει εσωτερικές τάσεις και κρυσταλλικές διαταραχές, με σκοπό την μείωση των τάσεων και την στερεοποίηση του υλικού σε μια κρυσταλλική κατάσταση χαμηλής ενέργειας.

6.4.1 Αλγόριθμος Προσομοιωμένης Ανόπτησης (Simulated annealing)

Εισαγωγή

Οι Kirkpatrick et al. (1983) εντοπίζοντας πολλές ομοιότητες μεταξύ των προβλημάτων βελτιστοποίησης και της διεργασίας annealing πρότεινε μια στοχαστική μέθοδο· τον αλγόριθμο προσομοιωμένης ανόπτησης (**S.A**.), συνδυάζοντας την στατιστική μηχανική (αλγόριθμος των Metropolis et al. (1953)) με την θερμοδυναμική (ανόπτηση). Το βασικό χαρακτηριστικό αυτού του αλγορίθμου είναι ότι επιτρέπει την μετάβαση τόσο σε καλύτερες όσο και σε χειρότερες καταστάσεις. Στην δεύτερη περίπτωση, η μετάβαση σε χειρότερες λύσεις γίνεται με κάποια πιθανότητα μικρότερη της μονάδας. Το γεγονός αυτό είναι ιδιαίτερα σημαντικό, διότι δίνει την δυνατότητα στον αλγόριθμο να απεγκλωβιστεί από τοπικά ακρότατα και να αναζητήσει την βέλτιστη λύση μέσα από μια ευρύτερη περιοχή λύσεων. Η δυνατότητα να μεταβεί σε λύσεις που δεν βελτιώνουν την αντικειμενική συνάρτηση και να γίνουν δεκτές ορίζεται από το κριτήριο των Metropolis et al. (1953) και στηρίζεται στην πιθανότητα του Boltzmann.Το κριτήριο των Metropolis et al. (1953) εξαρτάται από δύο παραμέτρους:

- Πρώτον, από την παράμετρο «θερμοκρασία» η οποία ξεκινάει από μια υψηλή τιμή ,όπως συμβαίνει στην ανόπτηση και με ορισμένο βήμα μείωσης καταλήγει στο ελάχιστο
- Δεύτερον, από την διαφορά που υπάρχει μεταξύ της τρέχουσας και της νέας λύσης ως προς την αντικειμενική συνάρτηση.

6.4.2 Αρχή λειτουργίας της Προσομοιωμένης Ανόπτησης

Ο αλγόριθμος αποτελείται από έναν αριθμό επαναλήψεων (Markov chain) ανα τιμή «θερμοκρασίας» (που στο εξής συμβολίζεται με β). Σε κάθε επανάληψη παράγεται μια νέα λύση και σύμφωνα με το κριτήριο των Metropolis et al. στην περίπτωση που είναι «καλύτερη», δηλαδή βελτιώνει την αντικειμενική συνάρτηση, τότε άμεσα ορίζεται ως η τρέχουσα λύση. Διαφορετικά, εάν η διαφορά που υπάρχει μεταξύ της τρέχουσας και της νέας λύσης στην αντικειμενική συνάρτηση είναι μεγαλύτερη από μηδέν, τότε δεχόμαστε την νέα λύση ως τρέχουσα με πιθανότητα:

$$B_{i,j} = exp\left(\frac{-\Delta E_{i,j}}{\beta}\right) \quad (6.1)$$

με $\Delta E_{i,j}>0$.

Όπου

i: τρέχουσα λύση

j: νεα δοκιμαστική λύση

$B_{i,j}$: πιθανότητα αποδοχής για μετάβαση από την τρέχουσα στην νέα λύση

β: θερμοκρασία προσομοιωμένης ανόπτησης (με τη μείωση της θερμοκρασίας εχουμε και μείωση της πιθανοτητας αποδοχής της νέας λύσης που δεν βελτιώνει την αντικειμενική συνάρτηση)

$\Delta E_{i,j}=E_i-E_j$

E_i:Η τιμή της ενέργειας στην κατάσταση i

E_j:Η τιμή της ενέργειας στηννέα κατάσταση j

Σε αντίθετη περίπτωση η τρέχουσα λύση παραμένει αμετάβλητη. Να σημειωθεί πως τα παραπάνω κριτήρια αφορούν προβλήματα ελαχιστοποίησης

Με την ολοκλήρωση κάθε αλυσίδας Markov πραγματοποιείται και μείωση της παραμέτρου ελέγχου β, δηλαδή της «θερμοκρασίας» και η αποδοχή νέων λύσεων που υποβαθμίζουν την αντικειμενική συνάρτηση γίνεται αυστηρότερη. Κριτήρια μείωσης της θερμοκρασίας (β):

$$\beta_{\kappa+1} = \left[1 + \frac{ln(1+\delta)\cdot\beta_\kappa}{\Delta E_{max}} \cdot \beta_\kappa\right]^{-1} \cdot \beta_\kappa \quad (6.2)$$

Όπου

β_κ= η θερμοκρασία (β) για την τρέχουσα αλυσίδα Markov (κ)

δ = στατιστική παράμετρος που ελέγχει την ταχύτητα της διαδικασίας ανόπτησης

και

$\Delta E_{max}(\beta_\kappa) = E_{max}(\beta_\kappa)-E_{min}$ (6.3)

Ο ρυθμός μείωσης της θερμοκρασίας εξαρτάται άμεσα από τηνμέγιστη απόκλιση $\Delta E_{max}(\beta_\kappa)$ και σύμφωνα με τους Aarts και vanLaarhoven (1985) το $\Delta E_{max}(\beta_\kappa)$ μπορεί να υπολογιστεί προσεγγιστικά ως :

$\Delta E_{max}(\beta_\kappa)=3\cdot\sigma(\beta_\kappa)$ (6.4)

με $\sigma(\beta_\kappa)$ = τυπική απόκλιση ενεργειών της β_κ

Στην περίπτωση, όμως, που το πλήθος των μετρήσεων δεν είναι αρκετό για να προκύψει μια ορθή τυπική απόκλιση, ο υπολογισμός της μέγιστης απόκλισης ενεργειών σύμφωνα με τους Marcoulaki και

Kokossis (1999) μπορεί να προκύψει από την μέγιστη ενέργεια σε κάθε θερμοκρασιακό επίπεδο ($E_{max}(\beta_к)$) και την ελάχιστη ενέργεια μέχρι εκείνη τη στιγμή (E_{min}*).

$$\Delta E_{max}(\beta_к) = min\{< E(\beta_к) > +3 \cdot \sigma(\beta_к), E_{max}(\beta_к)\} - E_{min}^{*} \quad (6.5)$$

Όπου

E^{*}_{min} = Είναι η ελάχιστη ενέργεια που έχει συναντήσει ο αλγόριθμος μέχρι τη στιγμή εκείνη

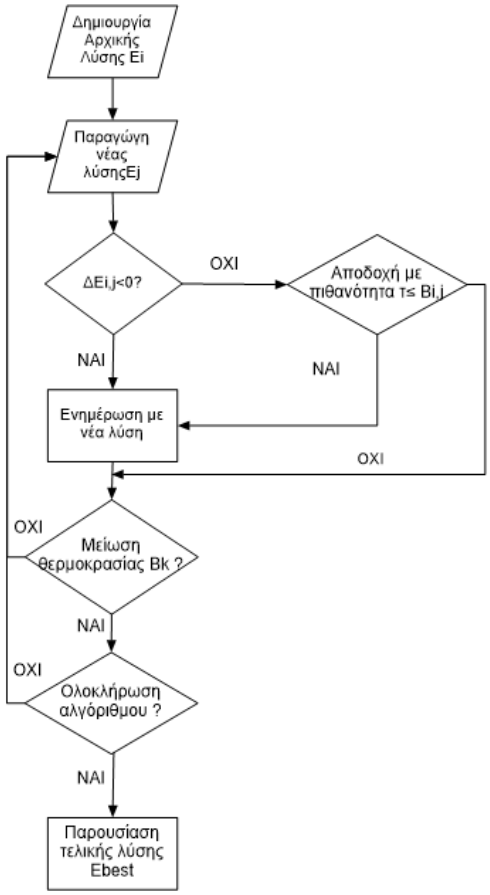
$\sigma(\beta_к)$ = τυπική απόκλιση ενεργειών κατά την αλυσίδα Tm θερμοκασίας $\beta_к$

$<E(\beta_к)>$=Μέση τιμή της ενέργειας για θερμοκρασία $\beta_к$

Η παραπάνω διαδικασία ολοκληρώνεται με κριτήριο τερματισμού την θερμοκρασία όταν αυτή πάρει τιμή πολύ κοντά στο μηδέν. Η τελική λύση μιάς διαδικασίας S.A. αποτελεί μέρος μιάς κατανομής πιθανότητας. Αυτή η κατανομή είναι πιο στενή και πιο μετατοπισμένη σε χαμηλές ενέργειες όσο πιο αργή είναι η διαδικασία.

6.4.3. Διάγραμμα Ροής Προσομοιωμένης Ανόπτησης

Ο αλγόριθμος προσομοιωμένης ανόπτησης όπως παρατίθεται παρακάτω στο διάγραμμα ροής ξεκινά από μια αρχική λύση (E_i) και προχωρά σε κάθε επανάληψη στην παραγωγή μιας νέας λύσης (E_j). Στην περίπτωση που η νέα λύση βελτιώνει την αντικειμενική συνάρτηση f, αυτομάτως γίνεται αποδεκτή και ενημερώνεται το σύστημα με την νέα λύση. Στην αντίθετη περίπτωση ένας τυχαίος αριθμός τα $\in[0,1]$ δημιουργείται από μια ομοιόμορφη κατανομή και αν ισχύσει $\tau \leq B_{i,j}$τότε η νέα λύση γίνεται αποδεκτή σαν η τρέχουσα λύση. Αν $\tau > B_i$, η λύση απορρίπτεται. Η επαναλληπτική αυτή διεργασιά ολακληρώνεται μόλις ένα κριτήριο τερματισμού ικανοποιηθεί. Στο σημείο αυτό αξίζει να σημειωθεί ότι το κριτήριο για την μείωση της θερμοκρασίας ορίζεται από τον χρήστη στον αλγόριθμο, με ένα δεδομένο μήκος αλυσίδας. Το μεγάλο μήκος προσφέρει καλύτερη αλλά πιο αργή σύγκλιση.



Σχήμα 6.1. Αλγόριθμος προσομοιωμένης ανόπτησης για επίλυση προβλημάτων βελτιστοποίησης

6.5 Περιγραφή μεθόδου βελτιστοποίησης μοριακών διαμορφώσεων

Στην παρούσα εργασία επιχειρείται σχεδιασμός χημικών προϊόντων με βέλτιστες ή/και επιθυμητές τιμές φυσικών ή/και θερμοδυναμικών ιδιοτήτων. Το πλαίσιο μοριακού σχεδιασμού που χρησιμοποιείται εδώ βασίζεται στις δημοσιεύσεις των Marcoulaki και Kokossis (1998, 2000α, 2000β), καθώς και το σχετικό λογισμικό που αναπτύχθηκε από τους Marcoulaki et al. (1997-2003). Σε αυτή την ενότητα της εργασίας επιχειρείται μια σύντομη περιγραφή της μεθόδου και του σχετικού αλγορίθμου.

Έστω:

- ένα σύνολο λειτουργικών ομάδων, των οποίων οι συνδυασμοί τους ορίζουν τις δομές,
- ένα σύνολο μαθηματικών μοντέλων για την πρόβλεψη των επιθυμητών ιδιοτήτων κάθε μορίου συναρτήσει των λειτουργικών ομάδων του
- μια συνάρτηση που να περιγράφει μαθηματικά

Βάσει τον παραπάνω δεδομένων το πρόβλημα της σύνθεσης νέων υλικών στοχεύει να καθορίσει:

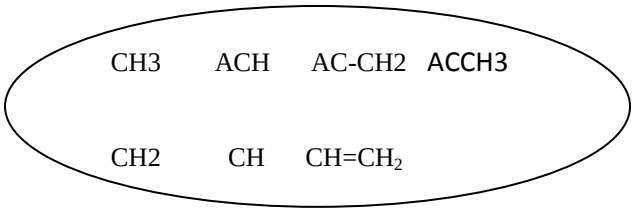
- τις βέλτιστες τιμές των αποδόσεων
- και τις κατάλληλες μοριακές δομές για την επίτευξη αυτών

Καθώς επίσης ο αριθμός των λειτουργικών ομάδων αντιπροσωπεύεται από έναν μεγάλο αριθμό διακριτών μεταβλητών, που ο συνδυασμός τους μπορεί να περιλαμβάνει μια συγκεκριμένη ομάδα περισσότερο από μία φορά. Βάσει των παραπάνω η συνδυαστική αναζήτηση γίνεται πιο πολύπλοκη.

6.6 Βασικοί ορισμοί

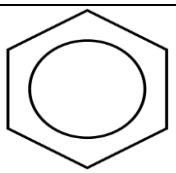
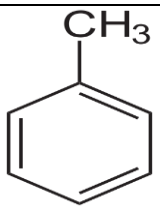
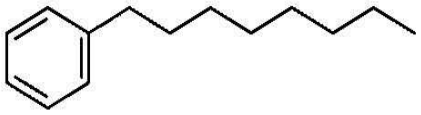
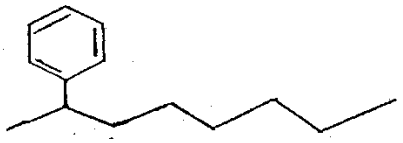
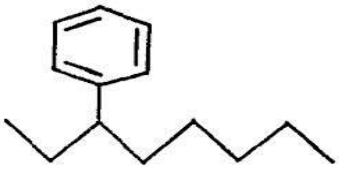
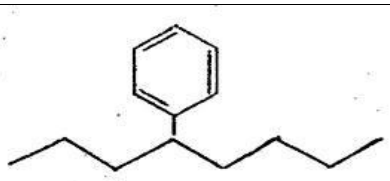
6.6.1 Λειτουργικές ομάδες

Στο σχήμα 6.2 δίνεται ένα σύνολο λειτουργικών ομάδων. Ο Πίνακας 6.1 δίνει παραδείγματα διαφόρων μορίων που θα επιχειρήσουμε να «αποδομήσουμε» στις λειτουργικές ομάδες που τα συγκροτούν.



Σχήμα 6.1 Σύνολο λειτουργικών ομάδων

Πινάκας 6.1. Παραδείγματα μορίων

Μόρια προς αποδόμηση	
Αιθάνιο	CH_3-CH_3
κανονικό-Οκτάνιο	CH_3-$(CH_2)_6$-CH_3
1-Οκτένιο	CH_2=CH_2-$(CH_2)_5$-CH_3
Βενζόλιο	
Τολουόλιο	
1-Φαινολοκτανιο	
2-Φαινολοκτανιο	
3-Φαινολοκτανιο	
4-Φαινολοκτανιο	

Στον πινάκα 6.2 στην πρώτη στήλη βρίσκεται το όνομα της ενώσεις και στις επόμενες παρατηρούμε τις επαναλήψεις κάθε λειτουργικής ομάδας σε κάθε ένωση.

Πινάκας 6.2. Λειτουργικές ομάδες

Ένωση	CH_3	CH_2	CH	ACH	$AC\text{-}CH_2$	$CH{=}CH_2$	$ACCH_3$
Αιθάνιο	2	0	0	0	0	0	0
κανονικό-Οκτάνιο	2	6	0	0	0	0	0
1-Οκτένιο	1	5	0	0	0	1	0
Βενζόλιο	0	0	0	6	0	0	0
Τολουόλιο	0	0	0	5	0	0	1
1-Φαινολοκτανιο	0	0	0	5	1	1	0
2-Φαινολοκτανιο	0	0	0	5	1	1	0
3-Φαινολοκτανιο	0	0	0	5	1	1	0
4-Φαινολοκτανιο	0	0	0	5	1	1	0

Για παράδειγμα, το μόριο του κανονικού οκτανίου αποτελείται από τις ομάδες CH_3 (2 φορές) και CH_2 (6 φορές). Αντίστοιχα, το μόριο του 1-φαινολοκτανίου αποτελείται από τις ομάδες ACH (5 φορές), $AC\text{-}CH_2$, CH, CH_2. Το μόριο του 2-φαινολοκτανίου αποτελείται από τις ομάδες $AC\text{-}CH_2$, CH, CH_2. Το ίδιο ακριβώς ισχύει και για τα μόρια των 3-φαινυλοκτανίου και 4-φαινυλοκτανίου.

6.6.2 Χαρακτηριστικά των λειτουργικών ομάδων

Οι λειτουργικές ομάδες έχουν διάφορα χαρακτηριστικά που τα χρησιμοποιεί ο αλγόριθμος των Marcoulaki και Kokossis (2000α) προκειμένου να εξασφαλίσει ότι ένα σύνολο λειτουργικών ομάδων μπορεί να συνιστά ένα εφικτό μόριο.

Αρωματική ομάδα

Αρωματική είναι μια ομάδα όταν περιέχει ένα άτομο άνθρακα που συμμετέχει σε αρωματικό δακτύλιο. Μια αρωματική ένωση μπορεί να περιέχει αρωματικές και μη αρωματικές ομάδες όπως είδαμε και παραπάνω με το παράδειγμα για το Φαινολοκτάνιο (5xACH, $AC\text{-}CH_2$,$CH{=}CH_2$) που έχει δυο αρωματικές ομάδες ACH και $AC\text{-}CH_2$ και δυο μη αρωματικές CH, CH_2

Μη-αρωματική ομάδα

Μη-αρωματική είναι μια ομάδα όταν δεν περιέχει κανένα άτομο άνθρακα συμμετέχον σε αρωματικό δακτύλιο. Μια μη αρωματική ένωση δεν μπορεί να περιέχει αρωματικές ομάδες . Για παράδειγμα, το κανονικό οκτάνιο (2xCH3, 6xCH2) περιχεί μόνο μη αρωματικές ομάδες.

Είναι σημαντικό να σημειωθεί ότι στο λογισμικό των Marcoulaki και Kokossis (2000α) δεν υπάρχει ασύνδετη η αρωματική ομάδα AC. Στον Πίνακα Β.1 του Παραρτήματος Β παρουσιάζονται όλες οι ομάδες που χρησιμοποιεί το λογισμικό των Marcoulaki και Kokossis (2000α).

6.6.3 Περιορισμοί αναπαράστασης και σύνθεσης μορίων

Όπως παρουσιάζεται στην Ενότητα 6.1, ο αλγόριθμος των Marcoulaki και Kokossis (2000α) κάνει μια αναζήτηση στο χώρο των δυνατών μοριακών διαμορφώσεων. Κάθε μόριο αναπαριστάται από ένα μοριακό διάνυσμα (molecular vector) που δηλώνει το πλήθος των ομάδων που συνιστούν το μόριο. Για παράδειγμα, τα μόρια του **Πινάκα 2.2** απεικονίζονται ως:

- Αιθάνιο: **M**3 = [2xCH3]
- κανονικό Οκτάνιο: **M**1 = [2xCH3 6xCH2]
- 1-Οκτενιο: **M**2 = [1xCH3 5xCH2 1xCH=CH3]
- Βενζόλιο: **M**4 = [6xACH]
- Τολουόλιο: **M**5 = [5xACH, 1xACCH3]
- 1-Φαινολοκτάνιο **M**6 = [5 x ACH, 1xAC-CH2, CH=CH2]
- 2-Φαινολοκτάνιο **M**7 = [5 x ACH, 1xAC-CH2, CH=CH2]
- 3-Φαινολοκτάνιο **M**8 = [5 x ACH, 1xAC-CH2, CH=CH2]
- 4-Φαινολοκτάνιο **M**9 = [5 x ACH, 1xAC-CH2, CH=CH2]

Η αναζήτηση που κάνει ο αλγόριθμος βασίζεται σε μια επαναληπτική διαδικασία που ξεκινάει από ένα μόριο, το «μεταλλάσει» σε δεύτερο μόριο, το δεύτερο «μεταλλάσσεται» σε τρίτο κ.ο.κ. Οι «μεταλλάξεις» αυτές στο εξής θα ονομάζονται *μετατοπίσεις*. Οι λειτουργικές ομάδες συμμετέχουν στις μετατοπίσεις αυτές και σε αυτό το πλαίσιο οι Marcoulaki και Kokossis (2000α) όρισαν ότι κάθε ομάδα διαθέτει τρεις μορφές συγγένειας. Συγκεκριμένα:

- Συγγένεια «χαμηλότερου σθένους» που προκύπτει από την προσθήκη ενός ατόμου υδρογόνου στην ομάδα. Π.χ. το CH_2 που έχει σθένος 2, συγγενεύει έτσι με το CH_3 που έχει ένα παραπάνω Η και σθένος 1. Για την κάθε ομάδα g η αντίστοιχη ομάδα χαμηλότερου σθένους συμβολίζεται με $f_{\sigma-}(g)$.

- Συγγένεια «υψηλότερου σθένους» που προκύπτει από την αφαίρεση ενός ατόμου Η στο group. Π.χ. το CH_2 που έχει σθένος 2, συγγενεύει έτσι με το CH που έχει ένα λιγότερο Η και σθένος 3. Για την κάθε ομάδα g η αντίστοιχη ομάδα υψηλότερου σθένους συμβολίζεται με $f_{\sigma+}(g)$.
- Συγγένεια τύπου, για τη συγγένεια μη-αρωματικής ομάδας με αρωματική ομάδα (και το αντίστροφο). Η αρωματική ομάδα προκύπτει από την κάλυψη ενός ελεύθερου δεσμού της μη-αρωματικής ομάδας από έναν αρωματικό άνθρακα. Π.χ η μη-αρωματική ομάδα CH2 συγγενεύει έτσι με την αρωματική ομάδα AC-CH2. Για την κάθε ομάδα g η συγγενής ομάδα τύπου συμβολίζεται με $f_t(g)$.
- Επιπλέον, κάθε λειτουργική ομάδα έχει ελεύθερους δεσμούς, το σθένος της, που επιτρέπουν τη σύνδεσή της με άλλες ομάδες ώστε κάθε ομάδα να αποτελεί μέρος ενός μορίου. Για την κάθε ομάδα g το σθένος της συμβολίζεται με $f_s(g)$.

Στο Πινάκα 2.3 στην $3^{η}$ στήλη είναι αριθμός ελεύθερων δεσμών κάθε λειτουργικής ομάδας, στην $4^{η}$ στήλη δίνεται ο τύπος της ομάδας (δηλαδή αν η λειτουργική ομάδα είναι αρωματική ή όχι), στην $5^{η}$ στήλη δίνεται η συγγενής ομάδα , στην $6^{η}$ στήλη δίνεται συγγενής «χαμηλότερου σθένους» και στην $7^{η}$ στήλη η συγγενής ομάδα «υψηλότερου σθένους».

Ας δούμε ένα σύντομο παράδειγμα για το σθένος και τον τύπο των λειτουργικών ομάδων ενός συνόλου G που ορίζεται ως G={CH4, CH3, CH2NH2, OH, ACCH3, ACOH, AC-COOH, AC-CH=CH2, ACCH2, ACCH, AC-CH=CH}

Το υποσύνολο του G που έχει όλες τις μη-αρωματικές ομάδες του G είναι

G_{na} ={CH4, CH3, CH2NH2,OH}

Το υποσύνολο του G_{na} που έχει όλες τις μη-αρωματικές ομάδες με σθένος ίσο με ένα είναι

$G_{na,1}$ = {CH3, CH2NH2, OH}

Το υποσύνολο του G που έχει όλες τις αρωματικές ομάδες του G είναι

G_a ={ACCH3, ACOH, AC-COOH, AC-CH=CH2, ACCH2, ACCH, AC-CH=CH}

Το υποσύνολο του G_a που έχει όλες τις αρωματικές ομάδες με σθένος ίσο με μηδέν είναι

$G_{a,0}$ = { ACCH3, ACOH, AC-COOH, AC-CH=CH2}

Πινάκας 6.3. Λειτουργικές ομάδες , σθένος ,συγγένεια χαμηλότερου σθένους ή υψηλότερου σθένους

α/α	Ομάδα (g)	fs(g)	ft(g)	fτ(g)	fσ-(g)	fσ+(g)
1	-CH3	1	Μη αρωμ.	ACCH3	-----	>CH2
2	>CH2	2	Μη αρωμ	ACCH2-	-CH3	>CH-
3	>CH-	3	Μη αρωμ	ACCH<	>CH2	>C<
4	>C<	4	Μη αρωμ	ACC<	>CH-	-----
5	CH2=CH-	1	Μη αρωμ	AC--CH2=CH	-----	-CH=CH-
6	-CH=CH-	2	Μη αρωμ	AC--CH=CH	CH2=CH-	-CH=C<
7	CH2=C<	2	Μη αρωμ	AC- >C= CH2	CH2=CH-	-CH=C<
8	-CH=C<	3	Μη αρωμ	AC- >C= CH-	-CH=CH-	-C=C<
9	>C=C<	4	Μη αρωμ	AC- >C=C-	-CH=C<	-----
10	-OH	1	Μη αρωμ	ACOH	-----	-----
11	ACH	0	Αρωμ.	-----	-----	-----
12	ACCH3	0	Αρωμ.	-CH3	-----	ACCH2-
13	ACCH2-	1	Αρωμ.	>CH2	ACCH3	ACCH<
14	ACCH<	2	Αρωμ.	>CH-	ACCH2-	ACC<
15	ACC<	3	Αρωμ.	>C<	ACCH<	-----
16	AC--CH=CH2	0	Αρωμ.	CH2=CH-	-----	AC--CH=CH
17	AC--CH=CH	1	Αρωμ.	-CH=CH-	AC--CH=CH	AC- >C= CH-
18	AC- >C=CH2	1	Αρωμ.	CH2=C<	AC--CH=CH	AC- >C= CH-
19	AC- >C=CH-	2	Αρωμ.	-CH=C<	AC--CH=CH	AC- >C=C<
20	AC- >C=C-	3	Αρωμ.	>C=C<	AC- >C= CH-	-----
21	ACOH	0	Αρωμ.	-OH	-----	-----

6.7 Πλαίσιο βελτιστοποίησης

Σε αυτή την ενότητα, παρουσιάζεται το πλαίσιο λειτουργίας του αλγορίθμου βελτιστοποίησης. Πιο συγκεκριμένα, περιγράφονται οι εναλλακτικές μετατοπίσεις που μπορούν να εφαρμοστούν σε ένα μόριο. Έστω το μόριο CH_3-$(CH_2)_4$-CH_3. Σε αυτό μπορούμε να:

- αντικαταστήσουμε την ομάδα CH3 με την ομάδα -CH=CH- που έχει το ίδιο σθένος
- προσθέσουμε την ομάδα -CH=CH- που έχει σθένος 2 και να επεκτείνουμε το μόριο.
- προσθέσουμε την ομάδα –CH< που έχει σθένος 3. Σε αυτή την περίπτωση καταλήγουμε σε μια διάταξη που έχει έναν ελεύθερο δεσμό. Για να διορθώσουμε αυτό το «πρόβλημα» μπορούμε να

αντικαταστήσουμε μια ομάδα CH_2 με την συγγενή της χαμηλότερου σθένους CH3.

- αφαιρέσουμε την ομάδα CH_3 που έχει σθένος 1 και να συρρικνώσουμε το μόριο
- αφαιρέσουμε την ομάδα CH_2 που έχει σθένος 2 και να συρρικνώσουμε το μόριο ακόμα περισσότερο

Στη συνέχεια παρουσιάζονται κάποια παραδείγματα για κάθε μια από τις αλλαγές που προαναφέρθηκαν, ξεκινώντας από την αντικατάσταση :

Ένα διάνυσμα M3= [2 CH_3,CH, CH=O]

Αρχικό μόριο ➔ CH(CH3)(CH=O)-CH3

Αντικατάσταση ➔ CH3-CH=O με COO

νέο μοριακό διάνυσμα M3'=[2CH3,CH,COO]

a) CH3-CH-(CH=O)-CH3 ➔ COO-CH(CH3)-CH3

b) CH3-⬡-CH2-CH(OH)-CH2-⬡-NH2

➔CH2Cl-⬡-CH2-CH(OH)-CH2-⬡-NH2

Όπως παρατηρούμε στο παραπάνω παράδειγμα στην αλλαγή με αντικατάσταση μία ομάδα δίνει την θέση της σε μία άλλη ομάδα ίδιου σθένους και ίδιου τύπου.

Παραδείγματα διαδοχικών επεκτάσεων ενός μορίου με προσθήκη μιας επιπλέον λειτουργικής ομάδας:

Αρχικό μόριο ➔⬡-CHCl-CH2-COOH

a) +Br ➔ ⬡-CH(Br)Cl-CH2-COOH

b) +CH2 ➔ ⬡-CHCl-CH2- CH2-COOH

c) +ACCH3 ➔CH3-⬡-CHCl-CH2-COOH

d) +ACCH< ➔[>CH]-⬡-CHCl-CH2-COOH

Στο παράδειγμα επέκτασης μια καινούργια λειτουργική ομάδα συμπεριλαμβάνεται στο μόριο.

Παράδειγμα συρρίκνωσης του μορίου:

Σταθερό μέρος : HO-⬡-CH(Cl)-CH(-CH2CN)-CH2-CH3

a) –CH2CN➔ HO-  -CHCl-CH-CH2-CH3

b) –CH2 ➔ HO-  -CH(Cl)-CH(CH2CN)-CH3

c) –ACOH➔  -CH(Cl)-CH(CH2CN)-CH2-CH3

d) –AC-CH(Cl)➔ OH-  -CH(-CH2CN)-CH2-CH3

e) –CH3➔ HO-  -CH(-Cl)-CH(-CH2CN)-CH2

Στο παράδειγμα συρρίκνωσης έχουμε την αφαίρεση μιας από τις υπάρχουσες λειτουργικές ομάδες του μορίου.

Κεφάλαιο 7. Εφαρμογή των νέων μοντέλων LCA στο Μοριακό Σχεδιασμό (σχεδιασμός καινοτόμων διαλυτών)

7.1 Χαρακτηριστικές ιδιότητες διαλυτών για διαχωρισμούς

Η εύρεση ενός κατάλληλου διαλύτη για διεργασίες διαχωρισμού όπου απαιτείται η προσθήκη ενός διαλυτικού μέσου, S. για τη διευκόλυνση του διαχωρισμού δύο ουσιών A και B αποτελεί είναι ένα τυπικό πρόβλημα που αντιμετωπίζεται με CAMD. Τέτοιες διεργασίες διαχωρισμού περιλαμβάνουν την εκχύλιση υγρού-υγρού (LL-extraction) (Marcoulaki, Kokossis, 2000 και Marcoulaki, Kokossis, 1998), την εκχυλιστική απόσταξη (extractive distillation) (Marcoulaki, Kokossis, 2000), εκχυλιστική ζύμωση (extractive fermentation) (Marcoulaki, Batzias, 2003) και απορρόφηση αερίου από υγρό (gas absorbtion) (Marcoulaki, Kokossis, 2000). Οι Marcoulaki et al.(2001) επέκτειναν το εργαλείο CAMD στο σχεδιασμό μίγματος διαλυτών και παρουσίασαν εφαρμογές σε διεργασίες εκχύλισης και εκχυλιστικής απόσταξης.

Σε αυτή την ενότητα περιγράφονται τα βασικά χαρακτηριστικά που χρησιμοποιούνται στην επιλογή διαλυτών για διεργασίες διαχωρισμού.

Η εκλεκτικότητα του διαλύτη εκφράζει τη διαχωριστική ικανότητα του διαλύτη, δηλαδή είναι η ικανότητα που έχει ώστε να διαλύει όσο το δυνατόν περισσότερα μόρια του εκχυλίσματος ως προς τα μόρια του υπολείμματος. Είναι ίση με το λόγο του συντελεστή κατανομής της διαλυμένης ουσίας που περιγράφεται παρακάτω προς το συντελεστή κατανομής της ουσίας που πρέπει να παραμείνει στο υπόλειμμα.

$$Εκλεκτικότητα\ διαλύτη: S_S = \frac{\gamma_{B,S^\infty}}{\gamma_{A,S^\infty}} \cdot \frac{MW_A}{MW_B} \quad (7.1)$$

Οι δείκτες A, B και S συμβολίζουν το εκχύλισμα (τη διαλυμένη ουσία), δηλαδή τη βουτανόλη, το υπόλειμμα, δηλαδή το νερό και το διαλύτη, αντίστοιχα.

Ο συντελεστής κατανομής διαλυμένης ουσίας εκφράζει την κατανομή της ποσότητας του εκχυλίσματος στη φάση του διαλύτη σε σχέση με την ποσότητα στη φάση του υπολείμματος.

$$Συντελεστής\ κατανομής\ διαλυμένης\ ουσίας\ (βουτανόλης): M = \frac{\gamma_{A,B^\infty}}{\gamma_{A,S^\infty}} \cdot \frac{MW_B}{MW_S} \quad (7.2)$$

Οι απώλειες διαλύτη εκφράζουν το ποσό του διαλύτη που απομακρύνεται με το υπόλειμμα. Αποτελεί και ένα μέτρο έκφρασης της εκλεκτικότητας του διαλύτη και της μη αναμιξιμότητας μεταξύ υπολείμματος και διαλύτη.

$$Απώλειες\ διαλύτη\ στο\ υπόλειμμα: S_l = \frac{1}{\gamma_{S,B}^{\infty}} \cdot \frac{MW_S}{MW_B} \quad (7.3)$$

7.2 Περιγραφή μελέτης περίπτωσης

Στην παρούσα εργασία, μελετάται ένα τυπικό πρόβλημα διαχωρισμού μίγματος που περιέχει κανονική βουτανόλη και νερό με χρήση εκχύλισης υγρού-υγρού. Σε αυτό το παράδειγμα, η διαλυμένη ουσία, Α, στις εξισώσεις 7.1 και 7.3 είναι η βουτανόλη και το υπόλειμμα Β είναι το νερό. Υποθέτουμε πως μετά την εκχύλιση ακολουθεί ανάκτηση του διαλύτη του διαλύτη με χρήση απλής κλασματικής απόσταξης. Τα δεδομένα του προβλήματος δίνονται στην εργασία των Marcoulaki και Kokossis (2000) που αντιμετώπισαν αυτό το πρόβλημα χωρίς περιορισμούς LCA.

Εξετάζονται λοιπόν οι παρακάτω περιπτώσεις:

- Περίπτωση Βάσης: το πρόβλημα επιλύεται χωρίς περιορισμούς LCA. Οι υπόλοιποι περιορισμοί είναι κατά την εργασία των Marcoulaki και Kokossis (2000).

- Περιπτώσεις LCAi: το πρόβλημα επιλύεται με περιορισμούς LCA, δηλαδή τίθενται τρία ανώτατα όρια για τις τιμές GWP, CED, EI 99 που συμβολίζονται με GWP_{max}, CED_{max}, EI 99_{max} αντίστοιχα. Παρουσιάζονται αποτελέσματα για δύο υποπεριπτώσεις, (1) με ελαστικούς περιορισμούς και (2) με αυστηρούς περιορισμούς ως προς τα LCA χαρακτηριστικά του διαλύτη.

Όπως έχει αναφερθεί και στο Κεφάλαιο 5, δεν υπάρχουν διαθέσιμες συνεισφορές για όλες τις ομάδες που αρχικά προτάθηκαν από τους Marcoulaki και Kokossis (2000) (106 ομάδες). Γι' αυτό, στις περιπτώσεις LCAi όπου είναι απαραίτητοι οι υπολογισμοί των τριών δεικτών LCA, εξαιρούνται από τη διαδικασία αναζήτησης όλες οι ομάδες που δεν συμπεριλήφθησαν στη διαδικασία ανάπτυξης των μοντέλων LCA (βλ. Κεφάλαιο 5).

Ο σχεδιασμός του διαλύτη πρέπει να επιτυγχάνει μέγιστο συντελεστή κατανομής (Μ) ώστε η μέγιστη ποσότητα βουτανόλης να βρίσκεται στη φάση του εκχυλίσματος. Ακόμα, η εκλεκτικότητα του διαλύτη (S_s) πρέπει να είναι επαρκώς υψηλή ώστε η ποσότητα του νερού που θα βρεθεί στη φάση του εκχυλίσματος να είναι ελάχιστη σε σχέση με την ποσότητα της βουτανόλης για να επιτευχθεί ο βέλτιστος διαχωρισμός. Είναι, επιπλέον απαραίτητο να εξασφαλισθεί η όσο το δυνατόν μικρότερη αναμιξιμότητα του διαλύτη και του

νερού. Γι' αυτό και οι απώλειες του διαλύτη στο υπόλειμμα (S_l) πρέπει να είναι χαμηλές. Τέλος, η θερμοκρασία βρασμού του διαλύτη πρέπει να έχει επαρκή διαφορά από τη θερμοκρασία βρασμού της βουτανόλης, ώστε να επιτυγχάνεται διαχωρισμός τους με απλή κλασματική απόσταξη. Τα παραπάνω εξασφαλίζουν το ελάχιστο μέγεθος εξοπλισμού των διεργασιών διαχωρισμού και της ανακύκλωσης.

Κατά συνέπεια, ο στόχος μας είναι η μεγιστοποίηση του συντελεστή κατανομής της βουτανόλης στο διαλύτη έναντι του νερού. Επομένως, η αντικειμενική συνάρτηση του προβλήματος βελτιστοποίησης διατυπώνεται μαθηματικά ως εξής:

Min(f): $f=\frac{1}{M}$ (7.4)

Πίνακας 7.1. Περιορισμοί για το πρόβλημα βελτιστοποίησης

Περιορισμοί	**Περίπτωση Βάσης**	**Περίπτωση LCA1**	**Περίπτωση LCA2**
$\mathbf{S_{s,min}}$	7	7	7
$\mathbf{M_{min}}$	1	1	1
$\mathbf{S_{l,max}}$	0,1	0,1	0,1
$\mathbf{T_{min}}$	320	320	320
$\mathbf{T_{max}}$	514,7	514,7	514,7
$\mathbf{GWP_{max}}$	-	10	8
$\mathbf{CED_{max}}$	-	200	150
$\mathbf{EI\ 99_{max}}$	-	0,5	0,4

7.3 Παρουσίαση αποτελεσμάτων

Τα αποτελέσματα που προέκυψαν στις περιπτώσεις βάσης, LCA1 και LCA2 φαίνονται στον πίνακα 7.2 και στον πίνακα 7.3 παρουσιάζονται για την κάθε λύση μερικές εναλλακτικές, προτεινόμενες δομές σε μορφή SMILES. Έπειτα, στον πίνακα 7.4 παρουσιάζονται οι τιμές που λαμβάνει η αντικειμενική συνάρτηση, η διαλυτότητα, οι απώλειες του διαλύτη, η θερμοκρασία βρασμού του διαλύτη και οι τρεις δείκτες LCA. Τέλος, στο διάγραμμα 7.1 παρουσιάζονται σε διαγραμματική μορφή τα αποτελέσματα της αντικειμενικής συνάρτησης για τις λύσεις κάθε περίπτωσης.

Πίνακας 7.2 Προτεινόμενοι διαλύτες για το πρόβλημα διαχωρισμού Βουτανόλης-Νερού

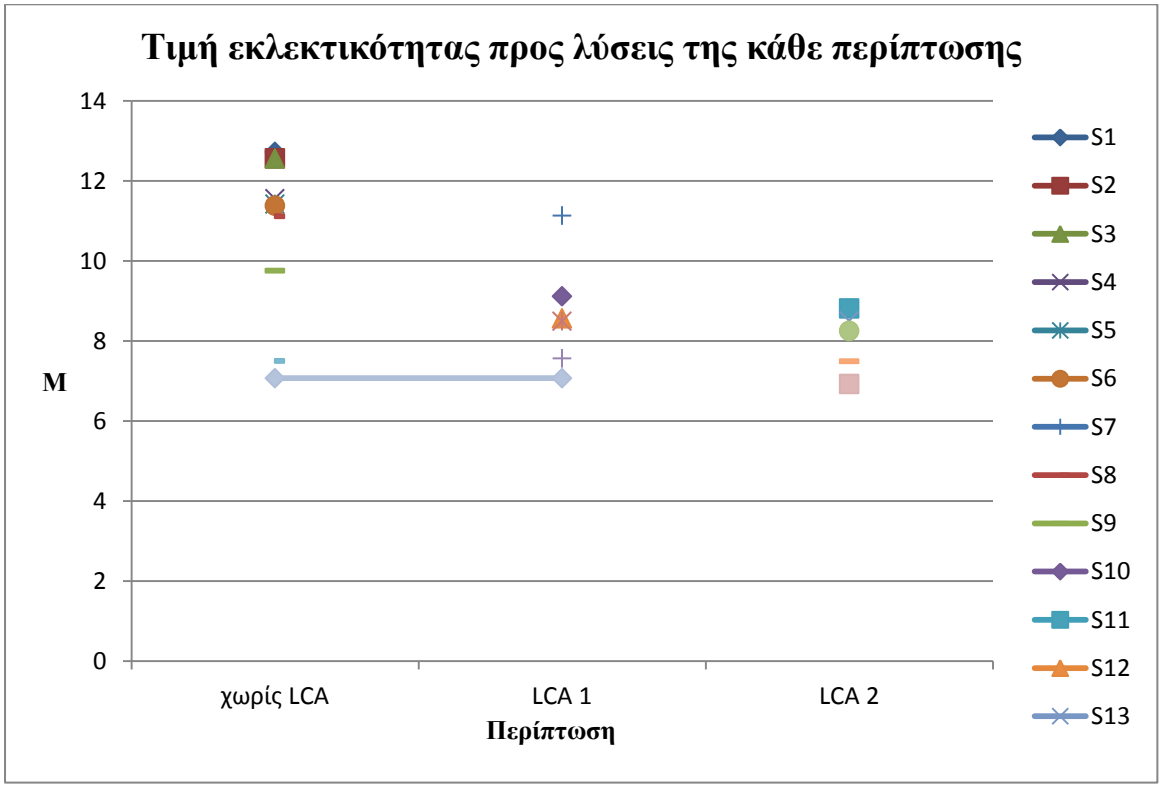
Περίπτωση	**Λύσεις αλγορίθμου CAMD**									
Βάση	S1	S2	S3	S4	S5	S6	S8	S9	S17	S19
LCA 1	S7	S10	S12	S14	S16	S19				
LCA 2	S11	S13	S15	S18	S20					

Πίνακας 7.3. Λειτουργικές ομάδες και προτεινόμενες δομές των λύσεων CAMD Διαλύτης

Διαλύτης	Λειτουργικές Ομάδες	Εναλλακτικά Μόρια
S1	5xCH3, 2xC, 1xCH2-CN	CC(C)(C)C(C)(C)CC#N CC(C#N)(C)C(C)(C)CC
S2	3xCH3, 3xCH2, 1xC, 1xCH2-CN	CCCC(C)(C)CCC#N CC(CC)(CCC)CC#N
S3	4xCH3, 1xCH2, 1xCH, 1xC, 1xCH2-CN	CC(C)CC(C)(C)CC#N CC(CC)(C)C(C)CC#N
S4	4xCH3, 1xC, 1xCH=C, 1xCH2-CN	CC(C)(C)C=C(C)CC#N CC=C(C)C(C)(C)CC#N
S5	2xCH3, 1xCH2, 1xC, 1xCH2=CH, 1xCH2-CN	CCC(C)(C=C)CC#N CC(C)(CC=C)CC#N
S6	3xCH3, 1xCH2, 1xCH, 1xCH=C, 1xCH2-CN	CCC(C)C=C(C)CC#N N#CCC=C(CC)C(C)C N#CCC(CC)C=C(C)C
S7	4xCH2, 1xCH2=CH, 1xCH2-CN	C=CCCCCCC#N
S8	1xCH3, 2xCH2, 1xCH, 1xCH2=CH, 1xCH2-CN	C=CC(C)CCCC#N C=CCC(CC)CC#N
S9	2xCH3, 1xCH2, 1xCH2=C, 1xCH=C, 1xCH2-CN	C=C(C)CC=C(C)CC#N C=C(CC#N)C=C(CC)C
S10	1xCH3, 2xCH2, 1xCH=CH, 1xCH2=C, 1xCH2-CN	CCC=CCC(CC#N)=C N#CCC=C(CC=CC)=C
S11	3xCH3, 4xCH2, 1xCH, 1xCH-NH2	NC(C)CC(C)CCCC NC(CC)C(CC)CCC
S12	3xCH2, 1xCH2=CH, 1xCH2=C, 1xCH2-CN	C=CCCCC(CC#N)=C C=CCC(CCCC#N)=C
S13	3xCH3, 2xCH2, 1xCH, 1xCH=CH, 1xCH-NH2	NC(C)CCC(C)C=CC CC=CC(CC)CC(C)N
S14	1xCH3, 2xCH2, 1xCH, 1xCH2-CN, 1xCH2-Cl	CCC(CC#N)CCCl
S15	2xCH3, 3xCH2, 1xCH, 1xCH2=CH, 1xCH-NH2	C=CCCCC(C)C(C)N CCC(C=C)CC(CC)N
S16	2xCH3, 1xCH2, 1xC, 2xCH2=CH, 1xCH-NH2	C=CCC(C=C)(C)C(C)N CC(C)(C=C)CC(C=C)N
S17	2xCH3, 4xACH, 1xACCH2, 1xAC->CHNH2	c1c(C(C)N)cc(CC)cc1 (meta δομή) c1(C(C)N)ccc(CC)cc1 (para δομή) c1c(C(C)N)c(CC)ccc1 (ortho δομή)
S18	1xCH3, 1xCH2, 1xCH-NH2, 5ACH, 1xACCH2	c1c(CCC(C)N)cccc1
S19	1xCH2-CN, 5xACH, 1xACCH2	c1c(CCC#N)cccc1
S20	2xCH3, 1xCH-NH2, 4xACH, 2xACCH2	c1c(C)cc(C(C)N)cc1 (meta δομή) c1(C(C)N)ccc(C)cc1 (para δομή) c1c(C(C)N)c(C)ccc1 (ortho δομή)

Πίνακας 7.4 Χαρακτηριστικές ιδιότητες των διαλυτών του Πίνακα 7.3

Διαλύτης	Αντικειμενική Συνάρτηση	M	S_s	S_l	T_b	Μορ. Βάρος	GWP	CED	ECO
S1	0,079	12,72	173,1	0,048	472,1	139,2	6,961	208,6	0,695
S2	0,080	12,56	174,9	0,044	485,8	139,2	5,263	156,8	0,570
S3	0,080	12,55	175,0	0,044	479,0	139,2	5,966	179,8	0,639
S4	0,087	11,55	102,3	0,054	484,3	137,2	19,61	434,8	0,740
S5	0,088	11,42	71,53	0,096	466,9	123,2	6,316	182,6	0,564
S6	0,088	11,38	103,3	0,049	490,8	137,2	18,61	405,9	0,684
S7	0,090	11,13	71,77	0,088	481,0	123,2	4,618	133,7	0,440
S8	0,090	11,11	71,77	0,087	474,1	123,2	5,321	153,7	0,508
S9	0,103	9,756	56,63	0,051	494,8	135,2	17,51	387,1	0,661
S10	0,110	9,116	51,97	0,041	496,3	135,2	7,144	195,4	0,475
S11	0,114	8,811	20,84	0,083	461,4	143,3	0,496	64,23	0,379
S12	0,117	8,562	44,84	0,036	495,8	135,2	4,267	139,8	0,490
S13	0,118	8,496	17,74	0,096	461,3	141,2	4,475	147,5	0,363
S14	0,118	8,489	120,6	0,097	496,5	145,6	4,251	123,7	0,475
S15	0,121	8,251	16,75	0,089	460,7	141,3	1,598	91,95	0,378
S16	0,132	7,564	13,23	0,096	452,5	139,2	3,695	148,5	0,432
S17	0,133	7,496	23,94	0,059	502,1	149,2	1,108	45,10	0,214
S18	0,134	7,491	23,88	0,062	490,2	149,2	0,868	52,71	0,327
S19	0,142	7,067	119,5	0,041	501,7	131,2	4,591	114,5	0,457
S20	0,144	6,925	26,71	0,018	507,7	163,3	-1,478	-5,437	0,215



Διάγραμμα 7.1. Διαγραμματική απεικόνιση του συντελεστή κατανομής, M, προς τις λύσεις κάθε περίπτωσης

7.4 Συμπεράσματα

Στον Πίνακα 7.4, παρατηρείται πως ο διαλύτης S20 έχει αρνητικές τιμές δεικτών GWP και CED. Από τα όσα έχουν προαναφερθεί στην Ενότητα 3 για τους δείκτες αυτούς, γίνεται αντιληπτό πως αρνητικές τιμές δεν εκφράζουν πραγματικές τιμές, εφόσον μια ένωση δεν δύναται να έχει «αρνητικό» περιβαλλοντικό αποτύπωμα. Αρνητικές τιμές είναι γενικότερα αναμενόμενες με τη χρήση των μοντέλων της Ενότητας 5 μιας και υπάρχουν ομάδες, διαφορετικές για το κάθε μοντέλο, οι οποίες έχουν αρνητικές συνεισφορές. Σε μόρια (κυρίως μικρά), στα οποία επαναλαμβάνονται αυτές οι ομάδες αρκετές φορές, ενδέχεται να προκύπτουν αρνητικοί δείκτες. Δεδομένου, όμως, του εύρους των σφαλμάτων μπορεί να γίνει η υπόθεση πως οι τιμές αυτών των δεικτών θα ήταν αρκετά χαμηλές.

Στον Πίνακα 7.4, παρατηρείται η προσθήκη όλο και αυστηρότερων περιορισμών LCA, οδηγεί σε όλο και μεγαλύτερες τιμές της αντικειμενικής συνάρτησης, της εκλεκτικότητας και της θερμοκρασίας βρασμού, ενώ οι απώλειες του διαλύτη αυξάνονται.

Οι λόγοι για τους οποίους η μεθοδολογία επίλυσης συγκλίνει σε υποδεέστερες λύσεις με την προσθήκη κριτηρίων LCA σε σχέση με τις λύσεις της περίπτωσης βάσης είναι κατά πρώτον, ο περιορισμός της αναζήτησης λύσεων μόνο προς λύσεις που ικανοποιούν τις συνθήκες LCA και κατά δεύτερον, η μείωση του πλήθους των ομάδων που συμμετέχουν στην αναζήτηση, κάτι που συντελεί στην επιπλέον μείωση του συνόλου των εφικτών, πιθανών λύσεων ανάμεσα στις οποίες γίνεται η αναζήτηση. Έτσι, λοιπόν, η μεθοδολογία για να καλύψει τους όλο και αυστηρότερους περιορισμούς LCA, πρέπει να κατευθυνθεί προς λιγότερο καλές λύσεις και από άποψη αντικειμενικής συνάρτησης..

Όσον αφορά τις λύσεις, φαίνεται πως στην περίπτωση βάσης η μεθοδολογία κατευθύνθηκε κυρίως προς μόρια στη κατηγορία των «κυανιούχων» (νιτρίλια), δηλαδή μόρια με τις ομάδες $–CH_2CN$, $-CH_3CN$, κορεσμένα και ακόρεστα, αλειφατικά και κατά περιπτώσεις αρωματικά. Το ίδιο συμβαίνει και στην περίπτωση LCA 1, ενώ στην περίπτωση LCA 2 η μεθοδολογία κατευθύνεται προς λύσεις στην κατηγορία των αμινών (κορεσμένες και ακόρεστες, αλειφατικές και αρωματικές), ενώ να σημειωθεί πως οι αμίνες ως λύσεις έχουν εμφανιστεί και στις περιπτώσεις βάσης και LCA 1. Η παραπάνω αλλαγή στην κατεύθυνση των λύσεων είναι πλήρως δικαιολογημένη: τα νιτρίλια και γενικώς οι κυανιούχες ενώσεις έχουν μεγάλη επίδραση στη ρύπανση του περιβάλλοντος και συγκεκριμένα, του υδροφόρου ορίζοντα, του εδάφους και του αέρα (Mao et al. 2010). Είναι δυνατόν δε μέσω αντιδράσεων τα νιτρίλια να μετατραπούν σε HCN, που είναι τοξικότατο. Όσο οι περιορισμοί LCA, είναι ανύπαρκτοι έως και χαλαροί, η αλγοριθμική διαδικασία δε λαμβάνει υπόψη τις περιβαλλοντικές επιπτώσεις και επιλέγει ενώσεις από αυτήν την κατηγορία ως βέλτιστες στο ρόλο διαλύτη. Όταν, όμως, εισέρχονται αυστηρότεροι περιβαλλοντικοί περιορισμοί, τότε τα νιτρίλια αποτελούν πλέον μια ασύμβατη περιβαλλοντικά λύση και αντικαθίστανται από τις φιλικότερες προς το περιβάλλον αμίνες.

Η προσθήκη των περιορισμών LCA έχει οδηγήσει σε λύσεις όχι μόνο λιγότερο καλές από άποψη αντικειμενικής συνάρτησης, αλλά και περιορισμών. Συγκεκριμένα, ενώ οι λύσεις της περίπτωσης βάσης κινούνται σε τιμές εκλεκτικότητας στην περιοχή από 23,34-175, φαίνεται πως στην περίπτωση LCA 1 κινούνται σε αρκετά χαμηλότερες τιμές και συγκεκριμένα, 13,23-120,6. Στη δε LCA 2 περίπτωση, με τους περισσότερο αυστηρούς περιορισμούς, στην περιοχή 16,75-26,71. Όσον αφορά στις απώλειες διαλύτη, φαίνεται πως οι τιμές των λύσεων είναι αρκετά κοντά στις περιπτώσεις βάσης και LCA 1, με εύρος 0,041-0,096 και 0,036-0,097 αντίστοιχα ενώ αντίθετα, οι τιμές απώλειας διαλύτη για την περίπτωση LCA 2, έχουν αυξηθεί αρκετά αφού κινούνται στην περιοχή 0,062-0,096, με μια μόνο λύση με τιμή 0,018.

8. Συμπεράσματα-Δυνατότητες για μελλοντική έρευνα

Στην παρούσα εργασία αναπτύχθηκαν μοντέλα Συνεισφοράς Ομάδων για τον υπολογισμό τριών δεικτών που εμπίπτουν στη μεθοδολογία Ανάλυσης Κύκλου Ζωής: του Global Warming Potential, του Cumulative Energy Demand και του Ecoindicator 99 με τη χρήση διαφορετικών στατιστικών μεθοδολογιών γραμμικής και μη γραμμικής παλινδρόμησης. Αφού συγκρίθηκαν οι προβλεπτικές δυνατότητες των μοντέλων της εκάστοτε μεθοδολογίας και προέκυψε και αναλύθηκε το καλύτερο μοντέλο για κάθε δείκτη, δοκιμάστηκαν τα παραπάνω μοντέλα, συμμετέχοντας σε διαδικασία Μοριακού Σχεδιασμού με χρήση Ηλεκτρονικού Υπολογιστή για την εύρεση βέλτιστου διαλύτη διεργασίας υγρής εκχύλισης με χαλαρά και αυστηρά περιβαλλοντικά κριτήρια. Τα αποτελέσματα συγκρίθηκαν με τα αποτελέσματα της ίδιας διαδικασίας χωρίς περιβαλλοντικά κριτήρια.

Ανάλυση των μοντέλων LCA

Αναπτύχθηκαν μοντέλα πρόβλεψης με διαφορετικές στατιστικές μεθοδολογίες: Γραμμική Παλινδρόμηση Πολλών Μεταβλητών, Ανάλυση Πρωτευόντων Συνιστωσών, Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων, Παρεμβολή τύπου «Kriging», Μέθοδοι Συναρτήσεων Ακτινικής Βάσης και Μέθοδοι Συναρτήσεων Ακτινικής Βάσης σε συνδυασμό με Ανάλυση Πρωτευόντων Συνιστωσών. Τα μοντέλα που αναπτύχθηκαν με την μεθοδολογία PLS με 10 συνιστώσες δίνουν μια πρώτη εκτίμηση των δεικτών GWP, CED, EI 99, αλλά σε καμία περίπτωση δε μπορούν να θεωρηθούν αξιόπιστα ως προς τα τελικά τους αποτελέσματα. Παρότι τα μοντέλα αυτά πετυχαίνουν χαμηλά μέσα σχετικά σφάλματα ως GC, δηλαδή 24% για το GWP, 11% για το CED και 15% για το EI 99, ενώ το απόλυτο μέσο σχετικό σφάλμα είναι 62% για το GWP, 40% για το CED και 38% για το EI 99, οι συσχετίσεις που επιτυγχάνονται δεν ξεπερνούν το 31%. Οι συσχετίσεις, λοιπόν, δε μπορούν πάντα να υποσχεθούν καλά αποτελέσματα και τα παραπάνω μοντέλα μπορούν να χρησιμοποιηθούν μόνο για κάποιες πρώτες εκτιμήσεις και όχι για λήψη αποφάσεων. Οι λόγοι που αυτά τα μοντέλα δε λειτούργησαν τόσο αποτελεσματικά είναι οι παρακάτω:

- Το αρχικό σύνολο τιμών που χρησιμοποιήθηκε δεν ήταν αρχικά μεγάλο
- Μετά το διαχωρισμό σε σύνολο εκπαίδευσης και δοκιμής, ο λόγος των μορίων του συνόλου εκπαίδευσης προς τους περιγραφείς ήταν 171/57=3, δηλαδή πολύ χαμηλός
- Είναι πολύ πιθανόν να μην είναι δυνατόν, γενικά, να συσχετιστούν καλά οι δείκτες GWP, CED και EI 99 με το σύνολο των ομάδων με γραμμικό τρόπο ή με τον μη γραμμικό τρόπο του πλέγματος που επιχείρησαν οι μεθοδολογίες παρεμβολή τύπου «kriging», RBF και RBF-PCA.

- Η κατανομή των τιμών εξόδου(output) των δεικτών που δόθηκαν δεν είχαν ομαλή κατανομή. Αυτό σημαίνει πως στο σύνολο των πειραματικών και υπολογισμένων τιμών των δεικτών LCA, που χρησιμοποιήθηκαν στο σχηματισμό του μοντέλου υπήρχαν αρκετές πολύ μεγάλες τιμές που ξεχώριζαν αρκετά από το γενικό μέσο όρο των τιμών εξόδου. Αυτό επιδεινώνει αρκετά το σχηματισμό του μοντέλου, καθώς στη διαδικασία προσαρμογής μιας γραμμικής σχέσης στα πειραματικά, η μεθοδολογία ανάπτυξης πρέπει να πετύχει βέλτιστη προσαρμογή σε δεδομένα που είναι αρκετά διεσπαρμένα.
- Δεν πρέπει να αγνοηθεί πως αρκετοί από τους λόγους δύο δεικτών, του Συντελεστή Προσαρμογής και της Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων έχουν αντικατασταθεί με 0 και αυτό έχει σίγουρα μεγάλη επίδραση στα τελικά δεδομένα.
- Ο αριθμός των 1000 διαμερίσεων δεν πετυχαίνει όλους τους πιθανούς συνδυασμούς μορίων στο σύνολο εκπαίδευσης και δοκιμής.

Υπάρχουν, λοιπόν, αρκετές δυνατότητες βελτίωσης, αν μάλιστα αυξηθεί ο αριθμός αυτών των διαμερίσεων και αν αντιμετωπιστεί το πρόβλημα της έλλειψης πειραματικών μετρήσεων. Στο μέλλον, ακόμα, μπορεί να επιχειρηθεί παλινδρόμηση με χρήση ενός μόνο μορίου για τον έλεγχο της απόδοσης του μοντέλου(leave-one-out validation),αν και θα είναι απαραίτητη η αύξηση του αριθμού των διαμερισμών κατά πολύ για να φανεί η τάση των μοντέλων σε όλες τις ομάδες μορίων. Εν τέλει, υπάρχει περιθώριο να δοκιμαστούν για τα μοντέλα σχέσεις μη γραμμικές (εκθετικές, λογαριθμικές, ανώτερης τάξης πολυωνυμικές κ.λπ.).

Χρήση CAMD με περιβαλλοντικούς περιορισμούς για σχεδιασμό καινοτόμων διαλυτών

Μελετήθηκαν στην παρούσα εργασία οι δυνατότητες εργαλείου Μοριακού Σχεδιασμού με χρήση των παραπάνω μοντέλων για το σχεδιασμό καινοτόμων χημικών ουσιών. Συγκεκριμένα, μελετήθηκε η περίπτωση εύρεσης βέλτιστου διαλύτη για τη διεργασία εκχύλισης νερού-κανονικής βουτανόλης. Το πρόβλημα επιλύθηκε αρχικά χωρίς περιβαλλοντικούς περιορισμούς και έπειτα, σε μια περίπτωση με χαλαρούς περιβαλλοντικούς περιορισμούς και σε μια δεύτερη περίπτωση με αυστηρότερους περιβαλλοντικούς περιορισμούς. Στην πρώτη περίπτωση επίλυσης του προβλήματος με χαλαρούς περιβαλλοντικούς περιορισμούς δόθηκαν μέγιστες τιμές περιβαλλοντικής επιβάρυνσης στους τρεις δείκτες LCA, με την απαίτηση ο σχεδιαζόμενος διαλύτης να μην τις ξεπερνά. Στη δεύτερη περίπτωση επίλυσης με αυστηρότερους περιβαλλοντικούς περιορισμούς, οι μέγιστες τιμές των δεικτών μειώθηκαν περαιτέρω.

Η προσθήκη περιορισμών έθεσε νέα κριτήρια στην αναζήτηση βέλτιστου διαλύτη, περιορίζοντας το χώρο των εφικτών λύσεων και όπως είναι λογικό, με την προσθήκη όλο και αυστηρότερων περιορισμών, οι λύσεις του προβλήματος μετατοπίζονται σε νέες ομάδες μορίων. Τα αποτελέσματα του προβλήματος χωρίς

περιορισμούς LCA, επιβεβαιώνονται και από ερευνητικά δεδομένα, να είναι λιγότερο φιλικά προς το περιβάλλον από τις φιλικότερες ενώσεις-λύσεις του προβλήματος με αυστηρούς περιορισμούς.

Συγκεκριμένα, παρατηρείται πως οι βέλτιστες λύσεις για το πρόβλημα χωρίς περιβαλλοντικούς περιορισμούς εντοπίζονται κυρίως στις ομάδες των νιτριλίων. Πειραματικά δεδομένα επιβεβαιώνουν τις σοβαρές επιπτώσεις των ενώσεων αυτών στη ρύπανση του περιβάλλοντος. Οι λύσεις των νιτριλίων παραμένουν, παρολαυτά, και ως βασική κατηγορία λύσεων και στην περίπτωση της επίλυσης με χαλαρούς περιβαλλοντικούς περιορισμούς. Με την προσθήκη των αυστηρών περιβαλλοντικών κριτηρίων, όμως, μετατοπίζονται οι λύσεις του προβλήματος σε διαφορετική κατηγορία ενώσεων, τις αμίνες, οι οποίες είναι περιβαλλοντικά συμβατές.

Είναι, λοιπόν, εμφανές πως τα μοντέλα πρόβλεψης δεικτών LCA συνεργάστηκαν ικανοποιητικά με τη μεθοδολογία Μοριακού Σχεδιασμού για την εξαγωγή αποτελεσμάτων στο πρόβλημα της αναζήτησης βέλτιστου διαλύτη που να υπακούει σε περιορισμούς LCA και συστήνεται η χρήση του εργαλείου αυτού, όχι μόνο για την επίλυση παλαιών περιπτώσεων προβλημάτων, αλλά και για την επίλυση καινούργιων προβλημάτων

9. Βιβλιογραφία

Ενότητα 2.2

Poling B.E., Prausnitz J.M., O' Connel J.P., Properties of Gases and Liquids, 5th edition, 2001, Mc Graw-Hill

Gao C., Govind R. , Henry H. T., Application of the Group Contribution method for predicting the toxicity of organic chemicals, Environmental Toxicology and Chemistry, 1991; 11: 631-636

Constantinou L.,Gani R., New Group Contribution Method for Estimating Properties of Pure Compounds, AIChE Journal 1994; 40(10): 1697-1710

Lydersen, A. L., Estimation of Critical Properties of Organic Compounds, Univ. Wisconsin, Coll. Eng., Eng. Exp. Stn. rept. 3, April, 1955, Madison, WI

Missenard C. Properties of Gases & Liquids, 4th edition, McGraw-Hill, 1988

Ενότητα 2.3

Bondi A., Physical Properties of Molecular Crystals, Liquids, and Glasses, Journal of Polymer Science Part A-1: Polymer Chemistry 1969; 7(8): 2466

Fredenslund Aa., Jones R.L., Prausnitz J.M., Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures, AIChE Journal 1975; 21 (6): 1086-1099

Gmehling J., Li J., Schiller M., Modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties, Industrial and Engineering Chemistry Research 1993; 32(1): 178-193

Hansen H.K., Rasmussen P., Fredenslund Aa., Schiller M., Gmehling J., Vapor-liquid equilibria by UNIFAC group contribution. 5. Revision and extension, Industrial and Engineering Chemistry Research1991; 30(10): 2352-2355

Larsen B.L., Rasmussen P., Fredenslund A., A modified UNIFAC group-contribution model for prediction of phase equilibria and heats of mixing, Industrial and Engineering Chemistry Research 1987; 26(11): 2274-2286

Lohmann, J., Gmehling, J., Modified UNIFAC (Dortmund): Reliable model for the development of thermal separation processes, Journal of Chemical Engineering of Japan 2001; 34(1): 43-54

Tassios D., Εφαρμοσμένη Θερμοδυναμική Χημικού Μηχανικού, Πανεπιστημιακές Εκδόσεις Ε.Μ.Π., Αθήνα 2001, Κεφάλαιο 13, σελίδα 499

Κεφάλαιο 3

Life Cycle Assessment: Principle and Practice, Scientific Applications International Corporation, May 2006: 1-6

Goldsmith E., Allen R., A Blueprint for Survival, 1972, The Economist 2(1)

Meadows, D.H., The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind, 1972, Universe Books, New York. pp. 205

Rolf F., Jungbluth N (Editors) , Althaus H.J., Bauer C., Doka G., Dones R., Hischier R., Hellweg S., Humbert S., Köllner T., Loerincik Y, Margni M., Nemecek T., Implementation of Life Cycle Impact Assessment Methods, Data v2.0 2007, ecoinvent report No. 3, 2007, Swiss Centre for Life Cycle Inventories & Ecoinvent Centre

Eco-indicator 99, Manual for Designers, Ministry of Housing, Spatial Planning and the Environment, October 2000, The Netherlands

Wernet G., Hellweg S. ,Ulrich Fischer, Papadokonstantakis S., Hungerbuehler K., Molecular-Structure-Based Models of Chemical Inventories using Neural Networks, Environmental Science and Technology, 2008; 42(17): 6717–6722

Wernet G., Papadokonstantakis S., Hellweg S, Hungerbuehler K., Bridging data gaps in environmental assessments: Modeling impacts of fine and basic chemical production, Green Chemistry, 2009; 11(11): 1826–1831

Frischknecht R. , Jungbluth N. Althaus H.J., Doka G., Dones, R., Heck T., Hellweg S., Hischier R.' Nemecek T., Rebitzer G., Spielmann M., The ecoinvent database: Overview and methodological framework, International Journal of Life Cycle Assessment 2005; 10 (1): 3-9

Κεφάλαιο 4

Edmund R. Malinowski, Factor Analysis in Chemistry, 3rd edition, Wiley Interscience, 2002, Κεφάλαιο 8

I.T. Joliffe, Principal Component Analysis, 2nd edition, Springer, 2002, p.1-6

Forrester Alexander, Sobester Andras, Keane Andy, Engineering Design via Surrogate Modelling, A practical guide, Wiley, 2008

Randall D. Tobias, An Introduction to Partial Least Squares Regression, SAS Institute Inc., Cary, NC

Alexandridis A., Sarimveis H., Bafas G., A new algorithm for online structure and parameter adaptation of RBF networks, Neural Networks, 2003; 16(7): 1003–1017

infoman.teikav.edu.gr/e_education/118/files/ANOVA.doc, ΤΕΙ Καβάλας, Τμήμα Διαχείρισης Πληροφοριών

http://www.statsoft.com/textbook/anova-manova

Κεφάλαιο 5

De Maesschalck R., Jouan-Rimbaud D., Massart D.L., The Mahalanobis distance, 2000, Chemometrics and Intelligent Laboratory Systems, 2000; 50(1): 1–18

Ενότητες 6.1 και 6.2

Tassios D.P., Extractive and Azeotropic Distillation, Advances in Chemistry Series, 1972; 115: 46.

Abrams D.S., Prausnitz J.M., Statistical thermodynamics of liquids mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems, AIChE Journal, 1975; 21 (1): 116

Cockrem M., Flatt J., Lightfoot E., Solvent Selection for Extraction from Dilute Solution, Separation Science and Technology, 1989; 24 (11): 769-807

Gani R., Nielsen B., Fredenslund Aa, A Group Contribution Approach to Computer-Aided Molecular Design, AIChE Journal, 1991; 37 (9): 1318

Joback K. G., Stephanopoulos G., Designing Molecules Possessing Desired Physical Property Values Proceedings, 1989, FOCAPD'89, Snowmass, CO

Kolbe B., Gmehling J. and Onken U., Selection of Solvents for Extractive Rectification by means of Predicted Equilibria Data, 1979; 83 (11): 1133-1136

Magnussen T., Michelsen M.L. , Fredenslund Aa., Molecular Design of Solvents for Liquid Extraction based on UNIFAC, Fluid Phase Equilibria, 1979; 13: 331

Porter K. E., Sitthiosoth S., and Jenkins J. D., Designing a Solvent for Gas Absorption, Chemical Engineering Research and Design, 1991; 69 (3): 229

Naser S. F., Fournier R.L., A System for the Design of an Optimum Liquid-Liquid Extractant molecule, Computers and Chemical Engineering, 1991; 15 (6): 397

Bansal V., Ross R., Perkins, J.D., Pistikopoulos E.N., Optimal design and control of double-effect distillation systems , Source of the Document IEE Conference Publication, 1998; 455:1096-1101

Folić M., Adjiman C.S., Pistikopoulos E.N., Computer-aided solvent design for reactions: Maximizing product formation, Source of the Document Industrial and Engineering Chemistry Research, 2008; 47 (15): 5190-5202

Pretel E. J., Lopez P. A. Lopez, Bottini S.B. , Brignole E. A., Computer-Aided Molecular Design of Solvents for Separation Processes, AIChE Journal, 1994; 40 (8): 1349

Vaidyanathan R**.**, El-Halwagi M**.**, Computer-aided design of high performance polymers, Journal of Elastomers and Plastics, 1994; 26 (3): 277-293

Venkatasubramanian V., Chan K., Caruthers J.M., Computer-aided molecular design using genetic algorithms, Source of the Document Computers and Chemical Engineering 1994; 18 (9): 833-844

Mitrofanov I., Sansonetti S., Abildskov J., Sin G., Gani R., The Solvent Selection framework: Solvents for organic synthesis, separation processes and ionic liquids solvents, Computer Aided Chemical Engineering, 2012; 30: 762-766

Gani R. , Jiménez-González C. , , Crafts P.A. , Jones M. , Powell L. , Atherton, J.H., Cordiner, J.L., A modern approach to solvent selection, Chemical Engineering, 2006; 113 (3): 30-43

Karunanithi A.T., Achenie L.E.K., Gani R. , A computer-aided molecular design framework for crystallization solvent design, Chemical Engineering Science, 2006; 61 (4): 1247-1260

Wang Y.P., Achenie L.E.K. , A CAPD approach for reaction solvent design, Computer Aided Chemical Engineering, 2001; 9: 585–590

Gani, R., Achenie, L.E.K., Venkatasubramanian, V., Chapter 16: Challenges and opportunities for CAMD Authors of Document, Computer Aided Chemical Engineering, 2002; 12 (C): 357-377

Satyanarayana K.C., Abildskov J., Gani R., Computer-aided polymer design using group contribution plus property models, Computers and Chemical Engineering, 2009; 33 (5): 1004-1013

Sinha M. , Achenie L.E.K., Gani R. ,Blanket wash solvent blend design using interval analysis, Industrial and Engineering Chemistry Research, 2003; 42 (3): 516-527

Roughton B.C., Christian B., White J., Camarda K.V., Gani R., Simultaneous design of ionic liquid entrainers and energy efficient azeotropic separation processes, Computers and Chemical Engineering, 2012; 42: 248-262

Karunanithi A.T., Achenie L.E.K. , Gani R., Optimal (Solvent) mixture design through a decomposition based CAMD methodology, Computer Aided Chemical Engineering; 2004, 18 (C): 217-222

Karunanithi A.T., Achenie L.E.K. Gani R., À new decomposition-based computer-aided molecular/mixture design methodology for the design of optimal solvents and solvent mixtures, Industrial and Engineering Chemistry Research, 2005; 44 (13): 4785-4797

Gani R. , Jiménez-González C., Constable D.J.C. , Method for selection of solvents for promotion of organic reactions, Computers and Chemical Engineering, 2005; 29 (7): 1661-1676

Kossack S., Kraemer K., Gani R. Marquardt W., A systematic synthesis framework for extractive distillation processes, Chemical Engineering Research and Design, 2008; 86 (7): 781-792

Modarresi H., Conte, E., Abildskov J., Gani R., Crafts P., Model-based calculation of solid solubility for solvent selection - A review, Industrial and Engineering Chemistry Research, 2008; 47 (15): 5234-5242

Gani R. , Gómez P.A., Folić M., Jiménez-González C., Constable D.J.C. , Solvents in organic synthesis: Replacement and multi-step reaction systems, Computers and Chemical Engineering, 2008; 32 (10): 2420-2444

Karunanithi A.T. , Mehrkesh A. , Computer-aided design of tailor-made ionic liquids, AIChE Journal, 2013; 59 (12): 4627-4640

Kim K.J., Diwekar U.M., Efficient combinatorial optimization under uncertainty. 2. Application to stochastic solvent selection, Industrial and Engineering Chemistry Research, 2002; 41 (5): 1276-1296

Sahinidis N.V., Tawarmalani M., Yu M., Design of alternative refrigerants via global optimization, AIChE Journal, 2003; 49 (7): 1761-1775

Samudra A., Sahinidis N.V., Design of heat-transfer media components for retail food refrigeration, Industrial and Engineering Chemistry Research, 2013; 52 (25): 8518-8526

Karunanithi A.T., Acquah C., Achenie L.E.K., Tuning the Morphology of Pharmaceutical Compounds via Model Based Solvent Selection, Chinese Journal of Chemical Engineering, 2008; 16 (3): 465-473

Song H. , Song J, The application of computer-aided molecular design in selecting solvents for extractive distillation, Progress in Chemistry, 2008; 18 (9): 1188-1193

Giovanoglou A., Barlatier J., Adjiman C.S., Pistikopoulos E.N., Cordiner J.L., Optimal Solvent Design for Batch Separation Based on Economic Performance, AIChE Journal, 2003; 49 (12): 3095-3109

Marcoulaki E.C., Kokossis A.C., On the development of novel chemicals using a systematic synthesis approach. Part I. Optimisation framework, Chemical Engineering Science , 2000; 55 (13): 2529-2546

Marcoulaki E.C., Kokossis A.C., On the development of novel chemicals using a systematic optimisation approach. Part II. Solvent design, Chemical Engineering Science, 2000; 55 (13): 2547-2561

Papadopoulos A.I., Stijepovic M., Linke P., On the systematic design and selection of optimal working fluids for Organic Rankine Cycles, Applied Thermal Engineering, 2010; 30 (6-7): 760-769

Papadopoulos A.I., Linke P., On the synthesis and optimization of liquid-liquid extraction processes using stochastic search methods, Computers and Chemical Engineering, 2004; 28 (11): 2391-2406

Papadopoulos A.I., Linke P., On the integrated design of solvents and processes using a decomposition based approach, Computer Aided Chemical Engineering, 2004; 18 (C): 259-264

Papadopoulos A.I., Linke P, A unified framework for integrated process and molecular design, Chemical Engineering Research and Design, 2005; 83 (6 A): 674-678

Marcoulaki E.C., Kokossis A.C., Batzias F.A., Novel chemicals for clean and efficient processes using stochastic optimization, Computers and Chemical Engineering, 2000; 24 (2-7): 705-710

Marcoulaki E.C. , Batzias F.A., Extractant design for enhanced biofuel production through fermentation of cellulosic wastes, Computer Aided Chemical Engineering, 2003; 14 (C): 1121-1126

Papadopoulos A.I., Linke P., Integrated design of optimal processes and molecules: A framework for solvent-based separation and reactive-separation systems, Computer Aided Chemical Engineering, 2005; 20 (C): 1645-1650

Papadopoulos A.I., Stijepovic M., Linke P., Seferlis P., Voutetakis S., Molecular design of working fluid mixtures for organic rankine cycles, Computer Aided Chemical Engineering, 2013; 32: 289-294

Papadopoulos A.I., Linke P., Efficient integration of optimal solvent and process design using molecular clustering, Chemical Engineering Science, 2006; 61 (19): 6316-6336

Κεφάλαιο 7

Marcoulaki E.C., Batzias F.A. , Extractant design for enhanced biofuel production through fermentation of cellulosic wastes, Computer Aided Chemical Engineering 2003; 14 (C): 1121-1126

Marcoulaki E.C., Kokossis A.C. , Molecular design synthesis using stochastic optimisation as a tool for scoping and screening Computers and Chemical Engineering 1998; 22 (SUPPL. 1): S11-S18

Marcoulaki E.C., Kokossis A.C., Batzias F.A., Computer - Aided synthesis of molecular mixtures and process streams, Computer Aided Chemical Engineering 2001; 9 (C): 451-456

Mao M. A., Scelza R., Scotti R., Gianfreda L, Role of Enzymes in the Remediation of polluted Environments, Journal of Soil Science and Plant Nutrition 2010; 10 (3): 333-353

Παραρτήματα

Παράρτημα Α

Το πρόβλημα των πολλαπλών αποσυνθέσεων

Στο κεφάλαιο αυτό επιχειρείται να αντιμετωπιστεί το πρόβλημα των πολλαπλών αποδομήσεων που εντοπίζεται στις μεθόδους GC. Όπως έχει γίνει ήδη γνωστό, όταν είναι επιθυμητό να προβλεφθούν για ένα μόριο ιδιότητες με χρήση GC, πρέπει να αποδομηθεί πρώτα στα επιμέρους groups που το απαρτίζουν. Να αναφερθεί ασφαλώς και πάλι, πως η κάθε μέθοδος χρησιμοποιεί διαφορετικές δομικές μονάδες, που καθορίζονται από τη φύση της προβλεπόμενης ιδιότητας και από τη σύνδεση της με τους περιγραφείς της κάθε μεθόδου. Παρλαυτά, σε αρκετές περιπτώσεις μορίων και σε συγκεκριμένα μοντέλα GC, υπάρχουν περισσότεροι του ενός τρόπου για να περιγραφεί μία ένωση. Αυτό σημαίνει, πως είναι δυνατόν να βρεθούν περισσότεροι του ενός συνδυασμοί ομάδων για να αποδομηθεί ένα μόριο (decomposition). Διαφορετικοί συνδυασμοί, όμως, σημαίνει και διαφορετικές συνεισφορές κάθε φορά της κάθε ομάδας στο μόριο και συνεπώς, διαφορετικά τελικά αποτελέσματα. Τίθεται, λοιπόν, το ερώτημα για το αν απέχουν αρκετά τα αποτελέσματα για την κάθε αποδόμηση και αν ναι, ποιο είναι αυτό που εξασφαλίζει αποτελέσματα με το μικρότερο σφάλμα. Μελετάται, τελικά, το μοτίβο που πρέπει να ακολουθεί η αποδόμηση μιας ένωσης για να βρεθεί για να είναι το πιο ακριβές. Προς αυτήν την κατεύθυνση εντοπίζονται δύο βασικές κατηγορίες, στις οποίες εντοπίζεται το παραπάνω πρόβλημα: οι αιθέρες και οι αμίνες. Αυτές οι δύο κατηγορίες χρησιμοποιούνται και γίνονται δοκιμές και συγκρίσεις με τις πειραματικές τους τιμές ώστε να βρεθεί ποια αποδόμηση είναι η βέλτιστο.

Χρησιμοποιούνται αιθέρες και αμίνες αυξανόμενης ανθρακικής αλυσίδας. Ευρίσκονται οι πειραματικές τιμές ιδιοτήτων, για τις οποίες υπάρχουν πειραματικές τιμές διαθέσιμες στο κοινό και με χρήση των GC μοντέλων των Gani και Constantinou (1994), υπολογίζονται οι τιμές των παραπάνω ιδιοτήτων. Εκτός από τις θερμοδυναμικές ιδιότητες, όμως, υπολογίζονται και οι συντελεστές ενεργότητας σε άπειρη αραίωση της ένωσης στο νερό, αλλά και του νερού στην ένωση, με δύο μεθόδους: NRTL και UNIFAC. Επίσης, για λόγους αναφοράς ευρίσκεται και από τη UNIFAC, ο συντελεστής ενεργότητας σε άπειρη αραίωση για το εκάστοτε μόριο (διαλύτη) στο εξάνιο (οργανικός διαλύτης). Έπειτα ευρίσκεται η διαφορά της εκτίμησης που λαμβάνεται από το ένα συνδυασμό από ομάδες με την εκτίμηση από το δεύτερο συνδυασμό. Το σχετικό σφάλμα από την πειραματική βρίσκεται:

$$\%\, error = \frac{\pi\varepsilon\iota\rho\alpha\mu\alpha\tau\iota\kappa\acute{\eta}\ \tau\iota\mu\acute{\eta} - \upsilon\pi\sigma\lambda\sigma\gamma\iota\sigma\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta}}{\upsilon\pi\sigma\lambda\sigma\gamma\iota\sigma\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta}}$$

,ενώ η διαφορά μεταξύ των παραπάνω δύο εκτιμήσεων βρίσκεται από:

$$difference = \frac{(\upsilon\pi o\lambda o\gamma\iota\sigma\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta}\ 1 - \upsilon\pi o\lambda o\gamma\iota\sigma\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta}\ 2)}{\frac{(\upsilon\pi o\lambda o\gamma\iota\sigma\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta}\ 1 + \upsilon\pi o\lambda o\gamma\iota\sigma\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta}\ 2)}{2}}$$

Οι αιθέρες που χρησιμοποιήθηκαν ήταν οι: αιθυλομεθυλαιθέρας (methyl ethyl ether), μεθυλοπροπυλαιθέρας (methyl propyl ether), ισοπροπυλομεθυλαιθέρας (isopropyl methyl ether), βουτυλομεθυλαιθέρας (butyl methyl ether), ισοβουτυλ-μεθυλαιθέρας (isobutyl methyl ether), μεθυλ-πεντυλαιθέρας (methyl pentyl ether), ισοπροπυλο-ισοβουτυλαιθέρας (isopropyl isobutyl ether), triglyme. Οι αμίνες που χρησιμοποιήθηκαν είναι οι διμεθυλο-αιθυλαμίνη (dimethylethylamine), διαιθυλο-μεθυλαμίνη (diethylmethylamine), διμεθυλο-βουτυλαμίνη (dimethylbutylamine). Παρακάτω δίνονται οι πίνακες με τα αποτελέσματα που προέκυψαν:

Πίνακας Α.1. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για μεθυλο-αιθυλαιθέρας

SMILES ένωσης		**COCC**	
Χημικός τύπος		CH3OCH2CH3	
Όνομα ένωσης		methyl ethyl ether	
Tm (K)			160
Tb (K)			281
Tc (K)			438
Pc (bar)			-
ΔHv (kJ/mol)			-
γ^{∞}του διαλύτη στο νερό/UNIQUAC			2,953
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,918
αποδόμηση		**[2xCH3,1xCH2O]**	**[1xCH3O,1xCH3,1xCH2]**
Tm(K)		114	150
Tb(K)		250	287
Tc(K)		406	445
Pc(bar)		-	-
ΔHv(kJ/mol)		-	-
γ^{∞} του διαλύτη στο νερό UNIFAC		49	25
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1,272	1,684
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		30%	7%
Tb(K) (%) error		11%	-2%
Tc(K) (%) error		7%	-2%
Pc(bar) (%) error		-	-
ΔHv(kJ/mol) (%) error		-	-
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		-1567%	-738%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-39%	-83%
Differences in GC (%)			
Tm(K) (%) diff		-7%	
Tb(K) (%) diff		-3%	
Tc(K) (%) diff		-2%	
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff			
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		17%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		-7%	
Τοξικότητα LC 50(mol/L)		0,0572	

Πίνακας Α.2. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για ισοπροπυλο-μεθυλαιθέρας

SMILES ένωσης		**COCCC**	
Χημικός τύπος		CH3OCH2CH2CH3	
Όνομα ένωσης		methyl propyl ether	
Tm(K)			134
Tb(K)			312
Tc(K)			476
Pc(bar)			-
ΔHv(kJ/mol)			-
γ^{∞}του διαλύτη στο νερό/UNIQUAC			527
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,974
γ^{∞} του διαλύτη στο νερό /NRTL			474
γ^{∞} του διαλύτη στο /NRTL			-
αποδόμηση		**[2xCH3,1xCH2,1xCH2O]**	**[1xCH3O,2xCH2,1xCH3]**
Tm(K)		140	170
Tb(K)		299	328
Tc(K)		463	492
Pc(bar)		-	-
ΔHv(kJ/mol)		-	-
γ^{∞} του διαλύτη στο νερό UNIFAC		162	78
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1,281	1,626
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		-4%	-27%
Tb(K) (%) error		4%	-5%
Tc(K) (%) error		3%	-3%
Pc(bar) (%) error		-	-
ΔHv(kJ/mol) (%) error		-	-
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		69%	85%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-32%	-67%
Differences in GC (%)			
Tm(K) (%) diff		-5%	
Tb(K) (%) diff		-2%	
Tc(K) (%) diff		-2%	
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff			
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		17%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		-6%	
Τοξικότητα LC 50(mol/L)		0,0292	

Πίνακας Α.3. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για ισοπροπυλο-μεθυλαιθέρας

SMILES ένωσης		**COC(C)C**	
Χημικός τύπος		CH3OCH(CH3)2	
Όνομα ένωσης		isopropyl methyl ether	
Tm(K)			-
Tb(K)			323
Tc(K)			-
Pc(bar)			-
ΔHv(kJ/mol)			26
γ^{∞}του διαλύτη στο νερό/UNIQUAC			269
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,973
γ^{∞} του διαλύτη στο νερό /NRTL			267
γ^{∞} του διαλύτη στο /NRTL			-
αποδόμηση		**[1xCH3O,1xCH,2xCH3]**	**[3xCH3,1xCH-O]**
Tm(K)		-	-
Tb(K)		313	274
Tc(K)		-	-
Pc(bar)		-	-
ΔHv(kJ/mol)		29	25
γ^{∞} του διαλύτη στο νερό UNIFAC		78	365
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1,626	1,079
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		-	-
Tb(K) (%) error		3%	15%
Tc(K) (%) error		-	-
Pc(bar) (%) error		-	-
ΔHv(kJ/mol) (%) error		-10%	4%
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		71%	-36%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-67%	-11%
Differences in GC (%)			
Tm(K) (%) diff			
Tb(K) (%) diff		3%	
Tc(K) (%) diff			
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff		4%	
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		-32%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		10%	
Τοξικότητα LC 50(mol/L)		0,0109	

Πίνακας Α.4. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για μεθυλο-βουτυλαιθέρας

SMILES ένωσης		**COCCCC**	
Χημικός τύπος		CH3O(CH2)3CH3	
Όνομα ένωσης		methyl butyl ether	
Tm(K)			158
Tb(K)			344
Tc(K)			-
Pc(bar)			-
ΔHv(kJ/mol)			30
γ^{∞}του διαλύτη στο νερό/UNIQUAC			296
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,998
γ^{∞} του διαλύτη στο νερό /NRTL			4012
γ^{∞} του διαλύτη στο /NRTL			-
αποδόμηση		**[1xCH3O,1xCH3,3xCH2]**	**[2xCH3,1xCH2O,2xCH2]**
Tm(K)		186	162
Tb(K)		363	339
Tc(K)		-	-
Pc(bar)		-	-
ΔHv(kJ/mol)		36	32
γ^{∞} του διαλύτη στο νερό UNIFAC		245	523
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1,557	1,265
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		-18%	-2%
Tb(K) (%) error		-6%	1%
Tc(K) (%) error		-	-
Pc(bar) (%) error		-	-
ΔHv(kJ/mol) (%) error		-20%	-6%
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		17%	-77%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-56%	-27%
Differences in GC (%)			
Tm(K) (%) diff		4%	
Tb(K) (%) diff		2%	
Tc(K) (%) diff			
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff		3%	
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		-18%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		5%	
Τοξικότητα LC 50(mol/L)		0,0148	

Πίνακας Α.5. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για ισοβουτυλο-μεθυλαιθέρας

SMILES ένωσης		**COCC(C)C**	
Χημικός τύπος		CH3OCH2CH(CH3)2	
Όνομα ένωσης		isobutyl methyl ether	
Tm(K)			-
Tb(K)			332
Tc(K)			497
Pc(bar)			34
ΔHv(kJ/mol)			-
γ^{∞}του διαλύτη στο νερό/UNIQUAC			219
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,998
γ^{∞} του διαλύτη στο νερό /NRTL			1269
γ^{∞} του διαλύτη στο /NRTL			-
αποδόμηση		**[1xCH3O,1xCH2,1xCH,2xCH3]**	**[3xCH3,1xCH2O,1xCH]**
Tm(K)		-	-
Tb(K)		351	325
Tc(K)		517	492
Pc(bar)		35	34
ΔHv(kJ/mol)		-	-
γ^{∞} του διαλύτη στο νερό UNIFAC		245	524
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1,557	1,265
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		-	-
Tb(K) (%) error		-6%	2%
Tc(K) (%) error		-4%	1%
Pc(bar) (%) error		-3%	2%
ΔHv(kJ/mol) (%) error		-	-
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		-12%	-140%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-56%	-27%
Differences in GC (%)			
Tm(K) (%) diff			
Tb(K) (%) diff		2%	
Tc(K) (%) diff		1%	
Pc(bar) (%) diff		1%	
ΔHv(kJ/mol) (%) diff			
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		-18%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		5%	
Τοξικότητα LC 50(mol/L)		0,0056	

Πίνακας Α.6. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για μεθυλο-πεντυλαιθέρας

SMILES ένωσης		**COCCCCC**	
Χημικός τύπος		CH3O(CH2)4CH3	
Όνομα ένωσης		methyl pentyl ether	
Tm(K)			-
Tb(K)			372
Tc(K)			-
Pc(bar)			-
ΔHv(kJ/mol)			-
γ^{∞}του διαλύτη στο νερό/UNIQUAC			4
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,999
γ^{∞} του διαλύτη στο νερό /NRTL			-
γ^{∞} του διαλύτη στο /NRTL			-
αποδόμηση		**[1xCH3O,4xCH2,1xCH3]**	**[2xCH3,1xCH2O,3xCH2]**
Tm(K)		-	-
Tb(K)		393	372
Tc(K)		-	-
Pc(bar)		-	-
ΔHv(kJ/mol)		-	-
γ^{∞} του διαλύτη στο νερό UNIFAC		760	1665
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1	1
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		-	-
Tb(K) (%) error		-6%	0,064%
Tc(K) (%) error		-	-
Pc(bar) (%) error		-	-
ΔHv(kJ/mol) (%) error		-	-
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		-16999%	-37348%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-48%	-23%
Differences in GC (%)			
Tm(K) (%) diff			
Tb(K) (%) diff		1%	
Tc(K) (%) diff			
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff			
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		-19%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		5%	
Τοξικότητα LC 50(mol/L)		0,00759	

Πίνακας Α.7. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για ισοβουτυλ-ισοπροπυλαιθέρας

SMILES ένωσης		**CC(C)OCC(C)C**	
Χημικός τύπος		(CH3)2CHOCH2CH(CH3)2	
Όνομα ένωσης		isobutyl isopropyl ether	
Tm(K)			-
Tb(K)			372
Tc(K)			-
Pc(bar)			-
ΔHv(kJ/mol)			32
γ^{∞}του διαλύτη στο νερό/UNIQUAC			4,914
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,982
γ^{∞} του διαλύτη στο νερό /NRTL			-
γ^{∞} του διαλύτη στο /NRTL			-
αποδόμηση		**[4xCH3,1xCH-O,1xCH,1xCH2]**	**[1xCH2O,2xCH,4xCH3]**
Tm(K)		-	-
Tb(K)		374	379
Tc(K)		-	-
Pc(bar)		-	-
ΔHv(kJ/mol)		36	36
γ^{∞} του διαλύτη στο νερό UNIFAC		12450	5250
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1,052	1,183
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		-	-
Tb(K) (%) error		-0,480%	-2%
Tc(K) (%) error		-	-
Pc(bar) (%) error		-	-
ΔHv(kJ/mol) (%) error		-12%	-12%
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		-253223%	-106715%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-7%	-20%
Differences in GC (%)			
Tm(K) (%) diff			
Tb(K) (%) diff		-0,322%	
Tc(K) (%) diff			
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff		0,075%	
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		20%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		-3%	
Τοξικότητα LC 50(mol/L)		0,000547	

Πίνακας Α.8. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για τριγλύμιο

SMILES ένωσης		**COCCOCCOCCOC**	
Χημικός τύπος		CH3O(CH2)2O(CH2)2O(CH2)2OC	
Όνομα ένωσης		triglyme	
Tm(K)			228
Tb(K)			489
Tc(K)			-
Pc(bar)			-
ΔHv(kJ/mol)			-
γ^{∞}του διαλύτη στο νερό/UNIQUAC			4,254
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,921
γ^{∞} του διαλύτη στο νερό /NRTL			-
γ^{∞} του διαλύτη στο /NRTL			-
αποδόμηση		**[4xCH2O,2xCH2,2xCH3]**	**[2xCH3O,2xCH2O,4xCH2]**
Tm(K)		246	268
Tb(K)		473	498
Tc(K)		-	-
Pc(bar)		-	-
ΔHv(kJ/mol)		-	-
γ^{∞} του διαλύτη στο νερό UNIFAC		207	66
γ^{∞} του διαλύτη στο C6H12/UNIFAC		7	15
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		-8%	-17%
Tb(K) (%) error		3%	-2%
Tc(K) (%) error		-	-
Pc(bar) (%) error		-	-
ΔHv(kJ/mol) (%) error		-	-
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		-4756%	-1446%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-652%	-1558%
Differences in GC (%)			
Tm(K) (%) diff		-2%	
Tb(K) (%) diff		-1%	
Tc(K) (%) diff			
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff			
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		26%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		-19%	
Τοξικότητα LC 50(mol/L)		0,0332	

Πίνακας Α.9. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για διμεθυλο-αιθυλαμίνη

SMILES ένωσης		**CCN(C)C**	
Χημικός τύπος		CH3CH2N(CH3)2	
Όνομα ένωσης		name	dimethylethylamine
Tm(K)			133
Tb(K)			310
Tc(K)			
Pc(bar)			
ΔHv(kJ/mol)			
γ^{∞}του διαλύτη στο νερό/UNIQUAC			3,238
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,993
γ^{∞} του διαλύτη στο νερό /NRTL			-
γ^{∞} του διαλύτη στο /NRTL			-
αποδόμηση		**[3xCH3,1xCH2N]**	**[1xCH3N,2xCH3,1xCH2]**
Tm(K)		87	182
Tb(K)		277	323
Tc(K)			
Pc(bar)			
ΔHv(kJ/mol)			
γ^{∞} του διαλύτη στο νερό UNIFAC		19	1,839
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1,254	1,133
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		8%	-37%
Tb(K) (%) error		-8%	-4%
Tc(K) (%) error			
Pc(bar) (%) error			
ΔHv(kJ/mol) (%) error			
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		69%	43%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-26%	-14%
Differences in GC (%)			
Tm(K) (%) diff		-10%	
Tb(K) (%) diff		0,738%	
Tc(K) (%) diff			
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff			
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		-15%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		3%	

Πίνακας Α.10. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για διαιθυλο-μεθυλαμίνη

SMILES ένωσης		**CN(CC)CC**	
Χημικός τύπος		CH3N(CH2CH3)2	
Όνομα ένωσης		name	diethylmethylamine
Tm(K)			77
Tb(K)			337
Tc(K)			
Pc(bar)			
ΔHv(kJ/mol)			30
γ^{∞}του διαλύτη στο νερό/UNIQUAC			5
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,999
γ^{∞} του διαλύτη στο νερό /NRTL			-
γ^{∞} του διαλύτη στο /NRTL			-
αποδόμηση		**[1xCH2N,3xCH3,1xCH2]**	**[1xCH3N,2xCH3,2xCH2]**
Tm(K)		121	197
Tb(K)		321	359
Tc(K)			
Pc(bar)			
ΔHv(kJ/mol)		31	31
γ^{∞} του διαλύτη στο νερό UNIFAC		62	6,017
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1,061	1,133
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		-57%	-155%
Tb(K) (%) error		5%	-6%
Tc(K) (%) error			
Pc(bar) (%) error			
ΔHv(kJ/mol) (%) error		-2%	-4%
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		-1173%	-23%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-6%	-13%
Differences in GC (%)			
Tm(K) (%) diff		-12%	
Tb(K) (%) diff		-3%	
Tc(K) (%) diff			
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff		-0,454%	
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		41%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		-2%	

Πίνακας Α.11. Σύγκριση πειραματικών-υπολογισμένων θερμοδυναμικών ιδιοτήτων για διμεθυλο-βουτυλαμίνη

SMILES ένωσης		**CCCCN(C)C**	
Χημικός τύπος		CH3(CH2)3N(CH3)2	
Όνομα ένωσης		name	dimethylbutylamine
Tm(K)			282
Tb(K)			368
Tc(K)			
Pc(bar)			
ΔHv(kJ/mol)			
γ^{∞}του διαλύτη στο νερό/UNIQUAC			5,408
γ^{∞}του διαλύτη στο C6H12/UNIQUAC			0,995
$\gamma^{\infty}$of solvent in Water/NRTL			-
$\gamma^{\infty}$of solvent in C6H12/NRTL			-
αποδόμηση		**[1xCH3N,2xCH3,3xCH2]**	**[1xCH2N,3xCH3,2xCH2]**
Tm(K)		210	147
Tb(K)		389	357
Tc(K)			
Pc(bar)			
ΔHv(kJ/mol)			
γ^{∞} του διαλύτη στο νερό UNIFAC		19	201
γ^{∞} του διαλύτη στο C6H12/UNIFAC		1,113	1,049
Πειρ. Σφάλματα-GC (%)			
Tm(K) (%) error		26%	48%
Tb(K) (%) error		-6%	3%
Tc(K) (%) error			
Pc(bar) (%) error			
ΔHv(kJ/mol) (%) error			
γ^{∞} του διαλύτη στο νερό UNIQUAC (%) error		-257%	-3613%
γ^{∞} του διαλύτη στο C6H12/UNIQUAC (%) error		-12%	-5%
Differences in GC (%)			
Tm(K) (%) diff		9%	
Tb(K) (%) diff		2%	
Tc(K) (%) diff			
Pc(bar) (%) diff			
ΔHv(kJ/mol) (%) diff			
γ^{∞} του διαλύτη στο νερό /UNIFAC (%) diff		-41%	
γ^{∞} του διαλύτη C6H12/UNIFAC (%) diff		1%	

Από τους πίνακες Α.1-Α.11 προκύπτει πως όσον αφορά στις θερμοφυσικές ιδιότητες των καθαρών ουσιών οι διαφορές μεταξύ των δύο εκτιμήσεων είναι παρόμοιες και δεν ξεπερνούν κατά πολύ το 10%. Αυτό το νούμερο κρίνεται ως σχετικά ικανοποιητικό δεδομένης της φύσης των μοντέλων (έλλειψης ενσωμάτωσης δομικής πληροφορίας στο μόριο). Όσον αφορά το συντελεστή ενεργότητας σε άπειρη αραίωση, εντοπίζονται αρκετά μεγάλες διαφορές μεταξύ των δύο εκτιμήσεων, που κυμαίνονται περί το 20% και μπορεί να ανέλθουν και μέχρι το 40%. Παρατηρούνται, δε, μεγαλύτερα σφάλματα στον υπολογισμό του συντελεστή για άπειρη αραίωση της ένωση στο νερό, έναντι στους συντελεστές για την ένωση στο εξάνιο (σπανίως η διαφορά να ξεπεράσει το 10%). Αυτά τα σφάλματα οφείλονται στις αποδομήσεις, όπου για τους αιθέρες χρησιμοποιείται η ομάδα >CH_2O και >CHO-, έναντι της –CH_3O και για τις αμίνες, όπου χρησιμοποιείται >CH_2N, έναντι της –CH_3N. Στις παραπάνω περιπτώσεις η τιμή του συντελεστή ενεργότητας στο νερό είναι κατά υπερτιμημένη σε αρκετά μεγάλο βαθμό.

Όσον αφορά στα σφάλματα από την πειραματική τιμή, στις ιδιότητες καθαρών ουσιών, φαίνεται πως αντίθετα με τις προβλέψεις για το συντελεστή ενεργότητας, οι αποδομήσεις που περιλαμβάνουν –CH_3O , έναντι των >CH_2O και >CHO- για τους αιθέρες και τις αποδομήσεις που περιλαμβάνουν –CH_3N, έναντι του >CH_2N για τις αμίνες, φαίνεται να υπολογίζουν τα σημεία τήξης με μεγαλύτερο σφάλμα από την πειραματική τιμή. Οι διαφορές στα σφάλματα μπορεί να αγγίζουν μέχρι και το 100%, ενώ κυμαίνονται για τις περισσότερες περιπτώσεις γύρω από το 20%. Όσον αφορά στις υπόλοιπες ιδιότητες καθαρών ουσιών, οι δύο αποδομήσεις έχουν εκτιμήσεις που είναι πολύ κοντά σε σχέση με τα πειραματικά. Οι αποκλίσεις από τις πειραματικές τιμές των συντελεστών ενεργότητας είναι αρκετά μεγάλες. Όπως είναι λογικό, αυτό συμβαίνει σε όλες τις παραπάνω περιπτώσεις, όπου δίνονται υπερτιμημένες τιμές στο συντελεστή ενεργότητας στο νερό. Τέλος, υπολογίζεται και η τοξικότητα LC 50, για τους αιθέρες με την μέθοδο του Gao et al. (1992). Η μέθοδος αυτή δεν επιτρέπει τον υπολογισμό της τοξικότητας αμινών, λόγω έλλειψης κατάλληλων ομάδων. Οι εκτιμήσεις για την τοξικότητα συμπίπτουν σε όλες τις αποδομήσεις. Πρέπει να σημειωθεί πως σε καμία περίπτωση δεν αποδεικνύουν τα παραπάνω δεδομένα την ύπαρξη τάσεων για την αποσύνθεση σε κατάλληλες αποδομήσεις, καθώς τα δεδομένα δεν επαρκούν για την εξαγωγή κατάλληλων συμπερασμάτων, παρά εμφανίζουν κάποιες πρώτες εντυπώσεις στο παραπάνω θέμα. Εννοείται πως σε καμία περίπτωση, όπου είναι δυνατόν να γίνουν πολλαπλές αποδομήσεις, δεν πρέπει να χωρίζεται μια χαρακτηριστική ομάδα σε δύο διαφορετικά groups π.χ. στο μεθυλο-αιθυλαιθέρα (CH_3COOCH_3), όπου υπάρχει το [1x CH_3, 1xCH_3COO] και το [1xCH_3CO, 1xCH_3O]. Σαφώς και η δεύτερη αποσύνθεση είναι λάθος. Επίσης, στην περίπτωση του οξικού οξέος (CH_3COOH), όπου υπάρχει [1xCH_3CO, 1xOH] και [1xCH_3, 1xCOOH], σαφώς και θα προτιμηθεί η δεύτερη περίπτωση. Να σημειωθεί πως στους υπολογισμούς αυτούς χρησιμοποιήθηκαν μόνο πρώτης τάξης ομάδες. Για τις συγκεκριμένες μεθόδους είναι αναμενόμενο να μειωθεί το σφάλμα των εκτιμήσεων αν στους υπολογισμούς συμπεριληφθούν δεύτερης τάξης ομάδες.

Να σημειωθεί πως η διαφορές στις εκτιμήσεις των συντελεστών ενεργότητας είναι δικαιολογημένες μιας και στις περισσότερες περιπτώσεις, τα δύο μόρια που συμμετέχουν στο μίγμα (νερό και αιθέρας ή

αμίνη) έχουν αρκετά μεγάλες διαφορές στο μέγεθος και σχήμα, καθώς είναι γνωστό πως η UNIFAC σε αυτές τις περιπτώσεις δίνει αναξιόπιστα αποτελέσματα. Αν, όμως, εξαιρεθεί το παραπάνω οι εκτιμήσεις για τις ιδιότητες των καθαρών ουσιών δε διαφέρουν περισσότερο από το αναμενόμενο σφάλμα των μοντέλων GC. Γενικά, λοιπόν, όσο δεν διασπάται μια χαρακτηριστική ομάδα του μορίου σε περισσότερες από μια ομάδες δεν υπάρχει σημαντική διαφορά ανάμεσα στις επιμέρους αποσυνθέσεις.

Παράρτημα Β. Ομάδες που συμμετέχουν στο εργαλείο CAMD και στη ανάπτυξη των μοντέλων

Οι ομάδες που συμμετέχουν στο εργαλείο CAMD των Marcoulaki και Kokossis (2000) φαίνονται στον πίνακα Β.1. Με αστερίσκο σημειώνονται οι ομάδες παρέμειναν μετά την απόρριψη όσων δε συμμετείχαν στο σύνολο δεδομένων:

Πίνακας Β.1. Ομάδες που συμμετέχουν στην αρχική μεθοδολογία CAMD και ομάδες που απέμειναν μετά την αφαίρεση

Ομάδες								
$-CH_3$*	$CH_3(C=O)$-*	CH_3-NH_2*	-COOH*	$-CH_2-NO_2$	$-C(Cl)_2(F)$	ACCH<*	AC-CH_2-O-	AC->CHCl
$>CH_2$*	$-CH_2(C=O)$-*	$-CH_2-NH_2$*	$-CH_2-Cl$*	$>CH-NO_2$	$-C(Cl)(F)_2$*	ACC<-*	AC->CH-O-	AC-->CCl
>CH-*	H(C=O)-*	$>CH-NH_2$*	>CH-Cl*	-I	CH_3-Cl*	AC--$CH=CH_2$*	AC--CH_2NH_2	AC--$CH(Cl)_2$*
>C<*	CH_3COO-*	CH_3-NH-*	->C-Cl	-Br	C_5H_4O	AC--CH=CH-	AC->$CH-NH_2$	AC--$C(Cl)_3$
$CH_2=CH$-*	$-CH_2COO$-*	$-CH_2-NH$-*	$CH_2-(Cl)_2$*	CH#C-*	C_5H_5N*	AC->C=CH_2	AC--CH_2-NH-	AC--CH_2NO
-CH=CH-*	HCOO-*	>CH-NH-	$-CH-(Cl)_2$*	-C#C-*	$-C_5H_4N$	AC->C=CH-	AC->CH-NH-	AC->$CHNO_2$
$CH_2=C<$*	CH_3-O-*	$CH_3-N<$*	$CH-(Cl)_3$*	$-C(F)_3$*	$>C_5H_3N$	AC->C=C<	AC--CH_2-N<	AC--I
-CH=C<*	$-CH_2-O$-*	$-CH_2-N<$*	$-C-(Cl)_3$*	CH_4	ACH*	ACOH*	AC--CH_2CN	AC--Br
>C=C<*	>CH-O-*	CH_3-CN*	$C-(Cl)_4$*	$CH_2=O$*	$ACCH_3$*	AC-CH_2CO	AC--COOH*	AC--C#CH
-OH*	F-CH_2-O-	$-CH_2-CN$*	CH_3-NO_2	$CH_2=C=CH$-	$ACCH_2$-*	AC-CH_2COO	AC--CH_2Cl*	AC--C#C-
AC--$C(F)_3$*	AC--$CH=C=CH_2$		AC-$C(Cl)_3$	AC-NO_2*	AC-NH_2*	AC-Cl	AC-$CCl(F)_2$	

Παράρτημα Γ. Μόρια που συμμετέχουν στην ανάπτυξη των μοντέλων

Τα παρακάτω μόρια και οι τιμές τους για τους τρεις μελετώμενους δείκτες συμμετείχαν στο σύνολο δεδομένων για το σχηματισμό των μοντέλων από τη βάση του Ecoinvent:

Πίνακας Γ.1. Μόρια που συμμετέχουν στην ανάπτυξη μοντέλων

Μόρια	
cyclohexanone, at plant	trichloromethane, at plant
dichloromethane, at plant	trichloropropane, from hypochlorination of allyl chloride, at plant
diethanolamine, at plant	trichloropropane, from hypochlorination of allyl chloride, at plant
diethyl ether, at plant	trichloropropane, from hypochlorination of allyl chloride, at plant
diethylene glycol, at plant	triethanolamine, at plant
dimethyl ether, at plant	triethylene glycol, at plant
dimethylacetamide, at plant	trimethylamine, at plant
dimethylamine, at plant	xylene, at plant
dioxane, at plant	styrene, at plant
dipropylene glycol monomethyl ether, at plant	vinyl acetate, at plant
epichlorohydrin, from hypochlorination of allyl chloride, at plant	butadiene, at plant
esters of versatic acid, at plant	butene, mixed, at plant
ethanol from ethylene, at plant	butene, mixed, at plant
propylene, at plant	tetrahydrofuran, at plant
cyclohexanol, at plant	toluene, liquid, at plant
benzoic-compounds, at regional storehouse	ethyl acetate, at plant
dinitroaniline-compounds, at regional storehouse	ethyl benzene, at plant
pyridine-compounds, at regional storehouse	ethylene dichloride, at plant
methanol, from synthetic gas, at plant	ethylene glycol diethyl ether, at plant
1,1-dimethylcyclopentane, from naphtha, at plant	ethylene glycol dimethyl ether, at plant
1-butanol, propylene hydroformylation, at plant	ethylene glycol monoethyl ether, at plant
1-pentanol, at plant	ethylene glycol, at plant
1-propanol, at plant	ethylene oxide, at plant
2,3-dimethylbutan, from naphtha, at plant	ethylenediamine, at plant
2-butanol, at plant	formaldehyde, production mix, at plant
2-methyl-1-butanol, at plant	formic acid, at plant
2-methyl-2-butanol, at plant	glycerine, from epichlorohydrin, at plant
2-methylpentane, from naphtha, at plant	heptane, at plant
3-methyl-1-butanol, at plant	hexafluorethane, at plant
3-methyl-1-butyl acetate, at plant	hexane, at plant
4-methyl-2-pentanone, at plant	isobutanol, at plant
DTPA, diethylenetriaminepentaacetic acid, at plant	isobutyl acetate, at plant
EDTA, ethylenediaminetetraacetic acid, at plant	isohexane, at plant

Πίνακας Γ.2. Μόρια που συμμετέχουν στην ανάπτυξη μοντέλων

Μόρια	
N,N-dimethylformamide, at plant	isopropanol, at plant
N-methyl-2-pyrrolidone, at plant	isopropyl acetate, at plant
acetaldehyde, at plant	latex, at plant
acetic acid , at plant	methanol, at plant
acetic anhydride, at plant	methyl acetate, at plant
acetone, liquid, at plant	methyl ethyl ketone, at plant
acetonitrile, at plant	methyl formate, at plant
acrylic acid, at plant	methyl tert-butyl ether, at plant
adipic acid, at plant	methyl-3-methoxypropionate, at plant
alkylbenzene, linear, at plant	methylchloride, at plant
allyl chloride, from reacting propylene and chlorine, at plant	methylcyclohexane, at plant
aniline, at plant	methylcyclopentane, from naphtha, at plant
benzal chloride, at plant	monochlorobenzene, at plant
benzene, at plant	monochloropentafluoroethane, at plant
benzyl chloride, at plant	monoethanolamine, at plant
butane-1,4-diol, at plant	nitrobenzene, at plant
butanes from butenes, at plant	dichlorobenzene, at plant
butyl acetate, at plant	penta-erythritol, at plant
butyrolactone	pentane, at plant
carbon tetrachloride, at plant	phenol, at plant
chloroacetic acid, at plant	propanal, at plant
chloromethyl methyl ether, at plant	propylene glycol, liquid, at plant
cumene, at plant	propylene oxide, liquid, at plant
cyclohexane, at plant	

Παράρτημα Δ. Επιλογή στατιστικού δείκτη για την αξιολόγηση μοντέλων

Συντελεστής Προσδιορισμού ή Μέσο σφάλμα :

Είναι μεγίστης σημασίας να σημειωθεί πως για την επικύρωση της αξιοπιστίας του μοντέλου θεωρείται σημαντικός μόνο ο δείκτης Συντελεστής Προσδιορισμού. Οι υπόλοιποι δείκτες είναι ασφαλώς χρήσιμοι για την περαιτέρω ανάλυση των αποτελεσμάτων, αλλά στην παρούσα εργασία από αυτόν τον δείκτη θεωρείται πως εξαρτάται η εγκυρότητα των σχέσεων που αναπτύσσονται. Ο δείκτης αυτός, ως γνωστόν, υποδεικνύει τη συσχέτιση των υπολογισμένων τιμών από ένα μοντέλο με τις πειραματικές τιμές. Αν θέλουμε να το παραστήσουμε διαγραμματικά, θα μπορούσαμε να φανταστούμε δυο κάθετους άξονες: ο

οριζόντιος (άξονας x) περιέχει για τα μόρια τις υπολογισμένες τιμές από το κάθε μοντέλο και ο κάθετος (άξονας y) τις πειραματικές. Το κάθε μόριο που συμμετέχει στο σύνολο δοκιμής και δοκιμάζεται θα αναπαρίσταται από ένα σημείο στο χώρο αυτό, που θα έχει σαν τετμημένη την υπολογισμένη τιμή του GWP/CED/EI 99 και σαν τεταγμένη την πειραματική τιμή. Το ιδανικό είναι κάθε σημείο (μόριο) να βρίσκεται πάνω στην ευθεία y=x. Δηλαδή, η τεταγμένη να είναι ίση με την τετμημένη του σημείου που αναπαρίσταται, άρα η υπολογισμένη τιμή να είναι ίση με την πειραματική. Σε αυτήν την περίπτωση η συσχέτιση που θα κάνει το μοντέλο μεταξύ πειραματικών και υπολογισμένων θα είναι τέλεια και ο Συντελεστής Προσδιορισμού θα λαμβάνει τη μέγιστη τιμή του, δηλαδή θα είναι ίσος με τη μονάδα (η ελάχιστη τιμή είναι 0). Όμως, είναι αδύνατο ένα μοντέλο να προβλέπει ακριβώς τις τιμές όλων των μορίων, γι' αυτό και αναγκαστικά θα είναι όλα τα σημεία (μόρια του συνόλου δοκιμής) διεσπαρμένα γύρω από αυτή την ευθεία. Το πόσο διεσπαρμένα είναι καθορίζει και την ποιότητα του συσχετισμού (correlation) που έχει επιτύχει το μοντέλο: αν είναι αρκετά «απλωμένα» τα σημεία γύρω από την y=x, τόσο πιο πολύ θα απέχει τις περισσότερες φορές η πειραματική από την υπολογισμένη τιμή και τόσο χειρότερη θα είναι η συσχέτιση, που θα έχει αναπτυχθεί, άρα και τόσο χαμηλότερη η τιμή του Συντελεστή Προσδιορισμού. Αντιθέτως, όσο μεγαλύτερος ο δείκτης, τόσο το «νέφος» των σημείων θα είναι πιο πυκνό γύρω από την ευθεία και συνεπώς καλύτερη η συσχέτιση. Είναι αυτονόητο πως, η μέση τιμή των δεικτών στο σύνολο δοκιμής έχει πολύ μεγαλύτερη σημασία από αυτή στο σύνολο εκπαίδευσης, καθώς οι δυνατότητες του μοντέλου αναδεικνύονται στο να προβλέψει ιδιότητες μορίων που δε «γνωρίζει» από την εκπαίδευση του.

Είναι μια αρκετά λογική απορία, γιατί δε χρησιμοποιείται το μέσο σφάλμα (ή οποιαδήποτε άλλη κατηγορία σφάλματος) για να αναδειχτεί η απόδοση ενός μοντέλου. Η απάντηση σε αυτό είναι πως το μέσο σφάλμα, όντως, σα δείκτης απαντά στο βασικότερο ερώτημα που αφορά σε ένα μοντέλο, δηλαδή πόσο αξιόπιστα αποτελέσματα έχει. Ασφαλώς και προτιμούμε να χρησιμοποιήσουμε ένα μοντέλο, του οποίου οι προβλέψεις είναι πιο κοντά στις πειραματικές από ένα άλλο που έχει μεγαλύτερο σφάλμα. Όμως, το σφάλμα, ως δείκτης, δεν ενσωματώνει πληροφορίες που αφορούν την εσωτερική λειτουργία του μοντέλου. Βασίζεται καθαρά στη διαφορά μεταξύ των πειραματικών και υπολογισμένων τιμών και δεν παρέχει καμία στατιστική ένδειξη για την ποιότητα της συσχέτισης που έχει αναπτυχθεί από το μοντέλο. Επίσης, δεν παρέχει με απόλυτη βεβαιότητα στοιχεία στην περίπτωση που είναι επιθυμητή μια νέα πρόβλεψη: στην πρόβλεψη μιας νέας τιμής είναι δυνατόν το σφάλμα να είναι και κατά πολύ μεγαλύτερο από το σφάλμα που συνοδεύει το μοντέλο. Αντίθετα, ο Συντελεστής Προσδιορισμού μας δίνει πληροφορίες για την εσωτερική λειτουργία του μοντέλου και δεν δίνει βάρος, μόνο στη διαφορά μεταξύ πειραματικής και υπολογισμένης τιμής, αλλά στατιστικά πληροφορεί για το είδος και την ποιότητα της συσχέτισης. Ασφαλώς, όταν αυξάνεται η ποιότητα της συσχέτισης (δηλαδή, να αυξάνεται ο Συντελεστής Προσδιορισμού) είναι λογικό να μειώνεται και το σφάλμα, αφού τα σημεία συγκεντρώνονται κατά πολύ περισσότερο γύρω από την ευθεία y=x.

Για να γίνει πιο σαφής η προτίμηση του Συντελεστή Προσδιορισμού, έναντι του σφάλματος αναφέρεται ως παράδειγμα το μοντέλο του «μέσου όρου». Αυτό το μοντέλο είναι το απλούστερο και το χειρότερο από άποψη συσχέτισης. Σε αυτό το μοντέλο, πολύ απλά, εξάγεται ο μέσος όρος των πειραματικών μετρήσεων όλων των μορίων που συμμετέχουν στο σύνολο για το κάθε μοντέλο (GWP,CED,EI 99) και αυτή η τιμή δίνεται σε οποιοδήποτε μόριο είναι επιθυμητό να προβλεφθεί η τιμή του δείκτη του . Το μοντέλο αυτό δεν έχει ικανότητα να δει την είσοδο ενός μορίου. Κοινώς, όποιο μόριο και να του δοθεί ως είσοδος (ανεξαρτήτως διανύσματος εισόδου) το μοντέλο αυτό δίνει πάντα το μέσο όρο. Φαινομενικά, αυτό το μοντέλο είναι αδύνατο να έχει αξιοπιστία, μιας και δε γίνεται να ξεχωρίσει το ένα μόριο από το άλλο. Το σφάλμα, όμως, του παραπάνω μοντέλου είναι μόλις μεταξύ 50-55% και για τους τρεις δείκτες. Λαμβάνοντας υπόψη πως τα μοντέλα GC μπορεί να έχουν 30% σφάλμα, το μοντέλο του «μέσου όρου» δε δίνει πολύ κατά τόσο πολύ χειρότερα αποτελέσματα όσο ήταν αναμενόμενο. Είναι βέβαιο πως για κάποια μόρια θα προβλέπει τιμές με μεγάλη απόκλιση, αλλά σίγουρα και για κάποιες άλλες ενώσεις θα προβλέπει αποτελέσματα με πολύ μικρή απόκλιση. Επικεντρώνοντας, λοιπόν, την προσοχή μας στο σφάλμα βλέπουμε ένα όχι και τόσο χειρότερο μοντέλο από τα GC. Αφού, λοιπόν, το σφάλμα αυτού του μοντέλου δεν απέχει πάρα πολύ από αυτό των GC, είναι επόμενο να τεθεί η ερώτηση, αν τελικά ένας χρήστης είναι λογικό να προτιμήσει αυτό έναντι του οποιουδήποτε GC, κυρίως μάλιστα αφού είναι και κατά πολύ απλούστερο στη χρήση του (δεν περιλαμβάνει κανενός είδους τύπο ή πράξη και αποτελείται αποκλειστικά από έναν αριθμό). Η απάντηση σε αυτή την ερώτηση βρίσκεται στη διαφορετική πληροφορία που δίνεται από το σφάλμα και από το Συντελεστή Προσδιορισμού: το μοντέλο του μέσου όρου μπορεί να έχει ένα σφάλμα όχι κατά πολύ μεγαλύτερο από αυτό των GC, αλλά δεν επιχειρεί κανένα συσχετισμό. Με άλλα λόγια ο Συντελεστής Προσδιορισμού του είναι σχεδόν μηδέν. Αντιθέτως, τα GC μοντέλα, που επιχειρούνται να αναπτυχθούν, συσχετίζουν την πειραματική με την υπολογισμένη τιμή μέσω της εισόδου, δηλαδή, με το ιδιαίτερο, χαρακτηριστικό διάνυσμα εισόδου κάθε μορίου. Θα μπορούσαμε να πούμε, πως τα μοντέλα GC έχουν έναν εσωτερικό μηχανισμό συσχέτισης που δεν εξετάζεται από το σφάλμα, παρά μόνον από το Συντελεστή Προσδιορισμού. Κρίνεται, λοιπόν, πως η χρήση μόνο του σφάλματος μπορεί να είναι μέχρι και παραπλανητική. Δε είναι δυνατόν να αποφανθούμε, πως το μοντέλο έφτασε στην τελική τιμή μόνο από το σφάλμα ούτε να εξετάσουμε την ποιότητα της συσχέτισης. Επειδή, μάλιστα, είναι σημαντικό να τονιστεί το είδος των διάφορων συσχετίσεων για τυχόν μελλοντική έρευνα, ο Συντελεστής Προσδιορισμού κρίνεται ως ποιο κατάλληλο.

Οι παραπάνω πράξεις εκτελούνται, ασφαλώς, και για τα τρία μοντέλα που αναπτύσσονται, δηλαδή, GWP, CED, EI 99.

Παράρτημα Ε. Συνεισφορές ομάδων για πρόβλεψη δεικτών LCA

Στους πίνακες Ε.1 έως Ε. φαίνονται οι συνεισφορές (μέσες συνεισφορές) των ομάδων μαζί με τη διακύμανση του συνόλου και την αβεβαιότητα (95%).

Πίνακας Ε.1. Συνεισφορές και αβεβαιότητα για το μοντέλο GWP(PLS, 10 συνιστώσες)

Ομάδες	Συνεισφορές(GWP)	Διακύμανση	Αβεβαιότητα (3·Διακύμανση)
intecept	**2,05**	0,12	0,35
CH3	**-0,46**	0,04	0,11
CH2	**0,14**	0,01	0,02
CH	**1,19**	0,17	0,51
C	**2,60**	0,54	1,63
CH2=CH	**0,84**	0,30	0,91
CH=CH	**4,21**	7,02	21,07
CH2=C	**0,22**	0,20	0,59
CH=C	**11,91**	8,82	26,45
C=C	**11,48**	15,61	46,82
OH	**0,07**	0,04	0,11
CH3-(C=O)-	**1,82**	0,34	1,03
CH2-(C=O)-	**0,49**	0,36	1,09
H(C=O)	**-0,50**	0,47	1,41
CH3COO-	**0,10**	0,15	0,44
CH2COO	**0,08**	0,39	1,16
HCOO	**4,35**	3,99	11,98
CH3O	**0,20**	0,56	1,67
CH2O	**0,93**	0,21	0,63
CHO	**3,79**	0,94	2,83
CH3-NH2	**0,48**	0,11	0,32
CH2-NH2	**1,24**	0,08	0,24
CH-NH2	**-0,81**	0,78	2,35
CH3-NH	**-0,42**	0,07	0,21
CH2-NH	**1,59**	0,15	0,45
CH3-N	**0,46**	0,22	0,66
CH2-N	**0,62**	1,17	3,50
CH3-CN	**0,43**	0,09	0,27
CH2-CN	**1,68**	0,27	0,80
COOH	**0,83**	0,80	2,40
CH2-Cl	**-0,59**	0,14	0,43
CH-Cl	**1,20**	0,24	0,71
CH2(CL)2	**0,61**	0,14	0,41
CH-Cl2	**0,38**	0,05	0,16
CH-Cl3	**0,62**	0,15	0,44
C-Cl3	**0,49**	0,10	0,30
C-(Cl)4	**-0,29**	0,06	0,19
CH#C	**1,80**	1,23	3,69
C#C	**3,69**	2,36	7,09
C(F)3	**4,83**	1,22	3,67
CH2=O	**-0,58**	0,14	0,41
C(Cl)F2	**2,41**	1,85	5,55

Πίνακας Ε.2. Συνεισφορές και αβεβαιότητα για το μοντέλο GWP (PLS, 10 συνιστώσες) (συνέχεια)

Ομάδες	Συνεισφορές(GWP)	Διακύμανση	Αβεβαιότητα (3·Διακύμανση)
AC-CH3	**-0,39**	0,10	0,31
AC-CH2	**-0,99**	0,33	0,98
AC-CH	**-0,87**	0,18	0,53
AC-C	**0,44**	0,63	1,88
AC-CH=CH2	**0,32**	0,06	0,18
ACOH	**2,72**	0,24	0,71
AC-COOH	**3,07**	1,97	5,90
AC-CH2Cl	**-0,67**	0,15	0,46
AC-CH(Cl)2	**-0,63**	0,13	0,39
AC-C(F)3	**1,30**	0,54	1,61
AC-NO2	**1,04**	0,32	0,97
AC-NH2	**0,92**	0,17	0,52
AC-Cl	**-0,53**	0,13	0,40
CH3-(Cl)	**0,43**	0,09	0,27
C5H5N	**4,11**	4,09	12,27
ACH	**0,33**	0,01	0,03

Πίνακας Ε.3.Συνεισφορές και αβεβαιότητα για το μοντέλο CED (PLS, 10 συνιστώσες)

Ομάδες	Συνεισφορές(CED)	Διακύμανση	Αβεβαιότητα (3·Διακύμανση)
intecept	**52,85**	**43,24**	**129,71**
CH3	**-2,78**	13,16	39,49
CH2	**1,89**	1,10	3,31
CH	**28,24**	45,75	137,24
C	**62,56**	253,47	760,40
CH2=CH	**27,66**	110,46	331,38
CH=CH	**102,55**	2914,25	8742,74
CH2=C	**18,26**	112,01	336,02
CH=C	**234,84**	3214,05	9642,16
C=C	**228,24**	6641,51	19924,52
OH	**2,59**	14,56	43,67
CH3-(C=O)-	**43,13**	147,53	442,59
CH2-(C=O)-	**8,18**	105,01	315,04
H(C=O)	**-0,98**	160,23	480,68
CH3COO-	**-2,90**	75,30	225,90
CH2COO	**-10,81**	133,41	400,22
HCOO	**61,23**	574,48	1723,43
CH3O	**9,03**	253,18	759,53
CH2O	**19,44**	64,98	194,95
CHO	**77,26**	455,59	1366,77
CH3-NH2	**18,28**	116,60	349,79
CH2-NH2	**37,50**	30,50	91,49
CH-NH2	**-7,15**	372,71	1118,14
CH3-NH	**-10,19**	36,05	108,15

Πίνακας Ε.4. Συνεισφορές και αβεβαιότητα για το μοντέλο CED (PLS, 10 συνιστώσες) (συνέχεια)

Ομάδες	Συνεισφορές(CED)	Διακύμανση	Αβεβαιότητα (3·Διακύμανση)
CH2-NH	**40,28**	**52,69**	**158,06**
CH3-N	**11,22**	96,62	289,85
CH2-N	**31,30**	136,56	409,69
CH3-CN	**14,45**	75,74	227,22
CH2-CN	**54,44**	129,12	387,36
COOH	**-6,31**	65,84	197,51
CH2-Cl	**-10,27**	51,77	155,30
CH-Cl	**15,58**	57,69	173,07
CH2(CL)2	**-8,46**	33,73	101,19
CH-Cl2	**2,10**	6,83	20,50
CH-Cl3	**-5,95**	24,22	72,66
C-Cl3	**2,54**	9,28	27,84
C-(Cl)4	**-11,45**	50,55	151,65
CH#C	**27,96**	347,82	1043,46
C#C	**80,33**	1133,40	3400,20
C(F)3	**88,05**	428,17	1284,51
CH2=O	**-5,99**	24,86	74,59
C(Cl)F2	**43,00**	595,77	1787,30
CH3-(Cl)	**-4,93**	21,10	63,29
C5H5N	**73,86**	1333,98	4001,93
ACH	**8,81**	3,91	11,72
AC-CH3	**-4,37**	67,10	201,29
AC-CH2	**-25,40**	91,16	273,47
AC-CH	**-16,43**	69,51	208,52
AC-C	**-11,26**	197,54	592,61
AC-CH=CH2	**-2,83**	13,55	40,64
ACOH	**82,28**	232,61	697,83
AC-COOH	**58,37**	469,10	1407,30
AC-CH2Cl	**-13,04**	58,93	176,78
AC-CH(Cl)2	**-14,89**	68,03	204,08
AC-C(F)3	**-14,90**	50,62	151,85
AC-NO2	**6,73**	124,05	372,15
AC-NH2	**21,78**	67,25	201,75
AC-Cl	**-11,02**	56,92	170,77

Πίνακας Ε.5. Συνεισφορές και αβεβαιότητα για το μοντέλο EI 99 (PLS, 10 συνιστώσες)

Ομάδες	Συνεισφορές(EI 99)	Διακύμανση	Αβεβαιότητα (3·Διακύμανση)
intecept	**0,277**	**0,000**	**0,001**
CH3	**-0,029**	0,000	0,000
CH2	**0,005**	0,000	0,000
CH	**0,106**	0,000	0,001
C	**0,196**	0,001	0,002
CH2=CH	**-0,026**	0,000	0,001
CH=CH	**-0,008**	0,001	0,003
CH2=C	**0,055**	0,001	0,003
CH=C	**0,212**	0,001	0,004
C=C	**0,028**	0,001	0,003
OH	**-0,057**	0,000	0,000
CH3-(C=O)-	**0,066**	0,001	0,004
CH2-(C=O)-	**0,029**	0,000	0,001
H(C=O)	**0,012**	0,000	0,001
CH3COO-	**-0,017**	0,000	0,001
CH2COO	**-0,001**	0,001	0,002
HCOO	**-0,030**	0,001	0,003
CH3O	**-0,001**	0,003	0,008
CH2O	**0,018**	0,000	0,001
CHO	**0,107**	0,001	0,003
CH3-NH2	**-0,026**	0,000	0,001
CH2-NH2	**0,159**	0,000	0,001
CH-NH2	**0,064**	0,003	0,009
CH3-NH	**-0,056**	0,001	0,002
CH2-NH	**0,143**	0,001	0,002
CH3-N	**-0,037**	0,001	0,002
CH2-N	**0,126**	0,001	0,003
CH3-CN	**0,019**	0,000	0,000
CH2-CN	**0,171**	0,004	0,011
COOH	**-0,035**	0,000	0,001
CH2-Cl	**-0,059**	0,000	0,001
CH-Cl	**0,020**	0,000	0,001
CH2(CL)2	**-0,040**	0,000	0,001
CH-Cl2	**-0,013**	0,000	0,000
CH-Cl3	**-0,041**	0,000	0,001
C-Cl3	**-0,017**	0,000	0,000
C-(Cl)4	**-0,077**	0,002	0,005
CH#C	**-0,031**	0,000	0,001
C#C	**0,099**	0,002	0,006
C(F)3	**0,330**	0,004	0,013
CH2=O	**-0,060**	0,001	0,003

Πίνακας Ε.6. Συνεισφορές και αβεβαιότητα για το μοντέλο EI 99 (PLS, 10 συνιστώσες) (συνέχεια)

Ομάδες	Συνεισφορές(EI 99)	Διακύμανση	Αβεβαιότητα (3·Διακύμανση)
C(Cl)F2	**0,166**	**0,008**	**0,024**
CH3-(Cl)	**-0,027**	0,000	0,001
C5H5N	**0,195**	0,011	0,032
ACH	**0,015**	0,000	0,000
AC-CH3	**-0,018**	0,001	0,002
AC-CH2	**-0,064**	0,001	0,002
AC-CH	**0,007**	0,001	0,002
AC-C	**0,014**	0,002	0,006
AC-CH=CH2	**-0,007**	0,000	0,000
ACOH	**0,231**	0,003	0,008
AC-COOH	**0,146**	0,014	0,042
AC-CH2Cl	**-0,050**	0,001	0,002
AC-CH(Cl)2	**-0,057**	0,001	0,003
AC-C(F)3	**-0,077**	0,001	0,004
AC-NO2	**-0,015**	0,001	0,003
AC-NH2	**0,206**	0,003	0,010
AC-Cl	**-0,049**	0,001	0,002

Παράρτημα ΣΤ. Δείκτες που χρησιμοποιούνται για την αξιολόγηση των μοντέλων

Στον πίνακα ΣΤ.1 φαίνονται οι 16 στατιστικοί δείκτες που χρησιμοποιήθηκαν για την αξιολόγηση του μοντέλου:

Πίνακας ΣΤ.1 Στατιστικοί δείκτες για την αξιολόγηση των μοντέλων

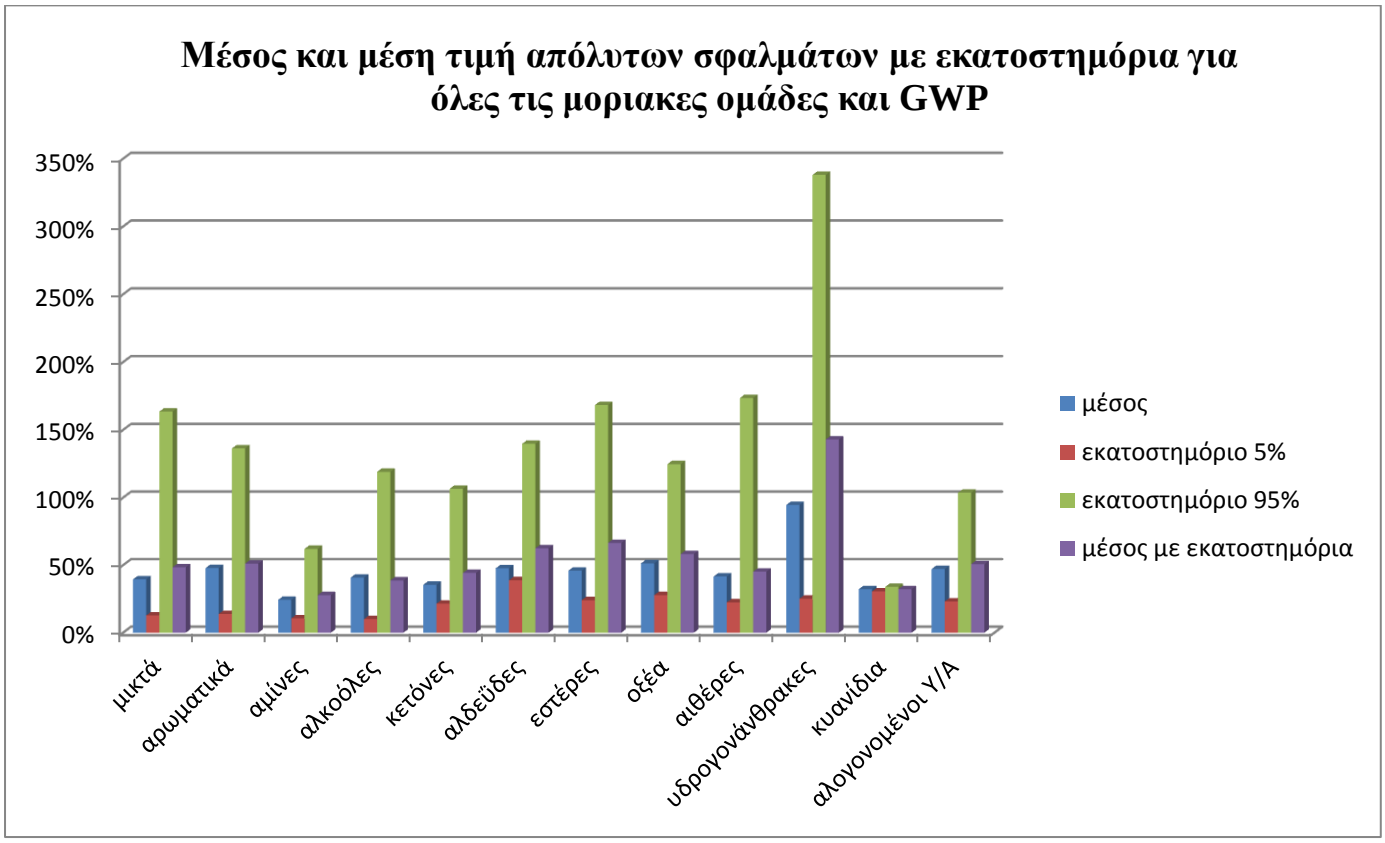
Στατιστικοί δείκτες
Τυπική Απόκλιση των διαφορών μεταξύ πειραματικών μετρήσεων και υπολογισμών (Difference of the Measurements and Predictions Standard Deviations)
R^2 μεταξύ Πειραματικών μετρήσεων και υπολογισμών (R-square of correlation between Measurements and Predictions)
Μέση Τιμή των Απόλυτων Σχετικών Σφαλμάτων (Mean of Absolute Relative Error)
Μέγιστη Τιμή των Απόλυτων Σχετικών Σφαλμάτων (Max of Absolute Relative Error)
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων (Standard Deviation of Absolute Relative Error)
Μέσο Σχετικό Σφάλμα (Mean Relative Error)
Τυπική Απόκλιση των Σχετικών Σφαλμάτων (Standard Deviation of Relative Error)
Ρίζα του μέσου τετραγώνου σφάλματος (Root Mean Square Error)
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων (Slope of the best least squares linear regression line),
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων (Intercept of the best least squares linear regression line),
Συστηματικό RMSE (Systematic RMSE),
Μη συστηματικό RMSE (Non-Systematic RMSE),
Συντελεστής Προσδιορισμού (Coefficient of Determination)
Συντελεστής Προσδιορισμού-D1 (Coefficient of Determination-D1),
Συντελεστής Προσδιορισμού-D2 (Coefficient of Determination-D2)

Παράρτημα Ζ. Βήματα για τον υπολογισμό της Απόστασης Mahalanobis

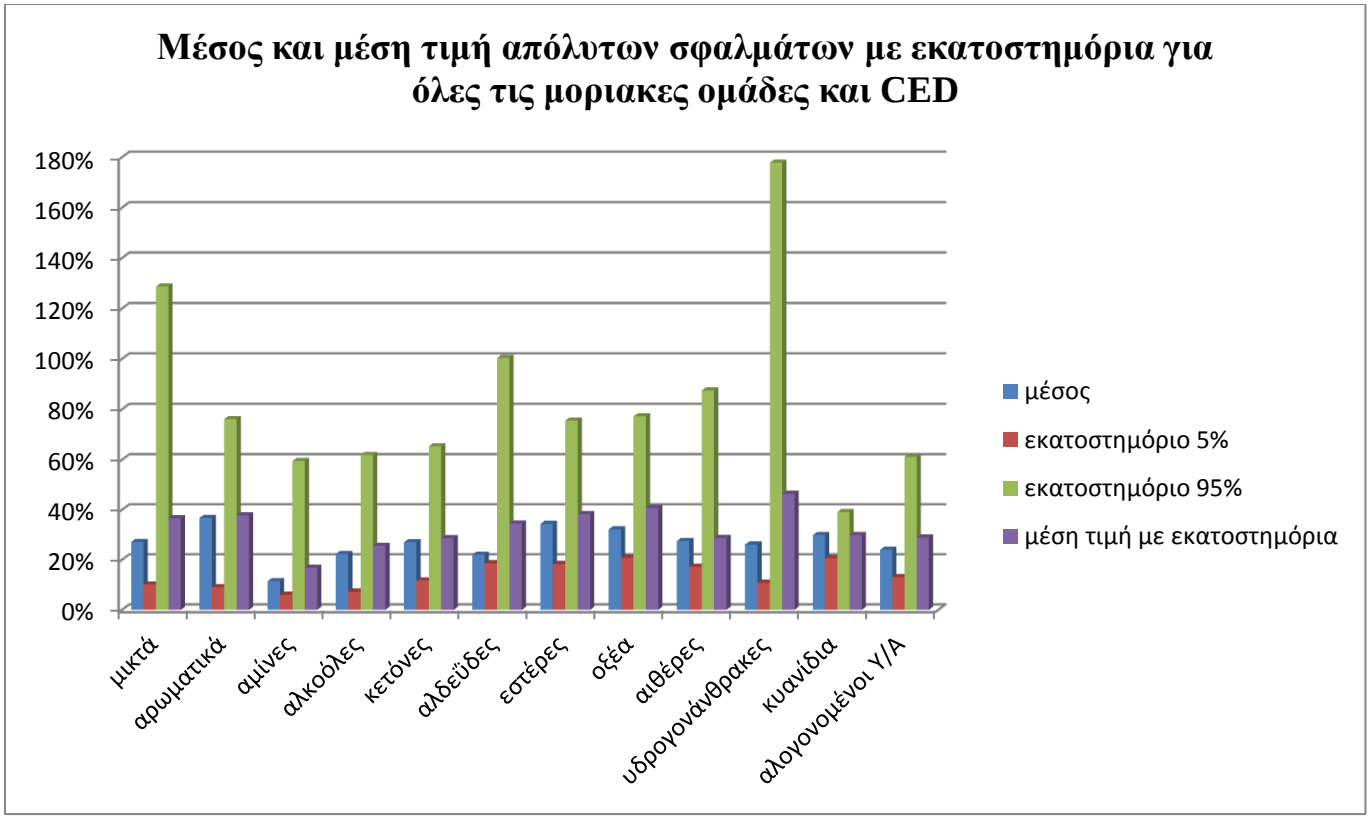
Το πρόβλημα που αντιμετωπίστηκε, κατά την εξαγωγή της απόστασης είναι πως σε αρκετά σύνολα εκπαίδευσης προκύπτει πως οι πίνακες 171x57 μπορεί να περιέχουν μια στήλη ίση με το 0. Αυτό προκύπτει στις περιπτώσεις, όπου υπάρχει κάποιο μόριο από αυτά που μπορούν να περιγραφούν από μια ομάδα έχει μεταφερθεί στο σύνολο δοκιμής. Έτσι δεν υπάρχει πλέον στη στήλη της αντίστοιχης ομάδας (που χρησιμοποιείται, ασφαλώς, μια μόνο φορά σε αυτό το μόριο) η μονάδα και γίνεται όλη η στήλη ίση με 0. Σε αυτές τις περιπτώσεις που η ορίζουσα του πίνακα είναι 0 και δε μπορεί να εξαχθεί ο αντίστροφος πίνακας είναι αδύνατο να υπολογιστεί η απόσταση mahalanobis. Για να αντιμετωπιστεί αυτό προστίθεται στα στοιχεία του κάθε πίνακα κάθε συνόλου δοκιμής «θόρυβος», δηλαδή τυχαίοι πολύ μικροί αριθμοί (π.χ. 0.000156). Κατά αυτόν τον τρόπο, αφενός διασπάται το πρόβλημα της μηδενικής ορίζουσας (singularity), αλλά αφετέρου οι αριθμοί είναι τόσο μικροί που δε μπορούν να επηρεάσουν την απόσταση mahalanobis.

(Εννοείται πως αργότερα στην ανάπτυξη του μοντέλου χρησιμοποιείται ο αρχικός πίνακας δεδομένων, χωρίς το «θόρυβο».)

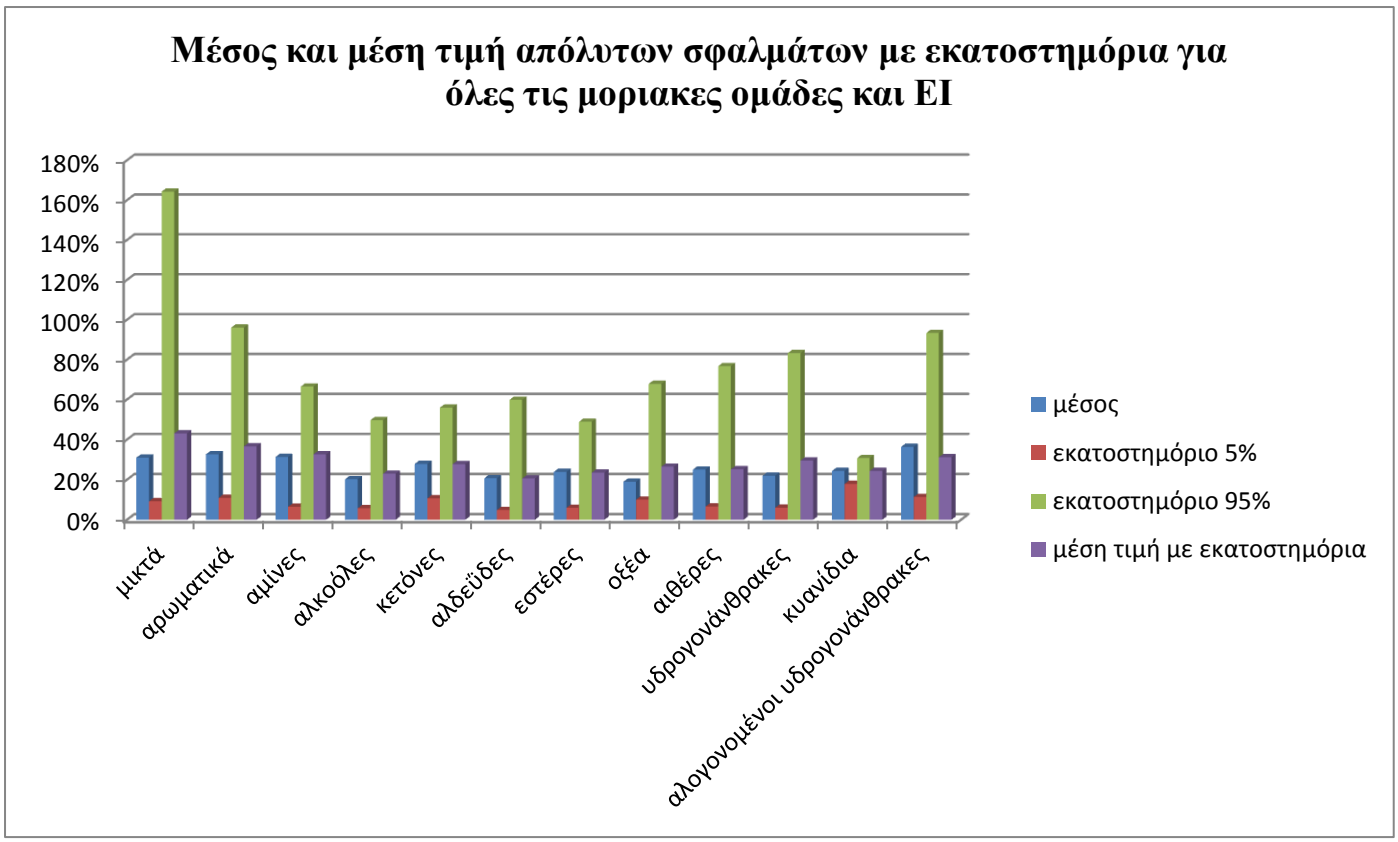
Παράρτημα Η. Εξαγωγή μέσων και μέσων όρων σφαλμάτων με χρήση εκατοστημορίων



Διάγραμμα Η.1. Στατιστικοί δείκτες για όλες τις ομάδες στο GWP



Διάγραμμα Η.2. Στατιστικοί δείκτες για όλες τις ομάδες στο CED



Διάγραμμα Η.3. Στατιστικοί δείκτες για όλες τις ομάδες στο EI 99

Παράρτημα Θ. Αποτελέσματα Ανάλυσης Διακύμανσης

Ακολουθούν οι πίνακες Θ.1-Θ.3 που προέκυψαν για τον δείκτη και για το κάθε σύνολο ομάδων για το χαμηλό σφάλμα:

Πίνακας Θ.1. ANOVA για τις ομάδες του μοντέλου GWP και χαμηλό σφάλμα

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ(GWP groups)						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Γραμμές	3,33E-06	10	3,33E-07	6,85E-06	1	2,347878
Στήλες	4,034622	2	2,017311	41,43879	7,71E-08	3,492828
Σφάλμα	0,973634	20	0,048682			
Σύνολο	5,008259	32				

Πίνακας Θ.2. ANOVA για τις ομάδες του μοντέλου CED και χαμηλό σφάλμα

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ(CED groups)						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Γραμμές	3,55E-15	26	1,37E-16	2,52E-15	1	1,70962
Στήλες	6,690746	2	3,345373	61,75791	1,84E-14	3,175141
Σφάλμα	2,816795	52	0,054169			
Σύνολο	9,507541	80				

Πίνακας Θ.3. ANOVA για τις ομάδες του μοντέλου EI 99 και χαμηλό σφάλμα

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ(EI 99 groups)						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Γραμμές	6,67E-09	31	2,15E-10	4,98E-09	1	1,636151
Στήλες	7,290706	2	3,645353	84,43608	1,99E-18	3,145258
Σφάλμα	2,676722	62	0,043173			
Σύνολο	9,967428	95				

Ακολουθεί η ANOVA για τις ομάδες των τριών δεικτών και μέσο σφάλμα (πίνακες Θ.4-Θ.6):

Πίνακας Θ.4. ANOVA για τις ομάδες του μοντέλου GWP και μεσό σφάλμα

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ(GWP groups)						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Γραμμές	-5,3E-15	22	-2,4223E-16	-6,8E-15		1,788887
Στήλες	9,94243	2	4,971215044	138,9231	9,72E-20	3,209278
Σφάλμα	1,574493	44	0,035783942			
Σύνολο	11,51692	68				

Πίνακας Θ.5. ANOVA για τις ομάδες του μοντέλου CED και μέσο σφάλμα

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ(CED groups)						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Γραμμές	0	21	0	0	1	1,812817
Στήλες	5,296263	2	2,64813171	35,9894	7,85E-10	3,219942
Σφάλμα	3,090392	42	0,073580763			
Σύνολο	8,386655	65				

Πίνακας Θ.6. ANOVA για τις ομάδες του μοντέλου EI 99 και μέσο σφάλμα

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ(EI 99 groups)						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Γραμμές	-4,4E-15	16	-2,77556E-16	-4,59068E-15		1,971683
Στήλες	3,785569	2	1,892784435	31,30604819	2,93E-08	3,294537
Σφάλμα	1,934741	32	0,060460663			
Σύνολο	5,72031	50				

Ακολουθεί η ANOVA για τις ομάδες των τριών δεικτών και υψηλό σφάλμα (πίνακες Θ.7-Θ.9):

Πίνακας Θ.7. ANOVA για τις ομάδες του μοντέλου GWP και υψηλό σφάλμα

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ(GWP groups)						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Γραμμές	6,43E-09	13	4,94E-10	1,18E-08	1	2,119166
Στήλες	3,521967	2	1,760984	42,00016	7,19E-09	3,369016
Σφάλμα	1,090129	26	0,041928			
Σύνολο	4,612096	41				

Πίνακας Θ.8. ANOVA για τις ομάδες του μοντέλου CED και υψηλό σφάλμα

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ(CED groups)						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Γραμμές	2,49E-10	8	3,11E-11	6,13E-10	1	2,591096
Στήλες	1,120853	2	0,560427	11,02907	0,000976	3,633723
Σφάλμα	0,813017	16	0,050814			
Σύνολο	1,93387	26				

Πίνακας Θ.9. ANOVA για τις ομάδες του μοντέλου EI 99 και υψηλό σφάλμα

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ(EI 99 groups)						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Γραμμές	4,44E-16	6	7,4E-17	2,25E-15	1	2,99612
Στήλες	3,372381	2	1,68619	51,31884	1,32E-06	3,885294
Σφάλμα	0,394286	12	0,032857			
Σύνολο	3,766667	20				

Παράρτημα Ι. Αποτελέσματα στατιστικών δεικτών μοντέλων

Για τη μέθοδο MLR προκύπτει:

Πίνακας Ι.1.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για MLR και GWP

Στατιστικοί δείκτες για GWP και MLR	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,50	0,13
R^2 μεταξύ Πειραματικών και υπολογισμών	0,81	0,37
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,40	0,60
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	4,84	4,19
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,56	0,71
Μέσο Σχετικό Σφάλμα	0,17	0,22
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,67	0,91
Ρίζα του μέσου τετραγώνου σφάλματος	2,19	4,74
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,81	0,62
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,81	1,89
Συστηματικό RMSE	0,95	3,01
Μη συστηματικό RMSE	1,97	3,34
Συντελεστής Προσδιορισμού	0,81	0,18
Συντελεστής Προσδιορισμού-D1	0,72	0,49
Συντελεστής Προσδιορισμού-D2	0,94	0,64

Πίνακας Ι.2.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για MLR και CED

Στατιστικοί δείκτες για CED και MLR	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	5,19	-4,41
R^2 μεταξύ Πειραματικών και υπολογισμών	0,90	0,48
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,24	0,39
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	5,40	5,43
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,47	0,66
Μέσο Σχετικό Σφάλμα	0,07	0,12
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,53	0,74
Ρίζα του μέσου τετραγώνου σφάλματος	32,54	85,82
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,90	0,77
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	10,77	20,99
Συστηματικό RMSE	10,08	49,65
Μη συστηματικό RMSE	30,91	61,00
Συντελεστής Προσδιορισμού	0,90	0,24
Συντελεστής Προσδιορισμού-D1	0,77	0,55
Συντελεστής Προσδιορισμού-D2	0,97	0,72

Πίνακας Ι.3.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για MLR και EI 99

Στατιστικοί δείκτες για EI 99 και MLR	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,03	0,02
R^2 μεταξύ Πειραματικών και υπολογισμών	0,70	0,28
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,26	0,40
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	3,12	2,43
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,36	0,46
Μέσο Σχετικό Σφάλμα	0,10	0,15
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,43	0,59
Ρίζα του μέσου τετραγώνου σφάλματος	0,10	0,18
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,70	0,46
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,10	0,19
Συστηματικό RMSE	0,06	0,11
Μη συστηματικό RMSE	0,09	0,14
Συντελεστής Προσδιορισμού	0,70	0,14
Συντελεστής Προσδιορισμού-D1	0,70	0,50
Συντελεστής Προσδιορισμού-D2	0,91	0,68

Για τη μέθοδο PCA/PCR προκύπτει:

Πίνακας Ι.4.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για PCA/PCR και GWP

Στατιστικοί δείκτες για GWP και PCA/PCR	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	2,58	2,62
R^2 μεταξύ Πειραματικών και υπολογισμών	0,25	0,28
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,82	0,88
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	12,50	8,44
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	1,37	1,40
Μέσο Σχετικό Σφάλμα	0,54	0,58
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	1,51	1,55
Ρίζα του μέσου τετραγώνου σφάλματος	4,48	5,16
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,25	0,26
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	3,21	3,53
Συστηματικό RMSE	3,86	4,39
Μη συστηματικό RMSE	2,25	2,23
Συντελεστής Προσδιορισμού	0,25	0,19
Συντελεστής Προσδιορισμού-D1	0,37	0,37
Συντελεστής Προσδιορισμού-D2	0,60	0,47

Πίνακας Ι.5.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για PCA/PCR και CED

Στατιστικοί δείκτες για CED και PCA/PCR	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	44,15	47,04
R^2 μεταξύ Πειραματικών και υπολογισμών	0,34	0,34
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,52	0,56
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	10,76	11,63
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,93	1,15
Μέσο Σχετικό Σφάλμα	0,27	0,30
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	1,03	1,17
Ρίζα του μέσου τετραγώνου σφάλματος	86,01	101,74
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,34	0,37
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	72,30	77,41
Συστηματικό RMSE	69,49	81,35
Μη συστηματικό RMSE	50,28	49,23
Συντελεστής Προσδιορισμού	0,34	0,24
Συντελεστής Προσδιορισμού-D1	0,42	0,40
Συντελεστής Προσδιορισμού-D2	0,69	0,57

Πίνακας Ι.6.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για PCA/PCR και EI 99

Στατιστικοί δείκτες για EI 99 και PCA/PCR	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,114317	0,110472
R^2 μεταξύ Πειραματικών και υπολογισμών	0,163509	0,10559
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,427479	0,452235
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	4,833011	3,547534
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,490749	0,514272
Μέσο Σχετικό Σφάλμα	0,218834	0,226999
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,614744	0,640718
Ρίζα του μέσου τετραγώνου σφάλματος	0,174538	0,189788
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,163509	0,126861
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,291594	0,309374
Συστηματικό RMSE	0,159615	0,171722
Μη συστηματικό RMSE	0,070207	0,076063
Συντελεστής Προσδιορισμού	0,163509	0,06262
Συντελεστής Προσδιορισμού-D1	0,322248	0,302872
Συντελεστής Προσδιορισμού-D2	0,4941	0,420843

Για τη μέθοδο PLS με 10 συνιστώσες προκύπτει:

Πίνακας Ι.7.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για PLS και GWP

Στατιστικοί δείκτες για GWP και PLS	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,62	0,50
R^2 μεταξύ Πειραματικών και υπολογισμών	0,77	0,43
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,46	0,62
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	5,02	4,00
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,61	0,75
Μέσο Σχετικό Σφάλμα	0,18	0,24
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,74	0,94
Ρίζα του μέσου τετραγώνου σφάλματος	2,42	4,40
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,77	0,64
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,98	1,79
Συστηματικό RMSE	1,15	2,78
Μη συστηματικό RMSE	2,12	3,02
Συντελεστής Προσδιορισμού	0,77	0,26
Συντελεστής Προσδιορισμού-D1	0,68	0,51
Συντελεστής Προσδιορισμού-D2	0,93	0,69

Πίνακας Ι.8.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για PLS και CED

Στατιστικοί δείκτες για CED και PLS	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	7,25	5,46
R^2 μεταξύ Πειραματικών και υπολογισμών	0,86	0,55
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,29	0,40
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	5,16	4,91
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,46	0,58
Μέσο Σχετικό Σφάλμα	0,07	0,11
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,54	0,68
Ρίζα του μέσου τετραγώνου σφάλματος	38,38	77,66
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,86	0,80
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	14,88	20,50
Συστηματικό RMSE	13,93	42,69
Μη συστηματικό RMSE	35,72	54,19
Συντελεστής Προσδιορισμού	0,86	0,32
Συντελεστής Προσδιορισμού-D1	0,73	0,57
Συντελεστής Προσδιορισμού-D2	0,96	0,78

Πίνακας Ι.9.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για PLS και EI 99

Στατιστικοί δείκτες για EI 99 και PLS	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,03	0,03
R^2 μεταξύ Πειραματικών και υπολογισμών	0,68	0,35
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,27	0,38
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	3,15	2,36
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,36	0,44
Μέσο Σχετικό Σφάλμα	0,10	0,15
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,44	0,56
Ρίζα του μέσου τετραγώνου σφάλματος	0,11	0,17
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,68	0,50
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,11	0,18
Συστηματικό RMSE	0,06	0,10
Μη συστηματικό RMSE	0,09	0,13
Συντελεστής Προσδιορισμού	0,68	0,24
Συντελεστής Προσδιορισμού-D1	0,68	0,53
Συντελεστής Προσδιορισμού-D2	0,90	0,73

Για τη μέθοδο παρεμβολής τύπου «kriging» προέκυψε:

Πίνακας Ι.10.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για παρεμβολή τύπου «kriging» και GWP

Στατιστικοί δείκτες για GWP και παρεμβολή τύπου «kriging»	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,16	1,27
R^2 μεταξύ Πειραματικών και υπολογισμών	0,97	0,18
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,32	0,53
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	6,67	7,01
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,65	0,95
Μέσο Σχετικό Σφάλμα	0,18	0,24
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,71	1,06
Ρίζα του μέσου τετραγώνου σφάλματος	0,90	4,47
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,96	0,27
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,18	2,92
Συστηματικό RMSE	0,23	3,30
Μη συστηματικό RMSE	0,87	2,78
Συντελεστής Προσδιορισμού	0,97	0,10
Συντελεστής Προσδιορισμού-D1	0,85	0,48
Συντελεστής Προσδιορισμού-D2	0,99	0,50

Πίνακας Ι.11.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για παρεμβολή τύπου «kriging» και CED

Στατιστικοί δείκτες για CED και παρεμβολή τύπου «kriging»	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	2,86	4,11
R^2 μεταξύ Πειραματικών και υπολογισμών	0,97	0,35
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,22	0,35
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	6,98	8,66
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,57	0,97
Μέσο Σχετικό Σφάλμα	0,09	0,11
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,60	0,96
Ρίζα του μέσου τετραγώνου σφάλματος	18,57	68,98
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,96	0,50
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	4,10	50,22
Συστηματικό RMSE	4,31	38,05
Μη συστηματικό RMSE	18,06	53,11
Συντελεστής Προσδιορισμού	0,97	0,22
Συντελεστής Προσδιορισμού-D1	0,85	0,53
Συντελεστής Προσδιορισμού-D2	0,99	0,67

Πίνακας Ι.12.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για παρεμβολή τύπου «kriging» και EI 99

Στατιστικοί δείκτες για EI 99 και παρεμβολή τύπου «kriging»	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,02	0,04
R^2 μεταξύ Πειραματικών και υπολογισμών	0,92	0,24
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,18	0,34
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	3,25	2,73
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,32	0,48
Μέσο Σχετικό Σφάλμα	0,07	0,12
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,36	0,57
Ρίζα του μέσου τετραγώνου σφάλματος	0,06	0,15
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,86	0,37
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,05	0,22
Συστηματικό RMSE	0,03	0,11
Μη συστηματικό RMSE	0,05	0,11
Συντελεστής Προσδιορισμού	0,91	0,15
Συντελεστής Προσδιορισμού-D1	0,83	0,51
Συντελεστής Προσδιορισμού-D2	0,98	0,66

Για τη μέθοδο RBF προέκυψε:

Πίνακας Ι.13.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για RBF και GWP

Στατιστικοί δείκτες για GWP και RBF	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,50	1,68
R^2 μεταξύ Πειραματικών και υπολογισμών	0,83	0,31
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,47	0,53
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	6,32	4,41
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,73	0,73
Μέσο Σχετικό Σφάλμα	0,25	0,20
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,84	0,88
Ρίζα του μέσου τετραγώνου σφάλματος	2,28	3,95
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,83	0,39
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,75	2,55
Συστηματικό RMSE	0,96	3,16
Μη συστηματικό RMSE	2,07	2,43
Συντελεστής Προσδιορισμού	0,83	0,22
Συντελεστής Προσδιορισμού-D1	0,69	0,49
Συντελεστής Προσδιορισμού-D2	0,95	0,64

Πίνακας Ι.14.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για RBF και CED

Στατιστικοί δείκτες για CED και RBF	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	6,43	8,02
R^2 μεταξύ Πειραματικών και υπολογισμών	0,89	0,39
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,30	0,38
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	6,08	7,10
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,53	0,68
Μέσο Σχετικό Σφάλμα	0,11	0,09
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,60	0,74
Ρίζα του μέσου τετραγώνου σφάλματος	37,93	61,32
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,89	0,56
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	11,85	44,85
Συστηματικό RMSE	12,47	33,96
Μη συστηματικό RMSE	35,81	48,01
Συντελεστής Προσδιορισμού	0,89	0,27
Συντελεστής Προσδιορισμού-D1	0,73	0,51
Συντελεστής Προσδιορισμού-D2	0,97	0,73

Πίνακας Ι.15.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για RBF και EI 99

Στατιστικοί δείκτες για EI 99 και RBF	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,05	0,08
R^2 μεταξύ Πειραματικών και υπολογισμών	0,53	0,16
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,30	0,34
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	3,83	2,61
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,44	0,48
Μέσο Σχετικό Σφάλμα	0,13	0,13
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,52	0,57
Ρίζα του μέσου τετραγώνου σφάλματος	0,13	0,16
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,53	0,21
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,16	0,26
Συστηματικό RMSE	0,09	0,14
Μη συστηματικό RMSE	0,10	0,08
Συντελεστής Προσδιορισμού	0,53	0,11
Συντελεστής Προσδιορισμού-D1	0,63	0,46
Συντελεστής Προσδιορισμού-D2	0,83	0,55

Για τη μέθοδο RBF-PCA προέκυψε:

Πίνακας Ι.16.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για RBF-PCA και GWP

Στατιστικοί δείκτες για GWP και RBF-PCA	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	2,54	2,99
R^2 μεταξύ Πειραματικών και υπολογισμών	0,27	0,31
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,79	0,85
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	11,81	8,14
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	1,32	1,36
Μέσο Σχετικό Σφάλμα	0,53	0,57
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	1,45	1,50
Ρίζα του μέσου τετραγώνου σφάλματος	4,44	5,18
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,27	0,23
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	3,17	3,67
Συστηματικό RMSE	3,80	4,59
Μη συστηματικό RMSE	2,26	1,98
Συντελεστής Προσδιορισμού	0,27	0,19
Συντελεστής Προσδιορισμού-D1	0,38	0,36
Συντελεστής Προσδιορισμού-D2	0,61	0,47

Πίνακας Ι.17.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για RBF-PCA και CED

Στατιστικοί δείκτες για CED και RBF-PCA	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	43,10	56,48
R^2 μεταξύ Πειραματικών και υπολογισμών	0,36	0,37
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,50	0,53
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	10,31	11,08
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,89	1,11
Μέσο Σχετικό Σφάλμα	0,26	0,29
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,99	1,12
Ρίζα του μέσου τετραγώνου σφάλματος	84,99	102,26
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,36	0,33
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	70,62	82,55
Συστηματικό RMSE	67,95	86,99
Μη συστηματικό RMSE	50,35	43,99
Συντελεστής Προσδιορισμού	0,36	0,25
Συντελεστής Προσδιορισμού-D1	0,43	0,40
Συντελεστής Προσδιορισμού-D2	0,70	0,56

Πίνακας Ι.18.Τιμές στατιστικών δεικτών για σύνολα δοκιμής και εκπαίδευσης για RBF-PCA και EI 99

Στατιστικοί δείκτες για EI 99 και RBF-PCA	Σύνολο εκπαίδευσης	Σύνολο δοκιμής
Τυπική Απόκλιση διαφορών πειραματικών και υπολογισμών	0,12	0,12
R^2 μεταξύ Πειραματικών και υπολογισμών	0,16	0,13
Μέση Τιμή Απόλυτων Σχετικών Σφαλμάτων	0,42	0,44
Μέγιστη Τιμή Απόλυτων Σχετικών Σφαλμάτων	5,21	3,52
Τυπική Απόκλιση των Απόλυτων Σχετικών Σφαλμάτων	0,53	0,55
Μέσο Σχετικό Σφάλμα	0,22	0,24
Τυπική Απόκλιση των Σχετικών Σφαλμάτων	0,64	0,67
Ρίζα του μέσου τετραγώνου σφάλματος	0,18	0,19
Κλίση της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,16	0,11
Σταθερά της βέλτιστης ευθείας ελαχίστων τετραγώνων	0,29	0,32
Συστηματικό RMSE	0,16	0,18
Μη συστηματικό RMSE	0,07	0,06
Συντελεστής Προσδιορισμού	0,16	0,08
Συντελεστής Προσδιορισμού-D1	0,32	0,30
Συντελεστής Προσδιορισμού-D2	0,48	0,41