MISSING DATA IMPUTATION, OPTIMIZATION AND LOCALIZED SOFTWARE DEVELOPMENT EFFORT ESTIMATION

MUHAMMAD ARIF SHAH

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/13)

UNIVERSITI TEKNOLOGI MALAYSIA

DECLARATION OF THESIS / UNDERGRADUATE PROJECT REPORT AND COPYRIGHT

Author's full name : MUHAMMAD ARIF SHAH

Date of Birth : 05 October 1986

Title : Missing Data Imputation, Optimization and Localized Software Development Effort Estimation

Academic Session : 2018-2019/2

I declare that this thesis is classified as:

□ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*

☑ **RESTRICTED** (Contains restricted information as specified by the organization where research was done)*

□ **OPEN ACCESS** I agree that my thesis to be published as online open access (full text)

1. I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:
2. The thesis is the property of Universiti Teknologi Malaysia
3. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
4. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

SIGNATURE OF STUDENT	**SIGNATURE OF SUPERVISOR**
MH5145682	ASSOC. PROF. DR. DAYANG NORHAYATI ABANG JAWAWI
Passport number	**NAME OF SUPERVISOR**
Date: 20 JUNE 2019	Date: 20 JUNE 2019

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

Status Declaration Letter

Librarian 20 JUNE 2019

Perpustakaan Sultanah Zanariah
UTM, Skudai
Johor

Respected Sir/Madam,

CLASSIFICATION OF THESIS AS RESTRICTED

MISSING DATA IMPUTATION, OPTIMIZATION AND LOCALIZED SOFTWARE DEVELOPMENT EFFORT ESTIMATION

Please be informed that the above-mentioned thesis entitled — **Missing Data Imputation, Optimization And Localized Software Development Effort Estimation** be classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are:

(i) COMMERCIALIZATION OF RESEARCH RESULTS.
(ii) COPYRIGHT RULES RELATED TO WOS JOURNALS.

Thank you.

Sincerely yours,
ASSOC.PROF.DR. DAYANG NORHAYATI ABANG JAWAWI
N28-305-11
07- 5538870

"We hereby declare that we have read this thesis and in our
opinion this thesis is sufficient in term of scope and quality for the
award of the degree of Doctor of Philosophy in (Computer Science)"

Signature : ____________________________________

Name of Supervisor I : ASSOC. PROF. DR. DAYANG NORHAYATI ABANG JAWAWI

Date : 20 JUNE 2019

Signature : ____________________________________

Name of Supervisor II : DR. MOHD ADHAM BIN ISA

Date : 20 JUNE 2019

BAHAGIAN A - Pengesahan Kerjasama*

Adalah disahkan bahawa projek penyelidikan tesis ini telah dilaksanakan melalui kerjasama antara ________________________dengan ________________________

Disahkan oleh:

Tandatangan : Tarikh :

Nama :

Jawatan :

(Cop rasmi)

** Jika penyediaan tesis atau projek melibatkan kerjasama.*

BAHAGIAN B - Untuk Kegunaan Pejabat Sekolah Pengajian Siswazah

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pcmeriksa Luar **:**

Nama dan Alamat Pcmeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di SPS:

Tandatangan : Tarikh : 15JULAI 2018

Nama :

MISSING DATA IMPUTATION, OPTIMIZATION AND LOCALIZED SOFTWARE DEVELOPMENT EFFORT ESTIMATION

MUHAMMAD ARIF SHAH

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

JUNE 2019

DECLARATION

I declare that this thesis entitled *“Missing Data Imputation, Optimization and Localized Software Development Effort Estimation”* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

Signature : ....
Name : Muhammad Arif Shah
Date : 20 JUNE 2019

DEDICATION

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time.

The sacrifices of my wife (Shahida Arif) and Daughter (Eshal Arif) can never be disregarded, therefore I dedicate my work to them too.

ACKNOWLEDGMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, **Assoc. Prof. Dr. Dayang Norhayati Abang Jawawi,** for encouragement, guidance, criticism, and friendship. I am also very thankful to my co-supervisor **Dr. Mohd Adham** Bin Isa for his guidance, advices, and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

My fellow postgraduate student should also be recognized for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips have been useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am also grateful to all my family members.

ABSTRACT

The accuracy of software development effort estimation is one of the vital factors that leads to successful or failed projects. Most of the current estimation models are not adaptable enough according to the nature of software projects due to their special characteristics, such as intangibility and non-normality of project attributes. Localization-based estimation models solve the issues, but are unable to compare software projects, based on Language Type for accurate effort estimation. Secondly, these models use Analogy-Based Estimation (ABE), which completely depends upon past projects and any missing values that may cause unrealistic estimation results. This study extended the domain of localization to estimate the development effort according to the nature of software and introduced accurate missing data imputation techniques to prevent losing the most similar project. This study focused on ABE, which is a widely accepted non-algorithmic model incorporated in localized estimation. Five estimation models such as Localized Analogy Based Estimation (LABE), artificial Bee colony guided Analogy Based Estimation (BABE), Localized BABE (LBABE), Imputation and Optimization based Effort Estimation (ImOEE), Localized Imputation and Optimization based Effort Estimation (LImOEE) and three missing data imputation techniques such as Median Imputation of the Nearest Neighbours (MINN), Localized Imputation Technique (LIT), and Identical Project based Imputation (IPI) were developed in this study for accurate and unbiased development effort estimation. The techniques accurately filled the missing information in the past projects and the models dealt with the attribute weight optimization, project attribute selection in local space and model comparison. Imputation techniques consist of distance calculation and impute value calibrations with localization. The models are commonly composed by introducing the missing data imputation and soft computing techniques, and ABE. In this research, the models were evaluated using six real datasets. The results were compared with prominent estimation models. A comparative study of the developed models was performed to further validate the accuracy of the results. LImOEE model outperformed the other developed models. LImOEE showed 51%, 25%, 11%, and 31% improvements on Mean Magnitude of Relative Error (MMRE), Percentage of Prediction (PRED), Standard Accuracy (SA) and Effect Size (Δ) respectively for the BABE model. For, LBABE, it showed 37% improvement on MMRE, 12% improvement on PRED (0.25), 5% improvement on SA and 18% improvement on Δ. For the ImOEE model, it showed 26%, 8 %, 9% and 28% improvements on MMRE, PRED, SA and Δ respectively. The results revealed that, accurately imputing the missing data for ABE, optimizing the attribute weights, and extending the scope of localization to the important attributes have significantly improved the accuracy of software development effort estimation.

ABSTRAK

Ketepatan anggaran pembangunan perisian adalah salah satu faktor penting yang menyumbang kepada kejayaan dan kegagalan sesebuah projek. Kebanyakan model angaran terkini tidak sesuai berdasarkan sifat sebenar sesebuah projek berdasarkan ciri-ciri khas seperti atribut projek yang tidak ketara dan tidak normal. Model anggaran berasaskan tempatan telah menyelesaikan isu ini, tetapi tidak dapat membandingkan projek perisian, berdasar kepada jenis bahasa, untuk anggaran usaha yang tepat. Kedua, model ini menggunakan Anggaran Berasaskan-Analogi (ABE) yang keseluruhannya bergantung kepada projek lepas dan sebarang nilai yang hilang yang mungkin menyebabkan hasil anggaran yang tidak realistik. Kajian ini memperluaskan domain tempatan untuk menganggarkan usaha pembangunan berdasarkan sifat semulajadi perisian dan memperkenalkan teknik mengenal pasti taksiran data hilang untuk mengelakkan dari kehilangan kebanyakan projek yang sama. Kajian ini memberi tumpuan kepada ABE, yang merupakan model bukan beralgoritma dan berkerjasama dengan angaran tempatan. Lima model anggaran seperti *Localized Analogy Based Estimation* (LABE), *artificial Bee colony guided Analogy Based Estimation* (BABE), *Localized BABE (LBABE), Imputation and Optimization based Effort Estimation* (ImOEE), *Localized Imputation and Optimization based Effort Estimation* (LImOEE) dan tiga teknik taksiran data hilang seperti *Median Imputation of the Nearest Neighbours* (MINN), *Localized Imputation Technique* (LIT), and *Identical Project based Imputation* (IPI) dicadangkan dalam kajian ini untuk ketepatan dan anggaran perkembangan usaha yang tidak berat sebelah. Teknik yang betul memenuhi taksiran data yang hilang dalam projek lepas dan model yang mengendalikan atribut mengurangkan berat, pemilihan atribut tempatan dan perbandingan model. Teknik taksiran terdiri daripada pengiraan jarak dan menaksir nilai dengan penempatan. Model ini biasanya dibuat dengan memperkenalkan teknik menaksir data hilang dan teknik komputeran lembut, ABE. Dalam kajian ini model telah dinilai menggunakan enam data set asal. Hasil kajian telah dibandingkan dengan model anggaran yang terkenal. Kajian perbandingan untuk model yang dibangunkan telah dilakukan untuk mengesahkan ketepatan hasil kajian. Model LImOEE telah mengatasi model lain yang telah dibangunkan. LImOEE menunjukkan peningkatan sebanyak 51%, 25%, 11% dan 31% untuk *Mean Magnitude of Relative Error* (MMRE), *Percentage of Prediction* (PRED), *Standard Accuracy* (SA) dan *Effect Size* (Δ) untuk model artificial Bee colony guided Analogy Based Estimation (BABE). Untuk LBABE, ia menunjukkan peningkatan 37% ke atas MMRE, peningkatan 12% ke atas PRED (0.25), peningkatan 5% ke atas SA dan peningkatan 18% ke atas Δ. Untuk model ImOEE, ia menunjukkan peningkatan sebanyak 26%, 8%, 9% dan 28% ke atas MMRE, PRED, SA dan Δ. Hasil kajian menunjukkan bahawa ketepatan taksiran data hilang untuk ABE, mengoptimumkan berat atribut dan memperluaskan skop tempatan kepada atribut penting telah meningkatkan ketepatan anggaran pembangunan perisian.

TABLE OF CONTENTS

TITLE		PAGE
DECLARATION		**ii**
DEDICATION		**iii**
Acknowledgment		**iv**
ABSTRACT		**v**
ABSTRAK		**vi**
TABLE OF CONTENTS		**vii**
LIST OF TABLES		**xv**
LIST OF FIGURES		**xvii**
LIST OF ABBREVIATIONS		**xix**
LIST OF APPENDICES		**xxii**
CHAPTER 1	**INTRODUCTION**	**1**
1.1	Overview	1
1.2	Research Background	2
1.3	Problem Statement	7
1.3.1	Domain Problems	7
1.3.2	Technical Problems	7
1.3.2.1	Missing Data in the Historical Project of Software Engineering Datasets	8
1.3.2.2	Project Attribute Weighting	9
1.3.2.3	Unreliable Comparisons	9
1.4	Research Questions	9
1.5	Research Goals	10
1.6	Research Objectives	10
1.7	Research Scope	11
1.8	Research Justification	13

1.8.1 Importance of Software Development Effort Estimation 13

1.8.2 Importance of Localization 14

1.8.3 Importance of dataset completeness 14

1.9 Thesis Organization 15

CHAPTER 2 LITERATURE REVIEW 17

2.1 Introduction 17

2.2 Algorithmic Models 19

2.2.1 Source Lines of Code (SLOC) 19

2.2.2 Function Point Analysis 20

2.2.3 Estimation Models 22

i. Putnam's Model 22

ii. SEER-SEM 22

iii. COCOMO 23

iv. Regression Models 23

2.3 Non-Algorithmic Estimation Models 24

2.3.1 Classification And Regression Tree (CART) 25

2.3.2 Expert Judgment 25

2.3.3 Models based on Soft Computing 27

2.3.3.1 Artificial Neural Network (ANN) 27

2.3.3.2 Fuzzy Logic 27

2.3.3.3 Neuro-fuzzy 28

2.3.4 Analogy-Based Estimation (ABE) 29

2.3.4.1 Similarity Function 29

2.3.4.2 Solution Function 30

2.3.4.3 K Nearest Neighborhood (KNN) 31

2.3.4.4 Previous Studies on ABE 32

2.3.4.5 Attribute Weighting and Attribute Selection 33

2.3.4.6 Grey Scale Analysis 35

2.3.4.7 Adjustment 37

2.3.5 Soft Computing and ABE 37

2.4 Evaluation of Soft Computing based ABE 39

2.5 Datasets Employed for Soft Computing based ABE studies 40

2.6 Performance Metric 41

2.7 Evaluation of the Related Studies for ABE 42

2.8 Missing Data (MD) 42

2.8.1 Missing Data Mechanisms 43

2.8.2 Treatment (Missing Data Techniques) 44

2.8.2.1 MD Ignoring (Deletion) 44

2.8.2.2 MD Toleration 45

2.8.2.3 MD Imputation 45

2.8.2.4 Euclidean Distance 45

2.8.2.5 Manhattan Distance 46

2.9 Segmentation and Localization for Software Development Effort Estimation 49

2.10 Discussion 51

2.11 Summary 53

CHAPTER 3 RESEARCH METHODOLOGY 55

3.1 Introduction 55

3.2 Research Phases 55

3.3 Research Framework 57

3.4 Research Process 58

3.5 Phase A: Primary Studies and Initial Planning 61

3.5.1 Phase A1: Analysis of Existing Literature 61

3.5.2 Phase A2: Problem Discovery 62

3.5.3 Phase A3: Description of Datasets 64

3.5.3.1 Desharnais Dataset 64

3.5.3.2 Maxwell Dataset 65

3.5.3.3 ISBSG Dataset 65

3.5.4 Phase A4: Determining Performance Metrics 67

3.6 Phase ‘B’: Design and Implementation of Missing Data Imputation Techniques for ABE (MINN and IPI) 70

3.6.1 B1: Missing Data Imputation 70

3.6.2 B2: Existing MDI Techniques used with ABE 71

3.6.3 B3: Select the Imputation Technique 72

3.6.4 B4: Evaluation and Comparison of MDI Techniques. 73

3.7 Phase ‘C’: Design and implementation of BABE and ImoEE 74

3.7.1 C1: Determining the Attribute Weight Optimization Method 74

3.7.2 C2: Ensemble ABE and the Optimization Method (BABE) 75

3.7.3 C3: Combining the Imputation Technique with the outcome model of C2 (ImOEE) 77

3.7.4 C3: Evaluation and Comparison 78

3.8 Phase ‘D’: Investigate the Effects of Localization of Projects on ABE (LBABE) 78

3.8.1 D1: Data Filtering and Validation 79

3.8.2 D2: Select the Key Attributes 80

3.8.3 D3: Classify the Projects based on the Key Attributes 81

3.8.4 D4: Evaluate the Impact of Project Classification 81

3.9 Phase ‘E’: Design and implementation of LBABE 81

3.9.1 Phase E1: Localize the Attribute Weight Optimization (LBABE Model) 81

3.9.2 Phase E2: Estimate the Effort 82

3.9.3 Phase E3: Evaluating and Comparison 82

3.10 Phase ‘F’: Design and Implement LIT and LImOEE 82

3.10.1 Phase F1: Localize the Imputation Process. 82

3.10.2 Phase F2: Amalgamate Localized Optimization and the Localized Imputation 83

3.10.3 Phase F3: Evaluation and Comparison 83

3.11 Phase ‘G’: Evaluation and comparison 84

3.11.1 Phase G1: Compare the proposed Models and Techniques 84

3.11.2 Phase G2: Determining the Comparison Scope 84

3.12 Model and Technique Selection 84

3.12.1 Implementation 85

3.12.2 G3: Comparative Analysis 86

3.13 Summary 87

CHAPTER 4 MISSING DATA IMPUTATION FOR ANALOGY BASED ESTIMATION 89

4.1 Introduction 89

4.2 Motivation 89

4.3 Experimental Setup 90

4.4 The proposed Techniques 90

4.4.1 Median Imputation of the Nearest Neighbor 91

4.4.2 Identical Project-based Imputation (IPI) 92

4.5 Experimental Results 95

4.5.1 MINN 95

4.5.2 IPI 95

4.5.3 Effects of IPI, MINN, Numeric Cleansing and KNNI on ABE for Desharnais Dataset 96

4.5.4 Effects of IPI, MINN, Numeric Cleansing and KNNI on ABE for ISBSG Dataset 96

4.6 Comparison between the Proposed and Existing Techniques 97

4.6.1 Percentage Improvement 98

4.7 Technical Analysis 99

4.7.1 Difference between the existing and proposed techniques 99

4.7.2 The connection between Missing Data Imputation and ABE 100

4.7.3 Connection between the Proposed Techniques and other Objectives 101

4.8 Summary 101

CHAPTER 5 ATTRIBUTE WEIGHT OPTIMIZATION FOR ANALOGY-BASED EFFORT ESTIMATION 103

5.1 Introduction 103

5.2 Motivation 103

5.3 Artificial Bee colony guided Analogy-Based Estimation (BABE) 104

5.3.1 Training Stage of BABE 104

5.3.2 Testing Stage 105

5.4 Experimental Results of BABE 107

5.5 Imputation and Optimization-based Effort Estimation (ImOEE) 111

5.5.1 Training Stage of ImOEE 111

5.5.2 Testing Stage of ImOEE 114

5.5.3 Experimental Results 114

5.5.4 SA Results 115

5.6 Technical Analysis 116

5.6.1 Statistical Analysis 116

5.6.2 Statistical Evaluation for BABE 117

5.6.3 Difference between the Existing and Proposed Models 119

5.6.4 Connection between the Proposed Models and other Objectives 119

5.7 Summary 119

CHAPTER 6 LOCALIZED SOFTWARE DEVELOPMENT EFFORT ESTIMATION 121

6.1 Introduction 121

6.2 Motivation 122

6.2.1 Problem 1 (P1): Unreliable Comparison of Target Project Against the Historical Projects for Effort Estimation 122

6.2.2 Problem 2 (P2): Global Attribute Weighting and Unreliable Comparisons 122

6.2.3 Problem 3 (P3): Looking for the Donor Projects in the Global Space for MDI in Analog based Estimation 123

6.3 Solutions to the Problems 124

6.3.1 SOLUTION to P1: Key Attribute based Classification or Localization of the project comparison for Analogy-Based Software Development Effort Estimation 125

6.3.1.1 Experimental design of LABE 125

6.3.1.2 Experimental Results of LABE 127

6.3.1.3 Language Type based Classification 127

6.3.1.4 Development Type based Classification 127

6.3.1.5 Organization Type based Classification 128

6.3.1.6 Development Platform based Classification 128

6.3.1.7 Percentage Improvement 130

6.3.2 SOLUTION to P2: Localized Comparison and Attribute Weight Optimization for ABE 132

6.3.2.1 Experimental Design of Localized-BABE 133

6.3.2.2 Results of LBABE 137

6.3.2.3 Percentage Improvement 138

6.3.3 SOLUTION to P3: Localized Imputation Technique 139

6.3.4 Localized Imputation, and Optimization based Effort Estimation (LImOEE) 141

6.3.4.1 Experimental Results of LImOEE 143

6.4 Technical Analysis 143

6.4.1 Difference between Existing and the proposed models 143

6.4.1.1 LBABE 144

6.4.1.2 LIT 144

6.4.1.3 LImOEE 144

6.4.2 Connection Between Localization and other Objectives 145

6.5 Summary 145

CHAPTER 7 EVALUATION AND COMPARATIVE ANALYSIS OF THE PROPOSED MODELS **147**

7.1 Introduction 147

7.2 Comparison of the Proposed Software Development Effort Estimation Models 147

7.2.1 Results After the Cross-Validation of The Proposed Models 148

7.2.2 MMRE based Comparison 149

7.2.3 PRED (0.25) based Comparison 150

7.2.4 Standardized Accuracy (SA) and Δ based Comparison 151

7.2.5 Percentage Improvement of LImOEE Against BABE, LBABE, and ImOEE 152

7.3 Strengths and Weaknesses of the Proposed Models 152

7.4 Summary 154

CHAPTER 8 CONCLUSION AND FUTURE WORK **155**

8.1 Conclusion 155

8.2 Threat to Validity 157

8.3 Research Contributions 157

8.4 Future Works 159

REFERENCES **163**

Appendix A **177**

Appendix B **179**

LIST OF RESEARCH ARTICLES PRODUCED **181**