Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Πληροφοριακά και Επικοινωνιακά Συστήματα

Μεταπτυχιακή Διατριβή

Ανάπτυξη Web-based περιβάλλοντος για Οπτικοποίηση της Απόδοσης Μοντέλων Πρόβλεψης Κόστους Λογισμικού

Ανδρέας Μιχαήλ

Επιβλέπων Καθηγητής
Δρ. Μήττας Α. Νικόλαος

Δεκέμβριος 2016

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Πληροφοριακά και Επικοινωνιακά Συστήματα

Μεταπτυχιακή Διατριβή

Ανάπτυξη Web-based περιβάλλοντος για Οπτικοποίηση της Απόδοσης Μοντέλων Πρόβλεψης Κόστους Λογισμικού

Ανδρέας Μιχαήλ

Επιβλέπων Καθηγητής

Δρ. Μήττας Α. Νικόλαος

Η παρούσα μεταπτυχιακή Διατριβή υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων για απόκτηση μεταπτυχιακού τίτλου σπουδών στα Πληροφοριακά και Επικοινωνιακά Συστήματα από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών του Ανοικτού Πανεπιστημίου Κύπρου.

Δεκέμβριος 2016

Περίληψη

Η εκτίμηση κόστους λογισμικού (ΕΚΛ) είναι ένα σημαντικό ζήτημα, το οποίο έχει προσελκύσει έντονα το ενδιαφέρον αρκετών ερευνητών τις τελευταίες δεκαετίες. Κύριος στόχος είναι η ακριβής εκτίμηση της απαιτούμενης προσπάθειας (κυρίως σε ανθρωπομήνες) για την ολοκλήρωση ενός έργου λογισμικού. Η διαδικασία αυτή μπορεί να χρησιμοποιηθεί για το σχεδιασμό, χρονοπρογραμματισμό καθώς και τη διαχείριση κινδύνων ενός έργου.

Για την ικανοποίηση της ανάγκης έγκυρων και ακριβών προβλέψεων, η γνώση των χαρακτηριστικών κάθε μοντέλου εκτίμησης είναι πολύ σημαντική. Ένας επαγγελατίας θα πρέπει να γνωρίζει την ακρίβεια κάθε μοντέλου και εάν οι εκτιμήσεις του είναι μεροληπτικές, δηλαδή αν έχει τη τάση να παράγει είτε υπερεκτιμημένες είτε υποεκτιμημένες προβλέψεις. Με την οπτικοποίηση του σφάλματος πρόβλεψης στον RROC space είμαστε σε θέση να παρατηρήσουμε το μη ισοζυγισμένο σφάλμα πρόβλεψης για ένα μεγάλο αριθμό έργων, όπως και την ακρίβεια πρόβλεψης κάθε μοντέλου.

Λόγω του ότι μια μικρή διαφοροποίηση στο σύνολο δεδομένων μπορεί να επηρεάσει σημαντικά την απόδοση των μοντέλων εκτίμησης, δεν μπορούμε να καταλήξουμε σε ασφαλή συμπεράσματα με τη χρήση ενός μεμονωμένου συνόλου δεδομένου. Επιπρόσθετα, η συνέπεια των εκτιμήσεων κάθε μοντέλου θα πρέπει να ερευνηθεί, δηλαδή εάν και εφόσον έχει την ίδια ή παρόμοια απόδοση πρόβλεψης για κάθε μελλοντικό έργο. Για την αντιμετώπιση του ζητήματος αυτού κάνουμε χρήση τεχνικών αναδειγματολειψίας με επανάθεση, κατασκευάζοντας ένα μεγάλο αριθμό δειγμάτων για εξαγωγή ασφαλών συμπερασμάτων.

Η μεταπτυχιακή Διατριβή προχωρά στην υλοποίηση ενός δυαδικτυακού εργαλείου, με το οποίο ο επαγγελματίας μπορεί να διεξάγει άμεση σύγκριση των μοντέλων ΕΚΛ με την οπτικοποίηση των αποδόσεων τους στον RROC space. Το εργαλείο αυτό είναι φιλικό προς το χρήστη και δεν απαιτεί εξειδικευμένες γνώσεις, αλλά απλά τη εισαγωγή ενός συνόλου δεδομένων το οποίο εμπεριέχει τις πραγματικές και εκτιμημένες τιμές κόστους κάθε μεθόδου ΕΚΛ.

Summary

Software cost estimation (SCE) is an important issue, which has strongly attracted the interest of several researchers over the past decades. The main objective is the accurate prediction of the effort required (particularly in manmonths) to complete a software project. This procedure can be used for planning, scheduling as well as risk management of a project.

The choce of the best software technique is a crusial matter that has attracted the interest of many researchers in recent decades. The main goal is to exact an accurate estimate of the effort (mainly in manmonths) needed for completing a software project. This procedure can be used for planning and time scheduling of each activity and also for the risk management of a project.

In order to fulfill the needs for reliable and accurate predictions, knowledge of the characteristics of each prediction model is very important Knowing the characteristics of each prediction model is very crucial to answer this issue. A practisioner should know the accuracy of each model and if the estimations derived are biased, thus if there is a tendency constantly overestimating or underestimating cost. Visualizing the prediction error in RROC space, we can observe the non balanced prediction error for a large number of projects and also the prediction accuracy of each model.

A small differentiation of the data set can have significant effect in the prediction abilities of the comparative models so we can not have secure results with the use of one and only data set Furthermore, we have to study the consistency of the estimations thus if the estimating abilities are similar for every future project. In order to deal with this matter we suggest the use of resampling techniques, producing a large number of samples for more accurate results.

This postgraduation Thesis proceeds to the implementation of web tool, with which the practitioner may carry out a direct comparison of SCE methods by visualizing their performance in RROC space. This tool is user friendly and does not require specialized knowledge, but simply the introduction of a data set that contains the actual and predicted cost values of each SCE method.

Ευχαριστίες

Θα ήθελα να εκφράσω την ευγνωμοσύνη μου προς τον επιβλέποντα καθηγητή μου κ.Νικόλαο Μήττα για την εμπιστοσύνη που μου έδειξε, καθώς και για την υπομονή του όλο αυτό το διάστημα. Η ευρεία γνώση του στα σχετικά θέματα, η στήριξη και καθοδήγηση του υπήρξε καθοριστικής σημασίας για την ολοκλήρωση της μεταπτυχιακής Διατριβής.

Επιπλέον θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς την οικογένεια μου για την υπομονή και κατανόηση τους, καθώς και τη στήριξη που μου παρείχαν στη ζωή μου.

Περιεχόμενα

Πίνακας με ακρωνύμια

Ακρωνύμιο	Περιγραφή
AAE	Assymetric Absolute Error
AE	Absolute Error
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
Bag	Bagging EbA
CART	Classification and Regression Trees
CBR	Cased Based Reasoning
EbA	Estimation by Analogy
IBRE	Inverted Balance Relative Error
KS	Kolmogorov Smirnov
LM	Linear Model
LOOCV	Leave one out cross validation
LS	Least squares
LSD	Logarithmic Standard Deviation
MAAE	Mean Assymetric Absolute Error
MAE	Mean Absolute Error
MdMER	Median Magnitude of Relative Error to the Estimate
MdMRE	Median Magnitude of Relative Error
ME	Mean Error
MIBRE	Mean Inverted Balance Relative Error
MMER	Mean Magnitude Relative Error to the Estimate
MMER	Median Magnitude of Relative Error
MSE	Mean Square Error
NB	Naïve Bayes
Pred(p)	Predictions between p% of the actual values
REC	Regression Error Characteristic
RobMM	Robust M-estimator
ROC	Receiver Operating Characteristic
RROC	Regression Receiver Operating Characteristic
RSD	Relative Standard Deviation

SCE	Software Cost Estimation
SD	Standard Deviation
SE	Simple Error
SIM	Similarity
SOE	Sum of Overestimations
SQE	Square Error
StatREC	Statistical Regression Error Characteristic
SUE	Sum of Underestimations
UFC	Unadjusted Function-point Counts
ΕΚΛ	Εκτίμηση Κόστους Λογισμικού

Κεφάλαιο 1

Εισαγωγή

Η Εκτίμηση Κόστους Λογισμικού-ΕΚΛ (Software Cost Estimation-SCE) είναι μια ερευνητική περιοχή, η οποία έχει αναπτυχθεί ραγδαία τις τελευταίες δεκαετίες. Ο κύριος στόχος των ερευνητών είναι να προβλέψουν τους απαραίτητους πόρους που απαιτούνται, όπως τον προϋπολογισμό και τον χρόνο για την ολοκλήρωση ενός έργου λογισμικού. Εξαιτίας της ραγδαίας αύξησης της ζήτησης για λογισμικό υψηλών προδιαγραφών και ποιότητας, αλλά και του γεγονότος ότι σχετίζεται άμεσα με όλες τις φάσεις ανάπτυξης, η ΕΚΛ αποτελεί μία από τις πιο κρίσιμες δραστηριότητες της διαχείρισης των έργων (project management) (Mittas & Angelis 2008a: 616). Οι ανακριβείς εκτιμήσεις μπορούν να επιφέρουν καταστροφικές συνέπειες, τόσο στους οργανισμούς ανάπτυξης (software organizations), όσο και στους τελικούς πελάτες (customers) με μεγάλες καθυστερήσεις και χαμηλή ποιότητα των παραγόμενων προϊόντων ή ακόμα και ακύρωση των έργων (Lederer & Prasad 1995: 125).

Παρά το γεγονός ότι η συστηματική βιβλιογραφική επισκόπηση φανερώνει την εισαγωγή και εφαρμογή πολλών μεθόδων εκτίμησης με καινοτόμα και ενδιαφέρουσα ευρήματα, ο τεράστιος αριθμός των προτεινόμενων μεθόδων έφερε στην επιφάνεια ένα μεγάλο πρόβλημα που ονομάζεται συμπερασματική αστάθεια “conclusion instability” (Myrtveit et al. 2005: 380) και σχετίζεται με τα αντιφατικά αποτελέσματα που εξάγονται από τις πειραματικές εργασίες (experimental studies) και την έλλειψη μηχανισμών για την εξαγωγή ασφαλών συμπερασμάτων (Mittas & Angelis 2008a: 617). Το πρόβλημα αυτό καθιστά απαραίτητη τη χρήση, εύκολα ερμηνεύσιμων στατιστικών εργαλείων, ικανών να αναδείξουν την καλύτερη τεχνική πρόβλεψης, καθώς η πληθώρα των εναλλακτικών μεθόδων και μοντέλων που παρουσιάζονται στη βιβλιογραφία, κάνουν δύσκολη την επιλογή του καταλληλότερου μοντέλου, για κάποιο συγκεκριμένο σύνολο δεδομένων.

Η ερευνητική προσπάθεια στα πλαίσια της μεταπτυχιακής Διατριβής επικεντρώνεται στη οπτική αναπαράσταση της ικανότητας πρόβλεψης των μοντέλων, με τη χρήση των

Regression Receiver Operating Curves (RROC) (Hernández-Orallo 2013: 5). Οι καμπύλες αυτές είναι ένα δισδιάστατο γράφημα, από το οποίο μπορούμε να εξάγουμε εύκολα χρήσιμες πληροφορίες, σχετικά με την ανωτερότητα ενός μοντέλου πρόβλεψης σε σχέση με άλλα. Το πιο σημαντικό όμως, είναι ότι η RROC ανάλυση προσφέρει τη δυνατότητα μελέτης των δύο τύπων σφάλματος που εμφανίζονται κατά τη διαδικασία πρόβλεψης ήτοι της υπό-εκτίμησης (under-estimation) και της υπέρ-εκτίμησης (over-estimation). Πρόκειται για έναν πολύ σημαντικό διαχωρισμό, καθώς οι δύο τύποι σφαλμάτων επιφέρουν εντελώς διαφορετικές επιπτώσεις στους οργανισμούς ανάπτυξης και τους πελάτες (Mittas & Angelis 2013 : 1).

1.1 Σύντομη Ιστορική Αναδρομή

Οι μέθοδοι ΕΚΛ μπορούν να χρησιμοποιηθούν για τον καθορισμό προϋπολογισμού, το σχεδιασμό και προγραμματισμό κάθε σταδίου, τον καθορισμό των απαιτήσεων τόσο σε εργαλεία όσο και σε ανθρωπινο δυναμικό για την ανάπτυξη ενός λογισμικού, καθώς και την βελτίωση ενός υφιστάμενου συστήματος. Οι ακριβείς εκτιμήσεις δινουν τη δυνατότητα στον οργανισμό ανάπτυξης να ταξινομήσει τα υπο-ανάπτυξη έργα ανάλογα με τη σημαντικότητα τους, την επεξεργασία τυχόν αλλαγών που μπορεί να προκύψουν καθώς και τον επανασχεδιασμό του έργου (Leung & Fan 2002: 1). Επιπρόσθετα, η διαχείριση του έργου είναι πιο εύκολη όταν οι διαθέσιμοι πόροι είναι σε συνάρτηση με τις πραγματικές ανάγκες, ενώ ο πελάτης θα αναμένει το πραγματικό κόστος να είναι παραπλήσιο του εκτιμώμενου (Leung & Fan 2002: 1). Πρόκειται να παρουσιάσουμε τις κυρίαρχες τεχνικές και πως αυτές υλοποιήθηκαν για την επίτευξη του σκοπού αυτού.

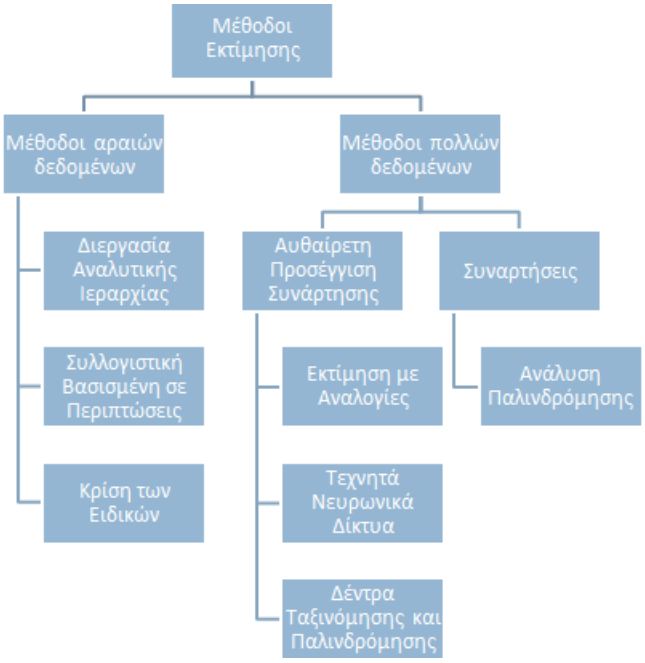
Το μέγεθος λογισμικού είναι ο πιο σημαντικός παράγοντας που επηρεάζει το κόστος λογισμικού (Leung & Fan 2002: 3). Οι κύριες μετρικές μεγέθους λογισμικού είναι οι γραμμές κώδικα και τα λειτουργικά σημεία. Οι γραμμές κώδικα αναφέρονται στις παραδοτέες γραμμές πηγαίου κώδικα λογισμικού, μη συμπεριλαμβανομένου σχολίων και κενών γραμμών (Leung & Fan 2002: 3). Η ακριβής μέτρηση των γραμμών κώδικα μπορεί να γίνει με την ολοκλήρωση του έργου. Για την εκτίμηση ενός κόστους λογισμικού θα πρέπει να χρησιμοποιήσουμε και μια εκτίμηση των γραμμών κώδικα. Η εκτίμηση αυτή συνήθως γίνεται με τεχνική κρίσης ειδικού, σε συνάρτηση με μια τεχνική που λέγεται PERT (Leung & Fan 2002: 3). Περιλαμβάνει κρίση ειδικού του χαμηλότερου, υψηλότερου και του μέσου πιθανού μεγέθους κώδικα. Η εκτίμηση μεγέθους κώδικα S υπολογίζεται ως εξής (Leung & Fan 2002: 3):

$S = \frac{S_l + S_h + 4S_m}{6}$ (Εξ. 1)

Η μέθοδος μέτρησης των λειτουργικών σημείων βασίζεται στη λειτουργία του προγράμματος και προτάθηκε από (Albrecht & Gaffney 1983). Ο συνολικός αριθμός λειτουργικών σημείων βασίζεται στις μετρήσεις διαφορετικών τύπων σε 5 κλάσεις: δεδομένων εισαγωγής από το χρήστη, δεδομένων εξαγωγής από το σύστημα στο χρήστη, δεδομένων αλληλεπίδρασης με το χρήστη, εσωτερικού τύπου τα οποία χρησιμοποιούνται από το σύστημα, εξωτερικού τύπου τα οποία χρησιμοποιούνται μεταξύ του συστήματος και άλλων συστημάτων (Leung & Fan 2002: 4). Για κάθε τύπο δεδομένων ορίζονται 3 επίπεδα πολυπλοκότητας {1= simple, 2=medium, 3=complex} και ένα επίπεδο βαρύτητας από 3 μέχρι 15 (για περίπλοκα εσωτερικά αρχεία) (Leung & Fan 2002: 4).Ο αριθμός των μη προσαρμοσμένων λειτουργικών σημείων δίνεται από:

$UFC = \sum_{i=1}^{5} \sum_{j=1}^{3} N_{ij} W_{ij}$ (Εξ. 2), (Leung & Fan 2002: 4), όπου N και W ο αριθμός και βάρος των τύπων της κλάσης i με πολυπλόκότητα j.

Οι μεθόδοι ΕΚΛ μπορούν να χωρισθούν σε 2 κύριες κατηγορίες, σε μεθόδους που κάνουν χρήση πολλών παλαιών δεδομένων και αραιών δεδομένων, δηλαδή μεθόδους που χρησιμοποιούν λίγα εώς καθόλου παλαιά δεδομένα (Myrtveit et al. 2005: 381). Στην πρώτη κατηγορία ανήκουν σχέσεις της γενικής μορφής, $y = ax^b$, δηλαδή μαθηματικές εξισώσεις που περιγράφουν τη συσχέτιση μεταξύ των μεταβλητών βάσει μίας προκαθορισμένης συνάρτησης (Myrtveit et al. 2005: 381). Ένα παράδειγμα αυτών είναι η ανάλυση παλινδρόμησης. Επιπλέον, περιλαμβάνουν μεθόδους στις οποίες δεν υπάρχει συσχέτιση μεταξύ των παραγόντων κόστους και εκτίμησης, όπως την εκτίμηση με αναλογίες (Estimation by Analogy-EbA), τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks-ANN) και τα δέντρα ταξινόμησης και παλινδρόμησης (Classification and Regression Trees-CART). Οι μεθόδοι αραιών δεδομένων μπορούν να κατηγοριοποιηθούν σε 3 κυρίως ομάδες όπως είναι η διεργασία αναλυτικής ιεραρχίας, η συλλογιστική βασισμένη σε περιπτώσεις και σε μεθόδους όπου γίνεται χρήση της κρίσης ειδικών.



Διάγραμμα 1. Διάγραμμα ταξινόμησης μεθόδων εκτίμησης (Myrtveit et al. 2005: 382)

Οι τεχνικές παλινδρόμησης είναι οι πιο δημοφιλείς για την κατασκευή μοντέλων και χρησιμοποιούνται σε συνδυασμό με τεχνικές βασισμένες σε μοντέλα. Η κανονική παλινδρόμηση βασίζεται στη μέθοδο κανονικών ελαχίστων τετραγώνων (Boehm et al. 2000: 31). Χρησιμοποιείται όταν έχουμε στη διάθεση μας αρκετά δεδομένα και ο αριθμός των παρατηρήσεων, είναι πολύ μεγαλύτερος από τον αριθμό των μεταβλητών πρόβλεψης (Boehm et al. 2000: 32). Δεν θα πρέπει να υπάρχουν ελλειπή δεδομένα και ακραίες τιμές, ενώ ανεξάρτητες μεταβλητές δεν θα πρέπει να συσχετίζονται (Boehm et al. 2000: 32).

Τα νευρωνικά δίκτυα είναι μοντέλα εκτίμησης, τα οποία μπορούν να εκπαιδευτούν χρησιμοποιώντας ιστορικά δεδομένα, για τη παραγωγή καλύτερων αποτελεσμάτων, προσαρμόζοντας αυτόματα τις τιμές των παραμέτρων για την μείωση των διαφορών, μεταξύ πραγματικών τιμών και των τιμών πρόβλεψης (Boehm et al. 2000: 25). Η ανάπτυξη ενός τέτοιου νευρωνικού δικτύου γίνεται αναπτύσσοντας ένα δίκτυο. Όταν το δίκτυο αναπτυχθεί, το μοντέλο θα πρέπει να εκπαιδευτεί παρέχοντας του ιστορικά δεδομένα και τις αντίστοιχες πραγματικές τιμές κόστους (Boehm et al. 2000: 25). Το μοντέλο επαναλαμβάνει τον αλγόριθμο εκπαίδευσης, προσαρμόζοντας αυτόματα τις παραμέτρους για τις συναρτήσεις εκτίμησης, μέχρι οι εκτιμήσεις και οι πραγματικές τιμές να βρίσκονται εντός προκαθορισμένης διαφοράς (Boehm et al. 2000: 25).

Οι επαγγελματίες χρησιμοποιούν συλλογιστική βασισμένη σε περιπτώσεις (CBR), για την έρευνα εργασιών που περιγράφουν συνθήκες και περιορισμούς που προέκυψαν κατά την ανάπτυξη προηγούμενων έργων λογισμικού, τις τεχνικές και διοικητικές αποφάσεις που πάρθηκαν και τις τελικές επιτυχίες ή αποτυχίες (Boehm et al. 2000: 24). Προσπαθούν να βρουν παρόμοιες περιπτώσεις, στις οποίες να αναλύσουν τις ενέργειες που εφαρμόσθηκαν και τα αποτελέσματα τους (Boehm et al. 2000: 24). Η συλλογιστική βασισμένη σε περιπτώσεις μπορεί να ανήκει τόσο στις μεθόδους αραιών δεδομένων όσο και στις μεθόδους πολλών δεδομένων (Myrtveit et al 2005: 381). Αν χρησιμοποιείται για τον εντοπισμό της κοντινότερης περίπτωσης, τότε είναι πολλών δεδομένων. Η μέθοδος εκτίμησης με αναλογία είναι ένα παράδειγμα μιας τέτοιας χρήσης CBR. Όταν γίνεται χρήση της CBR για μια ήδη επιλεγμένη περίπτωση, τότε είναι μέδοδος μεμονωμένων δεδομένων.

Οι τεχνικές βασισμένες στη κρίση ειδικού, συλλέγουν προηγούμενη γνώση και εμπειρία ειδικών ανάπτυξης λογισμικού και παρέχουν εκτιμήσεις βασισμένες σε μια σύνθεση γνωστών αποτελεσμάτων προηγούμενων έργων (Boehm et al. 2000: 20). Μια μορφή μιας τέτοιας τεχνικής είναι η Delphi, η τεχνική αυτή είναι χρήσιμη όταν δεν έχουμε εμπειρικά δεδομένα και θα πρέπει να εμπιστευτούμε κάποιο ειδικό (Boehm et al. 2000: 20). Στην αρχή ζητείται από τους συμμετέχοντες να γίνει μια εκτίμηση ενός θέματος, χωρίς να έχουν γνώση των απόψεων των υπολοίπων συμμετεχόντων (Boehm et al. 2000: 20). Στο δεύτερο γύρο, γίνεται επανεκτίμηση του θέματος, έχοντας υπ’ όψη τη γνώμη και γνώση των υπολοίπων συμμετεχόντων του πρώτου γύρου (Boehm et al. 2000: 25). Αυτό έχει ως αποτέλεσμα τον περιορισμό των εκτιμήσεων των συμμετεχόντων σε σχέση με το θέμα.

Η μέθοδος Αναλυτικής Ιεραρχίας είναι ένας τρόπος ιεράρχησης των στοιχείων ενός έργου, για την εκτίμηση και έλεγχο του προϋπολογισμού (Boehm et al. 2000: 21). Περιγράφει τη βασική δομή λογισμικού και τα κόστη που θα πρέπει να εκτιμηθούν. Μια Αναλυτική Ιεραρχία λογισμικού αποτελείται από δυο ιεραρχίες, μια που αναπαριστά το προϊόν λογισμικού και την άλλη που αναπαριστά τις απαιτούμενες δραστηριότητες για τη δημιουργία λογισμικού (Boehm et al. 2000: 22).

Η εκτίμηση κόστους λογισμικού συνιστάται από μια πληθώρα τεχνικών, εκ των οποίων αναφέραμε πιο πάνω. Κάθε τεχνική έχει ειδική χρήση και καμία δεν θα πρέπει να αγνοείται. Για να καταλήξουμε σε ασφαλή συμπεράσματα, θα πρέπει να γίνεται χρήση πληθώρας τεχνικών, και να ερευνηθούν οι λόγοι για τους οποίους υπάρχει σημαντική διαφορά μεταξύ

τους. Η ανάλυση αυτή, μπορεί να μας βοηθήσει στον εντοπισμό των παραγόντων που επηρεάζουν το κόστος.

1.2 Πολιτικές Σύγκρισης Μεθόδων Πρόβλεψης και Συμπερασματική Αστάθεια

Ο μεγάλος αριθμός ερευνητικών μελετών, στον τομέα της ΕΚΛ, έχει αναδείξει το πρόβλημα της συμπερασματικής αστάθειας. Το φαινόμενο αυτό αναφέρεται στην ασυνέπεια των αποτελεσμάτων που έχουν προκύψει από τον μεγάλο αριθμό εκτιμητών, οι οποίοι κάνουν χρήση διαφορετικών συνόλων δεδομένων (Myrtveit et al 2005: 381). Κατά τις δύο τελευταίες δεκαετίες, αρκετοί ερευνητές προσπαθούν να εντοπίσουν τις πηγές του φαινομένου και να αποτιμήσουν τις συνέπειες που έχει στη διαδικασία εκτίμησης.

Η έλλειψη καθορισμένης ερευνητικής μεθοδολογίας στην έρευνα λογισμικού, μπορεί να οδηγήσει σε ετερογενή δείγματα, μετρήσεις και τεχνικές αναφοράς (Mair & Shepperd 2005: 7). Πολλοί ερευνητές χειρίζονται διαφορετικά τις ακραίες τιμές, υπάρχει διαγραφή παρατηρήσεων και μεταβλητών (Myrtveit & Stensrud 2012: 25) και έτσι η επαλήθευση της αξιολόγησης του υποψήφιου μοντέλου, πολλές φορές, είναι δύσκολη εώς αδύνατη (Kitchenham & Mendes 2009: 2).

Έχει επισημανθεί, ότι το κύριο πρόβλημα με τις τεχνικές αξιολόγησης των συστημάτων πρόβλεψης είναι ότι τα σύνολα δεδομένων που χρησιμοποιεί ο κάθε ερευνητής, πολλές φορές, επιλέγονται μεροληπτικά σε σχέση με το αν είναι καλύτερα για μια συγκεκριμένη τεχνική (Kitchenham & Mendes 2009: 2). Αυτό όμως σημαίνει ότι η μέθοδος αξιολόγησης δεν είναι αμερόληπτη και έγκυρη. Επί πλέον, κάποια σύνολα δεδομένων δεν παραμένουν αμετάβλητα με το χρόνο ή μπορεί να έχουν ελλειπείς τιμές (Kitchenham & Mendes 2009: 2). Αυτό καθιστά αδύνατη την επανάληψη της πειραματικής διαδικασίας (Kitchenham & Mendes 2009: 2).

Ένα άλλο σημαντικό ζήτημα, είναι εάν πρέπει να συγκρίνουμε προβλέψεις βασισμένες σε ολόκληρο το σύνολο δεδομένων ή σε προβλέψεις οι οποίες χωρίζουν το σύνολο δεδομένων, σε σύνολα εκπαίδευσης (training set) και ελέγχου (test set) (Kitchenham & Mendes 2009: 2). Πολλοί ερενητές συμφωνούν ότι η τελευταία τεχνική είναι καλύτερη,

αλλά αν χρησιμοποιήσουμε κάτι άλλο από μια απλή leave-one-out cross-validation (LOOCV) διαδικασία, τα αποτελέσματα δεν είναι ελέγξιμα, αν δεν καθοριστούν τα μέρη των συγκεκριμένων συνόλων δεδομένων (Kitchenham & Mendes 2009: 2). Το πρόβλημα επιδεινώνεται ακόμα περισσότερο, αν λάβουμε υπόψη ότι η διαδικασία επικύρωσης (validation process) των προτεινόμενων μεθοδολογιών πραγματοποιείται σε μεμονωμένα σύνολα δεδομένων (single dataset) ή σε ένα πολύ μικρό αριθμό συνόλων δεδομένων (Kitchenham & Mendes 2009: 2) καθιστώντας τη γενίκευση των αποτελεσμάτων μία πολύ απαιτητική διαδικασία.

Οι (Menzies and Shepperd 2012: 5) αναφέρουν ότι υπάρχουν δυο κύριες αιτίες για την ύπαρξη του φαινομένου της συμπερασματικής αστάθειας, η μεροληψία και η διασπορά. Η μεροληψία μετρά την απόκλιση των εκτιμήσεων από τις πραγματικές τιμές. Ένας μεροληπτικός εκτιμητής θα παρουσιάζει απόκλιση των εκτιμήσεων του σταθέρα προς μια κατεύθυνση από τις πραγματικές τιμές. Αυτό μπορεί να προκύψει και από τη διαδικασία επαλήθευσης, όταν δεν γίνεται διαχωρισμός των συνόλων επικύρωσης από τα σύνολα προσαρμογής και συνεπεία αυτού να υπάρχουν σφάλματα υποεκτίμησης (Menzies and Shepperd 2012: 5). Κατά τη διαδικασία επιλογής θα πρέπει να λάβουμε υπόψη και τη διασπορά του σφάλματος πρόβλεψης. Αν έχουμε υψηλή διασπορά τότε είναι πολύ πιθανό να καταλήξουμε σε συμπερασματική αστάθεια, ειδικά όταν το σύνολο δεδομένων είναι μικρού μεγέθους (Menzies and Shepperd 2012: 5). Η υψηλή διασπορά μπορεί να μειωθεί επαναλαμβάνοντας τη διαδικασία επαλήθευσης αρκετές φορές (Menzies and Shepperd 2012: 5).

Ένα άλλο παράδειγμα, τέτοιων ευρημάτων σε αντίφαση, μπορούμε να δούμε από τους (Mair & Shepperd 2005: 4), όπου οι ερευνητές συνδύασαν τα ευρήματα από 20 ξεχωριστές μελέτες, από το 1997 μέχρι και το 2004. Στις μελέτες αυτές χρησιμοποιήθηκαν δυο μεθόδοι, η παλινδρόμηση και η εκτίμηση με ανάλογία (Estimation by Analogy-EbA). Οι συγγραφείς αποφάνθηκαν ότι στις 12 από τις 20 μελέτες η EbA λειτούργησε καλύτερα από τη παλινδρόμηση. Επίσης επεσήμαναν ότι μόνο 6 από τις 20 μελέτες (30%), έκαναν χρήση ελέγχου στατιστικών υποθέσεων ώστε να αξιολογήσουν την δυνατότητα πρόβλεψης των δυο συγκρινόμενων μοντέλων. Οι (Kitchenham and Mendes 2009: 5) υπέδειξαν ότι δεν είναι έγκυρο να αποφανθούμε για την ανωτερότητα μιας μεθόδου χωρίς την διεξαγωγή στατιστικών ελέγχων, για την διευρεύνηση της σημαντικότητας της διαφοράς στις τιμές των μέτρων ακρίβειας.

1.3 Συνεισφορά Μεταπτυχιακής Διατριβής & Επίλυση του προβλήματος της Αξιολόγησης των Εναλλακτικών Μοντέλων

Σε αυτήν τη μεταπτυχιακή Διατριβή, πρόκειται να μελετηθεί και να επεκταθεί η χρήση της Regression Receiver Operating Curves (RROC) ανάλυσης, για τη διερεύνηση της προβλεπτικής ικανότητας διαφόρων αλγορίθμων εκτίμησης κόστους και της εξαγωγής συμπερασμάτων σχετικά με τη μεροληψία (bias) και τη διασπορά (variance) του σφάλματος πρόβλεψης. Συνοπτικά, ο RROC space αναπαριστά την ικανότητα πρόβλεψης εναλλακτικών μοντέλων σε ένα δισδιάστατο γράφημα, όπου απεικονίζεται η αποδόμηση του σφάλματος πρόβλεψης υπό τη μορφή της υποεκτίμησης και υπερεκτίμησης. Με το γράφημα αυτό, δίνεται η δυνατότητα στους διοικητές έργων να αποκτήσουν σημαντικές πληροφορίες, οι οποίες θα τους οδηγήσουν με μεγαλύτερη ασφάλεια στη διαδικασία λήψης αποφάσεων, για το ερευνητικό πρόβλημα της σύγκρισης και επιλογής της καταλληλότερης μεθόδου για ένα συγκεκριμένο σύνολο δεδομένων.

Πιο συγκεκριμένα, στην μεταπτυχιακή Διατριβή στόχος είναι η μελέτη της μεταβλητότητας του μη ισοζυγισμένου σφαλμάτος πρόβλεψης. Στην προσπάθεια αυτή, προτείνεται μία παρόμοια προσέγγιση με εκείνη των (Mittas & Angelis 2016: 4), οι οποίοι εισάγουν τη χρήση της bootstrap αναδειγματοληψίας (resampling) για τη μελέτη των σφαλμάτων πρόβλεψης και της μεροληψίας των εναλλακτικών μοντέλων πρόβλεψης μέσω της RROC ανάλυσης. Η διαφορά της παρούσας μεταπτυχιακής Διατριβής με την προτεινόμενη μεθοδολογία των (Mittas & Angelis 2016: 4) έγκειται στον τρόπο διεξαγωγής της αναδειγματοληψίας και της εκτίμησης της κατανομής των σφαλμάτων. Αναλυτικότερα, οι (Mittas & Angelis 2016: 4) προτείνουν τη χρήση της bootstrap αναδειγματοληψίας των περιπτώσεων (cases) του συνόλου δεδομένων και την προσαρμογή ενός μεγάλου αριθμού μοντέλων στα bootstrap δείγματα, από όπου και υπολογίζονται τα τελικά διαστήματα εμπιστοσύνης για τα σφάλματα υποεκτιμήσης και υπερεκτίμησης. Στην μεταπτυχιακή Διατριβή, προτείνεται η χρήση της bootstrap αναδειγματοληψίας στις κατανομές των δύο τύπων σφάλματος, τα οποία προέρχονται από τη διαδικασία επικύρωσης του εκάστοτε μοντέλου με την τεχνική LOOCV και κατασκευή διαστημάτων εμπιστοσύνης με βάση τα bootstrap δείγματα για εξαγωγή περαιτέρω ασφαλών συμπερασμάτων.

Λόγω της ραγδαίας ανάπτυξης του τομέα της ΕΚΛ, οι ερευνητές έχουν ανάγκη εργαλείων εύκολων στη χρήση και προσβάσιμων σε αυτούς, ούτως ώστε να έχουν τη δυνατότητα διεξαγωγής έγκυρης και άμεσης σύγκρισης. Πρόκειται να αναπτύξουμε μια web-based εφαρμογή όπου θα διεξάγεται γραφική σύγκριση των εναλλακτικών μοντέλων πρόβλεψης, έτσι ώστε ο χρήστης να μπορεί να εξάγει εύκολα και γρήγορα συμπεράσματα σε σχέση με κάθε συγκρινόμενο μοντέλο.

Η κύρια συνεισφορά μπορεί να συνοψιστεί στα πιο κάτω θέματα:

- Ανάλυση και περιγραφή της RROC ανάλυσης και εφαρμογή της στην ΕΚΛ.
- Γραφική αξιολόγηση και σύγκριση της προβλεπτικής ικανότητας των εναλλακτικών μοντέλων στον RROC space.
- Χρησιμοποίηση τεχνικών bootstrap αναδειγματολειψίας, για τον εντοπισμό της εμπειρικής κατανομής, που ακολουθούν οι συναρτήσεις σφάλματος των μοντέλων και εύρεση των διαστημάτων εμπιστοσύνης στον RROC space.
- Ανάπτυξη web-based εφαρμογής, όπου ο χρήστης εισάγοντας ένα αρχείο συνόλου δεδομένων μπορεί να διεξάγει δυαδικτυακά σύγκριση εναλλακτικών μοντέλων.

Κεφάλαιο 2

Αξιολόγηση Μοντέλων Πρόβλεψης

2.1 Πολιτικές και μεθοδολογίες σύγκρισης μοντέλων

Η ΕΚΛ είναι ένα ερευνητικό πεδίο που βρίσκεται υπό ανάπτυξη και τις τελευταίες δεκαετίες, η βιβλιογραφία που σχετίζεται με το θέμα αυτό έχει αυξηθεί σε μεγάλο βαθμό (Jorgensen & Shepperd 2007: 43). Στις μελέτες αυτές, η ανάγκη για ακριβείς και έγκυρες εκτιμήσεις για ένα μελλοντικό έργο, είναι το κεντρικό ερευνητικό θέμα. Παρά την εξέλιξη και ανάπτυξη διάφορων τεχνικών, δεν υπάρχει κάποια συγκεκριμένη τεχνική η οποία να μπορεί να θεωρηθεί ως ανώτερη σε κάθε περίπτωση.

Κατά τη διαδικασία αξιολόγησης και σύγκρισης των εναλλακτικών μοντέλων πρόβλεψης, οι ερευνητές θα πρέπει να είναι σε θέση να απαντήσουν τρία βασικά ερωτήματα (Shepperd & MacDonell 2012: 2). Θα πρέπει να μελετηθεί κατά πόσο ένα σύστημα πρόβλεψης συμπεριφέρεται καλύτερα από κάποια βάση, τυχαίας τιμής πρόβλεψης. Εάν δεν συμπεριφέρεται καλύτερα από κάποια τυχαία τιμή πρόβλεψης, τότε δεν θα πρέπει καν να θεωρούμε ότι διεξάγεται πρόβλεψη (Shepperd & MacDonell 2012: 2). Επιπρόσθετα, η διαφορά στην ακρίβεια πρόβλεψης μεταξύ των εναλλακτικών συστημάτων πρόβλεψης, δεν θα πρέπει να οφείλεται στη τύχη αλλά να είναι στατιστικά σημαντική με βάση κάποια προκαθορισμένη τιμή α (Shepperd & MacDonell 2012: 3). Το αποτέλεσμα της πρακτικής εφαρμογής κάποιου μοντέλου σε σχέση με κάποιο άλλο θα πρέπει να μελετηθεί, ώστε να έχουμε ξεκάθαρα αποτέλεσματα ως προς την διάφορά απόδοσης τους (Shepperd & MacDonell 2012: 3).

Οι (Myrtveit et al. 2005: 381) αναφέρουν ότι η διαδικασία αξιολόγησης στις πλείστες ερευνητικές μελέτες γίνεται ως εξής: τα μοντέλα πρόβλεψης επικυρώνονται σε κάποιο σύνολο δεδομένων με βάση τη διαδικασία επικύρωσης cross-validation. Η διαδικασία Cross-Validation είναι ένας τρόπος εξαγωγής αμερόληπτων εκτιμητών του σφάλματος

πρόβλεψης (Myrtveit et al. 2005: 383). Η διαδικασία αυτή συνιστά τη διαγραφή μιας παρατήρησης κάθε φορά και τη προσαρμογή ενός μοντέλου στις υπόλοιπες n-1 (Myrtveit et al. 2005: 383). Η διαδικασία Cross-Validation μπορεί να διεξαχθεί, είτε χωρίζοντας το σύνολο δεδομένων σε n-1 παρατηρήσεις είτε σε k ίσες διαμερίσεις παρατηρήσεων (Myrtveit et al. 2005: 383). Δηλαδή αν χρησιμοποιούμε τη διαδικασία k-folds Cross-Validation τότε θα πρέπει να έχουμε k ίσες διαμερίσεις μεγέθους t παρατηρήσεων και σε κάθε επανάληψη της διαδικασίας προσαρμογής να διαγράφεται μια διαμέριση μεγέθους t (Myrtveit et al. 2005: 383). Ο ερευνητής, κάνοντας χρήση ενός τοπικού μέτρου ακρίβειας, βρίσκει τη ικανότητα πρόβλεψης ενός προσαρμοσμένου μοντέλου για τη διεγραμμένη παρατήρηση. Στη συνέχεια βρίσκει τη μέση ακρίβεια πρόβλεψης κάνοντας χρήση ενός καθολικού μέτρου ακρίβειας. Το μοντέλο με τη υψηλότερη ακρίβεια, δηλαδή τη χαμηλότερη τιμή ενός μέτρου ακρίβειας θεωρείται ως το καλύτερο.

Η πιο γνωστή μέθοδος σύγκρισης των εναλλακτικών μοντέλων είναι η χρήση των μέτρων ακρίβειας, τα οποία ουσιαστικά είναι συναρτήσεις του σφάλματος πρόβλεψης (prediction error) ανάμεσα στην πραγματική (actual) και την εκτιμώμενη (predicted) τιμή (Myrtveit et al. 2005: 383). Ένα παράδειγμα μιας τέτοιας αξιολόγησης με βάση τα μέτρα ακρίβειας είναι των Foss et al. (2003: 990, 991). Οι ερευνητές εφήρμοσαν 5 μοντέλα πρόβλεψης σε 1000 σύνολα δεδομένων, τα μοντέλα αξιολογήθηκαν με βάση τα μέτρα ακρίβειας MMRE, MdMRE, MMER, SD, RSD, LSD, MBRE και MIBRE. Έγινε σύγκριση κάθε μοντέλου με το πραγματικό, ενώ για κάθε σύγκριση μέτρησαν πόσες φορές κάθε μοντέλο παρουσίασε τις καλύτερες τιμές MMRE, MdMRE, MMER, SD, RSD, LSD, MBRE και MIBRE αντίστοιχα. Έδειξαν ότι το MMRE θεωρεί ως καλύτερα μοντέλα που υποεκτιμούν το πραγματικό κόστος, το MdMRE παρουσιάζει παρόμοια αποτελέσματα ενώ το MMER εντόπισε το καλύτερο μοντέλο σε 3 από τις 4 περιπτώσεις. Τα μέτρα MBRE και MIBRE παρουσίασαν παρόμοια αποτελέσματα με το MMER. Το μέτρο ακρίβειας τυπικής απόκλισης (SD) εντόπισε το πραγματικό μοντέλο ως καλύτερο κάθε φορά. Η τυπική απόκλιση σχετικού σφάλματος (RSD) θεώρησε ως καλύτερο το πραγματικό μοντέλο σε σχέση με τα υπόλοιπα 4 με μεγαλύτερη πιθανότητα από το SD. Η λογαριθμική τυπική απόκλιση (LSD) εντόπισε το πραγματικό μοντέλο με αρκετά υψηλή πιθανότητα (>0.7). Ένα άλλο παράδειγμα αξιολόγησης με βάση τα μέτρα ακρίβειας είναι των (Shepperd and Schofield 1997: 740), οι οποίοι συνέκριναν τη χρήση μεθόδων με αναλογία, με μοντέλα παλινδρόμησης για 9 σύνολα δεδομένων. Η εκτίμηση με αναλογία παρουσίασε τα καλύτερα αποτελέσματα σε όλες τις περιπτώσεις, όταν η σύγκριση έγινε με βάση το μέτρο MMRE, ενώ με βάση το pred(25) στις 7 από τις 9 περιπτώσεις. Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί σε περιπτώσεις όπου δεν μπορούμε να κατασκευάσουμε ένα αλγοριθμικό μοντέλο, όπως για παράδειγμα όταν όλα τα δεδομένα είναι διακριτής μορφής. Επιπρόσθετα, η χρήση της

συνηθίζεται στα αρχικά στάδια ανάπτυξης ενός έργου, όταν δεν υπάρχουν στατιστικά σημαντικές σχέσεις μεταξύ των δεδομένων.

Οι Jeffery et al. (2001: 4) διερεύνησαν την ακρίβεια των τεχνικών ordinary least squares regression, της εκτίμησης με Αναλογία, stepwise ANOVA, CART, robust regression και τα πλεονεκτήματα της χρήσης συνόλων δεδομένων που στηρίζονται στην ίδια εταιρεία και σε δεδομένα διάφορων εταιρείών με βάση το ISBSG. Όταν χρησιμοποιήθηκαν τα δεδομένα της υπό μελέτη εταιρείας, οι τεχνικές OLS, CART και η εκτίμηση με Αναλογία είχαν τα καλύτερα αποτελέσματα. Παράλληλα, οι (Shepperd and Kadoda 2001: 1015) διεξήγαγαν σύγκριση 4 τεχνικών πρόβλεψης, με σκοπό να διερευνήσουν ποια χαρακτηριστικά συνόλων δεδομένων, ευνοούν τις τεχνικές για την ανάπτυξη των μοντέλων πρόβλεψης. Η μέθοδος που χρησιμοποίησαν ήταν η τεχνητή δημιουργία συνόλων δεδομένων, χωρίς την γνώση των χαρακτηριστικών ανάπτυξης τους (Shepperd and Kadoda 2001: 1017), λόγω του ότι η χρήση πραγματικών συνόλων δεδομένων θα μπορούσε να υποσκελίσει την ανάλυση. Σε πραγματικά συνόλα δεδομένων δεν είναι ξεκάθαρο σε τι βαθμό υπάρχουν τα υπό μελέτη χαρακτηριστικά, ενώ με τη χρήση τεχνητών δεδομενών υπάρχει η δυνατότητα συνδυασμού των υπό μελέτη παραγόντων (Shepperd and Kadoda 2001: 1017). Επιπρόσθετα, η μη γνώση του 'πραγματικού' μοντέλου καθιστά την επεξεργασία σε μικρά σύνολα δεδομένων ακόμα πιο περίπλοκη (Shepperd and Kadoda 2001: 1017). Ένας άλλος λόγος χρήσης τεχνητών συνόλων δεδομένων είναι η δυνατότητα κατασκευής μεγάλων συνόλων επικύρωσης (validation sets).

Τα σύνολα δεδομένων συνήθως είναι μικρού μεγέθους και τα σφάλματα πρόβλεψης εμφανίζουν συχνά μεγάλη λοξότητα (skewness) (Mittas & Angelis 2008a: 617). Μικρές αλλαγές στα σύνολα δεδομένων μπορεί να επηρεάσουν δραστικά τα αποτελέσματα σύγκρισης. Για την εξάλειψη των περιορισμών αυτών, οι (Mittas και Angelis 2009: 223) προτείνουν τη χρήση τεχνικών αναδειγματολειψίας από το αρχικό δείγμα. Συγκεκριμένα προτείνουν τη χρήση της μη παραμετρικής bootstrap, η οποία δεν κάνει υποθέσεις για τη κατανομή του δείγματος. Από ένα δείγμα μεγέθους n, γίνεται εξαγωγή δειγμάτων με n περιπτώσεις με αντικατάσταση (sampling with replacement), για τον υπολογισμό της άγνωστης στατιστικής παραμέτρου $\hat{\theta}$ (π.χ μέσο ή διάμεσο των σφαλμάτων) (Mittas & Angelis 2009: 223). Η εμπειρική κατανομή (empirical distribution) που εξάγεται με τη μέθοδο bootstrap, μπορεί να χρησιμοποιηθεί για τον υπολογισμό του τυπικού σφάλματος, της μεροληψίας και τον υπολογισμό διαστημάτων εμπιστοσύνης για την άγνωστη παράμετρο $\hat{\theta}$ του πληθυσμού (Mittas & Angelis 2009: 223). Εάν τα δυο διαστήματα

εμπιστοσύνης δεν εμπεριέχουν τμήματα τους, το ένα σε σχέση με το άλλο, τότε υπάρχει στατιστικά σημαντική διαφορά σε σχέση με τα δυο μοντέλα (Mittas & Angelis 2009: 227).

2.2 Μεθοδολογίες γραφικής σύγκρισης μοντέλων

Οι παραπάνω πολιτικές σύγκρισεις έχουν συντελέσει αναμφισβήτητα στον περιορισμό, σε ένα βαθμό του φαινομένου της συμπερασματικής αστάθειας, καθώς παρέχουν χρήσιμες πληροφορίες που μπορούν να ληφθούν υπόψη στη διαδικασία επιλογής του καταλληλότερου μοντέλου πρόβλεψης. Από την άλλη μεριά, όμως, και παρά το γεγονός, ότι στατιστικοί έλεγχοι υποθέσεων συνιστούν μια ανεκτίμητη διαδικασία για τη σύγκριση των μοντέλων πρόβλεψης, πολλές φορές παραβλέπονται για χάρη απλότητας λόγω του γεγονότος ότι είναι δύσκολο να ερμηνευθούν από μη ειδικούς (Mittas & Angelis 2010: 622). Χωρίς αμφιβολία, οι διοικητές των έργων επιθυμούν να λάβουν σε πολύ σύντομο χρονικό περιθώριο σημαντικές αποφάσεις για την ανωτερότητα ενός μοντέλου πρόβλεψης σε σχέση με άλλα εναλλακτικά μοντέλα και για τον λόγο αυτόν επιθυμούν εργαλεία που οπτικοποιούν τις αποδόσεις των μοντέλων.

Προς την κατεύθυνση της οπτικής σύγκρισης της απόδοσης εναλλακτικών μοντέλων, οι (Kitchenham et al. 2001: 83) προτείνουν τη χρήση θηκογραμμάτων (boxplots) για τις κατανομές της συνάρτησης σφάλματος z ($z = \frac{Y_{E_i}}{Y_{A_i}}$) που σχετίζεται με τη μεροληψία των μοντέλων. Αν η διάμεσος είναι κοντά στο 1, οι προβλέψεις είναι αμερόληπτες (Kitchenham et al., 2001) ενώ τιμές μεγαλύτερες ή μικρότερες της μονάδας υποδεικνύουν υπερεκτίμηση και υποεκτιμήση, αντίστοιχα. Παράλληλα, τα θηκογράμματα φανερώνουν την ύπαρξη μεγάλης διασποράς ή λοξότητας της κατανομής και γενικά, παρέχουν ένα απλό μέσο για τη σύγκριση των προβλέψεων των εναλλακτικών μοντέλων. Παράλληλα, με τη οπτικοποίηση των σφαλμάτων μέσω του θηκογράμματος, οι (Kitchenham et al. 2001: 83) προτείνουν και τη χρήση του στατιστικού ελέγχου υποθέσεων t-test για να ελεγχθεί αν το απόλυτο σχετικό σφάλμα για κάθε σημείο δεδομένων που προκύπτει από ένα σημείο πρόβλεψης, είναι σημαντικά καλύτερο από ένα άλλο.

Οι Mittas & Angelis (2008c: 2) προτείνουν ένα εργαλείο οπτικοποίησης των ικανοτήτων πρόβλεψης των εναλλακτικών μοντέλων ΕΚΛ, τις καμπύλες Regression Error Characteristic (REC). Το εργαλείο αυτό παρουσιάζει μια εκτίμηση της συνάρτησης αθροιστικής κατανομής (cumulative distribution) του προβλεπόμενου σφάλματος. Ουσιαστικά είναι ένα

δισδιάστατο γράφημα, όπου ο οριζόντιος (ή x-άξονας) αναπαριστά την ανοχή σφάλματος (error tolerance) ενός προκαθορισμένου μέτρου ακρίβειας και ο κάθετος (ή y-άξονας), την ακρίβεια (accuracy) του μοντέλου πρόβλεψης (Bij et al. 2003: 43). Η ακρίβεια ορίζεται ως η αναλογία ή το ποσοστό των έργων που παρουσιάζουν σφάλμα πρόβλεψης μικρότερο ή ίσο από την προκαθορισμένη τιμή ανοχής σφάλματος (Mittas & Angelis 2008c: 2). Παρά το γεγονός, ότι REC καμπύλες παρέχουν σημαντικές πληροφορίες δεν είναι ικανές να ελέγξουν, κατά πόσο υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των εναλλακτικών συναρτήσεων σφάλματος που παράγονται από διάφορα μοντέλα. Για τον λόγο αυτόν, οι Mittas & Angelis (2012: 41) κάνουν χρήση του στατιστικού μέτρου Kolmogorov-Smirnov (KS) και της τεχνικής αντιμετάθεσης (permutation) προτείνοντας έναν στατιστικό έλεγχο υποθέσεων που συγκρίνει τις REC καμπύλες δύο εναλλακτικών μοντέλων πρόβλεψης. Ο προτεινόμενος αλγόριθμος χρησιμοποιεί αντιμετάθεση των κατανομών του σφάλματος με επαναληπτικό τρόπο και υπολογίζει σε κάθε επανάληψη το στατιστικό μέτρο KS εξάγοντας τελικά τη στάθμη σημαντικότητας του προτεινόμενου ελέγχου υποθέσεως.

Οι ίδιοι ερευνητές εισάγουν και τη χρήση της μερικής REC καμπύλης (Partial REC curve), η οποία επεκτείνει τα πλεονεκτήματα των καμπύλων REC, σε ένα συγκεκριμένο διάστημα των πραγματικών τιμών κόστους (Mittas & Angelis 2008b: 1). Οι καμπύλες αυτές είναι εύκολες στην ανάγνωση και ερμηνεία, και δείχνουν τη στρατηγική που πρέπει να ακολουθηθεί, για μια καλύτερη διαχείριση ενός έργου λογισμικού. Το πιο σημαντικό πλεονέκτημα που παρέχει η ανάλυση REC, είναι ότι όλα τα μέτρα ακριβείας, όπως τα MMRE, MdMRE και Pred(p), παρουσιάζονται με βάση τα γεωμετρικά χαρακτηριστικά και τις ιδιότητες των καμπύλων REC (Mittas & Angelis 2010: 636). Η ανάλυση αυτή είναι ικανή να επικεντρωθεί σε ορισμένα υποσύνολα του προβλεπόμενου σφάλματος. Αυτό είναι πολύ χρήσιμο, αφού μπορεί να αποδειχθεί ότι ένα μοντέλο πρόβλεψης, είναι καλύτερο για ορισμένα είδη έργων, στα οποία αντιστοιχεί ένα ορισμένο πλαίσιο των πραγματικών τιμών κόστους. Ακόμα οι καμπύλες REC, παρέχουν πολύτιμες πληροφορίες για τη συμμετρία της κατανομής του προβλεπόμενου σφάλματος και την ύπαρξη ακραίων τιμών (Mittas & Angelis 2010: 637). Μπορούν να χρησιμοποιηθούν για τη ρύθμιση της μεθοδολογίας πρόβλεψης, τον εντοπισμό των παραγόντων που επηρεάζουν την ακρίβεια πρόβλεψης, και τη διερεύνηση της για ορισμένες πραγματικές τιμές κόστους (Mittas & Angelis 2010: 637).

Όλα τα παραπάνω ερευνητικά ζητήματα που αφορούν την οπτικοποίηση των σφαλμάτων πρόβλεψης μαζί με ένα προτεινόμενο πλαίσιο εργασίας (framework) παρουσιάζονται σε ένα εργαλείο που αναπτύχθηκε από τους (Mittas et al. 2015: 10) και ονομάζεται StatREC. Το εργαλείο αυτό, είναι ένα λογισμικό το οποίο βασίζεται στη γλώσσα R (R Core Team 2016)

και υλοποιεί ορισμένες στατιστικές διαδικασίες, ικανές να παρέχουν ένα πλαίσιο για αποτελεσματική σύγκριση των διαφόρων τεχνικών. Σχεδιάστηκε με τέτοιο τρόπο που ο χρήστης απλά εισάγει ένα αρχείο, το οποίο περιέχει τις προβλέψεις πολλαπλών μοντέλων, χωρίς την ανάγκη της γνώσης αλγοριθμικών ή μαθηματικών λεπτομερειών για το πως σχεδιάστηκαν τα μοντέλα αυτά. Στη συνέχεια, είναι ικανό να διεξάγει στατιστική και γραφική ανάλυση, στις μετρικές σφάλματος που προκύπτουν από απλούς υπολογισμούς μεταξύ των προβλεπόμενων και πραγματικών τιμών. Παρέχει ένα φάσμα λειτουργιών, οι οποίες χρησιμοποιούνται για τη σύγκριση δύο ή περισσότερων μοντέλων πρόβλεψης. Πέρα από τις βασικές λειτουργίες, για αυτόματη αξιολόγηση των συναρτήσεων σφάλματος και σύγκριση με ένα μοντέλο αναφοράς (baseline model), παρέχει εντοπισμό των παραγόντων που επηρεάζουν τα σφάλματα, ερευνά πώς διαφέρουν τα σφάλματα σε ένα συγκεκριμένο πλαίσιο των τιμών κόστους και εντοπίζει αν το μοντέλο πρόβλεψης έχει τάση να κάνει υπερτιμήσεις ή υποεκτιμήσεις κόστους.

Αναλυτικότερα, οι (Mittas et al. 2015: 8) χρησιμοποιούν μια διαδικασία LOOCV, και στη συνέχεια εισάγουν τις πραγματικές και τις τιμές εκτίμησεις στο StatREC. Η μεθοδολογία που χρησιμοποιούν εντοπίζεται στα πιο κάτω σημεία. Ερευνούν κατά πόσο οι διαφορές των ανταγωνιστικών μοντέλων οφείλονται στη τύχη, κάνοντας χρήση του αλγόριθμου Scott-Knott στα μέτρα ακριβείας (Mittas et al. 2015: 9). Τα μοντέλα πρόβλεψης ομαδοποιούνται σε πέντε ομάδες, οι οποίες έχουν παρόμοια ικανότητα πρόβλεψης (Mittas et al. 2015: 9). Ελέγχονται οι διαφορές μεταξύ των μοντέλων και αν συνιστούν σημαντική διαφορά. Εντοπίζεται η χειρότερη ομάδα πρόβλεψης και συγκρίνεται με το χειρότερο μοντέλο (Mean Model). Αν δεν υπάρχει σημαντική διαφορά μεταξύ τους, η ομάδα αυτή διαγράφεται από τα ανταγωνιστικά μοντέλα. Το μοντέλο του μέσου ορίζεται από τη πιο κάτω συνάρτηση:

$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ (Εξ. 3) (Mittas et al. 2015: 5)

Το μοντέλο αυτό θεωρείται ως το χειρότερο, εφόσον η τιμή κάθε περίπτωσης, εκτιμάται από τις μέσες τιμές όλων των περιπτώσεων στο σύνολο δεδομένων χωρίς να λαμβάνουμε υπ' όψη τις ανεξάρτητες μεταβλήτες κάποιου μοντέλου (Mittas et al. 2015: 5). Το μοντέλο του μέσου έχει το μεγαλύτερο αρνητικό συντελεστή συσχέτισης ($r = -1$) σε σχέση με τις πραγματικές τιμές της συνάρτησης κόστους (Mittas et al. 2015: 5).

Για την επιλογή της καλύτερης ομάδας μοντέλων πρόβλεψης, το StatREC χρησιμοποιεί bootstrap resampling για τον εντοπισμό του τυπικού σφάλματος, της μεροληψίας και των διαστημάτων εμπιστοσύνης (Mittas et al. 2015: 7). Χρησιμοποιούνται οι Partial REC curves για τον εντοπισμό της ανωτερότητας των ανταγωνιστικών μοντέλων, σε συγκεκριμένα

πλαίσια όπως το μέγεθος έργου, ενώ παράλληλα γίνεται έλεγχος σχετικά με την ύπαρξη μεροληψίας προς υπερεκτίμηση ή υποεκτίμηση κόστους (Mittas et al. 2015: 16).

2.3 Αξιοπιστία και εγκυρότητα των ερευνητικών εργασιών στην ΕΚΛ

Όπως αναφέραμε και προηγουμένως, η αξιολόγηση των μοντέλων ΕΚΛ στις πλείστες ερευνητικές μελέτες γίνεται με βάση των μέτρων ακρίβειας. Τα μέτρα αυτά παρουσιάζουν μια γενική τάση της ικανότητας πρόβλεψης για ολόκληρο το σετ δεδομένων, χωρίς να λαμβάνονται υπόψη οι κατανομές του σφάλματος πρόβλεψης. Ένα μοντέλο πρόβλεψης Α μπορεί να λειτουργεί καλύτερα από κάποιο άλλο Β για κάποιες υψηλές τιμές κόστους, ενώ για κάποιες χαμηλές τιμές το Β μπορεί να είναι καλύτερο από το Α, το γεγονός αυτό αγνοείται. Επιπρόσθετα, η σειρά κατάταξης των ανταγωνιστικών μοντέλων μπορεί να διαφοροποιηθεί ανάλογα με το μέτρο ακριβείας (Myrtveit et al.2005: 388), ενώ η ύπαρξη ορισμένων ακραίων τιμών μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα. Η ικανότητα αξιολόγησης κάθε μέτρου ακρίβειας ερμηνεύεται διαφορετικά, γι' αυτό και υπάρχει ανάγκη να γίνεται χρήση διαφόρων μέτρων ακρίβειας. Με τη χρήση πέραν του ενός μέτρων ακριβείας, η διαδικασία αξιολόγησης των τεχνικών εκτίμησης γίνεται ακόμα πιο περίπλοκη, εφόσον τα αποτελέσματα αυτών είναι αντιφατικά και έτσι δεν μπορούμε να έχουμε μοναδική λύση. Λόγω του ότι κάθε μέτρο ακριβείας ευνοεί διαφορετικά μοντέλα, ο συνδυασμός τους δεν θα συνιστούσε πιο έγκυρα μέτρα (Myrtveit & Stensrud 2012: 26).

Οι (Shepperd and Kadoda 2001: 1015) ερεύνησαν κατά πόσον η επιλογή συνόλων δεδομένων επηρεάζει την απόδοση των τεχνικών πρόβλεψης. Η μέθοδος Stepwise Regression Procedure παρουσίασε τις πιο ακριβείς προβλέψεις για τα κανονικού τύπου και τα κανονικού + ακραίες τιμές τύπου συνόλων δεδομένων, ενώ τα σύνολα δεδομένων με χαρακτηριστικά αλληλοσυσχέτισης παρουσίασαν καλύτερα αποτελέσματα για τις μεθόδους μηχανικής μάθησης (Shepperd and Kadoda 2001: 1019). Με την εφαρμογή του Krushkal Wallis test επιβεβαίωσαν ότι η σχέση μεταξύ συνόλων δεδομένων και τεχνικής πρόβλεψης παρουσιάζει σημαντικά διαφορετικά αποτελέσματα (Shepperd and Kadoda 2001: 1019). Άρα η επιλογή του 'καλύτερου' μοντέλου εξαρτάται από τα χαρακτηριστικά του συνόλου δεδομένων. Μελέτησαν κατά πόσον τα αποτελέσματα σύγκρισης είναι τα ίδια σε αριθμό δειγμάτων training set από το αρχικό σύνολο δεδομένων. Για μικρά training set βρήκαν ότι οι διαφορές είναι σημαντικές, εφόσον παρουσίασαν διαφορετικά αποτελέσματα στις 27 από τις 32 περιπτώσεις, ενώ για μεγάλα training set τα αποτελέσματα ήταν λιγότερο

ευπαθή αφού διαφορετικά ήταν τα αποτελέσματα στις 14 από τις 32 περιπτώσεις (Shepperd and Kadoda 2001: 1019). Για τους λόγους αυτούς οι ερευνητές προτείνουν τη χρήση αρκετών training sets για την εξαγωγή πιο έγκυρων αποτελεσμάτων. Ερεύνησαν εάν το μέγεθος του training set επηρεάζει τα αποτελέσματα σύγκρισης. Για την απάντηση του ερωτήματος αυτού, διεξήγαγαν σύγκριση του απόλυτου σφάλματος πρόβλεψης μεταξύ training set με 20 περιπτώσεις και training set με 100 περιπτώσεις (Shepperd and Kadoda 2001: 1019). Τα αποτελέσματα τους ήταν ότι η χρήση μεγάλων training set ήταν θετική σε όλες τις περιπτώσεις (Shepperd and Kadoda 2001: 1019). Η μέθοδος SWR παρουσίασε θετικά αποτελέσματα μόνο για ένα είδος συνόλου δεδομένων, ενώ οι μεθόδοι μηχανικής μάθησης παρουσίασαν καλύτερα αποτελέσματα για κάθε είδος που μελετήθηκε (Shepperd and Kadoda 2001: 1019). Άρα η επιλογή του μοντέλου πρόβλεψης θα πρέπει να βασίζεται στα διαθέσιμα δεδομένα.

Οι Kitchenham και Mendes (2009:3) επισημαίνουν ότι τα δεδομένα, στα οποία βασίστηκε η κατασκευή ενός μοντέλου, θα πρέπει να αντιπροσωπεύουν τα έργα για τα οποία το μοντέλο θα χρησιμοποιηθεί. Κανένα μοντέλο εκτίμησης δεν θα κάνει ακριβή και έγκυρη πρόβλεψη για έργα διαφορετικά στη φύση, από έργα στα οποία βασίστηκε η κατασκευή του. Θα πρέπει να λαμβάνεται υπόψη η ανομοιογένεια μεταξύ των έργων και η κατηγοριοποίηση τους σε διαφορετικούς τύπους έργων, όπως καινούρια και περαιτέρω βελτίωσης (Kitchenham & Mendes 2009: 3). Το ταίριασμα των μεθόδων σε ένα σύνολο δεδομένων, ανεξάρτητο από πλευράς χρόνου και αμετάβλητο, δεν αποδεικνύει ότι η τεχνική πρόβλεψης θα λειτουργεί αξιόπιστα σε πραγματικές συνθήκες (Kitchenham & Mendes 2009: 3). Όταν γίνεται χρήση συνόλων δεδομένων που αλλάζουν με το χρόνο, θα πρέπει να υπάρχει ενδελεχής εξήγηση, για το πως επιλέχθηκαν τα έργα που έκαναν χρήση (Kitchenham & Mendes 2009: 2).

Οι (Mittas & Angelis 2008a: 616) αναφέρουν ότι κάθε μέτρο ακρίβειας είναι απλά μια τιμή κεντρικής τάσης (π.χ. μέση τιμή ή διάμεσος) που υπολογίζεται από κάποιο δείγμα σφαλμάτων και για τον λόγο αυτόν παρουσιάζει σημαντική μεταβλητότητα. Αν η διαδικασία επικύρωσης βασίζεται απλά σε σύγκριση των καθολικών μέτρων ακρίβειας, χωρίς τη χρήση τυπικών στατιστικών ελέγχων υποθέσεων, τότε υπάρχει ο κίνδυνος να θεωρηθεί ότι μία μέθοδος εκτίμησης υπερτερεί σε σχέση με κάποια άλλη, αλλά στην πραγματικότητα να μην υπάρχει στατιστικά σημαντική διαφορά στις συναρτήσεις σφάλματος (Mittas & Angelis 2008a: 617). Επιπρόσθετα, δυο εναλλακτικές μέθοδοι μπορεί να είναι πανομοιότυπες σε σχέση με το μέσο σφάλμα που παράγουν, αλλά να έχουν εντελώς διαφορετικές κατανομές. Οι (Mittas & Angelis 2008a: 620) σημειώνουν ότι οι

τεχνικές αναδειγματολειψίας υπόκεινται σε δυο μορφές διαφοροποίησης: τη τυχαιοποίηση της επιλογής του αρχικού δείγματος από το πληθυσμό και τη τυχαιοποίηση της επανελειμμένης επιλογής από το αρχικό δείγμα. Η διαφοροποίηση αυτή θεωρείται μικρή και μπορεί να ξεπεραστεί με την αύξηση του αριθμού των επιλεγμένων δειγμάτων.

Κεφάλαιο 3

Η Ανάλυση RROC

3.1 Ανάγκη χρήσης της ανάλυσης RROC

Η διαδικασία της σύγκρισης, όπως έχουμε ήδη αναφέρει, βασίζεται κυρίως σε μέτρα ακρίβειας που υπολογίζονται από συναρτήσεις σφάλματος, ενώ σημαντικό ρόλο διαδραματίζουν οι στατιστικοί έλεγχοι υποθέσεων για τη γενίκευση των αποτελεσμάτων και την επιλογή του καταλληλότερου μοντέλου. Παρόλα αυτά, το πρόβλημα της συμπερασματικής αστάθειας παραμένει άλυτο και γίνεται εκτεταμένη προσπάθεια από την πλευρά των ερευνητών για να κατανοήσουν τις πηγές του προβλήματος.

Εκτός από τα συμπεράσματα που έχουν προκύψει από τις διάφορες μελέτες για το φαινόμενο, οι διοικητές έργων θα πρέπει να λαμβάνουν υπόψη ότι οι ανακρίβειες των εκτιμήσεων δεν έχουν την ίδια βαρύτητα για τους πελάτες και τους διοικητές του έργου (Mittas & Angelis 2013: 3). Επιπρόσθετα, στο ερευνητικό πεδίο της ΕΚΛ, η κατάσταση περιπλέκεται ακόμα περισσότερο, καθώς υπάρχουν δυο είδη εσφαλμένων προβλέψεων και μπορούν να διακριθούν σε υποεκτιμήσεις, όπου το πραγματικό κόστος είναι υψηλότερο από το εκτιμώμενο και τις υπερεκτιμήσεις, όπου η εκτίμηση είναι υψηλότερη από το πραγματικό κόστος (Kemerer 1987: 420).

Η υποεκτίμηση μπορεί να έχει σοβαρές επιπλοκές στην υλοποίηση του έργου, καθώς υπάρχει πάντα ο κίνδυνος για λανθασμένο σχεδιασμό του χρονοπρογραμματισμού αλλά και της στελέχωσης της ομάδας ανάπτυξης (Kemerer 1987: 420). Σύμφωνα με το νόμο του Brooks, η εκ των υστέρων στελέχωση ενός έργου λογισμικού με περισσότερα άτομα, έχει ως αποτέλεσμα ακόμα περισσότερο χρόνο και εκπαίδευση έτσι ώστε το επιπλέον προσωπικό να γίνει παραγωγικό, για την ολοκλήρωση του έργου (Brooks, 1986). Αυτό μπορεί να σημαίνει μειωμένη ποιότητα του παραγόμενου έργου ή καθυστέρηση στην υλοποίηση, με πολύ σοβαρές συνέπειες στη λειτουργία και τη φήμη του οργανισμού

ανάπτυξης (Mittas & Angelis 2013: 1). Επιπρόσθετα, πολλά συμβόλαια προνοούν την επιβολή προστίμων για καθυστερημένη παράδοση ενός έργου και περισσότερα κέρδη για πρόωρη αποπεράτωση (Budd & Cooper 2005).Στη χειρότερη περίπτωση, οι απαιτούμενοι πόροι μπορεί να μην είναι αρκετοί, για την αποπεράτωση του έργου και ο διοικητής του έργου να αναγκαστεί να αναστείλει τη διαδικασία ανάπτυξης, με αποτέλεσμα την απώλεια χρημάτων και πόρων (Lederer & Prasad 1995: 125).

Από την άλλη πλευρά η υπερεκτίμηση έχει επίσης σοβαρές συνέπειες. Σύμφωνα με το νόμο του Parkinson η ομάδα ανάπτυξης επεκτείνει την εργασία, έτσι ώστε να συμπληρωθεί ο απαιτούμενος χρόνος για την ανάπτυξη του έργου, με αποτέλεσμα μειωμένο ρυθμό παραγωγής (Boehm 1981). Ο οργανισμός ανάπτυξης θα μπορούσε στο χρόνο αυτό να αναπτυξει περισσότερα έργα και να αποκομίσει περισσότερα κέρδη, ενώ η υπερβολική χρήση πόρων θα μπορούσε να οδηγήσει ακόμα και στην ακύρωση του έργου (Lederer & Prasad 1995: 125).

Σε κάθε περίπτωση, ο κύριος στόχος είναι η μείωση του κύκλου ζωής της διαδικασίας ανάπτυξης και η παροχή προϊόντων υψηλής ποιότητας εντός προϋπολογισμού, για την επίτευξη της ικανοποίησης των πελατών και του μεγαλύτερου κέρδους. Οι διοικητές έργων θα πρέπει να λάβουν υπόψη τις διαφορετικές επιπτώσεις που μπορεί να έχει στην διαδικασία ανάπτυξης ένα υποεκτιμημένο ή υπερεκτιμημένο έργο. Επιπρόσθετα, οι διοικητές έργων είναι περισσότερο πρόθυμοι να στηρίξουν τη διαδικασία λήψης αποφάσεων σε λύσεις που προέρχονται από εργαλεία οπτικοποίησης της προβλεπτικής ικανότητας εναλλακτικών μοντέλων πρόβλεψης (Mittas & Angelis 2013: 2).

Για την εξάλειψη των περιορισμών αυτών και τη μελέτη των δύο τύπων σφάλματος (υποεκτίμηση/υπερεκτίμηση), οι (Mittas & Angelis 2013: 1) προτείνουν την υιοθέτηση και χρήση της *Regression Receiver Operating Characteristic* (RROC) ανάλυσης, στη διαδικασία της σύγκρισης εναλλακτικών μοντέλων πρόβλεψης. Η ανάλυση RROC, η οποία έχει προταθεί από τον (Hernández-Orallo 2013: 5) μπορεί να θεωρηθεί ως επέκταση της κλασικής ανάλυσης Receiver Operating Characteristic (ROC) που χρησιμοποιείται ευρέως σε προβλήματα σύγκρισης εναλλακτικών κατηγοριοποιητών (classifiers). Γενικά, η ανάλυση RROC βασίζεται στην κατασκευή ενός δισδιάστατου γραφήματος, του RROC space, όπου ο οριζόντιος άξονας αναπαριστάνει το συνολικό σφάλμα υποεκτίμησης και ο κάθετος άξονας το συνολικό σφάλμα υπερεκτίμησης δίνοντας τη δυνατότητα για άμεση σύγκριση και

εξαγωγή χρήσιμης πληροφορίας για την απόδοση εναλλακτικών μοντέλων πρόβλεψης (Hernández-Orallo 2013 : 5, Mittas & Angelis 2013: 3).

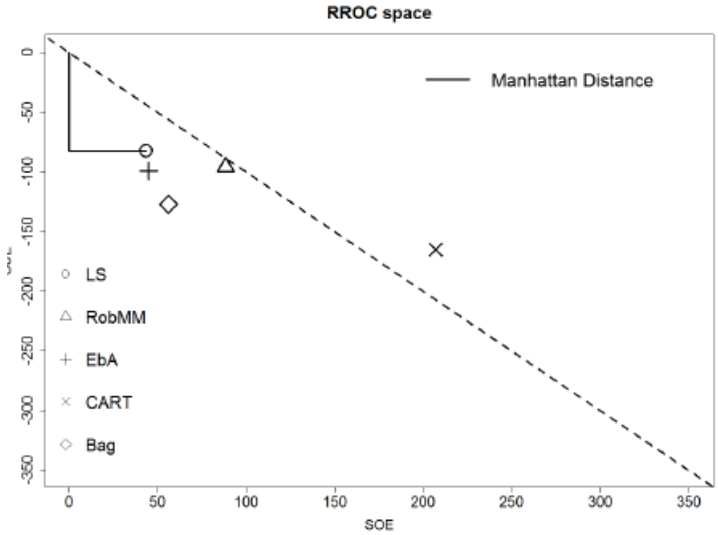
3.2 Γραφική αξιολόγηση και σύγκριση των μοντέλων στον RROC space

Η κατασκευή του RROC space, προκύπτει από το άθροισμα του συνολικού σφάλματος υπερεκτίμησης και υποεκτίμησης, το οποίο σφάλμα e_i ορίζεται ως η διαφορά μεταξύ της εκτιμώμενης Y_{E_i} κόστους και της πραγματικής Y_{A_i}. Η συνολική υπερεκτίμηση (sum of overestimation-SOE) και υποεκτίμηση (sum of underestimation-SUE) δίνονται από τις εξισώσεις 4 και 5 (Hernández-Orallo 2013: 6).

$$SOE = \sum_{i=1}^{n}\{e_i \,|e_i > 0\} \qquad \text{(Εξ. 4)}$$

$$SUE = \sum_{i=1}^{n}\{e_i \,|e_i < 0\} \qquad \text{(Εξ. 5)}$$

Στο Σχήμα 1 παρουσιάζεται ένα παράδειγμα του RROC space, από σφάλματα που προκύπτουν από πέντε εναλλακτικά μοντέλα πρόβλεψης του κόστους. Πιο συγκεκριμένα, έγινε χρήση *i*) της Ordinary Least Squares (LS) regression (Abran & Robillard 1996) , ii) της Robust M-estimator (RobMM) (Jeffery et al. 2001), iii) της εκτίμησης με αναλογίες (EbA) (Shepperd, & Schofield 1997: 737), iv) της bagging EbA (Bag) (Mittas et al. 2008: 7) και v) των δένδρων ταξινόμησης και παλινδρόμησης (CART) (Briand 1992: 933).



Σχήμα 1. RROC space (Mittas & Angelis 2013: 4)

Ένα ιδανικό μοντέλο πρόβλεψης θα πρέπει να έχει μηδενικές τιμές SOE και SUE, με το σημείο (0,0), να αναπαριστά τη βέλτιστη δυνατότητα πρόβλεψης (Hernandez- Orallo 2013: 6, Mittas & Angelis 2013: 4). Η διαγώνιος γραμμή αναφοράς (Σχήμα 1) αντιπροσωπεύει ισοζυγισμένες καταστάσεις, με ίσες τιμές SOE και SUE (Mittas & Angelis 2013: 4). Αυτό σημαίνει ότι η απόδοση των μοντέλων που αντιπροσωπεύεται με κάποιο σημείο στη διαγώνιο δεν παρουσιάζει μεροληψία (Mittas & Angelis 2013: 4). Τα σημεία που βρίσκονται πάνω από τη διαγώνιο, αντιστοιχούν σε αποδόσεις μοντέλων, τα οποία παρουσιάζουν μεροληψία υπερεκτίμησης κόστους, ενώ το αντίθετο ισχύει για τα μοντέλα των οποίων οι αποδόσεις τους βρίσκονται κάτω από τη διαγώνιο (Mittas & Angelis 2013: 4).

Για την αποτίμηση της συνολικής μεροληψίας μίας μεθόδου πρόβλεψης, το συνολικό σφάλμα υποεκτίμησης και υπερεκτίμησης μπορούν να συνδυαστούν σε ένα καθολικό μέτρο μεροληψίας, τη μέση τιμή σφάλματος (Mean Error), η οποία δίνεται από:

ME $= \frac{1}{n}(\sum_{i=1}^{n}(SOE + SUE)$ (Εξ. 6) (Mittas & Angelis 2013: 4), όπου n ο αριθμός περιπτώσεων εκτίμησης.

Μια σημαντική ιδιότητα του RROC space, είναι ότι διάφορα μέτρα ακρίβειας, όπως το μέσο απόλυτο σφάλμα (Mean Absolute Error-MAE) μπορεί να υπολογιστεί γεωμετρικά. Σχηματικά το MAE, ορίζεται ως η απόσταση Manhattan από το ιδανικό σημείο (0,0) μέχρι το δισδιάστατο σημείο, στο οποίο αντιστοιχεί η δυνατότητα πρόβλεψης του εκάστοτε μοντέλου (Hernández-Orallo 2013: 7). Η απόσταση αυτή είναι ένας τρόπος, να συγκρίνουμε την ακρίβεια ενός μοντέλου με ένα σύνολο εναλλακτικών υποψήφιων μοντέλων. Όσο μικρότερη είναι η απόσταση αυτή, τόσο το μοντέλο αυτό έχει καλύτερη δυνατότητα πρόβλεψης (Hernández-Orallo 2013: 7). Ακόμα, θα πρέπει να αναφερθεί ότι το καθολικό μέτρο ακρίβειας MAE μπορεί να υπολογιστεί προσθέτοντας τις απόλυτες τιμές των SOE και SUE και διαιρώντας με το συνολικό αριθμό έργων στο σετ δεδομένων (Hernández-Orallo, 2013: 6).

Από το Σχήμα 1, μπορούμε να δούμε ότι το μοντέλο LS έχει τη μικρότερη τιμή MAE, καθώς το σημείο που παριστάνει την προβλεπτική ικανότητα του μοντέλου βρίσκεται εγγύτερα στο σημείο (0,0), και άρα παρουσιάζει μεγαλύτερη ακρίβεια πρόβλεψης (Mittas & Angelis 2013: 4). Η απόδοση του μοντέλου RobMM βρίσκεται εγγύτερα στη διαγώνιο και επομένως έχει τη μικρότερη μεροληψία (Mittas & Angelis 2013: 4), ενώ παρουσιάζει τη δεύτερη

χειρότερη απόδοση σε όρους ακρίβειας. Αντίθετα, το μοντέλο Bag παρουσιάζει τη μεγαλύτερη μεροληψία, εφ' όσον το σημείο του βρίσκεται στη μεγαλύτερη απόσταση από τη διαγώνιο, σε σχέση με των υπολοίπων μοντέλων (Mittas & Angelis 2013: 4).

Από το Σχήμα 1 και τα χαρακτηριστικά των μοντέλων πρόβλεψης, είμαστε σε θέση να συμπεράνουμε ότι κάθε τεχνική πρόβλεψης μπορεί να είναι αμερόληπτη αλλά αυτό δεν σημαίνει ότι είναι και ακριβής (Mittas & Angelis 2013: 4). Αύτο θα μπορούσε να θεωρηθεί ως ένα γεγονός συμπερασματικής αστάθειας και για τον λόγο αυτόν θα πρέπει να γίνεται χρήση καταλλήλων μέτρων ακρίβειας ώστε να γίνεται επικύρωση και εκτίμηση της μεροληψίας της πειραματικής μελέτης (Menzies & Shepperd 2012: 5) και οι ερευνητές να γνωρίζουν τι πραγματικά μετρά το κάθε ένα από αυτά (Kitchenham et al. 2001: 81).

3.3 Ισομετρικά του RROC space

Η RROC ανάλυση βασίζεται στη συνάρτηση ασύμμετρου απόλυτου σφάλματος (Asymmetric Absolute Error-AAE), η οποία ορίζεται ως εξής (Hernández-Orallo 2013: 5):

$$AAE_i = \begin{cases} 2c\left(Y_{A_i} - Y_{E_i}\right) & , \alpha\nu\ Y_{E_i} < Y_{A_i} \\ 2(1-c)\left(Y_{E_i} - Y_{A_i}\right), & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}, \quad \text{(Εξ. 7)}$$

όπου Y_{A_i} είναι η πραγματική τιμή κόστους και Y_{E_i} η εκτιμώμενη τιμή κόστους, για κάθε περίπτωση. Ο όρος c παίρνει τιμές από 0 μέχρι και 1 και ορίζει την απώλεια κέρδους (loss) λόγω υπερεκτίμησης ή υποεκτίμησης προϋπολογισμού για ένα συγκεκριμένο πρόβλημα ΕΚΛ (Mittas & Angelis, 2013). Αν το $c = 0$, οι υποεκτιμήσεις δεν παρουσιάζουν απώλεια κέρδους και άρα δεν μπορούν να θεωρηθούν ως σφάλμα, ενώ από την άλλη μεριά, για $c = 1$ οι υπερεκτιμήσεις δεν παρουσιάζουν απώλεια κέρδους και δεν λογαριάζονται ως σφάλματα. Για $c = 0.5$, η συνάρτηση σφάλματος αναπαριστά το μέτρο της μέσης τιμής απόλυτου σφάλματος (MAE), όπου η απώλεια είναι συμμετρική και δεν έχουμε διάκριση μεταξύ υπερεκτίμησης και υποεκτίμησης (Hernández-Orallo 2013: 5). Να σημειωθεί στο σημείο αυτό, ότι το απόλυτο σφάλμα (AE) είναι ένα μέτρο, το οποίο έχει χρησιμοποιηθεί ευρέως στην ΕΚΤ για τη διαδικασία της σύγκρισης εναλλακτικών μοντέλων πρόβλεψης.

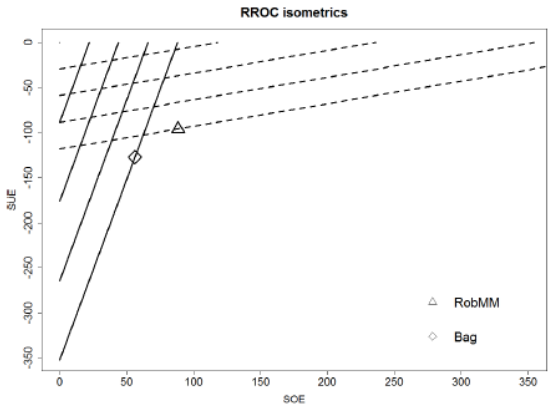
Το συνολικό ασύμμετρο απόλυτο σφάλμα υπολογίζεται από τη συνάρτηση $SAAE = -2cSUE + 2(1-c)SOE$ (Εξ. 8) (Hernández-Orallo 2013: 8), ενώ το Μέσο Ασύμμετρο Απόλυτο Σφάλμα υπολογίζεται από το $MAAE = SAAE/n$ (Εξ. 9) (Mittas & Angelis 2013: 5).

Ένα RROC ισομετρικό ορίζεται ως μια συλλογή σημείων στον RROC space με την ίδια τιμή, για μια ασύμμετρη συνάρτηση σφάλματος. Δηλαδή είναι μια ευθεία γραμμή με συγκεκριμένη κλίση λ, όπου τα μοντέλα που βρίσκονται σε αυτήν, έχουν ίση προσδοκώμενη προβλεπόμενη απόδοση, για μια συγκεκριμένη λειτουργική συνθηκη c. Αυτός είναι ο λόγος που τα ισομετρικά, αναφέρονται και ως γραμμές ίσης απόδοσης. Η ισομετρική γραμμή περιγράφεται από τη πιο κάτω παραμετρική εξίσωση:

$t = -2cSUE + 2(1-c)SOE$ (Εξ. 10) (Hernández-Orallo 2013: 9)

Η κλίση λ κάθε ισομετρικού βασίζεται στη λειτουργική συνθήκη c και δίνεται από $\lambda = \frac{1-c}{c}$ (Hernández-Orallo 2013: 5)

Ας υποθέσουμε ένα συμβόλαιο μεταξύ ενός οργανισμού και ενός πελάτη, ο οποίος δεν είναι διατεθειμένος να πληρώσει ολόκληρο το εκτιμωμένο κόστος για την ανάπτυξη του έργου, αν το έργο δεν ολοκληρωθεί στο συμφωνημένο χρόνο (Mittas & Angelis 2013: 5). Για το λόγο αυτό ο διοικητής του έργου θεωρεί ότι οι υποεκτιμήσεις είναι τέσσερεις φορές πιο ακριβές από τις αντίστοιχες υπερεκτιμήσεις, και οδηγεί στην υπόθεση ότι το $c = 0{,}8$ (Mittas & Angelis 2013: 5).



Σχήμα 2. Ισομετρικά RROC space (Mittas & Angelis 2013: 5)

Το Σχήμα 2, παρουσιάζει διαφορετικές ισομετρικές γραμμές (διακεκομμένες γραμμές) για τη λειτουργική συνθήκη της περίπτωσης ($c = 0.8$). Βλέπουμε ότι όλα τα ισομετρικά είναι παράλληλες ευθείες με κλίση ($\lambda = (1 - 0.8)/0.8 = 0.25$). Μπορούμε να εντοπίσουμε το καλύτερο μοντέλο πρόβλεψης μετατοπίζοντας τις παράλληλες γραμμές από το ιδανικό σημείο $(0{,}0)$ μέχρι τη πρώτη ευθεία που τέμνει το σημείο, το οποίο αντιστοιχεί στο μοντέλο

του RROC space. Για το Σχήμα 2, με κλίση $\lambda = 0.25$, μετατοπίζοντας τις ευθείες βρίσκουμε πρώτα το μοντέλο RobMM (Mittas & Angelis 2013: 5).

Σε μια άλλη περίπτωση, ο οργανισμός ανάπτυξης του έργου εφαρμόζει μια πιο χαλαρή πολιτική στα υποτιμημένα έργα, ενώ επιθυμεί να αυξήσει την παραγωγικότητα της ομάδας ανάπτυξης (Mittas & Angelis 2013: 5). Για τον λόγο αυτόν, εφαρμόζει μια στρατηγική κατά την οποία, η υπερεκτίμηση είναι τέσσερεις φορές πιο ακριβή σε σχέση με την υποεκτίμηση ($c = 0{,}2$). Στη περίπτωση αυτή, τα ισομετρικά είναι παράλληλες ευθείες με κλίση ($\lambda = (1 - 0.2)/0.2 = 4$), και μετατοπίζοντας παράλληλα τις ευθείες, αρχίζοντας από το σημείο $(0{,}0)$, βρίσκουμε το μοντέλο Bag (Mittas & Angelis 2013: 5).

Η χρήση των ισομετρικών παρέχει νέα δεδομένα, σχετικά με την λήψη αποφάσεων για το καλύτερο μοντέλο πρόβλεψης, όπου η υπερεκτίμηση και η υποεκτίμηση δεν παρουσιάζουν συμμετρική απώλεια κέρδους (Mittas & Angelis 2013: 5). Αν υποθέσουμε μια περίπτωση στην οποία δεν έχουμε συμμετρική απώλεια, όπου δεν υπάρχει διάκριση μεταξύ υπερεκτίμησης και υποεκτίμησης, τότε το μοντέλο με το μικρότερο MAE υπερισχύει των υπολοίπων.

3.4 Σύγκριση της δυνατότητας πρόβλεψης με τις καμπύλες RROC

Ο Hernandez- Orallo (2013:11), κάνει χρήση μιας σταθεράς μετατόπισης s, για τη παραγωγή επεξεργασμένων προβλέψεων. Δεδομένου ενός μοντέλου πρόβλεψης κόστους λογισμικού m, το οποίο μετατοπίζεται με μια σταθερά s για κάθε πρόβλεψη $\hat{y}$, έχουμε $\hat{y}' \leftarrow \hat{y} + s$ (Hernandez- Orallo 2013: 12). Η χρήση της μετατόπισης s μπορεί να οδηγήσει σε μοντέλα, με διαφορετικές τιμές SOE και SUE, και υψηλότερο ή χαμηλότερο κόστος (Hernandez-Orallo, 2013). Δεδομένου του ασύμμετρου απόλυτου σφάλματος, με μια λειτουργική συνθήκη c και υποθέτοντας ένα σταθερά μετατοπισμένο μοντέλο, η ιδανική μετατόπιση $\hat{s}$ προκύπτει από την Εξ. 11 ως εξής (Hernandez- Orallo 2013: 12):

$$\hat{s}(c) = \min_s(2(1-c).SOE(s) - 2cSUE(s) \qquad \text{(Εξ. 11)}$$

Αυτό σημαίνει ότι η αρχική μεροληψία του μοντέλου, μπορεί να μηδενιστεί κάνοντας χρήση της μετατόπισης $\hat{s}$. (Hernandez- Orallo, 2013). Αυτό οδηγεί στη κατασκευή της καμπύλης RROC. Βρίσκοντας το φάσμα των τιμών από $\hat{s}(0)$ μέχρι $\hat{s}(1)$ μπορούμε να σχεδιάσουμε τις

καμπύλες RROC, χρησιμοποιώντας τον απαιτούμενο αριθμό τιμών σε αυτό το φάσμα (Hernandez- Orallo 2013: 12).

Οι καμπύλες RROC μπορούν να σχεδιαστούν και να αναλυθούν με πιο ευθύ τρόπο. Μέσω μιας επανελειμμένης διαδικασίας, προσθέτουμε μια σταθερά μετατόπισης s από το διάνυσμα των σφαλμάτων για κάθε έργο n σε κάθε πρόβλεψη (Hernandez- Orallo 2013: 12, Mittas & Angelis 2013, 7).

Ο αλγόριθμος για τον υπολογισμό των RROC καμπυλών παρουσιάζεται ως εξής (Hernandez-Orallo 2013: 12):

1. Υπολογισμός των τιμών σφάλματος $\hat{e}$ μεγέθους n $(\hat{e} = \hat{y} - y)$. ($όπου\ \hat{y}\ εκτιμημένη\ τιμή$)
2. Ταξινόμηση των τιμών σφάλματος $\hat{e}$ σε φθίνουσα σειρά.
3. Αρχικοποίηση $SOE1 \leftarrow 0$ και $SUE \leftarrow -\infty$
4. Για κάθε επανάληψη i $(i = 1,2,..,n)$:
    a. Θέτουμε $s \leftarrow e_i$.
    b. Υπολογίζουμε $e' = e - s$.
    c. Υπολογίζουμε τις τιμές $SOE_i\ και\ SUE_i$ για την οπτικοποίηση του i διανύσματος του RROC.
5. Σχεδιασμός των ευθύγραμμων τμημάτων μεταξύ των $n + 2$ κορυφών στον RROC space.

Κεφάλαιο 4

Γραφική Σύγκριση Μοντέλων

4.1 Στόχοι

Η ανάλυση RROC, η οποία έχει παρουσιαστεί στο προηγούμενο κεφάλαιο, προσφέρει τη δυνατότητα οπτικοποίησης και διάκρισης των δύο τύπων σφάλματος, δηλαδή της υπερεκτίμησης και υποεκτίμησης κόστους. Από την άλλη μεριά, ένα μειονέκτημα της γραφικής απεικόνισης των παραπάνω εννοιών μέσω της RROC ανάλυσης είναι ότι δεν προσφέρει καμία πληροφορία σχετικά με την αβεβαιότητα, που αποτελεί ένα εγγενές χαρακτηριστικό των μοντέλων πρόβλεψης (Mittas & Angelis 2016: 3). Λαμβάνοντας υπόψη ότι μικρές αλλαγές στο σύνολο δεδομένων μπορεί να επηρεάσουν την απόδοση των μοντέλων, θα πρέπει να λάβουμε υπόψη την αβεβαιότητα που καλύπτει τη διαδικασία εκτίμησης και οι λήπτες αποφάσεων να μπορούν να στηρίξουν με κάποιο βαθμό εμπιστοσύνης την απόφαση τους (Mittas & Angelis 2016: 3). Το θέμα αυτό μπορεί να επιτευχθεί μέσα από στατιστικές διαδικασίες, οι οποίες θα μοντελοποιούν τη μεταβλητότητα των σφαλμάτων πρόβλεψης και επομένως θα ελαχιστοποιούν την αβεβαιότητα που είναι εγγενής στη διαδικασία επικύρωσης των μοντέλων.

Συχνάμ η επιλογή ενός μοντέλου ΕΚΛ δεν είναι και τόσο εύκολη υπόθεση. Για παράδειγμα μπορούμε να υποθέσουμε ένα μοντέλο Α με την ιδιότητα να παράγει αμερόληπτες εκτιμήσεις και μεγάλη διασπορά σφάλματος, και ένα μοντέλο Β το οποίο παράγει μεροληπτικές εκτιμήσεις με μικρή διασπορά σφάλματος. Το παράδειγμα αυτό φέρνει στην επιφάνεια το δίλημμα μεροληψίας-διασποράς (Friedman et al. 2001: 37). Με τη χρήση του τετραγωνικού σφάλματος $SQE_i = (\hat{Y}_i - Y_i)^2$ (Εξ. 12), είμαστε σε θέση να διασπάσουμε το μέσο τετραγωνικό σφάλμα (MSE) στα δυο συνθετικά μέρη, δηλαδή τη μεροληψία και διασπορά (πρώτο και δεύτερο συνθετικό αντίστοιχα) σύμφωνα με τη πιο κάτω εξίσωση (Geman et al. 1992:10):

$$MSE = (E[\hat{f}(x) - f(x)])^2 + E[\hat{f}(x) - E[\hat{f}(x)]]^2 \qquad \text{(Εξ. 13)}$$

Η κατανόηση των σύνθετων πτυχών της ικανότητας πρόβλεψης αποτελεί βασικό στοιχείο, αφού ένα υποψήφιο μοντέλο μπορεί να παρουσίαζει διαφορετική συμπεριφορά σε όρους ακρίβειας, μεροληψίας και διασποράς. Ένα μοντέλο μπορεί να θεωρηθεί ως ακριβές, εάν οι εκτιμήσεις που παράγει είναι αρκετά κοντά στις πραγματικές τιμές, ενώ η μεροληψία σχετίζεται με τη τάση ενός μοντέλου να παράγει υποεκτιμήσεις ή υπερεκτιμήσεις, σε σχέση με τις πραγματικές τιμές κόστους (Mittas & Angelis 2016: 2). Η διασπορά των εκτιμήσεων έχει να κάνει με τη συνέπεια πρόβλεψης, δηλαδή αν ένα μοντέλο παρουσιάζει παρόμοια ικανότητα πρόβλεψης για διαφορετικές περιπτώσεις, για παράδειγμα για διαφορετικά σύνολα δεδομένων (Mittas & Angelis 2016: 2). Για καλύτερη διεξαγωγή σύγκρισης των εναλλακτικών μοντέλων ΕΚΛ, τα χαρακτηριστικά αυτά θα πρέπει να εξετασθούν σε ένα πλαίσιο γενικής μορφής και όχι ξεχωριστά.

Ο RROC space παρέχει τη δυνατότητα γραφικής απεικόνισης του σφάλματος πρόβλεψης, ούτως ώστε ο επαγγελματίας να έχει τη δυνατότητα εξαγωγής συμπερασμάτων, όσον αφορά τη ακρίβεια και μεροληψία των μεθόδων εκτίμησης (Mittas & Angelis 2016: 2), ενώ η αναδειγματολειψία bootstrap μπορεί να χρησιμοποιηθεί για τη διευρεύνηση της διασποράς του σφάλματος πρόβλεψης (Mittas & Angelis 2016: 4). Με βάση τα πιο πάνω χαρακτηριστικά των μεθόδων που αναφέραμε οι (Mittas & Angelis 2016:4) προτείνουν τη χρήση του RROC space για bootstrapped σύνολα δεδομένων για εξαγωγή συμπερασμάτων, όσων αφορά το δίλημμα μεροληψίας και διασποράς.

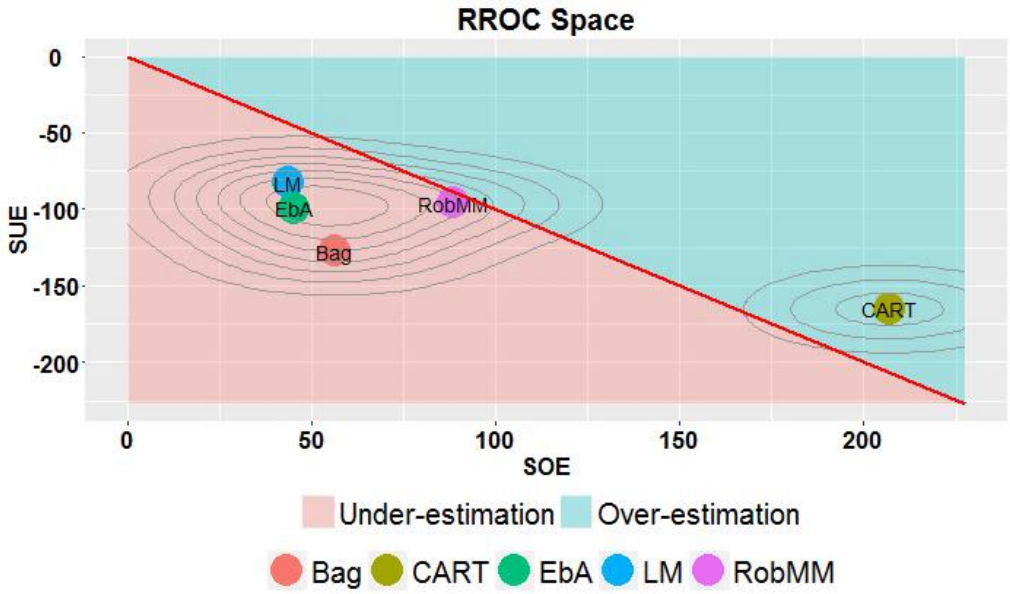
Η μεθοδολογία που ανέπτυξαν διαχωρίζει το αρχικό σύνολο δεδομένων σε σύνολα επικύρωσης (validation set) και εκπαίδευσης (training set), εξομοιώνει με επαναληπτικό τρόπο την παρατηρούμενη κατανομή των πραγματικών τιμών κόστους για το training set μέσω της δειγματοληψίας με επανάθεση (case sampling with replacement) και προσαρμογή ενός μοντέλου σε κάθε επαναληπτικό σύνολο δεδομένων (Mittas & Angelis 2016: 4). Στη συνέχεια, γίνεται υπολογισμός των σφαλμάτων υποεκτίμησης και υπερεκτίμησης για κάθε ένα από τα επαναληπτικά σύνολα δεδομένων και η κατασκευή διαστημάτων εμπιστοσύνης με βάση τις μονοδιάστατες κατανομές των SOE και SUE (Mittas & Angelis 2016: 4).

Στην μεταπτυχιακή Διατριβή προτείνεται ένας εναλλακτικός τρόπος κατασκευής διαστημάτων εμπιστοσύνης για τη μεροληψία ενός μοντέλου πρόβλεψης. Η μεθοδολογία

βασίζεται και πάλι στη χρήση της επαναληπτικής αναδειγματοληψίας bootstrap αλλά διαφοροποιείται ο τρόπος με τον οποίο γίνεται η προσαρμογή των μοντέλων. Πιο αναλυτικά, ο αλγόριθμος στηρίζεται στην αναδειγματοληψία της κατανομής των σφαλμάτων (error resampling) που έχουν υπολογιστεί από τη διαδικασία επικύρωσης LOOCV στο αρχικό σύνολο δεδομένων. Επομένως, διαφοροποιείται ο προτεινόμενος αλγόριθμος των (Mittas & Angelis 2016: 4) και αντί για την αναδειγματοληψία περιπτώσεων (case resampling) προτείνεται η επαναδειγματοληψία των σφαλμάτων πρόβλεψης που υπολογίζονται από το εκάστοτε μοντέλο. Στη συνέχεια, μετά την αναδειγματοληψία με επανάθεση των σφαλμάτων πρόβλεψης, κατασκευάζονται CIs κάνοντας χρήση των 2.5% και 97.5% ποσοστιαίων σημείων της bootstrap κατανομής. Θα πρέπει να σημειωθεί ότι σε περιπτώσεις όπου υπάρχει η πεποίθηση ότι το μοντέλο προβλέπει με ακρίβεια τη συνάρτηση κόστους, η αναδειγματοληψία των σφαλμάτων ενδέχεται να επιφέρει αρκετά ικανοποιητικά αποτελέσματα και είναι συγκρίσιμα με τα αποτελέσματα που παράγονται από την αναδειγματοληψία των περιπτώσεων σε ένα σύνολο δεδομένων.

4.2 Παρουσίαση Μεθοδολογίας

Για την περιγραφή της προτεινόμενης μεθοδολογίας, στην ενότητα αυτή κάνουμε χρήση του συνόλου δεδομένων Albrecht (Albrecht & Gaffney, 1983) πάνω στο οποίο έγινε η προσαρμογή πέντε εναλλακτικών μοντέλων πρόβλεψης (LS, RobMM, EbA, bagging EbA (Bag), CART) και έχουν παρουσιαστεί στο προηγούμενο κεφάλαιο. Για την οπτικοποίηση των αποδόσεων πρόβλεψης τους χρησιμοποιήθηκε ο RROC space (Σχήμα 3), από τον οποίο μπορούμε να δούμε τη μεροληψία σε σχέση με τα σφάλματα πρόβλεψης κάθε μοντέλου.



Σχήμα 3. RROC space για το σύνολο δεδομένων Albrecht,

Από Σχήμα 3, μπορούμε να δούμε ότι το μοντέλο RobMM εμφανίζει τη μικρότερη μεροληψία σε σχέση με τα ανταγωνιστικά μοντέλα καθώς όπως έχουμε προαναφέρει βρίσκεται πιο κοντά στη διαγώνιο της RROC καμπύλης. Από την άλλη μεριά, το μοντέλο Bag φαίνεται να εμφανίζει τη χειρότερη απόδοση σε όρους μεροληψίας. Τέλος, σε όρους ακρίβειας το μοντέλο LM φαίνεται να εμφανίζει τη μεγαλύτερη ακρίβεια, εφόσον έχει τη μικρότερη απόσταση από το ιδανικό σημείο (0,0) ενώ το μοντέλο CART παρουσιάζει τη χειρότερη απόδοση σε όρους ακρίβειας.

Όπως έχουμε προαναφέρει, τα παραπάνω σημεία προέρχονται από δείγματα σφαλμάτων και ως εκ τούτου εμφανίζουν σημαντική μεταβλητότητα. Στόχος, λοιπόν, είναι να μελετηθεί η μεταβλητότητα των σφαλμάτων υποεκτίμησης και υπερεκτίμησης και να αναπαρισταθεί στο RROC space ώστε να μπορούν οι ερευνητές και οι διοικητές έργου να λαμβάνουν χρήσιμη πληροφορία από τη μελέτη ενός τέτοιου γραφήματος, που θα διευκολύνει τη διαδικασία λήψη αποφάσεων για την επιλογή του καταλληλότερου μοντέλου πρόβλεψης.

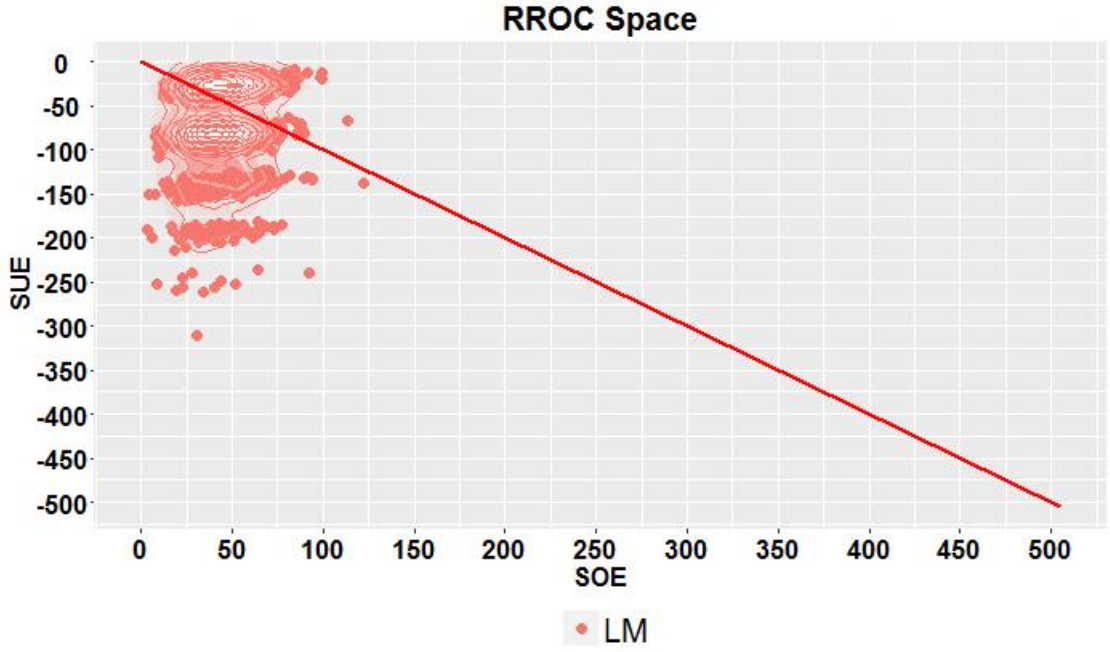
Λόγω του ότι δεν υπάρχουν αναλυτικές εκφράσεις για την μελέτη της μεροληψίας (Mittas & Angelis 2016: 4), πρόκειται να κάνουμε χρήση της τεχνικής προσομοίωσης bootstrap. Η τεχνική bootstrap είναι μια τεχνική προσομοίωσης βασισμένη στην αναδειγματοληψία, η οποία χρησιμοποιείται για την εξαγωγή της κατανομής ενός δείγματος και βασίζεται στη δειγματοληψία με επανάθεση. Πρόκειται να χρησιμοποίησουμε τη μη-παραμετρική (non-parametric) μέθοδο bootstrap για την εύρεση της εμπειρικής κατανομής του δείγματος. Η μέθοδος αυτή περιλαμβάνει τη δημιουργία ενός μεγάλου αριθμού δειγμάτων *B*, τα οποία

εξάγονται παίρνοντας κάθε φορά ένα αριθμό περιπτώσεων n, με επανάθεση, από το αρχικό δείγμα (Efron & Tibshirani 1993: 4). Στο ερευνητικό πρόβλημα που καλούμαστε να αντιμετωπίσουμε, ο αλγόριθμος προσαρμόζεται, έτσι ώστε το δείγμα να αποτελείται από τα σφάλματα πρόβλεψης, που παράγονται από την εφαρμογή της LOOCV σε ένα σύνολο δεδομένων.

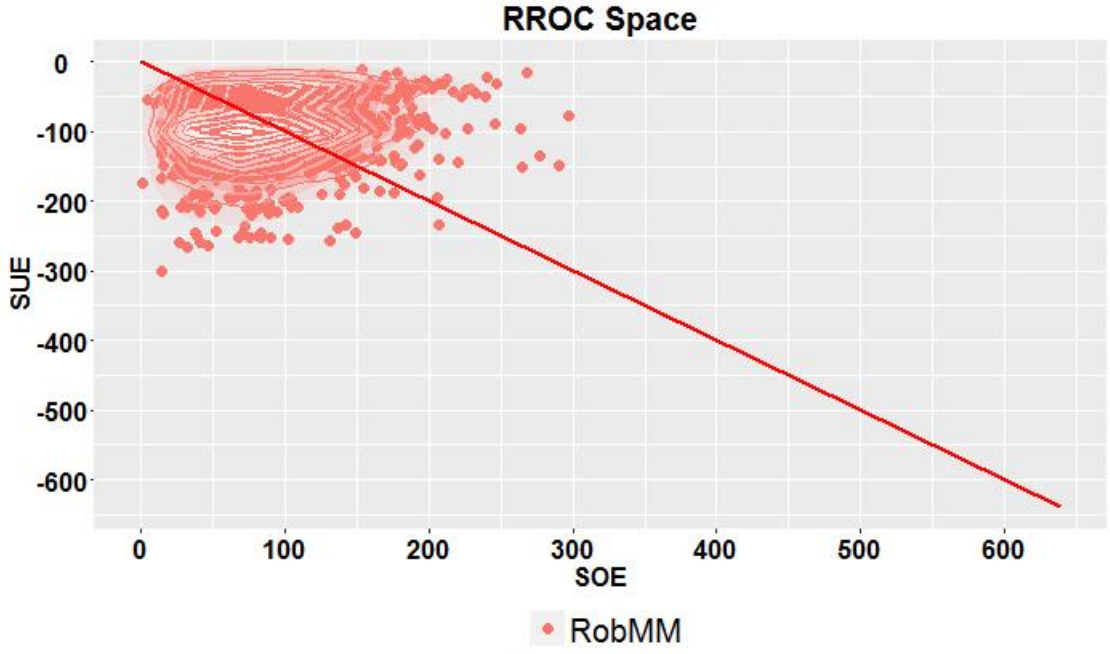
Για την κατασκευή διαστημάτων εμπιστοσύνης για τα σημεία του RROC space, τα βήματα του αλγορίθμου είναι τα ακόλουθα:

1. Προσαρμογή ενός μοντέλου στο αρχικό σύνολο δεδομένων μέσω της διαδικασίας leave-one-out cross-validation (LOOCV).
2. Εκτίμηση και απεικόνιση των τιμών SOE και SUE στον RROC space.
3. Από το αρχικό δείγμα σφαλμάτων λαμβάνεται με επανάθεση ένας μεγάλος αριθμός B bootstrap δειγμάτων με επανάθεση, δηλαδή με αντικατάσταση επιλέγεται ένα σετ περιπτώσεων $(j_1, .., j_n)$, δημιουργώντας κάθε φορά το δείγμα $x^{*i} = \left(x_{j1}, .., x_{jn}\right)$, μεγέθους n, όπου $i = 1,2, \dots B$.
4. Για κάθε δείγμα x^{*i} υπολογίζουμε το άθροισμα των υπερεκτιμημένων και υποεκτιμημένων περιπτώσεων.
5. Από τα δύο διανύσματα των υπερεκτιμημένων και υποεκτιμημένων περιπτώσεων, κατασκευάζουμε τον RROC space, για κάθε περίπτωση x^{*i}.
6. Από το σύνολο των τιμών του αθροίσματος υπερεκτίμησης και υποεκτίμησης, κατασκευάζουμε τα 95% διαστήματα εμπιστοσύνης για τις τιμές SOE και SUE με τη βοήθεια των 2.5% και 97.5% ποσοστιαίων σημείων της bootstrap κατανομής.

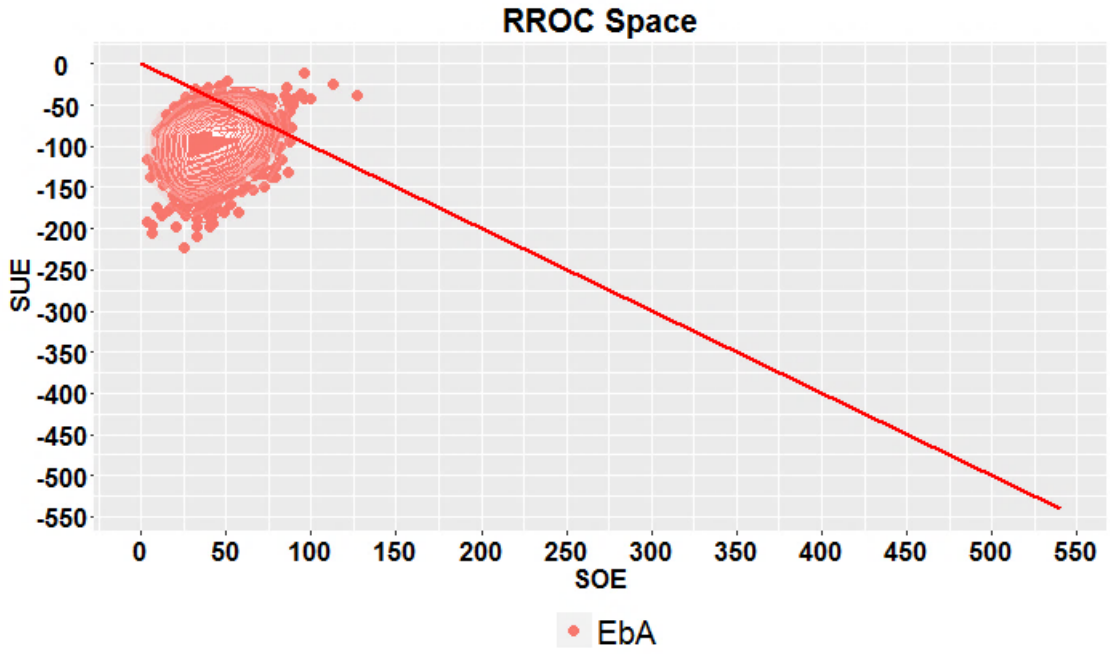
Στα Σχήματα 4, 5, 6, 7 και 8 παρουσιάζονται οι bootstrap κατανομές για τα σφάλματα της υπερεκτίμησης και υποεκτίμησης των μοντέλων LM,RobMM, EbA, Bag και CART αντίστοιχα.



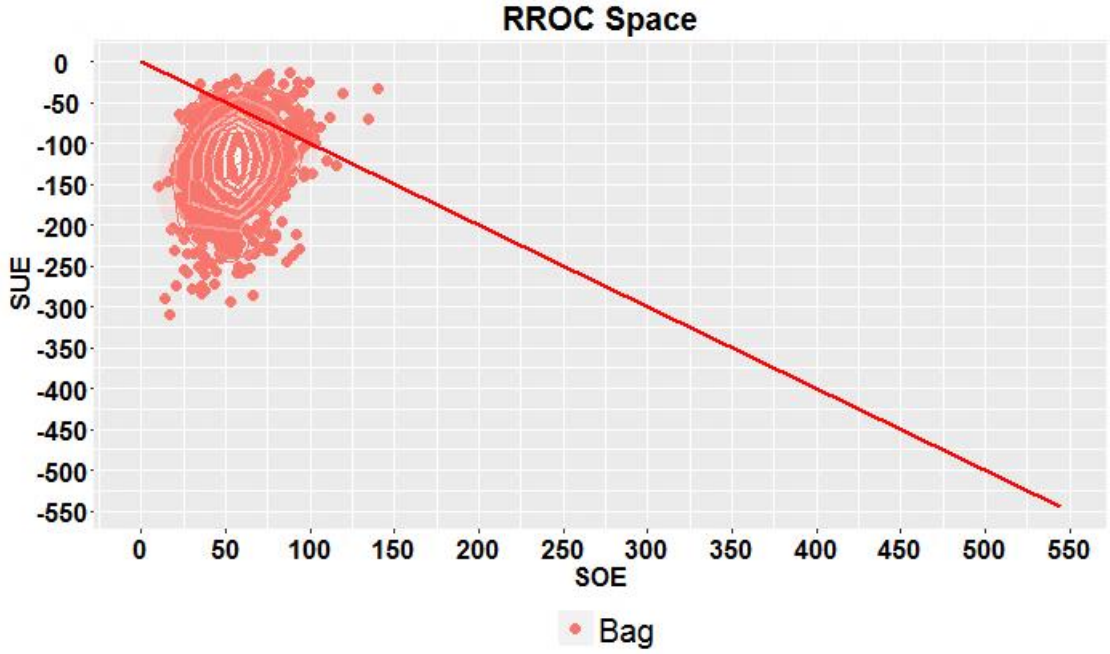
Σχήμα 4. RROC space των bootstrap δειγμάτων σφάλματος που προκύπτουν από το μοντέλο LM για το σύνολο δεδομένων Albrecht.



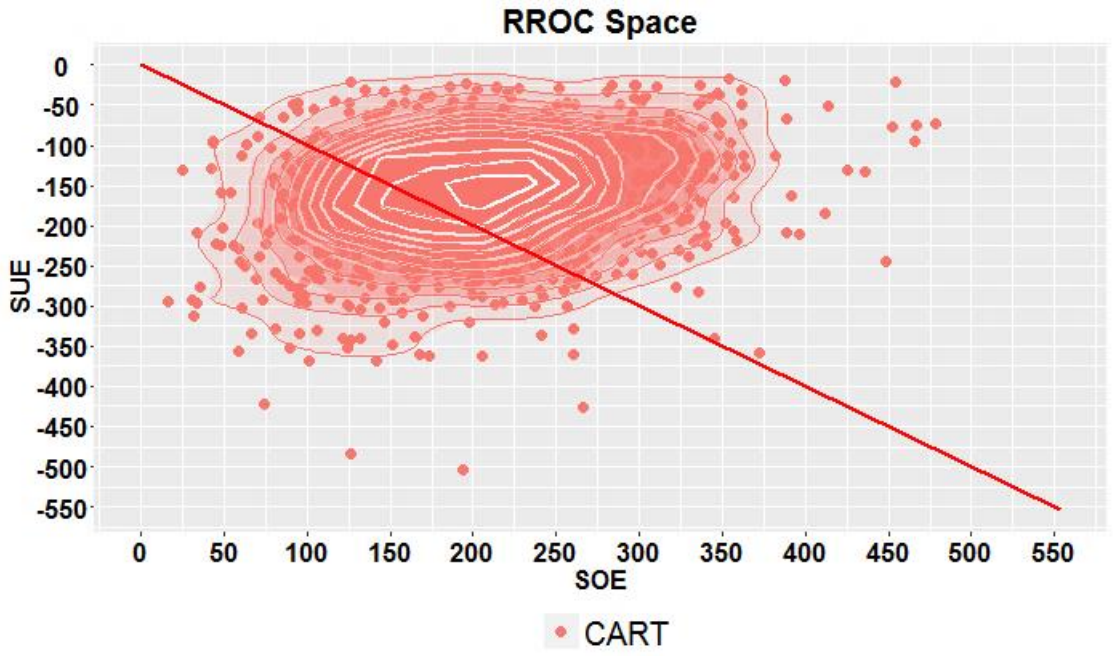
Σχήμα 5. RROC space των bootstrap δειγμάτων σφάλματος που προκύπτουν από το μοντέλο RobMM για το σύνολο δεδομένων Albrecht.



Σχήμα 6. RROC space των bootstrap δειγμάτων σφάλματος που προκύπτουν από το μοντέλο EbA για το σύνολο δεδομένων Albrecht.



Σχήμα 7. RROC space των bootstrap δειγμάτων σφάλματος που προκύπτουν από το μοντέλο EbA για το σύνολο δεδομένων Albrecht.



Σχήμα 8: RROC space των bootstrapp δειγμάτων σφάλματος για το μοντέλο EbA και το σύνολο δεδομένων Albrecht.

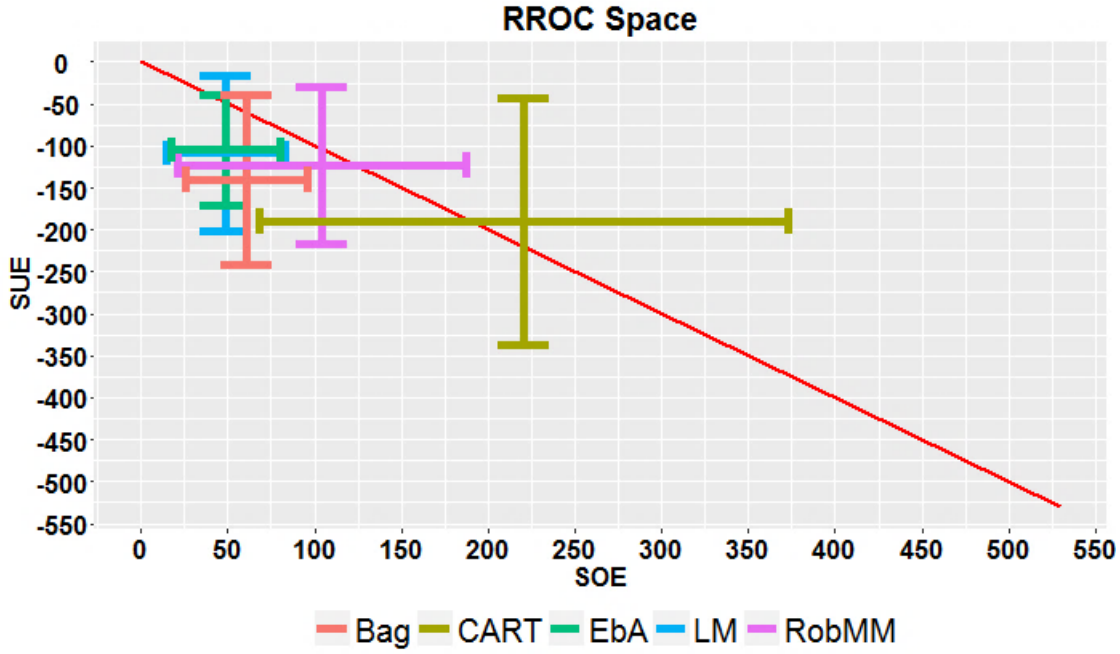
Συγκρίνοντας τις bootstrap κατανομές των σφαλμάτων υποεκτίμησης και υπερεκτίμησης με βάση τα πιο πάνω σχήματα, παρατηρούμε ότι το μοντέλο CART παρουσιάζει τη μεγαλύτερη διασπορά στις εκτιμήσεις του, με αρκετά μεγάλες τιμές μεροληψίας και προς τις δυο περιπτώσεις, ενώ η ακρίβεια που παρουσιάζει είναι αρκετά χαμηλή αφού οι εκτιμήσεις του βρίσκονται αρκετά μακριά από το ιδανικό σημείο (0,0). Το αντίθετο ισχύει για τα μοντέλα EbA και LM καθώς παρουσιάζουν παραπλήσιες δυνατότητες και βρίσκονται αρκετά κοντά στο σημείο (0,0) και η διασπορά των εκτιμήσεων τους βρίσκεται πλησίον της γραμμής αμεροληψίας. Επιπρόσθετα, είμαστε σε θέση να παρατηρήσουμε ότι το μοντέλο RobMM παρουσιάζει αρκετά μεγάλη διασπορά στις τιμές SOE ενώ τα μοντέλα Eba,Bag και LM παρουσιάζουν πολύ μικρότερες τιμές SOE σε σχέση με τα υπόλοιπα, ενώ η πληθώρα των παρατηρήσεων τους ευρίσκεται κάτω από τη γραμμή αμεροληψίας, το οποίο δείχνει τάση για υποεκτίμηση κόστους.

Για την επιλογή του ‘καλύτερου’ μοντέλου πρόβλεψης, θα πρέπει να γίνεται ισοζυγισμός των πλεονεκτημάτων και μειονεκτημάτων, σε σχέση με τα εναλλακτικά και οι λήπτες αποφάσεων να λαμβάνουν υπόψη τους τόσο την ακρίβεια πρόβλεψης και μεροληψία, αλλά και τη διασπορά του σφάλματος πρόβλεψης.

Με βάση τις μονοδιάστατες bootstrap κατανομές των σφαλμάτων υποεκτίμησης και υπερεκτίμησης και τη μέθοδο των ποσοστιαίων σημείων μπορούμε να υπολογίσουμε ένα 95% διάστημα εμπιστοσύνης για κάθε ένα μοντέλο χωριστά. Στο Σχήμα 9 παρουσιάζονται διαγραμματικά τα 95% διαστήματα εμπιστοσύνης και στον Πίνακα 1 παρουσιάζονται αναλυτικές λεπτομέρειες για κάθε μοντέλο.

95% BOOTSTRAP CONFIDENCE INTERVALS					
	LM	ROBMM	EBA	BAG	CART
SOE	[13.5, 82.7]	[17.5 , 195.0]	[14.2, 76.5]	[15170.0, 57684.7]	[24.8, 96.6]
SUE	[-212.6, -15.7]	[-219.1, -27.5]	[-241.3, -38.4]	[-386382.1, -145606.7]	[-334.0, -41.4]
ME	[-6.9, 2.1]	[-6.5, 5.6]	[-5.8, 0.9]	[-8.6, 1.6]	[-8.4, 11.6]
VAR	[14.1, 423.0]	[26.5, 562.7]	[34.3, 129.8]	[44.8, 324.7]	[231.4, 1114.8]
MSE	[8.0, 428.6]	[18.4, 557.3]	[29.0, 139.7]	[29.0, 359.9]	[221.7, 1051.2]

Πίνακας 1. Τα 95% διαστήματα εμπιστοσύνης για το bootstrapped σύνολο δεδομένων Albrecht.



Σχήμα 9. Δισδιάστατα 95% διαστήματα εμπιστοσύνης στον RROC space

4.3 Κώδικας Προγραμματισμού

Για την υλοποίηση της μεθόδου bootstrap και τον εντοπισμό των διαστημάτων εμπιστοσύνης του κάθε μοντέλου προχωρήσαμε στην υλοποίηση του πιο κάτω κώδικα, τον οποίο πρόκειται να περιγράψουμε στη συνέχεια. Υπολογίζουμε τον πίνακα σφάλματος πρόβλεψης (πραγματικές- εκτιμημένες τιμές) GAP. Από αυτόν κάνουμε ανάδειγματοληψία μεγέθους n (όσες είναι οι περιπτώσεις) k φορές, με αντικατάσταση. Το κάθε δείγμα που παίρνουμε αποθηκεύεται σε ένα διάνυσμα GAP.vector. Το άθροισμα των θετικών τιμών (υπερεκτίμήσεις) του κάθε δείγματος αποθηκεύεται στη πρώτη στήλη του πίνακα ROC, ενώ στη δεύτερη στήλη, οι αντίστοιχες αρνητικές (υποεκτιμήσεις). Συνολικά έχουμε 1000 αθροίσματα υπερεκτιμήσεων και υποεκτιμήσεων. Τα αθροίσματα υπερεκτιμήσεων και υποεκτιμήσεων τοποθετούνται σε δυο αντίστοιχες λίστες (listofSOQMatrices, listofSUQMatrices). Στη λίστα listofBootstrapModels τοποθετούνται τα 1000 δείγματα δεδομένων που αναπτύχθηκαν με τη μέθοδο bootstrap, για κάθε μοντέλο. Από τη λίστα listofSOQmatrices δημιουργούμε ένα πίνακα με τις συνολικές υπερεκτιμήσεις σε κάθε στήλη για κάθε μοντέλο, το αντίστοιχο γίνεται με τη λίστα listofSUQmatrices και το πίνακα SUQ για τις υποεκτιμήσεις.

Στη συνέχεια υπολογίζουμε τα 95% διαστήματα εμπιστοσύνης για κάθε μοντέλο, για κάθε περίπτωση (υπερεκτίμησης και υποεκτίμησης), με βάση τη διάμεσο τιμή. Για το σκόπο αυτό, ταξινομούμε τους πίνακες SOQ και SUQ σε αύξουσα σειρά και δημιουργούμε στους SOQ_sorted και SUQ_sorted. Με τη συνάρτηση apply και probs = c(.025, 0.975) βρίσκουμε τα τεταρτημόρια που αντιστοιχούν στο 2.5% και το 97.5% των πινάκων SOQ_sorted και SUQ_sorted και τα αποθηκεύουμε στο αντίστοιχο διάνυσμα quartilesx, quartilesy. Τα 2.5% τεταρτημόρια που αντιστοιχούν στις υπερεκτιμήσεις δηλαδή quartilesx αποθηκεύονται στο πίνακα lower1, ενώ τα αντίστοιχα τεταρτημόρια για τις υποεκτιμήσεις στο πίνακα quartilesy στον lower2. Τα 97.5% τεταρτημόρια που αντιστοιχούν στις υποεκτιμήσεις, δηλαδή quartilesy αποθηκεύονται στο πίνακα upper1, ενώ τα αντίστοιχα τεταρτημόρια για τις υπερεκτιμήσεις, δηλαδή για το πίνακα quartilesy στον upper2.

```
"BootRROCSpace" <- function (Dataset=albrecht, nboot=1000)
{  library(ggplot2)
  library(gridExtra)
  library(scales)
  library("klaR")
  library("psych")
  library("rpart")
  library("cluster")
    Actual <- Dataset [,1]
  Predicted <- Dataset [,c(-1)]
    NamesModels <- names(Predicted)
  group=factor(NamesModels)
  noErrors <- nrow(Predicted)
  noModels <- ncol(Predicted)
  Gap <- -(Actual-Predicted)
  Gap.Matrix <- matrix(,noErrors,nboot)
```

```
ROC.Matrix <- matrix(,nboot,2)
SOQ.Matrix<-matrix(,nboot,4)
SUQ.Matrix<-matrix(,nboot,4)
  listOfBootstrapModels <- vector(mode = "list", length = noModels)
listOfROCMatrices <- vector(mode = "list", length = noModels)
listOfSOQMatrices <- vector(mode = "list", length = noModels)
listOfSUQMatrices <- vector(mode = "list", length = noModels)
  for (k in 1:noModels){
  for (i in 1:nboot){
    indices <- sample(noErrors, replace=T)
    Gap.Vector <- Gap[indices,k]
         Gap.Matrix[,i]  <- Gap.Vector
         ROC.Matrix[i,1]<- sum(Gap.Vector[Gap.Vector>=0])
    ROC.Matrix[i,2]<-sum(Gap.Vector[Gap.Vector<0])
       }
     listOfSOQMatrices[[k]]<-ROC.Matrix[,1]
  listOfSUQMatrices[[k]]<-ROC.Matrix[,2]
  listOfBootstrapModels[[k]] <- Gap.Matrix
  listOfROCMatrices[[k]] <- ROC.Matrix
    }
  SOQ <- data.frame(do.call("cbind", listOfSOQMatrices))
  SUQ <- data.frame(do.call("cbind", listOfSUQMatrices))

SOQ_sorted <- apply(SOQ,2,sort,decreasing=F)
```

```
SUQ_sorted <- apply(SUQ,2,sort,decreasing=F)
 quartilesx<-apply(SOQ_sorted, 2, quantile, probs = c(.025, 0.975),  na.rm = TRUE)
 quartilesy<-apply(SUQ_sorted, 2, quantile, probs = c(.025, 0.975),  na.rm = TRUE)
  lower1<- quartilesx[1,]
upper1<-quartilesx[2,]
lower2<-quartilesy[1,]
upper2<-quartilesy[2,]
  d1=data.frame(NamesModels,lower1, upper1)
  d2=data.frame(NamesModels,lower2, upper2)
  for (k in 1:noModels){
     Gap.Matrix<-listOfBootstrapModels[[k]]
     Gap_new <- data.frame(apply(Gap.Matrix,2,sort,decreasing=F))
     meanerror<-c(colMeans(Gap_new))
  quart_meanerror<- data.frame(quantile(meanerror, c(.025, 0.975)))
    variance<-c(colMeans((Gap_new-colMeans(Gap_new))^2))
  quart_variance<- data.frame(quantile(variance, c(.025, 0.975)))
     Gap_new2<-Gap_new^2
  mean_sqerror<-c(colMeans(Gap_new2))
  quart_meansqerror<- data.frame(quantile(mean_sqerror, c(.025, 0.975)))
    }
  dfBootstrap <- data.frame(do.call("rbind", listOfBootstrapModels))
Models <- rep(NamesModels, each=noErrors)
  dfBootstrap <- data.frame(Models,dfBootstrap)
  dfROC <- data.frame(do.call("rbind", listOfROCMatrices))
```

```
 ROCModels <- rep(NamesModels, each=nboot)
 dfROC <- data.frame(ROCModels,dfROC)
names(dfROC)<- c("Models", "SOQ", "SUQ")
newdfROC<-dfROC[ROCModels=="CART",]
 AxisLimit <- max(max(dfROC[,"SOQ"]), max(abs(dfROC[,"SUQ"])))
AxisLimitX <- round(1.1*AxisLimit,1)
 d <- data.frame(x=c(1.1*AxisLimit,1.1*AxisLimit,0,0,0,1.1*AxisLimit),
            y=c(-1.1*AxisLimit,0,0,-1.1*AxisLimit,0,-1.1*AxisLimit),
            Qualification=c("Over-estimation","Over-estimation","Over-estimation",
                        "Under-estimation","Under-estimation","Under-estimation"))
 d$Qualification <- factor(d$Qualification, c("Under-estimation","Over-estimation"))
 BootstrapRROCSpace <- ggplot(data=dfROC,environment = environment()) +
  geom_point(aes(x=SOQ,
             y=SUQ,color=Models),size=3)+
    stat_density2d(aes(x=SOQ,
                y=SUQ,
                alpha= ..level..,
                fill=Models
 ), size=1,  n=20, geom="polygon",show.legend =FALSE)+
    geom_density2d(aes(x=SOQ,
                y=SUQ, color=Models), show.legend =FALSE)+
    geom_segment(aes(x = 0, y = 0,
              xend = 1.1*AxisLimit,
              yend =-1.1*AxisLimit),size=1, colour="red")+
```

```
      scale_x_continuous(breaks=pretty_breaks(n=10)) +
   scale_y_continuous(breaks=pretty_breaks(n=10)) +
     theme(legend.position="bottom",
         strip.text.x = element_text(size = 12, hjust = 0.5, vjust = 0.5, face = 'bold'),
        legend.title=element_blank(),
         legend.text = element_text(size=20),
        axis.title.x = element_text(face="bold", color="black", size=16),
        axis.title.y = element_text(face="bold", color="black", size=16),
        plot.title = element_text(face="bold", color = "black", size=20),
        axis.text.x = element_text(size = 16, hjust = 0.5, vjust = 0.5, face = 'bold',color =
"black"),
        axis.text.y = element_text(size = 16, hjust = 0.5, vjust = 0.5, face = 'bold',color =
"black")) +
   labs(x="SOE",
       y = "SUE",title= "RROC Space")
   errbars<-ggplot()+
     geom_segment(aes(x = 0, y = 0,
                 xend = 1.1*AxisLimit,
                yend =-1.1*AxisLimit),size=1, colour="red")+
      geom_errorbar(data=d1, mapping=aes( x=(lower1+upper1)/2,color=factor(group),
                              ymin=lower2, ymax=upper2),
             inherit.aes=FALSE,width=30, size=2,show.legend = TRUE )+
      geom_errorbarh(data=d2, mapping=aes(y=(lower2+upper2)/2,x=upper1,
                              xmin=lower1, xmax=upper1,color=factor(group)),
              inherit.aes=FALSE,      height=30, size=2,  show.legend = TRUE)+
```

```
      scale_x_continuous(breaks=pretty_breaks(n=10)) +
   scale_y_continuous(breaks=pretty_breaks(n=10)) +
      theme(legend.position="bottom",
            strip.text.x = element_text(size = 12, hjust = 0.5, vjust = 0.5, face = 'bold'),
       legend.title=element_blank(),
       legend.text = element_text(size=20),
       axis.title.x = element_text(face="bold", color="black", size=16),
       axis.title.y = element_text(face="bold", color="black", size=16),
       plot.title = element_text(face="bold", color = "black", size=20),
       axis.text.x = element_text(size = 16, hjust = 0.5, vjust = 0.5, face = 'bold',color = "black"),
       axis.text.y = element_text(size = 16, hjust = 0.5, vjust = 0.5, face = 'bold',color = "black")) +
   labs(x="SOE",
       y = "SUE",title= "RROC Space")
   list(BootstrapRROCSpace=BootstrapRROCSpace,errbars=errbars,
     listOfBootstrapModels=listOfBootstrapModels,
     listOfROCMatrices=listOfROCMatrices)

}
```

Κεφάλαιο 5

Πειραματική Μελέτη

Σε αυτήν την ενότητα, παρουσιάζονται τα αποτελέσματα των συγκρίσεων, τα οποία εξήχθησαν με εφαρμογή της διαδικασίας προσαρμογής (fitting) των μοντέλων, πιο συγκεκριμένα κάνοντας χρήση της LOOCV σε γνωστά σύνολα δεδομένων της ΕΚΛ. Παρουσιάζονται τα αποτελέσματα σύγκρισης για κάθε μοντέλο και σύνολο δεδομένων που χρησιμοποιήθηκε. Επιπρόσθετα, διεξάγεται σύγκριση με βάση διαφορετικά μέτρα ακρίβειας για τον εντοπισμό τυχόν διαφορετικών αποτελεσμάτων και τον εντοπισμό τυχούσης μεροληψίας αυτών. Επιπλέον, παρουσιάζεται ο πίνακας διαστημάτων εμπιστοσύνης για τις περιπτώσεις υπερεκτίμησης και υποεκτίμησης, της μέσης τιμής σφάλματος, της διασποράς καθώς και της μέσης τιμής τετραγωνικού σφάλματος.

5.1 Μοντέλα Πρόβλεψης

Για την υλοποίηση της πειραματικής διαδικασίας χρησιμοποιήσαμε 4 γνωστές μεθόδους που έχουν χρησιμοποιηθεί στο ερευνητικό πεδίο της ΕΚΛ και είναι: i) η μέθοδος least squares regression (LS) (Abran & Robillard 1996) ii) τα δένδρα ταξινόμησης και παλινδρόμησης (CART) (Briand et al. 1992: 933), iii) η μέθοδος εκτίμησης με αναλογίες (EbA) (Shepperd, & Schofield 1997: 737) και iv) η μέθοδος Naïve Bays Classifier (Ren et al. 2009: 945). Διεξάγαμε σύγκριση των μοντέλων αυτών στον RROC space για 5 σετ δεδομένων, αφού πρώτα υλοποιήθηκε η τεχνική bootstrap για την εύρεση της εμπειρικής κατανομής των σφάλματων πρόβλεψης.

5.1.1 Least Squares Regression

Η μέθοδος πρόβλεψης Least Squares Regression (Kitchenham et al., 2001), (Briand et al., 1999), (Boehm et al., 2000) έχει ως σκοπό τον εντοπισμό των ιδανικών παραμέτρων για την ελαχιστοποίηση του τετραγωνικού σχετικού σφάλματος. Αν υποθέσουμε ότι έχουμε n ζεύγη δεδομένων $(size_i,\ effort_i)$ $(i = 1,2,3,\dots,n)$, η συνάρτηση του μοντέλου Least Squares regression έχει τη μορφή Y_i=a$f(x_i)$+b (Εξ. 14),

όπου Y_i=$Effort_i$ η απαιτούμενη προσπάθεια για την ολοκλήρωση του έργου ,

x_i=$size_i$ το μέγεθος του έργου.

Σκόπός μας είναι η δημιουργία ενός μοντέλου εκτίμησης της μεταβλητής Y της μορφής: $\hat{Y} = \beta_0 + \beta_1 x_i + u_t$ (Εξ. 15). Για την εύρεση της συνάρτησης εκτίμησης, θα πρέπει να εκτιμηθούν οι παραμέτροι β_0, β_1 έτσι ώστε να γίνει ελαχιστοποίηση του αθροίσματος του σχετικού σφάλματος:

Min $(\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_i x_i)^2)$ (Εξ. 16)

5.1.2 Εκτίμηση με Αναλογία

Η Εκτίμηση με Αναλογία (Shepperd & Schofield 1997: 738), συγκρίνει το έργο πρόβλεψης με παρόμοια προηγούμενα με κοινά χαρακτηριστικά και καθορίζει τη προσπάθεια του προβλεπόμενου έργου ως συνάρτηση των γνωστών απαιτούμενων προσπαθειών για παρόμοια έργα. Για την υλοποίηση της μεθόδου αυτής θα πρέπει να βρεθούν τα πιο κοντινά ανάλογα του προβλεπόμενου έργου και να προβλεφθεί η προσπάθεια του έργου, προσαρμόζοντας τις απαιτούμενες προσπάθειες των πιο κοντινών αναλογιών. Η εκτίμηση με αναλογίες μπορεί να χειριστεί εύκολα κάθε πρόβλημα εκτίμησης λογισμικού, αφού οι εκτιμήσεις βασίζονται σε πραγματικές περιπτώσεις και όχι σε συτήματα που θα πρέπει να ακολουθούνται.

Για τον σχεδιασμό ενός νέου συστήματος λογικής βασισμένης σε περιπτώσεις, θα πρέπει να βρεθούν οι παράγοντες που είναι σημαντικοί για τον καθορισμό της ομοιότητας. Για την εύρεση της ομοιότητας μπορούμε να χρησιμοποιήσουμε τη μέθοδο των πιο κοντινών αλγορίθμων. Ένας κοινός αλγόριθμος δίνεται από (Aha 1991: 4)

$$SIM(C_1.C_2, P) = \frac{1}{\sqrt{\sum_{1 \in P} Feature\_dissimilarity(C_{1j}, C_{2j})}}$$ (Εξ. 17),

όπου P ένα σύνολο χαρακτηριστικών και C_1, C_2 2 διαφορετικές περιπτώσεις.

Τα χαρακτηριστικά ανομοιότητας των περιπτώσεων βρίσκονται ως εξής:

$$Feature\_dissimilarity(C_{1j}, C_{2j}) \begin{cases} (C_{1j} - C_{2j})^2 \\ 0 \\ 1 \end{cases} \quad \text{(Εξ. 18)}$$

Στην πρώτη περίπτωση τα χαρακτηριστικά έναι αριθμητική μεταβλητή, στη δεύτερη είναι κατηγορική μεταβλητή και ισχύει $C_{1j} = C_{2j}$ ενώ στη τρίτη είναι κατηγορική μεταβλητή και ισχύει $C_{1j} \neq C_{2j}$.

5.1.3 Classification and Regression Trees

Ένα δένδρο ταξινόμησης και παλινδρόμησης (Briand et al. 1992: 933) είναι ένα δυαδικό δένδρο, το οποίο δεδομένου μιας μεταβλητής εισαγωγής Χ, παράγει ένα αποτέλεσμα $\hat{Y}$, το οποίο προσεγγίζει κάποια τυχαία μεταβλητή Υ, στοχαστικά συσχετισμένη με τη Χ. Για κάθε εσωτερικό κόμβο του δένδρου, υπάρχει μια δυαδική συνάρτηση της μεταβλητής εισαγωγής Χ και για κάθε εξωτερικό κόμβο, υπάρχει ένα συγκεκριμένο αποτέλεσμα $\hat{Y}$. Αρχίζοντας από τον αρχικό κόμβο η δυαδική συνάρτηση είναι 0 και ακολουθεί το αριστερό κλαδί. Αν το αποτέλεσμα είναι 1 ακολουθεί το δεξί κλαδί. Η διαδικασία επαναλαμβάνεται μέχρι να φτάσουμε σε κάποιο εξωτερικό κόμβο ή φύλλο, όπου η ετικέτα $\hat{Y}$ είναι το αποτέλεσμα. Το δένδρο είναι φτιαγμένο έτσι ώστε να ελαχιστοποιείται η απώλεια μεταξύ Υ και $\hat{Y}$.

Αν ένα διάνυσμα Χ=$\{X_1, \dots, X_n\}$ περιλαμβάνει μια κατηγορική μεταβλητή x, σε ένα σετ Α=$\{x_1, \dots, x_n\}$, τότε η συλλογή των επιτρεπτών ελέγχων στο x περιλαμβάνει τους ελέγχους:

$$A(x) = \begin{cases} 0 \text{ αν } x \in A_0 \\ 1 \text{ αν } x \in A_1 \end{cases},$$

για κάθε διαμερισμό A_0, A_1 του Α.

$\{A_0 \cup A_1 = A$ και $A_0 \cap A_1 = 0\}$

Αν έχουμε x∈Α, τα σημεία $E[Y|X = x]$ σχηματίζονται σε πραγματική ευθεία, τότε υπάρχει ένα οριακό σημείο w τέτοιο ώστε τα x, κάτω από το w ανήκουν στο ιδανικό A_0 και τα x πάνω από το w ανήκουν στο ιδανικό A_1. Άρα θα πρέπει να επιλέξουμε κάθε ένα από τα N-1 πιθανά οριακά σημεία και να επιλέξουμε το διαμερισμό με τη καλύτερη απόδοση, δηλαδή

να ελαχιστοποιεί τη τετραγωνική διαφορά, μεταξύ ταξινομημένων και προβλεπόμενων τιμών.

5.1.4 Naïve Bayes Classifier

Οι Naïve Bayes ταξινομητές βασίζονται στο θεώρημα Bayes και έχουν τη δυνατότητα να προβλέψουν εάν ένα δείγμα ανήκει σε μια συγκεκριμένη κλάση, υποθέτουν ότι η επίδραση μιας τιμής ενός χαρακτηριστικού σε μια δεδομένη κλάση είναι ανεξάρτητη από τις τιμές άλλων χαρακτηριστικών (Murphy, 2006). Η υπόθεση αυτή καλείται ανεξαρτησία υπο συνθήκη και γίνεται για την απλοποίηση των υπολογισμών που περιλαμβάνονται και υπό αυτή την έννοια θεωρούνται αφελείς ("Naïve"). Έχουν δείξει υψηλή ακρίβεια στα δεδομένα κατάταξης για μεγάλα σετ δεδομένων.

Αν υποθέσουμε ένα σύνολο δεδομένων με n χαρακτηριστικά ο ταξινομητής Bayes αναθέτει πιθανότητες $p(C_k/x_1, \dots, x_n)$ για κάθε πιθανή κλάση C_k. Η ύπαρξη μεγάλου αριθμού χαρακτηριστικών καθιστά πολύ δύσκολη εώς αδύνατη η εύρεση μιας τέτοιας πιθανότητας, για τον λόγο αυτό γίνεται χρήση της πιθανοφάνειας Bayes $P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$. (Εξ. 19) (Ren et al. 2009: 945)

Η πιθανότητα $p(C_k/x_1, \dots, x_n)$ μπορεί να γραφεί ως εξής:

$p(C_k/x_1, \dots, x_n) \approx p(C_k, x_1, \dots, x_n) = p(C_k)p(x_1/C_k)p(x_2/C_k)p(x_3/C_k) \approx$
$p(C_k)\prod_{i=1}^{n} p(x_1/C_k)$ (Εξ. 20) (Ren et al. 2009: 945)

Ο ταξινομητής Bayes είναι μια συνάρτηση η οποία θέτει μια ετικέτα κλάσης $\hat{y} = C_k$ για κάποιο k ως εξής: $\hat{y} = arg\max_{k \in K} p(C_k)\prod_{i=1}^{n} p(x_i/C_k)$ (Εξ. 21) (Ren et al. 2009: 945)

5.2 Μέτρα Ακρίβειας

Για τις ανάγκες της πειραματικής διαδικασίας χρησιμοποιήθηκαν τα μέτρα ακρίβειας Mean Magnitude of Relative Error (MMRE), Mean Magnitude of Relative Error to the Estimate (MMER) και Mean of Absolute Error (MAE).

Ονομασία	Συνάρτηση

• Mean Magnitude of Relative Error (Conte et al., 1986)	$MMRE = \frac{1}{n}\sum_{i}^{n}\left\|\frac{Y_i - \hat{Y}_i}{Y_i}\right\|$
• Mean Magnitude of Relative Error to the Estimate (Kitchenham et al., 2001)	$MMER = \frac{1}{n}\sum_{i}^{n}\left\|\frac{Y_i - \hat{Y}_i}{\hat{Y}_i}\right\|$
• Mean of Absolute Error (Kitchenham et al., 2001)	$MAE = \frac{\sum_{i}^{n}\left\|Y_i - \hat{Y}_i\right\|}{n}$

Πίνακας 2. Ολικά μέτρα ακρίβειας που χρησιμοποιήθηκαν.

5.3 Σύνολα Δεδομένων

Η υλοποίηση της πειραματικής διαδικασίας που ακολουθήσαμε βασίστηκε σε 4 σετ δεδομένων, τα οποία είναι δημοσίως διαθέσιμα και έχουν χρησιμοποιηθεί ευρέως στον τομέα εκτίμησης κόστους λογισμικού. Χρησιμοποιήσαμε 4 σύνολα δεδομένων από το PROMISE repository (Shirabad & Menzies 2005). Κάθε σετ δεδομένων περιέχει ένα αριθμό έργων, ένα αριθμό ανεξαρτήτων μεταβλητών και μια εξαρτημένη μεταβλητή, τη προσπάθεια.

5.3.1 Σύνολο Δεδομένων COCOMO81

Το σύνολο δεδομένων COCOMO81 (Boehm, 1981) είναι μία ευρέως διαδεδομένη βάση εκτίμησης κόστους, η οποία έχει χρησιμοποιηθεί για τη βαθμονόμηση του γνωστού μοντέλου COCOMO και αποτελείται από 63 ολοκληρωμένα έργα λογισμικού. Το σύνολο περιλαμβάνει είκοσι δύο ανεξάρτητες μεταβλητές κόστους εκ των οποίων, οι δύο είναι συνεχείς, οι δεκαεπτά διάταξης και οι τρεις ονομαστικής κλίμακας (Πίνακας 3).

Μεταβλητή	Περιγραφή	Επίπεδα
productivity	Παραγωγικότητα ομάδας ανάπτυξης	Συνεχής
duration	Διάρκεια σε μήνες	Συνεχής
year	Χρονολογία ολοκλήρωσης του έργου	Συνεχής
type	Τύπος έργου	1=BUS 2=CTL 3=HMI 4=SCI

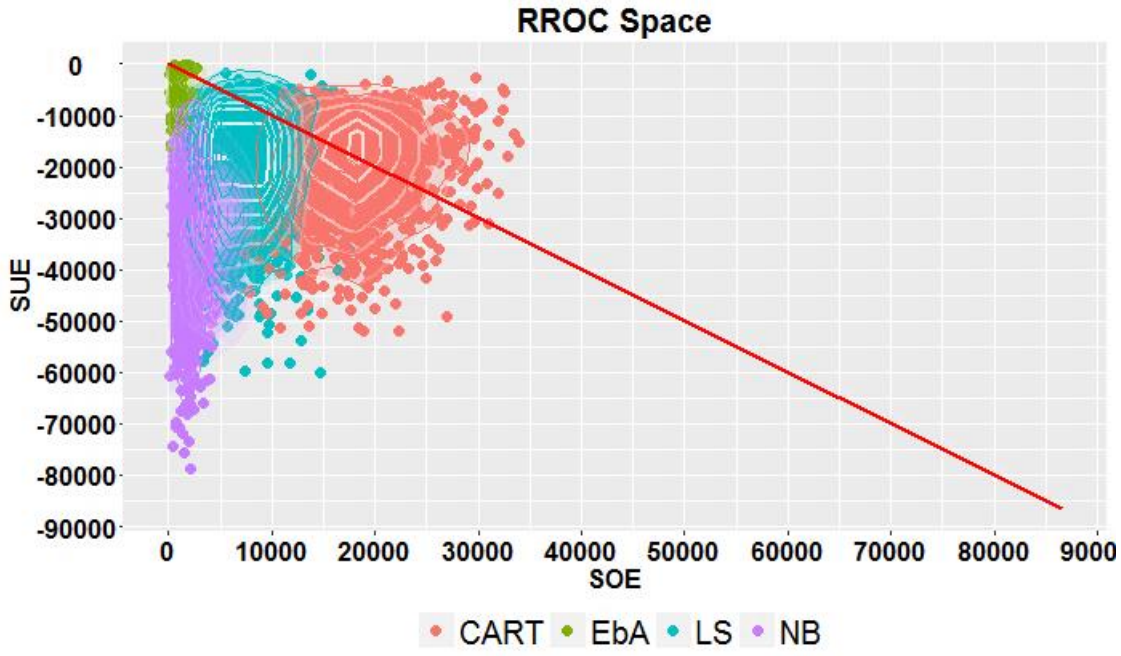
type_c	Τύπος υπολογιστή	5=SUP 6=SYS 1=MAX 2=MID 3=MIN 4=MIC
rmode	Τρόπος ανάπτυξης λογισμικού	1=Embedded
rely	Απαιτούμενη αξιοπιστία λογισμικού	
data	Μέγεθος βάσης δεδομένων	
cplx	Πολυπλοκότητα	
time	Χρονικοί περιορισμοί εκτέλεσης	
stor	Περιορισμοί στη κύρια μνήμη	
virt	Αλλαγές στο σύστημα HW/SW	
turn	Χρόνος απόκρισης υπολογιστή	1=Very low
acap	Ικανότητα αναλυτών	2=Low
aexp	Εμπειρία σε εφαρμογές	3=Nominal
pcap	Ικανότητα προγραμματιστών	4=High
vexp	Εμπειρία με το σύστημα HW/SW	5=Very high
lexp	Εμπειρία στις γλώσσες προγραμματισμού	6=Extra high
cont	Συνέχεια προσωπικού	
modp	Χρήση μοντέρνων πρακτικών προγραμματισμού	
tool	Χρήση εργαλείων προγραμματισμού	
sced	Πίεση από χρονοδιάγραμμα ανάπτυξης	
rvol	Αλλαγές στις απαιτήσεις	

Πίνακας 3. Περιγραφή του συνόλου δεδομένων COCOMO.

95% BOOTSTRAP CONFIDENCE INTERVALS (cocomo)				
	LS	EBA	CART	NB
SOE	[2727.0, 13882.6]	[276.0 , 2019.2]	[9183.2, 29420.6]	[589.5, 4305.3]
SUE	[-46372.3, -5826.6]	[-16483.7, -596.0]	[-42362.3, -5715.9]	[-62228.3, -12683.6]
ME	[-611.5, 68.0]	[-254.3, 10.8]	[-427.2, 286.4]	[-970.5, -144.7]
VAR	[237898, 5376555]	[14988.8, 1108888]	[486281.7, 4761249]	[335212.9, 6852036]
MSE	[176722.9, 5498163]	[3770.3, 1111279]	[431301.6, 4672992]	[169139.6, 7392738]

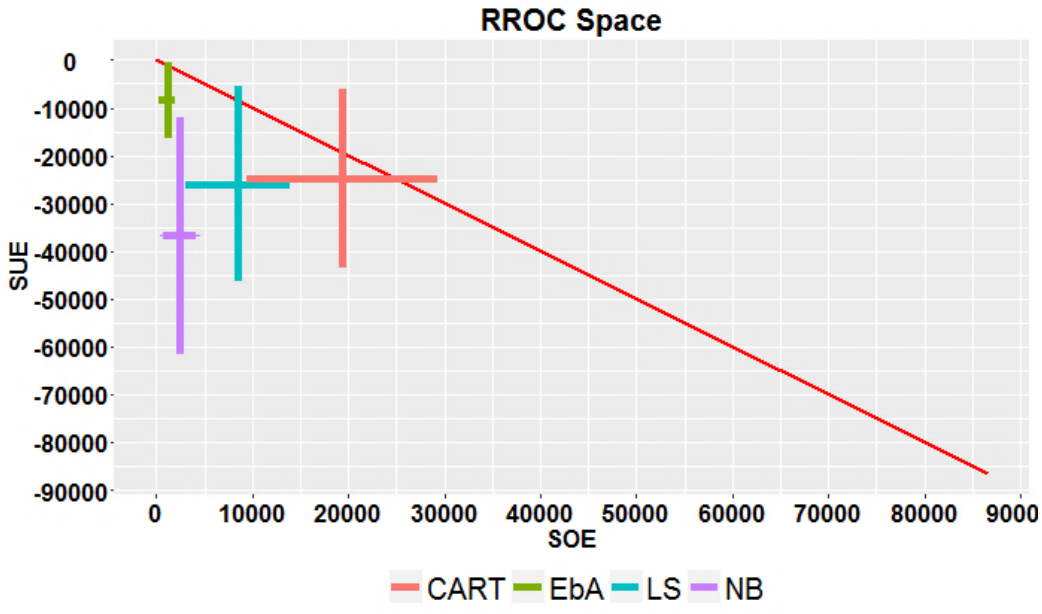
Πίνακας 4. Τα 95% διαστήματα εμπιστοσύνης για το bootstrapped σύνολο δεδομένων COCOMO.

Ο Πίνακας 4 παρουσιάζει τα διαστήματα εμπιστοσύνης για τις τιμές mean error (ME), variance (VAR), και Mean Square Error (MSE). Οι τιμές SOE για τις μεθόδους πρόβλεψης NB και EbA παρουσιάζουν επικαλυπτόμενα διαστήματα που υποδεικνύουν περιοχές με ίση ακρίβεια πρόβλεψης. Από το Πίνακα 4 παρατηρούμε ότι η μέθοδος EbA, για το σύνολο δεδομένων COCOMO, παρουσιάζει το μικρότερο διάστημα mean square error και άρα είναι η πιο έγκυρη μέθοδος πρόβλεψης.



Σχήμα 10. RROC space των bootstrapped παρατηρήσεων σφάλματος για το δείγμα COCOMO

Από το Σχήμα 10, βλέπουμε τις πυκνότητες των παρατηρήσεων σφάλματος για κάθε μοντέλο πρόβλεψης που χρησιμοποιήσαμε στον RROC space. Μπορούμε να διακρίνουμε ότι το μοντέλο EbA έχει τη καλύτερη δυνατότητα πρόβλεψης, σε όρους ακρίβειας, εφ'όσον έχει συσσωρευμένη πυκνότητα παρατηρήσεων κοντά στο ιδανικό σημείο (0,0). Το κέντρο της πυκνότητας παρατηρήσεων του μοντέλου CART, βρίσκεται εντός της γραμμής αμεροληψίας, ενώ η διασπορά των παρατηρήσεων είναι αρκετά μεγάλη και έτσι δεν μπορεί να θεωρηθεί αξιόπιστο. Οι παρατηρήσεις του μοντέλου NB βρίσκονται κάτω από τη γραμμή μεροληψίας, το οποίο συνιστά τάση για υποεκτίμηση κόστους.



Σχήμα 11. Δισδιάτατα 95% διαστήματα εμπιστοσύνης στον RROC space για το bootstrapped σύνολο δεδομένων COCOMO.

Από το Σχήμα 11, μπορούμε να δούμε τα 95 % διαστήματα εμπιστοσύνης για κάθε μοντέλο για το σετ δεδομένων COCOMO. Συμπερασματικά βλέπουμε ότι το μοντέλο EbA είναι το πιο συνεπές στις εκτιμήσεις του εφόσον έχει μικρότερο διάστημα εμπιστοσύνης και άρα μικρότερη διασπορά. Τα μοντέλα NB και EbA παρουσιάζουν πολύ μικρότερες τιμές SOE σε σχέση με τα υπόλοιπα. Το κέντρο παρατηρήσεων του μοντέλου CART βρίσκεται κοντά στη γραμμή αμεροληψίας, ενώ βρίσκεται σε μεγαλύτερη απόσταση από το ιδανικό σημείο (0,0).

Διεξάγαμε σύγκριση με 4 μέτρα ακρίβειας ήτοι τα MMRE, MMER, MAE και το μέτρο σχετικού σφάλματος, το οποίο παρουσιάσαμε στην αρχική σύγκριση.

Model	MMRE	MMER	MAE
LS	1.04	1.14	466.00
EbA	0.06	0.07	113.49
CART	6.0	1.08	627.73
NB	0.77	1.19	569.92

Πίνακας 5. Σύγκριση μέτρων ακρίβειας

για το σύνολο δεδομένων COCOMO.

Το μέτρο ακρίβειας MMRE για το σύνολο δεδομένων COCOMO, προκρίνει ως ιδανικό το μοντέλο EBA, αφού βρίσκεται στο ιδανικό σημείο (0,0), ενώ το μοντέλο CART θεωρείται πολύ χειρότερο σε σχέση με τα υπόλοιπα. Η σύγκριση της ικανότητας πρόβλεψης μπορεί να απεικονισθεί ως εξής: EbA > NB > LS > CART. Για το μέτρο MMER θεωρείται και πάλι ιδανικό το μοντέλο EBA ενώ χειρότερο θεωρείται το NB με μικρή διαφορά από τα LS, CART και NB. Για το μέτρο MAE θεωρείται πολύ καλύτερο το EBA σε σχέση με τα υπόλοιπα ενώ το CART θεωρείται ως το χειρότερο. Συμπερασματικά, βλέπουμε ότι η επίλογη του καλύτερου μοντέλου πρόβλεψης σχετίζεται άμεσα με το μέτρο ακρίβειας που έχει χρησιμοποιήθει.

5.3.2 Σύνολο δεδομένων Maxwell

Το σύνολο δεδομένων Maxwell, που χρησιμοποιήθηκε στο πειραματικό κομμάτι, αποτελείται από 63 ολοκληρωμένα έργα λογισμικού μίας εμπορικής φιλανδικής τράπεζας (Maxwell, 1996). Οι δεκαοκτώ ανεξάρτητες μεταβλητές περιλαμβάνουν όλους τους τύπους δεδομένων (1 συνεχής, 15 διατάξιμες και 2 ονομαστικές) ενώ η εξαρτημένη μεταβλητή είναι η προσπάθεια (effort) που καταβλήθηκε για την ολοκλήρωση των έργων (Πίνακας 6).

Μεταβλητή	Περιγραφή	Επίπεδα
effort	Προσπάθεια που καταβλήθηκε από την ομάδα ανάπτυξης, μετρημένη σε εργατοώρες	Συνεχής
size	Μέτρηση των βαθμών λειτουργίας του συστήματος	Συνεχής
app	Τύπος της εφαρμογής	1=Customer Service 2=M.I.S 3=Transaction Control 4=Production Control 5=Information/on-line service
har	Πλατφόρμα υλικού	1=Networked 2=Mainframe 3=PC 4=Mini Computer 5=Multi-platform
t01	Συμμετοχή πελάτη	

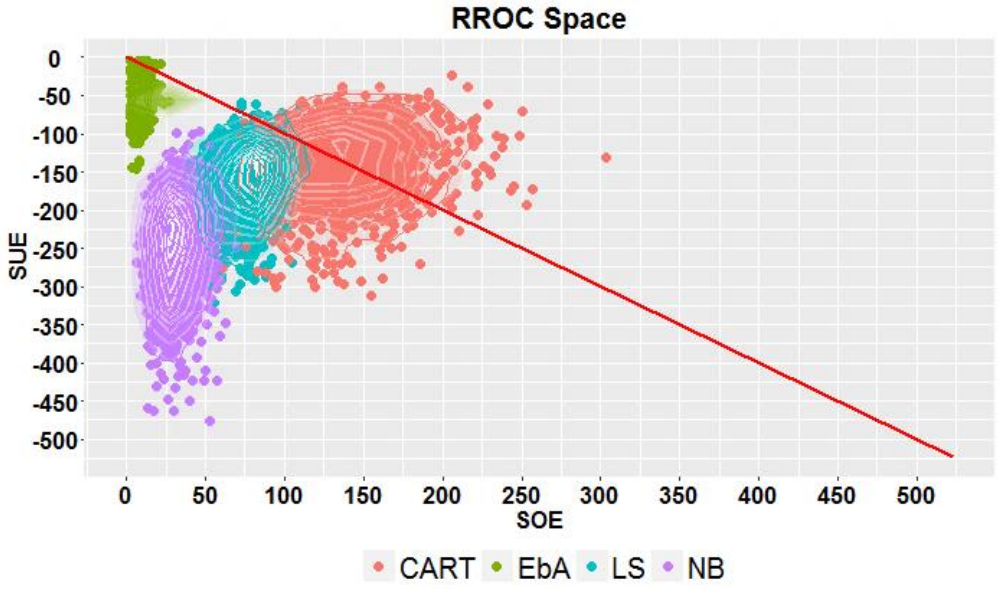
t02	Επάρκεια περιβάλλοντος ανάπτυξης	
t03	Διαθεσιμότητα προσωπικού	
t04	Χρήση προτύπων	
t05	Χρήση μεθόδων	1=Πολύ χαμηλός
t06	Χρήση εργαλείων	2=Χαμηλός
t07	Λογική πολυπλοκότητα λογισμικού	3=Ονομαστικός
		4=Υψηλός
t09	Προδιαγραφές ποιότητας	5=Πολύ υψηλός
t10	Προδιαγραφές αποτελεσματικότητας	
t11	Προδιαγραφές εγκατάστασης	
t12	Ικανότητες ανάλυσης προσωπικού	
t13	Γνώση του πεδίου εφαρμογής του προσωπικού	
t14	Ικανότητες προσωπικού στη χρήση εργαλείων	
t15	Ικανότητες ομάδας ανάπτυξης	

Πίνακας 6. Περιγραφή του συνόλου δεδομένων Maxwell.

95% BOOTSTRAP CONFIDENCE INTERVALS (MAXWELL)				
	LS	EBA	CART	NB
SOE	[49349.1, 106600.2]	[2890.6 , 16604.1]	[80703.73, 216955.1]	[15170.0, 57684.7]
SUE	[-278061, -84737.3]	[-93738.6, -7962.3]	[-261327.4, -64178.4]	[-386382.1, -145606.7]
ME	[-3106.1, 30.3]	[-1376.3, 40.1]	[-2091.5, 1876.4]	[-5760.8, -1657.1
VAR	[16295337, 89076340]	[676374.3, 28659855]	[22174509, 149153099]	[21030870, 149317748]
MSE	[14990807, 93608100]	[280437.4, 29465673]	[21796138, 145724348]	[21960478, 175806710]

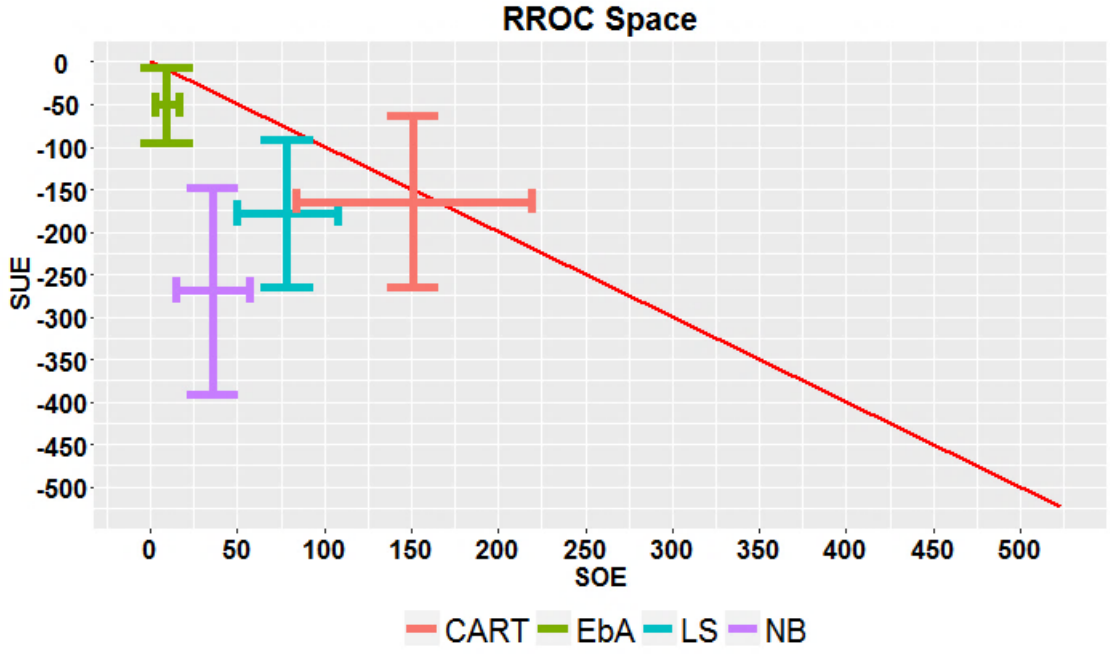
Πίνακας 7. Τα 95% διαστήματα εμπιστοσύνης για το bootstrapped σύνολο δεδομένων MAXWELL.

Από τον Πίνακα 7, βλέπουμε ότι το μοντέλο CART παρουσιάζει το μεγαλύτερο MSE και άρα δεν μπορεί να θεωρηθεί αξιόπιστο, σε αντίθεση το μοντέλο EBA. Οι περιοχές SUE για τα μοντέλα πρόβλεψης LS και CART παρουσιάζουν αλληλοεπικαλυπτόμενα διαστήματα και άρα μοντέλα με ίση ακρίβεια πρόβλεψης σε ορισμένες περιοχές.



Σχήμα 12. RROC space των bootstrapped παρατηρήσεων σφάλματος για το δείγμα Maxwell.

Από το Σχήμα 12, βλέπουμε ότι ο RROC space των bootstrapped παρατηρήσεων του φάλματος για το σετ δεδομένων Maxwell απεικονίζει και πάλι το μοντέλο EbA ως αυτό με τη μικρότερη διασπορά παρατηρήσεων, με μια μικρή τάση για υποεκτίμηση. Για τα NB και LS μπορούμε να συμπεράνουμε ότι παρουσιάζουν μια ξεκάθαρη τάση για υποεκτίμηση κόστους με αρκετά μεγάλη διασπορά εκτιμήσεων, ενώ το μοντέλο CART παρουσιάζει και πάλι τη μεγαλύτερη διασπορά εκτίμήσεων με το κέντρο της πυκνότητας των παρατηρήσεων του σφάλματος να βρίσκεται εντος της γραμμής αμεροληψίας.



Σχήμα 13. Δισδιάτατα 95% διαστήματα εμπιστοσύνης στον RROC space για το δείγμα Maxwell.

Από τα 95% διαστήματα εμπιστοσύνης, που παρουσιάζει το Σχήμα 13, μπορούμε να διακρίνουμε καλύτερα ότι το μοντέλο EbA παρουσιάζει τις μικρότερες τιμές υποεκτίμησης (SUE), ενώ τα μοντέλα EbA και NB τις μικρότερες τιμές υπερεκτίμησης (SOE). Το μοντέλο LS παρουσιάζει μια μικρή τάση για υποεκτίμηση κόστους, ενώ το CART παρουσιάζει μεγάλη διασπορά στις παρατηρήσεις του, ειδικά για υπερεκτίμηση κόστους (SOE).

Model	MMRE	MMER	MAE
LS	0.57	0.66	3916.232
EbA	0.05	0.05	772.1129
CART	0.84	0.59	4591.513
NB	0.48	1.229329	4563.339

Πίνακας 8. Σύγκριση των μοντέλων πρόβλεψης με βάση τα MMRE, MMER και MAE για το σύνολο δεδομένων Maxwell

Από το Πίνακα 8, βλέπουμε ότι το μέτρο ακρίβειας MMRE για το σύνολο δεδομένων MAXWELL, προκρίνει ως ιδανικό το μοντέλο EBA, αφού βρίσκεται στο ιδανικό σημείο (0,0), ενώ το μοντέλο CART θεωρείται χειρότερο με μικρή διαφορά από τα υπόλοιπα. Η σύγκριση της ικανότητας πρόβλεψης μπορεί να απεικονισθεί ως εξής: EbA > NB> LS > CART. Για το

μέτρο MMER θεωρείται και πάλι ιδανικό το μοντέλο EBA ενώ χειρότερο θεωρείται το NB. Στη περίπτωση αυτή η ικανότητα πρόβλεψης απεικονίζεται ως εξής: EbA > CART> LS > NB. Για το μέτρο MAE, θεωρείται πολύ καλύτερο το EBA σε σχέση με τα υπόλοιπα ενώ το NB θεωρείται ως το χειρότερο.

5.3.3 Σύνολο δεδομένων NASA93

Το αρχικό σύνολο περιλαμβάνει 93 έργα λογισμικού από διαφορετικά ερευνητικά κέντρα της NASA με είκοσι τρεις ανεξάρτητες μεταβλητές (22 κατηγορικές και 1 συνεχή), ενώ η εξαρτημένη μεταβλητή είναι η πραγματική προσπάθεια (actual effort) που καταβλήθηκε για την ανάπτυξη των έργων μετρημένη σε εργατομήνες. Μετά την απομάκρυνση πέντε ανεξάρτητων μεταβλητών που δε σχετίζονται με την προσπάθεια (για παράδειγμα unique id, project name κ.τ.λ.), το σύνολο δεδομένων αποτελείται από τις μεταβλητές που παρουσιάζονται στον Πίνακα 9.

Μεταβλητή	Κλίμακα	Περιγραφή	Επίπεδα
effort	Συνεχής	Προσπάθεια σε εργατομήνες	
SLOC	Συνεχής	Γραμμές Κώδικα (χιλιάδες)	
forg	Ονομαστική	Σύστημα αέρους ή εδάφους	Flight, Ground
mode	Ονομαστική	Τρόπος ανάπτυξης λογισμικού	Embedded, Organic, Semidetached
acap	Διατάξιμη	Ικανότητα αναλυτών	
pcap	Διατάξιμη	Ικανότητα προγραμματιστών	
aexp	Διατάξιμη	Εμπειρία σε εφαρμογές	
modp	Διατάξιμη	Χρήση μοντέρνων πρακτικών προγραμματισμού	
tool	Διατάξιμη	Χρήση εργαλείων προγραμματισμού	
vexp	Διατάξιμη	Εμπειρία με το σύστημα HW/SW	
lexp	Διατάξιμη	Εμπειρία στις γλώσσες προγραμματισμού	
sced	Διατάξιμη	Πίεση από χρονοδιάγραμμα ανάπτυξης	
stor	Διατάξιμη	Περιορισμοί στη κύρια μνήμη	
data	Διατάξιμη	Μέγεθος βάσης δεδομένων	

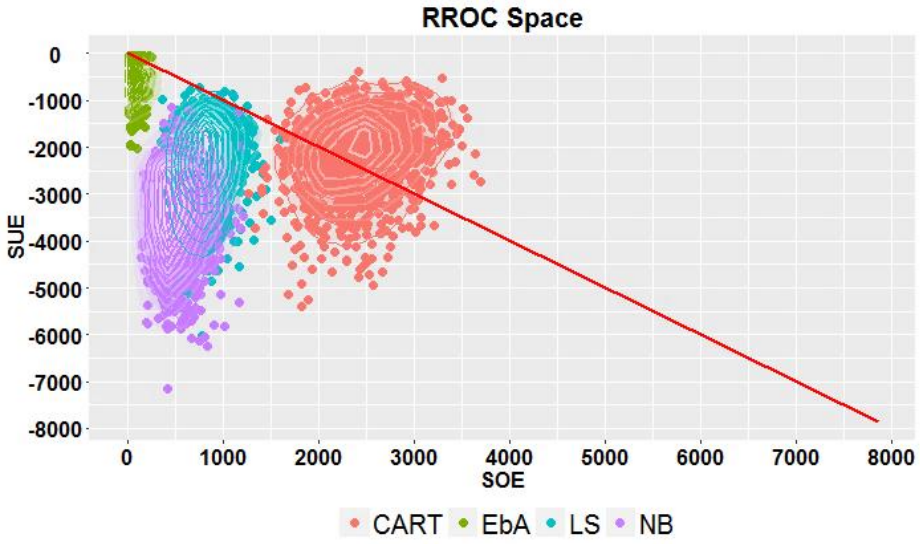
			Πολύ χαμηλός
time	Διατάξιμη	Χρονικοί περιορισμοί εκτέλεσης	Χαμηλός
turn	Διατάξιμη	Χρόνος απόκρισης υπολογιστή	Ονομαστικός
virt	Διατάξιμη	Αλλαγές στο σύστημα HW/SW	Υψηλός
cplx	Διατάξιμη	Πολυπλοκότητα	Πολύ υψηλός
rely	Διατάξιμη	Απαιτούμενη αξιοπιστία λογισμικού	Πάρα πολύ υψηλός

Πίνακας 9. Περιγραφή του συνόλου δεδομένων NASA93 (Idri& Abran, 2000)

95% BOOTSTRAP CONFIDENCE INTERVALS (NASA93)				
	LS	EBA	CART	NB
SOE	[436.6, 1234.7]	[19.3 , 176.3]	[1658.2, 3336.6]	[266.2, 1029.5]
SUE	[-4421.0, -1115.5]	[-1262.3, -61.7]	[-3760.3, -846.3]	[-5504.0, -1843.3]
ME	[-37.8, -1.4]	[-12.8, 0.5]	[-21.2, 21.2]	[-51.9, -12.6]
VAR	[1922.8, 20690.16]	[43.6, 4334.4]	[3274.4, 21061.8]	[2606.1, 22044.4]
MSE	[1619.8, 21471.1]	[12.3, 4363.7]	[3208.3, 20964.8]	[2404.7, 24076.1]

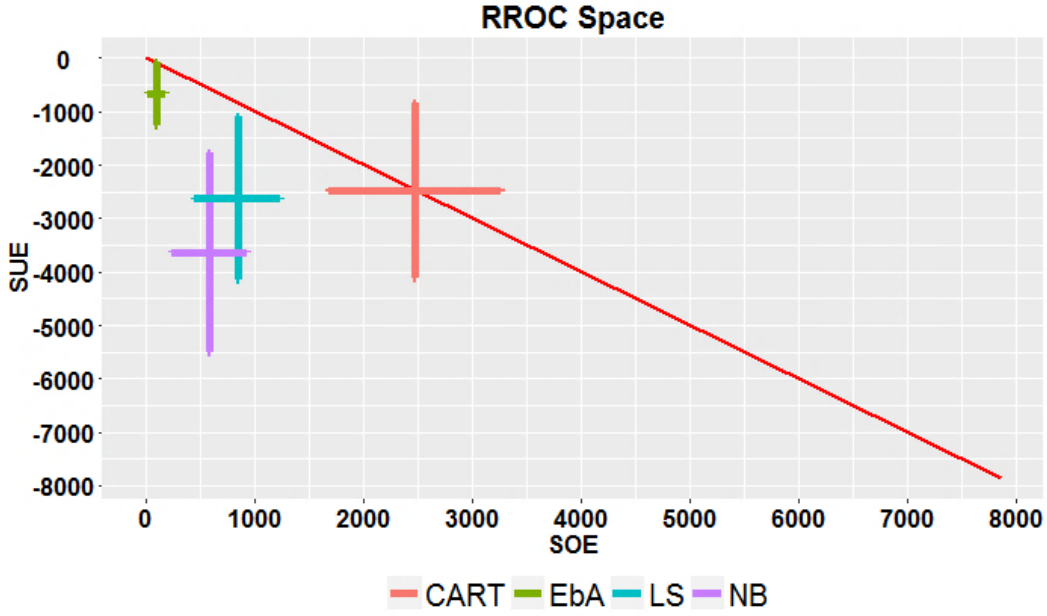
Πίνακας 10. Τα 95% διαστήματα εμπιστοσύνης για το bootstrapped σύνολο δεδομένων NASA93.

Από τον Πίνακα 10, βλέπουμε ότι οι τιμές SOE των NB και LS παρουσιάζουν αλληλοεπικαλυπτόμενα διαστήματα εμπιστοσύνης όπως και οι τιμές SUE των LS, CART και NB. Το μοντέλο EBA είναι αυτό με το μικρότερο MSE και άρα είναι το πιο συνεπές στις εκτιμήσεις του, ενώ το μοντέλο CART το πιο ασταθές.



Σχήμα 14: RROC space των bootstrapped παρατηρήσεων σφάλματος για το δείγμα Nasa93

Ο RROC space, για το δείγμα Nasa93, μας δίνει παραπλήσια με τα προηγούμενα αποτελέσματα. Πιο συγκεκριμένα, οι τιμές του μοντέλου EbA κυμαίνονται πλησίον του ιδανικού σημείου (0,0) με μια ελαφριά ταση υποεκτίμησης κόστους. Το επίκεντρο των παρατηρήσεων σφάλματος για το μοντέλο CART βρίσκεται ακριβώς πιο πάνω από τη ευθεία αμεροληψίας με τη μεγαλύτερη απόσταση Manhattan, σε σχέση με τα υπόλοιπα μοντέλα. Το μοντέλο LS παρουσιάζει ελαφρώς καλύτερα αποτελέσματα από το NB και πολύ καλύτερα από το CART σε όρους ακρίβειας και διασποράς των παρατηρήσεων σφάλματος.



Σχήμα 15: Τα 95% διαστήματα εμπιστοσύνης στον RROC space για το σύνολο δεδομένων Nasa93.

Από το Σχήμα 15 των 95% διαστημάτων εμπιστοσύνης για το δείγμα Nasa93, βλέπουμε ότι το επίκεντρο των παρατηρήσεων του μοντέλου CART βρίσκεται εντός της γραμμής αμεροληψίας, και το CART είναι πιο αμερόληπτο σε σχέση με τα υπόλοιπα. Τα μοντέλα LS, NB παρουσιάζουν ξεκάθαρη μεροληψία για υποεκτίμηση κόστους εφόσον οι παρατηρήσεις τους βρίσκονται στο αριστερό μέρος του σχήματος, ενώ το μοντέλο EbA παρουσιάζει τις μικρότερες τιμές SOE αλλά και SUE. Επιπλέον, βλέπουμε ότι το μοντέλο CART παρουσιάζει τις μεγαλύτερες τιμές SOE.

Model	MMRE	MMER	MAE
LS	0.67	0.93	352.6722
EbA	0.03	0.03	610.0968
CART	2.27	0.91	495.0413
NB	0.76	3.037.427	430.6914

Πίνακας 11. Σύγκριση των μοντέλων πρόβλεψης με βάση τα MMRE, MMER και MAE για το σύνολο δεδομένων Nasa93.

Το μέτρο ακρίβειας MMRE για το σύνολο δεδομένων NASA93, προκρίνει ως ιδανικό το μοντέλο EBA, αφού βρίσκεται στο ιδανικό σημείο (0,0), ενώ το μοντέλο CART θεωρείται πολύ χειρότερο σε σχέση με τα υπόλοιπα. Η σύγκριση της ικανότητας πρόβλεψης μπορεί να απεικονισθεί ως εξής: EbA > LS > NB > CART. Για το μέτρο MMER θεωρείται και πάλι ιδανικό το μοντέλο EBA, ενώ χειρότερο θεωρείται το NB με μικρή διαφορά από τα LS, CART και NB. Στη περίπτωση αυτή η ικανότητα πρόβλεψης απεικονίζεται ως εξής: EbA > CART> LS > NB. Για το μέτρο MAE θεωρείται πολύ καλύτερο το EBA σε σχέση με τα υπόλοιπα ενώ το CART θεωρείται ως το χειρότερο.

5.3.4 Σύνολο δεδομένων Desharnais

Το σετ δεδομένων έχει χρησιμοποιηθεί ευρέως στην εκτίμηση κόστους λογισμικού. Αποτελείται από 81 έργα, τα οποία αναπτύχθηκαν από ένα καναδέζικο οίκο λογισμικού το 1989. Σε 4 έργα υπάρχουν ελλειπείς τιμές και για τον λόγο αυτόν απαλήφθηκαν από τη πειραματικό σχεδιασμό που ακολουθήσαμε, ενώ δεν χρησιμοποιήθηκαν οι διακριτές μεταβλητές, Language και YearEnd.

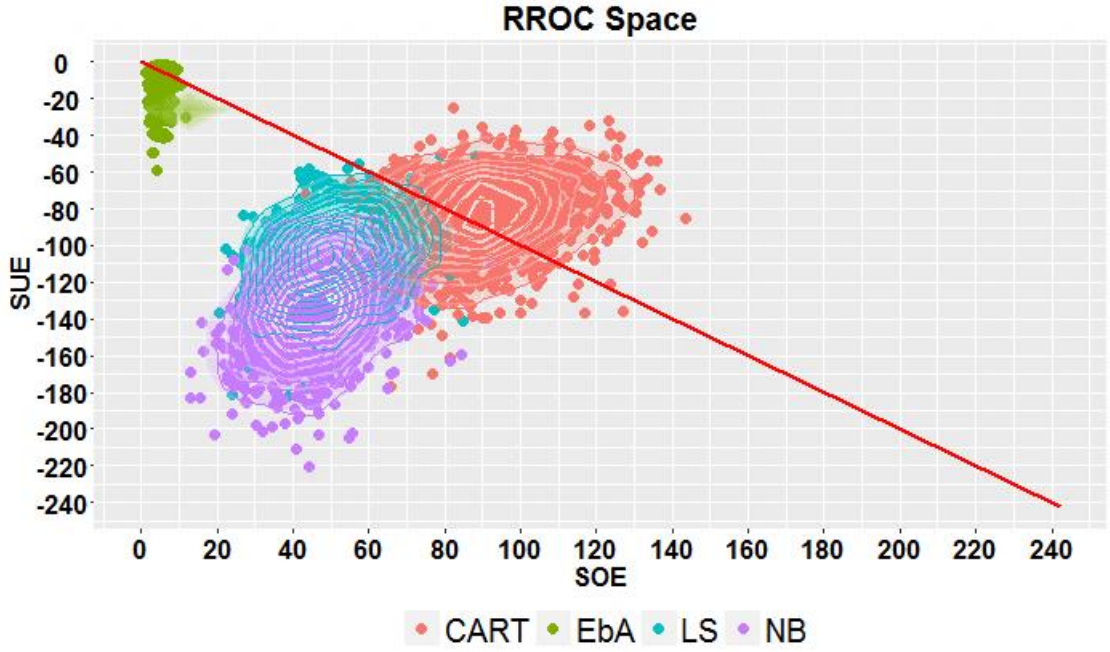
Μεταβλητή	Περιγραφή	Επίπεδα
TeamExp	Team experience	Numerical Measured in years
ManagerExp	Manager's experience	Numerical
YearEnd	Year of end	Numerical
ansactions	Transactions	Numerical Number of transactions
Entities	Entities	Numerical
PointsAdjust	Adjusted function points	Numerical
Envergure	Development environment	Numerical
PointsNonAjust	Unadjusted function points	Numerical
Effort	Development effort	Numerical

Πίνακας 12. Περιγραφή του συνόλου δεδομένων Desharnais. (Desharnais 1989)

95% BOOTSTRAP CONFIDENCE INTERVALS (desharnais)				
	LS	EBA	CART	NB
SOE	[30767.2, 76058.2]	[2281.7 , 8267.2]	[57309.4, 125145.2]	[25889.9, 71509.4]
SUE	[-148680.9, -68434.2]	[-32995.8, -2361.9]	[-127624.5, -49489.2]	[-182835.3, -95966.5]
ME	[-1414.3, -52.3]	[-354.9, 55.0]	[-739.6, 768.9]	[-1896.6, -454.4]
VAR	[5415491, 14223776]	[47708.4, 3179526]	[6311280, 18553295]	[5977676, 18000817]
MSE	[5308501, 15072262]	[25791.0, 3188117]	[6251561, 18520658]	[6474347, 20353304]

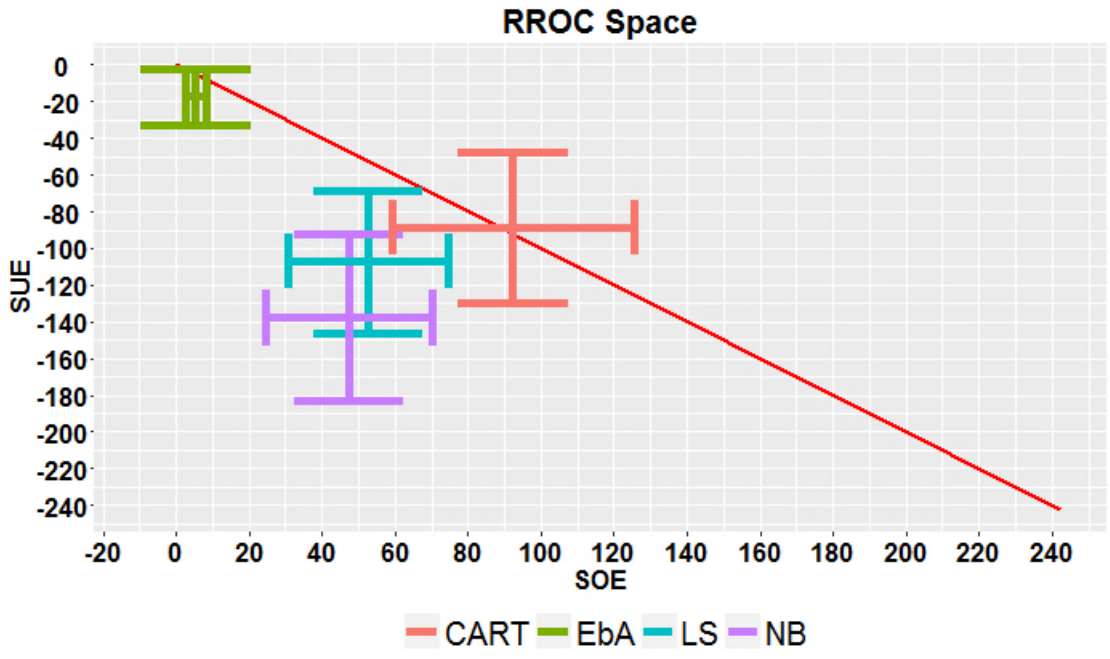
Πίνακας 13. Τα 95% διαστήματα εμπιστοσύνης για το bootstrapped σύνολο δεδομένων NASA93.

Από τον Πίνακα 13, βλέπουμε ότι οι τιμές SOE και SUE των LS και NB παρουσιάζουν παρόμοιες εκτιμήσεις, εφόσον οι αντίστοιχες περιοχές των διαστημάτων εμπιστοσύνης τους αλληλοεπικαλύπτονται.



Σχήμα 16: RROC space των bootstrapped παρατηρήσεων σφάλματος για το σετ δεδομένων Desharnais

Για το σετ δεδομένων Desharnais, το μοντέλο Eba παρουσιάζει και πάλι τη καλύτερη δυνατότητα πρόβλεψης και το CART αμεροληψία με μεγάλη διασπορά και κρίνεται ακατάλληλο για έγκυρες εκτιμήσεις. Το μοντέλο LS παρουσιάζει μεγαλύτερη ακρίβεια πρόβλεψης από το NB εφόσον οι παρατηρήσεις του βρίσκονται εγγύτερα στο (0,0), αν μετρήσουμε με απόσταση Manhattan.



Σχήμα 17: Δισδιάτατα 95% διαστήματα εμπιστοσύνης στον RROC space για το το σετ δεδομένων Desharnais

Από τα 95% διαστήματα εμπιστοσύνης, για το το σετ δεδομένων Desharnais, βλέπουμε ότι το μοντέλο EbA παρουσιάζει πολύ μικρές τιμές SOE, ενώ το CART τις μεγαλύτερες. Το μοντέλο NB παρουσιάζει μεγαλύτερη τάση υποεκτίμησης σε σχέση με το LS, με τη διασπορά των παρατηρήσεων τους να είναι παρόμοια.

Model	MMRE	MMER	MAE
LS	0.60	0.46	2039.13
EbA	0.03	0.03	230.96
CART	0.75	0.43	2271.96
NB	0.55	0.74	2359.60

Πίνακας 14. Σύγκριση των μοντέλων πρόβλεψης με βάση τα MMRE, MMER και MAE.

Το μέτρο ακρίβειας MMRE για το σύνολο δεδομένων DESHARNAIS, προκρίνει ως καλύτερο το μοντέλο EBA, αφού βρίσκεται στο ιδανικό σημείο (0,0), ενώ το μοντέλο CART θεωρείται πολύ χειρότερο σε σχέση με τα υπόλοιπα. Η σύγκριση της ικανότητας πρόβλεψης μπορεί να

απεικονισθεί ως εξής: EbA > NB> LS > CART. Για το μέτρο MMER θεωρείται και πάλι ιδανικό το μοντέλο EBA ενώ χειρότερο θεωρείται το NB με μικρή διαφορά από τα LS, CART και NB. Στη περίπτωση αυτή η ικανότητα πρόβλεψης απεικονίζεται ως εξής: EbA > CART> LS > NB. Για το μέτρο MAE, θεωρείται πολύ καλύτερο το EBA σε σχέση με τα υπόλοιπα, ενώ το NB θεωρείται ως το χειρότερο.

Ο Πίνακας 15 δημιουργήθηκε για την επίτευξη ολικής σύγκρισης των εναλλακτικών μοντέλων, που κάναμε χρήση στη πειραματική διαδικασία. Η σειρά κατάταξης γίνεται με βάση το μικρότερο διάστημα εμπιστοσύνης για τα μέτρα Mean Error και Variance, ενώ για τα μέτρα ακρίβειας MMRE, MMER και MAE, μοντέλα με καλύτερη απόδοση θεωρούνται αυτά με το μικρότερο μέσο.

Μοντέλο	ME	VAR	MMRE	MMER	MAE	Μέση Κατάταξη
COCOMO81						
LS	2	3	3	3	2	2.6
EbA	1	1	1	1	1	1.0
CART	3	2	4	2	4	3.0
NAIVE	4	4	3	4	3	3.6
Maxwell						
LS	2	2	3	3	1	2,2
EbA	1	1	1	1	4	1,6
CART	3	3	4	2	3	3
NAIVE	4	4	2	4	2	3,2
NASA93						
LS	2	3	2	3	1	2,2
EbA	1	1	1	1	4	1,6
CART	4	2	4	2	3	3
NAIVE	3	4	3	4	2	3,2
Desharnais						
LS	2	2	3	3	3	2,6
EbA	1	1	1	1	1	1
CART	4	4	4	2	2	3,2
NAIVE	3	3	2	4	4	3,2

Πίνακας 15. Σύγκριση των μοντέλων πρόβλεψης με βάση τα MMRE, MMER και MAE και τα σύνολα δεδομένων COCOMO81, NASA93, MAXWELL και DESHARNAIS.

Από το Πίνακα 15, μπορούμε να συμπεράνουμε ότι η ικανότητα πρόβλεψης διαφοροποιείται ανάλογα με το σύνολο δεδομένων και το μέτρο ακρίβειας που κάναμε χρήση. Για την εξαγωγή αμερόληπτων συμπερασμάτων δεν θα πρέπει να βασιστούμε σε ένα και μονο μέτρο ακρίβειας. Απόδειξη αυτού, το μοντέλο CART για το σύνολο δεδομένων Maxwell και με βάση το μέτρο MMER παρουσιάζει τις δεύτερες καλύτερες επιδόσεις ενώ με βάση το μέτρο MMRE παρουσιάζει τις χειρότερες. Ωστόσο, από τη μέση κατάταξη των μοντέλων ΕΚΛ που κάναμε χρήση στη πειραματική διαδικασία, προκύπτει ότι το μοντέλο EbA παρουσιάζει τις καλύτερες επιδόσεις, ενώ ως χειρότερο μοντέλο μπορεί να θεωρηθεί ο Naïve Bayes classifier. Η απόδοση πρόβλεψης με βάση τη μέση κατάταξη μπορεί να εκφραστεί μέσω της σχέσης EbA>LS>CART>Naïve.

Κεφάλαιο 6

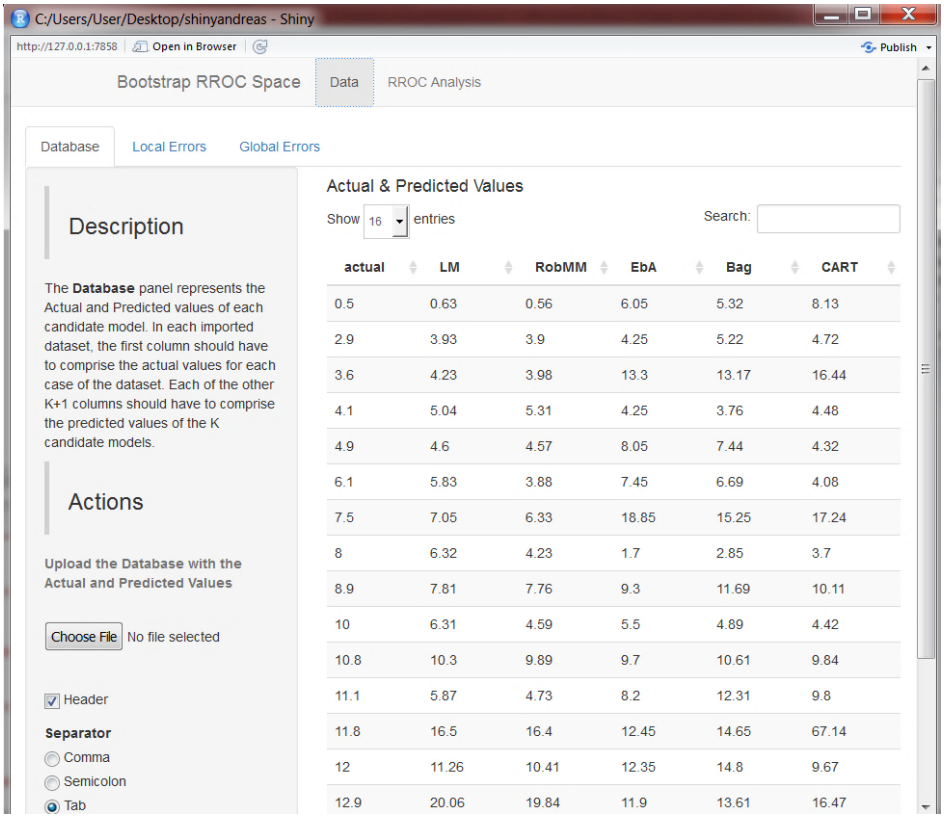
Υλοποίηση Web-Based Εφαρμογής

Η διεξαγωγή σύγκρισης των μοντέλων ΕΚΛ απαιτεί χρόνο και εξειδικευμένες γνώσεις για την υλοποίηση. Ένας επαγγελματίας ή ερευνητής, ο οποίος δεν έχει τις απαιτούμενες γνώσεις ή χρόνο μπορεί να καταφύγει σε όχι τόσο αξιόπιστες διαδικασίες σύγκρισης. Αυτό θα ήταν κρίσιμης σημασίας για μια ερευνητική μελέτη ή για την αποπεράτωση ενός έργου λογισμικού. Για τους λόγους αυτούς κρίνεται άμεση η ανάγκη ύπαρξης ενός εργαλείου έυκολα προσβάσιμου στο κοινό, με το οποίο ο επαγγελματίας να είναι σε θέση να διεξάγει έγκυρη και αξιόπιστη σύγκριση μεθόδων ΕΚΛ .

Για την ικανοποίηση της ανάγκης αυτής προχωρήσαμε στην υλοποίηση μιας web-based και φιλικής προς το χρήστη εφαρμογής, η οποία δεν απαιτεί εξειδικευμένες γνώσεις για την χρήση και λειτουργία της. Με το εργαλείο αυτό ο επαγγελματίας έχει τη δυνατότητα να διεξάγει άμεση σύγκριση μοντέλων ΕΚΛ από το χώρο του, με την απλή είσαγωγή ενός αρχείου. Με την οπτικοποίηση των ικανοτήτων πρόβλεψης κάθε μοντέλου ΕΚΛ στον RROC, ο χρήστης είναι σε θέση να κατανοήσει έυκολα και απλά τις ιδιότητες κάθε ενός από αυτά. Αξίζει να αναφέρουμε ότι δεν υπάρχει περιορισμός στον αριθμό των μοντέλων για τα οποία μπορεί να διεξαχθεί σύγκριση, ενώ η διαδικασία σύγκρισης γίνεται με βάση τις πιο γνωστές στατιστικές συναρτήσεις σφάλματος που έχουν χρησιμοποιηθεί στη βιβλιογραφία. Η εφαρμογή κάνει χρήση 6 στατιστικών συναρτήσεων σφάλματος ήτοι τα Absolute Error (AE), Simple Error (SE), Magnitude Relative Error (MRE), Magnitude Relative Error to the Estimate (MER), Balance Relative Error (BRE) και Inverted Balance Relative Error (IBRE), οι οποίες μπορούν να χρησιμοποιηθούν είτε για κάθε έργο ξεχωριστά είτε για ένα σύνολο έργων.

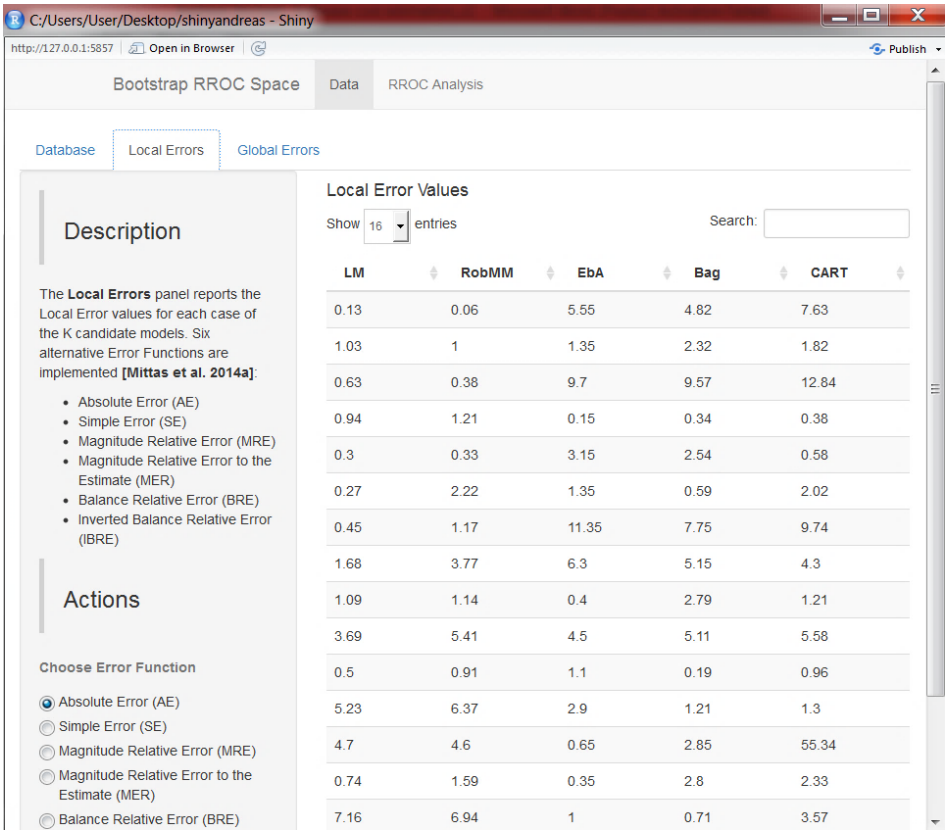
Η ύπαρξη της δυνατότητας εξαγωγής μεγάλου αριθμού δειγμάτων με αναδειγματολειψία bootstrap, κάνει εφικτή τη μελέτη της διαπορά του σφάλματος πρόβλεψης. Με το τρόπο αυτό ο επαγγελματίας είναι σε θέση να γνωρίζει κατά πόσο ένα μοντέλο είναι αξιόπιστο και συνεπές στις προβλέψεις του. Επιπρόσθετα, εξάγωντας τα 95% διαστήματα εμπιστοσύνης της κατανομής των σφαλμάτων πρόβλεψης, μπορεί να είναι πιο σίγουρος για τα αποτελέσματα σύγκρισης.

Η υλοποίηση της web-based εφαρμογής γίνεται με βάση το πακέτο Shiny της γλώσσας R. Στα πιο κάτω σχήματα παρουσιάζονται οι δυνατότητες της εφαρμογής.



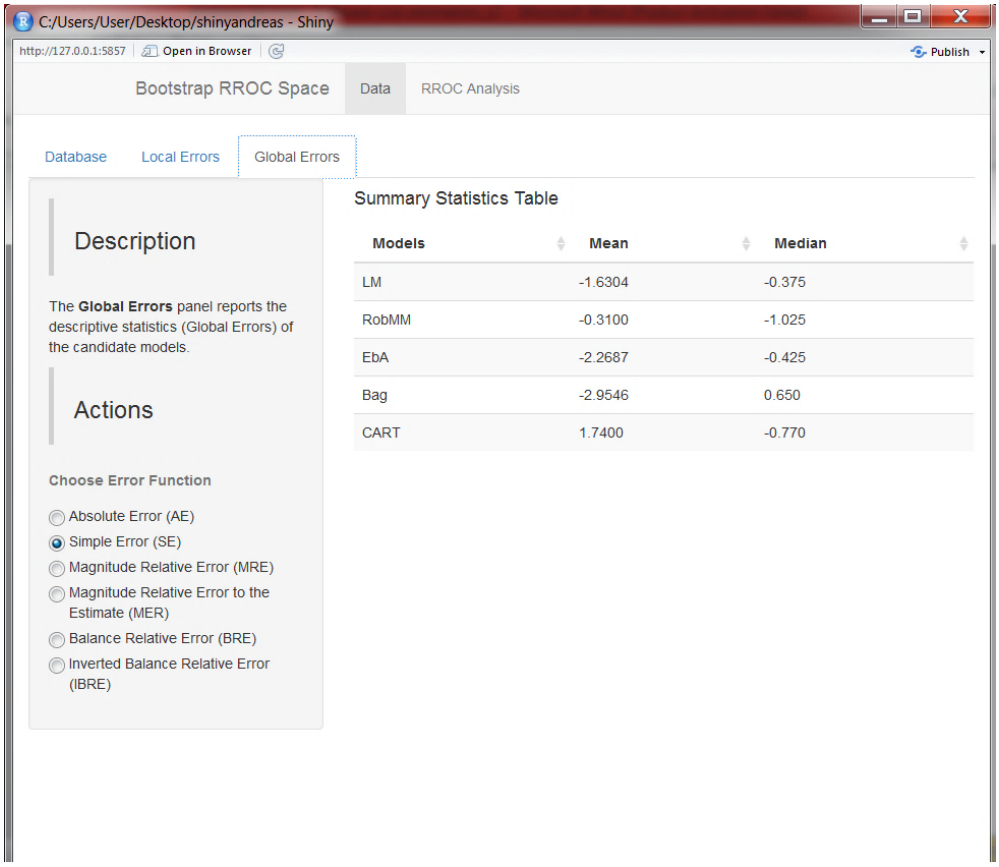
Σχήμα 18: Απεικόνιση της υλοποίησης της web-based εφαρμογής για το σκοπό του “uploading” συνόλου δεδομένων.

Η εφαρμογή χρησιμοποιεί αυτόματα το σύνολο δεδομένων Albrecht σε περίπτωση που δεν έχει επιλεχθεί κάποιο άλλο αρχείο. Ο χρήστης θα πρέπει να ανεβάσει ένα αρχείο στην εφαρμογή το οποίο θα εμπεριέχει τόσο τις πραγματικές τιμές κόστους για ένα σύνολο έργων αλλά και τις εκτιμώμενες τιμές πρόβλεψης για κάποιο αριθμό μοντέλων.



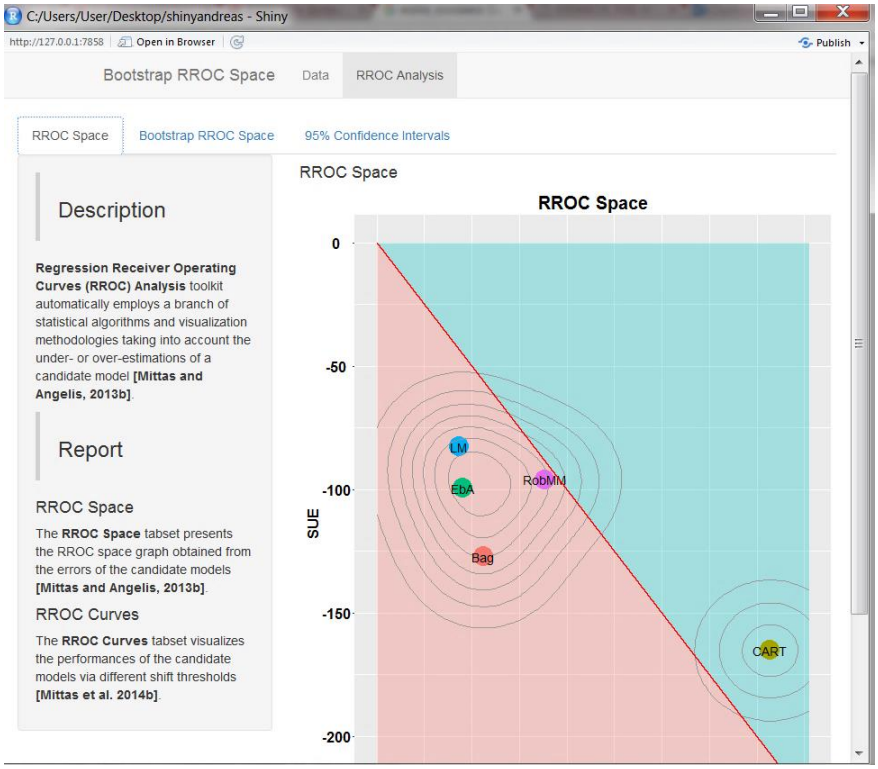
Σχήμα 19: Επιλογή τοπικής συναρτήσεως σφάλματος

Με την επιλογή μιας τοπικής συνάρτησης σφάλματος, ο χρήστης μπορεί να εντοπίσει άμεσα το καλύτερο μοντέλο ΕΚΛ για κάθε έργο ξεχωριστά.



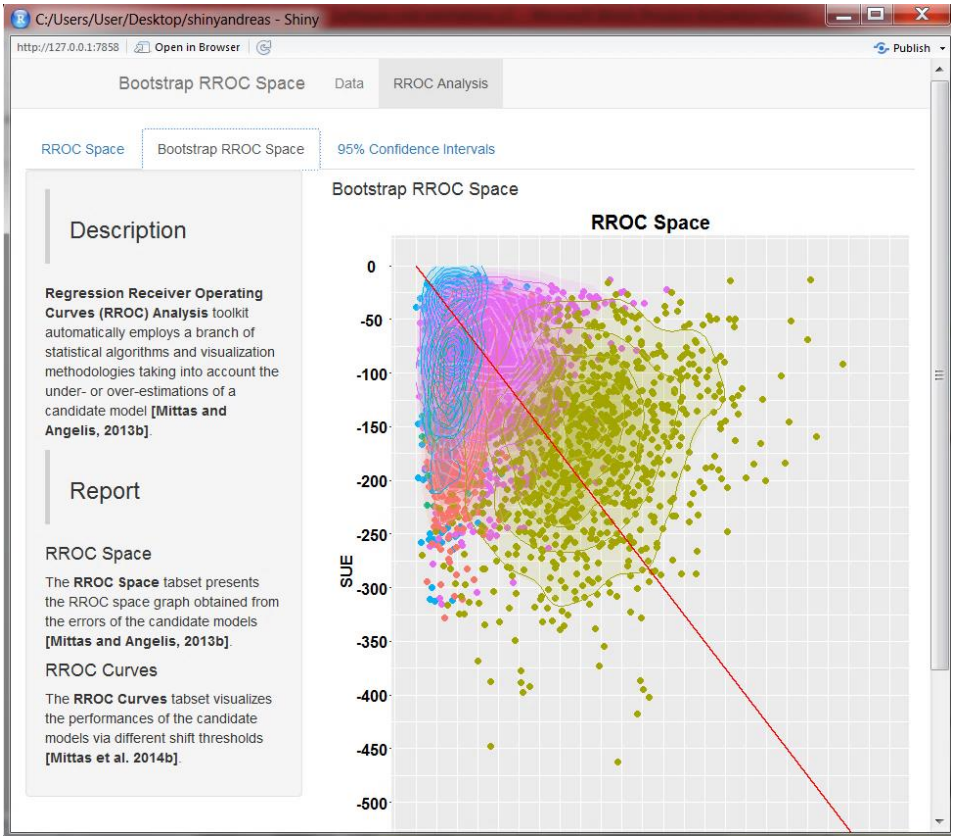
Σχήμα 20: Επιλογή καθολικού μέτρου ακρίβειας

Όπως αναφέραμε και πρωτύτερα υπάρχει η επιλογή της διαδικασίας σύγκρισης μέσω διάφορων καθολικών μέτρων ακριβείας. Ο χρήστης μπορεί να διεξάγει σύγκριση με όποιο μέτρο ακρίβειας θεωρεί ως καλύτερο. Τα αποτελέσματα σύγκρισης μπορεί να διαφέρουν ριζικά, για το λόγο αυτό θα πρέπει να χρησιμοποιούνται περισσότερα από ένα, ούτως ώστε ο χρήστης να έχει πλήρη εικόνα του σφάλματος πρόβλεψης.



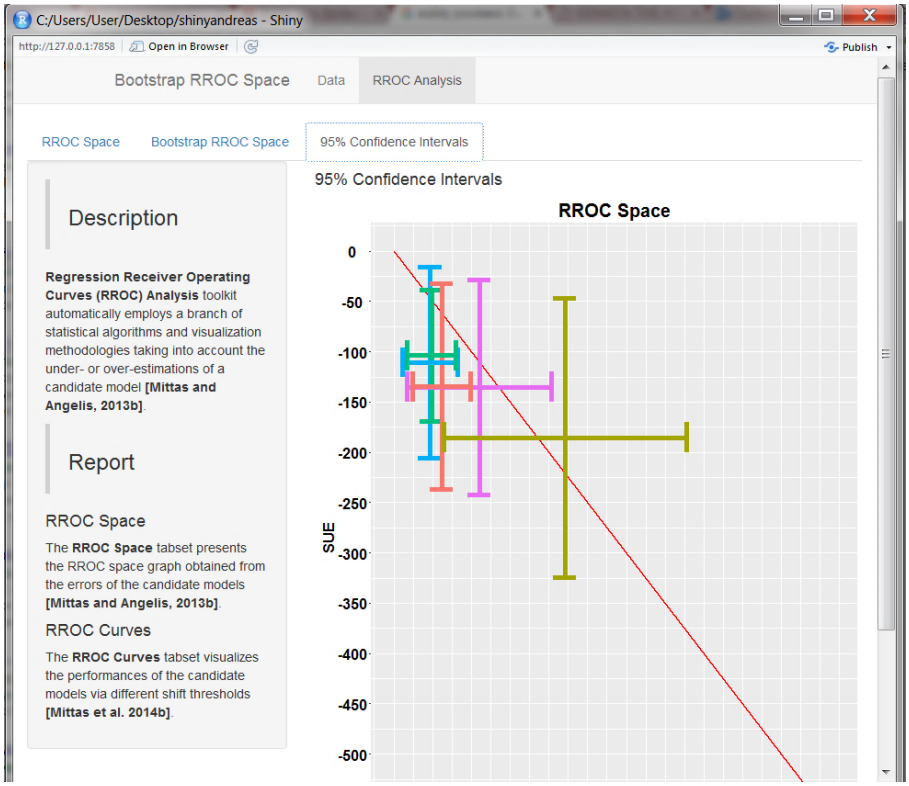
Σχήμα 21: Απεικονίζεται ο RROC space του συνόλου δεδομένων Albrecht.

Στο Σχήμα 21 απεικονίζεται ο RROC space για το σύνολο δεδομένων Albrecht. Λόγω του μικρου αριθμού έργων από τα οποία αποτελείται αλλά και της ύπαρξης ακραίων τιμών, ο χρήστης δεν είναι σε θέση να εξάγει αξιόπιστα και έγκυρα αποτελέσματα.



Σχήμα 22: Απεικονίζεται ο RROC space για τα bootstrapped σφάλματα του συνόλου δεδομένων Albrecht.

Η τεχνητή αναδειγματολειψία μπορεί να μας βοηθήσει στην εξαγωγή αποτελεσμάτων για ένα μεγάλο αριθμό δειγματών και να εντοπίσουμε τη διασπορά του σφάλματος πρόβλεψης. Μπορεί να θεωρηθεί ως αξιόπιστη διαδικασία εφόσον η διαδικασία γίνεται για μεγάλο αριθμό δειγμάτων και με επανάθεση. Επιπλέον ο χρήστης είναι σε θέση να εντοπίσει κατά πόσον ένα μοντέλο είναι συνεπές στις προβλέψεις του, δηλαδή εάν και εφόσον το σφάλμα πρόβλεψης που παράγει είναι της ίδιας τάξης.



Σχήμα 23: Απεικονίζονται τα 95% διαστήματα εμπιστοσύνης για τα bootstrapped σφάλματα του συνόλου δεδομένων Albrecht.

Με την εξαγωγή του 95% διαστήματα εμπιστοσύνης για το bootstrapped συνόλο δεδομένων στον RROC space, ο χρήστης είναι σε θέση να μειώσει την αβεβαιότητα που αφορά το εγγενές σφάλμα πρόβλεψης.

Κεφάλαιο 7

Επίλογος

7.1 Συμπεράσματα και μελλοντικές επεκτάσεις

Στη πειραματική μελέτη της μεταπτυχιακής Διατριβής, κύριος στόχος μας ήταν η εξαγωγή συμπερασμάτων σε σχέση με τις στατιστικές ιδιότητες και την ικανότητα πρόβλεψης των εναλλακτικών μοντέλων εκτίμησης κόστους λογισμικού. Για την υλοποίηση συγκριτικής δυνατότητας και οπτικοποίησης αυτής, κάναμε χρήση της ανάλυσης του RROC space. Διεξάγαμε σύγκριση 4 μοντέλων πρόβλεψης και οπτικοποίηση των ικανοτήτων πρόβλεψης τους, για 4 σετ δεδομένων. Εξετάσαμε την ικανότητα πρόβλεψης των μεθόδων εκτίμησης υπό 3 οπτικές γωνίες, την ακρίβεια, τη μεροληψία και τη διασπορά. Συμπεράναμε ότι για την έγκυρη επιλογή ενός μοντέλου ΕΚΛ θα πρέπει να λαμβάνουμε υπόψη και τους 3 πιο πάνω παράγοντες. Για παράδειγμα ένα μοντέλο μπορεί να είναι σχετικά ακριβές αλλά αυτό δεν σημαίνει ότι είναι και αμερόληπτο, ενώ ένα μοντέλο με μεγάλη διαπορά σφάλματος δεν μπορεί να θεωρηθεί ως αξιόπιστο. Πολύ συχνά, μια τάση για μεροληψία μπορεί να παρουσιάσει δραματικότερες συνέπειες σε σχέση με μια τάση για μεροληψία προς την άλλη πλευρά, είτε υπερεκτίμηση είτε υποεκτίμηση κόστους. Η ακριβής γνώση της μεροληπτικής τάσης ενός μοντέλου ΕΚΛ μπορεί να είναι καθοριστικός παράγοντας για την ορθή εξέλιξη του έργου. Επιπλέον, αν για παράδειγμα ο επαγγελματίας κοστολογήσει στον πελάτη ένα υποεκτιμημένο έργο, μπορεί να οδηγήσει σε ζημιές και κακή φήμη για τον οργανισμό ανάπτυξης. Για τους πιο πάνω λόγους, οι διοικητές έργων θα πρέπει να λαμβάνουν υπόψη τους ότι κάθε έργο λογισμικού είναι μοναδική περίπτωση και να μελετούνται όλα τα χαρακτηριστικά των συνόλων δεδομένων αλλά και των μοντέλων ΕΚΛ.

Για την υπερκέραση προβλημάτων της ύπαρξης ακραίων και ελλειπούσων τιμών, αλλά και τον εντοπισμό της διασποράς του σφάλματος πρόβλεψης, συνιστούμε τη χρήση τεχνικών αναδειγματολειψίας όπως η μη παραμετρική bootstrap. Με τον τρόπο αυτό, δημιουργείται τεχνητά ένα σύνολο δεδομένων αρκετά μεγάλο έτσι ώστε η ύπαρξη ορισμένων ακραίων ή

ελλειπούσων τιμών, να μην είναι καθοριστικής σημασίας για επιλογή ενός μοντέλου ΕΚΛ. Για τη μείωση της αβεβαιότητας σχετικά με τη προβλεπτική ικανότητα μιας τεχνικής, εντοπίσαμε τα 95% διαστήματα εμπιστοσύνης για κάθε μοντέλο και σύνολο δεδομένων.

Με την οπτικοποίηση του RROC space είμαστε σε θέση να εντοπίσουμε τις χαρακτηριστικές αδυναμίες κάθε μοντέλου ΕΚΛ. Με τη χρήση των RROC curves και τη κατασκευή ενός αλγόριθμου προσαρμοσμένου στο σφάλμα πρόβλεψης, θα μπορούσαμε να μελετήσουμε τη μείωση των αδυναμιών αυτών. Επιπλέον, θα μπορούσαμε να υλοποιήσουμε ένα web-based εργαλείο, το οποίο θα ελαχιστοποιεί το σφάλμα πρόβλεψης και θα παράγει πιο αξιόπιστες προβλέψεις.

Βιβλιογραφία

1. Abran, A., & Robillard, P. N. (1996). Function points analysis: An empirical study of its measurement processes. IEEE Transactions on Software Engineering, 22(12), 895-910.
2. Aha, D. W. (1991, May). Case-based learning algorithms. In Proceedings of the 1991 DARPA Case-Based Reasoning Workshop (Vol. 1, pp. 147-158).
3. Albrecht, A. J., & Gaffney, J. E. (1983). Software function, source lines of code, and development effort prediction: a software science validation. IEEE transactions on software engineering, (6), 639-648.
4. Bij, J. B., Edu, R. P. I., & Bennek, K. P. B. (2003). Regression error characteristic curves. In Twentieth international conference on machine learning (ICML-2003). Washington, DC.
5. Boehm, B. W. (1981). Software engineering economics (Vol. 197). Englewood Cliffs (NJ): Prentice-hall.
6. Boehm, B., Abts, C., & Chulani, S. (2000). Software development cost estimation approaches—A survey. Annals of software engineering, 10(1-4), 177-205.
7. Briand, L. C., Basili, V. R., & Thomas, W. M. (1992). A pattern recognition approach for software engineering data analysis. IEEE transactions on software engineering, 18(11), 931-942.
8. Briand, L. C., El Emam, K., Surmann, D., Wieczorek, I., & Maxwell, K. D. (1999, May). An assessment and comparison of common software cost estimation modeling techniques. In Proceedings of the 21st international conference on Software engineering (pp. 313-322). ACM.
9. Brooks, F. (1986). The mythical man-month. Tutorial on Softzuare Design Techniques, 35-42.
10. Budd, C. S., & Cooper, M. J. (2005). Improving on-time service delivery: The case of project as product. Human Systems Management, 24(1), 67-81.
11. Conte, S. D., Dunsmore, H. E., & Shen, V. Y. (1986). Software engineering metrics and models. Benjamin-Cummings Publishing Co., Inc..
12. Desharnais, J. M. (1989). Analyse statistique de la productivitie des projets informatique a partie de la technique des point des fonction. University of Montreal.
13. Efron, B., & Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. Genetic epidemiology, 23(1), 70-86.

14. Foss, T., Stensrud, E., Kitchenham, B., & Myrtveit, I. (2003). A simulation study of the model evaluation criterion MMRE. Software Engineering, IEEE Transactions on, 29(11), 985-995.
15. Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.
16. Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. Neural computation, 4(1), 1-58.
17. Hernández-Orallo, J. (2013). ROC curves for regression. Pattern Recognition, 46(12), 3395-3411.
18. Idri, A., & Abran, A. (2000). Towards a fuzzy logic based measures for software projects similarity. Proceedings of The MCSEAI.
19. Jeffery, R., Ruhe, M., & Wieczorek, I. (2001). Using public domain metrics to estimate software development effort. In Software Metrics Symposium, 2001. METRICS 2001. Proceedings. Seventh International (pp. 16-27). IEEE.
20. Jorgensen, M., & Shepperd, M. (2007). A systematic review of software development cost estimation studies. Software Engineering, IEEE Transactions on, 33(1), 33-53.
21. Kemerer, C. F. (1987). An empirical validation of software cost estimation models. Communications of the ACM, 30(5), 416-429.
22. Kitchenham, B. A., Pickard, L. M., MacDonell, S. G., & Shepperd, M. J. (2001, June). What accuracy statistics really measure [software estimation]. In Software, IEE Proceedings- (Vol. 148, No. 3, pp. 81-85). IET.
23. Kitchenham, B., & Mendes, E. (2009, May). Why comparative effort prediction studies may be invalid. In Proceedings of the 5th international Conference on Predictor Models in Software Engineering (p. 4). ACM.
24. Lederer, A. L., & Prasad, J. (1995). Causes of inaccurate software development cost estimates. Journal of systems and software, 31(2), 125-134.
25. Leung, H., & Fan, Z. (2002). Software cost estimation. Handbook of Software Engineering, Hong Kong Polytechnic University, 1-14.
26. Mair, C., & Shepperd, M. (2005, November). The consistency of empirical comparisons of regression and analogy-based software project cost prediction. In Empirical Software Engineering, 2005. 2005 International Symposium on (pp. 10-pp). IEEE.
27. Maxwell, K. D., Van Wassenhove, L., & Dutta, S. (1996). Software development productivity of European space, military, and industrial applications. Software Engineering, IEEE Transactions on, 22(10), 706-718.
28. Menzies, T., & Shepperd, M. (2012). Special issue on repeatable results in software engineering prediction. Empirical Software Engineering, 17(1), 1-17.
29. Mittas, N., & Angelis, L. (2008a). Comparing cost prediction models by resampling techniques. *Journal of Systems and Software*, *81*(5), 616-632.

30. Mittas, N., & Angelis, L., (2008b). Partial regression error characteristic curves for the comparison of software cost prediction models. In: Proceedings of the Artificial Intelligence Techniques in Software Engineering (AISEW'08), July, pp. 6–10
31. Mittas, N., & Angelis, L. (2008c, September). Comparing software cost prediction models by a visualization tool. In Software Engineering and Advanced Applications, 2008. SEAA'08. 34th Euromicro Conference (pp. 433-440). IEEE.
32. Mittas, N., & Angelis, L. (2009). Bootstrap Confidence Intervals for Regression Error Characteristic Curves Evaluating the Prediction Error of Software Cost Estimation Models. In AIAI Workshops (pp. 221-230).
33. Mittas, N., & Angelis, L. (2010). Visual comparison of software cost estimation models by regression error characteristic analysis. *Journal of Systems and Software*, *83*(4), 621-637.
34. Mittas, N., & Angelis, L. (2012). A permutation test based on regression error characteristic curves for software cost estimation models. *Empirical Software Engineering*, *17*(1-2), 34-61.
35. Mittas, N., & Angelis, L. (2013, September). Overestimation and Underestimation of Software Cost Models: Evaluation by Visualization. In*Software Engineering and Advanced Applications (SEAA), 2013 39th EUROMICRO Conference on* (pp. 317-324). IEEE.
36. Mittas, N., & Angelis, L. (2016, September). Managing the Uncertainty of Bias-Variance Tradeoff in Software Predictive Analytics. In *Software Engineering and Advanced Applications, 2016. SEAA'16. 42th Euromicro Conference* (to be announced). IEEE.
37. Mittas, N., Athanasiades, M., & Angelis, L. (2008). Improving analogy-based software cost estimation by a resampling method. Information and Software Technology, 50(3), 221-230.
38. Mittas, N., Mamalikidis, I., & Angelis, L. (2015). A framework for comparing multiple cost estimation methods using an automated visualization toolkit. Information and Software Technology, 57, 310-328.
39. Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia.
40. Myrtveit, I., & Stensrud, E. (2012). Validity and reliability of evaluation procedures in comparative studies of effort prediction models. Empirical software engineering, 17(1-2), 23-33.
41. Myrtveit, I., Stensrud, E., & Shepperd, M. (2005). Reliability and validity in comparative studies of software prediction models. Software Engineering, IEEE Transactions on, 31(5), 380-391.
42. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

43. Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., & Cheung, D. (2009, December). Naive bayes classification of uncertain data. In 2009 Ninth IEEE International Conference on Data Mining (pp. 944-949). IEEE.
44. Shepperd, M., & Kadoda, G. (2001). Comparing software prediction techniques using simulation. Software Engineering, IEEE Transactions on,27(11), 1014-1022.
45. Shepperd, M., & MacDonell, S. (2012). Evaluating prediction systems in software project estimation. Information and Software Technology, 54(8), 820-827.
46. Shepperd, M., & Schofield, C. (1997). Estimating software project effort using analogies. Software Engineering, IEEE Transactions on, 23(11), 736-743.
47. Shirabad, J. S., & Menzies, T. J. (2005). The PROMISE repository of software engineering databases. School of Information Technology and Engineering, University of Ottawa, Canada, 24.