Fuzzy Multiple Regression Model for Estimating Software Development Time

Venus Marza and Mir Ali Seyyedi
MSc student of Computer Engineering at Islamic Azad University of South Tehran Branch
Assistant Professor at the department of Computer Engineering, Islamic Azad University of South Tehran Branch
Corresponding author E-mail: Venus.Marza@gmail.com

Abstract: *As software becomes more complex and its scope dramatically increase, the importance of research on developing methods for estimating software development time has perpetually increased, so accurate estimation is the main goal of software managers for reducing risks of projects. The purpose of this article is to introduce a new Fuzzy Multiple Regression approach, which has the higher accurate than other methods for estimating. Furthermore, we compare Fuzzy Multiple Regression model with Fuzzy Logic model & Multiple Regression model based on their accuracy.*
Keywords: *Fuzzy Logic (FL), Multiple Regression Model, McCabe Complexity, Dhama Coupling, Development Time*

1. Introduction

Many study have already proposed models for size, effort, time and cost estimation. We just consider some of these studies: Regression analysis is a classical statistical technique for building estimation models. It is one of the most commonly used methods in econometric work. It is concerned with describing and evaluating the relationship between a dependent variable and one or more independent variables. The relationship is described as a model for estimating the dependent variable from independent variables. The model is built and evaluated through collecting sample data for these variables. This model was used first for estimating LOC of an information system (Kuan Tan et al. 2006).
Boehm was the first researcher to look at software from an economic point of view. Putnam also developed model known as SLIM, but both of COCOMO and SLIM are based on linear regression techniques (Moataz et al. 2005). Algorithmic models such as COCOMO, have failed to present suitable solutions that take into consideration technological advancements, because they are often unable to capture the complex set of relationships (e.g. the effect of each variable in a model to the overall prediction made using the model), they are not flexible enough to adapt to a new environment, and they can’t learn from their previous knowledge, also parametric models use a static predictive function for estimating (e.g. COCOMO use $Effort = A \cdot Size^{B}$ for estimating Effort (Xia et al. 2005)).
Their inability contributed to exploring non parametric methods Such as Fuzzy Logic, Soft computing which is consortium of methodologies centering in Fuzzy Logic, Artificial Neural Networks and Evolutionary Computation. Research of MacDonell has evolved into development of a FULSOME (FUzzy Logic for SOftware MEtrics), to assist project managers in making predictions (Moataz et al. 2005, MacDonell 2003]. Orginally, estimation was performed using only human expertise, but more recently, attention has turned to a variety of combining methods. Here we apply fuzzy concepts to regression model and compare their results with each other. The primary motivation of fuzzy set theory is the desire to build a formal quantitative structure capable of capturing the imprecision of human knowledge, that is, the manner in which knowledge is expressed in natural language. This theory seeks to bridge the gap that separates traditional mathematical models needed for physical systems, and the mental representation, generally imprecise, of such systems (Lima et al. 1999).
This paper is structured as follows. In section 2, multiple regression equation is considered and its result is shown in data subset, then in next section we apply fuzzy logic to the same data subset. In section 4 we introduce specific regression model with fuzzy concepts. In section 5, evaluation criteria are introduced for evaluating models, and in section 6, we apply these mentioned models to the same data subset for comparing them. Finally, conclusions are drawn in section 7.

2. Multiple Regression Equation

This model is the most common statistical technique for estimating. A linear equation with three independent variables (McCabe Complexity (MC), Dhama Coupling (DC), and physical Lines Of Code (LOC)) and a dependent one (Development Time (DT)) may be expressed as (Cuauhtemoc et al. 2005):

$$DT = b_0 + b_1 MC + b_2 DC + b_3 LOC \qquad (1)$$

Where b_0, b_1, b_2 and b_3 is obtained by solving follow equations:

Open Access Database www.intechweb.org

$$\begin{aligned}\sum y &= nb_0 + b_1(\sum x_1) + b_2(\sum x_2) + b_3(\sum x_3)\\ \sum x_1 y &= b_0(\sum x_1) + b_1(\sum x_1^2) + b_2(\sum x_1x_2) + b_3(\sum x_1x_3)\\ \sum x_2 y &= b_0(\sum x_2) + b_1(\sum x_1x_2) + b_2(\sum x_2^2) + b_3(\sum x_2x_3)\\ \sum x_3 y &= b_0(\sum x_3) + b_1(\sum x_1x_3) + b_2(\sum x_2x_3) + b_3(\sum x_3^2)\end{aligned} \quad (2)$$

For simplify we used x_1 as MC, x_2 as DC, x_3 as LOC and y as DT. By using data from Table 3 (Cuauhtemoc et al. 2005) and finding parameter b_0, b_1, b_2, b_3 we can give following linear equation (Cuauhtemoc et al. 2005):

$$DT' = 17.3097 + 2.06268 * MC - 32.9405 * DC - 0.0499692 * LOC \quad (3)$$

The result of each module is organized in Table 4.

3. Estimating by Fuzzy Logic Rules

A fuzzy model like any other model provides mapping from input to output. For obtaining a fuzzy model first the verbal expert knowledge, based on the correlation(r) between pairs of their variables, is translated into if-then rules. Parameters of this structure, such as membership functions and weights of rules, can be tuned by using input and output data.

Correlation is the degree that indicates two sets how much are related to each other, and is defined as follows (Cuauhtemoc et al. 2005):

$$r = \frac{n[\sum(X_i.Y_i)] - (\sum X_i)(\sum Y_i)}{\sqrt{[n(\sum X_i^2) - (\sum X_i)^2][n(\sum Y_i^2) - (\sum Y_i)^2]}} \quad (4)$$

Correlation between development time as dependent variable and McCabe complexity, Dhama Coupling and lines of code as independent variables are organized in Table 1 (Cuauhtemoc et al. 2005):

Pair	r	Pair	r
MC_DC	-0.3860	DT_MC	0.7078
MC_LOC	0.7653	DT_DC	-0.7051
DC_LOC	-0.4346	DT_LOC	0.5827

Table 1. Correlation between variables

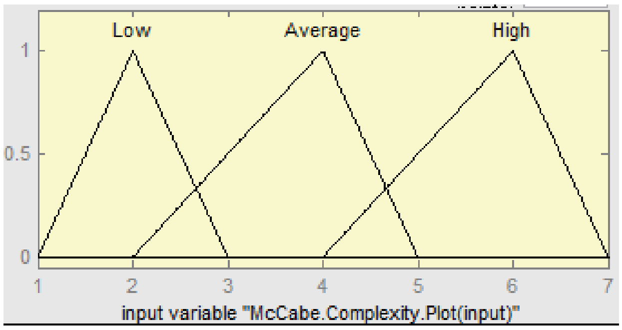
Variable Name	Range	MF	Parameters		
			a	b	c
McCabe	1-7	Low	1	2	3
		Average	2	4	5
		High	4	6	7
Dhama	0-0.4	Low	0.24	0.31	0.40
		Average	0.05	0.17	0.33
		High	0.00	0.12	0.21
Lines of Code	1-40	Low	3	8	15
		Average	10	16	25
		High	21	28	40

Variable Name	Range	MF	Parameters		
			a	b	c
Development Time (min)	1-27	Low	6.6	9.0	11.8
		Average	8.1	12.8	18.6
		High	14.0	20.0	27.0

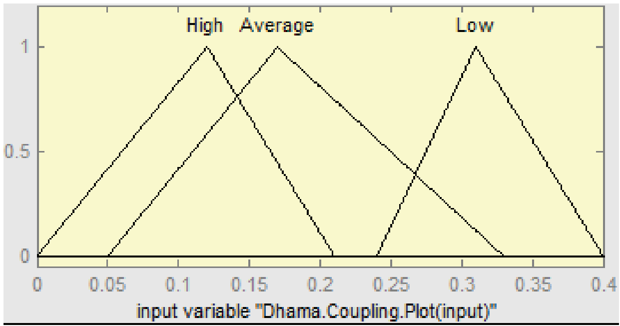
Table 2. Membership Function Parameters

So fuzzy rules were formulated as bellow:

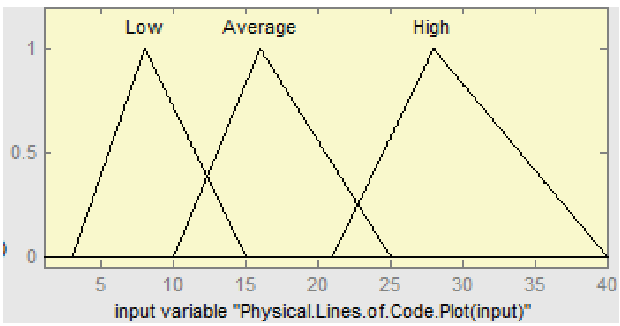
1. If *Complexity* is low and *Size(LOC)* is small then *DT* is low
2. If *Complexity* is average and *Size(LOC)* is medium then *DT* is average
3. If *Complexity* is high and *Size(LOC)* is big then *DT* is high
4. If *Coupling* is low then *DT* is low
5. If *Coupling* is average then *DT* is average
6. If *Coupling* is high then *DT* is high



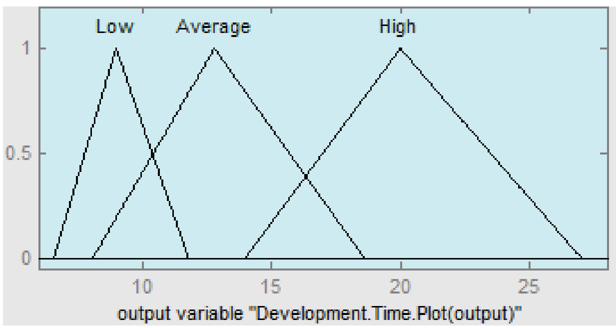
(a). McCabe Complexity Plot (input)



(b).Dhama Coupling Plot (input)



(c). Physical Lines of Code Plot (input)



(d). Development Time Plot (output)

Fig. 1. Membership functions for input & output

For triangular membership function, three parameters (a, b, c) are defined. In Table 2 (Cuauhtemoc et al. 2005), input and output membership function is shown for dependent and independent variables. Their scalar parameters (a, b, c) are defined as follows:
MF(x)=0 if x < a
MF(x)=1 if x = b
MF(x)=0 if x > c
The membership functions corresponding to Table 2 are shown in Fig.1(a), 1(b), 1(c), and 1(d).
Consequently, by using fuzzy rules and their memberships, DT is depicted in Table 4.

4. Multiple Regression Model with Fuzzy Concepts

Fuzzy concepts help us to find the deviation of each data from fitness equation, so we define a normal distribution membership function as follow:

$$U_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y_i-\mu}{\sigma})2} \qquad (5)$$

Where μ is average of sample points and σ is square root of variance math.
If we add fuzzy domain to Regression method, the effect of discrete data points on the fitness result will be reduced and the effect of concentrated data points on the fitness result will be enhanced.
For each data in Table 3, we obtain the membership function that is shown in column 7. A group of equations to obtain the fuzzy parameters are given as (Gu et al. 2006):

$$\begin{aligned} s_{11}b_1 + s_{12}b_2 + \ldots + s_{1k}b_k &= s_{1y} \\ s_{21}b_1 + s_{22}b_2 + \ldots + s_{2k}b_k &= s_{2y} \\ &. \\ &. \\ &. \\ s_{k1}b_1 + s_{k2}b_2 + \ldots + s_{kk}b_k &= s_{ky} \end{aligned} \qquad (6)$$

Here, $s_{ij} = \sum u \sum u\, x_i\, x_j - \sum u\, x_i \sum u\, x_j$ and $s_{iy} = \sum u \sum u\, x_i\, y - \sum u\, x_i \sum u\, y$, i,j=1,2,…,k.
Where y is each Development Time (DT) of mentioned projects, and here we have 41 projects for considering.
Then b_0 is calculated by:

$$b_0 = \frac{\sum uy}{\sum u} - b_1\frac{\sum ux_1}{\sum u} - b_2\frac{\sum ux_2}{\sum u} - \ldots - b_k\frac{\sum ux_k}{\sum u} \qquad (7)$$

By solving these equations, final equation is expressed as:

$$\begin{aligned} DT' &= 17.33532512 + 1.709789562 * MC - \\ & 29.55193896 * DC - 0.03700819852 * LOC \end{aligned} \qquad (8)$$

The result of this method is presented in Table 4.

5. Evaluating Techniques

A common criterion, which is calculated for each observation, is MRE and it is defined as follows:

$$MRE_i = \frac{|Actual\ Effort_i - \Pr edicted\ Effort_i|}{Actual\ Effort_i} \qquad (9)$$

With aggregation of MRE on all data, MMRE (Mean Magnitude of Relative Error) is achieved as follows (Burgess & Lefley, 2001):

$$MMRE = \frac{1}{n}\sum_{i=1}^{i=n}(\frac{|E_i - \overline{E}_i|}{E_i}) \qquad (10)$$

A complementary criterion that is used here is Pred(20). In general, Pred(*l*)=k/N where k is the number of observations where MRE is less than or equal to *l* (Cuauhtemoc et al. 2006), So Pred(20) gives the

	Module Description	MC	DC	LOC	DT	Ui
1	Calculates t Value	1	0.25	4	13	0.066575
2	Inserts a new element in a linked list	1	0.25	10	13	0.066575
3	Calculates a value according to normal distribution equation	1	0.333	4	9	0.012521
4	Calculates the variance	2	0.083	10	15	0.098389
5	Generates range square root	2	0.111	23	15	0.098389
6	Determines both minimum and maximum values from a sorted linked list	2	0.125	9	15	0.098389
7	Turns each linked list value into its *z* value	2	0.125	9	16	0.10702
8	Copies a list of values from a file to an array	2	0.125	14	16	0.10702
9	Determines parity of a number	2	0.167	7	16	0.10702
10	Defines segment limits	2	0.167	8	18	0.101369
11	From two lists (X and Y), returns the product of all Xi and Yi values	2	0.167	10	15	0.098389
12	Calculates a sum from a vector and its average	2	0.167	10	15	0.098389
13	Calculates *q* value	2	0.167	10	18	0.101369
14	Generates the sum of vector components	2	0.2	10	13	0.066575
15	Calculates the sum of a vector values square	2	0.2	10	14	0.08399
16	Calculates the average of the linked list values	2	0.2	10	15	0.098389
17	Counts the number of lines of code including blanks and comments	2	0.2	15	13	0.066575
18	Prints value non zero of a linked list	2	0.25	10	12	0.049
19	Stores values into a matrix	2	0.25	10	12	0.049
20	Generates range square root	3	0.083	17	22	0.037359
21	Returns the number of elements in a linked list	3	0.125	11	19	0.088273
22	Calculates the sum of odd segments (Simpson's formula)	3	0.125	15	18	0.101369
23	Calculates the sum of pair segments (Simpson's formula)	3	0.125	15	19	0.088273
24	Generates the standard deviation of the linked list values	3	0.143	13	21	0.053588
25	Returns the sum of square roots of a list values	3	0.143	14	20	0.071375
26	Prints a matrix	3	0.143	14	21	0.053588
27	Calculates the sum of odd segments (Simpson's formula)	3	0.143	15	19	0.088273
28	Calculates the sum of pair segments (Simpson's formula)	3	0.143	15	20	0.071375
29	Calculates the average of linked list values	3	0.167	13	15	0.098389
30	Returns the sum of a list of values	3	0.167	14	13	0.066575
31	Generates the standard deviation of linked list values	3	0.2	18	19	0.088273
32	Prints a linked list	3	0.25	9	13	0.066575
33	Calculates gamma value(G)	3	0.25	12	12	0.049
34	Calculates the average of vector components	3	0.25	17	12	0.049
35	Calculates the ranges standard deviation	4	0.077	16	21	0.053588
36	Calculates beta1 value	4	0.077	31	21	0.053588
37	Returns the product between values of two vectors and the number of these pairs	4	0.111	16	19	0.088273
38	Counts commented lines	4	0.2	24	18	0.101369
39	Reduces final matrix (according to Gauss method)	5	0.143	22	24	0.014536
40	Reduces a matrix (according to Gauss method)	5	0.143	22	25	0.008113
41	Counts blank lines	5	0.2	22	18	0.101369

Table 3. Modules description and metrics, MC (McCabe Complexity), DC (Dhama Coupling), LOC (Lines of Code), DT (Development Time *(minutes)*)

percentage of projects which were predicated with a MRE less or equal than 0.20. In general, the accuracy of an estimation technique is proportional to the Pred(20) and inversely proportional to the MMRE (Xia et al. 2005).

6. Experimental Results

Multiple Regression, fuzzy rules system and fuzzy multiple regression are applied to the same data subset. The MMRE & PRED(20) are shown in Table 4. Results are

Module	Actual DT	Multiple Regression		Fuzzy Logic		Fuzzy Multiple Regression	
		DT '	MRE i	DT '	MRE i	DT '	MRE i
1	13	10.9374	0.1587	13	0.0000	11.5097	0.114642
2	13	10.6376	0.1817	13	0.0000	11.2861	0.131836
3	9	8.2033	0.0885	9.15	0.0167	9.05681	0.006312
4	15	18.2013	0.2134	16.8	0.1200	17.9322	0.195481
5	15	16.6294	0.1086	17.8	0.1867	16.6204	0.108029
6	15	16.8678	0.1245	16.3	0.0867	16.7283	0.115219
7	16	16.8678	0.0542	16.3	0.0188	16.7283	0.045518
8	16	16.6179	0.0386	17.3	0.0813	16.542	0.033875
9	16	15.5842	0.0260	15.6	0.0250	15.5616	0.027401
10	18	15.5342	0.1370	15.5	0.1389	15.5243	0.137537
11	15	15.4343	0.0290	15.6	0.0400	15.4498	0.029988
12	15	15.4343	0.0290	15.6	0.0400	15.4498	0.029988
13	18	15.4343	0.1425	15.6	0.1333	15.4498	0.141677
14	13	14.3473	0.1036	14	0.0769	14.4746	0.11343
15	14	14.3473	0.0248	14	0.0000	14.4746	0.033899
16	15	14.3473	0.0435	14	0.0667	14.4746	0.035027
17	13	14.0974	0.0844	15.1	0.1615	14.2883	0.099101
18	12	12.7002	0.0584	12	0.0000	12.997	0.083081
19	12	12.7002	0.0584	12	0.0000	12.997	0.083081
20	22	19.9142	0.0948	17.6	0.2000	19.3823	0.118988
21	19	18.8305	0.0089	17.6	0.0737	18.3646	0.033442
22	18	18.6306	0.0350	17.6	0.0222	18.2156	0.011977
23	19	18.6306	0.0194	17.6	0.0737	18.2156	0.041285
24	21	18.1376	0.1363	17.3	0.1762	17.7581	0.154374
25	20	18.0877	0.0956	17.3	0.1350	17.7209	0.113956
26	21	18.0877	0.1387	17.3	0.1762	17.7209	0.156148
27	19	18.0377	0.0506	17.2	0.0947	17.6836	0.069283
28	20	18.0377	0.0981	17.3	0.1350	17.6836	0.115819
29	15	17.3471	0.1565	16.7	0.1133	17.0489	0.136593
30	13	17.2971	0.3305	16.7	0.2846	17.0116	0.308587
31	19	16.0102	0.1574	15.2	0.2000	15.8874	0.163822
32	13	14.8129	0.1395	13	0.0000	14.7451	0.134235
33	12	14.663	0.2219	13	0.0833	14.6333	0.219441
34	12	14.4131	0.2011	13	0.0833	14.447	0.203918
35	21	22.2245	0.0583	17	0.1905	21.3077	0.014651
36	21	21.475	0.0226	18.8	0.1048	20.7489	0.01196
37	19	21.1045	0.1108	17.2	0.0947	20.3029	0.068573
38	18	17.7731	0.0126	15.2	0.1556	17.3747	0.03474
39	24	21.8133	0.0911	17.3	0.2792	20.8445	0.131479
40	25	21.8133	0.1275	17.3	0.3080	20.8445	0.16622
41	18	19.9357	0.1075	15.2	0.1556	19.16	0.064446
MMRE		***0.1005***		***0.1057***		***0.0985***	
PRED(20)		***0.9024***		***0.9268***		***0.9268***	

Table 4. MMRE & PRED(20) comparison between estimation models

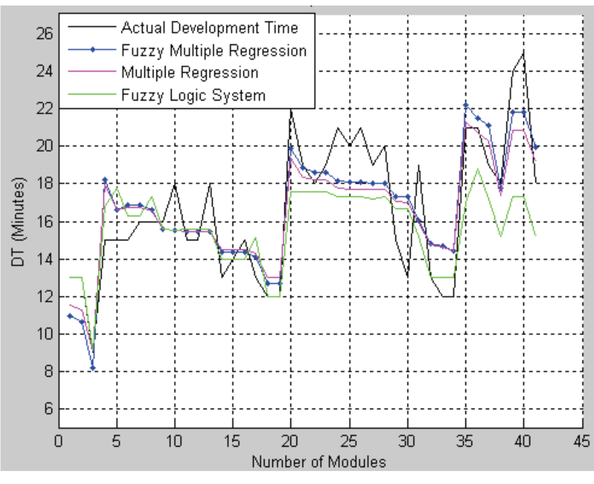


Fig. 2. comparison between estimation models

indicated that fuzzy multiple regression model is better than linear regression equations and fuzzy models in both evaluation criterion (PRED(20) & MMRE).

Comparison between actual development time, Multiple Regression Model, Fuzzy Logic and Fuzzy Multiple Regression Model is shown in Fig. 2. This figure is showed that fuzzy multiple regression output is close to actual development time in compare to the other models.

7. Conclusions and Future Research

The goal of this paper is to investigate the models for estimating software project. These techniques have been compared in terms of accuracy. Research demonstrates that fuzzy multiple regression models are better than linear regression equations and fuzzy models.

An ongoing research is related to apply neural network models using Bayesian Regularization training algorithm to data subset, because is more stable than fuzzy models that have membership functions whose derivatives have discontinuities at some points.

8. References

Moataz A. Ahmed, Moshood Omolade. Saliu, J. AlGhamdi, "Adaptive fuzzy logic-based framework for software development effort prediction", Elsevier Science, *Information and software technology*, pp 31-48, 2005.

V.Xia, D.Ho, L.F.Capretz, "Calibrating Function Points Using Neuro-Fuzzy Technique", 2005.

S.G.MacDonell, "Software source code sizing using fuzzy logic modelling", Elsevier Science, *Information and software technology*, pp 389-404, 2003.

L.M.Cuauhtemoc, J.Leboeuf, M.C.Yanez, T.Agustin Gutierrez, " Software Development Effort Estimation Using Fuzzy Logic: A Case Study", *IEEE, Proceedings of the Sixth Mexican International Conference on Computer Science (ENC'05)*, 2005.

X.Gu, G.Song, L.Xiao, "Design of a Fuzzy Decision-making Model and Its Application to Software Functional Size Measurement", *IEEE, International Conference on Computational Intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, 2006.

C.J.Burgess, M.Lefley, "Can genetic programming improve software effort estimation? A comparative evaluation", Elsevier Science, *Information and software technology*, pp 863-873, 2001.

L.M.Cuauhtemoc, Y.M Cornelio, G.T Agustin, "A Fuzzy Logic Model Based Upon Reused and New & Changed Code for Software Development Effort Estimation at Personal Level", *IEEE, Proceedings of the 15th International Conference on Computing (CIC'06)*, 2006.

H.B. Kuan Tan, Y.Zhao, H.Zhang, "Estimating LOC for Information Systems from their Conceptual Data Models", ACM, 2006.

O.Lima, P. P. Muniz Farias, A.D.Belchior," A Fuzzy Model for Function Point Analysis to Development and Enhancement Project Assessments", *CLEI Electronic Journal*, Vol.5, No.2, 1999.