Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Coordenação de Pós-Graduação em Ciência da Computação

Similarity-Based Test Suite Reduction in the Context of Model-Based Testing

Ana Emília Victor Barbosa Coutinho

Thesis submitted to Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I as partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Area: Computer Science

Field of Research: Software Engineering

Patrícia Duarte de Lima Machado

(Supervisor)

Emanuela Gadelha Cartaxo

(Co-Supervisor)

Campina Grande, Paraíba, Brazil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

C871s Coutinho, Ana Emília Victor Barbosa

Similarity-based test suite reduction in the context model-based testing / Ana Emília Victor Barbosa Coutinho. – Campina Grande, 2015.

175f: il.color.

Tese (Doutorado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2015.

"Orientação: Prof.ª PhD. Patrícia Duarte de Lima Machado, Prof.ª Dr.ª Emanuela Gadelha Cartaxo".

Referências.

1. Teste de Software. 2. Redução de Suítes de Teste. 3. Funções de Similaridade. 4. Engenharia de Software Empírica. I. Machado, Patrícia Duarte de Lima. II. Cartaxo, Emanuela Gadelha. III. Título.

CDU – 004.052.42(043)

"SIMILARITY-BASED TEST SUITE REDUCTION IN THE CONTEXT OF MODEL-BASED TESTING"

ANA EMILIA VICTOR BARBOSA COUTINHO

TESE APROVADA EM 20/03/2015

Patricia DL Machado

PATRICIA DUARTE DE LIMA MACHADO, Ph.D, UFCG
Orientador(a)

EMANUELA GADELHA CARTAXO, D.Sc, UFCG
Orientador(a)

ANAMARIA MARTINS MOREIRA, PhD, UFRJ
Examinador(a)

JULIANO MANABU IYODA, Ph.D., UFPE
Examinador(a)

JORGE CESAR ABRANTES DE FIGUEIREDO, D.Sc, UFCG
Examinador(a)

WILKERSON DE LUCENA ANDRADE, D.Sc, UFCG
Examinador(a)

CAMPINA GRANDE - PB

Universidade Federal
de Campina Grande

Declaro, para os devidos fins, que participei por videoconferência da apresentação da defesa da Tese de Doutorado de **Ana Emília Victor Barbosa Coutinho**, intitulada: "SIMILARITY-BASED TEST SUITE REDUCTION IN THE CONTEXT OF MODEL-BASED TESTING", em 20 de Março de 2015 e considero o trabalho aprovado.

Emanuela Gadelha Cartaxo

Emanuela Gadelha Cartaxo (UFCG)

Universidade Federal
de Campina Grande

Declaro, para os devidos fins, que participei por videoconferência da apresentação da defesa da Tese de Doutorado de **Ana Emília Victor Barbosa Coutinho**, intitulada: "SIMILARITY-BASED TEST SUITE REDUCTION IN THE CONTEXT OF MODEL-BASED TESTING", em 20 de Março de 2015 e considero o trabalho aprovado.

Juliano Manabu Iyoda (UFPE)

Resumo

Teste de software é uma importante e cara atividade do processo de desenvolvimento de software para avaliar a qualidade do produto. A fim de reduzir os custos na geração de casos de teste, abordagens de Teste Baseado em Modelos (*Model-Based Testing - MBT*) têm sido propostas. MBT fornece o benefício da geração automática de casos de teste a partir de modelos abstratos que captam, por exemplo, os requisitos do software. Apesar da automação ser fundamental na prática de MBT, a geração de casos de teste para aplicações industriais muitas vezes podem produzir grandes suítes de teste, tornando-as não rentáveis. Além disso, uma suíte de teste pode ter, eventualmente, vários casos de teste redundantes (em relação a um conjunto de requisitos de teste). A fim de lidar com este problema, diversos estudos têm sido desenvolvidos visando reduzir os custos relacionados ao tamanho de uma suíte de teste gerada automaticamente. Redução de suítes de teste (também conhecida como *minimização de suítes de teste*) tem como objetivo produzir um subconjunto representativo a partir de uma suíte de teste completa que satisfaça o mesmo conjunto de requisitos de teste. A maioria das estratégias de redução propostas na literatura são baseadas em heurística para maximizar a cobertura local e taxa de redução, no entanto, a capacidade de detecção de faltas é baixa. Além disso, poucas estratégias de redução consideram o grau de similaridade entre os casos de teste.

Neste sentido, o principal objetivo desta Tese é melhorar o processo de redução de suítes de teste, propondo uma estratégia baseada em similaridade e no uso de múltiplos critérios no contexto da MBT visando maximizar a cobertura de faltas. A ideia é identificar o grau de similaridade entre os casos de teste e manter na suíte de teste os mais diferentes casos de teste que juntos possam atender um conjunto de requisitos de teste, e ao mesmo tempo manter uma certa redundância na suíte de teste reduzida com a aplicabilidade dos vários critérios. Primeiro, investigamos a eficácia das funções de distância em nossa estratégia de redução de suítes de teste baseada em similaridade no contexto de MBT. Os resultados mostram que as funções de distância tem um comportamento semelhante em relação a redução do tamanho da suíte de teste. No entanto, como as suítes reduzidas são diferentes dependendo da função de distância aplicada, a escolha pode afetar significativamente a cobertura de faltas e a estabilidade. Depois, comparamos a nossa estratégia de redução com outras quatro heurísticas de redução de suítes de teste bem conhecidas usando simples ou múltiplos critérios de cobertura baseados em transição. Os resultados mostram que a escolha dos critérios de cobertura podem afetar significativamente o tamanho da suíte reduzida, a cobertura de faltas, e o espalhamento. Além disso, nossa estratégia apresentou resultados promissores em relação cobertura de faltas e de dispersão com o uso de *bi-critérios*.

Abstract

Software testing is an important and expensive activity of the software development process to evaluate product quality. In order to reduce the cost of test case generation, Model-Based Testing (MBT) approaches have been proposed. It provides the benefit of automatic test case generation from abstract models that capture, for instance, software requirements. Despite the fact that automation is critical to the practice of MBT, test case generation for industrial size applications can often produce large test suites that may not be cost-effective. Also, a test suite can have possibly several redundant test cases (in relation to one set of test requirements). In order to handle this problem, different studies have been developed aimed at reducing the costs related to the size of an automatically generated test suite. Test suite reduction (also known as *test suite minimization*) aims to produce a representative subset of the complete test suite that satisfies the same set of test requirements as the complete one. Most reduction strategies proposed in literature are based on heuristics to maximize local coverage and reduction rate, however the capability of fault detection is low. Furthermore, few reduction strategies consider the similarity degree among test cases.

In this sense, the main objective in this thesis is to improve the process of test suite reduction by proposing a strategy based on similarity and multi-criteria in the context of MBT aiming to maximize the fault coverage. The idea is to identify the degree of similarity among the test cases and keep in the suite the most different ones that together can meet a set of test requirements, and at the same time maintaining some redundancy in the reduced suite with the applicability of the multiple criteria. First, we investigate the effectiveness of distance functions for our similarity-based test suite reduction strategy in the context of MBT. Results show that the distance functions have similar behavior regarding suite size reduction. However, as reduced suites are different depending on the distance function applied, the choice can significantly affect fault coverage and stability. Afterwards, we compare our reduction strategy with other four well-known test suite reduction heuristics by using single or multiple coverage criteria for transition-based coverage criteria. Results show that choice of the coverage criteria can significantly affect suite size reduction, fault coverage, and scattering. Furthermore, our strategy showed promising results regarding fault coverage and scattering with *bi-criteria*.

“The greatest enemy of knowledge is not ignorance,
it is the illusion of knowledge.”

Stephen Hawking

Acknowledgment

First, I would like to thank God for the gift of life and for blessings all through my life that allowed me to get here.

I thank my parents, Benedito e Solange, for unconditional love, attention, affection, dedication, understanding, motivation and teachings always given to me. Thank you for all your effort extended to my personal and professional training!

To my beloved husband, Brauner, my companion and friend, who has constantly been by my side on this long journey, stimulating me to continue and never stop believing in myself. Thank you for your dedication, support, trust, and for showing me the meaning of love every day!

To my dear sister Ana Esther and brother Gustavo, friends and motivators, who always believed in my work. Thank you for your love and trust!

To my nephews, Leo and Brunninho, and nieces, Duda and Gabi, I thank you for your tenderness and for moments of happiness, making my life lighter and more joyous. Thank you for your affection!

I also thank my brothers-in-law and sisters-in-law for your incentive and support. Thank you for your friendship!

To my parents-in-law, Antonio e Maria do Carmo, who have been witnesses of my walk and have always cheered me on, I thank you for your affection, and all you support.

In short, I thank all my family, grandparents, uncles, aunts and cousins, who has always been a source happy and restful moments.

I thank my dear advisors, Patrícia and Emanuela, for the teachings given, experience shared, and direct responsibility in the construction of this Thesis. Thank you for the trust and friendship that you have always dedicated to me!

I would thank my dear English teacher, Betty, for her teachings and constant presence during this journey. Thank you for your friendship!

I thank my friends, Adriana Torres, Alana, Fofa, Francisco Eduardo, Larissa Ataíde, Paulo Eduardo and Vanessa, by personal support, trust and friendship. Thank you for everything!

To my Software Practices Laboratory (SPLab) colleagues, I thank you for your friendship and support during the entire Doctorate, especially Lilian, Marilene and Paloma. To my research and class colleagues, Adriana Carla, Alan, Catharine, Everton, Fabrício, João Felipe, Katyusco, Matheus and Taciano, thank you for moments of very important relaxation and conversation. To Francisco Neto, I thank you for your friendship and technical support since the beginning of this journey.

I thank the members of the examining board, professors Anamaria Moreira, Jorge Figueiredo, Juliano Iyoda and Wilkerson Andrade, for your precious contributions that contributed to the final result of this work.

To the State University of Paraíba (UEPB), especially Campus VI-Monteiro, which supplied financial support and allowed my leave from academic activities for the accomplishment of my professional training.

To the Post-Graduate Program in Computer Science of the Federal University at Campina Grande (PPGCC-UFCG), faculty and employees, for their welcoming, logistics and the opportunity for me to develop and grow as a professional and a human being.

I thank the Coordination for the Improvement of Higher Education Personnel (CAPES) and to the National Institute of Science and Technology for Software Engineering (INES) for financial support.

To all, despite not mentioned, who directly or indirectly have contributed to the accomplishment of this work.

Contents

1 Introduction **1**

1.1 Problem and Proposed Solution 3

1.2 Research Questions and Methodology 8

1.3 Contributions 9

1.4 Concluding Remarks 9

2 Background **11**

2.1 Software Testing 11

2.2 Model-Based Testing (MBT) 12

2.2.1 Labelled Transition System (LTS) 12

2.2.2 Annotated Labelled Transition System (ALTS) 14

2.3 Parameterized DFS Algorithm 15

2.4 Transition-Based Coverage Criteria 18

2.5 Test Suite Reduction 20

2.5.1 Heuristics for Test Suite Reduction 22

2.6 Distance Functions 26

2.6.1 Jaccard Index 27

2.6.2 Jaro Distance 27

2.6.3 Jaro-Winkler Distance 28

2.6.4 Levenshtein Distance 29

2.6.5 Sellers Algorithm 30

2.7 Experimental Studies in Software Engineering 31

2.8 Statistical Analysis 34

2.8.1 Descriptive Statistic 34

2.8.2 Hypothesis Testing 35

2.9 Concluding Remarks 36

3 Similarity-based Test Suite Reduction 37

3.1 The Proposed Strategy 37

3.2 Our Similarity Function 40

3.3 Example 42

3.4 Concluding Remarks 47

4 Investigating Distance Functions for Similarity-based Test Suite Reduction Strategy 49

4.1 Motivation 49

4.2 Experimental Studies 51

4.2.1 Experiment Planning 51

4.2.2 Analysis and Interpretation 60

4.3 Case Study 68

4.4 Concluding Remarks 72

5 Evaluation of the Similarity-based Test Suite Reduction Strategy 74

5.1 Experiment Definition 75

5.1.1 Definition 75

5.1.2 Planning 75

5.1.3 Operation 83

5.1.4 Threats to Validity 84

5.2 Experiment Analysis 84

5.2.1 Study Question 1 (SQ1) 85

5.2.2 Study Question 2 (SQ2) 89

5.2.3 Study Question 3 (SQ3) 92

5.2.4 Study Question 4 (SQ4) 93

5.3 Scattering 94

5.4 Concluding Remarks 96

6 Review on Test Suite Reduction 98

6.1 Heuristics and Clusters . . . 98

6.1.1 Comparative Studies . . . 99

6.1.2 Using Multiple Testing Criteria . . . 100

6.2 Specification-based Reduction . . . 102

6.3 Concluding Remarks . . . 103

7 Concluding Remarks 104

7.1 Conclusions . . . 104

7.2 Future Works . . . 106

A Results of Statistical Tests for the Evaluation of the Similary-based Test Suite Reduction Strategy 118

A.1 Configuration . . . 119

A.2 Normality test . . . 122

A.3 Kruskal-Wallis test . . . 124

A.3.1 Study Question 1 . . . 124

A.3.2 Study Question 2 . . . 125

A.3.3 Study Question 3 . . . 126

A.3.4 Study Question 4 . . . 126

A.4 Boxplot . . . 127

A.4.1 Study Question 1 . . . 127

A.4.2 Study Question 2 . . . 129

A.4.3 Study Question 3 . . . 131

A.4.4 Study Question 4 . . . 133

A.5 Mann-Whitney test and $\hat{A}_{12}$ effect size measurement . . . 135

A.5.1 Study Question 1 . . . 135

A.5.2 Study Question 2 . . . 139

A.5.3 Study Question 3 . . . 145

A.5.4 Study Question 4 . . . 147

A.6 The minimum, maximum, median and average . . . 149

A.7 Scattering (SSR_FC) . . . 155

A.7.1 Normality test 155
A.7.2 Kruskal-Wallis test 156
A.7.3 Boxplots 157
A.7.4 Mann-Whitney test and $\hat{A}_{12}$ effect size measurement 161
A.7.5 The minimum, maximum, median and average 168
A.7.6 Ordering of effectiveness 174

List of Symbols

ALTS - *Annotated Labelled Transition System*

CB - *Collector Biometrics*

DFS - *Depth-First Search*

EFSM - *Extended Finite State Machine*

FC - *Fault Coverage*

GE - *Greedy Essential*

GRE - *Greedy – 1-to-1 – Redundancy Essential*

HGS - *Harrold, Gupta, and Soffa*

IEEE - *Institute of Electrical and Electronics Engineers*

ILP - *Integer Linear Programing*

LTS - *Labelled Transition System*

LTS-BT - *Labelled Transitions System - Based Testing* MBT - *Model-Based Testing*

MC/DC - *Modified Condition/Decision Coverage*

PDFSam - *PDF Split and Merge*

PWIR - *Pairwise Interaction of Test Requeriments* RTS - *Reduction with tie-breacking*

SSR - *Suite Size Reduction*

SUT - *System Under Test*

TaRGeT - *Test and Requirements Generation Tool*

UML - *Unified Modeling Language*

UMLAUT - *Unified Modelling Language All pUrposes Transformer*

List of Figures

1.1 An example of an ALTS specification 4

2.1 Activities and artifacts of an MBT 13
2.2 An example of an ALTS specification 14
2.3 Tree obtained from a traditional DFS algorithm 16
2.4 Subtrees obtained from the ALTS of Figure 2.2 (b) 17
2.5 Tree generated from the ALTS of Figure 2.2 (b) with one expansion 17
2.6 The hierarchy of transition-based criteria 20
2.7 Examples of normal Q-Q plot and boxplot 35

4.1 Schema of the experimental study for each input specification 57
4.2 Boxplots for SSR and FC considering PDFSam configuration 61
4.3 Number of subsets of test cases and faults for the PDFSam configuration . . 63
4.4 Boxplots for SSR and FC considering TaRGeT configuration 64
4.5 Number of subsets of test cases and faults for the Target configuration . . . 66
4.6 Boxplots for SSR and FC considering the general average for CB$_{v_1}$ and CB$_{v_2}$ 69

5.1 Generation process of the synthetic specifications 81
5.2 Scheme to generate faults for each synthetic specification input 81
5.3 Overview of the experiment for each input specification 83
5.4 Boxplots considering SSR metric for SQ1 86
5.5 Boxplots considering FC metric for SQ1 87
5.6 Boxplots considering SSR metric for SQ2 90
5.7 Boxplots considering FC metric for SQ2 91

A.1 Boxplots considering SSR metric for SQ1 127

A.2 Boxplots considering FC metric for SQ1 128
A.3 Boxplots considering SSR metric for SQ2 129
A.4 Boxplots considering FC metric for SQ2 130
A.5 Boxplots considering SSR metric for SQ3 131
A.6 Boxplots considering FC metric for SQ3 132
A.7 Boxplots considering SSR metric for SQ4 133
A.8 Boxplots considering FC metric for SQ4 134
A.9 Boxplots considering SSR_FC metric for SQ1 157
A.10 Boxplots considering SSR_FC metric for SQ2 158
A.11 Boxplots considering SSR_FC metric for SQ3 159
A.12 Boxplots considering SSR_FC metric for SQ4 160

List of Tables

1.1 Test cases obtained with one expansion 4
1.2 Fault detection capability (%) 5
1.3 Scattering (%) 6
1.4 Frequency of detection of each fault (%) 7
2.1 Test suite obtained from the DFS algorithm in the tree 18
2.2 Satisfiability relations 21
2.3 Cardinality 25
2.4 Overview of statistical tests 36
3.1 Identical transition pairs 42
3.2 Frequency of detection of each fault for Sim (%) 48
4.1 Basic configuration of the two real-world specifications 55
4.2 Comparing test case and fault metrics of the synthetic LTS specifications to the corresponding real specification ones 56
4.3 Mean, standard deviation and the highest number of necessary replications for each metric and each application 58
4.4 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam configuration 62
4.5 Ordering of effectiveness for SSR and FC in PDFSam configuration . . . 63
4.6 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for general average in TaRGeT configuration 65
4.7 Ordering of effectiveness for SSR and FC in TaRGeT configuration 66
4.8 The configurations of the real-world specifications 68

4.9 Mann-Whitney and $\hat{A}_{12}$ effect size measurements when SSR and FC across the distance functions for CB_{v_1} . . . 70
4.10 Mann-Whitney and $\hat{A}_{12}$ effect size measurements when SSR and FC across the distance functions for CB_{v_2} . . . 71
4.11 Ordering of effectiveness for SSR and FC in CB_{v_1} and CB_{v_2} . . . 71
4.12 Number of different sets of test cases selected, number of distinct test cases, average frequency of inclusion of a test case in the reduced suite, number of different sets of faults detected, number of distinct faults and average frequency of inclusion of a fault detected by a reduced suite . . . 72

5.1 Null and alternative hypotheses considering SQ1 . . . 78
5.2 Null and alternative hypotheses considering SQ2 . . . 79
5.3 Basic configuration of the three real-world specifications . . . 81
5.4 Number of failures and faults of the three real-world specifications . . . 82
5.5 Ordering of effectiveness for each reduction strategy associated with all coverage criteria in terms of SSR and FC . . . 88
5.6 Ordering of effectiveness among reduction strategies for each coverege criteria in terms of SSR and FC . . . 89
5.7 Null and alternative hypotheses for SSR considering SQ3 . . . 92
5.8 Null and alternative hypotheses for FC considering SQ3 . . . 92
5.9 Ordering of effectiveness reduction strategies in combination with their best coverage criterion regarding the SSR and FC metrics . . . 93
5.10 Null and alternative hypotheses for SSR considering SQ4 . . . 94
5.11 Null and alternative hypotheses for FC considering SQ4 . . . 94
5.12 Ordering of effectiveness coverage criteria in combination with their best reduction strategy regarding the SSR and FC metrics . . . 95
5.13 Ordering of effectiveness for SQ3 and SQ4 regarding the SSR_FC . . . 96

A.1 Basic configuration for CB . . . 119
A.2 Basic configuration for PDFSam . . . 120
A.3 Basic configuration for TaRGeT . . . 121
A.4 Anderson-Darling normality test for CB real . . . 122

A.5 Anderson-Darling normality test for CB synthetics 122
A.6 Anderson-Darling normality test for PDFSam real 122
A.7 Anderson-Darling normality test for PDFSam synthetics 123
A.8 Anderson-Darling normality test for TaRGeT real 123
A.9 Anderson-Darling normality test for TaRGeT synthetics 123
A.10 Krukal-Wallis test for SQ1 124
A.11 Krukal-Wallis test for SQ2 125
A.12 Krukal-Wallis test for SQ3 126
A.13 Krukal-Wallis test for SQ4 126
A.14 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB real 135
A.15 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB synthetics 136
A.16 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam real 136
A.17 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam synthetics . 137
A.18 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT real 137
A.19 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT synthetics . . 138
A.20 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB real 139
A.21 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for general average in CB synthetics 140
A.22 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam real 141
A.23 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for general average in PDFSam synthetics 142
A.24 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT real 143
A.25 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for general average in TaRGeT synthetics 144
A.26 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB real 145
A.27 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB synthetics 145
A.28 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam real 145
A.29 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam synthetics . 146
A.30 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT real 146
A.31 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT synthetics . . 146
A.32 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB real 147

A.33 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB synthetics 147
A.34 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam real 147
A.35 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam synthetics . 147
A.36 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT real 148
A.37 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT synthetics . . 148
A.38 The minimum, maximum, median and average for CB real 149
A.39 The minimum, maximum, median and average for CB synthetics 150
A.40 The minimum, maximum, median and average for PDFSam real 151
A.41 The minimum, maximum, median and average for PDFSam synthetics . . . 152
A.42 The minimum, maximum, median and average for TaRGeT real 153
A.43 The minimum, maximum, median and average for TaRGeT synthetics . . . 154
A.44 Anderson-Darling normality test (ρ-value) for CB configuration 155
A.45 Anderson-Darling normality test (ρ-value) for PDFSam configuration . . . 155
A.46 Anderson-Darling normality test (ρ-value) for TaRGeT configuration 155
A.47 Krukal-Wallis test for SQ1 156
A.48 Krukal-Wallis test for SQ2 156
A.49 Krukal-Wallis test for SQ3 156
A.50 Krukal-Wallis test for SQ4 156
A.51 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB configuration . . . 161
A.52 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam configuration 162
A.53 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT configuration 162
A.54 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB configuration . . . 163
A.55 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam configuration 164
A.56 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT configuration 165
A.57 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB configuration . . . 166
A.58 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam configuration 166
A.59 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT configuration 166
A.60 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB configuration . . . 167
A.61 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam configuration 167
A.62 Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT configuration 167
A.63 The minimum, maximum, median and average for CB real 168

A.64 The minimum, maximum, median and average for CB synthetics 169
A.65 The minimum, maximum, median and average for PDFSam real 170
A.66 The minimum, maximum, median and average for PDFSam synthetics . . . 171
A.67 The minimum, maximum, median and average for TaRGeT real 172
A.68 The minimum, maximum, median and average for TaRGeT synthetics . . . 173
A.69 Ordering of effectiveness for each reduction strategy associated with all coverage criteria 174
A.70 Ordering of effectiveness among reduction strategies for each coverege criterion 175
A.71 Ordering of effectiveness reduction strategies in combination with their best coverage criterion 175
A.72 Ordering of effectiveness coverage criteria in combination with their best reduction strategy 175

Chapter 1

Introduction

Software Testing is an important activity along the software development cycle with different goals. Testing is a common validation approach in industry, and often used to evaluate the quality and reveal faults of applications [Binder 2000]. However, this activity is expensive, still largely *ad hoc*, unpredictably effective, and it generally consumes much of the overall development effort [Bertolino 2007; Utting and Legeard 2007]. In this sense, researchers have proposed several approaches aiming to decrease efforts in the activity of software testing.

In the last years, many approaches have been proposed in the context of Model-based Testing (MBT) to control software quality and to reduce costs [Pretschner 2005]. MBT is a *black box* approach that has raised interest from both academy and industry in the last years. It provides the benefit of automatic test case generation from a specification model of the system behavior, for instance, software requirements. In general, behavioral specifications of the System Under Test (SUT) can be constructed early in the development cycle. Thus, the test cases can be obtained before or during the development process. Furthermore, the development of a variety of strategies based on MBT has demonstrated its feasibility in the software process [Utting and Legeard 2007]. Despite the fact that automation is critical to the practice of MBT, test case generation for industrial size applications can often produce large test suites that may not be cost-effective, particularly for manual testing [Bertolino 2007]. The reason is that most of the times, automatic generation algorithms are based on a structural and systematic search for test cases constrained by test criteria. With the goal of improving the effectiveness of the suite by achieving coverage, algorithms may generate several similar test cases, depending on the model structure. In order to handle this problem, the testing team can perform additional test selection before test execution. However, test selection may profoundly impact on the success of the testing process as whole: important test cases such as the ones that uncover faults may not be selected [Pezzè and Young 2007]. Therefore, it is necessary to investigate approaches to deal with the costs related to the size of an automatically generated test suite in MBT.

Toward this purpose, researchers have investigated different approaches. Among the approaches proposed, we highlight *test suite reduction* (also known as *test suite minimization*) [Harrold et al. 1993; Chen and Lau 1998a]. Its goal is to produce a representative subset from the complete test suite that satisfies a set of test requirements with the same coverage as the complete test suite. The idea is to have in the subset the most representative test cases covering all set of test requirements faster. Generally, the automatically generated test suites may contain a considerable degree of redundancy among test cases [Cartaxo 2011]. Thus, the reduced test suite is formed by adding, one by one, the test cases that are not redundant with respect to the set of test requirements when compared to the ones already chosen. Another common approach in literature that can also be useful for addressing the test suite size problem is *test case selection*. Its goal is to select a subset of the complete test suite according to a specific objective. However, the test cases selected may or may not provide the same coverage of the set of test requirements as the complete test suite. In turn, *test suite reduction* is a test case selection that satisfies all test requirements of the complete test suite.

A number of test suite reduction strategies to be applied at code level have already been extensively investigated and experimented in literature [Rothermel et al. 2002]. These strategies are usually based on heuristics to maximize coverage, and test requirements are defined as a coverage criterion, such as statement, decision and so on. For instance, four well-known heuristics for code-based test suite reduction follow these ideas: Greedy [Chvätal 1979; Cormen et al. 2001], GE [Chen and Lau 1998b], GRE [Chen and Lau 1998a] and HGS [Harrold et al. 1993]. Empirical studies have shown that requirements-based reduction may be effective to reduce the size of the suite, however they may also reduce the capability of fault detection [Fraser and Wotawa 2007; Yoo and Harman 2012].

To address the test suite size problem, there are other approaches, present in the literature, based on test case classification according to a degree of similarity measured by a distance function [da Silva Simao et al. 2006; Kovács et al. 2009; Bertolino et al. 2010; Coutinho et al. 2013]. Empirical studies on test case selection based on similarity have shown that test case diversity may improve the rate of fault detection, and the choice of a distance function may directly influence on the fault detection ability of the test case selection strategies [Chen et al. 2010; Hemmati et al. 2013; Cartaxo et al. 2011].

On the other hand, in order to improve the fault coverage of the reduced test suite for code level, several researchers have investigated the combination of multiple coverage criteria to reduce test suites [Black et al. 2004; Jeffrey and Gupta 2007; Lin and Huang 2009; Selvakumar et al. 2010b; Khalilian and Parsa 2012]. However, investigation on reduction for specification-based test cases is recent, specially in the MBT field, with few strategies and experimental results. Furthermore, results are not conclusive and are divergent in comparison to the *white box* context.

1.1 Problem and Proposed Solution

As said before, different test suite reduction heuristics have been proposed to find a subset of test cases which satisfies the same set of test requirements as the complete test suite. These heuristics aim to maximize coverage with the elimination of redundant test cases from a test suite. For this, the heuristics make the best local choice at each step with the goal of finding the best global. According to Harrold et al. [Harrold et al. 1993], a test case is redundant if other test cases in the test suite provide the same coverage for a given coverage criterion. However, empirical studies have shown that a potential drawback of the test suite reduction is to decrease its fault detection capability [Rothermel et al. 2002].

To illustrate this drawback, let us consider the toy example of an *Annotated Labelled Transition System* (ALTS) specification presented in Figure 1.1 (a) that combines basic, alternate, and exception flows of a use case presented by Coutinho et al. [Coutinho et al. 2013]. The use case defines the behavior of a user account editing operation where: *i*) we can change user name and password and *ii*) we can delete a user account. As usual convention, labels beginning with "?" denote actor input actions, whereas labels beginning with "!" denote system output actions. However, for the sake of simplicity, we replace transition labels by letters as shown in Figure 1.1(b).

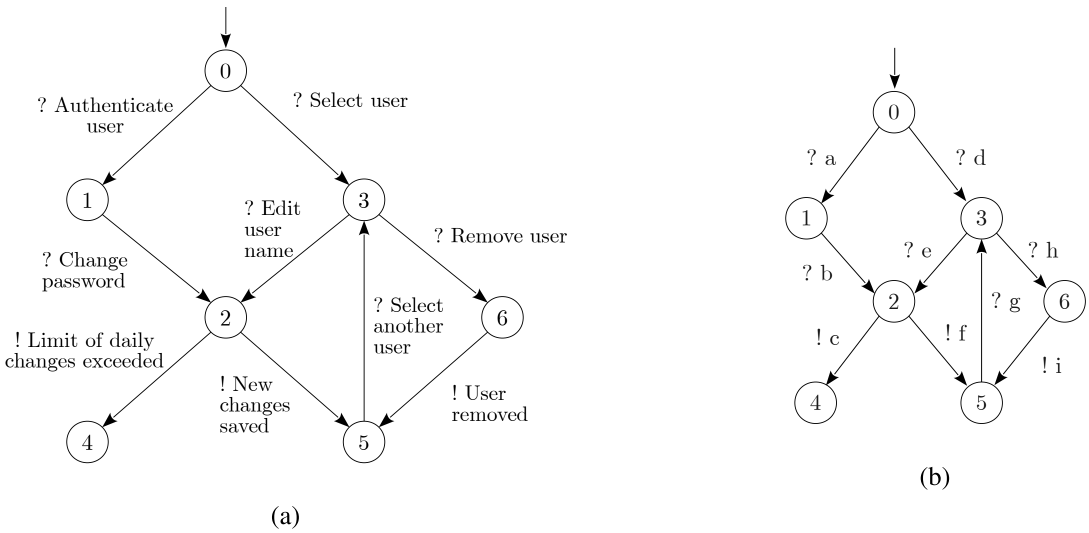


Figure 1.1: *An example of an ALTS specification*

In order to parameterize the number of times that paths with loop should be traversed, maximizing the exploration of different sequences, we opted by the generation algorithm proposed by Araújo et al. [Araújo et al. 2012]. Thus, we obtained thirteen test cases with one expansion, i.e., the paths with loop can be executed only one time, as shown in Table 1.1.

Table 1.1: *Test cases obtained with one expansion*

i	t_i	Test case
1	t_1	$\langle a, b, c \rangle$
2	t_2	$\langle a, b, f, g, h, i, g, h, i \rangle$
3	t_3	$\langle a, b, f, g, h, i \rangle$
4	t_4	$\langle a, b, f, g, e, c \rangle$
5	t_5	$\langle a, b, f, g, e, f, g, h, i \rangle$
6	t_6	$\langle a, b, f, g, e, f \rangle$
7	t_7	$\langle d, h, i, g, h, i \rangle$
8	t_8	$\langle d, h, i, g, e, c \rangle$
9	t_9	$\langle d, h, i, g, e, f \rangle$
10	t_{10}	$\langle d, e, c \rangle$
11	t_{11}	$\langle d, e, f, g, h, i \rangle$
12	t_{12}	$\langle d, e, f, g, e, c \rangle$
13	t_{13}	$\langle d, e, f, g, e, f \rangle$

And, we suppose three faults caused by three different test cases fail:

- **Fault 01:** t_5;
- **Fault 02:** t_7;
- **Fault 03:** t_{10}.

These test cases that failed (failures) have different characteristics, such as: the largest, middle and smallest test cases from the test suite. Then, we applied each heuristic 1,000 times in this scenario.

Considering the execution of the following test suite reduction heuristics: G, GE, GRE and HGS, we performed a case study to investigate the fault detection capability of the reduced test suite. From the results obtained, we observed some limitations of current heuristics for test suite reduction, particularly in the context of MBT, such as:

- **Fault missing:** Reduced test suite may lack test cases that fail because the heuristic chooses alternative test cases that cover the same requirements. This is often the case when the suite has a high degree of redundancy and, allowing reduction to be severe. Consider our running example, when we use *all-transitions* as coverage criterion we observe that all the heuristics present a high rate of reduction of 84.62% for all executions and the averages of fault coverage are low, range from 8.46% (HGS) to 38.69% (GE), as observed in Table 1.2. Now, using a stronger coverage criterion (*all-transition-pairs*), the rate of reduction is 69.23% for all heuristics in all executions. However, we observed that the average rate of fault coverage is greater than for *all-transitions*, but it's still low, range from 13.99% (HGS) to 41.93% (G), as presented in Table 1.2.

Table 1.2: *Fault detection capability (%)*

Heuristic	*All-transitions*					*All-transition-pairs*				
	Min.	Med.	Max.	Avg.	SD.	Min.	Med.	Max.	Avg.	SD.
G	33.33	33.33	66.67	37.26	14.40	33.33	33.33	66.67	41.93	14.59
GE	33.33	33.33	66.67	38.69	12.25	33.33	33.33	66.67	41.69	14.46
GRE	33.33	33.33	33.33	33.33	7.32E-013	33.33	33.33	33.33	33.33	7.32E-013
HGS	0.00	0.00	33.33	8.46	14.51	0.00	0.00	33.33	13.99	16.46

- **Fault scattering:** The heuristic applied does not directly favor the choice of the test cases that fail. For instance, a heuristic may select the biggest test cases and only the smallest ones actually fail. Even if we could choose a few more test cases to increase the size of the reduced suite, the test cases that fail would not be chosen. Consider our running example, in Table 1.3 presents the degree of scattering of the test cases that fail, and consequently the faults, in the selection order of the heuristics. Having a lower rate of scattering means that the strategy is more effective in adding test cases that fail to the reduced suite, particularly, if we decided to add a few more test cases to the reduced suite. Thus, when we use *all-transitions* and *all-transition-pairs* as coverage criteria we observe that G presents the best degree of scattering since the rate of reduction that reaches 100% fault coverage are 25.25% and 20.98%, respectively. However, we observed that the heuristics for both coverage criteria often has high rate of scattering.

Table 1.3: *Scattering (%)*

Heuristic	*All-transitions*					*All-transition-pairs*				
	Min.	Med.	Max.	Avg.	SD.	Min.	Med.	Max.	Avg.	SD.
G	0.00	23.08	76.92	25.25	20.21	0.00	15.38	61.54	20.98	17.09
GE	0.00	7.69	15.38	5.24	4.88	0.00	0.00	7.69	0.93	2.51
GRE	0.00	0.00	7.69	3.81	3.85	0.00	0.00	0.00	0.00	0.00
HGS	0.00	7.69	30.77	10.65	11.16	0.00	7.69	30.77	14.48	12.51

As shown in Table 1.4, t_5 is the longest test case, t_7 is the middle test case, and t_{10} is the smallest test case in the test suite. For both coverage criteria, we have that t_5 is generally chosen since it satisfies the maximum number of test requirements, except HGS. In turn, HGS tends to select t_4 for *all-transitions*, and t_8 and t_{11} for *all-transition-pairs* since these test case occurs most frequently among test requirements with lowest requirement cardinality (essentialness). Considering the use of *all-transition-pairs* as coverage criterion, we have that t_1 is a *essential test case*, and this will always be part of the reduced test suite. However, for the test cases fail t_7 and t_{10}, the heuristics tend to discard it since other test cases satisfy their and other unsatisfied test requirements for both coverage criteria.

Table 1.4: *Frequency of detection of each fault (%)*

Heuristic	*All-transitions*			*All-transition-pairs*		
	Fault 01 t_5	Fault 02 t_7	Fault 03 t_{10}	Fault 01 t_5	Fault 02 t_7	Fault 03 t_{10}
G	100	0.00	11.80	100	0.00	25.80
GE	100	0.00	16.10	100	0.00	25.10
GRE	100	0.00	0.00	100	0.00	0.00
HGS	0.00	25.40	0.00	42.00	0.00	0.00

In summary, the problem presented here is that different heuristics presented in the literature may have different performances when referred to size and fault detection depending on the selection criteria used. Usually, reduction is based on coverage of model elements. Moreover, it is also important to mention that experimental results presented in literature show that current strategies may have a performance comparable to a random choice strategy. Therefore, there is a need for an investigation on the weaknesses of the current strategies that may lead to the proposal of a more successful one.

In this sense, the main goal of this doctorate research is *to improve the process of test suite reduction by proposing a strategy based on similarity which allows the use of single or multiple coverage criteria in the MBT context aiming to maximize the fault coverage of the reduced test suite*. Based on the state-of-the-art limitations of current strategies, the focus of this work consists in developing a parameterized reduction strategy based on the use of a similarity function and multi-criteria that may be able to highlight different patterns of faults, and therefore improving fault detection capability of the reduced test suite. Similarity functions (also known as distance functions) have been largely considered for test selection strategies with promising results presented in the literature [Cartaxo et al. 2011; Fraser and Wotawa 2007]. Therefore, since MBT test suites are usually highly redundant, we believe that by considering similarity, we might be able to achieve a good balance between size and fault detection. Furthermore, we propose the use of multi-criteria with the aim of combining these criteria to find a reduced test suite that decreases the test suite size while increases fault coverage rate. Thus, our idea is to improve the rate of fault coverage while selecting a subset of the most different for each coverage criterion, however, even though extensive redundancy must be avoided, a little redundancy in the reduced suite may improve its chances of uncovering a fault from the use of multiple criteria.

1.2 Research Questions and Methodology

Based on the goal of this doctorate research, our research questions are:

Research Question 1 *In the context of MBT, how to address similarity among test cases to reduce the size of the test suite while simultaneously maintain a reasonable fault coverage?*

Research Question 2 *What influence does the choice of a similarity function have on the size and fault coverage of similarity-based test suite reduction techniques?*

Research Question 3 *What influence does the coverage criteria have on test suite reduction regarding size and fault coverage of a reduced test suite?*

From these research questions, we observe that:

- Very similar test cases may have a similar behavior in relation to fault coverage;
- Test suite reduction strategies may preserve the behavior of the reduced test suite size, and at the same time, may have a different behavior in relation to fault coverage;
- Coverage criteria that maintain some redundancy in the reduced suite may improve its chances of covering a fault.

The focus of this doctorate research is to define a subset of test cases from the complete test suite that satisfies a set of test requirements (coverage criteria) as the complete test suite aiming to maximize the fault coverage. This complete test suite is automatically generated in the context of MBT from a specification model, such as *Labelled Transition System* (LTS) and *Annotated Labelled Transition System* (ALTS). As answers to our research questions, we propose a parameterized reduction strategy based on similarity among test cases in the context of MBT. Furthermore, we extend similarity function proposed by Cartaxo et al. [Cartaxo et al. 2011] that calculates the similarity degree between test cases considering test cases with repeated transitions. To evaluate the effectiveness of the distance functions in our reduction strategy, we performe three empirical studies considering our function and other five well-known distance functions. Afterwards, we refine our reduction strategy allowing the use of single or multiple coverage criteria. Finally, we investigate the influence of the choice of coverage criteria, and the most appropriate coverage criteria applied for test suite reduction regarding size and fault coverage of reduced test suite.

1.3 Contributions

In order to answer these research questions, we propose a new reduction strategy based on similarity in the context of MBT. In this sense, our main contributions are:

- Extension of the similarity function proposed by Cartaxo et al. [Cartaxo et al. 2011] to consider repeated transitions (*path with loop*);
- Results of empirical studies investigating the use of different distance functions for test suite reduction;
- The use of single or multiple coverage criteria for the reduction of test suites. We investigate the use of three coverage criteria by reduction strategies;
- Tool support through the LTS-BT tool [Cartaxo et al. 2008] to execute the proposed strategy presented in this work[1].

1.4 Concluding Remarks

In this chapter, we presented an overview of this doctorate research. Further details are presented in the next chapters according to the following structure:

Chapter 2 In this chapter, we present the theoretical background, including some terms and concepts in order to make this thesis self-contained.

Chapter 3 This chapter presents our similarity function, and also our parameterized strategy for test suite reduction based on similarity.

[1] https://sites.google.com/a/computacao.ufcg.edu.br/lts-bt/

Chapter 4 This chapter presents an investigation on the effectiveness of distance functions for test suite reduction in the context of MBT. Three empirical studies executed to compare six distance functions by considering the size, fault coverage and stability (the number of different sets of faults produced by the selected suites, and the number of different sets of test cases selected) of the reduced test suite.

Chapter 5 This chaper presents six empirical studies in order to investigate the influence of the choice of coverage criteria used by reduction strategy, and also to evaluate and compare our reduction strategy proposed in Chapter 3 and four well-known reduction heuristics in literature.

Chapter 6 This chapter presents a review on test suite reduction related to this thesis;

Chapter 7 This chapter presents the answers to our research questions and future works related to our contributions.

Chapter 2

Background

This chapter presents some basic concepts and terminology that will be required to understand this document, i.e., to make this document self-contained. First, we introduce general concepts and differences between some terms used in software testing. Next, we present model-based testing concepts, and the transition-based notation used in this document. Afterwards, we present the common transition-based coverage criteria. Moreover, the algorithm to generate test suites is presented. In the sequence, we define the test suite reduction problem, and we briefly describe four well-known heuristics for reduction in literature. Furthermore, we present some candidate functions for measuring the similarity degree between two test cases. Finally, we briefly introduce the basic concepts experimentation in software engineering and statistical analysis adopted in our evaluation methodologies.

2.1 Software Testing

Software testing is an important and critical activity in the software development process, essential to evaluate the product quality by identifying problems in application under testing [Binder 2000].

In this work, we use the terms according to standard terminology defined by Institute of Electrical and Electronics Engineers (IEEE), such as: "*error*", "*fault*" and "*failure*". An *error* (also known as *mistake*) occurs because of a human action. This *error* provokes a *fault* (also known as *defect*) in the application that the person is using. In turn, when this *fault* is detected then the application fails, i.e., does not perform the functionality as required, then a *failure* occurs.

Therefore, testing is a set of activities aiming to find failures in a system in order to improve software quality. However, according to Utting and Legeard [Utting and Legeard 2007], software testing is very expensive and it generally consumes much of the overall development effort. In order to guide this activity, testing methods can be applied to identify a set of test cases (named *test suite*). Each test case is composed by a set of elements that describes a system behavior, such as the system's pre-conditions, a set of inputs, a set of expected outputs, among others. Sometimes the testing methods are classified as functional testing approach (also called *black box*) if the test cases rely only on the input/output behavior, or as structural testing approach (*white box*) who considers the implementation of the system to obtain the test cases [Abran et al. 2004].

2.2 Model-Based Testing (MBT)

Model-based Testing (MBT) is a functional testing approach (also called *black box*) based on automatic generation of test cases from behavioral specifications. Thus, test cases are designed from specification, i.e., they do not use information about their internal structure [Utting and Legeard 2007].

Figure 2.1 presents the main activities and artifacts of an MBT approach. Initially, the formal model is built from *software requirements*. From the formal model, the test cases are generated in order to exercise the system. According to Utting and Legeard [Utting and Legeard 2007], a test case is a finite structure composed of *test inputs* and *expected outputs*. Then, once test cases are defined and test infrastructure is built (scripts, adaptor, coverage tools), the *system under test* is executed, generating outputs. Next, these *generated outputs* are compared to the *expected outputs* to define the result of each test case as *pass* or *fail*.

2.2.1 Labelled Transition System (LTS)

The focus of MBT is on the system behavior that is an abstraction described by a specification model. In this work, the specifications used to generate the test cases are *Labelled Transition Systems* (LTSs). LTS is a common formalism considered by both fundamental and practical research on MBT that is also usually adopted as the semantics formalism of

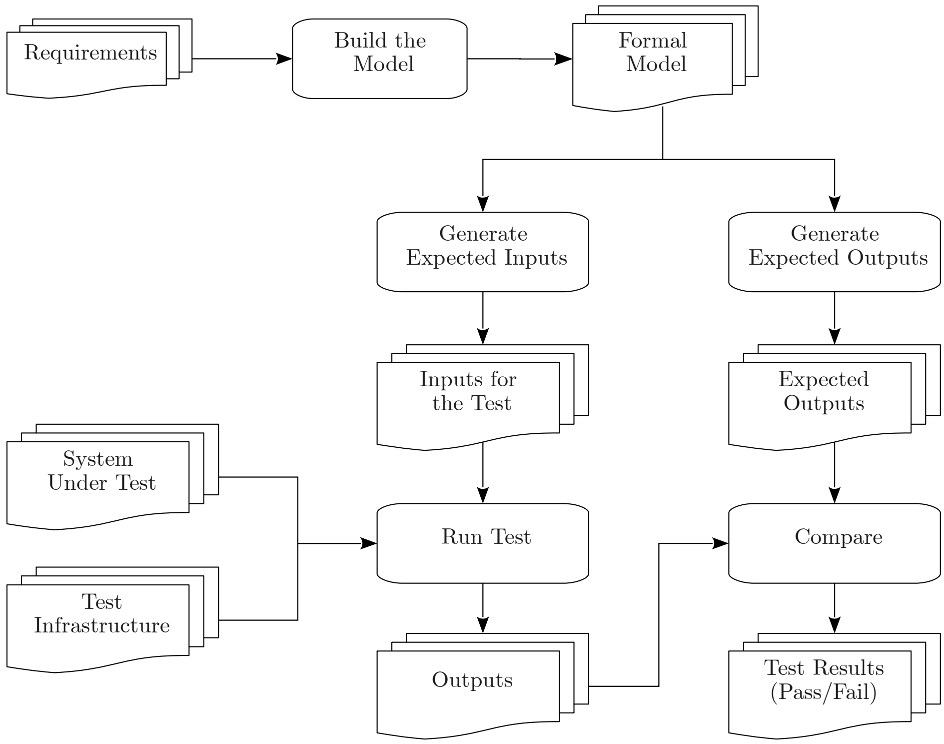


Figure 2.1: *Activities and artifacts of an MBT. Font: adapted from [Cartaxo 2011]*

specification notations [Tretmans 2008; Anand et al. 2013]. Furthermore, there are several tools to support test case generation from LTSs that are derived from abstract specifications, such as UMLAUT [Ho et al. 1999], TGV [Jard and Jéron 2005], TaRGeT [Nogueira et al. 2007], SPACES [Barbosa et al. 2007] and LTS-BT [Cartaxo et al. 2008].

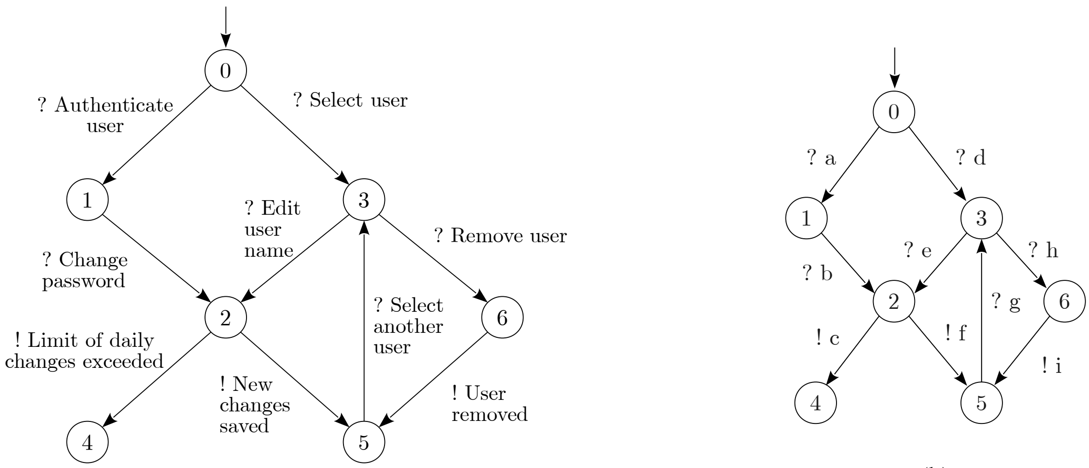
LTS is a directed graph defined in terms of states and labelled transitions between states to describe system behavior. According to Vries and Tretmans [de Vries and Tretmans 2000], an LTS can be formally defined as a 4-tuple $\langle S, L, T, s_0 \rangle$, where:

- S: is a finite, nonempty set of states;
- L: is a finite, nonempty set of labels;
- T: is a subset of $S \times L \times S$ (set of triples), called the transition relation;
- s_0: is the initial state, where $s_0 \in S$.

2.2.2 Annotated Labelled Transition System (ALTS)

Annotated Labelled Transition System (ALTS) is an extension of the LTS containing annotations to indicate special types of interactions [Cartaxo et al. 2007]. Since our focus is on functional testing, these annotations can represent for example the user actions, the system responses, among others types of interactions.

Figure 2.2 showns an example of an ALTS specification (this is the same example presented in Section 1.1).



(a) (b)

Figure 2.2: *An example of an ALTS specification*

For the ALTS specification presented in Figure 2.2 (b), we have:

- $S = \{0, 1, 2, 3, 4, 5, 6\}$;
- $L = \{a, b, c, d, e, f, g, h, i\}$;
- $T = \{(0, a, 1), (1, b, 2), (2, c, 4), (0, d, 3), (3, e, 2), (2, f, 5), (5, g, 3), (3, h, 6), (6, i, 5)\}$;
- $s_0 = \{0\}$.

Observing the example in Figure 2.2 (b), some concepts related to LTS or ALTS specifications can be considered, such as:

- **Path:** a *path* is a finite or infinite sequence of transitions from the initial state. In this work, a *test case* is defined as a *path*. *Paths* can be classified as:

  - *Simple path*: a path without repeated states or transitions, for example, the path d, e, c);
  - *Path with loop*: a path in which one or more states or transitions may be repeated, producing cycles. This ALTS has the following paths with loop, for example:
    * a, b, f, g, h, i;
    * a, b, f, g, e;
    * d, e, f, g;
    * d, h, i, g.

- **Depth:** the depth of the LTS or ALTS. It is calculated by considering the longest path (without repeated transitions - without loops). In this example, the depth of the ALTS specification is six defined by the simple path a, b, f, g, h, i;
- **Join:** is a state with more than one incoming transition (the example contains three joins: states 2, 3 and 5);
- **Transitions of joins:** it is the total number of incoming transitions of the joins, i.e., the total number of incoming transitions of the states with more than one incoming transition. Thereby, the ALTS specification has six transitions of joins (transitions b, d, e, f, g and i);
- **Fork:** is a state with more than one outgoing transition (the example contais three forks: states 0, 2 and 3);
- **Transitions of forks:** it is the total number of outgoing transitions of the forks, i.e., the total number of transitions of the states with more than one outgoing transition. This ALTS has six transitions of forks (transitions a, c, d, e, f and h);

2.3 Parameterized DFS Algorithm

In this work, we consider LTS and ALTS specifications as inputs to MBT approaches. Thus, test cases are obtained from a tree generated from an LTS or ALTS specification using the algorithm proposed by Araújo et al.[Araújo et al. 2012]. This algorithm allows us to parameterize the number of times the loops should be traversed from expansions, maximizing the exploration of different sequences. The idea is to transform an LTS or ALTS specification with loops in a tree, where each path is a test case. Therefore, considering the ALTS specification presented in Figure 2.2 (b), this algorithm is subdivided in three steps.

Step 01. Generating a tree T from an ALTS specification. This T is obtained using a traditional *Depth First Search* (DFS) algorithm from an ALTS specification. The algorithm stops when a vertex in the ALTS specification is visited for the second time, i.e., it uses *all-one-loop-paths* coverage as stop criterion. If the last vertex has already been previously visited in this path, then it is added to a list of marked nodes. Figure 2.3 shows the T obtained from the ALTS example. Note that the nodes 2, 5 and 3 are marked nodes because these are visited for the second time.

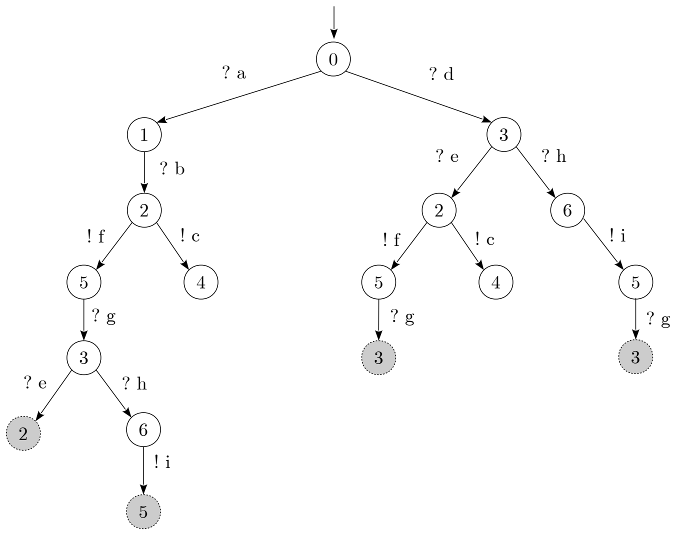


Figure 2.3: *Tree obtained from a traditional DFS algorithm from the ALTS of Figure 2.2 (b)*

Step 02. Generating subtrees of repetition. Each subtree represents a passage in a path with loop. The subtrees of expansions are obtained from the tree T by application of a traditional DFS, with *all-one-loop-paths* as stop criterion. The subtrees also keep a list of marked nodes. Figure 2.4 shows the subtrees A, B and C for the ALTS example. These subtrees are used in the next step.

Step 03. Expansion of the tree. This is a process of *collage* of subtrees of expansions. The subtrees will be placed along the tree obtained in Figure 2.4. If there is any path with loop in the ALTS specification, the number of replications (expansion of the tree) also needs to be informed. This number of expansions define the number

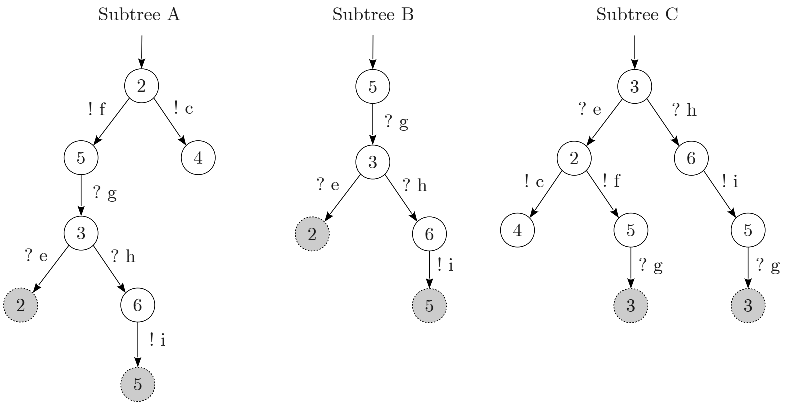


Figure 2.4: *Subtrees obtained from the ALTS of Figure 2.2 (b)*

of times that the path with loop is executed by the generated path, and consequently, the number of times that a subtree should be glued to the tree. Hence, for each node marked, it is added its subtree according to the node source. Figure 2.5 shows a test suite generated from the ALTS example based on the depth search test case generation algorithm, proposed by Araújo et al. [Araújo et al. 2012], with *all-one-loop-paths* as stop criterion.

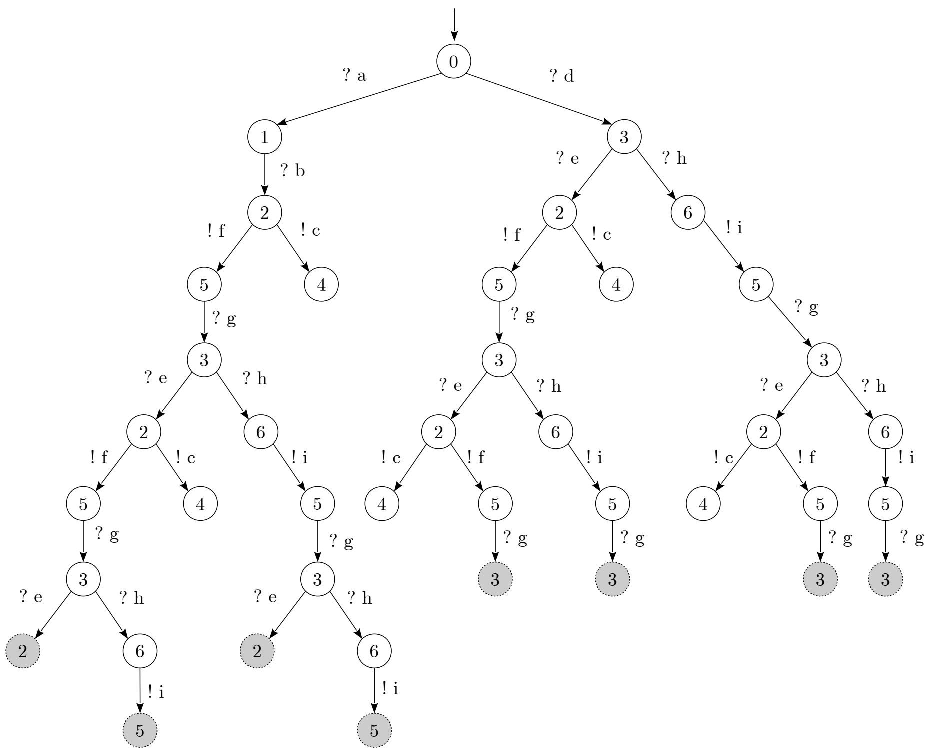


Figure 2.5: *Tree generated from the ALTS of Figure 2.2 (b) with one expansion*

Each path in the tree is a test case, and is associated with a system behavior. For ALTS, a test case is *valid* iff end up with a system output action denoted by labels beginning with "!". Therefore, we consider all transitions for each test case until the last transition beginning with "!". And, all possible paths comprise the test suite TS. Therefore, for the ALTS specification presented in Figure 2.5, we obtained 13 test cases using the DFS algorithm, as shown in Table 2.1.

Table 2.1: *Test suite obtained from the DFS algorithm in the tree presented in Figure 2.5*

i	t_i	Test case
1	t_1	$\langle a, b, c\rangle$
2	t_2	$\langle a, b, f, g, h, i, g, h, i\rangle$
3	t_3	$\langle a, b, f, g, h, i\rangle$
4	t_4	$\langle a, b, f, g, e, c\rangle$
5	t_5	$\langle a, b, f, g, e, f, g, h, i\rangle$
6	t_6	$\langle a, b, f, g, e, f\rangle$
7	t_7	$\langle d, h, i, g, h, i\rangle$
8	t_8	$\langle d, h, i, g, e, c\rangle$
9	t_9	$\langle d, h, i, g, e, f\rangle$
10	t_{10}	$\langle d, e, c\rangle$
11	t_{11}	$\langle d, e, f, g, h, i\rangle$
12	t_{12}	$\langle d, e, f, g, e, c\rangle$
13	t_{13}	$\langle d, e, f, g, e, f\rangle$

2.4 Transition-Based Coverage Criteria

According to Ammann and Offutt [Ammann and Offutt 2008], a coverage criterion is a set of rules that imposes test requirements on a test suite. LTS specifications are transition-based modeling notations, and many structural coverage criteria have been developed. Utting and Legeard [Utting and Legeard 2007] present the most common transition-based coverage criteria used in the context of MBT[1], such as presented below. As running example, we consider the test suite presented in Table 2.1.

- **All-states:** Every state of the specification is visited at least once. Thus, to reach this coverage only two test cases are required: t_1 and t_7;
- **All-transitions:** Every transition of the specification is visited at least once. So, this coverage needed of two test cases: t_3 and t_{10};
- **All-transition-pairs:** Every pair of adjacent transitions in the specification must be visited at least once. For example, the test cases t_1, t_5, t_8 and t_{12} are required to obtain *all-transition-pairs* coverage;
- **All-loop-free-paths:** Every loop-free path must be traversed at least once. A path is loop-free when it does not have repetitions. For the example, this coverage is reached with the test cases: t_1, t_3, t_4, t_8, t_9, t_{10} and t_{11};
- **All-one-loop-paths:** Every path containing at most a loop must be traversed at least once. In other words, this requires all the loop-free paths through the specification to be visited, plus all the paths that loop once. Thus, we need all test cases to reach this coverage criterion: t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{11}, t_{12} and t_{13};
- **All-round-trips:** This coverage criterion is similar to *all-one-loop-paths* since it requires that all loops are tested in the specification. However, it is a weaker criterion, since it only requires one path for testing one loop. In this coverage criterion, the test cases t_1, t_3, t_4, t_8, t_{10} and t_{12} are required;
- **All-paths:** Every path must be visited at least once. The *all-paths* criterion corresponds to an exhaustive testing in LTS or ALTS specifications. In practice, the generation algorithm should have a heuristic to avoid the state space explosion, enabling the proper use of this coverage.

[1]It is important to remark that, in this work, *all-configurations* is not applied since it is mostly used for statecharts.

Figure 2.6 presents the hierarchy of transition-based criteria. Note that, $A \rightarrow B$ means that criterion A is stronger than (subsumes) criterion B, i.e., the coverage of the criterion A also achieves 100% coverage of the criterion B.

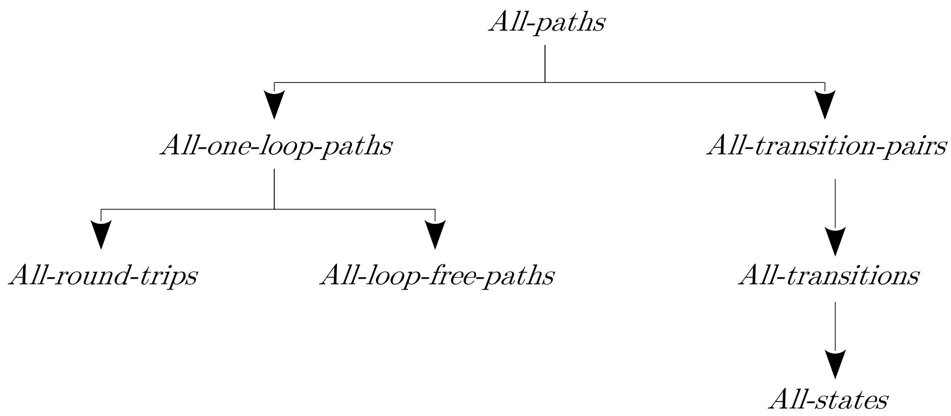


Figure 2.6: *The hierarchy of transition-based criteria. Font: adapted from [Utting and Legeard 2007]*

The use of some of these coverage criteria, such as: *all-loop-free-paths*, *all-one-loop-paths*, is not guarantee that all states, let alone all transitions, are covered [Utting and Legeard 2007]. Furthermore, they recommend the use of *all-transitions* coverage as a minimum measure of quality.

2.5 Test Suite Reduction

According to Harrold et al. [Harrold et al. 1993], the test suite reduction problem can be defined as follows:

Given: A test suite TS, a set $Req = \{req_1, req_2, \ldots, req_n\}$ of test requirements to be covered, and subsets of TS: $TS_1, TS_2, \ldots, TS_n$, where each test case of TS_i can be used to test req_i;
Problem: Find a minimal subset – the reduced set – $RS \subseteq TS$ that satisfies all of the *Req's*, that is, RS must have at least one test case for each req_i.

In general, finding RS is an NP-complete problem (minimization problems are NP-complete since they can be reduced to the *minimum set-covering* problem) [Cormen et al.

2001]. Therefore, heuristics and approximations are often applied to compute RS, such as the ones presented by Chen and Lau [Chen and Lau 1998b].

As mentioned before, in order to apply a reduction strategy it is necessary to define a *satisfiability relation* between TS and Req, relating each req_i to the set of test cases TS_i that cover it. Table 2.2 presents the *satisfiability relation* for the test suite presented in Table 2.1 for *all-transitions* and *all-transition-pairs* coverage criteria.

Table 2.2: *Satisfiability relations*

(a) *All-transitions* coverage

req_n	req	TS_n
req_1	(a)	$\{t_1, t_2, t_3, t_4, t_5, t_6\}$
req_2	(b)	$\{t_1, t_2, t_3, t_4, t_5, t_6\}$
req_3	(c)	$\{t_1, t_4, t_8, t_{10}, t_{12}\}$
req_4	(d)	$\{t_7, t_8, t_9, t_{10}, t_{11}, t_{12}, t_{13}\}$
req_5	(e)	$\{t_3, t_4, t_5, t_6, t_8, t_9, t_{10}, t_{11}, t_{12}, t_{13}\}$
req_6	(f)	$\{t_2, t_3, t_4, t_5, t_6, t_9, t_{11}, t_{12}, t_{13}\}$
req_7	(g)	$\{t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{11}, t_{12}, t_{13}\}$
req_8	(h)	$\{t_2, t_3, t_5, t_7, t_8, t_9, t_{11}\}$
req_9	(i)	$\{t_2, t_3, t_5, t_7, t_8, t_9, t_{11}\}$

(b) *All-transition-pairs* coverage

req_n	req	TS_n
req_1	(a, b)	$\{t_1, t_2, t_3, t_4, t_5, t_6\}$
req_2	(d, h)	$\{t_7, t_8, t_9\}$
req_3	(d, e)	$\{t_{10}, t_{11}, t_{12}, t_{13}\}$
req_4	(b, c)	$\{t_1\}$
req_5	(b, f)	$\{t_2, t_3, t_4, t_5, t_6\}$
req_6	(f, g)	$\{t_2, t_3, t_4, t_5, t_6, t_{11}, t_{12}, t_{13}\}$
req_7	(h, i)	$\{t_2, t_3, t_5, t_7, t_8, t_9, t_{11}\}$
req_8	(e, c)	$\{t_4, t_8, t_{10}, t_{12}\}$
req_9	(e, f)	$\{t_5, t_6, t_9, t_{11}, t_{12}, t_{13}\}$
req_{10}	(g, h)	$\{t_2, t_3, t_5, t_7, t_{11}\}$
req_{11}	(g, e)	$\{t_4, t_5, t_6, t_8, t_9, t_{12}, t_{13}\}$
req_{12}	(i, g)	$\{t_2, t_7, t_8, t_9\}$

For *all-transition-pairs* as coverage criterion, t_1 is *essential test case*. An essential test case is one that uniquely covers a given requirement, i.e., any test suite reduction strategy will keep it. However, it is important to remark that if we consider a weaker test criterion such as *all-transitions*, we would require at least one test case covering b and c, but not necessarily both in the same test case, i.e., (b, c). In this case, we cannot guarantee that the reduction strategy will select t_1. Therefore, the choice of coverage criteria that define test requirements can really influence fault detection capability of the reduced suite.

2.5.1 Heuristics for Test Suite Reduction

As said before, the goal of test suite reduction is to find a subset of the complete test suite (Reduced Set RS) that satisfies all test requirements.

In the sequence, we describe four well-known heuristics proposed for code-based reduction: Greedy (G) [Chvätal 1979; Cormen et al. 2001], GE [Chen and Lau 1998b], GRE [Chen and Lau 1998a] and HGS [Harrold et al. 1993]. These heuristics can also be applied on test suites obtained from MBT approaches.

As running example, we consider the test suite presented in Table 2.1 considering *all-transition-pairs* as coverage criterion shown in Table 2.2 (b).

Greedy Heuristic (G)

Greedy heuristic [Chvätal 1979; Cormen et al. 2001] repeatedly selects the test case t that satisfies the maximum number of unsatisfied test requirements while not all test requirements are satisfied. If there is a tie situation, a random choice is made. The test case t is added to Reduced Test Suite (RS) and all test requirements that can be satisfied by that test case are marked as a satisfied test requirement.

By applying this heuristic, we have:

1. t_5 satisfies the maximum number of unsatisfied test requirements. Then, $RS = \{t_5\}$ and the requirements req_1, req_5, req_6, req_7, req_9, req_{10} and req_{11} are marked as satisfied;

2. Now, t_8 satisfies the maximum number of unsatisfied test requirements. Then, $RS = \{t_5, t_8\}$ and the requirements req_2, req_8 and req_{12} are marked as satisfied;

3. Next, there is a tie situation: t_1, t_{10}, t_{11}, t_{12} and t_{13} satisfy the maximum number of unsatisfied test requirements. This way, an arbitrary choice is made. Then, t_1 is chosen, the reduced subset is $RS = \{t_5, t_8, t_1\}$ and the requirement req_4 is marked as satisfied;

4. Finally, the only requirement that it is not marked yet is req_3, and there is a tie situation among t_{10}, t_{11}, t_{12} and t_{13}. This way, an arbitrary choice is made. Then, t_{12} is chosen, the reduced subset $RS = \{t_5, t_8, t_1, t_{12}\}$, and req_3 is marked as satisfied.

Heuristic Greedy - Essential (GE)

This heuristic was defined by Chen and Lau and is based on the following concepts [Chen and Lau 1998b]:

- Essential concept - selects all essential test cases. A test case is essential when only this test case covers one specific requirement;
- Greedy heuristic.

Initially, all essential test cases are selected, and their respective requirements are marked as satisfied. Then, the greedy heuristic is applied.

By applying this heuristic, we have:

1. First of all, t_1 is selected, because they is essential test case. Then, $RS = \{t_1\}$ and the requirements req_1 and req_4 are marked as satisfied;
2. Since there are no more essential test cases, the greedy heuristic should be applied. Now, t_5 satisfies the maximum number of unsatisfied test requirements. Then, $RS = \{t_1, t_5\}$ and the requirements req_5, req_6, req_7, req_9, req_{10} and req_{11} are marked as satisfied;
3. Now, t_8 satisfy the maximum number of unsatisfied test requirements. Then, $RS = \{t_1, t_5, t_8\}$ and the requirements req_2, req_8 and req_{12} are marked as satisfied;
4. Finally, the only requirement that it is not marked yet is req_3, and there is a tie situation among t_{10}, t_{11}, t_{12} and t_{13}. Then, t_{12} is chosen, the reduced subset is $RS = \{t_1, t_5, t_8, t_{12}\}$, and req_3 is marked as satisfied.

Heuristic Greedy – 1-to-1 – Redundancy Essential (GRE)

This heuristic (also defined by Chen and Lau) is based on [Chen and Lau 1998a]:

- Greedy heuristic;

- 1-to-1 redundancy strategy - A test case $t_{1-1} \in TS$ is said to be 1-to-1 redundant if there is $t \neq t_{1-1}$ and $t \in TS$ such that $req(t_{1-1}) \subseteq req(t)$ [Chen and Lau 1998a], i.e., if all requirements that are covered by the test case t_{1-1} are covered by the test case t. Then the test case t_{1-1} is considered to be 1-to-1 redundant;

- Essential strategy.

The essential and 1-to-1 strategies are applied alternatively, until there is no essential and 1-to-1 redundant test cases. The greedy strategy is only applied if neither the essential nor 1-to-1 redundancy can be applied.

By applying this heuristic, we have:

1. t_1 is selected, because they are essential test cases. Then $RS = \{t_1\}$, and the requirements req_1 and req_4 are marked as satisfied;

2. Now, we do not have more essential test cases. So, we have to search 1-to-1 redundant test cases. (t_2, t_3), (t_2, t_5), (t_3, t_5), (t_5, t_6), (t_{10}, t_{12}) and (t_{12}, t_{13}) are 1-to-1 redundant test cases, since $req(t_3) \subseteq req(t_2) \subseteq req(t_5), req(t_6) \subseteq req(t_5), req(t_{10}) \subseteq req(t_{12})$ and $req(t_{13}) \subseteq req(t_{12})$. This way, t_2, t_3, t_6, t_{10} and t_{13} are not considered, since those are redundant in relation to t_5 and t_{12}, respectively;

3. Since there are no more essential test cases and 1-to-1 redundant test cases, the greedy heuristic should be applied. Now, t_5 satisfies the maximum number of unsatisfied test requirements. Then, $RS = \{t_1, t_5\}$ and the requirements req_5, req_6, req_7, req_9, req_{10} and req_{11} are marked as satisfied;

4. Now, t_8 satisfy the maximum number of unsatisfied test requirements. Then, $RS = \{t_1, t_5, t_8\}$ and the requirements req_2, req_8 and req_{12} are marked as satisfied;

5. Finally, the only requirement that it is not marked yet is req_3, and there is a tie situation among t_{11} and t_{12}. Then, t_{12} is chosen and $RS = \{t_1, t_5, t_8, t_{12}\}$, and req_3 is marked as satisfied.

Heuristic HGS

Harrold et al. [Harrold et al. 1993] present a test suite reduction strategy, which we call heuristic HGS. The idea is to select test cases according to their degree of *essentialness*. For this, it is necessary to calculate the cardinality of each test requirement. The cardinality is the number of test cases which satisfy the test requirement.

First, the test requirement with the lowest cardinality is considered. When a test case is added to the reduced set, all requirements covered by that test case are marked. Among the unmarked test requirements with lowest requirement cardinality, the heuristic selects the which covers more requirements. If there is a tie, the heuristic chooses the test case that occurs most frequently at the next highest requirement cardinality and so on (if there is a tie and the requirement cardinality is maximum, then the random choice is applied). This heuristic stops when the reduced set has test cases that cover all test requirements. In general, the main idea is to select test cases according to their essentialness, i.e., keeping in the reduced set the test cases in the order of most essential to least essential.

By applying this heuristic, we have:

1. Initially, we need to calculate the cardinality of each test requirement. The results can be seen in Table 2.3.

Table 2.3: *Cardinality*

Cardinality	Requirements
1	req_4
3	req_2
4	req_3, req_8, req_{12}
5	req_5, req_{10}
6	req_1, req_9
7	req_7, req_{11}
8	req_6

2. For the lowest cardinality (in this case it is one), there is only one test case t_1. Then, $RS = \{t_1\}$, and the requirements req_1 and req_4 are marked as satisfied;

3. Next, the lowest cardinality is 3 (req_2), there is a tie between t_7, t_8 and t_9. Then, we must see in the next highest requirement cardinality (in this case it is 4) which of them occurs most frequently. Then, t_8 is chosen because it occurs most frequently. Thus, $RS = \{t_1, t_8\}$ and the requirements req_2, req_7, req_8, req_{11} and req_{12} are marked as satisfied;

4. Now, the lowest cardinality is 4 (req_3), there is a tie between t_{10}, t_{11}, t_{12} and t_{13}. Then we must see in the next highest requirement cardinality (in this case it is 5) which of them occurs most frequently. Then, t_{11} is chosen because it occurs most frequently. Thus, $RS = \{t_1, t_8, t_{11}\}$ and the requirements req_3, req_6, req_9 and req_{10} are marked as satisfied;

5. Finally, the unique requirement that has not been yet marked is req_5, there is a tie between t_2, t_3, t_4, t_5 and t_6. However, we don't have any higher requirement cardinality, since all requirements of these cardinalities were already satisfied. Then, we apply a random choice between t_2, t_3, t_4, t_5 and t_6. Then, t_6 is chosen, the reduced subset is $RS = \{t_1, t_8, t_{11}, t_6\}$, and req_5 is marked as satisfied. Since all requirements are marked, and thus satisfied, the algorithm stops.

2.6 Distance Functions

In this section, we present five well-known distance functions and similarity functions to calculate the similarity degree between pairs of test cases. These functions are good candidates for detecting sequencing, matching, and/or repetition of transitions. We clarify that the terms distance functions and similarity functions are used without distinction in this thesis, since usually a similarity measure is the inverse of distance.

While other works have already applied these functions to similarity-based selection strategies [Heß2006; Vinson et al. 2007; Cartaxo et al. 2011; Hemmati et al. 2013; Fang et al. 2013], we apply these functions in the context of test suite reduction for MBT. It is important to remark that some of them needed to be slightly adapted to consider transition labels as the unit of comparison. Moreover, despite the fact that there are many other distance functions presented in the literature, our goal is to investigate the effect of distance functions, in general, on test suite reduction. For the sake of simplicity, we opt to choose a small set with the ones that are included in other studies in the general area of test case selection.

As running example, we consider test cases $t_2 = a, b, f, g, h, i, g, h, i$ and $t_5 = a, b, f, g, e, f, g, h, i$ from Table 2.1 that covers five *all-transition-pairs* (test requirements) in common: (g, h), (h, i), (f, g), (b, f) and (a, b) (see Table 2.2 (b)). These test cases start with editing user name, but differ by the subsequent operation, which is either another user name editing or removing a user.

2.6.1 Jaccard Index

The Jaccard's index, proposed by Jaccard [Jaccard 1901], is a similarity measure between sample sets. Let A and B be two sets of labels. The measure can be defined by the following function:

$$Jac(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where time complexity is $O(|A| + |B|)$.

In order to illustrate the Jaccard index, consider again test cases $t_2 = a, b, f, g, h, i, g, h, i$ and $t_5 = a, b, f, g, e, f, g, h, i$. Then, the calculation of Jaccard's index for test cases t_2 and t_5 is the following:

$$Jac(t_2, t_5) = \frac{|t_2 \cap t_5|}{|t_2 \cup t_5|} = \frac{|\{a, b, f, g, h, i\}|}{|\{a, b, f, g, e, h, i\}|} = \frac{6}{7} = 0.8571$$

Thus, the similarity degree between t_2 and t_5 calculated by using the Jaccard index is 85.71%.

2.6.2 Jaro Distance

The Jaro distance [Jaro 1989] is a measure of similarity between two strings. The idea of this measure is to calculate the similarity degree between two strings from the number of replacements of the position between characters (transpositions) and the number of different characters. Thus, given two strings $s_1 = a_1 \ldots a_k$ and $s_2 = b_1 \ldots b_l$ the Jaro distance is defined as:

$$Jaro(s_1, s_2) = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3} \times \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & otherwise \end{cases}$$

where:

- m is the number of matching characters;
- t is half the number of transpositions;
- $O(|s_1| + |s_2|)$ is the time complexity.

For instance, the number of matchings between test cases $t_2 = a, b, f, g, h, i, g, h, i$ and $t_5 = a, b, f, g, e, f, g, h, i$ is $m = 7$ and half the number of transpositions is $t = 1$, then:

$$Jaro(t_2, t_5) = \frac{1}{3} \times \left(\frac{7}{9} + \frac{7}{9} + \frac{7-1}{7} \right) = \frac{2.412}{3} = 0.8042$$

Thus, the similarity degree between t_2 and t_5 is 80.42%.

2.6.3 Jaro-Winkler Distance

The Jaro-Winkler distance [Winkler 1999], denoted JW, is a variant of the Jaro distance presented in Section 2.6.2, with the addition of the weighted prefix. Given two strings s_1 and s_2, the function is defined as:

$$JW(s_1, s_2) = Jaro(s_1, s_2) + \ell p(1 - Jaro(s_1, s_2))$$

Where:

- ℓ is the length of common prefix shared by the two strings with a maximum of four characters;
- p is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. p should not exceed 0.25, otherwise the distance can become larger than 1. The standard value for this constant in Winkler's work is $p = 0.1$;
- $O(|s_1| + |s_2|)$ is the time complexity.

The difference between Jaro and Jaro-Wrinkler is that Jaro-Winkler adds more weight in strings starting with the exact match characters. However, the maximum size for common prefix must be four, i.e, all matching characters past the first four have the same weight. The length of common prefix is multiplied by a constant, the standard being 0.1 for Jaro-Winkler distance.

For example, considering $p = 0.1$ and the test cases $t_2 = a, b, f, g, h, i, g, h, i$ and $t_5 = a, b, f, g, e, f, g, h, i$, then $\ell = 4$ and $Jaro(t_2, t_5) = 0.8042$. Then, the Jaro-Winkler distance is:

$$JW(t_2, t_5) = 0.8042 + 0.4(1 - 0.8042) = 0.8825$$

Thus, the similarity degree between t_2 and t_5 for Jaro-Winkler distance is 88.25%.

2.6.4 Levenshtein Distance

Levenshtein [Levenshtein 1966] proposes the distance function of editing, called *editDistance*. This function compares two strings and determines the minimum number of edit operations (deletion, insertion, and substitution) necessary to transform one string into another.

Consider two strings, A and B, where i and j are, respectively, their lengths. Firstly, a matrix M with $(i + 1) \times (j + 1)$ values is built, where the first row and the first column are initialized with values from 0 (incremented by 1) to the size of the test cases. The idea is to calculate the distances among all the prefixes of the first string A and all the prefixes of the second string B in a dynamic programming fashion. As the matrix is built, only the previous row (p) and the current row (q) are needed to calculate the current value of the matrix, where this value is the minimum of the three possible ways to do the transformation:

- *deletion:* $M[(p - 1, q)] + 1$;
- *insertion:* $M[(p, q - 1)] + 1$;
- *substitution:* $M[(p - 1, q - 1)] + cost$, where $cost = 0$ if $A[p] = B[q]$, otherwise $cost = 1$.

The value of $M[i+1,j+1]$ reflects the minimum number of operations necessary to convert one test case into another, i.e., the cost of the best sequence of edit operations. The degree similarity can be calculated in the interval of $[0,1]$ by the following function:

$$Lev(A,B) = 1 - \frac{M[i+1,j+1]}{max(i,j)}$$

where the time complexity is $O(|A| \times |B|)$.

For example, from Matrix 2.1, the similarity value between t_2 and t_5, calculated by Levenshtein distance is 77.78%, where $i = |t_2| = 9$, $j = |t_5| = 9$ and $M[9+1,9+1] = 2$ (*box contents*), obtained by calculation of:

$$Lev(t_2,t_5) = 1 - \frac{M[6+1,6+1]}{max(9,9)} = 1 - \frac{2}{9} = \frac{7}{9} = 0.7778$$

$$M = \begin{array}{c} \\ \\ a \\ b \\ f \\ g \\ e \\ f \\ g \\ h \\ i \end{array} \begin{array}{c} \begin{array}{cccccccccc} & a & b & f & g & h & i & g & h & i \end{array} \\ \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 3 & 2 & 1 & 1 & 2 & 3 & 3 & 4 \\ 6 & 5 & 4 & 3 & 2 & 2 & 2 & 3 & 4 & 3 \\ 7 & 6 & 5 & 4 & 3 & 3 & 3 & 2 & 3 & 4 \\ 8 & 7 & 6 & 5 & 4 & 4 & 4 & 3 & 2 & 3 \\ 9 & 8 & 7 & 6 & 5 & 5 & 5 & 4 & 3 & \boxed{2} \end{pmatrix} \end{array} \tag{2.1}$$

2.6.5 Sellers Algorithm

The algorithm proposed by Sellers [Sellers 1980] is a variation in the *editDistance* algorithm [Levenshtein 1966] (presented in Section 2.6.4) that modifies the way the matrix is created. The idea is to search for a string (sub-chain) in another string with a difference in at most k operations. Unlike the *editDistance* algorithm, the first row of the matrix is initialized with 0. This changes the calculation of the minimum number of operations to perform the transformation, from string A to string B, by ignoring any prefix of the string B. The degree of similarity is calculated by the same formula presented in Section 2.6.4. Also Sellers can be calculated in $O(|A| \times |B|)$ time.

$$Sel(A, B) = 1 - \frac{M[i+1, j+1]}{max(i, j)}$$

For example, considering test cases t_2 and t_5, the Sellers algorithm creates Matrix 2.2, where $i = |t_2| = 9$, $j = |t_5| = 9$ and $M[9+1, 9+1] = 2$ (*box contents*).

So, t_2 and t_5 are 77.78% redundant – the same value obtained by the Levenshtein distance (as shown bellow). But note that the base matrixes are different

$$Sel(t_2, t_5) = 1 - \frac{M[10, 10]}{Max(9, 9)} = 1 - \frac{2}{9} = \frac{7}{9} = 0.7778$$

$$M = \begin{array}{c} \\ \\ a \\ b \\ f \\ g \\ e \\ f \\ g \\ h \\ i \end{array} \begin{array}{c} \begin{array}{cccccccccc} & a & b & f & g & h & i & g & h & i \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 3 & 2 & 1 & 0 & 1 & 2 & 2 & 3 & 3 & 3 \\ 4 & 3 & 2 & 1 & 0 & 1 & 2 & 2 & 3 & 4 \\ 5 & 4 & 3 & 2 & 1 & 1 & 2 & 3 & 2 & 3 \\ 6 & 5 & 4 & 3 & 2 & 2 & 2 & 3 & 3 & 2 \\ 7 & 6 & 5 & 4 & 3 & 3 & 3 & 2 & 3 & 3 \\ 8 & 7 & 6 & 5 & 4 & 4 & 4 & 3 & 2 & 3 \\ 9 & 8 & 7 & 6 & 5 & 5 & 5 & 4 & 3 & \boxed{2} \end{pmatrix} \end{array} \tag{2.2}$$

2.7 Experimental Studies in Software Engineering

According to Wohlin et al. [Wohlin et al. 2012], there are four research methods in software engineering. These methods are:

- **The scientific method:** A model is built from observation of the world;
- **The engineering method:** Current solutions are studied, and the most appropriate changes are suggested, and then evaluated;
- **The empirical method:** A model is proposed and evaluated through empirical studies;

- **The analytical method:** A formal theory is proposed, and then it is compared with empirical observations.

In order to obtain objective and significant results, we opted for the application of appropriate empirical methods. These methods include surveys, case studies and experiments. Surveys aim to obtain descriptive and explanatory conclusions of a population from a sample based on forms, interviews and questionnaires. Case studies are used for monitoring projects or activities from the data collection for a specific purpose. In turn, experiments are rigorous, formal and controlled investigations providing a high level of control.

In this work, we focus in experimentation in software engineering aiming to increase the confidence in obtained results. For this, we use Wohlin et al. process for experimental studies in software engineering. The following activities are part of this process [Wohlin et al. 2012]:

- **Definition:** The purpose of this activity is to define the objective and goals of the experiment, i.e, the hypothesis has to be stated clearly;

- **Planning:** In the planning phase is defined how the experiment should be conducted. In this sense, several elements need to be specified, such as:

  - ***Context Selection:*** Define the context where the experiment will be conducted. Wohlin et al. [Wohlin et al. 2012] classifies the context of the experiment in four dimensions: *off-line vs. on-line*, *student vs. professional*, *toy vs. real problems* and *specific vs. general*;

  - ***Variables Selection:*** The dependent (observed) and independent (controlled and modified) variables characterize the experiment are chosen;

  - ***Hypothesis Formulation:*** Based on the dependent and independent variables, the null and alternative hypotheses of the experiment are defined;

  - ***Selection of Subjects:*** The subjects of an experiment are the people involved in it. Their selection of subjects is important for generalization of the results of the experiment;

  - ***Experiment Design:*** In this step, the experiment design is defined based on the statistical assumptions from the characteristics of the experiment, such as:

amount of object, subjects, factors and levels (treatments) [Jain 1991; Wohlin et al. 2012]. One factor (independent variable) can have varying values named treatments (or levels) that when changed will affect the dependent variables. The choice of the correct experiment design is crucial since misleading conclusions can appear. Most common experiments designs are [Wohlin et al. 2012]:

  * *One factor with two treatments:* To compare the two treatments against each other;
  * *One factor with more than two treatments:* To compare the treatments with each other, and each comparison is often performed on the treatment mean;
  * *Two factors with two treatments:* To compare the treatments in each factor with the others, for example 2*2 factorial design;
  * *More than two factors each with two treatments:* To compare the treatments in each one of the factors with those from the others. This type of designs is known as *factorial designs*;

- ***Instrumentation:*** According to Wohlin et al. [Wohlin et al. 2012], the instrumentation of an experiment can be characterized by three types of instruments: objects (the artifacts used), guidelines (to properly guide the subjects) and measurements (to conduct the data collection);
- ***Threats to Validity:*** The identification of potential threats to the validity is an important question concerning experiment results. Cook and Campbell [Cook and Campbell 1979] suggest that the threats can be identified according to the type of validation of results, such as:
  * *Conclusion validity:* The conclusion validity is concerned with the relationship between the treatment and the outcome;
  * *Internal validity:* This validity is concerned with the relationship between the treatments;
  * *Construct validity:* The construct validity is concerned with the relation between theory and observation;
  * *External validity:* This validity is concerned with the ability to generalize the results.

– **Operation:** In the operational phase the preparation, execution and data validation of an experiment are performed;

– **Analysis and interpretation:** The data collected during operation phase are analyzed and interpreted by using descriptive statistic to draw conclusions regarding the hypothesis;

– **Presentation and package:** The main concern of the last activity is to present the conclusions and artifacts of the experiment, so they can be properly presented to other researches.

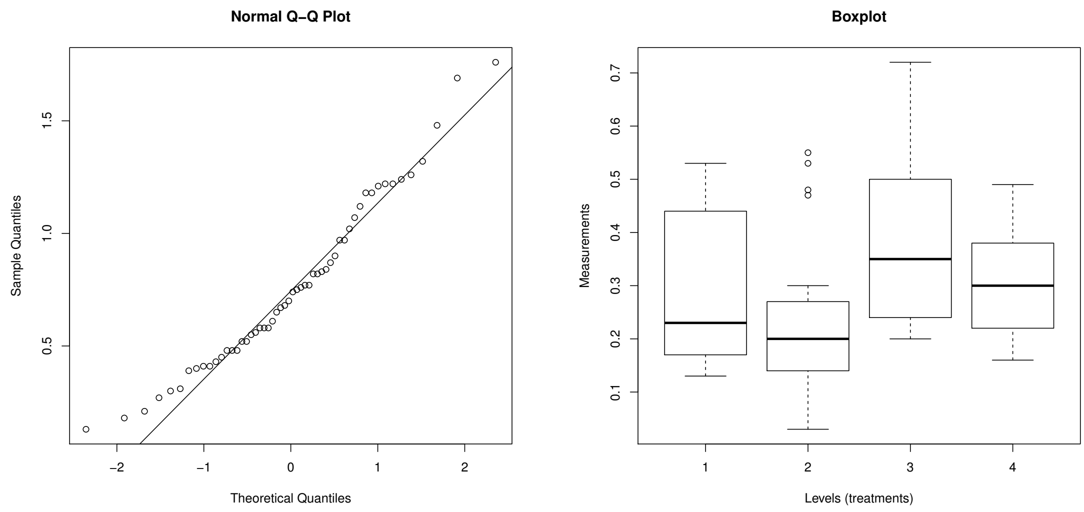
2.8 Statistical Analysis

After the data are collected, we can use descriptive statistics to describe and graphically present those data. Then, in order to obtain more significant conclusions, hypothesis testing allows researchers to verify if the null hypothesis can be rejected or not based on a sample from some statistical distribution according to a level of significance.

2.8.1 Descriptive Statistic

Descriptive statistic is used to better understand the data distribution (to identify abnormal data points), i.e, to check if data collected have or not a normal distribution [Jain 1991]. For this, graphical representation can be used to obtain information regarding the data collected such as mean, median, mode, variance, frequency, among others. There are several types of graphic representation. Among the graphics, we highlight normal quantile-quantile plot and boxplot. Figure 2.7 show examples of (a) a normal quantile-quantile plot and (b) a boxplot.

Normal quantile-quantile plot for is used to compare two probability distributions by plotting their quantiles against each other. It is a plot for a continuous variable that helps to determine if your data is close to being normally distributed. In turn, the boxplot are used to display the distribution of data based on the *five number* summary: minimum, first quartile, median, third quartile, and maximum. Graphically, the minimum and maximum are represented by whiskers above and below the box. The central rectangle spans the first quartile to the third quartile, and the thicker line shows the median. The outliers (unfilled dots) repre-



(a) (b)

Figure 2.7: *Examples of normal Q-Q plot and boxplot*

sent the individual values beyond the whiskers. Boxplots are also useful for comparing two or more variables and a visual interpretation of the data. When there is an interval overlap, this apparently indicates no statistical difference between them. When there is no overlap, we can state the statistical differences.

2.8.2 Hypothesis Testing

Based on a sample from some statistical distribution, hypothesis testing allows to check if it is possible to reject a null hypothesis with more significant conclusions than visual interpretation. There are several statistical tests in literature that are classified into parametric and nonparametric tests. The parametric tests are based on a known distribution. In turn, the nonparametric tests do not make the same type of assumption concerning the distribution of parametric test [Wohlin et al. 2012]. These statistical tests can be chosen according to the data distribution of the sample and the experiment design used as presented in Table 2.4.

All tests consider the ρ-*values* of the applied test to determine if the null hypothesis can be rejected according to an established confidence level, i.e., if ρ-*value* is smaller than the significance value (α) then the null hypothesis can be rejected in favour of the alternative hypothesis.

Table 2.4: *Overview of statistical tests*

Experimental Design	Parametric Test	Nonparametric Test
One factor with two treatments	$t - test$	Mann-Whitney test
One factor with more than two treatments	Paired $t - test$	Wilcoxon test
Two factors with two treatments	ANOVA	Kruskal-Wallis test
More than two factors each with two treatments	ANOVA	Friedman test

2.9 Concluding Remarks

In this chapter we presented the theoretical foundations necessary for the understanding of this work. Our investigation is focused on functional approaches, more specifically MBT approaches, to reduce the generated test suites from the complete test suite that covers a given set of test requirements. As the inputs are LTS and ALTS specifications, then we present the most common transition-based coverage criteria that can be used to define test requirements. Furthermore, we present five distance functions that may be used to calculate the similarity degree between pairs of test cases considering the repetition of transitions. In this work, we perform experimental studies in order to investigate the effectiveness of distance functions for our test suite reduction strategy based on similarity in the context of MBT, and to evaluate and analyse our proposal given other well-known reduction heuristics with respect to different coverage criteria.

Chapter 3

Similarity-based Test Suite Reduction

In this chapter, we present a new strategy for test suite reduction based on similarity of test cases which allows the use of single or multiple coverage criteria in the MBT context. The idea of this strategy is to keep the most different ones in the test suite that together can meet a set of test requirements from the identification of the degree of similarity among all the pairs of test cases. In turn, we present our function for measuring the similarity degree between two test cases that considers repetition of transitions. Additionally, further coverage criteria can be used to improve diversity of test cases of the reduced test suite by avoiding severe reduction.

3.1 The Proposed Strategy

Based on the problem presented in Section 1.1, the goal of our strategy is to reduce the test suite based on the degree of similarity among the test cases. The reduced test suite satisfies the same set of test requirements as the complete one. Our strategy to reduce the test suite size (abbreviated as *Sim*) is presented in Listing 3.1. Hence, to apply our reduction strategy, the following inputs are necessary:

- **Test Suite:** The set of test cases that should be reduced;
- **Coverage Criteria:** The ordered list of coverage criteria. This list must be defined from the weakest to the strongest for a same family of coverage criteria, for instance, *all-states*, *all-transitions*, and *all-transition-pairs*. In case of incomparable criteria, a

random choice is made. For each test coverage criterion, a set of test requirements that should be covered from reduced test suite is generated;

- **Similarity Function:** The function used to calculate the similarity degree between two test cases. To present an overview of the similarity degree between all test case pairs of a test suite, Cartaxo [Cartaxo 2011] proposed the *Similarity Matrix*. This matrix is assembled by applying the similarity function for each pair of test cases in the test suite;

- **Choice Function:** The function that defines the order of analysis of a pair of test cases, i.e., when a pair of test cases is chosen from the matrix, this function defines which of the two test cases will be analysed first.

The Listing 3.1 presents the steps of our reduction strategy. The first loop is used to select additional test cases into the reduced suite by using several test coverage criteria. For each test coverage criterion, a set of test requirements is obtained. Then, the idea is to analyze all the values on the matrix starting from the highest value, considering only the test cases that were not yet selected, and verify whether even with the removal of the chosen test case from the complete test suite, the coverage of test requirements remains 100% of the test requirements of the current coverage criterion. For this, the similarity matrix is created from the test cases that were not yet selected based on similarity function previously defined (lines 2 – 5). Then, in the second loop the `allMarkedPairs` method (line 6) verifies that all similarity degree existing on the matrix were already analysed.

Inside the repeating structure (lines 6 – 25), the first step is to find the two most similar test cases in the test suite from the highest value on the similarity matrix (line 7). Whenever a tie exists among highest value, one pair is randomly chosen. In the second step, the order of analysis of these two test cases is defined by the choice function (lines 8 and 9). For example, this function may be chosen based on assumptions. For instance, the test case to be analysed first should be that one that has more transitions because they may have the chance to uncover more failures. In the next step (lines 10 – 22), the first test case chosen is removed from the test suite. Afterwards, we check if the union of these test cases that were not yet removed from the reduced test suite satisfies all the test requirements of the current test coverage criterion (`satisfyAllTestRequirements` method). If all

Listing 3.1 Similarity-based test suite reduction strategy

input: complete test suite (TS), ordered list of coverage criteria ($criteria$), similarity function (sf) and choice function (cf)

output: reduced test suite (RS)

```
RS ← {}
for all c in criteria do
  reqs ← getTestRequirements(c, TS) {test requirements satisfied by complete test suite from the c coverage criterion}
  RS_c ← TS − RS {test suite to be reduced for c criterion}
  matrix ← createMatrix(RS_c, sf) {similarity matrix based on similarity function of the test suite to be reduced}
  while (!allMarkedPairs(matrix)) do
    pair ← matrix.getAllMaxValues().shuffle.get(0) {the most similar pair of test cases}
    firstTestCase ← getFirstTestCase(pair, cf)
    secondTestCase ← getSecondTestCase(pair, cf)
    RS_c.remove(firstTestCase)
    if (satisfyAllTestRequirements(RS ∪ RS_c, reqs)) then
      matrix.remove(firstTestCase)
    else
      RS_c.add(firstTestCase)
      RS_c.remove(secondTestCase)
      if (satisfyAllTestRequirements(RS ∪ RS_c, reqs)) then
        matrix.remove(secondTestCase)
      else
        RS_c.add(secondTestCase)
        matrix.markedPair(pair)
      end if
    end if
  end while
  RS ← RS ∪ RS_c
end for
return orderTestSuite(RS)
```

the requirements are satisfied, the first test case chosen is also removed from the similarity matrix. Otherwise, the first test case is added back to the test suite, and then the other one (the second test case chosen) is removed from the test suite in a similar way. If the two test cases cannot be removed from the test suite, then the pair of test cases is marked as analyzed. While all similarity matrix is not completely analyzed, new pairs of test cases continue to be selected, removed and tested in the similarity matrix. Afterwards, the additional test cases for these test requirements are added in the reduced test suite (line 24). Finally, the test cases of the reduced test suite are put in order from the smallest value related to the sum of the similarity degrees of one test case with all the other test cases to the largest value by `orderTestSuite` method (line 26).

Regarding the complexity analysis of Listing 3.1, we are able to observe a repeating structure (`forall` command in line 2) that repeats m times, where m is the number of coverage criteria. In line 5 (`while` command), we can observe a *loop* that is executed for the worst case $n-1$ times, where n is the number of test cases in the test suite. Furthermore, within each iteration this *loop*, the method `getAllMaxValue` in line 6 to search the matrix for the highest similarity values is executed for the worst case $\frac{n^2-n}{2}$ times, where n is the number of test cases in the test suite. In line 26, the sorting algorithm has a worst-case running time of $O(n^2)$. Therefore, the Listing 3.1 has a complexity of $O((m \times (n-1) \times (\frac{n^2-n}{2})) + n^2) = O((m \times (\frac{n^3-2n^2+n}{2})) + n^2) = O((m \times n^3) + n^2) = O(n^3)$.

3.2 Our Similarity Function

Cartaxo et al. [Cartaxo et al. 2011] define a redundancy measure that calculates the similarity degree between two test cases defined as *paths*. The degree is measured as the number of identical transitions divided by the average of path length as shown by following function:

$$SF(i,j) = \frac{nit(i,j)}{avg(|i|,|j|)}$$

where:

- $nit(i,j)$ is the number of identical transitions between the two test cases;
- $avg(|i|,|j|)$ is the average between the paths length.

In this work, we expect that the result of a similarity function considering two inputs is a real value normalized in the range $[0, 1]$, where 0 means that there is no similarity between inputs and 1 means that the inputs are equal. However, the result of this function can be greater than 1 for inputs considering repeated transitions, i.e., if a loop is traversed more than once. For instance, the similarity degree between $t_2 = a, b, f, g, h, i, g, h, i$ and $t_5 = a, b, f, g, e, f, g, h, i$ is calculated as follows:

$$SF(t_2, t_5) = \frac{nit(t_2, t_5)}{avg(|t_2|, |t_5|)} = \frac{12}{avg(9, 9)} = 1.333$$

To address this limitation, we present here an extension of this redundancy measure that ensures that the degree of similarity between test cases without repeated transitions is identical to the value calculated by the original function. The key idea is to consider the relation between the number of identical transitions of a path and their correspondent occurrences in both test cases (pairs) with average path lengths and set of distinct transitions. Thus, to calculate the similarity degree between two test cases i and j, considering repetition of transitions, we propose the following function:

$$SF(i, j) = \frac{|sit(i, j)|}{avg(|sdt(i)|, |sdt(j)|)} + \frac{nip(i, j)}{avg(|i|, |j|)}$$

where:

- $sit(i, j)$ is the set of identical transitions between two test cases, i.e., the intersection between $sdt(i)$ and $sdt(j)$;
- $nip(i, j)$ is the number of identical transition pairs between the two test cases;
- $sdt(i)$ is the set of distinct transitions in the i test case.

The average between the number of identical transition pairs (nip) plus the size of the set of identical transitions (sit) between two test cases calculates how much a test case is similar to another one considering repeated transitions. This value is divided by the average between the averages of the paths length and the set of distinct transitions (sdt) in order to balance the similarities between two test cases. The time complexity is $O(|i| + |j|)$.

For example, the similarity degree between $t_2 = a, b, f, g, h, i, g, h, i$ and $t_5 = a, b, f, g, e, f, g, h, i$ is calculated as follows:

- *Set of distinct transitions:*
  - $|sdt(t_2)| = |\{a, b, f, g, h, i\}| = 6$;
  - $|sdt(t_5)| = |\{a, b, f, g, e, h, i\}| = 7$;
- *Set of identical transitions:*
  - $|sit(t_2, t_5)| = |sdt(t_2) \ \cap \ sdt(t_5)| = |\{a, b, f, g, h, i\}| = 6$;
- *Number of identical transition pairs:*
  - $nip(t_2, t_5) = 4$, as presented in Table 3.1;

Table 3.1: *Identical transition pairs*

Identical transitions	Number of transitions t_2	t_5	Identical transition pairs $(minimun\ between\ t_2\ and\ t_5)$
a	1	1	1
b	1	1	1
f	1	2	1
g	2	2	2
h	2	1	1
i	2	1	1
Number of identical transition pairs			7

- *Paths length:* $|t_2| = 9$ and $|t_5| = 9$.

Then,

$$SF(t_2, t_5) = \frac{|sit(t_2, t_5)|}{avg(|sdt(t_2)|, |sdt(t_5)|)} + \frac{nip(t_2, t_5)}{avg(|t_2|, |t_5|)} = \frac{6}{avg(6, 7)} + \frac{7}{avg(9, 9)} = \frac{13}{15.5} = 0.8387$$

Hence, the similarity degree of the test cases t_2 and t_5 is 83.87%.

3.3 Example

In order to illustrate the strategy, we consider the following inputs to apply our reduction strategy:

- **Test Suite:** Test suite described in Table 2.1 in Section 2.3;
- **Coverage Criteria:** Ordered list with *all-transitions* and *all-transition-pairs* criteria (bi-criteria);
- **Similarity Function:** Our similarity function proposal in Section 3.2;
- **Choice Function:** The choice function used is based on the number of transitions. The key idea of this choice function is to compare the size of the test cases and to keep in the matrix the test case that has more transitions, since it can represent the highest functionality coverage. If the lengths are the same, one of them is taken to be analysed randomly. It is important to remark that the well-known reduction heuristics in literature often select the test cases to compose the reduced suite among the ones that cover more test requirements.

In turn, all similarity degrees among all pairs of test cases are presented by a similarity matrix, presented in *SM* 3.1.

$$SM = \begin{array}{cc} & \begin{array}{ccccccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & t_{11} & t_{12} & t_{13} \end{array} \\ \begin{array}{c} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \\ t_{10} \\ t_{11} \\ t_{12} \\ t_{13} \end{array} & \left(\begin{array}{ccccccccccccc} & 0.381 & 0.444 & 0.667 & 0.364 & 0.471 & 0.000 & 0.222 & 0.000 & 0.333 & 0.000 & 0.235 & 0.000 \\ & & 0.889 & 0.593 & 0.839 & 0.615 & 0.640 & 0.444 & 0.593 & 0.000 & 0.593 & 0.308 & 0.320 \\ & & & 0.667 & 0.857 & 0.696 & 0.545 & 0.500 & 0.667 & 0.000 & 0.667 & 0.348 & 0.364 \\ & & & & 0.714 & 0.870 & 0.182 & 0.500 & 0.500 & 0.444 & 0.500 & 0.696 & 0.545 \\ & & & & & 0.815 & 0.462 & 0.571 & 0.714 & 0.182 & 0.714 & 0.444 & 0.538 \\ & & & & & & 0.190 & 0.348 & 0.522 & 0.235 & 0.522 & 0.545 & 0.667 \\ & & & & & & & 0.727 & 0.727 & 0.250 & 0.727 & 0.381 & 0.400 \\ & & & & & & & & 0.833 & 0.667 & 0.833 & 0.696 & 0.545 \\ & & & & & & & & & 0.444 & 1.000 & 0.696 & 0.727 \\ & & & & & & & & & & 0.444 & 0.706 & 0.500 \\ & & & & & & & & & & & 0.696 & 0.727 \\ & & & & & & & & & & & & 0.857 \\ & & & & & & & & & & & & \end{array} \right) \end{array} \quad (3.1)$$

By applying of Listing 3.1, we have:

1° coverage criterion. Sim is applied considering *all-transitions* as coverage criterion (see Table 2.2 (a)).

1. The maximum value of the similarity matrix is 1.000 for the pair of test cases $\{t_9, t_{11}\}$. As the two test cases have the same length, an arbitrary choice is made.

t_{11} is chosen and removed. Then, we removed t_{11} from the similarity matrix since $RS' = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{12}, t_{13}\}$ satisfies all requirements;

2. Now, the maximum value of the similarity matrix is 0.889 for the test cases $\{t_2, t_3\}$, and we removed t_3 because it has less transitions. Then $RS' = \{t_1, t_2, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{12}, t_{13}\}$ satisfies all requirements, and we removed t_3 from the similarity matrix;

3. The next maximum value is 0.87 for the test cases $\{t_4, t_6\}$. As the two test cases have the same length, an arbitrary choice is made. t_4 is chosen and removed. Thus, $RS' = \{t_1, t_2, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{12}, t_{13}\}$ satisfies all requirements, and we removed t_4 from the similarity matrix;

4. Now, the maximum value of the similarity matrix is 0.839 for the test cases $\{t_2, t_5\}$. As the two test cases have the same length, an arbitrary choice is made. t_2 is chosen and removed. Then $RS' = \{t_1, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{12}, t_{13}\}$ satisfies all requirements, and we removed t_2 from the similarity matrix;

5. The next maximum value is 0.833 for the test cases $\{t_8, t_9\}$. As the two test cases have the same length, an arbitrary choice is made. t_9 is chosen and removed. Thus, $RS' = \{t_1, t_5, t_6, t_7, t_8, t_{10}, t_{12}, t_{13}\}$ satisfies all requirements, and we removed t_9 from the similarity matrix;

6. Now, the maximum value of the similarity matrix is 0.815 for the test cases $\{t_5, t_6\}$, and we removed t_6 because it has less transitions. Then $RS' = \{t_1, t_5, t_7, t_8, t_{10}, t_{12}, t_{13}\}$ satisfies all requirements, and we removed t_6 from the similarity matrix;

7. The next maximum value is 0.727 for the test cases $\{t_7, t_8\}$. As the two test cases have the same length, an arbitrary choice is made. t_8 is chosen and removed. Thus, $RS' = \{t_1, t_5, t_7, t_{10}, t_{12}, t_{13}\}$ satisfies all requirements, and we removed t_8 from the similarity matrix;

8. Now, the maximum value of the similarity matrix is 0.706 for the test cases $\{t_{10}, t_{12}\}$, and we removed t_{10} because it has less transitions. Then $RS' = \{t_1, t_5, t_7, t_{12}, t_{13}\}$ satisfies all requirements, and we removed t_{10} from the similarity matrix;

9. The next maximum value is 0.538 for the test cases $\{t_5, t_{13}\}$, and we removed t_{13} because it has less transitions. Then $RS' = \{t_1, t_5, t_7, t_{12}\}$ satisfies all requirements, and we removed t_{13} from the similarity matrix;
10. Now, the maximum value of the similarity matrix is 0.462 for the test cases $\{t_5, t_7\}$, and we removed t_7 because it has less transitions. Then $RS' = \{t_1, t_5, t_{12}\}$ satisfies all requirements, and we removed t_7 from the similarity matrix;
11. The next maximum value is 0.444 for following test cases $\{t_5, t_{12}\}$, and we removed t_{12} because it has less transitions. However, $RS' = \{t_1, t_5\}$ does not satisfy all requirements, then t_{12} is added in RS and not removed from the similarity matrix. Subsequently, t_5 is removed, and as $RS' = \{t_1, t_{12}\}$ does not satisfy all requirements, then t_5 is added in RS' and not removed from the similarity matrix, and this pair ($\{t_5, t_{12}\}$) is marked. Then, $RS' = \{t_1, t_5, t_{12}\}$;
12. Finally, the last maximum value is 0.364 for the test cases $\{t_1, t_5\}$, and we removed t_1 because it has less transitions. Then $RS' = \{t_5, t_{12}\}$ satisfies all requirements, and we removed t_1 from the similarity matrix. Since all pairs are marked, and thus satisfied, the algorithm stops;

2° coverage criterion. Sim is applied considering *all-transition-pairs* as coverage criterion (see Table 2.2 (b)), considering the test suite $TS = \{t_1, t_2, t_3, t_4, t_6, t_7, t_8, t_9, t_{10}, t_{11}, t_{13}\}$ to be reduced, and only the test requirements not coverage by $RS' = \{t_5, t_{12}\}$, in this case are (b, c), (d, h) and (i, g).

1. The maximum value of the similarity matrix is 1.000 for the pair of test cases $\{t_9, t_{11}\}$. As the two test cases have the same length, an arbitrary choice is made. t_9 is chosen and removed. Then, we removed t_9 from the similarity matrix since $RS' = \{t_1, t_2, t_3, t_4, t_6, t_7, t_8, t_{10}, t_{11}, t_{13}\}$ satisfies all requirements;
2. Now, the maximum value of the similarity matrix is 0.889 for the test cases $\{t_2, t_3\}$, and we removed t_3 because it has less transitions. Then $RS' = \{t_1, t_2, t_4, t_6, t_7, t_8, t_{10}, t_{11}, t_{13}\}$ satisfies all requirements, and we removed t_3 from the similarity matrix;

3. The next maximum value is 0.870 for the test cases $\{t_4, t_6\}$. As the two test cases have the same length, an arbitrary choice is made. t_4 is chosen and removed. Thus, $RS' = \{t_1, t_2, t_6, t_7, t_8, t_{10}, t_{11}, t_{13}\}$ satisfies all requirements, and we removed t_4 from the similarity matrix;

4. Now, the maximum value of the similarity matrix is 0.833 for the test cases $\{t_8, t_{11}\}$. As the two test cases have the same length, an arbitrary choice is made. t_8 is chosen and removed. Then $RS' = \{t_1, t_2, t_6, t_7, t_{10}, t_{11}, t_{13}\}$ satisfies all requirements, and we removed t_8 from the similarity matrix;

5. The maximum value of the similarity matrix is 0.727 for the pair of test cases $\{t_7, t_{11}\}$ and $\{t_{11}, t_{13}\}$. As there was a tie, we randomly chose $\{t_{11}, t_{13}\}$. As the two test cases have the same length, an arbitrary choice is made. t_{11} is chosen and removed. Then $RS' = \{t_1, t_2, t_6, t_7, t_{10}, t_{13}\}$ satisfies all requirements, and we removed t_{11} from the similarity matrix;

6. Now, the maximum value of the similarity matrix is 0.667 for the test cases $\{t_6, t_{13}\}$. As the two test cases have the same length, an arbitrary choice is made. t_{13} is chosen and removed. Then $RS' = \{t_1, t_2, t_6, t_7, t_{10}\}$ satisfies all requirements, and we removed t_{13} from the similarity matrix;

7. The maximum value of the similarity matrix is 0.640 for the test cases $\{t_2, t_7\}$, and we removed t_7 because it has less transitions. As $RS'' = \{t_1, t_2, t_6, t_{10}\}$ does not satisfy all requirements ((b, c), (d, h) and (i, g)), then t_7 is added in RS'' and not removed from the similarity matrix. After this, t_2 is removed, and $RS'' = \{t_1, t_6, t_7, t_{10}\}$ satisfies all requirements, then we removed t_2 from the similarity matrix;

8. Now, the maximum value of the similarity matrix is 0.471 for the test cases $\{t_1, t_6\}$, and we removed t_1 because it has less transitions. However, $RS'' = \{t_6, t_7, t_{10}\}$ does not satisfy all requirements, then t_1 is added in RS'' and not removed from the similarity matrix. After this, t_6 is removed, and $RS'' = \{t_1, t_7, t_{10}\}$ satisfies all requirements, then we removed t_6 from the similarity matrix. Then,$RS'' = \{t_1, t_7, t_{10}\}$;

9. The next maximum value is 0.333 for the test cases $\{t_1, t_{10}\}$. As the two test cases

have the same length, an arbitrary choice is made. t_{10} is chosen and removed. However, $RS'' = \{t_1, t_7\}$ satisfies all requirements, and we removed t_{10} from the similarity matrix;;

10. Now, the next maximum value is 0.000 for the test cases $\{t_1, t_7\}$, and we removed t_1 because it has less transitions. As $RS'' = \{t_7\}$ does not satisfy all requirements, then t_1 is added in RS'' and not removed from the similarity matrix. After this, t_7 is removed, and $RS'' = \{t_1\}$ does not satisfy all requirements, then t_7 is added in RS'' and not removed from the similarity matrix, and this pair ($\{t_1, t_7\}$) is marked. Since all pairs are marked, and thus satisfied, the algorithm stops;

Reduced test suite. Finally, $RS = RS' \cup RS''$ is ordered from the test case with least similarity to the test case with most similarity. Thus, $RS = \{t_1, t_7, t_{12}, t_5\}$.

3.4 Concluding Remarks

This chapter presented a new strategy for similarity-based test suite reduction which allows the application of multiple criteria in the MBT context, in order to obtain the reduction of a test suite while simultaneously trying to maximize the fault coverage. For this, the key idea is to reduce the test suite from the removal of the most similar test cases with the use of multiple criteria to improve diversity of the test cases selected, and maintain 100% of the test requirements covered.

Consider our running example presented in Section 1.1, we applied our reduction strategy $1{,}000$ times. The rate of reduction for our reduction strategy considering *all-transitions* and *all-transition-pairs* as coverage criterion is similar to the G, GE, GRE and HGS heuristics (84.62 and 69.23%, respectively). In turn, considering *bi-criteria* as coverage criterion presents the best percentage for fault coverage, in average 60.10%, as opposed to 29.43 and 52.46% for *all-transitions* and *all-transition-pairs*, respectively, as presented in Table 3.2. Furthermore, it is important to say that our strategy have a lower rate of scattering for all coverage criteria, when compared to the heuristics G, GE, GRE and HGS. The rate of reduction that reaches 100% fault coverage for *all-transitions*, *all-transition-pairs* and *bi-criteria* are, respectively, 34.94, 46.61 and 46.89%.

Table 3.2: *Frequency of detection of each fault for Sim (%)*

All-transitions			*All-transition-pairs*			*Bi-criteria*		
Fault 01 t_5	Fault 02 t_7	Fault 03 t_{10}	Fault 01 t_5	Fault 02 t_7	Fault 03 t_{10}	Fault 01 t_5	Fault 02 t_7	Fault 03 t_{10}
50.6	35.0	34.2	50.5	67.6	39.3	70.0	64.1	46.2

Hence, the similarity-based reduction strategy aims at addressing the limitations discussed in Chapter 1 by applying the following:

- **Fault missing:** According to Black et al. [Black et al. 2004], when applying test suite reduction there is the possibility that a test case considered redundant from a coverage perspective is not included in the reduced suite, even though this test case failed. Therefore, many times the reduction strategies may eliminate desirable test cases. In this sense, Sim is a multi-criteria strategy where a weaker criteria and a stronger criteria are applied in other to improve diversity by avoiding severe reduction. The idea is that, even though extensive redundancy must be avoided, a little redundancy in the reduced suite may improve its chances of covering a fault from the use of multiple criteria;

- **Fault scattering:** Primarily, Sim applies a similarity function in order to guide the choice of the most different test cases, instead of metrics such as size and essentialness. The use of similarity functions to compare test cases during test selection has already proved to be effective in promoting diversity and therefore improving fault detection [Hemmati and Briand 2010]. The reason is that the most redundant test cases w.r.t. the reduced suite, will be the less likely ones to be included, in crescent order of the degree of similarity.

In the following chapters, we investigate whether the choice of a distance function can influence on the performance of our reduction strategy (Sim). Afterwards, we conduct an experimental study with Sim and other redution heuristics by varying the coverage criteria from single to bi-criteria.

Chapter 4

Investigating Distance Functions for Similarity-based Test Suite Reduction Strategy

In this chapter, we present an investigation about the effectiveness of distance functions for test suite reduction in the context of MBT with respect to suite size reduction and fault coverage. Moreover, we observe the *stability* of the strategy when considering different functions according to different subsets of test cases and faults. In this sense, we apply our reduction strategy based on similarity presented in Chapter 3 by considering six distance functions: our function (*Similarity Function*) presented in Section 3.2, and five well-known functions in literature, *Levenshtein distance*, *Sellers algorithm*, *Jaccard index*, *Jaro distance*, and *Jaro-Winkler distance*, presented in Section 2.6. This chapter summarizes the study presented in [Coutinho et al. 2014].

4.1 Motivation

Intuitively, the choice of a distance function may directly influence on the performance of our reduction strategy. For instance, the function can tune a strategy to an extent in which it becomes capable of revealing differences that may speed up the achievement of coverage and at the same time diversifying the choice of test cases for improving fault coverage. Another important issue is that since reduction strategies often face draws and handle them by random selection, distance functions may also influence on the *stability* of the strategy, that is, how variable are the results obtained in relation to selected test cases and fault coverage by subsequent runs of the strategy.

Applications of distance functions spread across different contexts such as medicine [Felipe et al. 2003], speech [Thakur and Sahayam 2013] and image [Felipe et al. 2006] recognition. Moreover, there are many distance functions proposed in the literature, usually applied to specific applications or contexts where they are recognized as more effective [Akleman and Chen 1999]. For instance, the use of distance functions and equivalence relations is the basis of several fault localization strategies [Renieres and Reiss 2003; Xie et al. 2013].

More specifically, in the context of software testing, efforts have already been made to compare distance functions for both test case selection [Hemmati et al. 2013] and prioritization [Ledru et al. 2009]. On the one hand, empirical studies have already shown that the choice of the function may influence on fault detection capability for the general test selection and test case prioritization problems [Yoo and Harman 2012; Hemmati et al. 2013]. Particularly, Hemmati et al. [Hemmati et al. 2013] present a study on test selection strategies based on similarity where they consider the choice of different distance functions combined with other parameters to decide on the best strategy for test case selection. Among the results on 320 variants applied to two industrial case studies, top candidates emerge, even though differences found are minor. Generally, studies point to the need for more investigation. On the other hand, to the best of our knowledge, there are no studies comparing the effectiveness of distance functions applied to test suite reduction strategies for MBT. Different from test selection strategies where the tester may decide on the number of test cases to select, test suite reduction strategies rely on requirements coverage. In this sense, the choice of a distance function may influence on the size of the reduced suite as it may or not optimize coverage.

In order to investigate the influence in the choice of distance functions to reduce test suites, we perform three empirical studies. The first two, that will be presented in Section 4.2, are controlled experiments focusing on two real-world applications with real faults, and 10 synthetic specification models automatically generated from the configuration of each application with the sets of faults randomly defined for each generated model according to the obtained percentage of faults from each correspondent real-world specification. As coverage criterion for the reduction strategy, we choose *all-transition-pairs* criterion [Utting and Legeard 2007]. This criterion is satisfied if all pairs of adjacent transitions in the specification are traversed at least once [Utting and Legeard 2007]. In the third study, presented in Section 4.3, we apply the reduction strategy to two versions of a real-world industrial application with real faults collected from manual execution of test cases.

4.2 Experimental Studies

This section presents the experimental studies and the obtained results of the execution. The next subsection describes the activities performed to define and execute the studies. Afterwards, the results and analysis are presented.

4.2.1 Experiment Planning

In this section, we present the definition of two empirical studies to assess the effectiveness of different distance functions applied in the scope of the similarity-based strategy for test suite reduction presented in Section 3. Both studies focus on considering a real-world application model and real faults experienced during test execution. The idea is to consider two different real settings of application model and fault detection percentage in order to investigate the functions in a controlled way.

The first empirical study focus on a version of the PDFSam tool[1]. This application has few essential test cases and, consequently, a great potential for reduction. The second empirical study focus on a version of the TaRGeT tool [Nogueira et al. 2007; Ferreira et al. 2010] composed mostly of essential test cases, making the reduction task harder.

Definition

As mentioned before, the goal of these empirical studies is to investigate *distance functions to measure similarity between two test cases* to assess the effectiveness when applied in a

[1] http://www.pdfsam.org/

test suite reduction strategy based on similarity. For this, we observe, for the reduced suite, the *size and fault coverage*. Based on this goal, our general hypothesis is that

> *"Test suite reduction strategies based on similarity show a different performance regarding size and fault coverage of the reduced suite depending on the distance function used."*

Furthermore, we analyze the results considering the point of view of the tester (responsible for the testing process) in the context of MBT.

Planning

In the phase of planning, we define context selection, variables, hypothesis, instrumentation, design, and threats to validity as follows.

Context Selection Following the dimensions proposed by Wohlin et al., presented in Section 2.7, the studies are *off-line*, i.e., we perform them in laboratory, which is not a real industrial environment. For more general results, an experiment should be performed in real settings (*online*). Each empirical study has as inputs to the reduction strategy (with the different distance functions) one real-world application (*real problems*) and 10 synthetic automatically generated specifications. These specifications are randomly generated by considering the same configuration of the respective real-world application such as depth, number of forks, number of transitions of forks, number of joins, number of transitions of joins, and number of paths with loop. Since these empirical studies focus only on two sets of different configurations, those studies can be characterized as a *specific*.

Variables Selection The dependent and independent variables that compose our studies are defined as follows:

- *Independent variables*
  - *Test requirements*: *all-transition-pair coverage*;
  - *Test suite reduction strategy*: *Similarity-Based Test Suite Reduction Strategy* (*Sim*);

- *Distance functions*: functions to measure the similarity degree between two test cases applied in the reduction strategy. In this work, we analyze the functions:
  * *Jac:* Jaccard index (Section 2.6.1);
  * *Jaro:* Jaro distance (Section 2.6.2);
  * *JW:* Jaro-Winkler distance (Section 2.6.3);
  * *Lev:* Levenshtein distance (Section 2.6.4);
  * *Sel:* Sellers algorithm (Section 2.6.5);
  * *SF:* Similarity function (Section 3.2).
- *Choice function*: the order of analysis of these two test cases is defined according to their path lengths. Then, the test case with the lower number of transitions is the first to be analyzed. If the test cases have the same length, one of them is chosen randomly;
- *Faults:* the faults revealed by the test suite. For the synthetic models, faults are automatically defined considering the same pattern of the real models: a test case fails due to one fault (one-to-one relationship);

• *Dependent variables*

- *Suite Size Reduction (SSR)*: percentage of the number of test cases removed from the complete test suite.

$$SSR = \frac{|TS| - |RS|}{|TS|} \times 100\%$$

  where $|TS|$ is the number of test cases in the complete test suite and $|RS|$ is the number of test cases in the reduced test suite;
- *Fault Coverage (FC)*: percentage of the total number of faults uncovered by the reduced test suite:

$$FC = \frac{|F_{RS}|}{|F_{TS}|} \times 100\%$$

  where $|F_{TS}|$ is the number of faults revealed by the complete test suite and $|F_{RS}|$ is the number of faults revealed by the reduced test suite.

Hypothesis Formulation The experiment definition is formalized into hypotheses that are tested during the analysis of the experiment. Based on the goal of the empirical studies, for each dependent variable (SSR and FC), we define two hypotheses as follows[2]:

1. SSR*: A null hypothesis (*H_1^0*):* all distance functions have the same behavior regarding suite size reduction; *An alternative hypothesis (*H_1^1*):* all distance functions have a different behavior regarding suite size reduction.

$$H_1^0 : SSR_{Jac} = SSR_{Jaro} = SSR_{JW} = SSR_{Lev} = SSR_{Sel} = SSR_{SF}$$
$$H_1^1 : SSR_{Jac} \neq SSR_{Jaro} \neq SSR_{JW} \neq SSR_{Lev} \neq SSR_{Sel} \neq SSR_{SF}$$

2. FC*: A null hypothesis (*H_2^0*):* all reduction strategies have the same behavior regarding the rate of fault coverage; *An alternative hypothesis (*H_2^1*):* all reduction strategies have a different behavior regarding the rate of fault coverage.

$$H_2^0 : FC_{Jac} = FC_{Jaro} = FC_{JW} = FC_{Lev} = FC_{Sel} = FC_{SF}$$
$$H_2^1 : FC_{Jac} \neq FC_{Jaro} \neq FC_{JW} \neq FC_{Lev} \neq FC_{Sel} \neq FC_{SF}$$

Instrumentation The instruments of the experiments are defined as follows.

1. *Objects:* 1 real-world and 10 synthetic automatically generated LTS specifications for each empirical study (22 specification models in total);
2. *Guidelines:* since the strategy does not require people (subjects) to configure them, no guideline is used;
3. *Measurements:* the LTS-BT tool [Cartaxo et al. 2008] is used to support the experiments execution and data collection.

The two real-world specifications selected for each empirical study are briefly described as follows:

- *PDFSam:* an open-source tool used to split and merge pdf documents;
- *TaRGeT:* an application that generates test cases from use case documents in a MBT process.

[2]Note that equations expressed as $a = b = c$, represent $a = b \wedge b = c \wedge a = c$, and $a \neq b \neq c$, represent $a \neq b \vee b \neq c \vee a \neq c$.

In these studies, we consider a specific version of each of the real-world applications in which faults can be observed. For these versions, in order to generate the specification models, we consider a specification of software requirements written as use cases, by experienced testers, using the use case template of the TaRGeT tool. As output, the TaRGeT tool returns an LTS model that represents the execution flows of the use cases. It is important to remark that the version of TaRGeT we consider as object of the study is different from the one we use for generating the models. The latter is a stable and deployed one. Furthermore, we collect the faults considered in the studies by manually executing the version under testing and manually identifying faults from failures.

Table 4.1 presents the configuration of the real specification models, defined as: *i*) structural measures (based on the concepts presented in Section 2.2.2); *ii*) the number of test cases generated by the LTS-BT tool considering *all-one-loop-paths* coverage criteria; *iii*) the number of essential test cases; *iv*) the number of faults detected. Notice that the two real-world specifications have a different number of faults. This is due to the fact that we consider only and exactly the real faults detected in order to make the results resemble the practice. Moreover, it is important to remark that for each real-world specification, each fault is revealed by a distinct failure (test case).

Table 4.1: *Basic configuration of the two real-world specifications*

	PDFSam	TaRGeT
Depth	18	8
Paths with loop	5	0
Forks	15	26
Transitions of forks	41	101
Joins	11	16
Transitions of joins	34	42
Test cases (one expansion)	137	82
Essentials Test Cases	0	62
Faults	5	13
Failures	5	13

From the configurations of each real-world model, we generate 10 synthetic LTS models based on the strategy presented by Oliveira *et al.* [Oliveira Neto et al. 2013]. The LTS generator receives as input the depth, the number of transitions of joins, joins, transitions of forks, forks, and paths of loops for each real-world specification. Then, it generates a number of different models (10 in this study) for each configuration.

Table 4.2 presents the number of test cases generated, essential test cases, and faults generated for each synthetic model. Notice that they resemble the correspondent real one.

Table 4.2: *Comparing test case and fault metrics of the synthetic LTS specifications to the corresponding real specification ones*

Configuration	#	Test Cases	Essentials (%)	Faults (%)
	real	137	0 (0.00)	5 (3.65)
	01	181	0 (0.00)	7 (3.86)
	02	189	1 (0.53)	7 (3.70)
	03	181	1 (0.55)	7 (3.86)
	04	150	1 (0.67)	5 (3.33)
PDFSam	05	110	2 (1.82)	2 (3.63)
	06	155	4 (2.58)	6 (3.87)
	07	105	3 (2.86)	4 (3.80)
	08	103	4 (3.88)	4 (3.88)
	09	97	4 (4.12)	4 (4.12)
	10	100	7 (7.00)	4 (4.00)
	real	82	62 (75.60)	13 (15.85)
	01	88	50 (43.48)	18 (15.65)
	02	103	46 (44.66)	16 (15.53)
	03	99	57 (57.58)	16 (16.16)
	04	94	55 (58.51)	15 (15.95)
TaRGeT	05	88	57 (64.77)	14 (15.90)
	06	87	57 (65.52)	14 (16.09)
	07	88	59 (67.05)	14 (15.90)
	08	86	64 (74.42)	14 (16.27)
	09	84	63 (75.00)	13 (15.47)
	10	88	67 (76.14)	14 (15.90)

For the synthetic models, we randomly selected a number of test cases that fail and associated each failure with a fault to follow the same pattern of the real models. Moreover, the number of failures/faults approximates the percentage of faults of the real applications w.r.t. the number of test cases (PDFSam configuration: 3.65% and TaRGeT configuration: 15.85%). Likewise, the percentage of essential test cases is also an approximation, but it lacks a little bit of precision due to the fact that distribution of essential test cases depends on the model and we did not control it directly. However, variation is low: the percentage of essential test cases ranges from 0% and 7% for PDFSam configuration and from 44.66% and 89.87% for TaRGeT configuration.

Experimental Design In this investigation, there is one experimental study of one factor (distance function applied in the reduction strategy) with more than two treatments (the six distance functions investigated) for each specification. Thus, there are 11 experimental studies for each empirical study (10 synthetic specifications and 1 real specification). These experimental studies are structured in two experimental designs, i.e., one experimental design for each metric observed (SSR and FC) as illustrated in Figure 4.1.

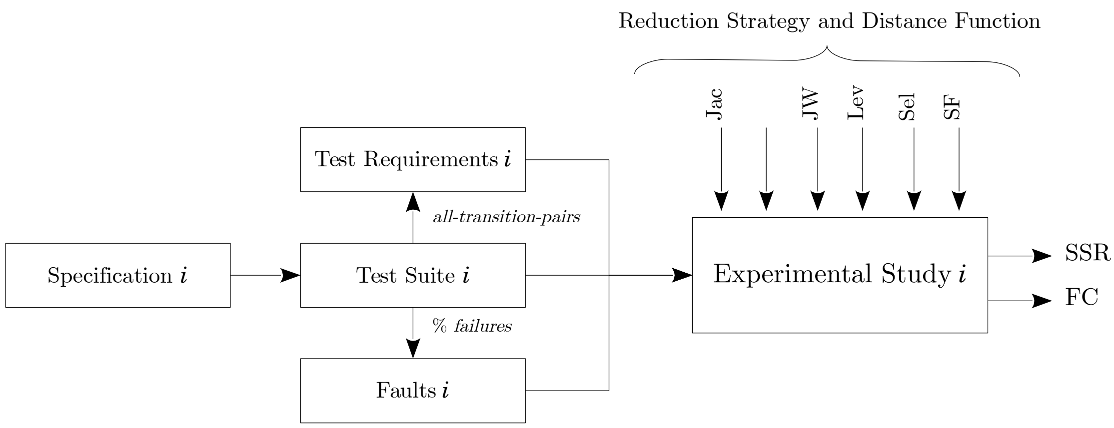


Figure 4.1: *Schema of the experimental study for each input specification*

As suggested in literature for experimental studies, we choose a confidence level of 95%. Then, we use $\alpha = 0.05$ whenever referring to statistical significance. Moreover, in order to obtain conclusions with statistical significance, the minimum sample size must be calculated. Thus, we execute the six distance functions 40 times to calculate the number of replications required (n), according to Jain [Jain 1991], for each metric in each one of the experimental study as follows:

$$n = \left(\frac{100.z.s}{r.\overline{x}}\right)^2$$

Where:

- z is a standard value from the normal distribution table, for a 95% confidence level $z = 1.96$;
- s is the standard deviation from the sample;
- r is the desired accuracy ($\alpha = 0.05$, then $r = 5$);
- $\overline{x}$ is the mean of the sample.

For each empirical study, we consider that the number of necessary replications is the highest value defined between the metrics SSR and FC for all specifications, as summarized in Table 4.3. Note that only the highest value for each metric in each empirical study is presented. For the configuration of the PDFSam application, the number of replications required is defined by the *JW* (*Jaro-Winkler distance*) function for Specification 02, observing the FC metric. In this case, we consider 62,000 replications of each distance function for each specification. In the configuration of TaRGeT, the highest value defined among the metrics (SSR and FC) of all specifications defined by *SF* (*Similarity Function*) for Specification 02, observing the FC metric. Therefore, for the configuration of TaRGeT, we consider approximately *40* replications of each distance function for each specification.

Table 4.3: *Mean, standard deviation and the highest number of necessary replications for each metric and each application*

	PDFSam		TaRGeT	
Metric	SSR	FC	SSR	FC
Distance Function	*Jaro*	*JW*	*SF*	*SF*
Specification	10	02	08	02
Mean ($\overline{x}$)	80.225	0.3571	15.813	68.281
Standard Deviation (s)	0.7675	2.2587	0.7354	9.3227
Necessary Replications (n)	0.1406	*61465.6*	3.3232	*28.645*

Operation

To execute these empirical studies, we implemented an LTS generator as proposed by Oliveira et al. [Oliveira Neto et al. 2013] to automatically generate the different specifications according to specific configurations. Furthermore, it was necessary to implement the distance functions and the *code to collect the data during the execution of the experiment*. We implemented them in the Java programming language[3]. Following this, we use the LTS-BT tool to generate test cases. Furthermore, we perform each step of the experiment for the maximum number of times defined among the metrics, using a machine with Intel Core(TM) i5 3.10 GHz, 8GB RAM running GNU Linux.

Threats to Validity

An important question concerning the results of the empirical studies are the potential threats to the validity that may negatively influence on the results, presented in Section 2.7.

The statistical tests used represent the main threat to *conclusion validity*. To deal with this threat, the number of executions of the experiments for each specification is eventually higher than the amount defined in the sample size. In order to maintain the statistical significance of the data, all analysis consider the confidence level of 95%, according to the suggestions for conducting experiments on the statistical literature [Jain 1991]. Thus, this ensures that we have a good conclusion validity.

A threat to *internal validity* is related to the control of the experiment. To make execution of the reduction strategy for each distance function automatic, during the implementation and execution of the strategies, we add control so that the execution environment would not be influenced by other processes, programs, or the machine on which the experiment is running. In these empirical studies, there are no people involved, and the same inputs (LTS specifications) are applied for all the distance functions. Thus, this internal validity is not considered critical.

For *construct validity*, the experiments setting is the main threat. To maintain the construct validity, the experimenter cannot influence on the measures. To handle this, the synthetic specifications are automatically generated from the configuration of real-world applications. Furthermore, our results rely on input specifications that have given a set of faults that were randomly generated. The number of faults for each configuration is defined according to the percentage of the real specification previously executed. In this real specification, the set of real faults is identified after each test case is manually executed by experienced software engineers.

[3]http://www.sun.com/java/

Another threat to construct validity is when the measurements of the metrics (SSR and FC) are not adequate. For this, these metrics are implemented according to the concepts proposed in the literature. Moreover, the implementation of the distance functions is another threat to validity. To deal with this, the distance functions are implemented according to the algorithms described in Section 2.6. In order to maintain the validity of the data, it is necessary to adapt the distance functions to calculate the degree of similarity between two test cases for these empirical studies.

The objects used in these experiments are the main threat to *external validity*, particularly, synthetic LTS specifications that are automatically generated, not representing a real behavior, even though they are randomly generated considering the same configuration of real applications. However, automation makes it possible to consider a number of specifications in a controlled way.

4.2.2 Analysis and Interpretation

The first step is to check whether data collected have a normal distribution for all specifications, considering the SSR and FC metrics. For this, we apply the Anderson-Darling normality test, using the R tool[4], considering the confidence level at 95% (significance level is $\alpha = 0.05$) [Jain 1991]. For the two empirical studies and all specifications, ρ-*values* are smaller than the significance value ($\alpha = 0.05$). Thus, we need to apply nonparametric tests. Since each experimental design has a unique factor with more than two treatments, we apply the nonparametric Kruskal-Wallis test to check the null hypotheses. This test is used to determine whether there are "*significant*" differences among the population medians. In the next subsections, we present and discuss these results, considering each empirical study.

First Empirical Study – PDFSam Configuration

For specifications 4 and 8, we obtain ρ-*values*$= 1.000$ by executing the Kruskal-Wallis test for both SSR and FC metrics. In other words, for these specifications and metrics, the distance functions have the same behavior with 95% confidence level. For the other specifications, we obtain ρ-*value* $= 0.0001$ by executing the Kruskal-Wallis test. These values are

[4] http://www.r-project.org/

smaller than the significance level ($\alpha = 0.05$) for all data. Thus, all null hypotheses can be rejected (H_1^0 and H_2^0), that is, for SSR and FC, the distance functions do not present the same behavior.

Figure 4.2 presents the boxplots for SSR and FC considering the general average in PDFSam configuration.

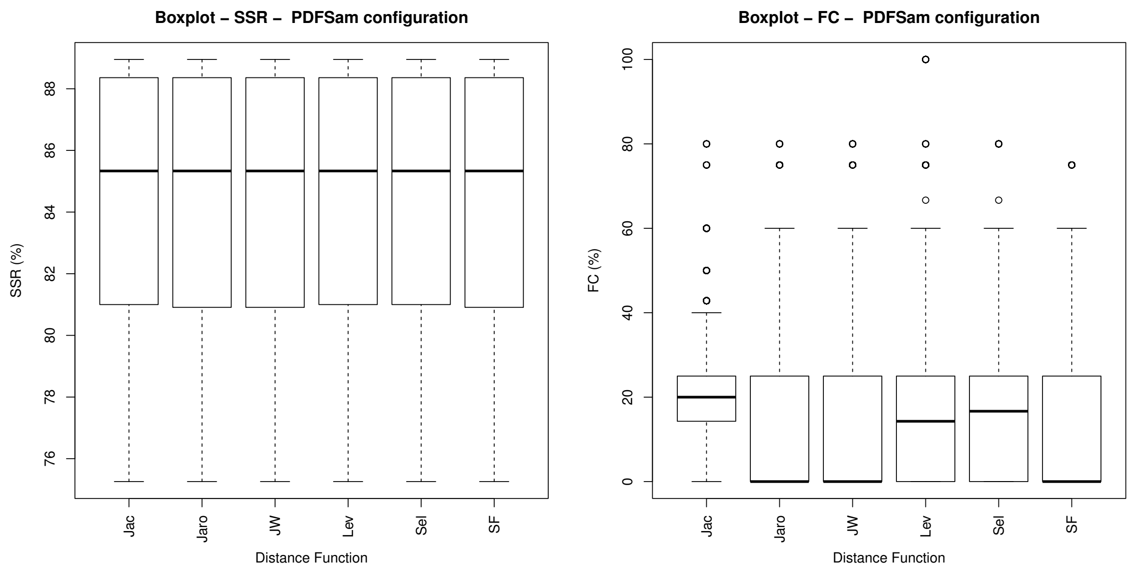


Figure 4.2: *Boxplots for SSR and FC considering the general average for PDFSam configuration*

As there are overlaps in the boxplots, we apply the Mann-Whiney test (Wilcoxon-Mann-Whitney test in *R*) between each pair of distance function. If the *ρ-value* $< \alpha$ for Mann-Whitney tests, then null hypothesis can be rejected in favor of the alternative hypothesis. In this case, the response variable tends to be either greater or smaller for one group in spite of the other group. For the other cases, such as *ρ-value* $\geq \alpha$, the null hypothesis cannot be rejected, and we conclude that the distance functions have similar behavior.

However, the Mann-Whitney test shows only whether there is a statistically significant difference between two treatments. In order to clarify the magnitude of the treatment effect, we use the $\hat{A}_{12}$ effect size measure proposed by Vargha and Delaney [Vargha and Delaney 2000]. Considering two treatments X and Y, $\hat{A}_{12} = 0.5$ indicates that there is no difference between the treatments X and Y, whereas $\hat{A}_{12} > 0.5$ indicates that X is superior to Y, and $\hat{A}_{12} < 0.5$ indicates that Y is superior to X. Note that $\hat{A}_{12}$ is between 0 and 1 and the larger the effect size, the further away the value from 0.5 is. We follow the categories used by Rogstad, Briand and Torkar [Rogstad et al. 2013], where they categorize the effect into

$Small < 0.10$, $0.10 < Medium < 0.17$ and $Large > 0.17$, the value being the distance from 0.5. Table 4.4 shows the Mann-Whitney U-tests and $\hat{A}_{12}$ effect size for each comparison considering the general average in PDFSam configuration.

Table 4.4: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for general average in PDFSam configuration*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
Jac and Jaro	0.000	Jac	Small (0.5127)	0.000	Jac	Large (0.6945)
Jac and JW	0.000	Jac	Small (0.5153)	0.000	Jac	Large (0.6962)
Jac and Lev	0.000	Lev	Small (0.4966)	0.000	Jac	Medium (0.6365)
Jac and Sel	0.000	Sel	Small (0.4967)	0.000	Jac	Small (0.5993)
Jac and SF	0.000	Jac	Small (0.5094)	0.000	Jac	Medium (0.6626)
Jaro and JW	0.000	Jaro	Small (0.5041)	1.493e-07	Jaro	Small (0.5017)
Jaro and Lev	0.000	Lev	Small (0.4824)	0.000	Lev	Small (0.4415)
Jaro and Sel	0.000	Sel	Small (0.4825)	0.000	Sel	Medium (0.3929)
Jaro and SF	0.000	SF	Small (0.4974)	0.000	SF	Small (0.4699)
JW and Lev	0.000	Lev	Small (0.4794)	0.000	Lev	Small (0.4397)
JW and Sel	0.000	Sel	Small (0.4795)	0.000	Sel	Medium (0.391)
JW and SF	0.000	SF	Small (0.4933)	0.000	SF	Small (0.4682)
Lev and Sel	0.000	Lev	Small (0.5001)	0.000	Sel	Small (0.4538)
Lev and SF	0.000	Lev	Small (0.5146)	0.000	Lev	Small (0.5285)
Sel and SF	0.000	Sel	Small (0.5144)	0.000	Sel	Small (0.575)

In all cases for SSR in Table 4.4, the effect size is classified as small between the distance functions. The results indicate that there is a small difference when applying different distance function combined with similarity-based reduction strategy, considering SSR. In terms of FC, the results show that when *Jac* is compared to others, its behavior is clearly better, with an effect size mostly from medium to large.

From the boxplots, Mann-Whitney tests and $\hat{A}_{12}$ effect size measurement, we calculate the average position of each distance function regarding effectiveness, as presented in Table 4.5. This table presents the performance order of the distance functions for the SSR and FC metrics.

Finally, by analyzing the data obtained in the 62,000 executions of the technique when considering each function, we can also observe the stability of the reduction technique with

Table 4.5: *Ordering of effectiveness for SSR and FC in PDFSam configuration*

	Suite Size Reduction (SSR)	Faults Coverage (FC)
real	$Jac > SF > Jaro = JW > Sel > Lev$	$Jac > SF > Jaro = JW = Sel > Lev$
01	$Sel > Lev > SF > Jaro = JW > Jac$	$Jac > Jaro = JW > Sel > Lev > SF$
02	$Jaro = JW = SF > Sel > Lev > Jac$	$Jac > Sel > Lev > Jaro = JW > SF$
03	$SF > Jac > Lev = Sel > Jaro = JW$	$Jac > SF > Lev = Sel > Jaro = JW$
04	$Jac = Jaro = JW = Lev = Sel = SF$	$Jac > Jaro = JW > Sel > Lev = SF$
05	$Jaro = JW > SF > Sel > Lev > Jac$	$Sel > Lev > SF > Jaro = JW > Jac$
06	$Jac > Lev > SF > Sel > JW > Jaro$	$Jac > Sel > Jaro > JW > Lev > SF$
07	$SF > Sel > Jac = Lev > Jaro > JW$	$SF > Jac > Lev > Sel > Jaro > JW$
08	$Jac > Lev > Sel > Jaro = JW = SF$	$Jac = Jaro = JW = Lev = Sel = SF$
09	$Lev > Jaro = JW > Sel = SF > Jac$	$Jac > Sel > Lev > Jaro = JW = SF$
10	$Lev > Sel > Jaro > Jac > SF > JW$	$SF > Jac > Jaro = JW = Lev = Sel$
All	$Lev > Sel > Jac > SF > Jaro > JW$	$Jac > Sel > Lev > SF > Jaro > JW$

respect to two measures: i) the number of different sets of faults produced by the selected suites; ii) the number of different sets of test cases selected (different suites). Ideally, the technique should be as stable as possible by presenting a low number of different sets in each case, making its performance more predictable.

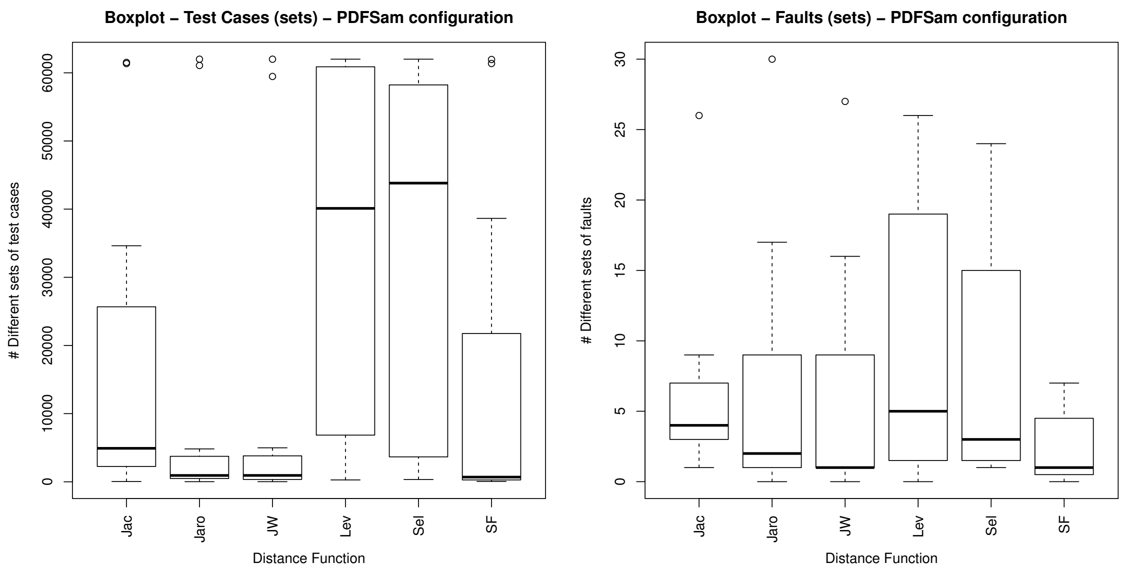


Figure 4.3: *Number of subsets of test cases and faults for the PDFSam configuration*

Figure 4.3 presents the boxplots obtained for each function. For the sets of test cases, *Jaro* and *JW* present the best stability in relation to the set of test cases, because the distance between each pair of test cases obtained by applying those distances is not equal generally. So, it is not necessary to frequently apply random selection. On the other hand, note that for the different sets of faults, *SF* is the most stable one, whereas *Lev* and *Sel* are the less stable. The reason is that *SF* is more precise in this context due to the presence of loops. It can more effectively detect the difference between a test case that is (contains) a subset of the other.

Second Empirical Study – TaRGeT Configuration

Considering the SSR metric and the specification of the real application, Specification 3 and Specification 10, we obtain ρ-*values* which are bigger than 0.05 by executing the Kruskal-Wallis test. For FC and Specification 3, we obtain ρ-*values* bigger than 0.05. Thus, not all null hypotheses can be rejected. In other words, for these specifications and metric, the distance functions have the same behavior with 95% confidence level. For the other cases, the ρ-*values* obtained are smaller than the significance level ($\alpha = 0.05$). Thus, all null hypotheses can be rejected (H_1^0 and H_2^0). So, with 95% confidence level, the distance functions can be considered different for SSR and FC.

Figure 4.4 shows the boxplots of the SSR and FC metrics considering the general average in the TaRGeT configuration. By observing the boxplots, we can see that behavior is only slightly different, generally making it impossible to rank the performance of the functions.

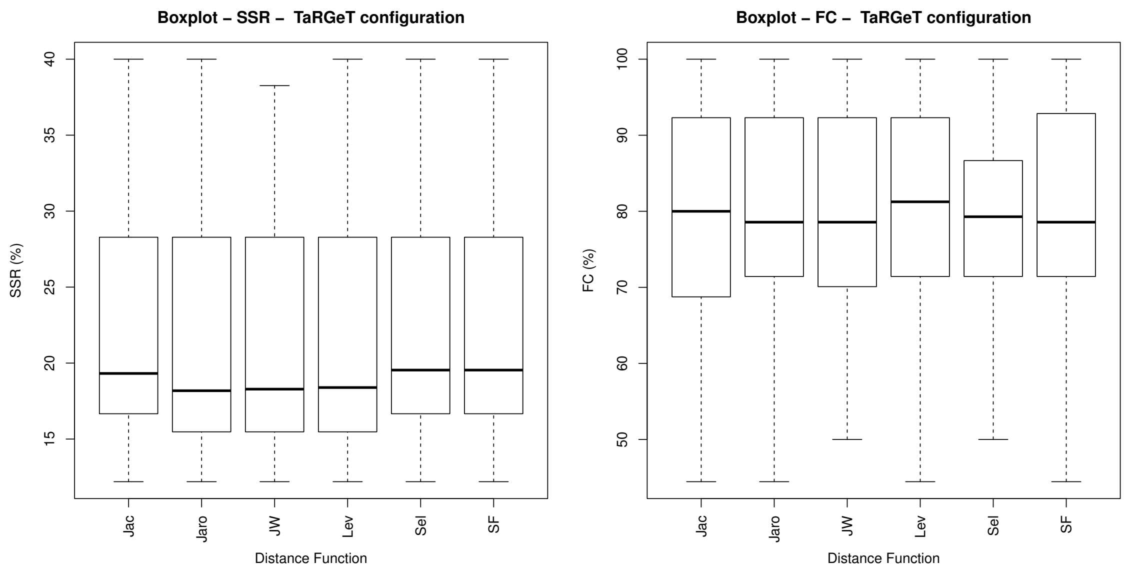


Figure 4.4: *Boxplots for SSR and FC considering the general average in TaRGeT configuration*

To uncover differences that might exist, we evaluate the pairs of distance functions by applying the Mann-Whitney tests and $\hat{A}_{12}$ effect size measurements (as defined in Section 4.2.2). Table 4.6 shows the Mann-Whitney U-tests and $\hat{A}_{12}$ effect size for each comparison considering the general average in the TaRGeT configuration.

Table 4.6: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for general average in TaRGeT configuration*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
Jac and Jaro	1.018e-26	Jac	Small (0.5239)	0.004913	Jac	Small (0.5163)
Jac and JW	1.731e-31	Jac	Small (0.5237)	0.006837	Jac	Small (0.5184)
Jac and Lev	6.685e-23	Jac	Small (0.5186)	0.2099	Jac	Small (0.5053)
Jac and Sel	0.1369	Sel	Small (0.4971)	0.00086	Jac	Small (0.5318)
Jac and SF	1.482e-09	SF	Small (0.4899)	0.3835	Jac	Small (0.5149)
Jaro and JW	0.004666	Jaro	Small (0.5015)	0.5968	Jaro	Small (0.5031)
Jaro and Lev	0.08457	Lev	Small (0.4948)	0.08643	Lev	Small (0.4901)
Jaro and Sel	2.462e-21	Sel	Small (0.4762)	0.3322	Jaro	Small (0.5186)
Jaro and SF	1.284e-29	SF	Small (0.4662)	0.4854	SF	Small (0.4944)
JW and Lev	3.907e-06	Lev	Small (0.494)	0.02606	Lev	Small (0.486)
JW and Sel	7.02e-27	Sel	Small (0.4761)	0.6267	JW	Small (0.515)
JW and SF	2.031e-32	SF	Small (0.4661)	0.2141	SF	Small (0.4928)
Lev and Sel	1.111e-17	Sel	Small (0.4812)	0.00253	Lev	Small (0.5308)
Lev and SF	5.281e-27	SF	Small (0.4715)	0.4833	Lev	Small (0.5098)
Sel and SF	9.641e-09	SF	Small (0.495)	0.04787	SF	Small (0.487)

As can be seen, for both metrics – SSR and FC – the effect size between the pairs of distance functions is considered small. This means that the behavior of one is better than the other one, even though the difference is small. Moreover, again Jac is prevalent for FC.

From the boxplots, Mann-Whitney tests and $\hat{A}_{12}$ effect size measurements, we can observe their performance, as presented in Table 4.7. In most cases, the performance of the functions can be considered similar. However, we can also note that, for both SSR and FC, *Lev* and *Sel* are the most closely related since either they present the same behavior or they are at subsequent levels of equality, except when the average is considered.

Finally, as in the first experiment, by analyzing the data obtained in the 40 executions of the technique when considering each function, we can also observe the stability of the

Table 4.7: *Ordering of effectiveness for SSR and FC in TaRGeT configuration*

	Suite Size Reduction (SSR)	Faults Coverage (FC)
real	$Jac = Jaro = JW = Lev = Sel = SF$	$Lev = Sel > Jac = Jaro = SF > JW$
01	$Jac = Lev = Sel > Jaro = SF > JW$	$JW > Jaro = Lev = Sel > Jac > SF$
02	$Jac = SF > Jaro = JW = Lev = Sel$	$Jaro = JW = Lev > Jac = Sel = SF$
03	$Jac = Jaro = JW = Lev = Sel = SF$	$SF > Jac = Jaro = JW = Lev = Sel$
04	$Jaro = JW = Lev = Sel = SF > Jac$	$Jaro > JW = Lev = Sel > Jac > SF$
05	$Sel = SF > Jac > Lev > Jaro = JW$	$Jac > Lev > Sel = SF > Jaro = JW$
06	$Jac = SF > Jaro = JW = Lev = Sel$	$Jac = Jaro = JW = Lev = SF > Sel$
07	$Jac = JW = Sel = SF > Jaro = Lev$	$Jac > Jaro = JW = SF > Lev = Sel$
08	$SF > Jac > Sel > Jaro = JW = Lev$	$SF > Jac = Jaro = JW = Lev > Sel$
09	$Jac = Sel = SF > Jaro = JW = Lev$	$SF > Jac > Jaro = JW = Lev = Sel$
10	$Jac = Jaro = JW = Lev = Sel = SF$	$Jac = Sel > Jaro = JW = Lev = SF$
All	$SF > Jac = Sel > Jaro = Lev > JW$	$Jac = Lev > Jaro = JW = SF > Sel$

reduction strategy by considering the same measures defined in Section 4.2.2. Figure 4.5 presents the boxplots obtained for each function. Note that, *SF* is the most stable one for both the different sets of test cases and faults. For the sets of faults, *Jac*, *Lev* and *Sel* are the less stable, even though differences here are less significant. The reason is that the TaRGeT configuration presents less redundancy.

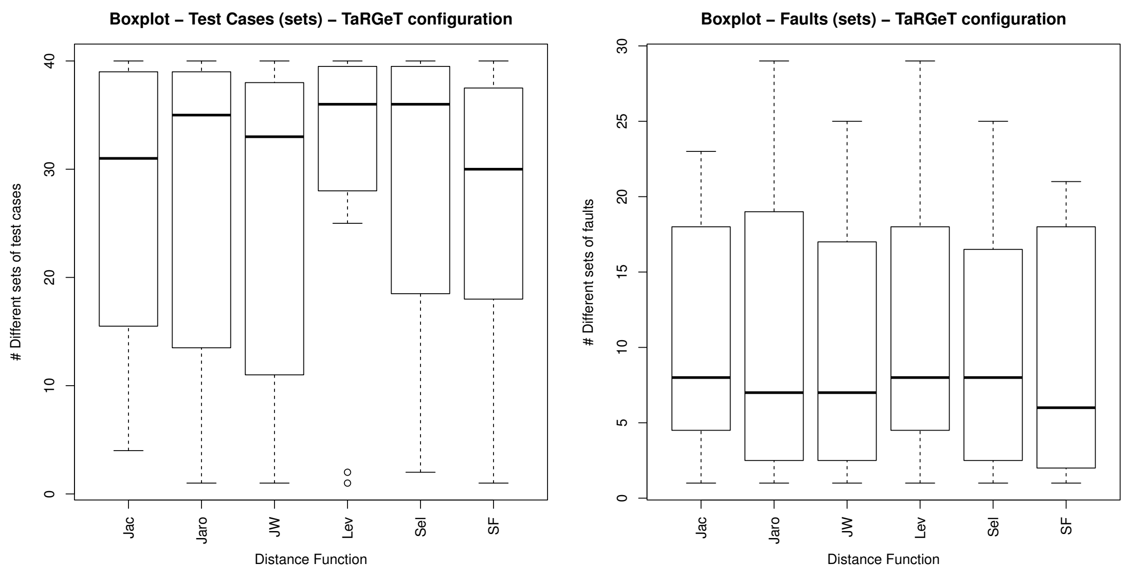


Figure 4.5: *Number of subsets of test cases and faults for the Target configuration*

General Remarks

In the presented experiments, we exercise and analyze distance functions in the context of a test suite reduction strategy. We consider two different scenarios by grouping specifications with a comparable configuration: *i*) in the PDFSam configuration group, reduction is more likely due to the presence of structures that may lead to higher degree of similarity between test cases; *ii*) in the TaRGeT configuration group, reduction is harder due to the prevalence of structures that do not directly lead to a higher degree of similarity, making the occurrence of essential test cases more likely.

It can be noticed that the configurations of the applications are different and the differences may impact directly on the results. For example, the number of *paths with loop* is a significant difference. This may have direct impact on the number of generated test cases and the degree of redundancy among test cases. As the PDFSam configuration has five *paths with loop*, then the generated test cases may contain a high degree of redundancy among them. Thus, we observe that the strategy presents a high rate of reduction. On the other hand, for the TaRGeT configuration, with no *paths with loop* and a big number of essential test cases, the reduction rate is low.

Results show that the PDFSam configuration presents differences that are more significant on performance between the functions since their influence on the overall result of the reduction technique is higher: the choice of the test case to be included depends on the function. However, we can conclude, for the investigated context, that the influence is mostly related to the FC metric rather than the SSR metric. Reduction percentage is quite similar in all cases, whereas fault coverage is more or less successful for different functions. *Jac* is in average the best function, particularly for the PDFSam configuration. This fact confirms a similar result presented by Hemmati et al. [Hemmati et al. 2013] in the context of test selection, where *Jac* and two of its variants are the distance functions with best performance for FC.

Regarding stability, the results obtained indicate that the average stability of the number of different sets of faults is usually related to better fault coverage. For instance, consider the results obtained by the *Jac* function. This may indicate that less precision can make the function more effective to cover different faults. Moreover, note that, in the PDFSam configuration, there are cases where the *SF* function, the more stable one, detected 0 faults.

Furthermore, there is a limit to stability: the less stable functions, *Sel* and *Lev*, cannot supersede *Jac* in general.

4.3 Case Study

The goal of this case study is to provide further investigation into the performance of the distance functions in a different context from the two experiments discussed so far. The study is based on an industrial application developed in the context of a cooperation between our research laboratory and Ingenico[5]. The application is a software for collecting and processing biometrics. From use cases, LTS specification models are automatically generated for two subsequent versions of the application, where one is a baseline version – CB_{v_1} – and the other is a delta version – CB_{v_2} – obtained from CB_{v_1} by two progressive modifications. From the models, we generate two test suites and manually executed them. From execution, we collect faults and failures. Table 4.8 describes the configurations of the two specification models.

Table 4.8: *The configurations of the real-world specifications*

	CB_{v_1}	CB_{v_2}
Depth	19	19
Paths with loop	16	15
Forks	26	25
Transitions of forks	63	60
Joins	10	8
Transitions of joins	25	21
Test cases (one expansion)	69	66
Essential test cases	15	17
Faults	10	6
Failures	12	7

Note that the rates of fault of the specifications based on size of the generated test suites are 14.49% for CB_{v_1} and 9.09% for CB_{v_2}. The number of essential test cases that fail for CB_{v_1} and CB_{v_2} are 2 and 3, respectively. All essential test cases that fail are associated to a

[5] www.ingenico.com

distinct fault. We expect that they are always included in the reduced suite, as they uniquely cover a requirement by definition.

For each specification, we execute 1,000 replications for each distance function. In order to draw observations based on these data, we apply a statistical analysis similar to that used in the other empirical studies.

Figure 4.6 presents the boxplots considering SSR and FC. Note that there are many overlaps; then, it is necessary to perform the Mann-Whitney test.

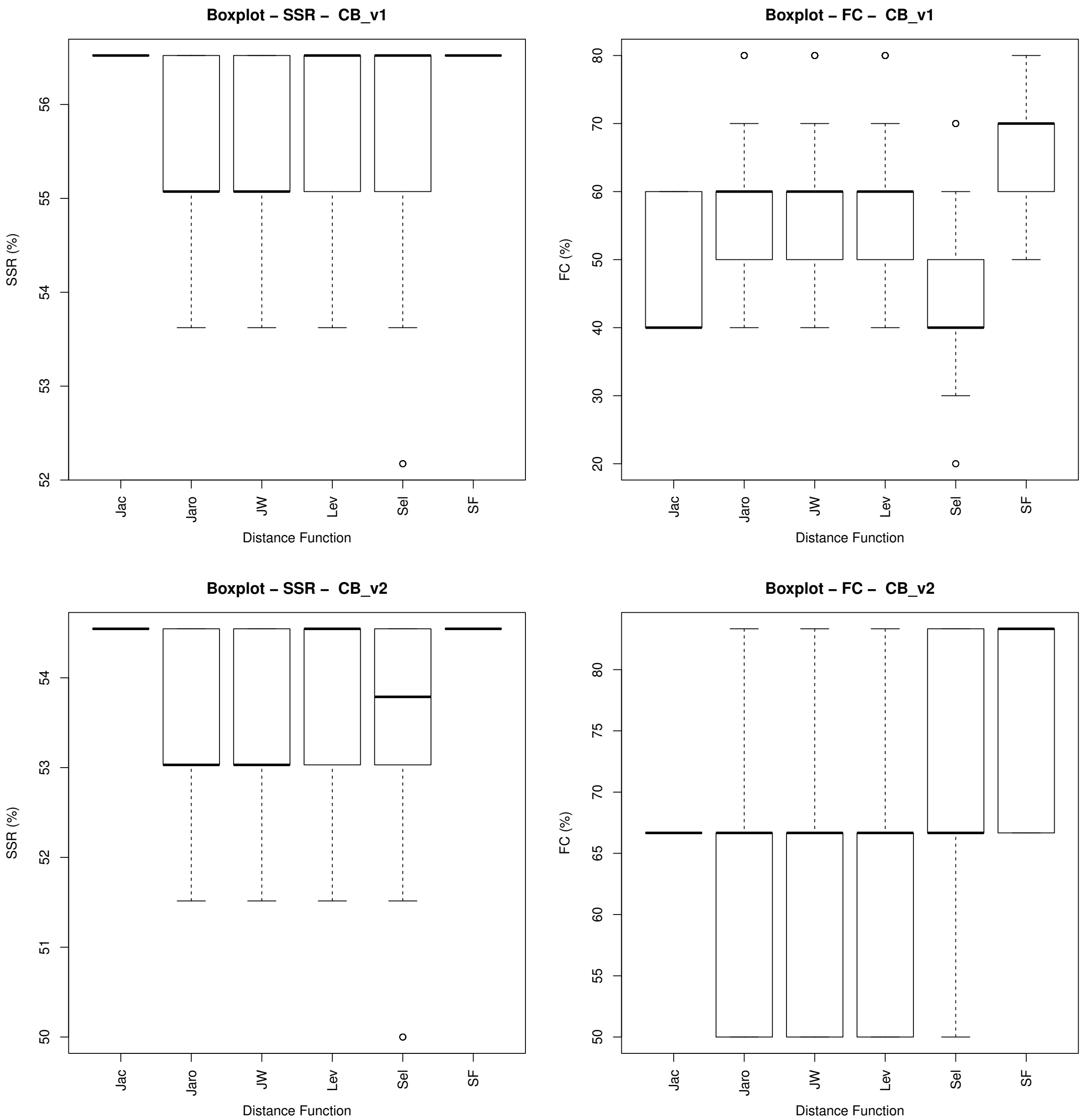


Figure 4.6: *Boxplots for SSR and FC considering the general average for CB_{v_1} and CB_{v_2}*

In order to clarify the magnitude of the difference between the distance functions, we perform the $\hat{A}_{12}$ effect size. The results of the Mann-Whitney U-tests and $\hat{A}_{12}$ effect size measurement for each distance function comparison is reported in Table 4.9 for CB$_{v_1}$, and in Table 4.10 for CB$_{v_2}$. Considering SSR for CB$_{v_1}$ and CB$_{v_2}$, we can see that Jac and SF present that best behavior, and there is no difference between them, whereas considering FC, the difference between them is large and SF is better.

Table 4.9: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements when SSR and FC across the distance functions for CB$_{v_1}$*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
Jac and Jaro	5.343e-115	Jac	Large (0.7865)	7.14e-93	Jaro	Large (0.2739)
Jac and JW	7.597e-111	Jac	Large (0.778)	7.14e-93	JW	Large (0.2616)
Jac and Lev	2.142e-86	Jac	Large (0.7055)	7.14e-93	Lev	Large (0.2763)
Jac and Sel	7.14e-93	Jac	Large (0.7455)	7.14e-93	Jac	Medium (0.6004)
Jac and SF	NaN	None	NO effect (0.5)	7.14e-93	SF	Large (0.1105)
Jaro and JW	0.7516	JW	Small (0.4932)	7.14e-93	JW	Small (0.4865)
Jaro and Lev	8.714e-14	Lev	Small (0.4127)	7.14e-93	Jaro	Small (0.503)
Jaro and Sel	0.8354	Sel	Small (0.4746)	7.14e-93	Jaro	Large (0.8236)
Jaro and SF	5.343e-115	SF	Large (0.2135)	7.14e-93	SF	Large (0.2658)
JW and Lev	9.654e-13	Lev	Small (0.4201)	7.14e-93	JW	Small (0.5164)
JW and Sel	0.6551	Sel	Small (0.4812)	7.14e-93	JW	Large (0.834)
JW and SF	7.597e-111	SF	Large (0.222)	7.14e-93	SF	Large (0.2773)
Lev and Sel	8.338e-12	Lev	Small (0.5565)	7.14e-93	Lev	Large (0.82)
Lev and SF	2.142e-86	SF	Large (0.2945)	7.14e-93	SF	Large (0.2663)
Sel and SF	7.14e-93	SF	Large (0.2545)	7.14e-93	SF	Large (0.0538)

From the boxplots, Mann-Whitney tests and $\hat{A}_{12}$ effect size measurements, we obtain the ordering of effectiveness for SSR and FC behavior in Table 4.11.

The fact that the choice of the distance function may influence on fault coverage follows the results obtained in the previous experiments to a certain extent. However, *Jac* did not performed as good as in the experiments regarding FC. By closely analysing the reduced suite, we can see that the measurement made by *Jac* made some failing test cases to be discarded, because each of them was considered similar to another that was selected.

As mentioned before, the distance function may influence on the order pairs of test cases are considered. Particularly, for the CB application, *SF* is more successful when comparing

Table 4.10: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements when SSR and FC across the distance functions for CB$_{v_2}$*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
Jac and Jaro	2.078e-114	Jac	Large (0.787)	1.228e-94	Jac	Medium (0.623)
Jac and JW	1.151e-117	Jac	Large (0.7965)	1.228e-94	Jac	Medium (0.6115)
Jac and Lev	5.652e-87	Jac	Large (0.712)	1.228e-94	Jac	Medium (0.621)
Jac and Sel	1.228e-94	Jac	Large (0.75)	1.228e-94	Sel	Small (0.488)
Jac and SF	NaN	None	NO effect (0.5)	1.228e-94	SF	Large (0.2435)
Jaro and JW	0.3932	Jaro	Small (0.5105)	1.228e-94	JW	Small (0.4906)
Jaro and Lev	3.542e-11	Lev	Small (0.421)	1.228e-94	Lev	Small (0.4959)
Jaro and Sel	0.01952	Sel	Small (0.4771)	1.228e-94	Sel	Medium (0.3994)
Jaro and SF	2.078e-114	SF	Large (0.213)	1.228e-94	SF	Large (0.2175)
JW and Lev	8.753e-15	Lev	Small (0.4106)	1.228e-94	JW	Small (0.5055)
JW and Sel	0.003858	Sel	Small (0.4672)	1.228e-94	Sel	Small (0.4079)
JW and SF	1.151e-117	SF	Large (0.2035)	1.228e-94	SF	Large (0.2243)
Lev and Sel	1.73e-06	Lev	Small (0.5519)	1.228e-94	Sel	Small (0.4007)
Lev and SF	5.652e-87	SF	Large (0.288)	1.228e-94	SF	Large (0.2136)
Sel and SF	1.228e-94	SF	Large (0.25)	1.228e-94	SF	Large (0.3178)

Table 4.11: *Ordering of effectiveness for SSR and FC in CB$_{v_1}$ and CB$_{v_2}$*

	Suite Size Reduction (SSR)	Fault Coverage (FC)
CB$_{v_1}$	$Jac > SF > Lev > Jaro = JW = Sel$	$SF > Jaro = JW = Lev > Jac > Sel$
CB$_{v_2}$	$Jac = SF > Lev > Sel > Jaro = JW$	$SF > Jac > Sel > Jaro = JW = Lev$

the total number of distinct faults and the frequency in which that faults are detected. However, *Jac* presented the more stable behavior, that is, less variance in reduced suite for 1,000 executions when considering the subset of test cases that fail, followed by *SF* (Table 4.12). This confirms the results obtained in the experiments: the function with best stability may not be the one with best performance for fault coverage. With the presence of essential test cases, *SF* becomes less stable than when applied in the PDFSam configuration.

For both specifications, we observed that *Lev* and *Sel* have a bigger variation in the sets of test cases that make up the reduced suite, when compared to *Jac* and *SF*. Moreover, in general, the number of faults detected at least once is greater than by other functions, that is,

Table 4.12: *Number of different sets of test cases selected, number of distinct test cases, average frequency of inclusion of a test case in the reduced suite, number of different sets of faults detected, number of distinct faults and average frequency of inclusion of a fault detected by a reduced suite*

	Metric	*Jac*	*Jaro*	*JW*	*Lev*	*Sel*	*SF*
CB_{v_1}	#Different Sets of Test Cases	20	80	80	421	553	48
	#Distinct Test Cases	38	43	43	51	47	39
	%Test Case Frequency	78.94	71.25	71.22	59.68	65.12	76.92
	#Different Sets of Faults	3	12	12	16	40	8
	#Distinct Faults	6	8	8	8	8	8
	%Fault Frequency	81.18	71.70	72.28	71.61	55.86	81.38
CB_{v_2}	#Different Sets of Test Cases	10	40	40	321	410	24
	#Distinct Test Cases	36	41	41	49	45	37
	%Test Case Frequency	83.33	74.74	74.81	62.17	68.04	81.08
	#Different Sets of Faults	1	4	4	5	6	2
	#Distinct Faults	4	5	5	6	6	5
	%Fault Frequency	100	75.08	75.54	75.16	80.48	90.26

they may eventually achieve a much better FC. However, on average, the number of covered faults by each reduced test suite is small, making them less reliable.The variation is due to the large number of draws among similarity degrees in the matrix, making it possible for test cases that fail not selected by the *SF* and *Jac* reduction, to be selected as a result of a random choice. As in the experiments, *Lev* and *Sel* present a comparable behavior.

4.4 Concluding Remarks

In this chapter, we presented three empirical studies with the goal of comparing distance functions when applied to a strategy of test suite reduction based on similarity in the context of MBT. These studies provide evidence on the impacts that the choice of a function can have on the performance of our reduction strategy regarding suite size reduction, fault coverage, and stability. In turn, results show that the choice of the distance function has little influence on suite size reduction, but it can more significantly influence fault coverage and stability. The reason is that each distance function leads to selection of a different suite and it is possible to have significant variations on this selection since reduction strategy often faces draws and handles them by random selection. To provide further evidence and deeper observation, we conduct a case study in the scope of a real-world application under development that has a different configuration from the ones previously considered in the experiment. The results from this study are comparable to the ones obtained in the experiment regarding the effect produced by the functions on suite size reduction, fault coverage and stability as well as on the pattern of related behavior of some functions (*Lev* and *Sel*). Additionally, in the case study, we can also observe stability of the number of different sets of faults and fault frequency of the reduction strategy when considering different functions.

Even though no definite conclusions can be reached, as the context of the experiments and case study are specific, for the model configurations investigated, the *SF* function promotes the best stability, followed by *Jac*, *Jaro* and *JW*. On the other hand, *Lev* and *Sel* present a relatively lower stability. Moreover, *Jac* often presents the best performance by optimizing the relationship between stability and fault coverage.

It is important to highlight that the number of paths with loops and the number of essential test cases in the each specification have also impact on the results of the reduction technique. When the number of paths with loops is high, probably the degree of redundancy in the test suite is high. Therefore, the reduction strategy can be more effective w.r.t. to size and consequently less effective w.r.t. fault coverage. When the number of essential test cases is high, observations are the opposite. Nevertheless, this is a behavior expected from the strategy of reduction based on similarity, as the average changes of rate are relatively similar when considering all functions.

Another interesting issue is the difference in the similarity degree for a given pair of test cases provided by the different distance functions. The differences have direct influence on the order in which the strategy evaluates pairs of test cases by considering the set of test cases that make up the reduced test suite. This might explain why a given test case is never part of the reduced test suites for a distance function, but it is always for another function.

Chapter 5

Evaluation of the Similarity-based Test Suite Reduction Strategy

This chapter presents six empirical studies to evaluate our strategy presented in Chapter 3. In these empirical studies, we compare our strategy with other four well-known test suite reduction heuristics that can be applied in the MBT context by using different transition-based coverage criteria. For this, we used 3 real-world specification models with real faults, and three sets of 30 synthetic specification models automatically generated from the configuration of each real-world application with sets of faults defined according to the fault model from each correspondent real-world specification. The reduction heuristics investigated are G, GE, GRE and HGS presented in Section 2.5.1, and our reduction strategy presented in Chapter 3. In these studies, we compare the effectiveness of the reduction strategies by application of the following coverage criteria *all-transitions*, *all-transition-pairs* and *bi-criteria*, regarding suite size reduction and fault coverage. Moreover, we also observe the *scattering* of the reduction strategies for 100% fault coverage when considering different coverage criteria. Although related works show very promising strategies, we chose to investigate the heuristics G, GE, GRE and HGS, because in general context they have the best behavior for reduced test suite size and fault coverage. Also, in this investigation, we do not consider *Dissimilarity* strategy proposed by Cartaxo [Cartaxo 2011] since in preliminary studies[1] [Coutinho 2011; Coutinho 2012a; Coutinho 2012b; Coutinho 2012c; Coutinho 2013] this strategy does not presented the good results for reduction size and fault

[1] http://splab.computacao.ufcg.edu.br/publications/technical-reports

coverage when compared to the heuristics and our reduction strategy.

5.1 Experiment Definition

From the six empirical studies, three are focused on three real-world application models and real faults detected during test execution. The other three are based on the average of a set of synthetic specification models. In this sense, for each real-world application model, 30 synthetic specification models are automatically generated based on structural measures of the real-world model and the set of faults are defined according to the obtained percentage of failures and faults from each correspondent real-world specification. From the relation between the number of failures and faults, a fault model is generated based on the identification of cliques of test cases that are likely to uncover the same potential fault. Thus, we have three empirical studies, one for each of the real-world application models, and three empirical studies for each set of synthetic specification models. Basically, the studies follow the same definition and planning, but they are run and analysed separately.

5.1.1 Definition

The goal of these empirical studies is to investigate *test suite reduction strategies* considering different *coverage criteria*, observing *reduced test suite size and fault coverage*. Based on this goal, our general hypothesis is that:

> *"Test Suite Reduction Strategies based on different coverage criteria show a different performance for the measures size and fault coverage of the reduced test suite"*.

Furthermore, the results are analyzed from the point of view of the tester (responsible for the testing process) in the context of MBT.

5.1.2 Planning

In the phase of planning, we plan how the experiment should be conducted according to six steps defined by Wohlin et al. [Wohlin et al. 2012]: context selection, variables, hypothesis, instrumentation, design and threats to validity as follows.

Context Selection

According to the dimensions proposed by Wohlin et al. [Wohlin et al. 2012], these empirical studies are *off-line* since we perform them in laboratory, i.e., not in an industrial environment.

For each empirical study, the inputs are three real-world specifications (*real problems*) and three sets of 30 synthetic automatically generated specifications (one set for each real-world application). These synthetic specifications are randomly generated by considering the same configuration of the respective real-world specification. Therefore, as these empirical studies focus only on three sets of different configurations, those studies can be characterized as *specific*. Furthermore, these studies did not involve human interaction.

Variables Selection

The definition of the variables characterize the experiment through the elements that are observed (dependent variables), and modified and controlled (independent variables). For these studies, the variables are defined as follows.

- *Independent variables:*
  - *Coverage criteria* (*test requirements* – see Section 2.4):
    - *all-transitions* (*T*);
    - *all-transition-pairs* (*P*);
    - *bi-criteria* (*B*): *all-transitions and all-transition-pairs*;
  - *Test suite reduction strategy*:
    - *Greedy Heuristic* (*G*);
    - *Heuristic Greedy-Essential* (*GE*);
    - *Heuristic Greedy* – $1 - to - 1$ – *Redundancy Essential* (*GRE*);
    - *Heuristic of the Harrold Gupta Soffa* (*HGS*);
    - *Similarity-Based Test Suite Reduction Strategy* (*Sim*);
  - *Faults:* the faults revealed by the test suite;

- *Dependent variables:*
  - *Suite Size Reduction (SSR)*: percentage of the number of test cases removed from the complete test suite.

    $$SSR = \frac{|TS| - |RS|}{|TS|} \times 100\%$$

    where $|TS|$ is the number of test cases in the complete test suite and $|RS|$ is the number of test cases in the reduced test suite;
  - *Fault Coverage (FC)*: percentage of the total number of faults uncovered by the reduced test suite:

    $$FC = \frac{|F_{RS}|}{|F_{TS}|} \times 100\%$$

    where $|F_{TS}|$ is the number of faults revealed by the complete test suite and $|F_{RS}|$ is the number of faults revealed by the reduced test suite.

To apply Sim, a similarity function and a choice function are requested. In these studies, our similarity function presented in Section 3.2 is used.

Regarding the choice function, to define the order of analysis between test cases, we opted for the function based on the number of transitions. The key idea is to compare the size of the test cases and to keep in the matrix that one more transitions, since it can represent the highest functionality coverage. If the lengths are the same, the analysis of order of the test cases is randomly chosen.

Hypothesis Formulation

Based on the response variables, we formulated the following study questions for these empirical studies:

- **SQ1:** For each reduction strategy, which coverage criterion is more effective in terms of SSR and FC?
- **SQ2:** For each coverage criterion, which reduction strategy is more effective in terms of SSR and FC?
- **SQ3:** When used in combination with their best coverage criterion from SQ1, which reduction strategy is more effective in terms of SSR and FC?

- **SQ4:** When used in combination with their best reduction strategy from SQ2, which coverage criterion is more effective in terms of SSR and FC?

Based on the study question SQ1, we define null and alternative hypotheses for each empirical study, as follows[2].

- *SSR (Table 5.1 (a)):*
  - *A null hypothesis (H_1^0, H_2^0, H_3^0, H_4^0, H_5^0):* for each reduction strategy, there are no differences among coverage criterion regarding suite size reduction;
  - *An alternative hypothesis (H_1^1, H_2^1, H_3^1, H_4^1, H_5^1):* for each reduction strategy, all coverage criteria have a different behavior regarding suite size reduction.
- *FC (Table 5.1 (b)):*
  - *A null hypothesis (H_6^0, H_7^0, H_8^0, H_9^0, H_{10}^0):* for each reduction strategy, there are no differences among coverage criterion regarding the rate of fault coverage;
  - *An alternative hypothesis (H_6^1, H_7^1, H_8^1, H_9^1, H_{10}^1):* for each reduction strategy, all coverage criteria have a different behavior regarding the rate of fault coverage.

Table 5.1: *Null and alternative hypotheses considering SQ1*

(a) SSR	(b) FC
$H_1^0 : SSR_{G_T} = SSR_{G_P} = SSR_{G_B}$	$H_6^0 : FC_{G_T} = FC_{G_P} = FC_{G_B}$
$H_1^1 : SSR_{G_T} \neq SSR_{G_P} \neq SSR_{G_B}$	$H_6^1 : FC_{G_T} \neq FC_{G_P} \neq FC_{G_B}$
$H_2^0 : SSR_{GE_T} = SSR_{GE_P} = SSR_{GE_B}$	$H_7^0 : FC_{GE_T} = FC_{GE_P} = FC_{GE_B}$
$H_2^1 : SSR_{GE_T} \neq SSR_{GE_P} \neq SSR_{GE_B}$	$H_7^1 : FC_{GE_T} \neq FC_{GE_P} \neq FC_{GE_B}$
$H_3^0 : SSR_{GRE_T} = SSR_{GRE_P} = SSR_{GRE_B}$	$H_8^0 : FC_{GRE_T} = FC_{GRE_P} = FC_{GRE_B}$
$H_3^1 : SSR_{GRE_T} \neq SSR_{GRE_P} \neq SSR_{GRE_B}$	$H_8^1 : FC_{GRE_T} \neq FC_{GRE_P} \neq FC_{GRE_B}$
$H_4^0 : SSR_{HGS_T} = SSR_{HGS_P} = SSR_{HGS_B}$	$H_9^0 : FC_{HGS_T} = FC_{HGS_P} = FC_{HGS_B}$
$H_4^1 : SSR_{HGS_T} \neq SSR_{HGS_P} \neq SSR_{HGS_B}$	$H_9^1 : FC_{HGS_T} \neq FC_{HGS_P} \neq FC_{HGS_B}$
$H_5^0 : SSR_{Sim_T} = SSR_{Sim_P} = SSR_{Sim_B}$	$H_{10}^0 : FC_{Sim_T} = FC_{Sim_P} = FC_{Sim_B}$
$H_5^1 : SSR_{Sim_T} \neq SSR_{Sim_P} \neq SSR_{Sim_B}$	$H_{10}^1 : FC_{Sim_T} \neq FC_{Sim_P} \neq FC_{Sim_B}$

[2]Note that equations expressed as $a = b = c$, represent $a = b \wedge b = c \wedge a = c$, and $a \neq b \neq c$, represent $a \neq b \vee b \neq c \vee a \neq c$.

For SQ2, the null and alternative hypotheses investigated are:

- *SSR (Table 5.2 (a)):*
  - *A null hypothesis (H^0_{11}, H^0_{12}, H^0_{13}):* for each coverage criterion, there are no differences among test suite reduction strategies regarding suite size reduction;
  - *An alternative hypothesis (H^1_{11}, H^1_{12}, H^1_{13}):* for each coverage criterion, all test suite reduction strategies have a different behavior regarding suite size reduction.
- *FC (Table 5.2 (b)):*
  - *A null hypothesis (H^0_{14}, H^0_{15}, H^0_{16}):* for each coverage criterion, there are no differences among test suite reduction strategies regarding the rate of fault coverage;
  - *An alternative hypothesis (H^1_{14}, H^1_{15}, H^1_{16}):* for each coverage criterion, all test suite reduction strategies have a different behavior regarding the rate of fault coverage.

Table 5.2: *Null and alternative hypotheses considering SQ2*

(a) SSR

$H^0_{11} : SSR_{G_T} = SSR_{GE_T} = SSR_{GRE_T} = SSR_{HGS_T} = SSR_{Sim_T}$
$H^1_{11} : SSR_{G_T} \neq SSR_{GE_T} \neq SSR_{GRE_T} \neq SSR_{HGS_T} \neq SSR_{Sim_T}$
$H^0_{12} : SSR_{G_P} = SSR_{GE_P} = SSR_{GRE_P} = SSR_{HGS_P} = SSR_{Sim_P}$
$H^1_{12} : SSR_{G_P} \neq SSR_{GE_P} \neq SSR_{GRE_P} \neq SSR_{HGS_P} \neq SSR_{Sim_P}$
$H^0_{13} : SSR_{G_B} = SSR_{GE_B} = SSR_{GRE_B} = SSR_{HGS_B} = SSR_{Sim_B}$
$H^1_{13} : SSR_{G_B} \neq SSR_{GE_B} \neq SSR_{GRE_B} \neq SSR_{HGS_B} \neq SSR_{Sim_B}$

(b) FC

$H^0_{14} : FC_{G_T} = FC_{GE_T} = FC_{GRE_T} = FC_{HGS_T} = FC_{Sim_T}$
$H^1_{14} : FC_{G_T} \neq FC_{GE_T} \neq FC_{GRE_T} \neq FC_{HGS_T} \neq FC_{Sim_T}$
$H^0_{15} : FC_{G_P} = FC_{GE_P} = FC_{GRE_P} = FC_{HGS_P} = FC_{Sim_P}$
$H^1_{15} : FC_{G_P} \neq FC_{GE_P} \neq FC_{GRE_P} \neq FC_{HGS_P} \neq FC_{Sim_P}$
$H^0_{16} : FC_{G_B} = FC_{GE_B} = FC_{GRE_B} = FC_{HGS_B} = FC_{Sim_B}$
$H^1_{16} : FC_{G_B} \neq FC_{GE_B} \neq FC_{GRE_B} \neq FC_{HGS_B} \neq FC_{Sim_B}$

From the answers of SQ1 and SQ2, we will define null and alternative hypotheses for each empirical study for SQ3 and SQ4. For SQ3, we are interested in comparing the five reduction strategies considering the best coverage criterion from SQ1 in relation to SSR and FC. Regarding SQ4, we will compare the best reduction strategy for each coverage criterion from SQ2 in terms of SSR and FC. These null and alternative hypotheses for SQ3 and SQ4 will be presented, respectively, in Sections 5.2.3 and 5.2.4.

Instrumentation

The instruments for these empirical studies are defined as follows.

1. *Objects:* 3 real-world and 90 synthetic automatically generated *Labelled Transition System* (LTS) specifications (30 synthetic specifications for each real-world specification models);
2. *Guidelines:* since the reduction strategies are automatic, and do not require people to configure them, no guideline is used;
3. *Measurements:* the LTS-BT tool [Cartaxo et al. 2008] is used to support the experiments execution and data collection.

The real-world LTS specifications are obtained by the TaRGeT tool from use cases written by experienced testers of three real-world applications. LTS is a common formalism used by research on MBT [Tretmans 2008; Anand et al. 2013]. These applications are briefly described as follows:

- *CB*: an industrial application for collecting and processing biometrics;
- *PDFSam:* an open source tool used to split and merge pdf documents;
- *TaRGeT:* an application that generates test cases from use case documents in a MBT process.

From the configuration of each real LTS specifications (*structural measures*) presented in Section 2.2.2, 30 synthetic LTS specifications are automatically generated based on the strategy presented by Oliveira et al. [Oliveira Neto et al. 2013], as illustrated in Figure 5.1.

It is important to remark that the version of TaRGeT we consider as object of the study is different from the one we use for generating the models.

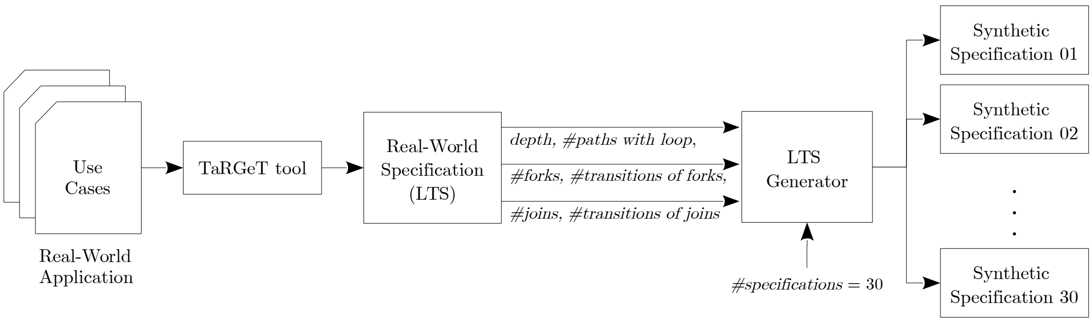


Figure 5.1: *Generation process of the synthetic specifications*

Table 5.3: *Basic configuration of the three real-world specifications*

	CB	PDFSam	TaRGeT
Depth	19	18	8
Paths with loop	16	5	0
Forks	26	15	26
Transitions of forks	63	41	101
Joins	10	11	16
Transitions of joins	25	34	42

Table 5.3 presents the configuration for each real-world LTS specification.

For PDFSam and TaRGeT synthetic models, we randomly selected a number of test cases that fail and associated each failure with a fault according to the percentage of faults of the real applications (PDFSam: 3.65% and TaRGeT: 15.85%).

For CB synthetic models, we generate the faults from the automatic fault model generation for each synthetic LTS specification, as illustrated in Figure 5.2.

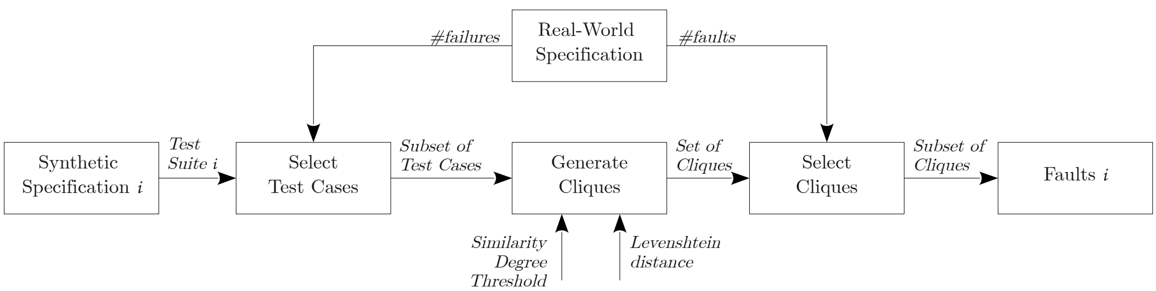


Figure 5.2: *Scheme to generate faults for each synthetic specification input*

The faults are defined from a subset of test cases of the complete test suite automatically generated (step 01). This subset of test cases is randomly selected according to the percentage of failures of 17.39% for CB real-world (step 02). Next, a graph G is created based on similarity degree threshold, where the vertices are test cases and the edges represent the cases where the similarity degree between two test cases is above the threshold – meaning that test cases may have similar capability of fault detection. This threshold is a similarity value indicating which test cases (among previously selected) may reveal the same (or different) fault. Thus, we defined the threshold as 75% of the largest similarity value between pairs of test cases of the complete test suite (TS) from the use of Levenshtein distance (presented in Section 2.6.4) as similarity measure based on threshold of the real specification model (step 03).

From the graph G, all possible cliques are identified, i.e., a subset of test cases that are likely to uncover the same potential fault (step 04). The result is a subset of cliques randomly chosen from the percentage of faults of 14.49% for CB real-world specification (step 05).

Table 5.4 presents the number of failures and faults detected by each real-world LTS specification. Appendix A.1 presents the number of test cases generated, essential test cases for each coverage criterion, and number of faults and failures generated for each synthetic model.

Table 5.4: *Number of failures and faults of the three real-world specifications*

	CB	PDFSam	TaRGeT
Failures	12	5	13
Faults	10	5	13

Experimental Design

The experimental design determines the number of experiments, the factor levels (treatments) combinations for each experiment, and the number of replications [Jain 1991][Wohlin et al. 2012]. In these empirical studies, for each coverage criterion there is one factor (*test suite reduction strategy*) with more than two treatments for each specification (5 treatments: G, GE, GRE, HGS and Sim).

Overall, we replicate each experiment 1,000 times as suggested by Arcuri and Briand [Arcuri and Briand 2011], leading to a total of 15,000 observations for each object (specification).

Thus, the total number of observations is 1,395,000 for all empirical studies (93 objects – 3 real-world and 90 synthetic specifications). The response variables are the metrics observed: SSR and FC. Figure 5.3 presents an overview of the experiment.

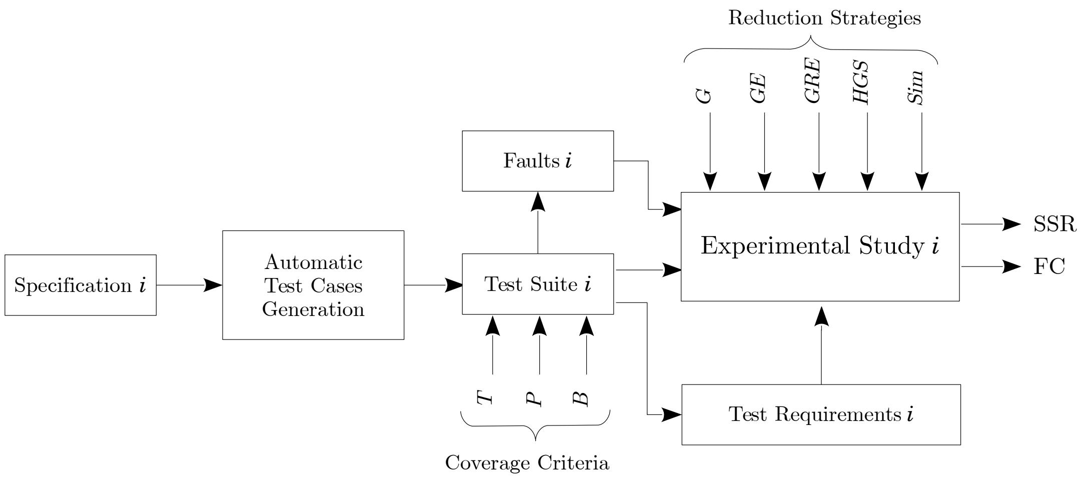


Figure 5.3: *Overview of the experiment for each input specification*

5.1.3 Operation

To automatically generate the different synthetic specifications from specific configurations, we implemented an LTS generator as proposed by Oliveira et al. [Oliveira Neto et al. 2013]. Furthermore, it was necessary to extend the heuristic algorithms to use an additional coverage criterion (*bi-criteria*), based on the idea of our strategy presented in Chapter 3. We implement these in Java programming language[3]. Following this, we used the LTS-BT tool to generate test cases.

Furthermore, we perform each step of these empirical studies, using a machine with Intel Core(TM) i5 3.10 GHz, 8GB RAM running GNU Linux.

[3]http://www.sun.com/java/

5.1.4 Threats to Validity

Aiming toward a good *conclusion validity*, on each strategy for each specification in the empirical studies is executed $1{,}000$ times. According to Arcuri and Briand [Arcuri and Briand 2011], for samples of size at least $1{,}000$, there are no practical difference between them regarding power and accuracy. Furthermore, all analysis consider the confidence level of 95%, as suggested by Jain [Jain 1991] for conducting experiments.

To make the control of the empirical studies and, consequently, the *internal validity*, there are no people involved, and the same inputs (objects) are applied for all the experiments. Thus, this internal validity is not considered critical.

The main threat is *construct validity* of our empirical studies. In order for the experimenter not to influence the measures, the synthetic specifications and its faults are automatically generated from the configuration of real-world specifications. Furthermore, we implemented the metrics (SSR and FC) and heuristics according to concepts proposed in literature. Another threat to validity is the implementation of the heuristics for multi-criteria. To deal with this, it is necessary to adapt the heuristics to apply the multi-criteria similiar to our strategy.

Regarding *external validity*, the synthetic specifications are the main threat. However, these specifications are automatically generated in a controlled way considering the same configuration of real specification models.

5.2 Experiment Analysis

After experiment definition of the empirical studies, we executed the experiments and collected the data for analysis. As suggested by Wohlin et al. [Wohlin et al. 2012], we check if the data collected have a normal distribution for all reduction strategies for each empirical study, considering the SSR and FC metrics. For this, we apply the Anderson-Darling normality test, using the R tool[4], considering the confidence level at 95% (significance level is $\alpha = 0.05$) [Jain 1991]. The results of the statistical tests applied are presented in Appendix A. In this Appendix, by applying the normality test in all empirical studies considering all study questions, we observe that the *ρ-values* are smaller than the significance value ($\alpha = 0.05$).

[4] http://www.r-project.org/

Thus, we need to apply nonparametric statistical tests.

Since each experimental design has a unique factor with more than two treatments, we apply the nonparametric Kruskal-Wallis test to check the null hypothesis. This test is used to determine if there are "*significant*" differences in treatments across multiple test attempts. For all empirical studies, we obtain $\rho - value < 0.05$ for SSR and FC metrics for all study questions (see Appendix A.3). Thus, all null hypotheses can be rejected, that is, for SSR and FC.

In the next subsections, we present and discuss the results for each study question, considering each empirical study.

5.2.1 Study Question 1 (SQ1)

To view the distribution of data, we generate boxplots for each empirical study considering SSR and FC metrics.

By looking at the boxplots for SSR in Figure 5.4, it is possible to compare the size of the reduced test suites by each reduction strategy and coverage criteria for each empirical study. The results of empirical studies suggest that T as coverage criterion can dramatically reduce the sizes of test suites for all reduction strategies.

On the other hand, the effectiveness of FC presented in Figure 5.5, is significantly harmed. Furthermore, the best reduction strategies and coverage criteria for SSR are the worst for FC, and vice versa.

CB and *PDFSam* (*real and synthetics*) have a high rate of reduction. The most plausible reason is due to the number of *paths with loop*, which can influence the number of generated test cases and the degree of redundancy among test cases. On the other hand, for *TaRGeT* (*real and synthetics*) we observe that the reduction rate is low due to the structural characteristics of the specification model with no *paths with loop* that do not lead to a higher degree of similarity, hence the occurrence of essential test cases is more likely.

As there are overlaps in the boxplots, we apply the Mann-Whiney test (Wilcoxon-Mann-Whitney test in R) to compare each pair of overlap. However, the Mann-Whitney test shows that a statistically significant difference exists between two treatments, but not the magnitude of this difference. Thus, we use the $\hat{A}_{12}$ effect size measure proposed by Vargha and Delaney [Vargha and Delaney 2000] to assess the differences between each pair of treatment

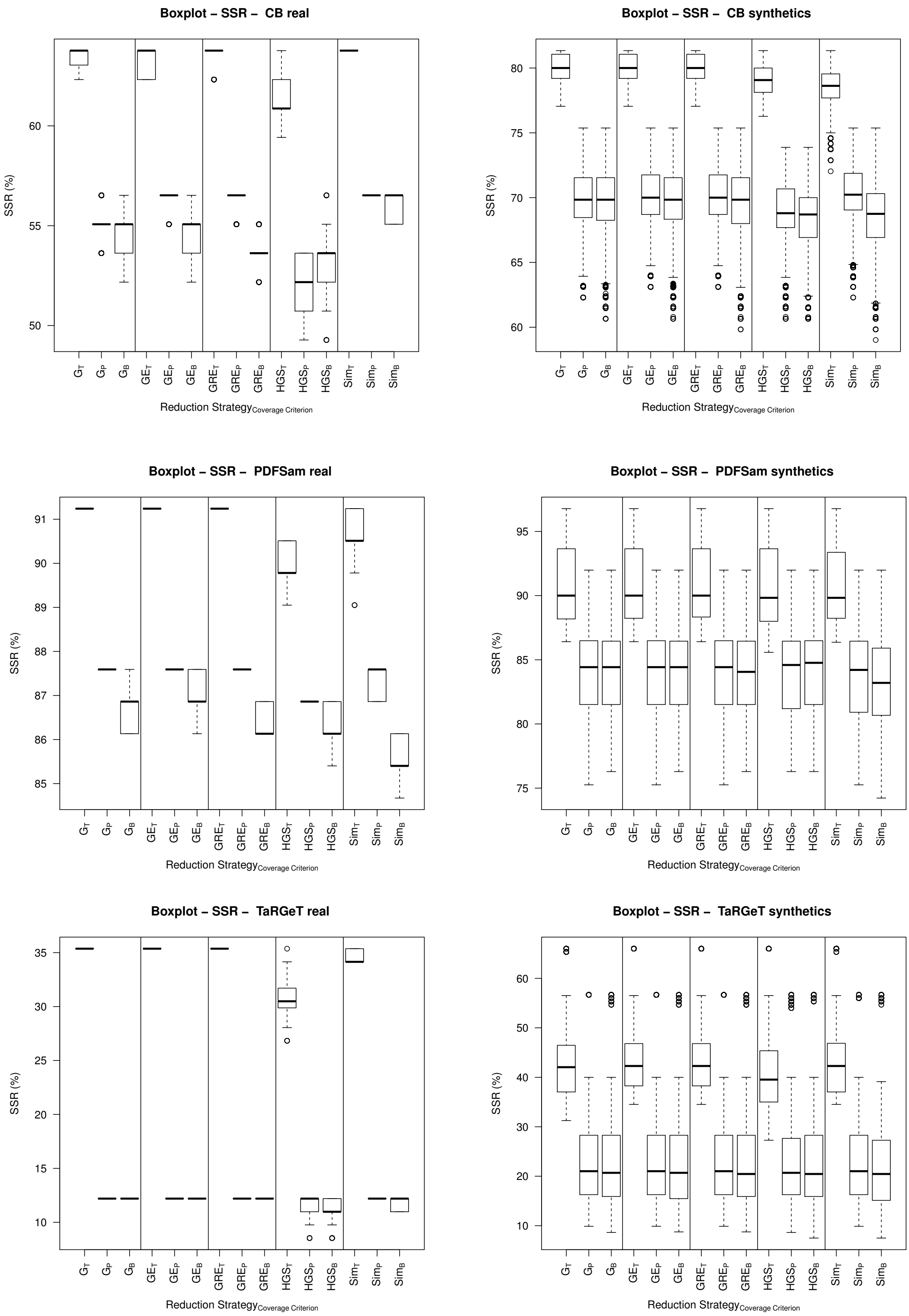


Figure 5.4: *Boxplots considering SSR metric for SQ1*

combinations. The $\hat{A}_{12}$ effect can be categorized as $Small < 0.10$, $0.10 < Medium < 0.17$ and $Large > 0.17$, the value being the distance from 0.5.

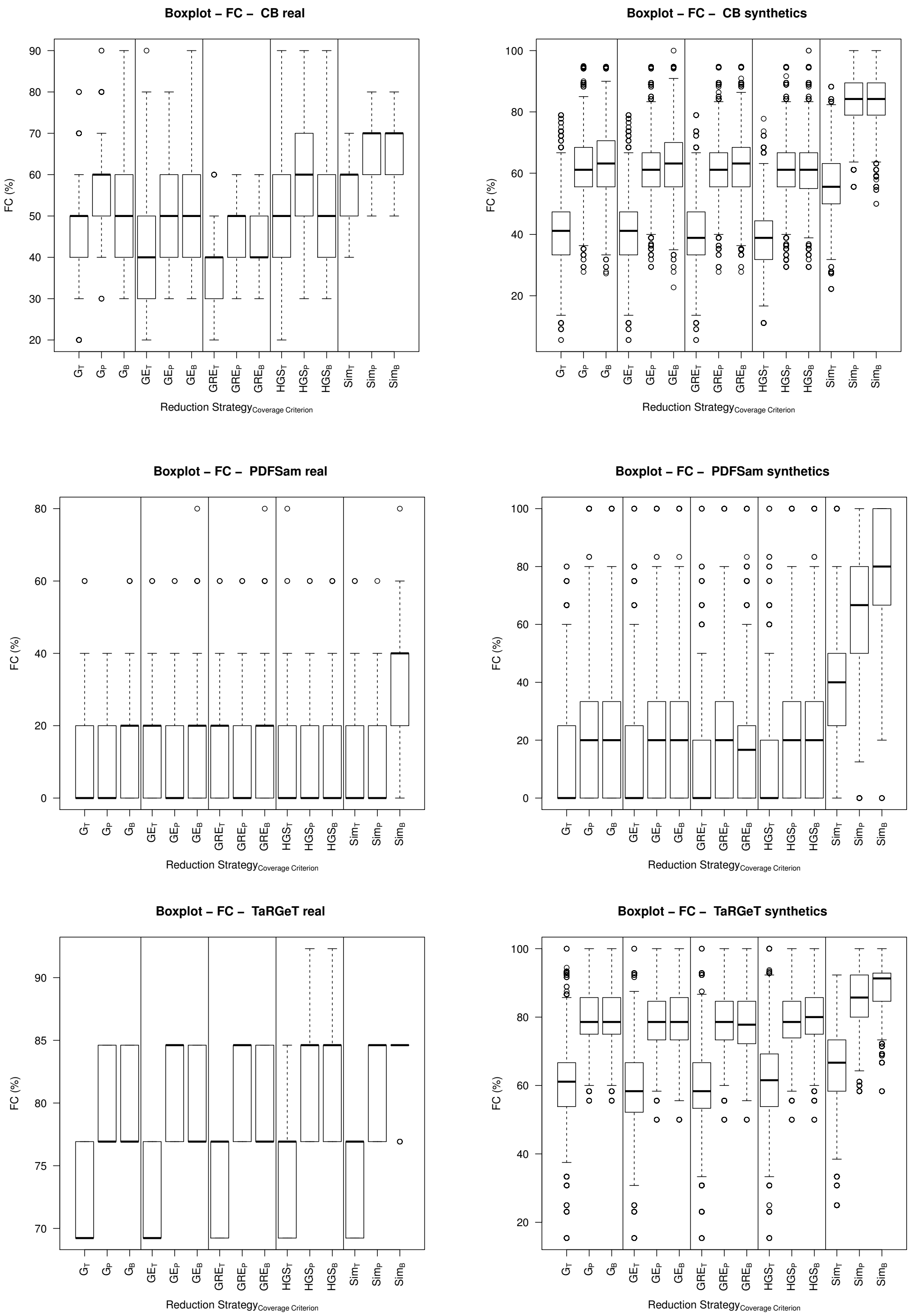


Figure 5.5: *Boxplots considering FC metric for SQ1*

Table 5.5 reports the performance order for each reduction strategy combined with different coverage criteria considering the SSR and FC metrics. These results are obtained from

Mann-Whitney tests and $\hat{A}_{12}$ effect size measurement (as can be seen in Appendix A.5.1).

Table 5.5: *Ordering of effectiveness for each reduction strategy associated with all coverage criteria in terms of SSR and FC*

	SSR	FC
CB real	$G_T > G_P > G_B$ $GE_T > GE_P > GE_B$ $GRE_T > GRE_P > GRE_B$ $HGS_T > HGS_B > HGS_P$ $Sim_T > Sim_P > Sim_B$	$G_P > G_B > G_T$ $GE_P > GE_B > GE_T$ $GRE_P > GRE_B > GRE_T$ $HGS_P > HGS_B > HGS_T$ $Sim_B > Sim_P > Sim_T$
CB synthetics	$G_T > G_P > G_B$ $GE_T > GE_P > GE_B$ $GRE_T > GRE_P > GRE_B$ $HGS_T > HGS_P > HGS_B$ $Sim_T > Sim_P > Sim_B$	$G_B > G_P > G_T$ $GE_B > GE_P > GE_T$ $GRE_B > GRE_P > GRE_T$ $HGS_P > HGS_B > HGS_T$ $Sim_P > Sim_B > Sim_T$
PDFSam real	$G_T > G_P > G_B$ $GE_T > GE_P > GE_B$ $GRE_T > GRE_P > GRE_B$ $HGS_T > HGS_P > HGS_B$ $Sim_T > Sim_P > Sim_B$	$G_B > G_P > G_T$ $GE_B > GE_T > GE_P$ $GRE_B > GRE_T > GRE_P$ $HGS_B > HGS_T > HGS_P$ $Sim_B > Sim_T > Sim_P$
PDFSam synthetics	$G_T > G_P > G_B$ $GE_T > GE_P > GE_B$ $GRE_T > GRE_P > GRE_B$ $HGS_T > HGS_B > HGS_P$ $Sim_T > Sim_P > Sim_B$	$G_P > G_B > G_T$ $GE_P > GE_B > GE_T$ $GRE_P > GRE_B > GRE_T$ $HGS_P > HGS_B > HGS_T$ $Sim_B > Sim_P > Sim_T$
TaRGeT real	$G_T > G_P = G_B$ $GE_T > GE_P = GE_B$ $GRE_T > GRE_P = GRE_B$ $HGS_T > HGS_P > HGS_B$ $Sim_T > Sim_P > Sim_B$	$G_P > G_B > G_T$ $GE_P > GE_B > GE_T$ $GRE_P > GRE_B > GRE_T$ $HGS_B > HGS_P > HGS_T$ $Sim_B > Sim_P > Sim_T$
TaRGeT synthetics	$G_T > G_P > G_B$ $GE_T > GE_P > GE_B$ $GRE_T > GRE_P > GRE_B$ $HGS_T > HGS_P > HGS_B$ $Sim_T > Sim_P > Sim_B$	$G_B > G_P > G_T$ $GE_B > GE_P > GE_T$ $GRE_P > GRE_B > GRE_T$ $HGS_B > HGS_P > HGS_T$ $Sim_B > Sim_P > Sim_T$

In general, the empirical studies present the similar behavior for SSR and FC. For SSR, the statistical tests suggest that T is more effective at reducing test suite for all reduction strategies investigated in all empirical studies. Futhermore, the results show that the effect size are large between T and P, and T and B, for all cases.

In terms of FC, in most of the cases P and B are the coverage criteria with best behavior, and generally the effect size between P and B are small, except for ***CB real*** (GE and Sim), ***PDFSam real PDFSam synthetics*** and ***TaRGeT real*** (Sim).

5.2.2 Study Question 2 (SQ2)

Figures 5.6 and 5.7, respectively, shows the boxplots of the SSR and FC metrics for each empirical study considering the SQ2. Since the boxplots are overlapped, we apply the Mann-Whiney test to compare pair of overlap. From Mann-Whitney tests and $\hat{A}_{12}$ effect size measurement (see Appendix A.5.2), we can observe the performance order of the reduction strategies for each coverage criteria considering the SSR and FC metrics, as presented in Table 5.6.

Table 5.6: *Ordering of effectiveness among reduction strategies for each coverege criteria in terms of SSR and FC*

	SSR	FC
CB real	$Sim_T > GRE_T > G_T > GE_T > HGS_T$ $Sim_P > GRE_P > GE_P > G_P > HGS_P$ $Sim_B > GE_B > G_B > GRE_B > HGS_B$	$Sim_T > HGS_T > G_T > GE_T > GRE_T$ $Sim_P > HGS_P > G_P > GE_P > GRE_P$ $Sim_B > HGS_B > G_B > GE_B > GRE_B$
CB synthetics	$GRE_T > G_T > GE_T > HGS_T > Sim_T$ $Sim_P > GE_P > GRE_P > G_P > HGS_P$ $G_B > GE_B > GRE_B > Sim_B > HGS_B$	$Sim_T > G_T > GE_T > GRE_T > HGS_T$ $Sim_P > G_P > GRE_P > GE_P > HGS_P$ $Sim_B > G_B > GRE_B > GE_B > HGS_B$
PDFSam real	$G_T = GE_T = GRE_T > Sim_T > HGS_T$ $G_P = GE_P = GRE_P > Sim_P > HGS_P$ $GE_B > G_B > GRE_B > HGS_B > Sim_B$	$GRE_T > GE_T > Sim_T > G_T > HGS_T$ $GRE_P > GE_P > Sim_P > G_P > HGS_P$ $Sim_B > GRE_B > GE_B > G_B > HGS_B$
PDFSam synthetics	$GRE_T > GE_T > G_T > Sim_T > HGS_T$ $G_P > GRE_P > GE_P > HGS_P > Sim_P$ $G_B > HGS_B > GE_B > GRE_B > Sim_B$	$Sim_T > G_T > GE_T > HGS_T > GRE_T$ $Sim_P > G_P > GE_P > HGS_P > GRE_P$ $Sim_B > G_B > GE_B > HGS_B > GRE_B$
TaRGeT real	$G_T = GE_T = GRE_T > Sim_T > HGS_T$ $G_P = GE_P = GRE_P = Sim_P > HGS_P$ $G_B = GE_B = GRE_B > Sim_B > HGS_B$	$HGS_T > Sim_T > GRE_T > GE_T > G_T$ $Sim_P > HGS_P > GE_T = GRE_T > G_P$ $Sim_B > HGS_B > GE_B > G_B > GRE_B$
TaRGeT synthetics	$GE_T > GRE_T > G_T > Sim_T > HGS_T$ $G_P > GRE_P > GE_P > Sim_P > HGS_P$ $G_B > GE_B > GRE_B > HGS_B > Sim_B$	$Sim_T > HGS_T > G_T > GRE_T > GE_T$ $Sim_P > G_P > HGS_P > GRE_P > GE_P$ $Sim_B > HGS_B > G_B > GE_B > GRE_B$

In terms of SSR, the heuristics G, GE and GRE present the best behavior, except for *CB real* and *CB synthetics* (only for T as coverage criterion). Furthermore, the effect size

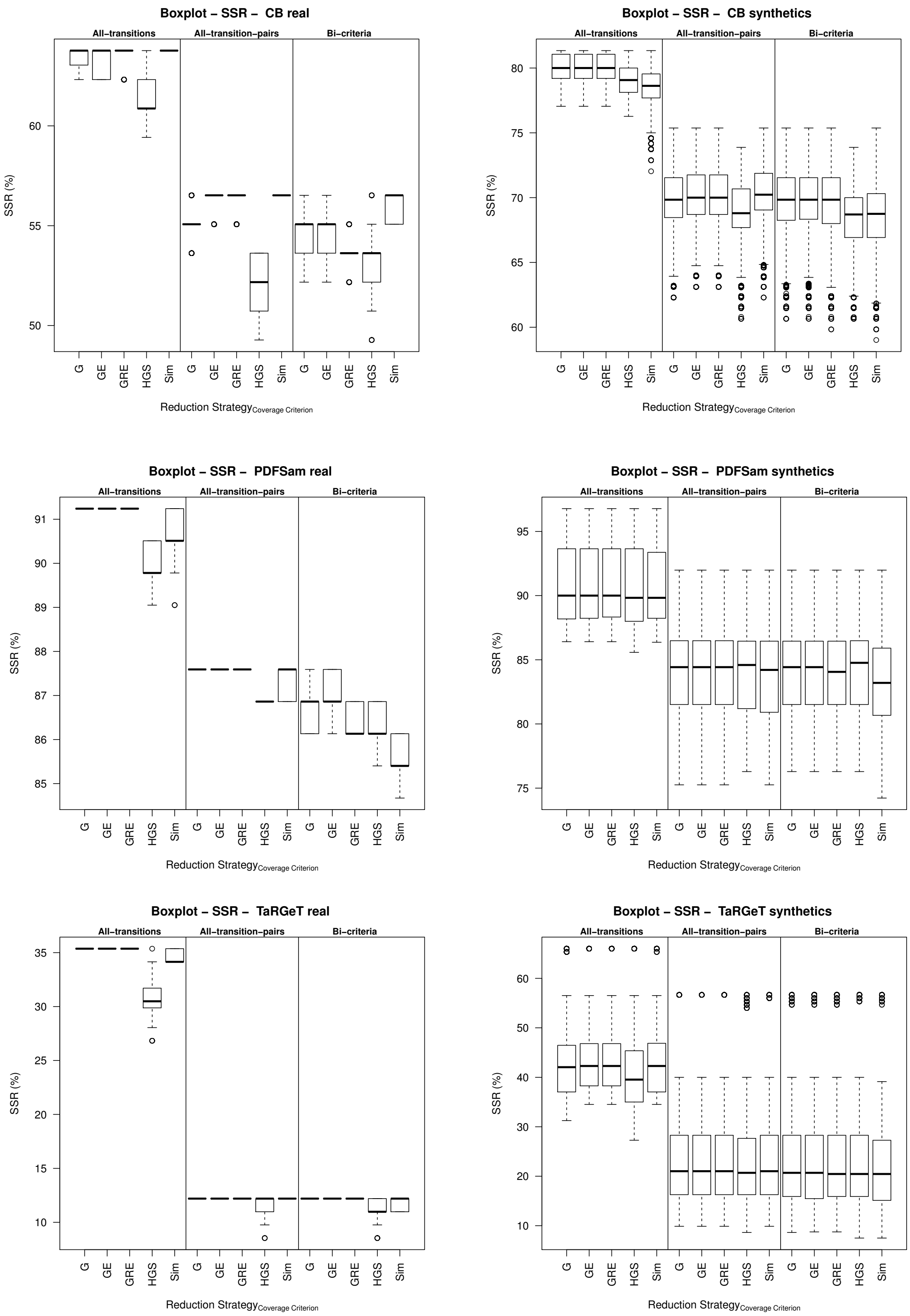


Figure 5.6: *Boxplots considering SSR metric for SQ2*

is classified as small among G, GE and GRE for all coverage criteria, i.e., there is no significant difference among heuristics, except for ***CB real*** considering P and B as coverage

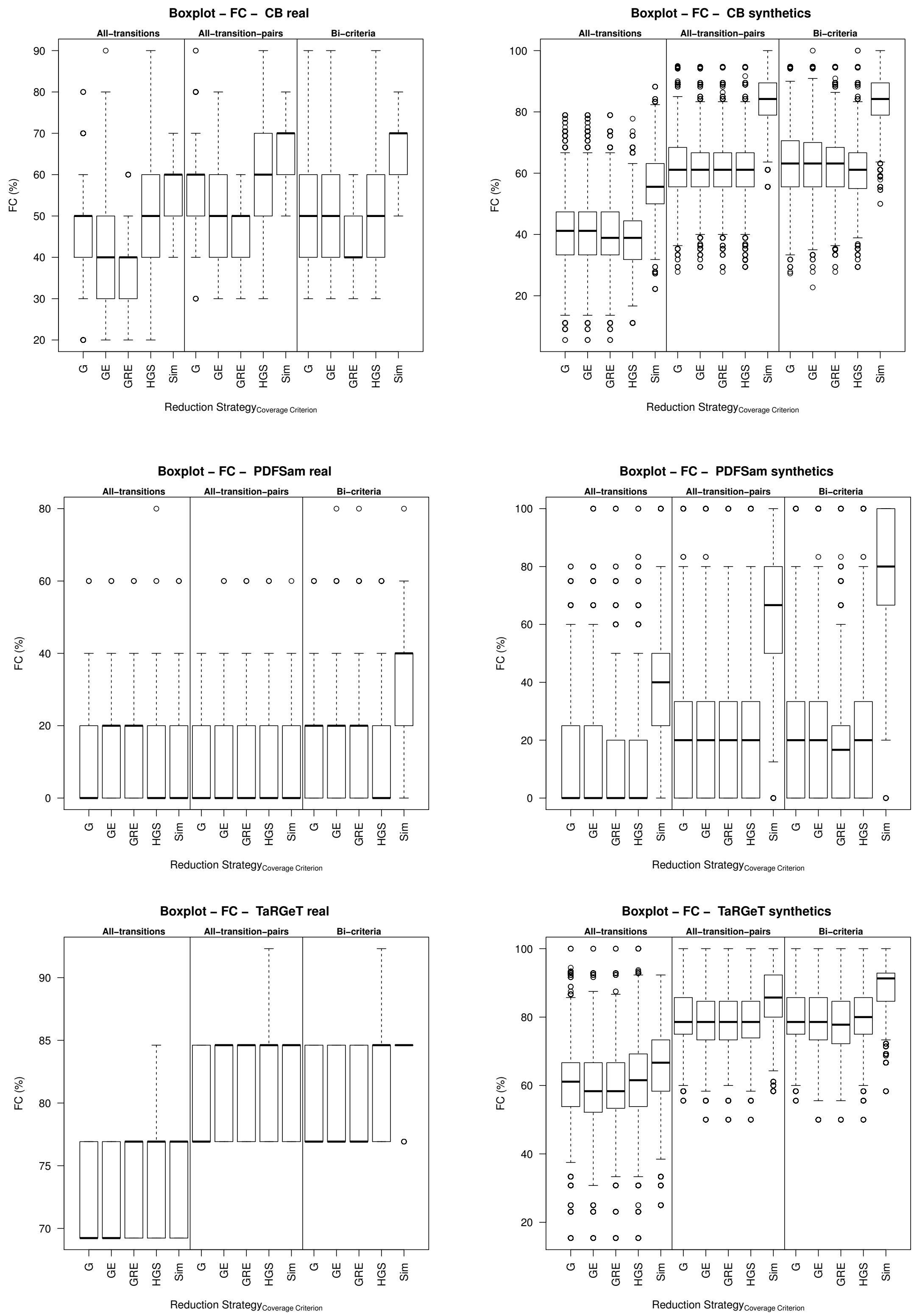


Figure 5.7: *Boxplots considering FC metric for SQ2*

criteria, and *PDFSam real* for B as coverage criterion. However, the percentage difference between the average is insignificant (maximum difference of 1.16%). For FC, in most cases

Sim presents the best behavior, except for *PDFSam real* (T and P) and *TaRGeT real* (T).

5.2.3 Study Question 3 (SQ3)

To address SQ3, we consider the reduction strategies in combination with the best coverage criterion in terms of SSR and FC according to results presented in Table 5.5 for each empirical study. Thus, the null and alternative hypotheses investigated are presented in Tables 5.7 and 5.8.

Table 5.7: *Null and alternative hypotheses for SSR considering SQ3*

CB real, CB synthetics, PDFSam real, PDFSam synthetics, TaRGeT real and TaRGeT synthetics	$H^0_{17}: SSR_{G_T} = SSR_{GE_T} = SSR_{GRE_T} = SSR_{HGS_T} = SSR_{Sim_T}$ $H^1_{17}: SSR_{G_T} \neq SSR_{GE_T} \neq SSR_{GRE_T} \neq SSR_{HGS_T} \neq SSR_{Sim_T}$

Table 5.8: *Null and alternative hypotheses for FC considering SQ3*

CB real	$H^0_{18}: FC_{G_P} = FC_{GE_P} = FC_{GRE_P} = FC_{HGS_P} = FC_{Sim_B}$ $H^1_{18}: FC_{G_P} \neq FC_{GE_P} \neq FC_{GRE_P} \neq FC_{HGS_P} \neq FC_{Sim_B}$
CB synthetics	$H^0_{19}: FC_{G_B} = FC_{GE_B} = FC_{GRE_B} = FC_{HGS_P} = FC_{Sim_P}$ $H^1_{19}: FC_{G_B} \neq FC_{GE_B} \neq FC_{GRE_B} \neq FC_{HGS_P} \neq FC_{Sim_P}$
PDFSam real	$H^0_{20}: FC_{G_B} = FC_{GE_B} = FC_{GRE_B} = FC_{HGS_B} = FC_{Sim_B}$ $H^1_{20}: FC_{G_B} \neq FC_{GE_B} \neq FC_{GRE_B} \neq FC_{HGS_B} \neq FC_{Sim_B}$
PDFSam synthetics	$H^0_{21}: FC_{G_P} = FC_{GE_P} = FC_{GRE_P} = FC_{HGS_P} = FC_{Sim_B}$ $H^1_{21}: FC_{G_P} \neq FC_{GE_P} \neq FC_{GRE_P} \neq FC_{HGS_P} \neq FC_{Sim_B}$
TaRGeT real	$H^0_{22}: FC_{G_P} = FC_{GE_P} = FC_{GRE_P} = FC_{HGS_B} = FC_{Sim_B}$ $H^1_{22}: FC_{G_P} \neq FC_{GE_P} \neq FC_{GRE_P} \neq FC_{HGS_B} \neq FC_{Sim_B}$
TaRGeT synthetics	$H^0_{23}: FC_{G_B} = FC_{GE_B} = FC_{GRE_P} = FC_{HGS_B} = FC_{Sim_B}$ $H^1_{23}: FC_{G_B} \neq FC_{GE_B} \neq FC_{GRE_P} \neq FC_{HGS_B} \neq FC_{Sim_B}$

In order to draw observations based on these hypotheses, we apply a statistical analysis similar to that used in the previous study questions. In turn, to clarify the magnitude of the difference between the reduction strategies combined with the different coverage criteria, we perform the Mann-Whitney tests and $\hat{A}_{12}$ effect size, and results are presented in Appendix A.5.3. From these results, we obtain the ordering of effectiveness for SSR and FC

behavior in Table 5.9.

Table 5.9: *Ordering of effectiveness reduction strategies in combination with their best coverage criterion regarding the SSR and FC metrics*

	SSR	FC
CB real	$Sim_T > GRE_T > G_T > GE_T > HGS_T$	$Sim_B > HGS_P > G_P > GE_P > GRE_P$
CB synthetics	$GRE_T > G_T > GE_T > HGS_T > Sim_T$	$Sim_P > G_B > GRE_B > GE_B > HGS_P$
PDFSam real	$G_T = GE_T = GRE_T > Sim_T > HGS_T$	$Sim_B > GRE_B > GE_B > G_B > HGS_B$
PDFSam synthetics	$GRE_T > GE_T > G_T > Sim_T > HGS_T$	$Sim_B > G_P > GE_P > HGS_P > GRE_P$
TaRGeT real	$G_T = GE_T = GRE_T > Sim_T > HGS_T$	$Sim_B > HGS_B > GE_P = GRE_P > G_P$
TaRGeT synthetics	$GE_T > GRE_T > G_T > Sim_T > HGS_T$	$Sim_B > HGS_B > G_B > GE_B > GRE_B$

By observing in Table 5.9, for SSR the results show that reduction strategies with T have a high reduction rate because T is weaker than P. Furthermore, the heuristics G, GE and GRE present the best behavior. On the other hand, the FC of the reduced test suite is adversely affected. In most cases, Sim is best reduction strategy considering B as coverage criterion, except for *CB synthetics*. However, in this case, the effect size is small between Sim_P and Sim_B.

5.2.4 Study Question 4 (SQ4)

For SQ4, we are interested in comparing the coverage criteria (T, P and B) with the best reduction strategies in terms of SSR and FC from SQ2 (Table 5.6). Based on the answers of SQ2, we define null and alternative hypotheses for each empirical study, as presented in Tables 5.10 and 5.11.

From Mann-Whitney tests and $\hat{A}_{12}$ effect size measurement (Appendix A.5.4), we can observe the performance order of the reduction strategies for each coverage are presented in Table 5.12.

In terms of SSR, it is clear that combining T with the heuristics (G, GE and GRE) presents the best behavior, i.e., a high reduction rate. For FC, the results show that Sim with B as coverage criterion has a better fault detection rate, except for *CB synthetics*.

Table 5.10: *Null and alternative hypotheses for SSR considering SQ4*

CB real	$H^0_{24} : SSR_{Sim_T} = SSR_{Sim_P} = SSR_{Sim_B}$ $H^1_{24} : SSR_{Sim_T} \neq SSR_{Sim_P} \neq SSR_{Sim_B}$
CB synthetics	$H^0_{25} : SSR_{GRE_T} = SSR_{Sim_P} = SSR_{G_B}$ $H^1_{25} : SSR_{GRE_T} \neq SSR_{Sim_P} \neq SSR_{G_B}$
PDFSam real	$H^0_{26} : SSR_{(G_T=GE_T=GRE_T)} = SSR_{(G_P=GE_P=GRE_P)} = SSR_{GE_B}$ $H^1_{26} : SSR_{(G_T=GE_T=GRE_T)} \neq SSR_{(G_P=GE_P=GRE_P)} \neq SSR_{GE_B}$
PDFSam synthetics	$H^0_{27} : SSR_{GRE_T} = SSR_{G_P} = SSR_{G_B}$ $H^1_{27} : SSR_{GRE_T} \neq SSR_{G_P} \neq SSR_{G_B}$
TaRGeT real	$H^0_{28} : SSR_{(G_T=GE_T=GRE_T)} = SSR_{(G_P=GE_P=GRE_P=Sim_P)} = SSR_{(G_B=GE_B=GRE_B)}$ $H^1_{28} : SSR_{(G_T=GE_T=GRE_T)} \neq SSR_{(G_P=GE_P=GRE_P=Sim_P)} \neq SSR_{(G_B=GE_B=GRE_B)}$
TaRGeT synthetics	$H^0_{29} : SSR_{GE_T} = SSR_{G_P} = SSR_{G_B}$ $H^1_{29} : SSR_{GE_T} \neq SSR_{G_P} \neq SSR_{G_B}$

Table 5.11: *Null and alternative hypotheses for FC considering SQ4*

CB real	$H^0_{30} : FC_{Sim_T} = FC_{Sim_P} = FC_{Sim_B}$ $H^1_{30} : FC_{Sim_T} \neq FC_{Sim_P} \neq FC_{Sim_B}$
CB synthetics	$H^0_{31} : FC_{Sim_T} = FC_{Sim_P} = FC_{Sim_B}$ $H^1_{31} : FC_{Sim_T} \neq FC_{Sim_P} \neq FC_{Sim_B}$
PDFSam real	$H^0_{32} : FC_{GRE_T} = FC_{GRE_P} = FC_{Sim_B}$ $H^1_{32} : FC_{GRE_T} \neq FC_{GRE_P} \neq FC_{Sim_B}$
PDFSam synthetics	$H^0_{33} : FC_{Sim_T} = FC_{Sim_P} = FC_{Sim_B}$ $H^1_{33} : FC_{Sim_T} \neq FC_{Sim_P} \neq FC_{Sim_B}$
TaRGeT real	$H^0_{34} : FC_{HGS_T} = FC_{Sim_P} = FC_{Sim_B}$ $H^1_{34} : FC_{HGS_T} \neq FC_{Sim_P} \neq FC_{Sim_B}$
TaRGeT synthetics	$H^0_{35} : FC_{Sim_T} = FC_{Sim_P} = FC_{Sim_B}$ $H^1_{35} : FC_{Sim_T} \neq FC_{Sim_P} \neq FC_{Sim_B}$

5.3 Scattering

By analyzing the data obtained, we can also observe the *scattering* of the faults for all reduction strategies regarding the different coverage criteria. The idea is to apply the reduction strategy with a new stop criterion, in this case 100% test requirements and 100% fault coverage, and observe the percentage of the number of test cases removed from the complete test

Table 5.12: *Ordering of effectiveness coverage criteria in combination with their best reduction strategy regarding the SSR and FC metrics*

	SSR	FC
CB real	$Sim_T > Sim_P > Sim_B$	$Sim_B > Sim_P > Sim_T$
CB synthetics	$GRE_T > Sim_P > G_B$	$Sim_P > Sim_B > Sim_T$
PDFSam real	$G_T = GE_T = GRE_T > G_P = GE_P = GRE_P > GE_B$	$Sim_B > GRE_T > GRE_P$
PDFSam synthetics	$GRE_T > G_P > G_B$	$Sim_B > Sim_P > Sim_T$
TaRGeT real	$G_T = GE_T = GRE_T > G_P = GE_P = GRE_P = Sim_P = G_B = GE_B = GRE_B$	$Sim_B > Sim_P > Sim_T$
TaRGeT synthetics	$GE_T > G_P > G_B$	$Sim_B > Sim_P > Sim_T$

suite that reaches this goal. This percentage can be calculated by the following metric:

$$SSR\_FC = \frac{|TS| - |TS'|}{|TS|} \times 100\%$$

where $|TS|$ is the number of test cases in the complete test suite and $|TS'|$ is the number of test cases that reaches 100% fault coverage.

To evaluate this metric, we consider all previous study questions. Based on results for SQ1 and SQ2, as can be seen in Appendix A.7.4 (Tables A.51 to A.56), respectively, in order to address SQ3 and SQ4, we generate boxplots for each empirical study, and we apply Mann-Whitney tests and $\hat{A}_{12}$ effect size measurement (see Tables A.57 to A.62). For SQ3, the performance order for each of the reduction strategies combined with the different coverage criteria from SQ1 are depicted in Table 5.13. In this case, *Sim* is in average the best reduction strategy combined with B as coverage criterion. The similar result is obtained for SQ4.

Another important point to highlight is that for four of the six empirical studies, the best reduction strategy combined with the best coverage criterion for SSR_FC is similar to the results obtained for FC, except for *PDFSam real* and *TaRGeT real*, for both questions (SQ3 and SQ4).

Table 5.13: *Ordering of effectiveness for SQ3 and SQ4 regarding the* SSR_FC

	SQ3	SQ4
CB real	$Sim_B > GRE_T > GE_T > G_T > HGS_T$	$Sim_B > GRE_T > G_P$
CB synthetics	$Sim_P > HGS_B > G_P > GE_P > GRE_P$	$Sim_P > Sim_B > HGS_T$
PDFSam real	$GRE_P > GE_B > G_B > Sim_P > HGS_T$	$GRE_P > G_B > GE_T$
PDFSam synthetics	$Sim_B > HGS_B > G_P > GE_T > GRE_T$	$Sim_B > Sim_P > Sim_T$
TaRGeT real	$Sim_T > GRE_T > GE_T > HGS_T > G_T$	$Sim_T > G_P > GE_B$
TaRGeT synthetics	$Sim_B > HGS_T > G_T > GRE_T > GE_T$	$Sim_B > Sim_P > Sim_T$

5.4 Concluding Remarks

This chapter presented six empirical studies aiming to compare five reduction strategies (G, GE, GRE, HGS and Sim) with three different coverage criteria (T, P and B). These empirical studies provide evidence on the impacts that the choice of a coverage criterion can have on the performance of reduction strategies regarding suite size reduction, fault coverage, and scattering. In the presented empirical studies, we consider three different real-world specification models, and three groups of 30 synthetic specification models with a comparable configuration to each real-world specification model. It is important to highlight that these configurations are different, and the differences may impact directly on the number of generated test cases and the degree of redundancy among test cases.

Results clearly show that the choice of coverage criteria can influence suite size reduction, fault coverage and scattering. The reason is that each coverage criterion leads to a different set of test requirements and it is possible to have significant differences on choice of the test cases for the reduced test suite.

For SSR, reduction rate presents differences that are more significant on performance between coverage criteria, such as between T and P, and T and B, since T is weaker than P and B. In turn, we can conclude that the combination among the heuristics G, GE, GRE with T as coverage criterion proved to be the most efficient for test suite reduction. These results confirm the widely expected fact that weaker coverage criteria indeed tends to favor reduction size whereas compromising fault coverage.

In terms of FC, results show that among all alternatives the combination of the Sim

with B (*all-transitions* and *all-transition-pairs*), can significantly increase the fault coverage rate with not a very significant loss on reduction size. In other words, when combining two criteria we may add a few more test cases, improving fault coverage a little further. These results obtained indicate that to select a subset of the most different and non-redundant test cases that covers all requirements, and while maintaining some redundancy in the reduced test suite by using of multi-criteria may improve the rate of fault coverage. Thus, while reduction strategies can be indeed effetive in reducing size, the studies (along with other studies presented in the literature) show that the choice of coverage criteria is key to effective fault coverage. Therefore, the use of B is a promissing approach.

Regarding scattering, the results show that Sim can be more effective than the heuristics for 100% test requirements and 100% fault coverage, and in most cases, the best coverage criterion is B. Therefore, Sim_B is a promissing approach to be further investigated and applied in practice.

Furthermore, different circunstances may lead a test manager to apply one or the other strategy when reducing MBT test suites. For instance, if there are few resources for test case execution, particularly manual execution, and there is low expectation of failure, G_T may be a good choice, since it is effective on reducing the suite and the costs of applying it may be lower than the others. On the other hand, if there is high expectation of failure, Sim can be applied with T, P or B, depending the availability of resources to run the reduced suite.

Chapter 6

Review on Test Suite Reduction

In this chapter, we present related work on strategies for test suite reduction. In literature, different studies have been developed to produce a reduced test suite from the complete test suite that covers a given set of test requirements. Our goal is to present strategies to reduce test suites that can be automated or/and used in the MBT context. First, we present some heuristics and clusters for code-based reduction, followed by comparative studies and strategies that allow the use of multiple testing criteria. Afterwards, strategies for specification-based reduction are presented.

6.1 Heuristics and Clusters

A number of test suite reduction strategies based on classical greedy algorithm have been reported. Tallam and Gupta [Tallam and Gupta 2005] proposed a greedy heuristic called *delayed-greedy strategy* inspired on a *concept analysis* framework. Concept analysis is a hierarchical clustering technique for classifying objects with discrete attributes. In their experiments, the reduced suites produced by this strategy were consistently of the same size or smaller size than prior heuristics, such as HGS and G.

Also inspired by the greedy algorithm, Parsa and Khalilian [Parsa and Khalilian 2009] present a strategy to minimize the test suite with two objectives: to generate the smaller reduced test suite and improve the fault detection effectiveness compared to other strategies. For this, the new heuristic algorithm combined the ideas of coverage-based and distribution-based approaches. Thus, to compose the reduced test suite the test cases should satisfy two objectives simultaneously: they must satisfy the maximum number of unsatisfied test requirements and it must have the minimum overlap in requirements coverage with other test cases.

Xu et al. [Xu et al. 2012] presented a strategy to reduce the size of a test suite and to decrease the total cost at the same time, called *Modified Greedy Algorithm*. This solution was inspired by the *weighted set covering problem* (WSC). The use of WSC techniques allowed to eliminate the redundancy and dynamically determine the priority of test cases to lower costs.

Based on *cluster analysis*, Parsa et al. [Parsa et al. 2009] proposed a strategy for test suite reduction. The *cluster analysis* define groups of objects with similar attributes. After clustering, the test cases are sampled from each group (cluster). According to this strategy, these clusters of test cases are based on similarity according to a certain coverage criterion. The most different test cases are chosen to form the reduced test suite.

Selvakumar et al. [Selvakumar et al. 2010a] suggested an algorithm to reduce the test suite and to improve the fault detection based on integration of *concept analysis* and *genetic algorithm*. Initially, the *concept analysis* is used to generate clusters of test cases. Later, these groups of test cases (initial population) are used by a genetic algorithm. Finally, a method was suggested and adopted to handle tie breaking conditions from the choice of the group of test cases having the larger number of intersections with others in the same level.

To maintain or even improve fault detection in test suite reduction, Zhang et al. [Zhang et al. 2010] proposed a strategy that adds some redundant test cases in the reduced set. For this, they proposed the concept of *relative redundancy* for test suite reduction. They show that this strategy increases the size of the reduced test suite a little, and it can retain or improve fault detection effectiveness.

6.1.1 Comparative Studies

Several studies have been conducted to compare different test suite reduction strategies proposed in literature, such as the ones proposed by Chen and Lau [Chen and Lau 1998b] and Zhong et al. [Zhong et al. 2006; Zhong et al. 2008]. The goal of both studies was to provide guidelines for choosing the most appropriate test suite reduction strategy.

According to Chen and Lau [Chen and Lau 1998b], the result of a simulation study with four heuristics: G, GE, GRE and HGS (presented in Section 2.5.1), was presented. In this study, the choice of the most apropriate reduction strategy depends on the satisfiability relation and the ratio of overlapping (a relative measure of the average number of test cases that satisfy each test requirement). However, fault detection capability was not considered since the heuristics are solely judged by the sizes of the reduced test suite and by execution time of the heuristics.

Zhong et al. [Zhong et al. 2006; Zhong et al. 2008] present an experimental study comparing HGS, GRE, genetic algorithm-based strategy and a strategy based on *Integer Linear Programing* (ILP). This study observes that all the four strategies can dramatically reduce the size of test suites, but the ILP-based strategy always produce the smallest representative sets. This study suggests the heuristic HGS as the first choice although the ILP-based strategy is recommended when the smallest reduced test suite is required or the fault detection capability needs to be ensured. Furthermore, the context of this experiment is in regression testing, where error detection information is required.

Finally, Rothermel et al. [Rothermel et al. 2002] presented empirical studies in order to evaluate the size and fault detection capability of the reduced test suites for heuristic HGS. They concluded that test suite reduction can drastically reduce the fault detection capability.

6.1.2 Using Multiple Testing Criteria

Black et al. [Black et al. 2004] present a strategy for test suite reduction by simultaneously combining bi-criteria from the use of binary *Integer Linear Programing* (ILP). An empirical study was performed by using the Siemens suite [1], and the results show that the suite size reduction and fault coverage could vary according to particular weighting factor used.

The strategy presented by Jeffrey and Gupta [Jeffrey and Gupta 2007] for reduction uses multiple testing criteria to improve the effectiveness in the fault coverage. The key idea is to add test cases in the reduced test suite that are redundant with respect to a particular coverage criterion, if the test cases are not redundant according to one or more other coverage criteria (*selective redundancy*). The results of the experimental study with Siemens suite and the Space program show that this strategy generates reduced test suites with less fault detection loss at the expense of only a relatively small increase in the sizes of the reduced suites.

[1] http://sir.unl.edu/portal/index.html

To treat a tie situation in traditional test suite reduction strategies, Lin and Huang [Lin et al. 2008; Lin and Huang 2009] proposed the use of additional testing criterion instead of a random choice. This strategy is called *reduction with tie-breacking* (RTB). To illustrate RTB, the HGS and GRE strategies were modified and evaluated in an experimental study. In this experiment, they concluded that RTB can improve the fault detection effectiveness with a negligible increase in the sizes of the reduced suites.

Hsu and Orso [Hsu and Orso 2009] developed a test suite reduction framework in a tool called MINTS. This tool permits encoding multi-criteria as binary *Integer Linear Programing* (ILP) problems and leveraged existing modern ILP solvers aiming to find optimal minimal solutions and increase fault detection or decrease cost.

In another work, Selvakumar et al. [Selvakumar et al. 2010b] modified GRE heuristic with selective redundancy ($GSRE$) for test suite reduction with the use of multiple testing criteria. In this strategy, some additional redundant test cases are added in the reduced test suite through *selective redundancy*. Hence, the reduced test suite is slightly bigger and the fault detection capability can be larger.

Chen et al. [Chen et al. 2011] proposed a strategy to test suite reduction based on pairwise interaction of test requirements, called $PWIR$. The idea is that covering all the pairwise interactions of requirements may improve fault detection without much increase of the size of the reduced test suite.

Using the *cluster analysis*, Khalilian and Parsa [Khalilian and Parsa 2012] proposed the use of two different coverage criteria during the reduction process to improve the fault detection capability of the reduced test suite. This algorithm is divided into two steps: $1)$ the test suite is reduced in order to satisfy the first coverage criterion and $2)$ the reduced test suite must be modified to satisfy the second criterion.

Pan Liu [Liu 2014] presents a novel reduction strategy for regression testing from the selection of test cases. In this paper, the heuristic HGS is extended by replacing the random choice for the same rank test cases according to the boundary coverage capability as second coverage criterion. However, only a simple case study is presented. The results shown that the use of a second coverage criterion leads to increases the fault detection rate.

6.2 Specification-based Reduction

In the MBT context, Heimdahl and George [Heimdahl and George 2004] investigated the use of several coverage criteria to significantly reduce the sizes of the test suites generated, such as: transition coverage, decision coverage, *Modified Condition/Decision Coverage* (MC/DC), MC/DC usage, among others. To reduce the test suite, the strategy chooses the test cases randomly. If this test case improves the coverage criterion, then it is added to the reduced set. On the other hand, the reduced test suite has a decrease in fault detection capability.

Jourdan et al. [Jourdan et al. 2006] propose the identification of patterns of interaction among the elements of the Extended Finite State Machine (EFSM) model that affect a requirement under test from the analysis of control and data dependencies. Based on this, the equivalent test cases are identified, and test suites can be reduced, keeping only one test case per equivalence class and eliminating the others.

Fraser and Wotawa [Fraser and Wotawa 2007] present a strategy to reduce the test suite with respect to the number of test cases and the total length of all test cases based on model-checker concepts. For this, they present a measure for redundancy that only considers a common prefix among the test cases, where the value of redundancy can be illustrated by representing a set of test cases as a tree. Hence, based on this measure of redundancy of the test suite, the test cases are transformed in order to avoid the redundancy.

Cichos and Heinze [Cichos and Heinze 2011] propose a strategy for similarity-based test suite reduction in the MBT context. This strategy identifies test case pairs that are especially suitable for merging, based on their similarity. They show that the size of the reduced test suite can be very close to the optimum.

In another study, Cartaxo [Cartaxo 2011] presents a strategy to reduce test suites based on dissimilarity in the MBT context. The idea is keep in the reduced test suite the most different test cases while providing 100% coverage of one defined test requirements. The case study presented shows that this strategy presents the worst rate of reduction compared to well-known heuristics in literature, however it presents the best percentage for faults coverage.

6.3 Concluding Remarks

This chapter discussed some studies for test suite reduction. A number of experimental studies have been proposed to investigate and to evaluate the test suite reduction strategies based on comparison of different strategies proposed in literature. These studies demonstrated that the fault detection capability can be significantly decreased by the reduction of the test suite. To increase the fault detection effectiveness, many strategies have been proposed and extensively experimented, but the results cannot be generalized and sometimes they are divergent.

However, most of the related works focus on code-based criteria, and few test suite reduction strategies for MBT context have been proposed. Moreover, these works do not evaluate the fault detection effectiveness of reduced test suites, considering only the rate of reduction, i.e., the reduced test suite size. Among these works, several present only simple case studies.

Notice that despite the use of specialized analysis and multiple criteria, most of them are based on the classical heuristics presented in Section 2.5.1 and/or their combination. Therefore, for the sake of simplicity of experimental design, these heuristics are used for comparison with our proposed strategy.

Chapter 7

Concluding Remarks

This chapter summarizes the main results of this work and presents some suggestions for future work. First, the conclusions are drawn in Section 7.1 and the possible future works are presented in Section 7.2.

7.1 Conclusions

The main objective of this doctorate research is to improve the process of test suite reduction by proposing a reduction strategy based on similarity in the context of MBT aiming to maximize the fault coverage of the reduced test suite. Considering the research questions defined in Section 1.2, the following results were achieved:

Research Question 1 *In the context of MBT, how to address similarity among test cases to reduce the size of the test suite while simultaneously maintain a reasonable fault coverage?*

In order to answer this first research question, we propose a new parametrized reduction strategy based on the use of a similarity function and multi-criteria presented in Chapter 3. The key idea is to reduce the test suite with the removal of the most similar test cases and at the same time maintaining a little redundancy in the reduced suite with the use of multiple criteria to improve diversity of the test cases selected and increase its chances of covering more faults, while maintaining 100% of the testing requirements covered. Our reduction strategy allows testers to choose the parameters, such as similarity function, coverage criteria and choice function. These choices may influence the selection of the test cases for the reduced test suite. A preliminary study is published in [Coutinho et al. 2013] and indicates that the reduced test suite sizes obtained by applying our strategy with single coverage criterion is in average similar than that one obtained by applying the heuristics. Regarding fault coverage, it can be observed that in average our strategy presents an improvement on fault detection effectiveness compared to other heuristics. Other studies presented in Chapter 5 show that our reduction strategy (Sim) with the use of multiple criteria (*bi-criteria* – B), can significantly increase the fault coverage rate while maintaining similar the reduced test suite size, when compared to the heuristics G, GE, GRE and HGS, and also with Sim considering the use of a single coverage criterion.

Research Question 2 *What influence does the choice of a similarity function have on the size and fault coverage of similarity-based test suite reduction techniques?*

In order to answer this second research question, three empirical studies were performed to investigate the effectiveness of our distance function and other 5 well-known distance functions applied in our similarity-based test suite reduction strategy in the context of MBT considering *all-transition-pairs* as coverage criterion. Our distance function, presented in Section 3.2, is inspired by the redundancy measure presented by Cartaxo et al. [Cartaxo et al. 2011]. This function, named *similarity function*, considers the number of repetitions of a transition, for instance, if a loop is traversed more than once, to calculate the redundancy measure between two test cases. As a result, we observed in Chapter 4 that the choice of a distance function may directly influence on the performance of the reduction strategy regarding suite size reduction, fault coverage and stability.

Research Question 3 *What influence does the coverage criteria have on test suite reduction regarding size and fault coverage of a reduced test suite?*

In order to answer this third question, we performed six empirical studies to investigate the influence of the choice of coverage criteria used by reduction strategies regarding suite size reduction and fault coverage. For this, the following coverage criteria were considered: *all-transitions*, *all-transition-pairs* and *bi-criteria*. For *bi-criteria*, first we apply a weaker criterion (*all-transitions*) followed by a stronger criterion (*all-transition-pairs*). In the results presented in Chapter 5, we observed that the choice of the coverage criteria has direct influence on reduction rate, and in the fault coverage. Thus, different circunstances may lead a test manager to apply one or the other strategy when reducing MBT test suites. The results suggest that the application of *all-transitions* as coverage criterion can dramatically reduce the size of test suites for all reduction strategies. Thus, if there are few resources for test case execution, particularly manual execution, and there is low expectation of failure, G_T, GE_T or GRE_T may be a good choice, since it is effective on reducing the suite and the costs of applying it may be lower than the others. On the other hand, if there is high expectation of failure, choice of coverage criteria and reduction strategy are keys to effective fault coverage. In most of the cases, the results indicate that among all alternatives the combination of the our reduction strategy Sim with *bi-criteria* (*all-transitions* and *all-transition-pairs*) as coverage criterion, can significantly increase the fault coverage rate. Furthermore, this combination (Sim_B) can be more effective than the heuristics considering 100% test requirements and 100% fault coverage.

7.2 Future Works

There are several problems that need to be solved and improved in future works. Next, some work proposals are presented:

Distance Function From the results obtained in the empirical studies presented in Chapter 4, we can have an overview of the distance functions behavior and effect on similarity-based test suite reduction, even though no definite conclusions can be reached yet. Besides, the results can motivate further investigation in the area, for instance regarding improvements on distance functions to suit the test suite reduction problem. The choice of a distance function can clearly influence fault coverage and also stability of a similarity-based strategy. On the other hand, fault coverage seems to be related to suite size reduction independently of the choice of the function. This motivates further investigation of how to improve the reduction strategy as well. Furthermore, executing more case studies and experiments using other configurations as input of the LTS generator is part of our future work.

Choice Function In our reduction strategy, a choice function is applied to define the order of analysis of a pair of test cases. In this thesis, the choice function used is based on their path lengths aiming to be similar the choice of test cases by well-known heuristics in literature. Thus, the first test case to be analyzed has the lowest number of transitions. If the test cases have the same length, one of them is chosen randomly. The influence of choice functions for our test suite reduction can be investigated in future works.

Coverage Criteria Test suite reduction is based on coverage criteria that determine which test cases to remove from the complete test suite with the aim of creating a smaller set of test cases that satisfies all of the test requirements as the complete test suite. In this work, we propose and investigate the use of multiple criteria for test suite reduction. Further investigation is necessary to evaluate the relationship of multiple criteria for our reduction strategy on industrial applications and generated specification models with different characteristics from the execution of more experiments.

Test Suite Reduction During the application of our reduction strategy, when there is a tie between values of the similarity matrix then a random choice is applied. Our idea is to apply and investigate other choice methods such as length, coverage criteria, etc.

Bibliography

[Abran et al. 2004] Abran, A., Bourque, P., Dupuis, R., and Moore, J. W., editors (2004). *Guide to the Software Engineering Body of Knowledge - SWEBOK*. IEEE Press, Piscataway, NJ, USA.

[Akleman and Chen 1999] Akleman, E. and Chen, J. (1999). Generalized distance functions. In *Shape Modeling International*, pages 72–79. IEEE Computer Society.

[Ammann and Offutt 2008] Ammann, P. and Offutt, J. (2008). *Introduction to Software Testing*. Cambridge University Press, New York, NY, USA, 1 edition.

[Anand et al. 2013] Anand, S., Burke, E. K., Chen, T. Y., Clark, J., Cohen, M. B., Grieskamp, W., Harman, M., Harrold, M. J., and Mcminn, P. (2013). An orchestrated survey of methodologies for automated software test case generation. *J. Syst. Softw.*, 86(8):1978–2001.

[Araújo et al. 2012] Araújo, J. D. S., Cartaxo, E. G., Neto, F. G. O., and Machado, P. D. L. (2012). Controlando a diversidade e a quantidade de casos de teste na geração automática a partir de modelos com loop. In *6th Brazilian Workshop on Systematic and Automated Software Testing, 2012*, Natal, RN, Brazil.

[Arcuri and Briand 2011] Arcuri, A. and Briand, L. (2011). A practical guide for using statistical tests to assess randomized algorithms in software engineering. In *Software Engineering (ICSE), 2011 33rd International Conference on*, pages 1–10.

[Barbosa et al. 2007] Barbosa, D. L., , Lima, H. S., Machado, P. D. L., Figueiredo, J. C. A., Jucá, M. A., and Andrade, W. L. (2007). Automating functional testing of components from UML specifications. *International Journal of Software Engineering and Knowledge Engineering*, 17(03):339–358.

[Bertolino 2007] Bertolino, A. (2007). Software testing research: Achievements, challenges, dreams. In *Future of Software Engineering, 2007. FOSE '07*, pages 85–103.

[Bertolino et al. 2010] Bertolino, A., Cartaxo, E., Machado, P., Marchetti, E., and ao Ouriques, J. (2010). Test suite reduction in good order: Comparing heuristics from a new viewpoint. In *Proceedings of the 22nd IFIP International Conference on Testing Software and Systems: Short Papers*, pages 13–18. CRIM.

[Binder 2000] Binder, R. V. (2000). *Testing object-oriented systems: models, patterns, and tools*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Black et al. 2004] Black, J., Melachrinoudis, E., and Kaeli, D. (2004). Bi-criteria models for all-uses test suite reduction. In *Software Engineering, 2004. ICSE 2004. Proceedings. 26th International Conference on*, pages 106–115.

[Cartaxo et al. 2007] Cartaxo, E., Neto, F., and Machado, P. (2007). Test case generation by means of UML sequence diagrams and labeled transition systems. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 1292–1297.

[Cartaxo 2011] Cartaxo, E. G. (2011). *Estratégias para Controlar o Tamanho da Suíte de Teste Gerada a partir de Abordagens MBT*. PhD thesis, Universidade Federal de Campina Grande, Campina Grande, Paraíba.

[Cartaxo et al. 2008] Cartaxo, E. G., Andrade, W. L., Neto, F. G. O., and Machado, P. D. L. (2008). LTS-BT: a tool to generate and select functional test cases for embedded systems. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC'08, pages 1540–1544, New York, NY, USA. ACM.

[Cartaxo et al. 2011] Cartaxo, E. G., Machado, P. D. L., and Neto, F. G. O. (2011). On the use of a similarity function for test case selection in the context of model-based testing. *Software Testing, Verification and Reliability*, 21(2):75–100.

[Chen et al. 2010] Chen, T. Y., Kuo, F.-C., Merkel, R. G., and Tse, T. H. (2010). Adaptive random testing: The ART of test case diversity. *Journal of Systems and Software*, 83(1):60–66.

[Chen and Lau 1998a] Chen, T. Y. and Lau, M. F. (1998a). A new heuristic for test suite reduction. *Information & Software Technology*, 40(5-6):347–354.

[Chen and Lau 1998b] Chen, T. Y. and Lau, M. F. (1998b). A simulation study on some heuristics for test suite reduction. *Information & Software Technology*, 40(13):777–787.

[Chen et al. 2011] Chen, X., Zhang, L., Gu, Q., Zhao, H., Wang, Z., Sun, X., and Chen, D. (2011). A test suite reduction approach based on pairwise interaction of requirements. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 1390–1397, New York, NY, USA. ACM.

[Chvätal 1979] Chvätal, V. (1979). A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235.

[Cichos and Heinze 2011] Cichos, H. and Heinze, T. (2011). Efficient test suite reduction by merging pairs of suitable test cases. In Dingel, J. and Solberg, A., editors, *Models in Software Engineering*, volume 6627 of *Lecture Notes in Computer Science*, pages 244–258. Springer Berlin Heidelberg.

[Cook and Campbell 1979] Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin.

[Cormen et al. 2001] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. MIT Press, Cambridge, MA.

[Coutinho 2011] Coutinho, A. E. V. B. (2011). An experimental investigation of the selection order established by test suite reduction strategies. Technical report, Software Practices Laboratory, Federal University of Campina Grande.

[Coutinho 2012a] Coutinho, A. E. V. B. (2012a). An experimental investigation of distance functions for test suite reduction strategies based on similarity. Technical report, Software Practices Laboratory, Federal University of Campina Grande.

[Coutinho 2012b] Coutinho, A. E. V. B. (2012b). An experimental investigation of the selection order established by test suite reduction strategies in two real-world specifications. Technical report, Software Practices Laboratory, Federal University of Campina Grande.

[Coutinho 2012c] Coutinho, A. E. V. B. (2012c). Experimental investigation of the selection order of test cases established by test suite reduction strategies. Technical report, Software Practices Laboratory, Federal University of Campina Grande.

[Coutinho 2013] Coutinho, A. E. V. B. (2013). Defining a hybrid strategy for the choice of a test suite reduction strategy. Technical report, Software Practices Laboratory, Federal University of Campina Grande.

[Coutinho et al. 2013] Coutinho, A. E. V. B., Cartaxo, E. G., and de Lima Machado, P. D. (2013). Test suite reduction based on similarity of test cases. In *SAST 2013*.

[Coutinho et al. 2014] Coutinho, A. E. V. B., Cartaxo, E. G., and de Lima Machado, P. D. (2014). Analysis of distance functions for similarity-based test suite reduction in the context of model-based testing. *Software Quality Journal*, pages 1–39.

[da Silva Simao et al. 2006] da Silva Simao, A., de Mello, R., and Senger, L. (2006). A technique to reduce the test case suites for regression testing based on a self-organizing neural network architecture. In *Computer Software and Applications Conference, 2006. COMPSAC '06. 30th Annual International*, volume 2, pages 93–96.

[de Vries and Tretmans 2000] de Vries, R. G. and Tretmans, J. (2000). On-the-fly conformance testing using SPIN. *International Journal on Software Tools for Technology Transfer*, 2(4):382–393.

[Fang et al. 2013] Fang, C., Chen, Z., Wu, K., and Zhao, Z. (2013). Similarity-based test case prioritization using ordered sequences of program entities. *Software Quality Journal*, pages 1–27.

[Felipe et al. 2006] Felipe, J. C., Marques, P. M. A., Balan, A. G. R., Traina, C. J., and Traina, A. J. M. (2006). Comparing images with distance functions based on attribute interaction. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC'06, pages 1398–1399, New York, NY, USA. ACM.

[Felipe et al. 2003] Felipe, J. C., Traina, A. J. M., and Jr., C. T. (2003). Retrieval by content of medical images using texture for tissue identification. In *CBMS*, pages 175–180. IEEE Computer Society.

[Ferreira et al. 2010] Ferreira, F., Neves, L., Silva, M., and Borba, P. (2010). TaRGeT: a model based product line testing tool. In *CBSOFT 2010: Tools Session*.

[Fraser and Wotawa 2007] Fraser, G. and Wotawa, F. (2007). Redundancy based test-suite reduction. In *Proceedings of the 10th international conference on Fundamental approaches to software engineering*, FASE'07, pages 291–305, Berlin, Heidelberg. Springer-Verlag.

[Harrold et al. 1993] Harrold, M. J., Gupta, R., and Soffa, M. L. (1993). A methodology for controlling the size of a test suite. *ACM Trans. Softw. Eng. Methodol.*, 2(3):270–285.

[Heimdahl and George 2004] Heimdahl, M. and George, D. (2004). Test-suite reduction for model based tests: effects on test quality and implications for testing. In *Automated Software Engineering, 2004. Proceedings. 19th International Conference on*, pages 176–185.

[Hemmati et al. 2013] Hemmati, H., Arcuri, A., and Briand, L. (2013). Achieving scalable model-based testing through test case diversity. *ACM Transactions Software Engineering Methodology*, 22(1):1–42.

[Hemmati and Briand 2010] Hemmati, H. and Briand, L. (2010). An industrial investigation of similarity measures for model-based test case selection. In *Software Reliability Engineering (ISSRE), 2010 IEEE 21st International Symposium on*, pages 141–150.

[Heß2006] Heß, A. (2006). An iterative algorithm for ontology mapping capable of using training data. In *Proceedings of the 3rd European Conference on The Semantic Web: Research and Applications*, ESWC'06, pages 19–33, Berlin, Heidelberg. Springer-Verlag.

[Ho et al. 1999] Ho, W. M., Jezequel, J.-M., Le Guennec, A., and Pennaneac'h, F. (1999). UMLAUT: an extendible UML transformation framework. In *Automated Software Engineering, 1999. 14th IEEE International Conference on.*, pages 275–278.

[Hsu and Orso 2009] Hsu, H.-Y. and Orso, A. (2009). MINTS: A general framework and tool for supporting test-suite minimization. In *Proceedings of the 31st International Conference on Software Engineering*, ICSE '09, pages 419–429, Washington, DC, USA. IEEE Computer Society.

[Jaccard 1901] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles 37*, pages 547–579.

[Jain 1991] Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling*. John Wiley.

[Jard and Jéron 2005] Jard, C. and Jéron, T. (2005). TGV: theory, principles and algorithms. *International Journal on Software Tools for Technology Transfer*, 7(4):297–315.

[Jaro 1989] Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

[Jeffrey and Gupta 2007] Jeffrey, D. and Gupta, N. (2007). Improving fault detection capability by selectively retaining test cases during test suite reduction. *IEEE Transactions on Software Engineering*, 33(2):108–123.

[Jourdan et al. 2006] Jourdan, G.-V., Ritthiruangdech, P., and Ural, H. (2006). Test suite reduction based on dependence analysis. In *Proceedings of the 21st international conference on Computer and Information Sciences*, ISCIS'06, pages 1021–1030, Berlin, Heidelberg. Springer-Verlag.

[Khalilian and Parsa 2012] Khalilian, A. and Parsa, S. (2012). Bi-criteria test suite reduction by cluster analysis of execution profiles. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7054 LNCS:243–256.

[Kovács et al. 2009] Kovács, G., Németh, G., Subramaniam, M., and Pap, Z. (2009). Optimal string edit distance based test suite reduction for sdl specifications. In *SDL 2009: Design for Motes and Mobiles*, volume 5719 of *Lecture Notes in Computer Science*, pages 82–97. Springer Berlin Heidelberg.

[Ledru et al. 2009] Ledru, Y., Petrenko, A., and Boroday, S. (2009). Using string distances for test case prioritisation. In *Proceedings of the 2009 IEEE/ACM International Confer-*

ence on Automated Software Engineering, ASE '09, pages 510–514, Washington, DC, USA. IEEE Computer Society.

[Levenshtein 1966] Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

[Lin and Huang 2009] Lin, J.-W. and Huang, C.-Y. (2009). Analysis of test suite reduction with enhanced tie-breaking techniques. *Information and Software Technology*, 51(4):679–690.

[Lin et al. 2008] Lin, J.-W., Huang, C.-Y., and Lin, C.-T. (2008). Test suite reduction analysis with enhanced tie-breaking techniques. In *Management of Innovation and Technology, 2008. ICMIT 2008. 4th IEEE International Conference on*, pages 1228–1233.

[Liu 2014] Liu, P. (2014). An efficient reduction approach to test suite. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2014 15th IEEE/ACIS International Conference on*, pages 1–5.

[Nogueira et al. 2007] Nogueira, S., Cartaxo, E., Torres, D., Aranha, E., and Marques, R. (2007). Model based test generation: An industrial experience. In *1st Brazilian Workshop on Systematic and Automated Software Testing - SBBD/SBES 2007*, João Pessoa, PB, Brazil.

[Oliveira Neto et al. 2013] Oliveira Neto, F. G., Feldt, R., Torkar, R., and Machado, P. D. L. (2013). Searching for models to test software technology. In *Proceedings of First International Workshop on Combining Modelling and Search-Based Software Engineering, CMSBSE/ICSE'2013*.

[Parsa and Khalilian 2009] Parsa, S. and Khalilian, A. (2009). A bi-objective model inspired greedy algorithm for test suite minimization. In *Proceedings of the 1st International Conference on Future Generation Information Technology*, FGIT '09, pages 208–215, Berlin, Heidelberg. Springer-Verlag.

[Parsa et al. 2009] Parsa, S., Khalilian, A., and Fazlalizadeh, Y. (2009). A new algorithm to test suite reduction based on cluster analysis. In *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, pages 189–193.

[Pezzè and Young 2007] Pezzè, M. and Young, M. (2007). *Software Testing and Analysis: Process, Principles and Techniques*. Wiley.

[Pretschner 2005] Pretschner, A. (2005). Model-based testing. In *Proceedings of the 27th international conference on Software engineering*, ICSE '05, pages 722–723, New York, NY, USA. ACM.

[Renieres and Reiss 2003] Renieres, M. and Reiss, S. (2003). Fault localization with nearest neighbor queries. In *Automated Software Engineering, 2003. Proceedings. 18th IEEE International Conference on*, pages 30–39.

[Rogstad et al. 2013] Rogstad, E., Briand, L., and Torkar, R. (2013). Test Case Selection for Black-box Regression Testing of Database Applications. *Information and Software Technology*, 55(10):1781–1795.

[Rothermel et al. 2002] Rothermel, G., Harrold, M. J., Ronne, J. V., and Hong, C. (2002). Empirical studies of test-suite reduction. *Journal of Software Testing, Verification, and Reliability*, 12:219–249.

[Sellers 1980] Sellers, P. H. (1980). The theory and computation of evolutionary distances: Pattern recognition. *Journal of Algorithms*, 1(4):359–373.

[Selvakumar et al. 2010a] Selvakumar, S., Dinesh, M., Dhineshkumar, C., and Ramaraj, N. (2010a). Reducing the size of the test suite by genetic algorithm and concept analysis. *Communications in Computer and Information Science*, 90 CCIS:153–161.

[Selvakumar et al. 2010b] Selvakumar, S., Dinesh, M., Dhineshkumar, C., and Ramaraj, N. (2010b). Test suite diminuition using GRE heuristic with selective redundancy approach. *Communications in Computer and Information Science*, 90 CCIS:563–571.

[Tallam and Gupta 2005] Tallam, S. and Gupta, N. (2005). A concept analysis inspired greedy algorithm for test suite minimization. *SIGSOFT Software Engineering Notes*, 31(1):35–42.

[Thakur and Sahayam 2013] Thakur, A. S. and Sahayam, N. (2013). Speech recognition using euclidean distance. *International Journal of Emerging Technology and Advanced Engineering*, 3(2):587–590.

[Tretmans 2008] Tretmans, J. (2008). Model based testing with labelled transition systems. In Hierons, R. M., Bowen, J. P., and Harman, M., editors, *Formal Methods and Testing*, pages 1–38. Springer-Verlag, Berlin, Heidelberg.

[Utting and Legeard 2007] Utting, M. and Legeard, B. (2007). *Practical Model-Based Testing: A Tools Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Vargha and Delaney 2000] Vargha, A. and Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132.

[Vinson et al. 2007] Vinson, A. R., Heuser, C. A., da Silva, A. S., and de Moura, E. S. (2007). An approach to xml path matching. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management*, WIDM '07, pages 17–24, New York, NY, USA. ACM.

[Winkler 1999] Winkler, W. E. (1999). The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Census Bureau.

[Wohlin et al. 2012] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer.

[Xie et al. 2013] Xie, X., Chen, T. Y., Kuo, F.-C., and Xu, B. (2013). A theoretical analysis of the risk evaluation formulas for spectrum-based fault localization. *ACM Trans. Softw. Eng. Methodol.*, 22(4):31:1–31:40.

[Xu et al. 2012] Xu, S., Miao, H., and Gao, H. (2012). Test suite reduction using weighted set covering techniques. In *Software Engineering, Artificial Intelligence, Networking and Parallel Distributed Computing (SNPD), 2012 13th ACIS International Conference on*, pages 307–312.

[Yoo and Harman 2012] Yoo, S. and Harman, M. (2012). Regression testing minimization, selection and prioritization: A survey. *Software Testing, Verification and Reliability*, 22(2):67–120.

[Zhang et al. 2010] Zhang, X., Gu, Q., Chen, X., Qi, J., and Chen, D. (2010). A study of relative redundancy in test-suite reduction while retaining or improving fault-localization effectiveness. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC'10, pages 2229–2236, New York, NY, USA. ACM.

[Zhong et al. 2006] Zhong, H., Zhang, L., and Mei, H. (2006). An experimental comparison of four test suite reduction techniques. In *Proceedings of the 28th international conference on Software engineering*, ICSE '06, pages 636–640, New York, NY, USA. ACM.

[Zhong et al. 2008] Zhong, H., Zhang, L., and Mei, H. (2008). An experimental study of four typical test suite reduction techniques. *Information and Software Technology*, 50(6):534–546.

Appendix A

Results of Statistical Tests for the Evaluation of the Similary-based Test Suite Reduction Strategy

This Appendix contains data regarding stastitical tests for the empirical studies presented in Chapter 5.

A.1 Configuration

Table A.1: *Basic configuration for CB*

Specification	#TCs	Essentials Test Cases				#Faults	%Faults	#Failures	%Failures
		#All-transitions	%All-transitions	#All-transition-pairs	%All-transition-pairs				
cb	69	8	11.594	15	21.739	12	17.391	12	17.391
001	122	10	8.197	11	9.016	17	13.934	21	17.213
002	131	7	5.344	12	9.160	22	16.794	23	17.557
003	120	11	9.167	12	10.000	17	14.167	19	15.833
004	134	7	5.224	8	5.970	19	14.179	23	17.164
005	131	7	5.344	15	11.450	18	13.740	21	16.031
006	134	7	5.224	8	5.970	19	14.179	21	15.672
007	130	7	5.385	11	8.462	18	13.846	22	16.923
008	125	7	5.600	14	11.200	18	14.400	20	16.000
009	129	7	5.426	8	6.202	18	13.953	20	15.504
010	128	10	7.813	13	10.156	18	14.063	22	17.188
011	134	7	5.224	9	6.716	19	14.179	23	17.164
012	122	11	9.016	18	14.754	17	13.934	17	13.934
013	128	10	7.813	12	9.375	18	14.063	23	17.969
014	128	10	7.813	11	8.594	18	14.063	20	15.625
015	132	7	5.303	10	7.576	19	14.394	20	15.152
016	134	7	5.224	8	5.970	19	14.179	23	17.164
017	124	10	8.065	11	8.871	18	14.516	23	18.548
018	130	7	5.385	12	9.231	18	13.846	21	16.154
019	125	10	8.000	13	10.400	18	14.400	19	15.200
020	134	7	5.224	8	5.970	19	14.179	21	15.672
021	134	7	5.224	11	8.209	20	14.925	21	15.672
022	131	7	5.344	9	6.870	18	13.740	20	15.267
023	133	7	5.263	10	7.519	19	14.286	22	16.541
024	126	11	8.730	14	11.111	18	14.286	20	15.873
025	130	7	5.385	13	10.000	18	13.846	19	14.615
026	130	7	5.385	8	6.154	18	13.846	22	16.923
027	131	7	5.344	12	9.160	18	13.740	22	16.794
028	122	11	9.016	13	10.656	18	14.754	21	17.213
029	118	13	11.017	14	11.864	12	10.169	13	11.017
030	126	11	8.730	12	9.524	18	14.286	18	14.286

Table A.2: *Basic configuration for PDFSam*

Specification	#TCs	Essentials Test Cases				#Faults	%Faults	#Failures	%Failures
		#All-transitions	%All-transitions	#All-transition-pairs	%All-transition-pairs				
pdfsam	137	0	0.000	0	0.000	5	3.650	5	3.650
001	100	5	5.000	7	7.000	3	3.000	3	3.000
002	125	0	0.000	1	0.800	4	3.200	4	3.200
003	237	0	0.000	0	0.000	8	3.376	8	3.376
004	124	4	3.226	4	3.226	4	3.226	4	3.226
005	103	4	3.883	4	3.883	3	2.913	3	2.913
006	122	0	0.000	0	0.000	4	3.279	4	3.279
007	181	0	0.000	0	0.000	6	3.315	6	3.315
008	133	0	0.000	0	0.000	4	3.008	4	3.008
009	117	0	0.000	0	0.000	4	3.419	4	3.419
010	150	0	0.000	0	0.000	5	3.333	5	3.333
011	110	1	0.909	2	1.818	4	3.636	4	3.636
012	103	0	0.000	4	3.883	3	2.913	3	2.913
013	150	1	0.667	1	0.667	5	3.333	5	3.333
014	97	0	0.000	4	4.124	3	3.093	3	3.093
015	104	8	7.692	8	7.692	3	2.885	3	2.885
016	120	2	1.667	2	1.667	4	3.333	4	3.333
017	138	0	0.000	0	0.000	5	3.623	5	3.623
018	145	0	0.000	0	0.000	5	3.448	5	3.448
019	189	0	0.000	1	0.529	6	3.175	6	3.175
020	155	0	0.000	4	2.581	5	3.226	5	3.226
021	116	3	2.586	3	2.586	4	3.448	4	3.448
022	149	4	2.685	4	2.685	5	3.356	5	3.356
023	119	0	0.000	3	2.521	4	3.361	4	3.361
024	105	0	0.000	3	2.857	3	2.857	3	2.857
025	178	0	0.000	0	0.000	6	3.371	6	3.371
026	171	0	0.000	0	0.000	6	3.509	6	3.509
027	117	2	1.709	2	1.709	4	3.419	4	3.419
028	154	0	0.000	0	0.000	5	3.247	5	3.247
029	148	1	0.676	1	0.676	5	3.378	5	3.378
030	181	1	0.552	1	0.552	6	3.315	6	3.315

Table A.3: *Basic configuration for TaRGeT*

Specification	#TCs	Essentials Test Cases				#Faults	%Faults	#Failures	%Failures
		#All-transitions	%All-transitions	#All-transition-pairs	%All-transition-pairs				
target	82	39	47.561	62	75.610	13	15.854	13	15.854
001	88	46	52.273	67	76.136	13	14.773	13	14.773
002	99	33	33.333	47	47.475	15	15.152	15	15.152
003	87	38	43.678	54	62.069	13	14.943	13	14.943
004	96	32	33.333	48	50.000	15	15.625	15	15.625
005	97	36	37.113	49	50.515	15	15.464	15	15.464
006	115	26	22.609	45	39.130	18	15.652	18	15.652
007	115	30	26.087	50	43.478	18	15.652	18	15.652
008	88	43	48.864	57	64.773	13	14.773	13	14.773
009	103	29	28.155	44	42.718	16	15.534	16	15.534
010	86	44	51.163	64	74.419	13	15.116	13	15.116
011	94	26	27.660	55	58.511	14	14.894	14	14.894
012	79	37	46.835	61	77.215	12	15.190	12	15.190
013	90	33	36.667	53	58.889	14	15.556	14	15.556
014	81	37	45.679	65	80.247	12	14.815	12	14.815
015	86	32	37.209	58	67.442	13	15.116	13	15.116
016	84	44	52.381	63	75.000	14	16.667	14	16.667
017	93	37	39.785	54	58.065	14	15.054	14	15.054
018	83	35	42.169	61	73.494	13	15.663	13	15.663
019	86	34	39.535	57	66.279	13	15.116	13	15.116
020	89	37	41.573	52	58.427	14	15.730	14	15.730
021	103	27	26.214	46	44.660	16	15.534	16	15.534
022	87	28	32.184	57	65.517	13	14.943	13	14.943
023	83	39	46.988	61	73.494	13	15.663	13	15.663
024	80	41	51.250	64	80.000	12	15.000	12	15.000
025	94	41	43.617	56	59.574	14	14.894	14	14.894
026	88	47	53.409	59	67.045	13	14.773	13	14.773
027	150	22	14.667	35	23.333	23	15.333	23	15.333
028	88	38	43.182	54	61.364	13	14.773	13	14.773
029	76	43	56.579	74	97.368	12	15.789	11	14.474
030	99	38	38.384	57	57.576	15	15.152	15	15.152

A.2 Normality test

Table A.4: *Anderson-Darling normality test for CB real*

Strategy	ρ-value					
	SSR			FC		
	T	P	B	T	P	B
G	7.071e-130	9.469e-166	1.395e-127	4.342e-65	1.077e-85	1.176e-71
GE	9.937e-139	3.188e-118	3.334e-131	2.732e-71	2.002e-79	1.806e-72
GRE	3.447e-52	2.59e-108	1.923e-177	2.951e-103	5.613e-156	6.776e-121
HGS	1.296e-139	6.656e-132	1.241e-92	4.842e-61	4.909e-70	2.925e-66
Sim	NaN	NaN	1.013e-142	1.667e-148	2.195e-144	5.129e-163

Table A.5: *Anderson-Darling normality test for CB synthetics*

Strategy	ρ-value					
	SSR			FC		
	T	P	B	T	P	B
G	∞	7.796e-169	4.886e-177	1.133e-159	6.585e-173	5.997e-172
GE	∞	4.483e-83	5.992e-177	7.007e-162	3.216e-174	5.31e-171
GRE	∞	6.376e-79	2.551e-190	5.016e-175	1.405e-182	5.035e-179
HGS	2.1e-185	7.426e+18	1.166e-91	4.666e-176	1.8e-182	7.012e-177
Sim	3.187e+40	4.917e-100	1.013e-185	1.842e-178	1.201e+243	2.459e+98

Table A.6: *Anderson-Darling normality test for PDFSam real*

Strategy	ρ-value					
	SSR			FC		
	T	P	B	T	P	B
G	NaN	NaN	3.608e-149	1.813e-176	1.023e-184	4.297e-164
GE	NaN	NaN	9.343e-150	5.028e-176	2.357e-180	2.597e-162
GRE	NaN	NaN	3.635e-183	4.813e-172	2.194e-179	7.909e-154
HGS	2.337e-170	NaN	3.261e-160	2.152e-190	5.39e-190	2.202e-186
Sim	2.163e-182	9.106e-171	9.952e-177	2.453e-173	7.042e-183	7.728e-123

Table A.7: *Anderson-Darling normality test for PDFSam synthetics*

Strategy	ρ-value					
	SSR			FC		
	T	P	B	T	P	B
G	∞	2.602e+12	6.479e-58	∞	∞	∞
GE	∞	497500000	1.385e-44	∞	∞	∞
GRE	∞	0.0007434	1.579e-52	∞	∞	∞
HGS	∞	1.614e-73	2.551e-50	∞	∞	∞
Sim	∞	3.943e-62	4.31e-183	5.325e+124	∞	∞

Table A.8: *Anderson-Darling normality test for TaRGeT real*

Strategy	ρ-value					
	SSR			FC		
	T	P	B	T	P	B
G	NaN	NaN	NaN	8.868e-185	6.784e-185	1.067e-184
GE	NaN	NaN	NaN	7.291e-185	7.422e-185	7.92e-185
GRE	NaN	NaN	NaN	7.57e-185	7.422e-185	2.039e-184
HGS	8.062e-60	5.407e-175	1.694e-161	4.674e-171	1.151e-188	1.005e-173
Sim	8.868e-185	NaN	9.851e-185	3.418e-184	1.881e-162	1.87e+241

Table A.9: *Anderson-Darling normality test for TaRGeT synthetics*

Strategy	ρ-value					
	SSR			FC		
	T	P	B	T	P	B
G	∞	∞	∞	7.146e-156	1.954e-179	2.058e-174
GE	∞	∞	∞	4.887e-163	2.578e-187	9.121e-185
GRE	∞	∞	∞	4.784e-165	1.997e-184	2.032e-182
HGS	∞	∞	∞	7.552e-156	1.248e-185	3.969e-183
Sim	∞	∞	∞	3.159e+68	1.286e+180	∞

A.3 Kruskal-Wallis test

A.3.1 Study Question 1

Table A.10: *Krukal-Wallis test for SQ1*

Specification	Comparison	ρ-value	
		SSR	FC
CB real	$G_T = G_P = G_B$	0.000	1.564e-82
	$GE_T = GE_P = GE_B$	0.000	1.564e-82
	$GRE_T = GRE_P = GRE_B$	0.000	1.564e-82
	$HGS_T = HGS_P = HGS_B$	0.000	1.564e-82
	$Sim_T = Sim_P = Sim_B$	0.000	1.564e-82
CB synthetics	$G_T = G_P = G_B$	0.000	0.000
	$GE_T = GE_P = GE_B$	0.000	0.000
	$GRE_T = GRE_P = GRE_B$	0.000	0.000
	$HGS_T = HGS_P = HGS_B$	0.000	0.000
	$Sim_T = Sim_P = Sim_B$	0.000	0.000
PDFSam real	$G_T = G_P = G_B$	0.000	1.792e-05
	$GE_T = GE_P = GE_B$	0.000	1.792e-05
	$GRE_T = GRE_P = GRE_B$	0.000	1.792e-05
	$HGS_T = HGS_P = HGS_B$	0.000	1.792e-05
	$Sim_T = Sim_P = Sim_B$	0.000	1.792e-05
PDFSam synthetics	$G_T = G_P = G_B$	0.000	0.000
	$GE_T = GE_P = GE_B$	0.000	0.000
	$GRE_T = GRE_P = GRE_B$	0.000	0.000
	$HGS_T = HGS_P = HGS_B$	0.000	0.000
	$Sim_T = Sim_P = Sim_B$	0.000	0.000
TaRGeT real	$G_T = G_P = G_B$	0.000	7.1e-296
	$GE_T = GE_P = GE_B$	0.000	7.1e-296
	$GRE_T = GRE_P = GRE_B$	0.000	7.1e-296
	$HGS_T = HGS_P = HGS_B$	0.000	7.1e-296
	$Sim_T = Sim_P = Sim_B$	0.000	7.1e-296
TaRGeT synthetics	$G_T = G_P = G_B$	0.000	0.000
	$GE_T = GE_P = GE_B$	0.000	0.000
	$GRE_T = GRE_P = GRE_B$	0.000	0.000
	$HGS_T = HGS_P = HGS_B$	0.000	0.000
	$Sim_T = Sim_P = Sim_B$	0.000	0.000

A.3.2 Study Question 2

Table A.11: *Krukal-Wallis test for SQ2*

Specification	Comparison	ρ-value SSR	ρ-value FC
CB real	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000	2.834e-295
	$G_P = GE_P = GRE_P = HGS_P = Sim_P$	0.000	0.000
	$G_B = GE_B = GRE_B = HGS_B = Sim_B$	0.000	0.000
CB synthetics	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000	0.000
	$G_P = GE_P = GRE_P = HGS_P = Sim_P$	0.000	0.000
	$G_B = GE_B = GRE_B = HGS_B = Sim_B$	0.000	0.000
PDFSam real	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000	1.228e-17
	$G_P = GE_P = GRE_P = HGS_P = Sim_P$	0.000	1.766e-11
	$G_B = GE_B = GRE_B = HGS_B = Sim_B$	0.000	8.402e-246
PDFSam synthetics	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	1.346e-17	0.000
	$G_P = GE_P = GRE_P = HGS_P = Sim_P$	3.65e-94	0.000
	$G_B = GE_B = GRE_B = HGS_B = Sim_B$	5.728e-284	0.000
TaRGeT real	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000	8.019e-08
	$G_P = GE_P = GRE_P = HGS_P = Sim_P$	0.000	3.84e-20
	$G_B = GE_B = GRE_B = HGS_B = Sim_B$	0.000	3.194e-170
TaRGeT synthetics	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000	0.000
	$G_P = GE_P = GRE_P = HGS_P = Sim_P$	1.449e-10	0.000
	$G_B = GE_B = GRE_B = HGS_B = Sim_B$	9.232e-06	0.000

A.3.3 Study Question 3

Table A.12: *Krukal-Wallis test for SQ3*

Specification	Metric	Comparison	ρ-value
CB real	SSR	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000
	FC	$G_P = GE_P = GRE_P = HGS_P = Sim_B$	0.000
CB synthetics	SSR	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000
	FC	$G_B = GE_B = GRE_B = HGS_P = Sim_P$	0.000
PDFSam real	SSR	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000
	FC	$G_B = GE_B = GRE_B = HGS_B = Sim_B$	8.402e-246
PDFSam synthetics	SSR	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	1.346e-17
	FC	$G_P = GE_P = GRE_P = HGS_P = Sim_B$	0.000
TaRGeT real	SSR	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000
	FC	$G_P = GE_P = GRE_P = HGS_B = Sim_B$	6.3e-153
TaRGeT synthetics	SSR	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000
	FC	$G_B = GE_B = GRE_P = HGS_B = Sim_B$	0.000

A.3.4 Study Question 4

Table A.13: *Krukal-Wallis test for SQ4*

Specification	Metric	Comparison	ρ-value
CB real	SSR	$Sim_T = Sim_P = Sim_B$	0.000
	FC	$Sim_T = Sim_P = Sim_B$	6.993e-224
CB synthetics	SSR	$GRE_T = Sim_P = G_B$	0.000
	FC	$Sim_T = Sim_P = Sim_B$	0.000
PDFSam real	SSR	$G_T = GE_T = GRE_T = G_P = GE_P = GRE_P = GE_B$	0.000
	FC	$GRE_T = GRE_P = Sim_B$	1.695e-215
PDFSam synthetics	SSR	$GRE_T = G_P = G_B$	0.000
	FC	$Sim_T = Sim_P = Sim_B$	0.000
TaRGeT real	SSR	$G_T = GE_T = GRE_T = G_P = GE_P = GRE_P = Sim_P = G_B = GE_B = GRE_B$	0.000
	FC	$HGS_T = Sim_P = Sim_B$	0.000
TaRGeT synthetics	SSR	$GE_T = G_P = G_B$	0.000
	FC	$Sim_T = Sim_P = Sim_B$	0.000

A.4 Boxplot

A.4.1 Study Question 1

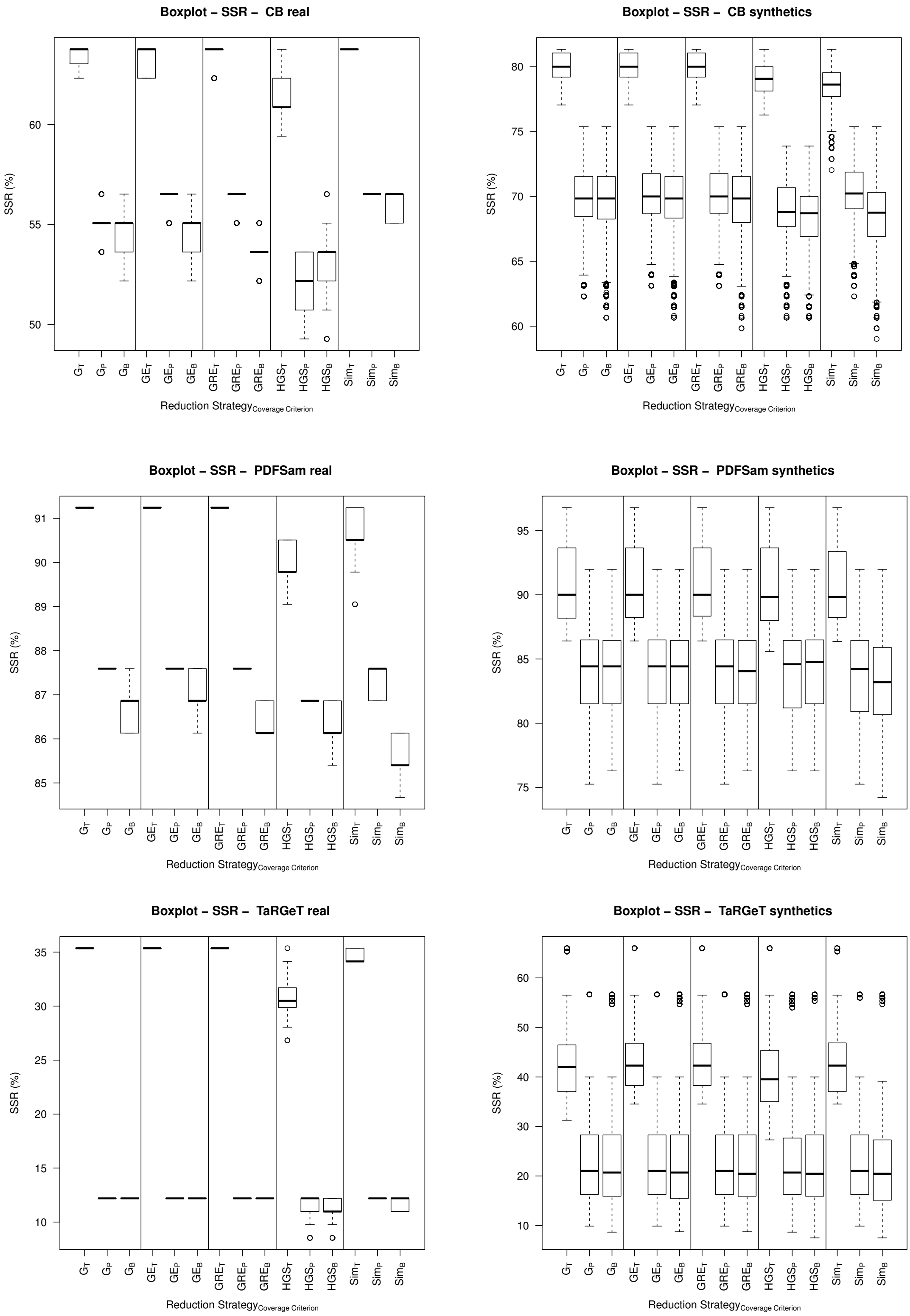


Figure A.1: *Boxplots considering SSR metric for SQ1*

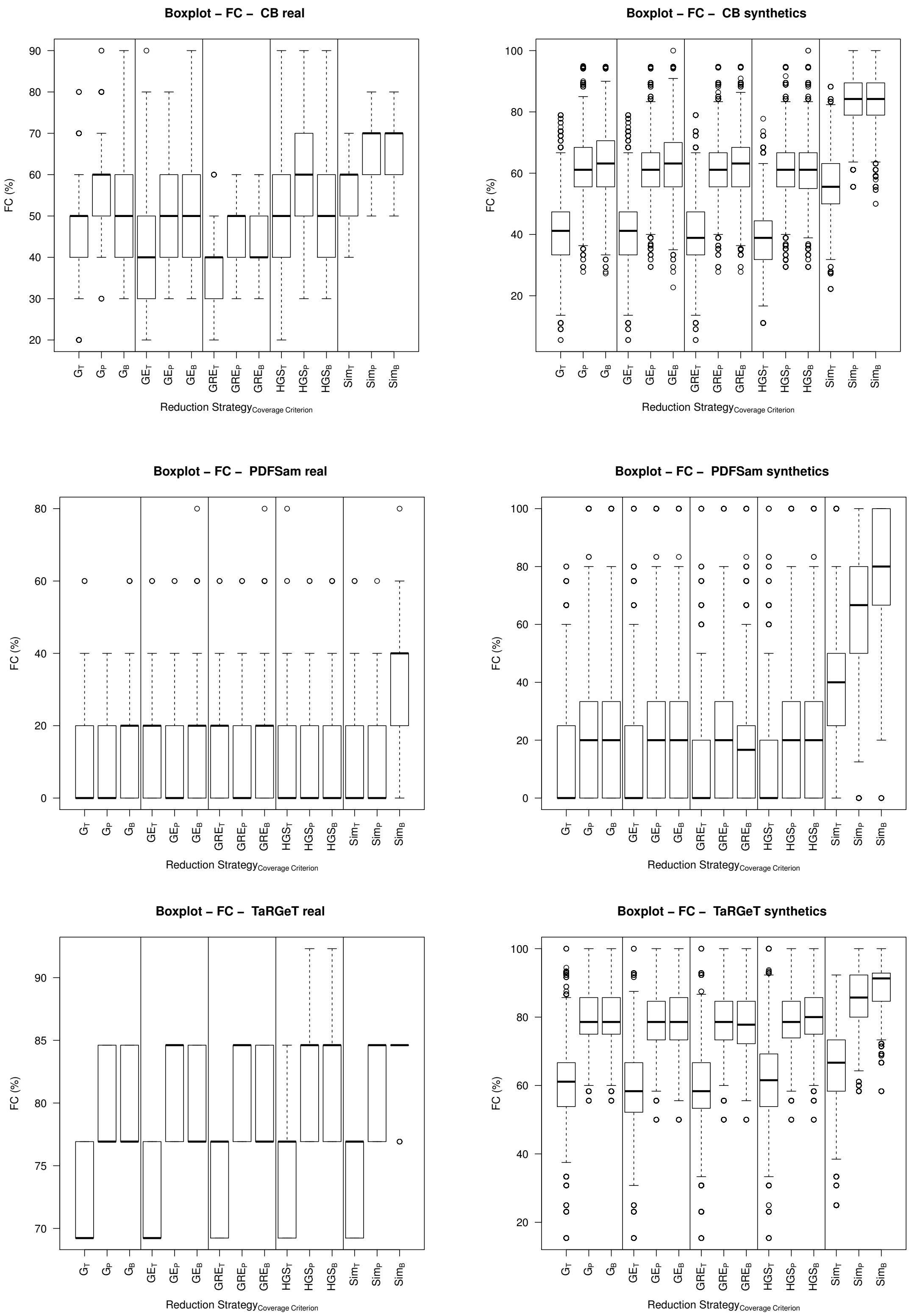


Figure A.2: *Boxplots considering FC metric for SQ1*

A.4.2 Study Question 2

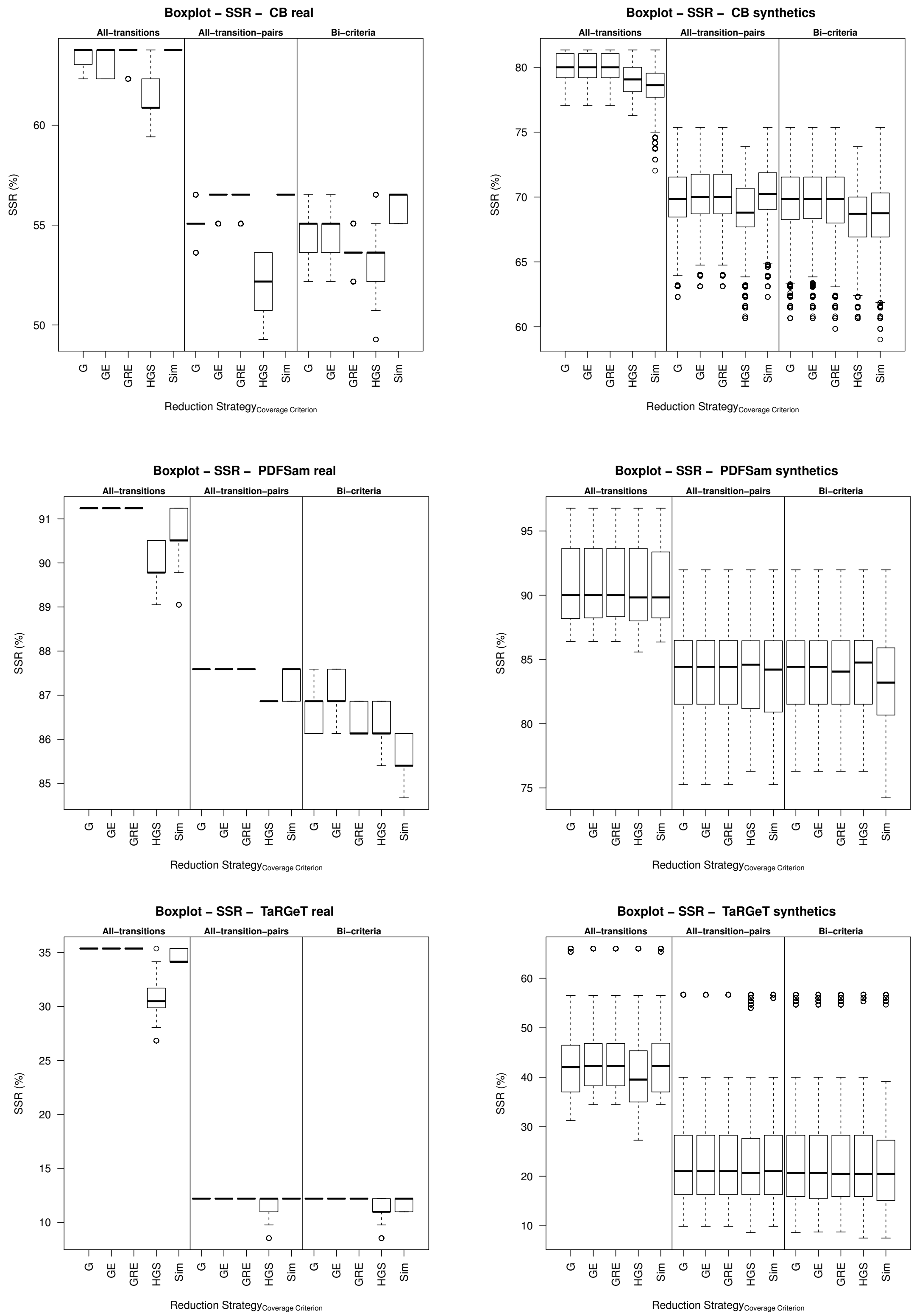


Figure A.3: *Boxplots considering* SSR *metric for SQ2*

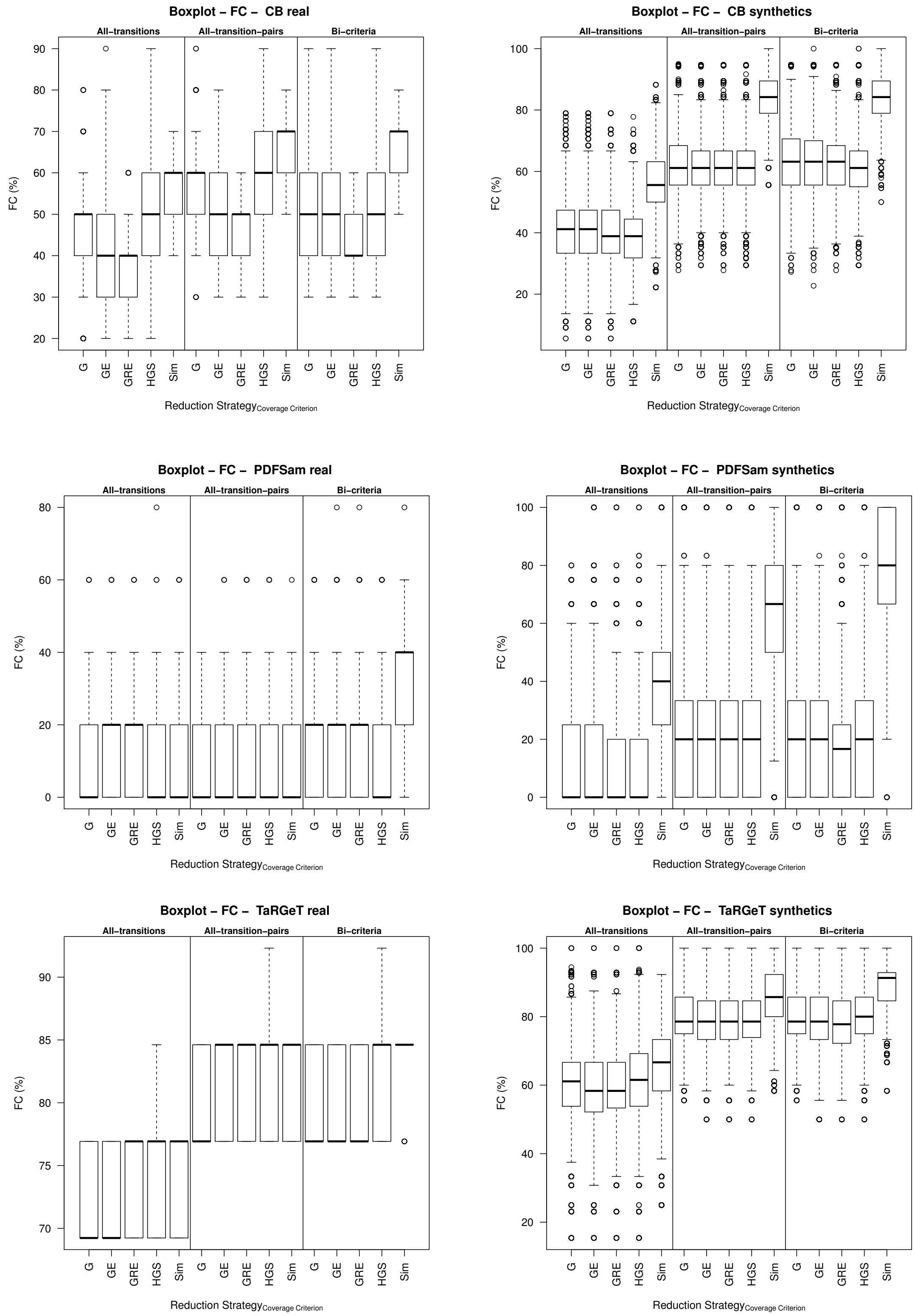


Figure A.4: *Boxplots considering FC metric for SQ2*

A.4.3 Study Question 3

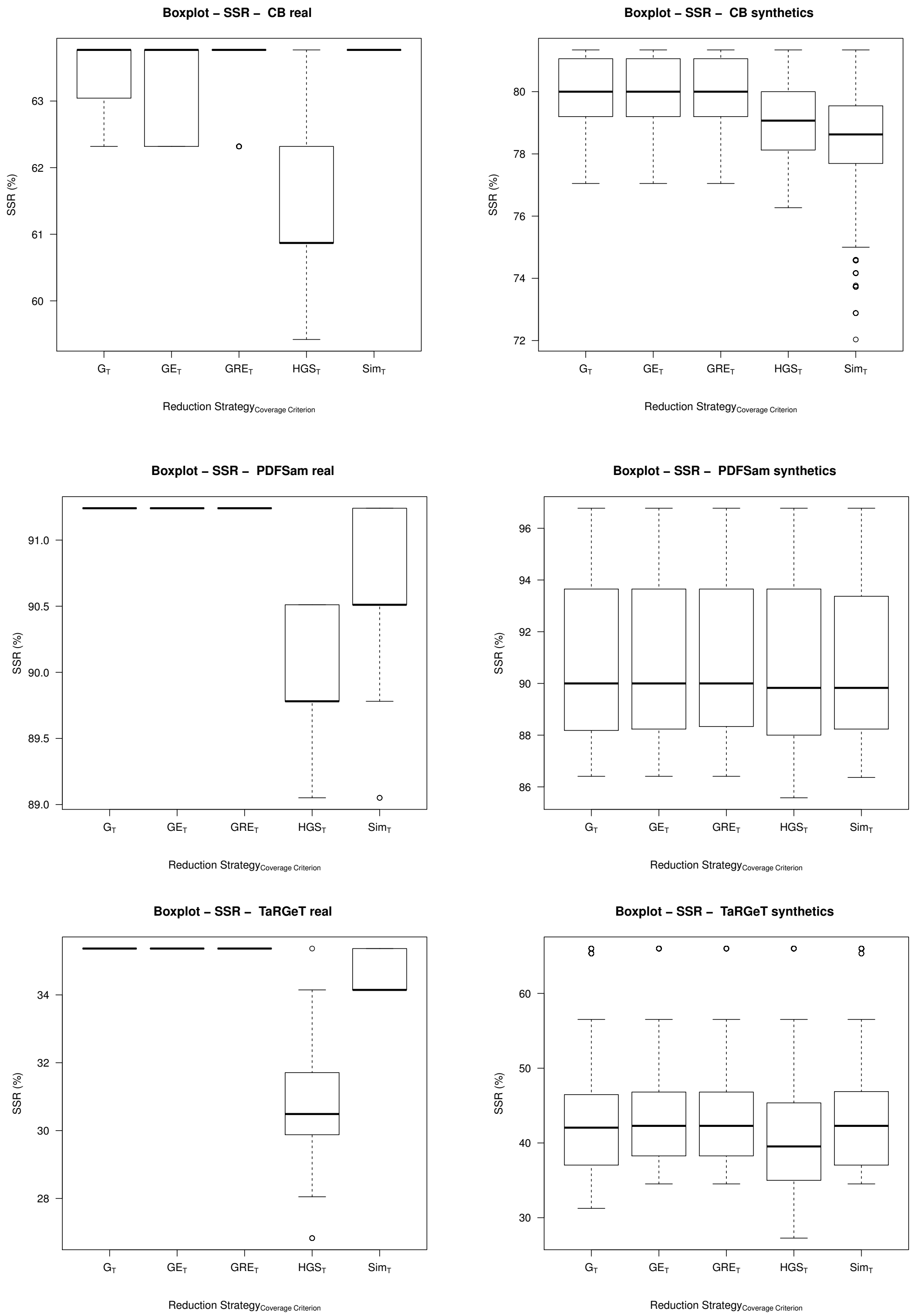


Figure A.5: *Boxplots considering SSR metric for SQ3*

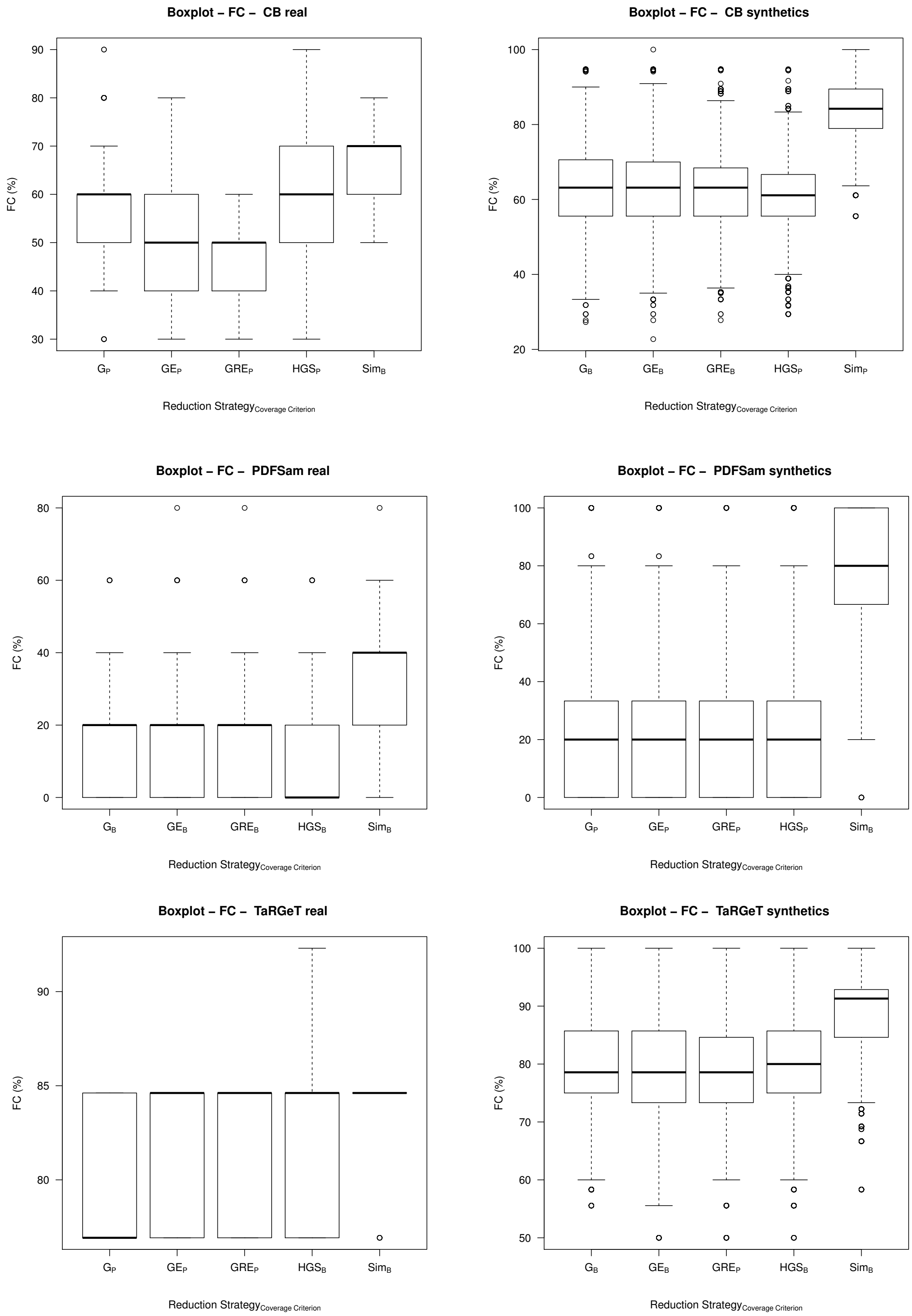


Figure A.6: *Boxplots considering FC metric for SQ3*

A.4.4 Study Question 4

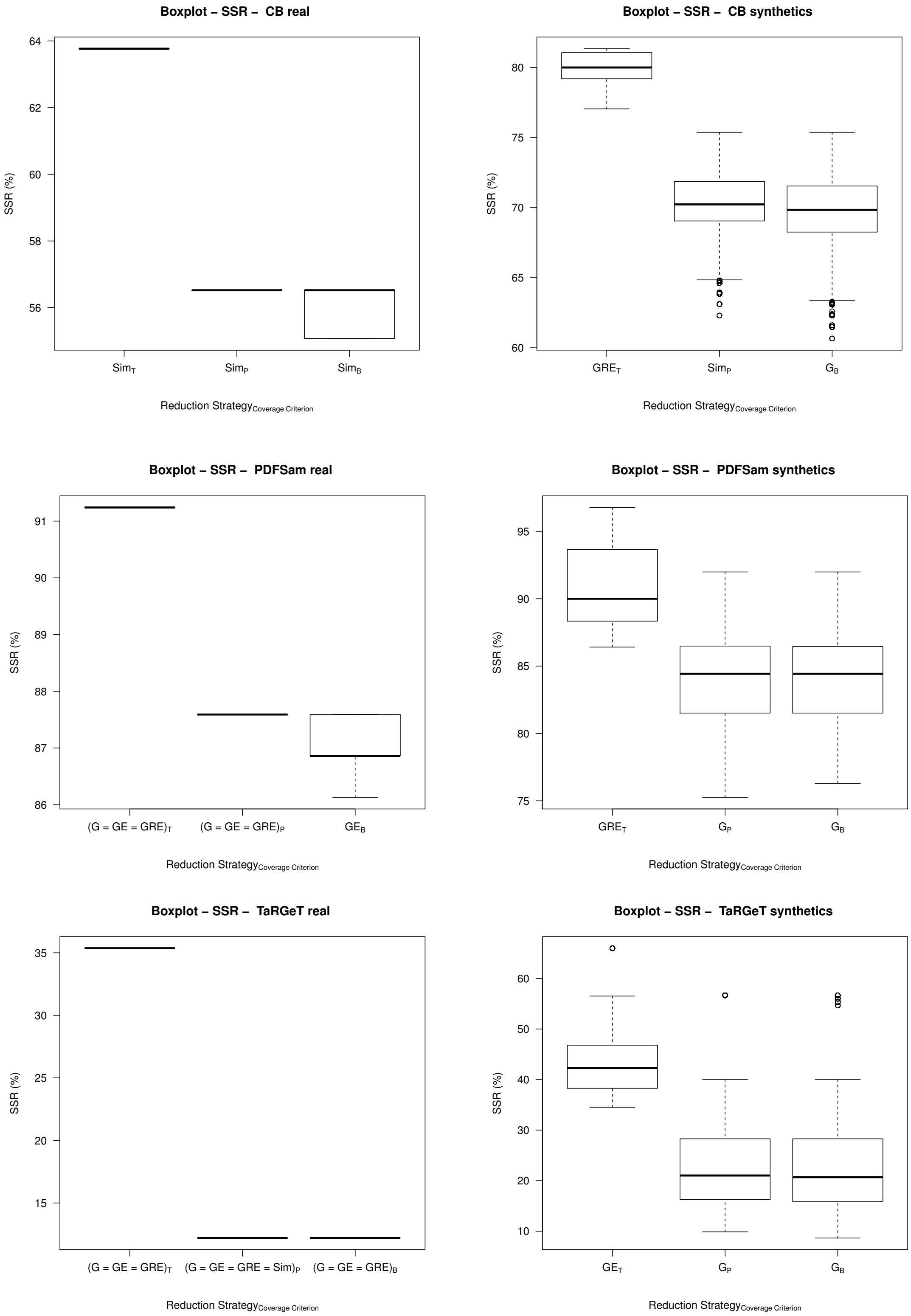


Figure A.7: *Boxplots considering SSR metric for SQ4*

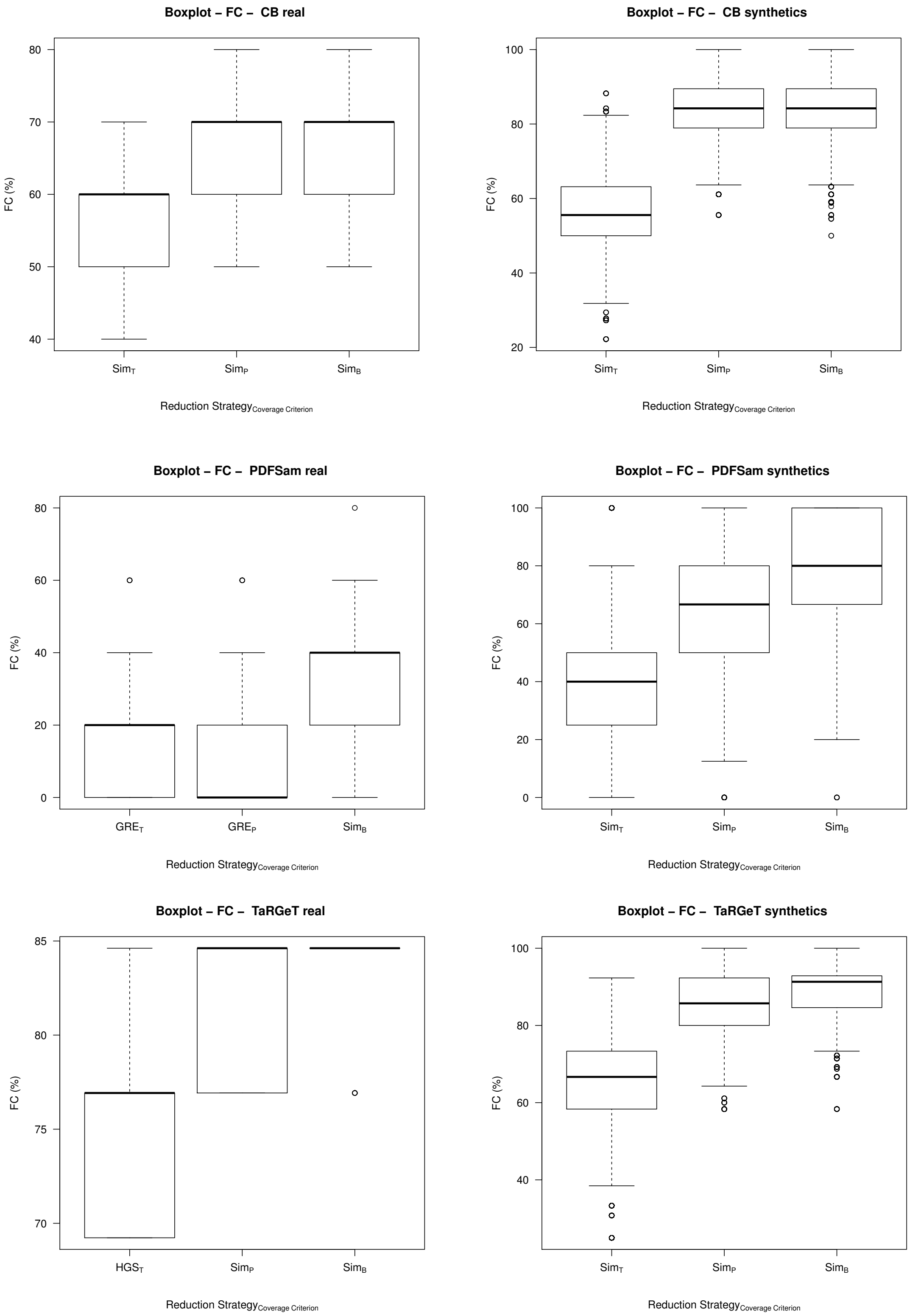


Figure A.8: *Boxplots considering FC metric for SQ4*

A.5 Mann-Whitney test and $\hat{A}_{12}$ effect size measurement

A.5.1 Study Question 1

Table A.14: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB real*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and G_P	2.787e-169	G_T	Large (1)	2.052e-73	G_P	Large (0.2608)
G_T and G_B	5.907e-168	G_T	Large (1)	2.738e-3	G_B	Medium (0.3567)
G_P and G_B	2.497e-38	G_P	Medium (0.6565)	1.431e-15	G_P	Medium (0.6023)
GE_T and GE_P	1.704e-173	GE_T	Large (1)	3.705e-61	GE_P	Large (0.2803)
GE_T and GE_B	8.651e-168	GE_T	Large (1)	1.784e-32	GE_B	Medium (0.3432)
GE_P and GE_B	8.726e-134	GE_P	Large (0.8892)	3.515e-09	GE_P	Small (0.5696)
GRE_T and GRE_P	1.376e-177	GRE_T	Large (1)	8.941e-96	GRE_P	Large (0.2099)
GRE_T and GRE_B	1.775e-172	GRE_T	Large (1)	6.186e-53	GRE_B	Large (0.3003)
GRE_P and GRE_B	2.672e-164	GRE_P	Large (0.9766)	5.2e-18	GRE_P	Medium (0.6013)
HGS_T and HGS_P	2.833e-166	HGS_T	Large (1)	1.068e-7	HGS_P	Large (0.2631)
HGS_T and HGS_B	7.183e-166	HGS_T	Large (1)	1.648e-28	HGS_B	Medium (0.3581)
HGS_P and HGS_B	6.94e-51	HGS_B	Large (0.3029)	3.376e-16	HGS_P	Medium (0.6058)
Sim_T and Sim_P	1.799e-219	Sim_T	Large (1)	2.045e-107	Sim_P	Large (0.1834)
Sim_T and Sim_B	1.139e-183	Sim_T	Large (1)	1.473e-122	Sim_B	Large (0.1464)
Sim_P and Sim_B	6.921e-61	Sim_P	Medium (0.6355)	0.000115	Sim_B	Small (0.4609)

Table A.15: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB synthetics*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and G_P	0.000	G_T	Large (1)	0.000	G_P	Large (0.06985)
G_T and G_B	0.000	G_T	Large (1)	0.000	G_B	Large (0.06501)
G_P and G_B	2.394e-37	G_P	Small (0.5006)	2.779e-41	G_B	Small (0.4699)
GE_T and GE_P	0.000	GE_T	Large (1)	0.000	GE_P	Large (0.07616)
GE_T and GE_B	0.000	GE_T	Large (1)	0.000	GE_B	Large (0.06709)
GE_P and GE_B	0.000	GE_P	Small (0.5145)	2.106e-91	GE_B	Small (0.4548)
GRE_T and GRE_P	0.000	GRE_T	Large (1)	0.000	GRE_P	Large (0.05693)
GRE_T and GRE_B	0.000	GRE_T	Large (1)	0.000	GRE_B	Large (0.05275)
GRE_P and GRE_B	0.000	GRE_P	Small (0.5208)	2.639e-58	GRE_B	Small (0.4659)
HGS_T and HGS_P	0.000	HGS_T	Large (1)	0.000	HGS_P	Large (0.04894)
HGS_T and HGS_B	0.000	HGS_T	Large (1)	0.000	HGS_B	Large (0.05192)
HGS_P and HGS_B	0.000	HGS_P	Small (0.5449)	1.982e-07	HGS_P	Small (0.5114)
Sim_T and Sim_P	0.000	Sim_T	Large (0.9981)	0.000	Sim_P	Large (0.01373)
Sim_T and Sim_B	0.000	Sim_T	Large (0.9998)	0.000	Sim_B	Large (0.01459)
Sim_P and Sim_B	0.000	Sim_P	Large (0.6717)	3.726e-06	Sim_P	Small (0.51)

Table A.16: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam real*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and G_P	1.799e-219	G_T	Large (1)	0.05014	G_T	Small (0.5193)
G_T and G_B	1.099e-171	G_T	Large (1)	0.004555	G_B	Small (0.4659)
G_P and G_B	3.892e-136	G_P	Large (0.878)	1.193e-06	G_B	Small (0.4464)
GE_T and GE_P	1.799e-219	GE_T	Large (1)	0.04026	GE_T	Small (0.5263)
GE_T and GE_B	8.748e-172	GE_T	Large (1)	0.02592	GE_B	Small (0.4788)
GE_P and GE_B	1.074e-133	GE_P	Large (0.869)	2.824e-05	GE_B	Small (0.4535)
GRE_T and GRE_P	1.799e-219	GRE_T	Large (1)	0.04307	GRE_T	Small (0.5229)
GRE_T and GRE_B	1.288e-176	GRE_T	Large (1)	0.0001771	GRE_B	Small (0.4624)
GRE_P and GRE_B	1.288e-176	GRE_P	Large (1)	9.816e-09	GRE_B	Small (0.4404)
HGS_T and HGS_P	6.354e-173	HGS_T	Large (1)	0.4599	HGS_T	Small (0.5046)
HGS_T and HGS_B	2.427e-169	HGS_T	Large (1)	0.006281	HGS_B	Small (0.4705)
HGS_P and HGS_B	2.448e-114	HGS_P	Large (0.8045)	0.0004722	HGS_B	Small (0.4657)
Sim_T and Sim_P	1.002e-169	Sim_T	Large (1)	0.0194	Sim_T	Small (0.5226)
Sim_T and Sim_B	2.483e-169	Sim_T	Large (1)	3.776e-111	Sim_B	Large (0.1749)
Sim_P and Sim_B	1.386e-172	Sim_P	Large (1)	1.653e-116	Sim_B	Large (0.1552)

Table A.17: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam synthetics*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and G_P	0.000	G_T	Large (0.9125)	0.000	G_P	Medium (0.3556)
G_T and G_B	0.000	G_T	Large (0.9182)	0.000	G_B	Medium (0.3624)
G_P and G_B	0.000	G_P	Small (0.5121)	4.788e-09	G_P	Small (0.508)
GE_T and GE_P	0.000	GE_T	Large (0.9165)	0.000	GE_P	Medium (0.3601)
GE_T and GE_B	0.000	GE_T	Large (0.9232)	0.000	GE_B	Medium (0.363)
GE_P and GE_B	0.000	GE_P	Small (0.5118)	0.5529	GE_P	Small (0.5028)
GRE_T and GRE_P	0.000	GRE_T	Large (0.917)	0.000	GRE_P	Medium (0.358)
GRE_T and GRE_B	0.000	GRE_T	Large (0.9228)	0.000	GRE_B	Medium (0.3881)
GRE_P and GRE_B	0.000	GRE_P	Small (0.5146)	1.263e-3	GRE_P	Small (0.5295)
HGS_T and HGS_P	0.000	HGS_T	Large (0.9085)	0.000	HGS_P	Medium (0.3573)
HGS_T and HGS_B	0.000	HGS_T	Large (0.9096)	0.000	HGS_B	Medium (0.3602)
HGS_P and HGS_B	0.000	HGS_B	Small (0.493)	0.0063	HGS_P	Small (0.5045)
Sim_T and Sim_P	0.000	Sim_T	Large (0.9255)	0.000	Sim_P	Large (0.2272)
Sim_T and Sim_B	0.000	Sim_T	Large (0.937)	0.000	Sim_B	Large (0.1144)
Sim_P and Sim_B	0.000	Sim_P	Small (0.542)	0.000	Sim_B	Medium (0.3382)

Table A.18: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT real*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and G_P	1.799e-219	G_T	Large (1)	1.928e-128	G_P	Large (0.1212)
G_T and G_B	1.799e-219	G_T	Large (1)	1.567e-13	G_B	Large (0.1261)
G_P and G_B	NaN	None	NO effect (0.5)	0.3702	G_P	Small (0.51)
GE_T and GE_P	1.799e-219	GE_T	Large (1)	4.231e-129	GE_P	Large (0.1203)
GE_T and GE_B	1.799e-219	GE_T	Large (1)	4.215e-128	GE_B	Large (0.1259)
GE_P and GE_B	NaN	None	NO effect (0.5)	0.3081	GE_P	Small (0.5115)
GRE_T and GRE_P	1.799e-219	GRE_T	Large (1)	1.313e-125	GRE_P	Large (0.1252)
GRE_T and GRE_B	1.799e-219	GRE_T	Large (1)	8.861e-125	GRE_B	Large (0.1364)
GRE_P and GRE_B	NaN	None	NO effect (0.5)	0.04552	GRE_P	Small (0.522)
HGS_T and HGS_P	5.984e-167	HGS_T	Large (1)	4.717e-118	HGS_P	Large (0.1705)
HGS_T and HGS_B	6.065e-167	HGS_T	Large (1)	5.163e-123	HGS_B	Large (0.1595)
HGS_P and HGS_B	0.0005367	HGS_P	Small (0.5391)	0.000107	HGS_B	Small (0.4704)
Sim_T and Sim_P	4.198e-176	Sim_T	Large (1)	1.456e-141	Sim_P	Large (0.08548)
Sim_T and Sim_B	8.215e-168	Sim_T	Large (1)	8.032e-176	Sim_B	Large (0.001623)
Sim_P and Sim_B	2.146e-106	Sim_P	Large (0.74)	1.099e-67	Sim_B	Medium (0.345)

Table A.19: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT synthetics*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and G_P	0.000	G_T	Large (0.941)	0.000	G_P	Large (0.0832)
G_T and G_B	0.000	G_T	Large (0.9424)	0.000	G_B	Large (0.0821)
G_P and G_B	0.000	G_P	Small (0.5222)	3.389e-07	G_B	Small (0.4897)
GE_T and GE_P	0.000	GE_T	Large (0.9444)	0.000	GE_P	Large (0.07939)
GE_T and GE_B	0.000	GE_T	Large (0.9464)	0.000	GE_B	Large (0.07877)
GE_P and GE_B	0.000	GE_P	Small (0.525)	0.06795	GE_B	Small (0.4953)
GRE_T and GRE_P	0.000	GRE_T	Large (0.9444)	0.000	GRE_P	Large (0.08135)
GRE_T and GRE_B	0.000	GRE_T	Large (0.9467)	0.000	GRE_B	Large (0.08904)
GRE_P and GRE_B	0.000	GRE_P	Small (0.525)	9.896e-47	GRE_P	Small (0.5289)
HGS_T and HGS_P	0.000	HGS_T	Large (0.9286)	0.000	HGS_P	Large (0.1001)
HGS_T and HGS_B	0.000	HGS_T	Large (0.9304)	0.000	HGS_B	Large (0.09743)
HGS_P and HGS_B	0.000	HGS_P	Small (0.5123)	6.644e-13	HGS_B	Small (0.4848)
Sim_T and Sim_P	0.000	Sim_T	Large (0.9422)	0.000	Sim_P	Large (0.0739)
Sim_T and Sim_B	0.000	Sim_T	Large (0.9426)	0.000	Sim_B	Large (0.04572)
Sim_P and Sim_B	0.000	Sim_P	Small (0.532)	0.000	Sim_B	Small (0.4152)

A.5.2 Study Question 2

Table A.20: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB real*

All-transitions

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and GE_T	0.4693	G_T	Small (0.507)	2.438e-11	G_T	Small (0.5906)
G_T and GRE_T	7.757e-06	GRE_T	Small (0.46)	3.661e-63	G_T	Large (0.7244)
G_T and HGS_T	1.496e-140	G_T	Large (0.9219)	0.2007	HGS_T	Small (0.4881)
G_T and Sim_T	2.618e-56	Sim_T	Medium (0.375)	9.725e-75	Sim_T	Large (0.2602)
GE_T and GRE_T	5.055e-07	GRE_T	Small (0.453)	2.317e-29	GE_T	Medium (0.6356)
GE_T and HGS_T	3.828e-141	GE_T	Large (0.9188)	2.6e-15	HGS_T	Medium (0.3994)
GE_T and Sim_T	2.322e-59	Sim_T	Medium (0.368)	5.798e-103	Sim_T	Large (0.1805)
GRE_T and HGS_T	1.274e-151	GRE_T	Large (0.9392)	1.663e-69	HGS_T	Large (0.2682)
GRE_T and Sim_T	7.486e-39	Sim_T	Small (0.415)	1.737e-151	Sim_T	Large (0.06318)
HGS_T and Sim_T	1.812e-164	Sim_T	Large (0.024)	5.267e-62	Sim_T	Large (0.2787)

All-transition-pairs

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_P and GE_P	2.561e-107	GE_P	Large (0.1901)	1.015e-19	G_P	Medium (0.6155)
G_P and GRE_P	6.442e-108	GRE_P	Large (0.1855)	1.137e-91	G_P	Large (0.7768)
G_P and HGS_P	2.988e-156	G_P	Large (0.9692)	0.02106	HGS_P	Small (0.4704)
G_P and Sim_P	1.809e-147	Sim_P	Large (0.099)	3.186e-71	Sim_P	Large (0.2493)
GE_P and GRE_P	0.5202	GRE_P	Small (0.494)	2.006e-34	GE_P	Medium (0.6408)
GE_P and HGS_P	8.707e-168	GE_P	Large (1)	9.358e-26	HGS_P	Medium (0.3641)
GE_P and Sim_P	8.068e-53	Sim_P	Medium (0.383)	1.05e-108	Sim_P	Large (0.1751)
GRE_P and HGS_P	7.926e-168	GRE_P	Large (1)	3.878e-93	HGS_P	Large (0.2137)
GRE_P and Sim_P	3.342e-50	Sim_P	Medium (0.389)	3.514e-157	Sim_P	Large (0.04022)
HGS_P and Sim_P	7.167e-171	Sim_P	Large (0)	5.18e-56	Sim_P	Large (0.2913)

Bi-criteria

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_B and GE_B	0.3998	GE_B	Small (0.4917)	6.566e-11	G_B	Small (0.5816)
G_B and GRE_B	9.791e-42	G_B	Medium (0.6633)	1.342e-65	G_B	Large (0.7239)
G_B and HGS_B	1.056e-62	G_B	Large (0.7209)	0.1673	HGS_B	Small (0.4838)
G_B and Sim_B	1.144e-132	Sim_B	Large (0.1173)	5.272e-117	Sim_B	Large (0.1542)
GE_B and GRE_B	2.826e-46	GE_B	Large (0.6735)	1.024e-30	GE_B	Medium (0.6384)
GE_B and HGS_B	1.427e-64	GE_B	Large (0.7289)	1.602e-14	HGS_B	Small (0.4045)
GE_B and Sim_B	7.872e-131	Sim_B	Large (0.1205)	2.205e-136	Sim_B	Large (0.1069)
GRE_B and HGS_B	1.663e-13	GRE_B	Small (0.5996)	1.649e-70	HGS_B	Large (0.266)
GRE_B and Sim_B	1.725e-160	Sim_B	Large (0.02859)	4.118e-163	Sim_B	Large (0.01696)
HGS_B and Sim_B	9.19e-156	Sim_B	Large (0.03531)	3.324e-109	Sim_B	Large (0.1748)

Table A.21: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for general average in CB synthetics*

All-transitions

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and GE_T	0.704	G_T	Small (0.5004)	0.8939	G_T	Small (0.5001)
G_T and GRE_T	0.0002906	GRE_T	Small (0.4961)	6.71e-47	G_T	Small (0.5275)
G_T and HGS_T	0.000	G_T	Large (0.7032)	0	G_T	Small (0.5847)
G_T and Sim_T	0.000	G_T	Large (0.7718)	0	Sim_T	Large (0.1389)
GE_T and GRE_T	0.000119	GRE_T	Small (0.4957)	1.297e-49	GE_T	Small (0.5272)
GE_T and HGS_T	0.000	GE_T	Large (0.7029)	0	GE_T	Small (0.584)
GE_T and Sim_T	0.000	GE_T	Large (0.7716)	0	Sim_T	Large (0.1411)
GRE_T and HGS_T	0.000	GRE_T	Large (0.7064)	1.156e-212	GRE_T	Small (0.558)
GRE_T and Sim_T	0.000	GRE_T	Large (0.774)	0	Sim_T	Large (0.122)
HGS_T and Sim_T	0.000	HGS_T	Small (0.5892)	0	Sim_T	Large (0.09)

All-transition-pairs

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_P and GE_P	3.662e-272	GE_P	Small (0.4848)	3.147e-14	G_P	Small (0.5171)
G_P and GRE_P	4.38e-252	GRE_P	Small (0.4854)	0.05012	G_P	Small (0.5036)
G_P and HGS_P	0.000	G_P	Medium (0.6218)	1.98e-99	G_P	Small (0.5455)
G_P and Sim_P	0.000	Sim_P	Small (0.4583)	0.000	Sim_P	Large (0.03597)
GE_P and GRE_P	0.1937	GE_P	Small (0.5007)	9.159e-10	GRE_P	Small (0.4859)
GE_P and HGS_P	0.000	GE_P	Medium (0.6396)	1.343e-41	GE_P	Small (0.5284)
GE_P and Sim_P	3.625e-295	Sim_P	Small (0.4714)	0.000	Sim_P	Large (0.03285)
GRE_P and HGS_P	0.000	GRE_P	Medium (0.6389)	1.407e-88	GRE_P	Small (0.5432)
GRE_P and Sim_P	0.000	Sim_P	Small (0.4707)	0.000	Sim_P	Large (0.03261)
HGS_P and Sim_P	0.000	Sim_P	Medium (0.3376)	0.000	Sim_P	Large (0.02826)

Bi-criteria

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_B and GE_B	0.7723	G_B	Small (0.5002)	0.7273	G_B	Small (0.5012)
G_B and GRE_B	7.153e-13	G_B	Small (0.5063)	0.7899	G_B	Small (0.5006)
G_B and HGS_B	0.000	G_B	Medium (0.6471)	0.000	G_B	Small (0.5844)
G_B and Sim_B	0.000	G_B	Medium (0.6251)	0.000	Sim_B	Large (0.05058)
GE_B and GRE_B	2.075e-13	GE_B	Small (0.5061)	0.9786	GRE_B	Small (0.4993)
GE_B and HGS_B	0.000	GE_B	Medium (0.647)	0.000	GE_B	Small (0.5831)
GE_B and Sim_B	0.000	GE_B	Medium (0.625)	0.000	Sim_B	Large (0.05082)
GRE_B and HGS_B	0.000	GRE_B	Medium (0.6393)	0.000	GRE_B	Small (0.5855)
GRE_B and Sim_B	0.000	GRE_B	Medium (0.6183)	0.000	Sim_B	Large (0.04875)
HGS_B and Sim_B	8.119e-150	Sim_B	Small (0.4788)	0.000	Sim_B	Large (0.02867)

Table A.22: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam real*

All-transitions

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and GE_T	NaN	None	NO effect (0.5)	0.3779	GE_T	Small (0.4858)
G_T and GRE_T	NaN	None	NO effect (0.5)	0.2248	GRE_T	Small (0.485)
G_T and HGS_T	6.354e-173	G_T	Large (1)	1.025e-09	G_T	Small (0.5745)
G_T and Sim_T	2.063e-146	G_T	Large (0.873)	0.5262	Sim_T	Small (0.4947)
GE_T and GRE_T	NaN	None	NO effect (0.5)	0.6902	GRE_T	Small (0.4989)
GE_T and HGS_T	6.354e-173	GE_T	Large (1)	1.478e-12	GE_T	Small (0.5899)
GE_T and Sim_T	2.063e-146	GE_T	Large (0.873)	0.9069	GE_T	Small (0.5086)
GRE_T and HGS_T	6.354e-173	GRE_T	Large (1)	2.118e-13	GRE_T	Small (0.5891)
GRE_T and Sim_T	2.063e-146	GRE_T	Large (0.873)	0.6158	GRE_T	Small (0.5095)
HGS_T and Sim_T	5.363e-98	Sim_T	Large (0.2152)	5.535e-11	Sim_T	Small (0.4215)

All-transition-pairs

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_P and GE_P	NaN	None	NO effect (0.5)	0.4007	GE_P	Small (0.4927)
G_P and GRE_P	NaN	None	NO effect (0.5)	0.24	GRE_P	Small (0.4884)
G_P and HGS_P	1.799e-219	G_P	Large (1)	8.058e-07	G_P	Small (0.562)
G_P and Sim_P	5.23e-77	G_P	Large (0.6725)	0.7588	Sim_P	Small (0.4983)
GE_P and GRE_P	NaN	None	NO effect (0.5)	0.8274	GRE_P	Small (0.4958)
GE_P and HGS_P	1.799e-219	GE_P	Large (1)	9.173e-09	GE_P	Small (0.5684)
GE_P and Sim_P	5.23e-77	GE_P	Large (0.6725)	0.5667	GE_P	Small (0.5056)
GRE_P and HGS_P	1.799e-219	GRE_P	Large (1)	1.428e-09	GRE_P	Small (0.573)
GRE_P and Sim_P	5.23e-77	GRE_P	Large (0.6725)	0.4395	GRE_P	Small (0.5099)
HGS_P and Sim_P	1.832e-144	Sim_P	Large (0.1725)	1.968e-07	Sim_P	Small (0.4369)

Bi-criteria

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_B and GE_B	0.2315	GE_B	Small (0.485)	0.843	GE_B	Small (0.4995)
G_B and GRE_B	2.124e-62	G_B	Large (0.7038)	0.05683	GRE_B	Small (0.4811)
G_B and HGS_B	1.67e-78	G_B	Large (0.7399)	8.109e-10	G_B	Small (0.5768)
G_B and Sim_B	3.007e-155	G_B	Large (0.956)	1.626e-99	Sim_B	Large (0.1934)
GE_B and GRE_B	1.724e-72	GE_B	Large (0.7187)	0.04481	GRE_B	Small (0.4816)
GE_B and HGS_B	6.322e-85	GE_B	Large (0.7529)	1.678e-09	GE_B	Small (0.5776)
GE_B and Sim_B	1.852e-157	GE_B	Large (0.9597)	1.585e-101	Sim_B	Large (0.1941)
GRE_B and HGS_B	4.009e-12	GRE_B	Small (0.5633)	7.805e-15	GRE_B	Small (0.5924)
GRE_B and Sim_B	2.809e-145	GRE_B	Large (0.9054)	2.759e-92	Sim_B	Large (0.2142)
HGS_B and Sim_B	9.323e-114	HGS_B	Large (0.8266)	9.479e-118	Sim_B	Large (0.1466)

Table A.23: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for general average in PDFSam synthetics*

All-transitions

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	FC ρ-value	FC Superior	FC Effect Size
G_T and GE_T	1.018e-174	GE_T	Small (0.489)	5.547e-05	G_T	Small (0.5068)
G_T and GRE_T	2.067e-286	GRE_T	Small (0.4873)	1.419e-97	G_T	Small (0.5415)
G_T and HGS_T	0.2282	G_T	Small (0.5051)	1.012e-91	G_T	Small (0.5366)
G_T and Sim_T	9.033e-12	G_T	Small (0.5008)	0.000	Sim_T	Large (0.1759)
GE_T and GRE_T	1.748e-25	GRE_T	Small (0.4982)	7.393e-70	GE_T	Small (0.535)
GE_T and HGS_T	3.536e-163	GE_T	Small (0.5159)	3.584e-59	GE_T	Small (0.5299)
GE_T and Sim_T	2.132e-114	GE_T	Small (0.51)	0.000	Sim_T	Large (0.1715)
GRE_T and HGS_T	1.351e-250	GRE_T	Small (0.5176)	0.7757	HGS_T	Small (0.4946)
GRE_T and Sim_T	5.713e-187	GRE_T	Small (0.512)	0.000	Sim_T	Large (0.1546)
HGS_T and Sim_T	7.204e-07	Sim_T	Small (0.4916)	0.000	Sim_T	Large (0.152)

All-transition-pairs

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	FC ρ-value	FC Superior	FC Effect Size
G_P and GE_P	1.266e-14	G_P	Small (0.5014)	1.019e-12	G_P	Small (0.5124)
G_P and GRE_P	3.235e-10	G_P	Small (0.5011)	1.416e-106	G_P	Small (0.5433)
G_P and HGS_P	0.000	G_P	Small (0.5182)	2.582e-72	G_P	Small (0.5387)
G_P and Sim_P	0.000	G_P	Small (0.5408)	0.000	Sim_P	Large (0.09703)
GE_P and GRE_P	0.007528	GRE_P	Small (0.4997)	3.424e-51	GE_P	Small (0.5312)
GE_P and HGS_P	0.000	GE_P	Small (0.5173)	3.873e-32	GE_P	Small (0.5265)
GE_P and Sim_P	0.000	GE_P	Small (0.5394)	0.000	Sim_P	Large (0.08884)
GRE_P and HGS_P	0.000	GRE_P	Small (0.5175)	0.01515	HGS_P	Small (0.4955)
GRE_P and Sim_P	0.000	GRE_P	Small (0.5395)	0.000	Sim_P	Large (0.07624)
HGS_P and Sim_P	4.921e-253	HGS_P	Small (0.5203)	0.000	Sim_P	Large (0.08089)

Bi-criteria

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	FC ρ-value	FC Superior	FC Effect Size
G_B and GE_B	0.007884	G_B	Small (0.5011)	0.004659	G_B	Small (0.5071)
G_B and GRE_B	5.672e-43	G_B	Small (0.5036)	2.641e-159	G_B	Small (0.5644)
G_B and HGS_B	0.008151	G_B	Small (0.5006)	5.028e-62	G_B	Small (0.5358)
G_B and Sim_B	0.000	G_B	Small (0.5682)	0.000	Sim_B	Large (0.03377)
GE_B and GRE_B	2.58e-26	GE_B	Small (0.5025)	3.149e-131	GE_B	Small (0.5572)
GE_B and HGS_B	0.733	HGS_B	Small (0.4996)	1.135e-45	GE_B	Small (0.5284)
GE_B and Sim_B	0.000	GE_B	Small (0.5676)	0.000	Sim_B	Large (0.03371)
GRE_B and HGS_B	2.327e-20	HGS_B	Small (0.4973)	8.213e-21	HGS_B	Small (0.4702)
GRE_B and Sim_B	0.000	GRE_B	Small (0.5665)	0.000	Sim_B	Large (0.02595)
HGS_B and Sim_B	0.000	HGS_B	Small (0.5679)	0.000	Sim_B	Large (0.02597)

Table A.24: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT real*

Comparison	SSR			FC		
	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
All-transitions						
G_T and GE_T	NaN	None	NO effect (0.5)	0.7146	GE_T	Small (0.496)
G_T and GRE_T	NaN	None	NO effect (0.5)	0.2133	GRE_T	Small (0.486)
G_T and HGS_T	2.326e-167	G_T	Large (0.9995)	7.374e-07	HGS_T	Small (0.4396)
G_T and Sim_T	1.909e-114	G_T	Large (0.7585)	0.01008	Sim_T	Small (0.471)
GE_T and GRE_T	NaN	None	NO effect (0.5)	0.3796	GRE_T	Small (0.49)
GE_T and HGS_T	2.326e-167	GE_T	Large (0.9995)	5.664e-06	HGS_T	Small (0.4432)
GE_T and Sim_T	1.909e-114	GE_T	Large (0.7585)	0.02595	Sim_T	Small (0.475)
GRE_T and HGS_T	2.326e-167	GRE_T	Large (0.9995)	0.0001295	HGS_T	Small (0.4522)
GRE_T and Sim_T	1.909e-114	GRE_T	Large (0.7585)	0.1832	Sim_T	Small (0.485)
HGS_T and Sim_T	6.416e-163	Sim_T	Large (0.01136)	0.02042	HGS_T	Small (0.5343)
All-transition-pairs						
Comparison	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_P and GE_P	NaN	None	NO effect (0.5)	0.596	GE_P	Small (0.494)
G_P and GRE_P	NaN	None	NO effect (0.5)	0.5995	GRE_P	Small (0.494)
G_P and HGS_P	1.442e-96	G_P	Large (0.7495)	1.227e-06	HGS_P	Small (0.4595)
G_P and Sim_P	NaN	None	NO effect (0.5)	7.771e-17	Sim_P	Small (0.407)
GE_P and GRE_P	NaN	None	NO effect (0.5)	1.000	None	NO effect (0.5)
GE_P and HGS_P	1.442e-96	GE_P	Large (0.7495)	2.805e-05	HGS_P	Small (0.4653)
GE_P and Sim_P	NaN	None	NO effect (0.5)	1.033e-14	Sim_P	Small (0.413)
GRE_P and HGS_P	1.442e-96	GRE_P	Large (0.7495)	3.353e-05	HGS_P	Small (0.4653)
GRE_P and Sim_P	NaN	None	NO effect (0.5)	2.597e-15	Sim_P	Small (0.413)
HGS_P and Sim_P	1.442e-96	Sim_P	Large (0.2505)	0.01404	Sim_P	Small (0.4508)
Bi-crietria						
Comparison	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_B and GE_B	NaN	None	NO effect (0.5)	0.6817	GE_B	Small (0.4955)
G_B and GRE_B	NaN	None	NO effect (0.5)	0.596	G_B	Small (0.506)
G_B and HGS_B	1.732e-107	G_B	Large (0.7835)	3.045e-18	HGS_B	Small (0.422)
G_B and Sim_B	2.146e-106	G_B	Large (0.74)	6.15e-113	Sim_B	Large (0.242)
GE_B and GRE_B	NaN	None	NO effect (0.5)	0.3333	GE_B	Small (0.5105)
GE_B and HGS_B	1.732e-107	GE_B	Large (0.7835)	1.846e-18	HGS_B	Small (0.4261)
GE_B and Sim_B	2.146e-106	GE_B	Large (0.74)	2.09e-111	Sim_B	Large (0.2465)
GRE_B and HGS_B	1.732e-107	GRE_B	Large (0.7835)	1.173e-19	HGS_B	Small (0.4165)
GRE_B and Sim_B	2.146e-106	GRE_B	Large (0.74)	2.095e-116	Sim_B	Large (0.236)
HGS_B and Sim_B	1.854e-14	Sim_B	Small (0.4272)	3.177e-12	Sim_B	Medium (0.3432)

Table A.25: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for general average in TaRGeT synthetics*

All-transitions

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	FC ρ-value	FC Superior	FC Effect Size
G_T and GE_T	0.000	GE_T	Small (0.4701)	2.842e-66	G_T	Small (0.536)
G_T and GRE_T	0.000	GRE_T	Small (0.4704)	1.03e-53	G_T	Small (0.5303)
G_T and HGS_T	0.000	G_T	Small (0.5894)	1.336e-53	HGS_T	Small (0.475)
G_T and Sim_T	0.000	Sim_T	Small (0.4784)	0.000	Sim_T	Medium (0.3718)
GE_T and GRE_T	0.01189	GE_T	Small (0.5002)	0.08555	GRE_T	Small (0.4944)
GE_T and HGS_T	0.000	GE_T	Medium (0.613)	9.711e-219	HGS_T	Small (0.4393)
GE_T and Sim_T	0.000	GE_T	Small (0.5103)	0.000	Sim_T	Medium (0.3432)
GRE_T and HGS_T	0.000	GRE_T	Medium (0.6128)	1.863e-198	HGS_T	Small (0.4449)
GRE_T and Sim_T	7.272e-280	GRE_T	Small (0.51)	0.000	Sim_T	Medium (0.3477)
HGS_T and Sim_T	0.000	Sim_T	Medium (0.3952)	0.000	Sim_T	Medium (0.3951)

All-transition-pairs

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	FC ρ-value	FC Superior	FC Effect Size
G_P and GE_P	0.1963	G_P	Small (0.5001)	1.553e-14	G_P	Small (0.516)
G_P and GRE_P	0.7439	G_P	Small (0.5001)	1.258e-10	G_P	Small (0.5134)
G_P and HGS_P	0.000	G_P	Small (0.514)	0.4774	G_P	Small (0.5029)
G_P and Sim_P	1.716e-298	G_P	Small (0.5021)	0.000	Sim_P	Large (0.3004)
GE_P and GRE_P	0.3414	GRE_P	Small (0.5)	0.4292	GRE_P	Small (0.4975)
GE_P and HGS_P	0.000	GE_P	Small (0.5138)	1.515e-09	HGS_P	Small (0.4869)
GE_P and Sim_P	2.598e-288	GE_P	Small (0.502)	0.000	Sim_P	Large (0.29)
GRE_P and HGS_P	0.000	GRE_P	Small (0.5139)	8.852e-07	HGS_P	Small (0.4896)
GRE_P and Sim_P	1.35e-293	GRE_P	Small (0.502)	0.000	Sim_P	Large (0.2927)
HGS_P and Sim_P	0.000	Sim_P	Small (0.4887)	0.000	Sim_P	Large (0.2992)

Bi-criteria

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	FC ρ-value	FC Superior	FC Effect Size
G_B and GE_B	9.422e-42	G_B	Small (0.503)	4.866e-22	G_B	Small (0.521)
G_B and GRE_B	7.343e-52	G_B	Small (0.5031)	1.328e-140	G_B	Small (0.5524)
G_B and HGS_B	9.001e-84	G_B	Small (0.5037)	0.5043	HGS_B	Small (0.4979)
G_B and Sim_B	0.000	G_B	Small (0.512)	0.000	Sim_B	Large (0.2382)
GE_B and GRE_B	0.03936	GE_B	Small (0.5001)	2.348e-53	GE_B	Small (0.5311)
GE_B and HGS_B	3.777e-11	GE_B	Small (0.5007)	2.931e-28	HGS_B	Small (0.4771)
GE_B and Sim_B	0.000	GE_B	Small (0.5088)	0.000	Sim_B	Large (0.2229)
GRE_B and HGS_B	3.46e-06	GRE_B	Small (0.5006)	1.876e-151	HGS_B	Small (0.4458)
GRE_B and Sim_B	0.000	GRE_B	Small (0.5084)	0.000	Sim_B	Large (0.1946)
HGS_B and Sim_B	0.000	HGS_B	Small (0.5085)	0.000	Sim_B	Large (0.2411)

A.5.3 Study Question 3

Table A.26: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB real*

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	Comparison	FC ρ-value	FC Superior	FC Effect Size
G_T and GE_T	0.4693	G_T	Small (0.507)	G_P and GE_P	0.4693	G_P	Small (0.507)
G_T and GRE_T	7.757e-06	GRE_T	Small (0.46)	G_P and GRE_P	7.757e-06	GRE_P	Small (0.46)
G_T and HGS_T	1.496e-140	G_T	Large (0.9219)	G_P and HGS_P	1.496e-140	G_P	Large (0.9219)
G_T and Sim_T	2.618e-56	Sim_T	Medium (0.375)	G_P and Sim_B	2.618e-56	Sim_B	Medium (0.375)
GE_T and GRE_T	5.055e-07	GRE_T	Small (0.453)	GE_P and GRE_P	5.055e-07	GRE_P	Small (0.453)
GE_T and HGS_T	3.828e-141	GE_T	Large (0.9188)	GE_P and HGS_P	3.828e-141	GE_P	Large (0.9188)
GE_T and Sim_T	2.322e-59	Sim_T	Medium (0.368)	GE_P and Sim_B	2.322e-59	Sim_B	Medium (0.368)
GRE_T and HGS_T	1.274e-151	GRE_T	Large (0.9392)	GRE_P and HGS_P	1.274e-151	GRE_P	Large (0.9392)
GRE_T and Sim_T	7.486e-39	Sim_T	Small (0.415)	GRE_P and Sim_B	7.486e-39	Sim_B	Small (0.415)
HGS_T and Sim_T	1.812e-164	Sim_T	Large (0.024)	HGS_P and Sim_B	1.812e-164	Sim_B	Large (0.024)

Table A.27: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB synthetics*

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	Comparison	FC ρ-value	FC Superior	FC Effect Size
G_T and GE_T	0.704	G_T	Small (0.5004)	G_B and GE_B	0.704	G_B	Small (0.5004)
G_T and GRE_T	0.0002906	GRE_T	Small (0.4961)	G_B and GRE_B	0.0002906	GRE_B	Small (0.4961)
G_T and HGS_T	0.000	G_T	Large (0.7032)	G_B and HGS_P	0.000	G_B	Large (0.7032)
G_T and Sim_T	0.000	G_T	Large (0.7718)	G_B and Sim_P	0.000	G_B	Large (0.7718)
GE_T and GRE_T	0.000119	GRE_T	Small (0.4957)	GE_B and GRE_B	0.000119	GRE_B	Small (0.4957)
GE_T and HGS_T	0.000	GE_T	Large (0.7029)	GE_B and HGS_P	0.000	GE_B	Large (0.7029)
GE_T and Sim_T	0.000	GE_T	Large (0.7716)	GE_B and Sim_P	0.000	GE_B	Large (0.7716)
GRE_T and HGS_T	0.000	GRE_T	Large (0.7064)	GRE_B and HGS_P	0.000	GRE_B	Large (0.7064)
GRE_T and Sim_T	0.000	GRE_T	Large (0.774)	GRE_B and Sim_P	0.000	GRE_B	Large (0.774)
HGS_T and Sim_T	0.000	HGS_T	Small (0.5892)	HGS_P and Sim_P	0.000	HGS_P	Small (0.5892)

Table A.28: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam real*

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	Comparison	FC ρ-value	FC Superior	FC Effect Size
G_T and GE_T	NaN	None	NO effect (0.5)	G_B and GE_B	NaN	None	NO effect (0.5)
G_T and GRE_T	NaN	None	NO effect (0.5)	G_B and GRE_B	NaN	None	NO effect (0.5)
G_T and HGS_T	6.354e-173	G_T	Large (1)	G_B and HGS_B	6.354e-173	G_B	Large (1)
G_T and Sim_T	2.063e-146	G_T	Large (0.873)	G_B and Sim_B	2.063e-146	G_B	Large (0.873)
GE_T and GRE_T	NaN	None	NO effect (0.5)	GE_B and GRE_B	NaN	None	NO effect (0.5)
GE_T and HGS_T	6.354e-173	GE_T	Large (1)	GE_B and HGS_B	6.354e-173	GE_B	Large (1)
GE_T and Sim_T	2.063e-146	GE_T	Large (0.873)	GE_B and Sim_B	2.063e-146	GE_B	Large (0.873)
GRE_T and HGS_T	6.354e-173	GRE_T	Large (1)	GRE_B and HGS_B	6.354e-173	GRE_B	Large (1)
GRE_T and Sim_T	2.063e-146	GRE_T	Large (0.873)	GRE_B and Sim_B	2.063e-146	GRE_B	Large (0.873)
HGS_T and Sim_T	5.363e-98	Sim_T	Large (0.2152)	HGS_B and Sim_B	5.363e-98	Sim_B	Large (0.2152)

Table A.29: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam synthetics*

Comparison	SSR			Comparison	FC		
	ρ-value	Superior	Effect Size		ρ-value	Superior	Effect Size
G_T and GE_T	1.018e-174	GE_T	Small (0.489)	G_P and GE_P	1.018e-174	GE_P	Small (0.489)
G_T and GRE_T	2.067e-286	GRE_T	Small (0.4873)	G_P and GRE_P	2.067e-286	GRE_P	Small (0.4873)
G_T and HGS_T	0.2282	G_T	Small (0.5051)	G_P and HGS_P	0.2282	G_P	Small (0.5051)
G_T and Sim_T	9.033e-12	G_T	Small (0.5008)	G_P and Sim_B	9.033e-12	G_P	Small (0.5008)
GE_T and GRE_T	1.748e-25	GRE_T	Small (0.4982)	GE_P and GRE_P	1.748e-25	GRE_P	Small (0.4982)
GE_T and HGS_T	3.536e-163	GE_T	Small (0.5159)	GE_P and HGS_P	3.536e-163	GE_P	Small (0.5159)
GE_T and Sim_T	2.132e-114	GE_T	Small (0.51)	GE_P and Sim_B	2.132e-114	GE_P	Small (0.51)
GRE_T and HGS_T	1.351e-250	GRE_T	Small (0.5176)	GRE_P and HGS_P	1.351e-250	GRE_P	Small (0.5176)
GRE_T and Sim_T	5.713e-187	GRE_T	Small (0.512)	GRE_P and Sim_B	5.713e-187	GRE_P	Small (0.512)
HGS_T and Sim_T	7.204e-07	Sim_T	Small (0.4916)	HGS_P and Sim_B	7.204e-07	Sim_B	Small (0.4916)

Table A.30: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT real*

Comparison	SSR			Comparison	FC		
	ρ-value	Superior	Effect Size		ρ-value	Superior	Effect Size
G_T and GE_T	NaN	None	NO effect (0.5)	G_P and GE_P	NaN	None	NO effect (0.5)
G_T and GRE_T	NaN	None	NO effect (0.5)	G_P and GRE_P	NaN	None	NO effect (0.5)
G_T and HGS_T	2.326e-167	G_T	Large (0.9995)	G_P and HGS_B	2.326e-167	G_P	Large (0.9995)
G_T and Sim_T	1.909e-114	G_T	Large (0.7585)	G_P and Sim_B	1.909e-114	G_P	Large (0.7585)
GE_T and GRE_T	NaN	None	NO effect (0.5)	GE_P and GRE_P	NaN	None	NO effect (0.5)
GE_T and HGS_T	2.326e-167	GE_T	Large (0.9995)	GE_P and HGS_B	2.326e-167	GE_P	Large (0.9995)
GE_T and Sim_T	1.909e-114	GE_T	Large (0.7585)	GE_P and Sim_B	1.909e-114	GE_P	Large (0.7585)
GRE_T and HGS_T	2.326e-167	GRE_T	Large (0.9995)	GRE_P and HGS_B	2.326e-167	GRE_P	Large (0.9995)
GRE_T and Sim_T	1.909e-114	GRE_T	Large (0.7585)	GRE_P and Sim_B	1.909e-114	GRE_P	Large (0.7585)
HGS_T and Sim_T	6.416e-163	Sim_T	Large (0.01136)	HGS_B and Sim_B	6.416e-163	Sim_B	Large (0.01136)

Table A.31: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT synthetics*

Comparison	SSR			Comparison	FC		
	ρ-value	Superior	Effect Size		ρ-value	Superior	Effect Size
G_T and GE_T	0.000	GE_T	Small (0.4701)	G_B and GE_B	0.000	GE_B	Small (0.4701)
G_T and GRE_T	0.000	GRE_T	Small (0.4704)	G_B and GRE_P	0.000	GRE_P	Small (0.4704)
G_T and HGS_T	0.000	G_T	Small (0.5894)	G_B and HGS_B	0.000	G_B	Small (0.5894)
G_T and Sim_T	0.000	Sim_T	Small (0.4784)	G_B and Sim_B	0.000	Sim_B	Small (0.4784)
GE_T and GRE_T	0.01189	GE_T	Small (0.5002)	GE_B and GRE_P	0.01189	GE_B	Small (0.5002)
GE_T and HGS_T	0.000	GE_T	Medium (0.613)	GE_B and HGS_B	0.000	GE_B	Medium (0.613)
GE_T and Sim_T	0.000	GE_T	Small (0.5103)	GE_B and Sim_B	0.000	GE_B	Small (0.5103)
GRE_T and HGS_T	0.000	GRE_T	Medium (0.6128)	GRE_P and HGS_B	0.000	GRE_P	Medium (0.6128)
GRE_T and Sim_T	7.272e-280	GRE_T	Small (0.51)	GRE_P and Sim_B	7.272e-280	GRE_P	Small (0.51)
HGS_T and Sim_T	0.000	Sim_T	Medium (0.3952)	HGS_B and Sim_B	0.000	Sim_B	Medium (0.3952)

A.5.4 Study Question 4

Table A.32: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB real*

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	Comparison	FC ρ-value	FC Superior	FC Effect Size
Sim_T and Sim_P	1.799e-219	Sim_T	Large (1)	Sim_T and Sim_P	1.799e-219	Sim_T	Large (1)
Sim_T and Sim_B	1.139e-183	Sim_T	Large (1)	Sim_T and Sim_B	1.139e-183	Sim_T	Large (1)
Sim_P and Sim_B	6.921e-61	Sim_P	Medium (0.6355)	Sim_P and Sim_B	6.921e-61	Sim_P	Medium (0.6355)

Table A.33: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB synthetics*

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	Comparison	FC ρ-value	FC Superior	FC Effect Size
GRE_T and Sim_P	0.000	GRE_T	Large (1)	Sim_T and Sim_P	0.000	Sim_T	Large (1)
GRE_T and G_B	0.000	GRE_T	Large (1)	Sim_T and Sim_B	0.000	Sim_T	Large (1)
Sim_P and G_B	0.000	Sim_P	Small (0.5396)	Sim_P and Sim_B	0.000	Sim_P	Small (0.5396)

Table A.34: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam real*

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	Comparison	FC ρ-value	FC Superior	FC Effect Size
$G_T = GE_T = GRE_T$ and $G_P = GE_P = GRE_P$	1.799e-219	$G_T = GE_T = GRE_T$	Large (1.0)	GRE_T and GRE_P	1.799e-219	GRE_T	Large (1.0)
$G_T = GE_T = GRE_T$ and GE_B	8.748e-172	$G_T = GE_T = GRE_T$	Large (1.0)	GRE_T and Sim_B	8.748e-172	GRE_T	Large (1.0)
$G_P = GE_P = GRE_P$ and GE_B	1.074e-133	$G_P = GE_P = GRE_P$	Large (0.869)	GRE_P and Sim_B	1.074e-133	GRE_P	Large (0.869)

Table A.35: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam synthetics*

Comparison	SSR ρ-value	SSR Superior	SSR Effect Size	Comparison	FC ρ-value	FC Superior	FC Effect Size
GRE_T and G_P	0.000	GRE_T	Large (0.917)	Sim_T and Sim_P	0.000	Sim_T	Large (0.917)
GRE_T and G_B	0.000	GRE_T	Large (0.9233)	Sim_T and Sim_B	0.000	Sim_T	Large (0.9233)
G_P and G_B	0.000	G_P	Small (0.5121)	Sim_P and Sim_B	0.000	Sim_P	Small (0.5121)

Table A.36: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT real*

Comparison	SSR			Comparison	FC		
	ρ-value	Superior	Effect Size		ρ-value	Superior	Effect Size
$G_T = GE_T = GRE_T$ and $G_P = GE_P = GRE_P = Sim_P$	1.799e-219	$G_T = GE_T = GRE_T$	Large (1.0)	HGS_T and Sim_P	1.799e-219	HGS_T	Large (1.0)
$G_T = GE_T = GRE_T$ and $G_B = GE_B = GRE_B$	1.799e-219	$G_T = GE_T = GRE_T$	Large (1.0)	HGS_T and Sim_B	1.799e-219	HGS_T	Large (1.0)
$G_P = GE_P = GRE_P = Sim_P$ and $G_B = GE_B = GRE_B$	NaN	None	NO effect (0.5)	Sim_P and Sim_B	NaN	None	NO effect (0.5)

Table A.37: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT synthetics*

Comparison	SSR			Comparison	FC		
	ρ-value	Superior	Effect Size		ρ-value	Superior	Effect Size
GE_T and G_P	0.000	GE_T	Large (0.9444)	Sim_T and Sim_P	0.000	Sim_T	Large (0.9444)
GE_T and G_B	0.000	GE_T	Large (0.9461)	Sim_T and Sim_B	0.000	Sim_T	Large (0.9461)
G_P and G_B	0.000	G_P	Small (0.5222)	Sim_P and Sim_B	0.000	Sim_P	Small (0.5222)

A.6 The minimum, maximum, median and average

Table A.38: *The minimum, maximum, median and average for CB real*

All-transitions (T)						
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_T	62.32	63.77	63.77	63.41	0.6279
	GE_T	62.32	63.77	63.77	63.39	0.6392
	GRE_T	62.32	63.77	63.77	63.52	0.5447
	HGS_T	59.42	60.87	63.77	61.4	1.083
	Sim_T	63.77	63.77	63.77	63.77	0.00
FC	G_T	20	50	80	45.71	11.93
	GE_T	20	40	90	42.1	11.72
	GRE_T	20	40	60	36.32	9.063
	HGS_T	20	50	90	46.44	12.36
	Sim_T	40	60	70	55.28	7.195
All-transition-pairs (P)						
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_P	53.62	55.07	56.52	55.04	0.938
	GE_P	55.07	56.52	56.52	56.18	0.6139
	GRE_P	55.07	56.52	56.52	56.2	0.6026
	HGS_P	49.28	52.17	53.62	52.17	1.143
	Sim_P	56.52	56.52	56.52	56.52	0.00
FC	G_P	30	60	90	56.47	10.3
	GE_P	30	50	80	51.93	11.27
	GRE_P	30	50	60	46.3	6.896
	HGS_P	30	60	90	57.57	11.43
	Sim_P	50	70	80	65.42	7.326
Bi-criteria (B)						
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_B	52.17	55.07	56.52	54.37	1.151
	GE_B	52.17	55.07	56.52	54.41	1.129
	GRE_B	52.17	53.62	55.07	53.7	0.8801
	HGS_B	49.28	53.62	56.52	53.18	1.441
	Sim_B	55.07	56.52	56.52	56.13	0.6445
FC	G_B	30	50	90	52.3	11.35
	GE_B	30	50	90	48.95	11.5
	GRE_B	30	40	60	43.27	8.214
	HGS_B	30	50	90	53.1	11.99
	Sim_B	50	70	80	66.67	6.682

Table A.39: *The minimum, maximum, median and average for CB synthetics*

	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
			All-transitions (T)			
SSR	G_T	77.05	80	81.34	79.89	1.209
	GE_T	77.05	80	81.34	79.88	1.209
	GRE_T	77.05	80	81.34	79.9	1.213
	HGS_T	76.27	79.07	81.34	78.99	1.152
	Sim_T	72.03	78.63	81.34	78.48	1.46
FC	G_T	5.556	41.18	78.95	41.24	10.37
	GE_T	5.556	41.18	78.95	41.27	10.45
	GRE_T	5.556	38.89	78.95	40.32	10.09
	HGS_T	11.11	38.89	77.78	38.23	9.538
	Sim_T	22.22	55.56	88.24	56.57	9.569
			All-transition-pairs (P)			
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_P	62.3	69.84	75.37	69.84	2.686
	GE_P	63.11	70	75.37	69.99	2.642
	GRE_P	63.11	70	75.37	69.99	2.647
	HGS_P	60.66	68.8	73.88	68.78	2.62
	Sim_P	62.3	70.23	75.37	70.2	2.682
FC	G_P	27.78	61.11	95	62.09	9.371
	GE_P	29.41	61.11	94.74	61.54	9.323
	GRE_P	27.78	61.11	94.74	61.97	8.92
	HGS_P	29.41	61.11	94.74	60.58	9.283
	Sim_P	55.56	84.21	100	85.15	8.01
			Bi-criteria (B)			
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_B	60.66	69.84	75.37	69.78	3.046
	GE_B	60.66	69.84	75.37	69.78	3.043
	GRE_B	59.84	69.84	75.37	69.73	3.057
	HGS_B	60.66	68.7	73.88	68.35	2.717
	Sim_B	59.02	68.75	75.37	68.54	2.795
FC	G_B	27.27	63.16	94.74	63.13	9.872
	GE_B	22.73	63.16	100	63.09	9.903
	GRE_B	27.78	63.16	94.74	63.13	9.554
	HGS_B	29.41	61.11	100	60.22	9.344
	Sim_B	50	84.21	100	84.86	8.213

Table A.40: *The minimum, maximum, median and average for PDFSam real*

	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
All-transitions (T)						
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_T	91.24	91.24	91.24	91.24	0.00
	GE_T	91.24	91.24	91.24	91.24	0.00
	GRE_T	91.24	91.24	91.24	91.24	0.00
	HGS_T	89.05	89.78	90.51	90.03	0.4804
	Sim_T	89.05	90.51	91.24	90.62	0.4324
FC	G_T	0.00	0.00	60	11.7	13.13
	GE_T	0.00	20	60	12.24	12.77
	GRE_T	0.00	20	60	12.46	13.35
	HGS_T	0.00	0.00	80	8.2	12.33
	Sim_T	0.00	0.00	60	12.1	13.62
All-transition-pairs (P)						
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_P	87.59	87.59	87.59	87.59	0.00
	GE_P	87.59	87.59	87.59	87.59	0.00
	GRE_P	87.59	87.59	87.59	87.59	0.00
	HGS_P	86.86	86.86	86.86	86.86	0.00
	Sim_P	86.86	87.59	87.59	87.34	0.3471
FC	G_P	0.00	0.00	40	10.56	12.22
	GE_P	0.00	0.00	60	11.08	12.86
	GRE_P	0.00	0.00	60	11.26	12.84
	HGS_P	0.00	0.00	60	7.84	11.81
	Sim_P	0.00	0.00	60	10.74	12.57
Bi-criteria (B)						
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_B	86.13	86.86	87.59	86.85	0.5193
	GE_B	86.13	86.86	87.59	86.88	0.5172
	GRE_B	86.13	86.13	86.86	86.45	0.3622
	HGS_B	85.4	86.13	86.86	86.31	0.502
	Sim_B	84.67	85.4	86.13	85.64	0.3526
FC	G_B	0.00	20	60	13.56	13.95
	GE_B	0.00	20	80	13.66	14.18
	GRE_B	0.00	20	80	14.86	15.15
	HGS_B	0.00	0.00	60	9.8	13.36
	Sim_B	0.00	40	80	34.72	17.52

Table A.41: *The minimum, maximum, median and average for PDFSam synthetics*

	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
All-transitions (T)						
SSR	G_T	86.41	90	96.77	90.73	3.061
	GE_T	86.41	90	96.77	90.81	3.011
	GRE_T	86.41	90	96.77	90.82	3.015
	HGS_T	85.58	89.83	96.77	90.72	3.128
	Sim_T	86.36	89.83	96.77	90.73	2.919
FC	G_T	0.00	0.00	80	12.4	16.11
	GE_T	0.00	0.00	100	11.98	15.87
	GRE_T	0.00	0.00	100	10.03	15.13
	HGS_T	0.00	0.00	100	10.06	14.56
	Sim_T	0.00	40	100	40.17	24.26
All-transition-pairs (P)						
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_P	75.26	84.43	91.98	84.3	3.531
	GE_P	75.26	84.43	91.98	84.3	3.515
	GRE_P	75.26	84.43	91.98	84.3	3.52
	HGS_P	76.29	84.59	91.98	84.08	3.625
	Sim_P	75.26	84.21	91.98	83.93	3.615
FC	G_P	0.00	20	100	22.75	20.8
	GE_P	0.00	20	100	21.62	20.12
	GRE_P	0.00	20	100	19.36	19.29
	HGS_P	0.00	20	100	19.89	20.03
	Sim_P	0.00	66.67	100	65.84	23.46
Bi-criteria (B)						
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_B	76.29	84.43	91.98	84.2	3.467
	GE_B	76.29	84.43	91.98	84.19	3.465
	GRE_B	76.29	84.06	91.98	84.17	3.479
	HGS_B	76.29	84.76	91.98	84.21	3.502
	Sim_B	74.23	83.2	91.98	83.44	3.716
FC	G_B	0.00	20	100	21.91	20.19
	GE_B	0.00	20	100	21.54	20.34
	GRE_B	0.00	16.67	100	17.69	19.59
	HGS_B	0.00	20	100	19.31	19.15
	Sim_B	0.00	80	100	79.11	19.82

Table A.42: *The minimum, maximum, median and average for TaRGeT real*

	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
			All-transitions (T)			
SSR	G_T	35.37	35.37	35.37	35.37	0.00
	GE_T	35.37	35.37	35.37	35.37	0.00
	GRE_T	35.37	35.37	35.37	35.37	0.00
	HGS_T	26.83	30.49	35.37	30.94	1.516
	Sim_T	34.15	34.15	35.37	34.74	0.6098
FC	G_T	69.23	69.23	76.92	72.95	3.846
	GE_T	69.23	69.23	76.92	73.01	3.847
	GRE_T	69.23	76.92	76.92	73.16	3.847
	HGS_T	69.23	76.92	84.62	74.28	5.033
	Sim_T	69.23	76.92	76.92	73.39	3.835
			All-transition-pairs (P)			
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_P	12.2	12.2	12.2	12.2	0.00
	GE_P	12.2	12.2	12.2	12.2	0.00
	GRE_P	12.2	12.2	12.2	12.2	0.00
	HGS_P	8.537	12.2	12.2	11.47	0.8104
	Sim_P	12.2	12.2	12.2	12.2	0.00
FC	G_P	76.92	76.92	84.62	80.75	3.848
	GE_P	76.92	84.62	84.62	80.85	3.847
	GRE_P	76.92	84.62	84.62	80.85	3.847
	HGS_P	76.92	84.62	92.31	81.52	4.303
	Sim_P	76.92	84.62	84.62	82.18	3.578
			Bi-criteria (B)			
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_B	12.2	12.2	12.2	12.2	0.00
	GE_B	12.2	12.2	12.2	12.2	0.00
	GRE_B	12.2	12.2	12.2	12.2	0.00
	HGS_B	8.537	10.98	12.2	11.34	0.8579
	Sim_B	10.98	12.2	12.2	11.61	0.6096
FC	G_B	76.92	76.92	84.62	80.6	3.844
	GE_B	76.92	76.92	84.62	80.67	3.847
	GRE_B	76.92	76.92	84.62	80.51	3.839
	HGS_B	76.92	84.62	92.31	82.16	4.851
	Sim_B	76.92	84.62	84.62	84.57	0.5944

Table A.43: *The minimum, maximum, median and average for TaRGeT synthetics*

	All-transitions (T)					
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_T	31.25	42.05	66	42.92	7.166
	GE_T	34.52	42.29	66	43.5	7.099
	GRE_T	34.52	42.29	66	43.5	7.104
	HGS_T	27.27	39.53	66	41.02	7.705
	Sim_T	34.52	42.29	66	43.33	7.168
FC	G_T	15.38	61.11	100	60.11	11.54
	GE_T	15.38	58.33	100	58.85	11.42
	GRE_T	15.38	58.33	100	59.01	11.57
	HGS_T	15.38	61.54	100	61.16	11.69
	Sim_T	25	66.67	92.31	64.76	12.79
	All-transition-pairs (P)					
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_P	9.876	21.02	56.67	23.22	10.22
	GE_P	9.876	21.02	56.67	23.22	10.23
	GRE_P	9.876	21.02	56.67	23.22	10.23
	HGS_P	8.642	20.69	56.67	22.8	10.06
	Sim_P	9.876	21.02	56.67	23.15	10.23
FC	G_P	55.56	78.57	100	79.83	8.43
	GE_P	50	78.57	100	79.35	8.742
	GRE_P	50	78.57	100	79.42	8.763
	HGS_P	50	78.57	100	79.76	8.592
	Sim_P	58.33	85.71	100	85.94	8.357
	Bi-criteria (B)					
	Strategy	Minimum	Median	Maximum	Average	Standard Deviation
SSR	G_B	8.642	20.69	56.67	22.59	10.19
	GE_B	8.75	20.69	56.67	22.51	10.18
	GRE_B	8.75	20.45	56.67	22.49	10.16
	HGS_B	7.5	20.45	56.67	22.49	10.22
	Sim_B	7.5	20.45	56.67	22.23	10.22
FC	G_B	55.56	78.57	100	80.16	8.657
	GE_B	50	78.57	100	79.5	8.814
	GRE_B	50	77.78	100	78.48	8.458
	HGS_B	50	80	100	80.2	8.826
	Sim_B	58.33	91.3	100	88.53	7.777

A.7 Scattering (SSR_FC)

A.7.1 Normality test

Table A.44: *Anderson-Darling normality test (ρ-value) for CB configuration*

Strategy	CB real			CB synthetics		
	T	P	B	T	P	B
G	3.894e-63	1.854e-64	1.412e-145	∞	∞	NA
GE	1.243e-176	6.613e-45	1.116e-147	∞	NA	NA
GRE	1.613e-103	1.592e-63	3.093e-57	∞	NA	NA
HGS	2.123e-117	4.334e-134	4.754e-134	3.799e-130	2.089e-165	9.541e-94
Sim	3.242e-131	NaN	4.364e-170	∞	∞	∞

Table A.45: *Anderson-Darling normality test (ρ-value) for PDFSam configuration*

Strategy	PDFSam real			PDFSam synthetics		
	T	P	B	T	P	B
G	7.535e-70	4.266e-63	5.092e-61	∞	∞	∞
GE	8.18e-28	1.401e-59	NA	NA	∞	∞
GRE	4.781e-37	1.53e-79	9.31e-79	NA	∞	∞
HGS	1.266e-43	2.489e-59	2.934e-44	∞	∞	1.333e-75
Sim	6.646e-117	2.31e-20	5.654e-178	∞	∞	∞

Table A.46: *Anderson-Darling normality test (ρ-value) for TaRGeT configuration*

Strategy	TaRGeT real			TaRGeT synthetics		
	T	P	B	T	P	B
G	1.074e-69	8.819e-80	2.408e-158	∞	∞	∞
GE	2.042e-47	1.119e-167	1.285e-156	NA	∞	∞
GRE	1.326e-66	2.855e-157	1.395e-161	NA	∞	∞
HGS	3.454e-53	6.229e-134	1.72e-151	∞	∞	∞
Sim	8.868e-185	NaN	NaN	∞	∞	∞

A.7.2 Kruskal-Wallis test

Table A.47: *Krukal-Wallis test for SQ1*

	ρ-value					
Comparison	**CB real**	**CB synthetics**	**PDFSam real**	**PDFSam synthetics**	**TaRGeT real**	**TaRGeT synthetics**
$G_T = G_P = G_B$	3.108e-21	0.000	4.774e-80	0.000	7.377e-92	0.000
$GE_T = GE_P = GE_B$	3.108e-21	0.000	4.774e-80	0.000	7.377e-92	0.000
$GRE_T = GRE_P = GRE_B$	3.108e-21	0.000	4.774e-80	0.000	7.377e-92	0.000
$HGS_T = HGS_P = HGS_B$	3.108e-21	0.000	4.774e-80	0.000	7.377e-92	0.000
$Sim_T = Sim_P = Sim_B$	3.108e-21	0.000	4.774e-80	0.000	7.377e-92	0.000

Table A.48: *Krukal-Wallis test for SQ2*

	ρ-value					
Comparison	**CB real**	**CB synthetics**	**PDFSam real**	**PDFSam synthetics**	**TaRGeT real**	**TaRGeT synthetics**
$G_T = GE_T = GRE_T = HGS_T = Sim_T$	2.999e-77	0.000	3.73e-90	0.000	0.000	0.000
$G_P = GE_P = GRE_P = HGS_P = Sim_P$	9.237e-104	0.000	4.683e-105	0.000	0.000	0.000
$G_B = GE_B = GRE_B = HGS_B = Sim_B$	0.000	0.000	4.066e-121	0.000	0.000	0.000

Table A.49: *Krukal-Wallis test for SQ3*

Specification	Comparison	ρ-value
CB real	$G_P = GE_T = GRE_T = HGS_T = Sim_B$	1.05e-256
CB synthetics	$G_P = GE_P = GRE_P = HGS_B = Sim_P$	0.000
PDFSam real	$G_B = GE_P = GRE_P = HGS_T = Sim_P$	3.438e-35
PDFSam synthetics	$G_P = GE_T = GRE_T = HGS_B = Sim_B$	0.000
TaRGeT real	$G_T = GE_T = GRE_T = HGS_T = Sim_T$	0.000
TaRGeT synthetics	$G_T = GE_T = GRE_T = HGS_T = Sim_B$	0.000

Table A.50: *Krukal-Wallis test for SQ4*

Specification	Comparison	ρ-value
CB real	$GRE_T = G_P = Sim_B$	5.165e-175
CB synthetics	$HGS_T = Sim_P = Sim_B$	0.000
PDFSam real	$GRE_T = GRE_P = G_B$	0.005333
PDFSam synthetics	$Sim_T = Sim_P = Sim_B$	0.000
TaRGeT real	$Sim_T = G_P = GE_B$	0.000
TaRGeT synthetics	$Sim_T = Sim_P = Sim_B$	2.193e-189

A.7.3 Boxplots

Study Question 1

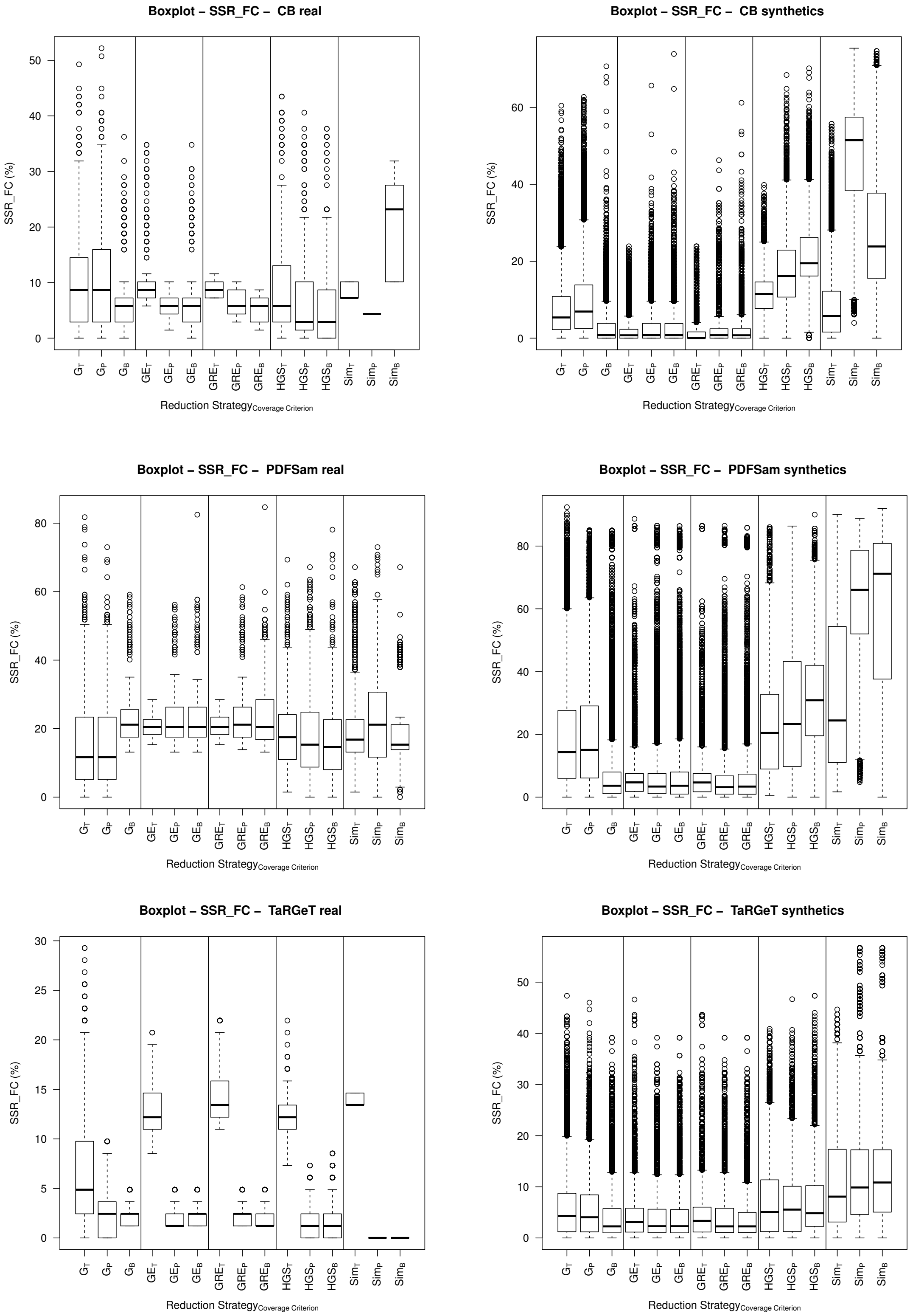


Figure A.9: *Boxplots considering SSR_FC metric for SQ1*

Study Question 2

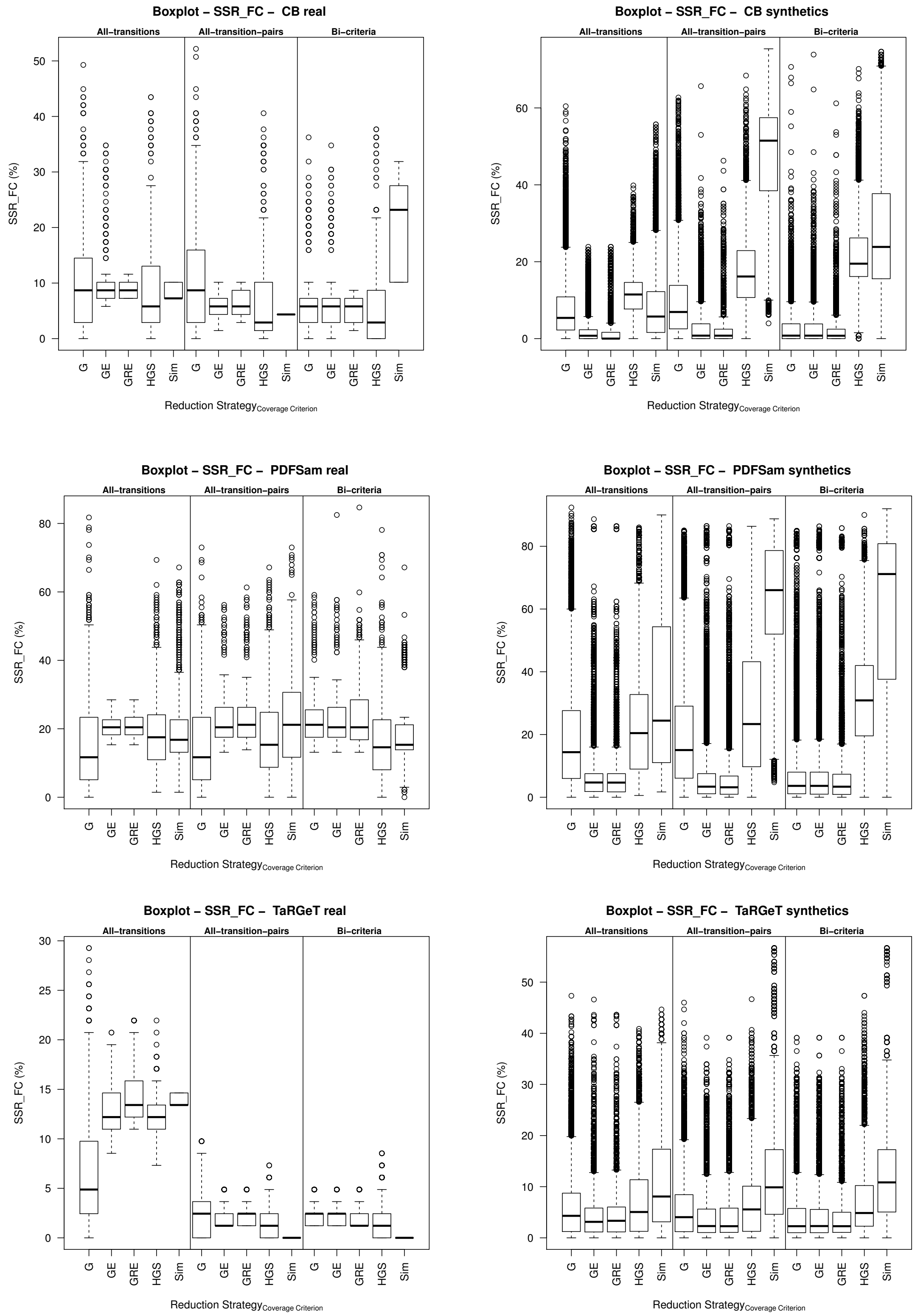


Figure A.10: *Boxplots considering SSR_FC metric for SQ2*

Study Question 3

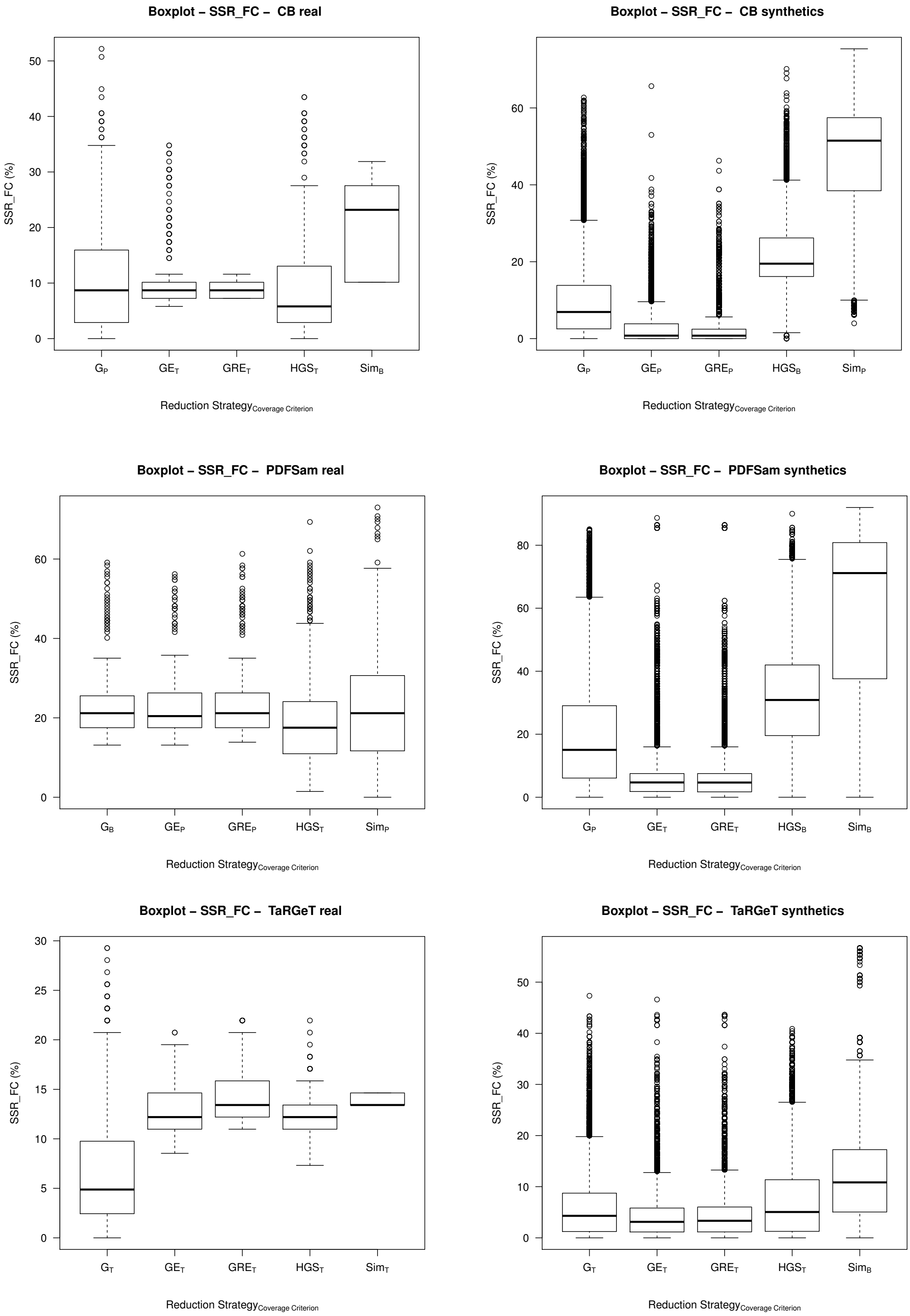


Figure A.11: *Boxplots considering SSR_FC metric for SQ3*

Study Question 4

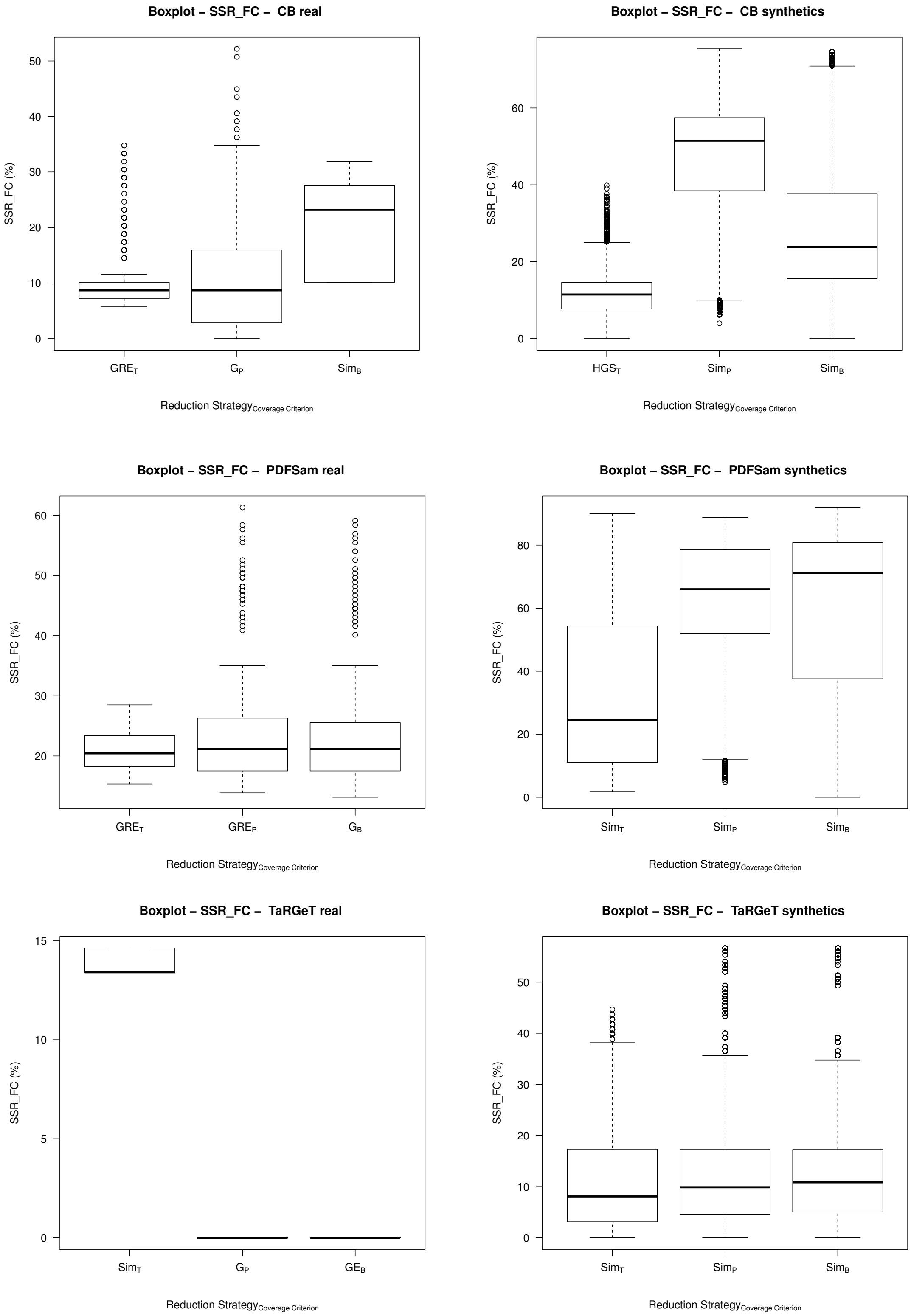


Figure A.12: *Boxplots considering SSR_FC metric for SQ4*

A.7.4 Mann-Whitney test and $\hat{A}_{12}$ effect size measurement

Study Question 1

Table A.51: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB configuration*

	CB real			CB synthetics		
Comparison	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and G_P	0.432	G_P	Small (0.4916)	4.565e-164	G_P	Small (0.4395)
G_T and G_B	2.658e-26	G_T	Medium (0.6039)	0.000	G_T	Large (0.7386)
G_P and G_B	6.402e-29	G_P	Medium (0.6122)	0.000	G_P	Large (0.7758)
GE_T and GE_P	1.552e-121	GE_T	Large (0.8397)	0.000	GE_P	Small (0.4446)
GE_T and GE_B	1.172e-66	GE_T	Large (0.8151)	0.000	GE_B	Small (0.4509)
GE_P and GE_B	0.2859	GE_P	Small (0.5097)	8.147e-1	GE_P	Small (0.5069)
GRE_T and GRE_P	5.46e-118	GRE_T	Large (0.8165)	0.000	GRE_P	Small (0.4542)
GRE_T and GRE_B	2.994e-144	GRE_T	Large (0.9142)	0.000	GRE_B	Small (0.4573)
GRE_P and GRE_B	2.468e-19	GRE_P	Medium (0.62)	1.616e-19	GRE_P	Small (0.5038)
HGS_T and HGS_P	1.36e-2	HGS_T	Medium (0.6533)	0.000	HGS_P	Large (0.2972)
HGS_T and HGS_B	4.241e-29	HGS_T	Large (0.6858)	0.000	HGS_B	Large (0.1426)
HGS_P and HGS_B	0.02356	HGS_P	Small (0.542)	0.000	HGS_B	Medium (0.3644)
Sim_T and Sim_P	4.722e-185	Sim_T	Large (1.000)	0.000	Sim_P	Large (0.0249)
Sim_T and Sim_B	1.834e-15	Sim_B	Large (0.05431)	0.000	Sim_B	Large (0.1833)
Sim_P and Sim_B	1.127e-169	Sim_B	Large (0)	0.000	Sim_P	Large (0.7811)

Table A.52: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam configuration*

	PDFSam real			**PDFSam synthetics**		
Comparison	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and G_P	0.1127	G_T	Small (0.5145)	1.189e-1	G_P	Small (0.4879)
G_T and G_B	5.423e-35	G_B	Large (0.2946)	0.000	G_T	Large (0.7596)
G_P and G_B	1.005e-46	G_B	Large (0.2789)	0.000	G_P	Large (0.7664)
GE_T and GE_P	3.189e-08	GE_P	Small (0.4608)	0.000	GE_T	Small (0.545)
GE_T and GE_B	3.219e-07	GE_B	Small (0.4629)	1.67e-273	GE_T	Small (0.5417)
GE_P and GE_B	0.9564	GE_P	Small (0.5034)	0.3629	GE_B	Small (0.4981)
GRE_T and GRE_P	2.086e-08	GRE_P	Small (0.4608)	0.000	GRE_T	Small (0.5646)
GRE_T and GRE_B	3.492e-08	GRE_B	Small (0.4861)	0.000	GRE_T	Small (0.5592)
GRE_P and GRE_B	0.9229	GRE_P	Small (0.5174)	0.3854	GRE_B	Small (0.4957)
HGS_T and HGS_P	0.02755	HGS_T	Small (0.5408)	2.409e-89	HGS_P	Small (0.4584)
HGS_T and HGS_B	4.143e-06	HGS_T	Small (0.5662)	0.000	HGS_B	Medium (0.3523)
HGS_P and HGS_B	0.02249	HGS_P	Small (0.5254)	0.000	HGS_B	Small (0.4127)
Sim_T and Sim_P	0.001127	Sim_P	Small (0.4493)	0.000	Sim_P	Large (0.1944)
Sim_T and Sim_B	0.2842	Sim_T	Small (0.5235)	0.000	Sim_B	Large (0.2474)
Sim_P and Sim_B	2.647e-05	Sim_P	Small (0.5607)	7.766e-2	Sim_B	Small (0.4814)

Table A.53: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT configuration*

	TaRGeT real			**TaRGeT synthetics**		
Comparison	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
G_T and G_P	5.483e-63	G_T	Large (0.714)	5.976e-09	G_T	Small (0.511)
G_T and G_B	3.726e-94	G_T	Large (0.7359)	0.000	G_T	Medium (0.6106)
G_P and G_B	8.323e-07	G_P	Small (0.5068)	0.000	G_P	Medium (0.6008)
GE_T and GE_P	7.015e-166	GE_T	Large (1.000)	6.608e-67	GE_T	Small (0.541)
GE_T and GE_B	7.249e-166	GE_T	Large (1.000)	6.949e-77	GE_T	Small (0.5441)
GE_P and GE_B	0.05218	GE_B	Small (0.4763)	0.02462	GE_P	Small (0.5035)
GRE_T and GRE_P	7.013e-166	GRE_T	Large (1.000)	1.814e-89	GRE_T	Small (0.5507)
GRE_T and GRE_B	5.055e-166	GRE_T	Large (1.000)	1.802e-183	GRE_T	Small (0.561)
GRE_P and GRE_B	0.477	GRE_P	Small (0.5076)	2.95e-13	GRE_P	Small (0.5077)
HGS_T and HGS_P	2.131e-166	HGS_T	Large (1.000)	4.57e-42	HGS_T	Small (0.5123)
HGS_T and HGS_B	2.556e-166	HGS_T	Large (0.9999)	2.349e-2	HGS_T	Small (0.504)
HGS_P and HGS_B	0.003626	HGS_B	Small (0.4536)	0.07147	HGS_B	Small (0.4932)
Sim_T and Sim_P	4.198e-176	Sim_T	Large (1.000)	4.799e-86	Sim_P	Small (0.4567)
Sim_T and Sim_B	4.198e-176	Sim_T	Large (1.000)	2.438e-17	Sim_B	Small (0.4317)
Sim_P and Sim_B	NaN	None	NO effect (0.5)	0.01233	Sim_B	Small (0.4734)

Study Question 2

Table A.54: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB configuration*

	Comparison	CB real			CB synthetics		
		ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
All-transitions (T)	G_T and GE_T	0.2219	GE_T	Small (0.4455)	0.000	G_T	Large (0.7931)
	G_T and GRE_T	0.8146	GRE_T	Small (0.4479)	0.000	G_T	Large (0.8194)
	G_T and HGS_T	7.595e-05	G_T	Small (0.5388)	0.000	HGS_T	Large (0.2928)
	G_T and Sim_T	5.938e-06	G_T	Small (0.5064)	6.599e-24	Sim_T	Small (0.4877)
	GE_T and GRE_T	0.01419	GRE_T	Small (0.4443)	3.251e-52	GE_T	Small (0.5316)
	GE_T and HGS_T	5.152e-12	GE_T	Large (0.6711)	0.000	HGS_T	Large (0.07919)
	GE_T and Sim_T	6.832e-50	GE_T	Medium (0.6644)	0.000	Sim_T	Large (0.2007)
	GRE_T and HGS_T	4.492e-09	GRE_T	Large (0.6729)	0.000	HGS_T	Large (0.0636)
	GRE_T and Sim_T	6.509e-81	GRE_T	Large (0.7457)	0.000	Sim_T	Large (0.1746)
	HGS_T and Sim_T	0.2563	Sim_T	Medium (0.377)	0.000	HGS_T	Large (0.6759)
All-transition-pairs (P)	G_P and GE_P	8.653e-38	G_P	Medium (0.6223)	0.000	G_P	Large (0.7695)
	G_P and GRE_P	2.592e-27	G_P	Small (0.5851)	0.000	G_P	Large (0.8071)
	G_P and HGS_P	6.835e-35	G_P	Medium (0.6638)	0.000	HGS_P	Large (0.2483)
	G_P and Sim_P	7.164e-80	G_P	Large (0.679)	0.000	Sim_P	Large (0.02664)
	GE_P and GRE_P	8.946e-11	GRE_P	Small (0.4179)	1.174e-80	GE_P	Small (0.5392)
	GE_P and HGS_P	0.002813	GE_P	Medium (0.6172)	0.000	HGS_P	Large (0.07896)
	GE_P and Sim_P	2.666e-73	GE_P	Medium (0.661)	0.000	Sim_P	Large (0.005023)
	GRE_P and HGS_P	1.207e-10	GRE_P	Medium (0.656)	0.000	HGS_P	Large (0.06048)
	GRE_P and Sim_P	1.487e-115	GRE_P	Large (0.7385)	0.000	Sim_P	Large (0.003678)
	HGS_P and Sim_P	0.000365	Sim_P	Small (0.407)	0.000	Sim_P	Large (0.05977)
Bi-criteria (B)	G_B and GE_B	0.4647	GE_B	Small (0.4894)	0.4564	G_B	Small (0.5013)
	G_B and GRE_B	0.03687	G_B	Small (0.5098)	5.018e-95	G_B	Small (0.5379)
	G_B and HGS_B	3.981e-11	G_B	Medium (0.6481)	0.000	HGS_B	Large (0.03116)
	G_B and Sim_B	2.459e-142	Sim_B	Large (0.04454)	0.000	Sim_B	Large (0.0668)
	GE_B and GRE_B	0.002379	GE_B	Small (0.5211)	1.126e-88	GE_B	Small (0.5366)
	GE_B and HGS_B	1.858e-12	GE_B	Medium (0.6535)	0.000	HGS_B	Large (0.03071)
	GE_B and Sim_B	2.847e-139	Sim_B	Large (0.0482)	0.000	Sim_B	Large (0.06636)
	GRE_B and HGS_B	3.318e-06	GRE_B	Medium (0.6353)	0.000	HGS_B	Large (0.02158)
	GRE_B and Sim_B	2.272e-165	Sim_B	Large (0)	0.000	Sim_B	Large (0.05381)
	HGS_B and Sim_B	1.424e-143	Sim_B	Large (0.08941)	0.000	Sim_B	Small (0.4106)

Table A.55: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam configuration*

	Comparison	PDFSam real ρ-value	PDFSam real Superior	PDFSam real Effect Size	PDFSam synthetics ρ-value	PDFSam synthetics Superior	PDFSam synthetics Effect Size
All-transitions (T)	G_T and GE_T	1.566e-26	GE_T	Large (0.3126)	0.000	G_T	Large (0.7553)
	G_T and GRE_T	5.114e-28	GRE_T	Large (0.3082)	0.000	G_T	Large (0.7547)
	G_T and HGS_T	1.617e-08	HGS_T	Medium (0.3927)	1.172e-276	HGS_T	Small (0.4122)
	G_T and Sim_T	2.567e-15	Sim_T	Medium (0.3703)	0.000	Sim_T	Medium (0.3389)
	GE_T and GRE_T	0.1498	GRE_T	Small (0.4819)	0.08026	GE_T	Small (0.5031)
	GE_T and HGS_T	7.578e-12	GE_T	Medium (0.6199)	0.000	HGS_T	Large (0.1595)
	GE_T and Sim_T	4.564e-07	GE_T	Medium (0.6547)	0.000	Sim_T	Large (0.1271)
	GRE_T and HGS_T	3.398e-13	GRE_T	Medium (0.6278)	0.000	HGS_T	Large (0.1612)
	GRE_T and Sim_T	2.144e-07	GRE_T	Medium (0.6615)	0.000	Sim_T	Large (0.1293)
	HGS_T and Sim_T	0.09227	Sim_T	Small (0.4825)	0.000	Sim_T	Small (0.4103)
All-transition-pairs (P)	G_P and GE_P	4.638e-46	GE_P	Large (0.2772)	0.000	G_P	Large (0.77)
	G_P and GRE_P	2.139e-48	GRE_P	Large (0.2718)	0.000	G_P	Large (0.7849)
	G_P and HGS_P	2.926e-08	HGS_P	Small (0.4102)	0.000	HGS_P	Medium (0.3859)
	G_P and Sim_P	3.9e-30	Sim_P	Medium (0.3424)	0.000	Sim_P	Large (0.06251)
	GE_P and GRE_P	0.5146	GRE_P	Small (0.4866)	1.008e-07	GE_P	Small (0.5238)
	GE_P and HGS_P	1.17e-23	GE_P	Large (0.6737)	0.000	HGS_P	Large (0.1533)
	GE_P and Sim_P	0.2397	GE_P	Small (0.5222)	0.000	Sim_P	Large (0.01828)
	GRE_P and HGS_P	1.082e-25	GRE_P	Large (0.6813)	0.000	HGS_P	Large (0.1419)
	GRE_P and Sim_P	0.07364	GRE_P	Small (0.53)	0.000	Sim_P	Large (0.01516)
	HGS_P and Sim_P	8.42e-12	Sim_P	Small (0.408)	0.000	Sim_P	Large (0.1027)
Bi-criteria (B)	G_B and GE_B	0.8799	G_B	Small (0.5021)	0.9922	G_B	Small (0.5016)
	G_B and GRE_B	0.1018	G_B	Small (0.5006)	3.128e-10	G_B	Small (0.5224)
	G_B and HGS_B	1.103e-36	G_B	Large (0.6972)	0.000	HGS_B	Large (0.1041)
	G_B and Sim_B	1.229e-11	G_B	Large (0.7003)	0.000	Sim_B	Large (0.06339)
	GE_B and GRE_B	0.2069	GRE_B	Small (0.4991)	6.299e-10	GE_B	Small (0.5209)
	GE_B and HGS_B	9.854e-37	GE_B	Large (0.6969)	0.000	HGS_B	Large (0.1031)
	GE_B and Sim_B	4.26e-11	GE_B	Large (0.6983)	0.000	Sim_B	Large (0.06301)
	GRE_B and HGS_B	5.968e-40	GRE_B	Large (0.701)	0.000	HGS_B	Large (0.08912)
	GRE_B and Sim_B	4.707e-13	GRE_B	Large (0.6865)	0.000	Sim_B	Large (0.05843)
	HGS_B and Sim_B	1.576e-09	Sim_B	Small (0.4147)	0.000	Sim_B	Large (0.2053)

Table A.56: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT configuration*

		TaRGeT real			**TaRGeT synthetics**		
	Comparison	ρ-value	Superior	Effect Size	ρ-value	Superior	Effect Size
All-transitions (T)	G_T and GE_T	2.725e-98	GE_T	Large (0.1731)	0.000	G_T	Small (0.5815)
	G_T and GRE_T	2.682e-114	GRE_T	Large (0.1402)	0.000	G_T	Small (0.5723)
	G_T and HGS_T	9.196e-95	HGS_T	Large (0.1755)	7.519e-147	HGS_T	Small (0.4519)
	G_T and Sim_T	7.541e-120	Sim_T	Large (0.1315)	0.000	Sim_T	Medium (0.3459)
	GE_T and GRE_T	5.105e-21	GRE_T	Medium (0.3666)	1.198e-22	GRE_T	Small (0.4876)
	GE_T and HGS_T	0.001992	GE_T	Small (0.5148)	0.000	HGS_T	Medium (0.375)
	GE_T and Sim_T	2.874e-27	Sim_T	Large (0.3096)	0.000	Sim_T	Large (0.2719)
	GRE_T and HGS_T	8.884e-45	GRE_T	Medium (0.6664)	0.000	HGS_T	Medium (0.3828)
	GRE_T and Sim_T	0.08093	Sim_T	Small (0.4412)	0.000	Sim_T	Large (0.2784)
	HGS_T and Sim_T	2.923e-78	Sim_T	Large (0.2006)	0.000	Sim_T	Medium (0.3953)
All-transition-pairs (P)	G_P and GE_P	2.087e-09	G_P	Small (0.5181)	0.000	G_P	Small (0.5992)
	G_P and GRE_P	2.323e-07	G_P	Small (0.5067)	0.000	G_P	Small (0.5973)
	G_P and HGS_P	7.648e-44	G_P	Medium (0.6644)	8.047e-115	HGS_P	Small (0.451)
	G_P and Sim_P	9.003e-125	G_P	Large (0.872)	0.000	Sim_P	Large (0.2813)
	GE_P and GRE_P	0.112	GRE_P	Small (0.4784)	0.1711	GRE_P	Small (0.4976)
	GE_P and HGS_P	1.224e-40	GE_P	Large (0.7092)	0.000	HGS_P	Medium (0.3588)
	GE_P and Sim_P	7.232e-173	GE_P	Large (1.000)	0.000	Sim_P	Large (0.2051)
	GRE_P and HGS_P	5.721e-46	GRE_P	Large (0.7209)	0.000	HGS_P	Medium (0.3609)
	GRE_P and Sim_P	9.904e-172	GRE_P	Large (1.000)	0.000	Sim_P	Large (0.2063)
	HGS_P and Sim_P	5.292e-99	HGS_P	Large (0.7845)	0.000	Sim_P	Large (0.3283)
Bi-criteria (B)	G_B and GE_B	0.9461	GE_B	Small (0.4971)	0.2561	G_B	Small (0.5009)
	G_B and GRE_B	0.4447	G_B	Small (0.5066)	1.878e-05	G_B	Small (0.5026)
	G_B and HGS_B	5.835e-26	G_B	Large (0.6892)	0.000	HGS_B	Medium (0.3413)
	G_B and Sim_B	8.504e-172	G_B	Large (1.000)	0.000	Sim_B	Large (0.1856)
	GE_B and GRE_B	0.4314	GE_B	Small (0.5097)	0.0003883	GE_B	Small (0.5018)
	GE_B and HGS_B	2.701e-26	GE_B	Large (0.6917)	0.000	HGS_B	Medium (0.3392)
	GE_B and Sim_B	1.257e-171	GE_B	Large (1.000)	0.000	Sim_B	Large (0.1833)
	GRE_B and HGS_B	9.064e-26	GRE_B	Large (0.6857)	0.000	HGS_B	Medium (0.3322)
	GRE_B and Sim_B	4.412e-172	GRE_B	Large (1.000)	0.000	Sim_B	Large (0.1747)
	HGS_B and Sim_B	7.523e-124	HGS_B	Large (0.8485)	0.000	Sim_B	Large (0.3049)

Study Question 3

Table A.57: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB configuration*

CB real				CB synthetic			
Comparison	ρ-value	Superior	Effect Size	Comparison	ρ-value	Superior	Effect Size
G_P and GE_T	0.7304	GE_T	Small (0.4519)	G_P and GE_P	0.000	G_P	Large (0.7695)
G_P and GRE_T	0.5743	GRE_T	Small (0.4542)	G_P and GRE_P	0.000	G_P	Large (0.8071)
G_P and HGS_T	1.557e-06	G_P	Small (0.5481)	G_P and HGS_B	0.000	HGS_B	Large (0.1601)
G_P and Sim_B	5.256e-74	Sim_B	Large (0.229)	G_P and Sim_P	0.000	Sim_P	Large (0.02664)
GE_T and GRE_T	0.01419	GRE_T	Small (0.4443)	GE_P and GRE_P	1.174e-80	GE_P	Small (0.5392)
GE_T and HGS_T	5.152e-12	GE_T	Large (0.6711)	GE_P and HGS_B	0.000	HGS_B	Large (0.03706)
GE_T and Sim_B	1.405e-109	Sim_B	Large (0.1428)	GE_P and Sim_P	0.000	Sim_P	Large (0.005023)
GRE_T and HGS_T	4.492e-09	GRE_T	Large (0.6729)	GRE_P and HGS_B	0.000	HGS_B	Large (0.02707)
GRE_T and Sim_B	2.839e-126	Sim_B	Large (0.1567)	GRE_P and Sim_P	0.000	Sim_P	Large (0.003678)
HGS_T and Sim_B	2.327e-100	Sim_B	Large (0.1474)	HGS_B and Sim_P	0.000	Sim_P	Large (0.08563)

Table A.58: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam configuration*

PDFSam real				PDFSam synthetic			
Comparison	ρ-value	Superior	Effect Size	Comparison	ρ-value	Superior	Effect Size
G_B and GE_P	0.5339	GE_P	Small (0.499)	G_P and GE_T	0.000	G_P	Large (0.7624)
G_B and GRE_P	0.2047	GRE_P	Small (0.4859)	G_P and GRE_T	0.000	G_P	Large (0.7619)
G_B and HGS_T	1.645e-20	G_B	Medium (0.6437)	G_P and HGS_B	0.000	HGS_B	Large (0.2891)
G_B and Sim_P	0.497	G_B	Small (0.5197)	G_P and Sim_B	0.000	Sim_B	Large (0.136)
GE_P and GRE_P	0.5146	GRE_P	Small (0.4866)	GE_T and GRE_T	0.08026	GE_T	Small (0.5031)
GE_P and HGS_T	1.034e-20	GE_P	Medium (0.6444)	GE_T and HGS_B	0.000	HGS_B	Large (0.0917)
GE_P and Sim_P	0.2397	GE_P	Small (0.5222)	GE_T and Sim_B	0.000	Sim_B	Large (0.05781)
GRE_P and HGS_T	8.794e-22	GRE_P	Medium (0.6523)	GRE_T and HGS_B	0.000	HGS_B	Large (0.0915)
GRE_P and Sim_P	0.07364	GRE_P	Small (0.53)	GRE_T and Sim_B	0.000	Sim_B	Large (0.05754)
HGS_T and Sim_P	2.545e-08	Sim_P	Small (0.4314)	HGS_B and Sim_B	0.000	Sim_B	Large (0.2053)

Table A.59: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT configuration*

TaRGeT real				TaRGeT synthetic			
Comparison	ρ-value	Superior	Effect Size	Comparison	ρ-value	Superior	Effect Size
G_T and GE_T	2.725e-98	GE_T	Large (0.1731)	G_T and GE_T	0.000	G_T	Small (0.5815)
G_T and GRE_T	2.682e-114	GRE_T	Large (0.1402)	G_T and GRE_T	0.000	G_T	Small (0.5723)
G_T and HGS_T	9.196e-95	HGS_T	Large (0.1755)	G_T and HGS_T	7.519e-147	HGS_T	Small (0.4519)
G_T and Sim_T	7.541e-120	Sim_T	Large (0.1315)	G_T and Sim_B	0.000	Sim_B	Large (0.2677)
GE_T and GRE_T	5.105e-21	GRE_T	Medium (0.3666)	GE_T and GRE_T	1.198e-22	GRE_T	Small (0.4876)
GE_T and HGS_T	0.001992	GE_T	Small (0.5148)	GE_T and HGS_T	0.000	HGS_T	Medium (0.375)
GE_T and Sim_T	2.874e-27	Sim_T	Large (0.3096)	GE_T and Sim_B	0.000	Sim_B	Large (0.191)
GRE_T and HGS_T	8.884e-45	GRE_T	Medium (0.6664)	GRE_T and HGS_T	0.000	HGS_T	Medium (0.3828)
GRE_T and Sim_T	0.08093	Sim_T	Small (0.4412)	GRE_T and Sim_B	0.000	Sim_B	Large (0.1952)
HGS_T and Sim_T	2.923e-78	Sim_T	Large (0.2006)	HGS_T and Sim_B	0.000	Sim_B	Large (0.3223)

Study Question 4

Table A.60: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for CB configuration*

CB real				CB synthetic			
Comparison	ρ-value	Superior	Effect Size	Comparison	ρ-value	Superior	Effect Size
GRE_T and G_P	0.7304	GRE_T	Small (0.5481)	HGS_T and Sim_P	0.000	Sim_P	Large (0.01975)
GRE_T and Sim_B	1.405e-109	Sim_B	Large (0.1428)	HGS_T and Sim_B	0.000	Sim_B	Large (0.211)
G_P and Sim_B	5.256e-74	Sim_B	Large (0.229)	Sim_P and Sim_B	0.000	Sim_P	Large (0.7811)

Table A.61: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for PDFSam configuration*

PDFSam real				PDFSam synthetic			
Comparison	ρ-value	Superior	Effect Size	Comparison	ρ-value	Superior	Effect Size
GRE_T and GRE_P	2.086e-08	GRE_P	Small (0.4608)	Sim_T and Sim_P	0.000	Sim_P	Large (0.1944)
GRE_T and G_B	2.94e-05	G_B	Small (0.4703)	Sim_T and Sim_B	0.000	Sim_B	Large (0.2474)
GRE_P and G_B	0.2047	GRE_P	Small (0.5141)	Sim_P and Sim_B	7.766e-20	Sim_B	Small (0.4814)

Table A.62: *Mann-Whitney and $\hat{A}_{12}$ effect size measurements for TaRGeT configuration*

TaRGeT real				TaRGeT synthetic			
Comparison	ρ-value	Superior	Effect Size	Comparison	ρ-value	Superior	Effect Size
Sim_T and G_P	2.756e-166	Sim_T	Large (1.000)	Sim_T and Sim_P	4.799e-86	Sim_P	Small (0.4567)
Sim_T and GE_B	4.225e-169	Sim_T	Large (1.000)	Sim_T and Sim_B	2.438e-170	Sim_B	Small (0.4317)
G_P and GE_B	4.986e-07	G_P	Small (0.506)	Sim_P and Sim_B	0.01233	Sim_B	Small (0.4734)

A.7.5 The minimum, maximum, median and average

Table A.63: *The minimum, maximum, median and average for CB real*

All-transitions (T)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_T	0.00	8.696	49.28	10.21	9.136
GE_T	5.797	8.696	34.78	9.935	4.656
GRE_T	7.246	8.696	11.59	9.386	1.613
HGS_T	0.00	5.797	43.48	8.649	7.911
Sim_T	7.246	7.246	10.14	7.977	1.259
All-transition-pairs (P)					
G_P	0.00	8.696	52.17	10.54	9.36
GE_P	1.449	5.797	10.14	5.672	2.475
GRE_P	2.899	5.797	10.14	6.441	2.476
HGS_P	0.00	2.899	40.58	5.878	6.964
Sim_P	4.348	4.348	4.348	4.348	0.00
Bi-criteria (B)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_B	0.00	5.797	36.23	6.361	5.081
GE_B	0.00	5.797	34.78	6.525	5.215
GRE_B	1.449	5.797	8.696	5.326	2.301
HGS_B	0.00	2.899	37.68	5.161	6.468
Sim_B	10.14	23.19	31.88	19.3	8.253

Table A.64: *The minimum, maximum, median and average for CB synthetics*

All-transitions (T)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_T	0.00	5.385	60.45	7.602	7.43
GE_T	0.00	0.7463	23.88	2.184	3.825
GRE_T	0.00	0.00	23.88	1.787	3.422
HGS_T	0.00	11.48	39.84	11.23	5.222
Sim_T	0.00	5.738	55.73	8.724	9.198
All-transition-pairs (P)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_P	0.00	6.923	62.71	9.587	9.042
GE_P	0.00	0.7692	65.67	3.36	5.281
GRE_P	0.00	0.7634	46.27	2.65	4.701
HGS_P	0.00	16.15	68.42	17.08	8.809
Sim_P	3.968	51.49	75.37	47.26	14.67
Bi-criteria (B)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_B	0.00	0.7692	70.68	3.189	5.016
GE_B	0.00	0.7692	73.88	3.161	4.97
GRE_B	0.00	0.7463	61.19	2.521	4.378
HGS_B	0.00	19.49	70.15	21.05	8.402
Sim_B	0.00	23.85	74.63	27.73	18.88

Table A.65: *The minimum, maximum, median and average for PDFSam real*

All-transitions (T)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_T	0.00	11.68	81.75	16.34	14.43
GE_T	15.33	20.44	28.47	20.61	3.108
GRE_T	15.33	20.44	28.47	20.84	3.182
HGS_T	1.46	17.52	69.34	19.15	11.26
Sim_T	1.46	16.79	67.15	20.78	12.93
All-transition-pairs (P)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_P	0.00	11.68	72.99	15.36	13.28
GE_P	13.14	20.44	56.2	22.49	7.038
GRE_P	13.87	21.17	61.31	22.95	7.643
HGS_P	0.00	15.33	67.15	18.41	12.71
Sim_P	0.00	21.17	72.99	22.49	14.16
Bi-criteria (B)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_B	13.14	21.17	59.12	22.37	6.913
GE_B	13.14	20.44	82.48	22.41	7.151
GRE_B	13.14	20.44	84.67	23.03	8.279
HGS_B	0.00	14.6	78.1	17.06	11.83
Sim_B	0.00	15.33	67.15	20.34	12.03

Table A.66: *The minimum, maximum, median and average for PDFSam synthetics*

All-transitions (T)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_T	0.00	14.36	92.38	18.91	16.47
GE_T	0.00	4.698	88.67	7.081	9.18
GRE_T	0.00	4.667	86.41	7.084	9.048
HGS_T	0.5618	20.44	86	24.05	18.45
Sim_T	1.685	24.42	90	33.3	25.94
All-transition-pairs (P)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_P	0.00	15.04	85.08	19.88	17.4
GE_P	0.00	3.371	86.49	7.489	11.81
GRE_P	0.00	3.175	86.45	6.834	11
HGS_P	0.00	23.33	86.36	27.15	19.55
Sim_P	4.762	66.02	88.76	62.87	18.1
Bi-criteria (B)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_B	0.00	3.623	84.96	7.659	12.02
GE_B	0.00	3.623	86.36	7.601	11.91
GRE_B	0.00	3.361	85.81	6.798	10.91
HGS_B	0.00	30.87	90	31.56	16.22
Sim_B	0.00	71.15	91.98	59.61	26.77

Table A.67: *The minimum, maximum, median and average for TaRGeT real*

All-transitions (T)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_T	0.00	4.878	29.27	6.76	6.095
GE_T	8.537	12.2	20.73	12.79	2.683
GRE_T	10.98	13.41	21.95	14.07	2.726
HGS_T	7.317	12.2	21.95	12.44	1.832
Sim_T	13.41	13.41	14.63	14	0.6097
All-transition-pairs (P)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_P	0.00	2.439	9.756	2.612	2.432
GE_P	1.22	1.22	4.878	2.054	1.057
GRE_P	1.22	2.439	4.878	2.133	1.072
HGS_P	0.00	1.22	7.317	1.237	1.446
Sim_P	0.00	0.00	0.00	0.00	0.00
Bi-criteria (B)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_B	1.22	2.439	4.878	2.137	1.09
GE_B	1.22	2.439	4.878	2.134	1.056
GRE_B	1.22	1.22	4.878	2.095	1.039
HGS_B	0.00	1.22	8.537	1.495	1.671
Sim_B	0.00	0.00	0.00	0.00	0.00

Table A.68: *The minimum, maximum, median and average for TaRGeT synthetics*

All-transitions (T)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_T	0.00	4.301	47.33	5.999	6.033
GE_T	0.00	3.125	46.6	4.127	4.212
GRE_T	0.00	3.333	43.62	4.2	4.104
HGS_T	0.00	5.05	40.87	7.552	7.516
Sim_T	0.00	8.081	44.66	10.84	9.433
All-transition-pairs (P)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_P	0.00	4.04	46	5.713	5.755
GE_P	0.00	2.299	39.13	4.068	4.965
GRE_P	0.00	2.273	39.13	4.091	4.982
HGS_P	0.00	5.556	46.67	6.817	6.417
Sim_P	0.00	9.876	56.67	12.08	9.982
Bi-criteria (B)					
Strategy	Minimum	Median	Maximum	Average	Standard Deviation
G_B	0.00	2.273	39.13	4.058	4.982
GE_B	0.00	2.299	39.13	3.997	4.874
GRE_B	0.00	2.273	39.13	3.796	4.54
HGS_B	0.00	4.854	47.33	6.93	6.505
Sim_B	0.00	10.84	56.67	12.51	9.171

A.7.6 Ordering of effectiveness

Study Question 1

Table A.69: *Ordering of effectiveness for each reduction strategy associated with all coverage criteria*

CB real	**CB synthetics**
$G_P > G_T > G_B$	$G_P > G_T > G_B$
$GE_T > GE_P > GE_B$	$GE_P > GE_B > GE_T$
$GRE_T > GRE_P > GRE_B$	$GRE_P > GRE_B > GRE_T$
$HGS_T > HGS_P > HGS_B$	$HGS_B > HGS_P > HGS_T$
$Sim_B > Sim_T > Sim_B$	$Sim_P > Sim_B > Sim_T$
PDFSam real	**PDFSam synthetics**
$G_B > G_T > G_P$	$G_P > G_T > G_B$
$GE_P > GE_B > GE_T$	$GE_T > GE_B > GE_P$
$GRE_P > GRE_B > GRE_T$	$GRE_T > GRE_B > GRE_P$
$HGS_T > HGS_P > HGS_B$	$HGS_B > HGS_P > HGS_T$
$Sim_P > Sim_T > Sim_B$	$Sim_B > Sim_P > Sim_T$
TaRGeT real	**TaRGeT synthetics**
$G_T > G_P > G_B$	$G_T > G_P > G_B$
$GE_T > GE_B > GE_P$	$GE_T > GE_P > GE_B$
$GRE_T > GRE_P > GRE_B$	$GRE_T > GRE_P > GRE_B$
$HGS_T > HGS_B > HGS_P$	$HGS_T > HGS_B > HGS_P$
$Sim_T > Sim_P = Sim_B$	$Sim_B > Sim_P > Sim_T$

Study Question 2

Table A.70: *Ordering of effectiveness among reduction strategies for each coverege criterion*

CB real	CB synthetics
$GRE_T > GE_T > G_T > Sim_T > HGS_T$	$HGS_T > Sim_T > G_T > GE_T > GRE_T$
$G_P > GRE_P > GE_P > Sim_P > HGS_P$	$Sim_P > HGS_P > G_P > GE_P > GRE_P$
$Sim_B > GE_B > G_B > GRE_B > HGS_B$	$Sim_B > HGS_B > G_B > GE_B > GRE_B$
PDFSam real	**PDFSam synthetics**
$GE_T > GRE_T > Sim_T > HGS_T > G_T$	$Sim_T > HGS_T > G_T > GE_T > GRE_T$
$GRE_P > GE_P > Sim_P > HGS_P > G_P$	$Sim_P > HGS_P > G_P > GE_P > GRE_P$
$G_B > GRE_B > GE_B > Sim_B > HGS_B$	$Sim_B > HGS_B > G_B > GE_B > GRE_B$
TaRGeT real	**TaRGeT synthetics**
$Sim_T > GRE_T > GE_T > HGS_T > G_T$	$Sim_T > HGS_T > G_T > GRE_T > GE_T$
$G_P > GRE_P > GE_P > HGS_P > Sim_P$	$Sim_P > HGS_P > G_P > GRE_P > GE_P$
$GE_B > G_B > GRE_B > HGS_B > Sim_B$	$Sim_B > HGS_B > G_B > GE_B > GRE_B$

Study Question 3

Table A.71: *Ordering of effectiveness reduction strategies in combination with their best coverage criterion*

CB real	$Sim_B > GRE_T > GE_T > G_T > HGS_T$
CB synthetics	$Sim_P > HGS_B > G_P > GE_P > GRE_P$
PDFSam real	$GRE_P > GE_B > G_B > Sim_P > HGS_T$
PDFSam synthetics	$Sim_B > HGS_B > G_P > GE_T > GRE_T$
TaRGeT real	$Sim_T > GRE_T > GE_T > HGS_T > G_T$
TaRGeT synthetics	$Sim_B > HGS_T > G_T > GRE_T > GE_T$

Study Question 4

Table A.72: *Ordering of effectiveness coverage criteria in combination with their best reduction strategy*

CB real	$Sim_B > GRE_T > G_P$
CB synthetics	$Sim_P > Sim_B > HGS_T$
PDFSam real	$GRE_P > G_B > GE_T$
PDFSam synthetics	$Sim_B > Sim_P > Sim_T$
TaRGeT real	$Sim_T > G_P > GE_B$
TaRGeT synthetics	$Sim_B > Sim_P > Sim_T$