



## **University of Bradford eThesis**

This thesis is hosted in [Bradford Scholars](#) – The University of Bradford Open Access repository. Visit the repository for full metadata or to contact the repository team



© University of Bradford. This work is licenced for reuse under a [Creative Commons Licence](#).

# Analogy-Based Software Project Effort Estimation

Mohammad Y. A. Azzeh

PhD

2010

# Analogy-Based Software Project Effort Estimation

Contributions to Projects Similarity Measurement, Attribute Selection & Attribute Weighting algorithms for Analogy-Based Effort Estimation

Mohammad Y. A. Azzeh

A thesis submitted for the degree of  
Doctor of Philosophy

Department of Computing  
School of Computing, Informatics & Media  
University of Bradford

2010

# DECLARATION

I hereby declare that this thesis has not been submitted, either in the same or different form, to this or any other university for a degree.

Signature\_\_\_\_\_

## Abstract

Software effort estimation by analogy is a viable alternative method to other estimation techniques, and in many cases, researchers found it outperformed other estimation methods in terms of accuracy and practitioners' acceptance. However, the overall performance of analogy based estimation depends on two major factors: similarity measure and attribute selection & weighting. Current similarity measures such as nearest neighborhood techniques have been criticized that have some inadequacies related to attributes relevancy, noise and uncertainty in addition to the problem of using categorical attributes.

This research focuses on improving the efficiency and flexibility of analogy-based estimation to overcome the abovementioned inadequacies. Particularly, this thesis proposes two new approaches to model and handle uncertainty in similarity measurement method and most importantly to reflect the structure of dataset on similarity measurement using Fuzzy modeling based Fuzzy C-means algorithm. The first proposed approach called Fuzzy Grey Relational Analysis method employs combined techniques of Fuzzy set theory and Grey Relational Analysis to improve local and global similarity measure and tolerate imprecision associated with using different data types (Continuous and Categorical). The second proposed approach presents the use of Fuzzy numbers and its concepts to develop a practical yet efficient approach to support analogy-based systems especially at early phase of software development. Specifically, we propose a new similarity measure and adaptation technique based on Fuzzy numbers.

We also propose a new attribute subset selection algorithm and attribute weighting technique based on the hypothesis of analogy-based estimation that assumes projects that are similar in terms of attribute value are also similar in terms of effort values, using row-wise Kendall rank correlation between similarity matrix based project effort values and similarity matrix based project attribute values. A literature review of related software engineering studies revealed that the existing attribute selection techniques (such as brute-force, heuristic algorithms) are restricted to the choice of performance indicators such as (*Mean of Magnitude Relative Error* and *Prediction Performance Indicator*) and computationally far more intensive. The proposed algorithms provide sound statistical basis and justification for their procedures.

The performance figures of the proposed approaches have been evaluated using real industrial datasets. Results and conclusions from a series of comparative studies with conventional estimation by analogy approach using the available datasets are presented. The studies were also carried out to statistically investigate the significant differences between predictions generated by our approaches and those generated by the most popular techniques such as: conventional analogy estimation, neural network and stepwise regression. The results and conclusions indicate that the two proposed approaches have potential to deliver comparable, if not better, accuracy than the compared techniques. The results also found that Grey Relational Analysis tolerates the uncertainty associated with using different data types. As well as the original contributions within the thesis, a number of directions for further research are presented.

Most chapters in this thesis have been disseminated in international journals and highly refereed conference proceedings.

# Acknowledgments

I would first and foremost like to thank Dr. Daniel Neagu and Prof. Peter Cowling for their dedicated supervision and unwavering support, without which this thesis would never have got off the ground.

I would especially like to express my gratitude to my lovely wife Tamara for her love, patience, and support. I also dedicate this work to my parents who have been supportive and understanding throughout the years of my PhD study. This work is also dedicated to my lovely daughter Toleen, who provided an additional and joyful dimension to my life mission.

I would also like to record my thanks to my friends at University of Bradford who have made contributions to this thesis. I would also like to acknowledge the support and funding of Applied Science University, Jordan.

# Publications

## *Published papers:*

- [1] Azzeh, M., Cowling, P. & Neagu, D. (2010), Software Stage-Effort Estimation Based on Association Rule Mining and Fuzzy Set Theory, in proceedings of the 10<sup>th</sup> IEEE international conference on computer and information technology (Accepted).
- [2] Azzeh, M., Neagu, D. & Cowling, P. (2010), *Fuzzy grey relational analysis for software effort estimation*, Empirical Software Engineering, Springer, Vol. 15, Issue 1, 60-90.
- [3] Azzeh, M., Neagu, D. & Cowling, P. (2009), Software Effort Estimation Based on Weighted Fuzzy Grey Relational Analysis, in proceedings of 5<sup>th</sup> international conference on software predictor models, Co-located with ICSE'31, Article number 8.
- [4] Azzeh, M., Neagu, D. & Cowling, P. (2008c), Adjusting Analogy Software Effort Estimation Based on Fuzzy Logic, in Proceedings of 3<sup>rd</sup> International conference on software and data technology ICSoft (SE/MUSE/GSDCA), 127-132.
- [5] Azzeh, M., Neagu, D. & Cowling, P. (2008b), Software Project Similarity Measurement Based On Fuzzy C-Means, in proceedings of 2<sup>nd</sup> International Conference on software process, Leipzig, Germany, 123-134.
- [6] Azzeh, M., Neagu, D. & Cowling, P. (2008a), Fuzzy Feature Subset Selection for Software Effort Estimation, in proceedings of 4<sup>th</sup> international workshop on software predictors PROMISE'08 (part of ICSE'08), Leipzig, Germany, 71-78.
- [7] Azzeh, M., Neagu, D. & Cowling, P. (2007), An Overview of Web Cost Estimation, in proceedings of the 8<sup>th</sup> Informatics Workshop, University of Bradford, 96-99.

## *Papers under review:*

- [8] Azzeh, M., Neagu, D. & Cowling, P., (2010), *FGRA+: An Improved Method For Software Effort Estimation*, *Journal of Information & Software Technology*, Elsevier, Accepted subject to corrections.
- [9] Azzeh, M., Neagu, D. & Cowling, P., *Analogy-Based Software Effort Estimation based on Fuzzy Numbers*, *Journal of System & Software*, Elsevier, Submitted (under review)
- [10] Azzeh, M., Neagu, D. & Cowling, P. (2010), Model Tree-Based Adaptation Strategy for Analogy Based Effort Estimation, 6<sup>th</sup> international conference on software predictor models (under review).

# Table of Contents

<b>1</b>	<b>Introduction</b>	
1.1	Background	2
1.2	Research motivations and aims	3
1.3	Scope of work	5
1.4	Thesis contributions	6
1.4.1	Issue 1: Attribute subset selection using Kendall's row-wise correlation	6
1.4.2	Issue 2: Attribute weighting using Kendall coefficient of concordance	7
1.4.3	Issue 3: Software project similarity measurement based on Fuzzy Set Theory using Fuzzy C-means	7
1.4.4	Issue 4: Grey Relational Analysis to increase the flexibility of similarity measure	8
1.4.5	Issue 5: Software project similarity measurement using fuzzy numbers	9
1.4.6	Issue 6: Adaptation technique based on mathematical operations of Fuzzy numbers	9
1.5	Thesis structure	9
<b>2</b>	<b>Background and Related Work</b>	
2.1	Terminology	12
2.1.1	Effort	12
2.1.2	Effort estimate	12
2.1.3	Estimation by analogy	13
2.2	Software effort estimation	14
2.2.1	Expert judgment technique	15
2.2.2	Model based techniques	16
2.2.3	Learning oriented techniques	18
2.2.4	The choice of estimation method	18
2.2.5	Software size estimation	19
2.3	Software effort estimation by analogy	21
2.4	Advantages of estimation by analogy	25
2.5	Analogy based systems	27
2.5.1	ANGEL	27
2.5.2	ESTOR	29
2.5.3	ACE	30
2.6	Issues in estimation by analogy	31
2.6.1	Similarity measurement	31
2.6.2	Attribute selection and weighting	36
2.7	Uncertainty in software effort estimation	38
2.8	Early stage software effort estimation	40
2.9	Evaluation criteria	40
2.10	Summary	43
<b>3</b>	<b>An Overview of Fuzzy Set Theory, Fuzzy Numbers and Grey Relational Analysis</b>	
3.1	Fuzzy set theory	46



3.1.1	Fuzzy membership functions	47
3.1.2	Fuzzy model construction	49
3.2	Generalized Fuzzy numbers	53
3.3	Grey Relational Analysis	55
3.3.1	Grey relational coefficient	57
3.3.2	Grey relational grade	57
3.3.3	Grey relational ranking	58
3.4	Chapter summary	58
<b>4</b>	<b>Attribute Selection &amp; Weighting Algorithms Based on Kendall Row-Wise Correlation For Analogy Based Estimation</b>	
4.1	Introduction	60
4.2	Estimation by analogy: Formal problem description	61
4.3	Assumption of analogy based estimation	62
4.4	Similarity matrix	64
4.5	Kendall rank correlation	65
4.5.1	Kendall's row-wise correlation between similarity matrix based nominal attribute and similarity matrix based effort	71
4.5.2	Significance test for Kendall row-wise rank correlation	73
4.6	Similarity degree ranking	73
4.7	Attribute selection	74
4.8	Attribute weighting	77
4.9	Experimental results	80
4.9.1	Albrecht data set	81
4.9.2	Kemerer data set	84
4.9.3	Desharnais data set	87
4.9.4	COCOMO data set	88
4.9.5	ISBSG data set	91
4.9.6	Empirical evaluation of attribute weighting	94
4.10	Chapter summary	96
<b>5</b>	<b>Analogy-Based Software Effort Estimation Using Fuzzy Set Theory and Grey Relational Analysis</b>	
5.1	Introduction	99
5.2	The proposed similarity measure	100
5.2.1	local similarity measures	100
5.2.1.1	Numerical Scale	101
5.2.1.2	Nominal Scale	106
5.2.1.3	Ordinal Scale	107
5.2.1.4	Set scale	107
5.2.2	Global similarity measure based on GRA	108
5.3	FGRA-The basic approach	110
5.3.1	Data preparation phase	111
5.3.2	Attribute selection phase	111
5.3.3	Effort prediction phase	112
5.4	Empirical evaluation	114
5.4.1	Experimental procedure	115
5.4.2	Measuring prediction accuracy of FGRA	118
5.4.3	Comparison of FGRA to Case-based reasoning, Artificial neural networks and stepwise regression	122

5.4.3.1	Comparison over ISBSG data set	123
5.4.3.2	Comparison over Desharnais data set	125
5.4.3.3	Comparison over COCOMO data set	127
5.4.3.4	Comparison over Kemerer data set	129
5.4.3.5	Comparison over Albrecht data set	131
<b>5.4.4</b>	<b>Discussion</b>	<b>133</b>
<b>5.5</b>	<b>Chapter summary</b>	<b>136</b>
<b>6</b>	<b>Analogy-Based Software Effort Estimation Using Fuzzy Numbers</b>	
<b>6.1</b>	<b>Introduction</b>	<b>139</b>
<b>6.2</b>	<b>The proposed similarity measure</b>	<b>139</b>
<b>6.2.1</b>	<b>Properties of the similarity measure</b>	<b>142</b>
<b>6.2.2</b>	<b>Comparison between the proposed similarity measure and other similarity measures based on Fuzzy numbers</b>	<b>143</b>
<b>6.3</b>	<b>GFNSE software prediction model</b>	<b>147</b>
<b>6.3.1</b>	<b>Fuzzy numbers construction</b>	<b>147</b>
<b>6.3.2</b>	<b>Finding similarity between target project and each comparative project at each attribute</b>	<b>150</b>
<b>6.3.3</b>	<b>Ranking closest projects</b>	<b>150</b>
<b>6.3.4</b>	<b>Deriving a new estimate for target project</b>	<b>151</b>
<b>6.4</b>	<b>Experimental results</b>	<b>153</b>
<b>6.4.1</b>	<b>Design of experiments</b>	<b>153</b>
<b>6.4.2</b>	<b>Results for ISBSG data set</b>	<b>154</b>
<b>6.4.3</b>	<b>Results for COCOMO data set</b>	<b>158</b>
<b>6.4.4</b>	<b>Results for Desharnais data set</b>	<b>161</b>
<b>6.4.5</b>	<b>Results for Albrecht data set</b>	<b>163</b>
<b>6.4.6</b>	<b>Results for Kemerer data set</b>	<b>166</b>
<b>6.5</b>	<b>Discussion</b>	<b>168</b>
<b>7</b>	<b>Conclusions and Further Work</b>	
<b>7.1</b>	<b>Research summary and conclusions</b>	<b>171</b>
<b>7.2</b>	<b>Synopsis of research findings</b>	<b>175</b>
<b>7.3</b>	<b>Limitation of Work</b>	<b>177</b>
<b>7.3.1</b>	<b>Analysis limitations</b>	<b>177</b>
<b>7.3.2</b>	<b>Approach limitations</b>	<b>178</b>
<b>7.4</b>	<b>Future works</b>	<b>178</b>
<b>7.4.1</b>	<b>Stage software effort estimation</b>	<b>178</b>
<b>7.4.2</b>	<b>Sensitivity analysis of the proposed estimation techniques</b>	<b>180</b>
<b>7.4.3</b>	<b>Determining optimal number of analogies</b>	<b>180</b>
<b>7.4.4</b>	<b>Improving construction of Fuzzy numbers</b>	<b>181</b>
	<b>Bibliography</b>	<b>182</b>
	Appendix 1: Fuzzy C-means algorithm	193
	Appendix 2: Albrecht data set	195
	Appendix 3: Kemerer data set	196
	Appendix 4: COCOMO data set	197
	Appendix 5: Desharnais data set	199
	Appendix 6 : ISBSG data set	201

# List of Figures

Figure 2.1: CBR process lifecycle	22
Figure 2.2: The generic framework for case-based prediction	23
Figure 2.3: Process of estimation by analogy	25
Figure 2.4: Process of ANGEL	28
Figure 2.5: Logical Framework of ESTOR	30
Figure 2.6: Process of ACE tool	31
Figure 2.7: Representation of Euclidean distance	32
Figure 3.1: Fuzzy sets for variable Team Experience	47
Figure 3.2: Crisp set for variable Team Experience	47
Figure 3.3: Different Types of membership functions	48
Figure 3.4: Algorithm of finding appropriate number of clusters	51
Figure 3.5: Algorithm of Constructing Fuzzy membership functions	52
Figure 3.6: Raw-data	52
Figure 3.7: Clustered data	52
Figure 3.8: Membership functions for attribute FA.	52
Figure 3.9: Membership functions for attribute FB	52
Figure 3.10: Membership functions for attribute FC.	52
Figure 3.11: Two Generalized trapezoidal Fuzzy numbers	53
Figure 3.12: Two Generalized triangular Fuzzy numbers	53
Figure 4.1: Scatter plot for similarity rankings in terms of attribute $X$ and $E$	64
Figure 4.2: Similarity matrix based project attribute(s) and effort respectively	65
Figure 4.3: General form of two symmetric square matrices	69
Figure 4.4: $SM(X)$ vs. $SM(E)$	70
Figure 4.5: $SM(Y)$ vs. $SM(E)$	70
Figure 4.6: Kendall's row-wise correlation between $SM(Z)$ vs. $SM(E)$	72
Figure 4.7: $SM(X)$ and its rankings on the right side	74
Figure 4.8: Attribute subset selection algorithm	76
Figure 4.9: Kendall's $W$ between $SM(X)$ vs. $SM(E)$	80
Figure 4.10: Box-plot of absolute residuals of prediction using $KFS$ , $Analogy-X$ , and $ANGEL$ for Albrecht data set	84
Figure 4.11: Box-plot of absolute residuals of prediction using $KFS$ , $Analogy-X$ , and $ANGEL$ for Kemerer data set	86
Figure 4.12: Box-plot of absolute residuals of prediction using $KFS$ , $Analogy-X$ , and $ANGEL$ for Desharnais	89
Figure 4.13: Box-plot of absolute residuals of prediction using $KFS$ , $Analogy-X$ , and $ANGEL$ for COCOMO	91
Figure 4.14: Box-plot of absolute residuals of prediction using $KFS$ , $Analogy-X$ , and $ANGEL$ for ISBSG	94
Figure 5.1: Fuzzy sets for $a_j$	102
Figure 5.2: Special case of similarity measure	103
Figure 5.3: generic framework of FGRA software effort estimation model	110
Figure 5.4: Artificial Neural Network	118
Figure 5.5: Box-plots of absolute residuals for each data set	121
Figure 5.6: Boxplot of Absolute residuals for ISBSG	125
Figure 5.7: Boxplot of Absolute residuals for Desharnais	127

Figure 5.8: Boxplot of Absolute residuals for COCOMO	129
Figure 5.9: Boxplot of Absolute residuals for Kemerer	131
Figure 5.10: Boxplot of Absolute residuals for Albrecht	133
Figure 5.11: The investigation into the sensitivity of FGRA to number of Clusters	135
Figure 6.1: Three generalized triangular Fuzzy numbers	142
Figure 6.2: Fifteen sets of generalized Fuzzy numbers	144
Figure 6.3: Fuzzy number coefficient determination	149
Figure 6.4: Boxplots of absolute residuals of GFNSE, CBR and SR over ISBSG	157
Figure 6.5: Boxplots of absolute residuals of GFNSE , CBR and SR over COCOMO	160
Figure 6.6: Boxplots of absolute residuals of GFNSE, CBR and SR over Desharnais	163
Figure 6.7: Boxplots of absolute residuals of GFNSE, CBR and SR over Albrecht	166
Figure 6.8: Boxplots of absolute residuals of GFNSE, CBR and SR over Kemerer	168

## List of Tables

Table 2.1: Illustration of <i>MMRE</i> problem	41
Table 3.1: Fuzzy partition matrix for software projects	50
Table 4.1: Similarity degrees between target project and other source projects	63
Table 4.2: Ranking of similarity degrees	63
Table 4.3: Rankings given by two experts <i>X</i> and <i>Y</i>	68
Table 4.4: Simple hypothetical data set	69
Table 4.5: Similarity degrees between target project $p_i$ and other source projects	71
Table 4.6: Ranks of similarity degrees in Table 4.5	71
Table 4.7: Adding Nominal Attribute <i>Z</i> to Simple data set	72
Table 4.8: Descriptive statistics of the data sets	81
Table 4.9: Albrecht data set description	82
Table 4.10: Prediction accuracy comparison for Albrecht data set	83
Table 4.11: Wilcoxon sum rank test for Albrecht data set	84
Table 4.12: Kemerer data set description	85
Table 4.13: Prediction accuracy comparison for Kemerer data set	85
Table 4.14: Wilcoxon sum rank test for Kemerer data set	86
Table 4.15: Desharnais data set description	87
Table 4.16: Prediction accuracy comparison for Kemerer data set	88
Table 4.17: Wilcoxon sum rank test for Desharnais	88
Table 4.18: COCOMO data set description	90
Table 4.19: Prediction accuracy comparison for COCOMO data set	90
Table 4.20: Wilcoxon sum rank test for Desharnais for COCOMO	91
Table 4.21: ISBSG data set description	92
Table 4.22: Prediction accuracy comparison for ISBSG data set	93
Table 4.23: Wilcoxon sum rank test for ISBSG	93
Table 4.24: <i>MMRE</i> and <i>PRED(25)</i> improvement when using attribute weighting	95
Table 5.1: FGRA-1 prediction accuracy results	120
Table 5.2: FGRA-2 prediction accuracy results	120
Table 5.3: FGRA-3 prediction accuracy results	120
Table 5.4: FGRA Statistical results based on residuals	122
Table 5.5: Formula for stepwise regression models	123
Table 5.6: Comparison on ISBSG Data set	124
Table 5.7: Mann-Whitney test of absolute residuals for ISBSG data set	125
Table 5.8: Comparison on Desharnais Data set	126
Table 5.9: Mann-Whitney test of absolute residuals for Desharnais data set	127
Table 5.10: Comparison on COCOMO Data set	128
Table 5.11: Mann-Whitney test of absolute residuals for COCOMO data set:	128
Table 5.12: comparison on Kemerer Data set	130
Table 5.13: Mann-Whitney test of absolute residuals for Kemerer data set	131
Table 5.14: Comparison on Albrecht Data set	132
Table 5.15: Mann-Whitney test of absolute residuals for Albrecht data set	133
Table 6.1: Comparison between the proposed similarity method and existing similarity methods	145
Table 6.2: Example of adaptation technique	152
Table 6.3: Performance figures for ISBSG with different number of analogies	155

Table 6.4: Wilcoxon signed rank test for paired absolute residuals over ISBSG	155
Table 6.5: Performance figures of comparing GFNSE to CBR and SR over ISBSG	157
Table 6.6: Comparison of techniques over ISBSG, using Mann Whitney U test	157
Table 6.7: Performance figures for COCOMO with different analogy numbers	158
Table 6.8: Wilcoxon signed rank test for paired absolute residuals over COCOMO	159
Table 6.9: Performance figures of comparing GFNSE to CBR and SR over COCOMO	160
Table 6.10: Comparison of techniques over COCOMO, using Mann Whitney U test	160
Table 6.11: Performance figures for Desharnais with different analogy numbers	161
Table 6.12: Wilcoxon signed rank test for paired absolute residuals over Desharnais	161
Table 6.13: Performance figures of comparing GFNSE to CBR and SR over Desharnais	162
Table 6.14: Comparison of techniques over Desharnais, using Mann Whitney U test	162
Table 6.15: Performance figures for Albrecht with different analogy numbers	164
Table 6.16: Wilcoxon signed rank test for paired absolute residuals over Albrecht	164
Table 6.17: Performance figures of comparing GFNSE to CBR and SR over Albrecht	165
Table 6.18: Comparison of techniques over Albrecht, using Mann Whitney U test	165
Table 6.19: Performance figures for Kemerer with different analogy numbers	166
Table 6.20: Wilcoxon signed rank test for paired absolute residuals over Kemerer	167
Table 6.21: Comparison of GFNSE to CBR and SR over Kemerer	167
Table 6.22: Comparison of techniques over Kemerer, using Mann Whitney U test	167

# List of Abbreviations

ANGEL	Analogy based effort estimation tool
ANN	Artificial Neural Network
CBR	Case Based Reasoning
EBA	Estimation by Analogy
FCM	Fuzzy C- Mean
FN	Fuzzy Number
FGRA	Fuzzy Grey Relational Analysis
GFNSE	Generalized Fuzzy Number Software Estimation
FST	Fuzzy Set Theory
GRA	Grey Relational Analysis
ISBSG	International Software Benchmarking Standard Group
KFS	Kendall based Feature Selection
MMRE	Mean of Magnitude Relative Error
MdMRE	Median of Magnitude Relative Error
PRED	Prediction Performance indicator
SR	Stepwise Regression