Published in the Proceedings of Smef 2006

Estimating with COSMIC-FFP Functional Size Measurement Using Fuzzy Sets

Gu Xunmei, Song Guoxin, Xiao Lizhong

Abstract

Functional Size Measurement (FSM) methods are intended to measure the size of software by quantifying the functional user requirements (FUR) of the software. The recent development of COSMIC-FFP as a functional size measurement metric has greatly improved the functional sizing of real-time systems. This paper attempts to achieve a fuzzy logic based generalisation of effort models for COSMIC-FFP data. The data sets are used to develop two different fuzzy models, based on the f-COCOMO and the fuzzy linear regression. Finally, a conclusion is drawn according to the two models.

1. Introduction

In order to manage a software project, it is of great importance to make accurate estimates of what effort (i.e. cost, time etc.) it will take to complete the project. Having accurate estimates enables qualified decisions about whether it is economically viable to carry it through. It also allows more efficient resource planning and utilisation.

Functional Size Measurement (FSM) methods are intended to measure the size of software by quantifying the Functional User Requirements (FUR) of the software. The capability to accurately quantify the size of software in an early or late stage of the development lifecycle is very important to software project indicators. The COSMIC-FFP (Common Software Measurement International Consortium) is a candidate method of FSM. The recent development of COSMIC-FFP as a functional size measurement metric has greatly improved the functional sizing of real-time systems. Furthermore, fuzzy logic models have been widely applied in estimation because their fuzzy output reduces the reliance on a single value for estimation, which can be very misleading if the value is erroneous. Their application in software metrics environment is valuable because it can take into account the contribution of linguistic factors (‘very low’, ’low’, ’nominal’, ’high’, ’very high’, ’extra-high’).

In spite of the presence of many fuzzy logic estimation models, none exists for a COSMIC-FFP estimation model. This paper attempts to achieve a fuzzy logic [1][2] based generalisation of effort models for COSMIC-FFP data [6]. The data sets are used to develop two different fuzzy models, based on the f-COCOMO and the fuzzy linear regression. The f-COCOMO model is extended and a fuzzy set model representation of COSMIC-FFP size and effort is developed. Its performance in comparison with the fuzzy regression model is considered. The fuzzy regression model is used to investigate the independence of the effort from other factors of the software project. Usually, estimating formulas are elaborated from databases of past projects that serve as experimental data for classical regression techniques.

This paper is structured as follows. Section 1 only makes an introduction. Related work about software metrics is described in section 2. The fuzzy set and fuzzy regression models developed are outlined in Section 3.The preliminary result of software metrics using fuzzy sets are presented in Section 4 and a conclusion is drawn in Section 5.

2. Related work

The sources of fuzziness include both the imprecision of the limited set of linguistic labels that are used to assess them, and also the unavoidable uncertainty of human approximate judgments about highly abstract concepts (like required reliability, complexity or technological volatility) or collective human capabilities (like experience or overall ability in performing specific work roles).

Related work on fuzziness regarding software estimation includes the generalisation of input model parameters through fuzzy numbers [3], the use of similarity relations in estimating by analogy [4], and the fuzzy estimation of function points [5].

When the COCOMO (Constructive Cost Model) was published at the beginning of eighties [7], fuzzy logic was not grounded on solid theoretical foundations. This was now been achieved; Zadeh and others did so in the nineties. Thus, it is not surprising that some of the concepts defined or used in COCOMO are somewhat incompatible with the fuzzy logic. A fundamental limitation of the original COCOMO model is that it forces the use of crisp real numbers to express the values of the variables, despite the fact that they are inherently fuzzy.

The f-COCOMO model, developed by Musilek, introduces the concept of fuzziness into the original COCOMO. The original inputs of the COCOMO model include a project size estimate and parameters, which affect the effort. The f-COCOMO accepts the fuzzified form of project size and provides a fuzzy estimate of effort [3].

3. The estimation principle using fuzzy set

The COSMIC-FFP size estimation method accounts for all the functionality within a real-time system. The efficacy of the COSMIC-FFP can be completely exploited only when it is applied in effort estimation models early in the software lifecycle. But of late, only linear and non-linear regression models for effort estimation with COSMIC-FFP have been explored [6]. In this paper, the data sets in [6] are used to develop the f-COCOMO and fuzzy linear regression models. In the linear and non-linear regression models, there is also a lot of uncertainty in the modelling because the other factors affecting the software effort have not been identified. Thus the applicability of fuzzy models is justified.

3.1. Linear and non-linear regression models

The simple linear regression model with one independent variable size is presented in [6]. Then multiple regression models were produced and these models took into account additional variables that could have an influence on the productivity of software projects.

With the full data set of 15 observations, the simple linear regression model with one independent variable, functional size, was constructed in Figure 1. The model yielded $Effort=91.16+0.768\times Cfsu$. Here, Cfsu is a standard unit of measurement, i.e. one COSMIC-FFP functional size unit. This model has a coefficient of determination (R^2) of 0.38. Here, the coefficient R^2 describes the percentage of variability and the value is between 0 and 1; when an R^2 is close to 1, it indicates that variability in the response to the predictive variable can be explained by the model, i.e. there is a strong relationship between the independent and dependent variables. So when the R^2 is 0.38, it indicates that only 38% of the total variability of the dependent variable effort is explained by the variability of the independent variable functional size. This provides an initial indication that either the relationship might not be linear or that another variable might have a significant influence on the relationship between size and work effort.

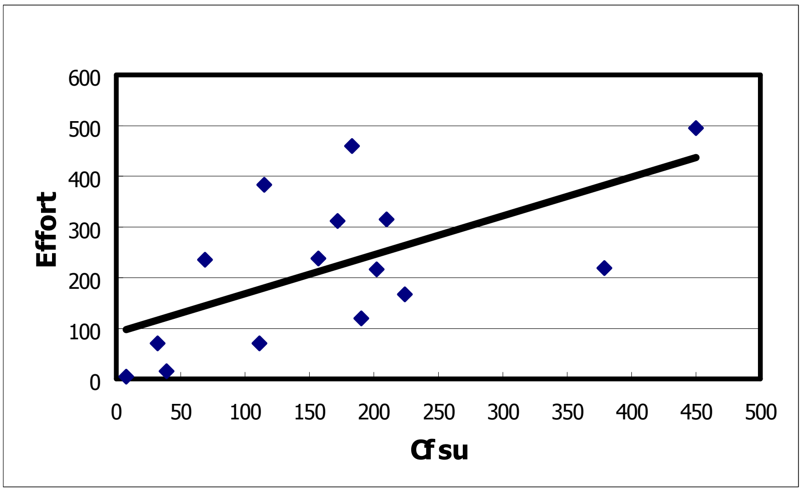

Figure 1: Linear regression model on functional size of organisation A

Besides the linear regression models, various other non-linear regression models were investigated, as illustrated in Table 1. Like the above, this table is based on the same 15 observations. It can be observed that a model with an exponential line ($Y=A\times X^{B}$) gives a good R^2 (0.716). This implies logarithmic transformations of the data to allow for the linearisation required by simple linear regression. This is recognized in the literature as a 'linearizable' function based on appropriate transformations. Therefore, the equation for this data set is: $log\ Effort=1.046\times log(Cfsu^{-0.02687})$ with an $R^2 = 0.72$. From the Table 1, we conclude that the relationship between the increase in effort due to an increase in size is not strong.

Table 1: Non-linear regression models

Equation	A	B	R	R^2
$Y=A\times X^{B}$	0.94	1.046	0.846	0.716
$Y=A\times e^{(B\times X)}$	45.95	0.006	0.662	0.387
Y=A+B×ln(X)	-241.20	96.792	0.664	0.442
Y=A+B/X	265.16	-2570.562	0.524	0.274
Y=1/(A+B×X)	0.057	-0.0002	0.477	0.228

3.2. f-COCOMO model combining the fuzzy set

The COSMIC-FFP size-effort relation is modelled with two models based on fuzzy logic, the fuzzy set model and the fuzzy linear regression. The fuzzy set model was developed as an extension of the f-COCOMO model [3]. The f-COCOMO model uses the fuzzy extension principle to represent the effort in terms of the size by including the equation being fuzzified as a constraint.

Let $C(x)$ be the fuzzy representation of the COSMIC-FFP size and $E(e)$ be the fuzzy effort output. Let $f(x)$ be a crisp function relating the effort and size. The fuzzy set representation is given as [3]:

$$E(e) = \sup_{x\in R; e=f(x)} [C(x)] \tag{1}$$

Applying the extension principle, the constraint on the above relation, given by $e=f(x)$ is eliminated and represented as follows[3]:

$$E(e) = [C(f^{-1}(e))] \tag{2}$$

Here, the fuzzy effort and size are assumed to have a triangular membership function. In the f-COCOMO model, the constraint *f(x)* is given by the basic COCOMO equation given by $E = aK^b$, where E is the effort and K is the size in LOC. The coefficients, 'a' and 'b' are assigned by certain specified values based on the type of software [3]. The f-COCOMO model has a simple version, where the coefficients are assumed to be crisp in the model and the augmented version, wherein 'a' and 'b' are considered as discrete fuzzy sets [3].

This paper uses the fuzzy extension principle; to represent crisp models developed from the enhancement data sets in [6] in a fuzzy set. The simple f-COCOMO [9] cannot be applied since it would not take into consideration the variability in the coefficients. The augmented f-COCOMO doesn't correspond to the COSMIC-FFP equations since the coefficient values are not specific. Therefore, the fuzzy set model defined by f-COCOMO is extended by representing the coefficients as continuous triangular fuzzy numbers. This is represented as follows:

$$E(e) = \sup_{x,a,b \in R; e = f(x)} [t\{C(x), A(a), B(b)\}] \tag{3}$$

Here, A (a) and B (b) represent the fuzzy coefficients (assuming the equation under consideration has two coefficients). Using the extension principle to eliminate the constraint, we can have:

$$E(e) = [t\{C(f^{-1}(e), A(a), B(b))\}] \tag{4}$$

The models that are fuzzified are those developed from the two data sets [6] using the average unit cost model and regression. And the average unit cost model [6] is defined as the ratio of effort to functional size:

$$Average\ unit\ cost = \frac{1}{n}\sum_{i=1}^{n}\frac{Effort_i}{Size_i} \tag{5}$$

This equation gives the average effort (in hours) to produce one unit of functional size. An example of a productivity/estimation model built with *a unit cost (estimated)=Average cost×Size*.

An equation derived from the data sets was:

$$Effort = a \times Cfsu \tag{6}$$

This is an average unit cost model, which was calculated on that sample of 15 projects [6] carried out within organisation A: 1.47 hours/Cfsu. So the estimation model built from this average unit cost is then: *Effort=1.47×Cfsu*. Each of the COSMIC-FFP data movement types (e.g. entry, exit, read and write) is assigned 1 Cfsu.

From the above context, another equation can be derived as follows:

$$Effort = a \times Cfsu + b \tag{7}$$

In fact, this equation is a result of a simple linear regression with size as the only independent variable. The coefficient is given above.

The two given equations of (6) and (7) are fuzzified by representing the size and the coefficients in terms of triangular fuzzy numbers and deriving the fuzzy effort output of organisation A. The triangular fuzzy numbers are given by:

$$C(x) = \begin{cases} \dfrac{x-\alpha}{m-\alpha} & x \in [\alpha, m] \\ \dfrac{\beta - x}{\beta - m} & x \in [m, \beta] \\ 0 & x \notin [\alpha, \beta] \end{cases} \tag{8}$$

Here, α and β indicate the range in which the triangular fuzzy number lies and m is the modal value i.e. the value which has the degree of membership in Figure 2.

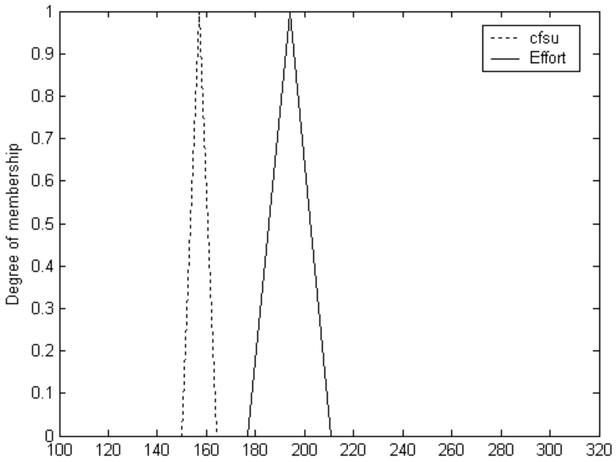

Figure 2: Triangular fuzzy size and effort

3.3. Fuzzy Linear Regression

In the modelling data sets, which are not normally distributed, the regression model lost most of its efficiency. Generally, software engineering data sets are not normal. It has been established that there is a lot of uncertainty existing with software engineering data. Therefore, using regression may not be completely efficient. Furthermore, in the fuzzy regression methods, the unfitted errors between the model and the observed data are viewed as the fuzziness of the model structure. The fuzzy regression model is required to find a regression model that fits the data within a particular fitting criterion that the fuzziness should be at its minimum. According to the fuzzy linear regression model, the regression coefficients are assumed to be triangular fuzzy numbers. Therefore, the estimated dependent variable represented by $\hat{Y}$ is also a fuzzy number. But Y represents the value of the observed variable.

A fuzzy regression analysis with a single dependent variable is given by:

$$\hat{Y} = \tilde{A}_0 + \tilde{A}_1 X \qquad (9)$$

Here, $\tilde{A}_0$ is a fuzzy intercept coefficient. Each fuzzy coefficient, is expressed as $\tilde{A} = (m_i, c_i)$, where m_i is the fuzzy centre and c_i is the fuzzy half-width. This approach defines the determination of the fuzzy coefficients such that the fuzzy output has minimum width, while a specific target of degree of belief h is satisfied. This value is also defined as the goodness of fit or measure of compatibility between the data and the model. This means that each of $\hat{Y}_i$ or crisp Y_i must fall within the estimated at h level. Here, h is the goodness of fit of the model with the data [9]. The main assumption in all the cases of the fuzzy regression is that the coefficients are symmetric fuzzy numbers. The formulae can also be extended for the asymmetric case, which is illustrated by [10]. Also, the inputs in this model, i.e. the independent variables are assumed to be crisp.

The fuzzy linear regression equations for the crisp data are as follows:

Minimize: $S = \sum_{j=1}^{n} W_j \bullet S_j$ (10)

Subject to:

$$\begin{cases} \sum_{j=1}^{n} x_{ij} a_j + (1-h_0) \sum_{j=1}^{n} | x_{ij} | e_j \geq y_i - (1-h_0) c_i \\ \sum_{j=1}^{n} x_{ij} a_j - (1-h_0) \sum_{j=1}^{n} | x_{ij} | e_j \leq y_i - (1-h_0) c_i \\ \quad e_j \succ 0; j = 1, 2, \cdots, n; i = 1, 2, \cdots, m \end{cases} \tag{11}$$

The fuzzy regression model for the crisp data can be represented by:

$$\hat{Y} = \tilde{A}_0 + \tilde{A}_1 X_1 + \tilde{A}_2 X_2 \tag{12}$$

This equation for a particular data set is developed by defining the minimisation problem subject to some certain constraints, which are generated from the individual elements in the data set. Each element gives rise to two constraints. Here, these regression coefficients are also expressed as $\tilde{A} = (e_i, c_i)$.

Therefore, according to the data set of organisation A, if the value of h_0 is given as 0.5,the fuzzy equation was derived as follows:

$$\hat{Y} = (170.9, 358.6) + (0.6, 0) X \tag{13}$$

$$S = 358.6 \tag{14}$$

From the above equations, we know that the lower the fuzziness of the system, the smaller is the indefiniteness of the system and also the smaller are the deviations between the observed and the estimated value.

On the basis of the above premise, we can propose a fuzzy linear regression model for the data sets including the difficulty parameter as an additional independent variable. The fuzzy effort model for two independent variables of COSMIC-FFP size and project difficulty variable is established as:

$$\hat{Y} = (0, 2721624) + (0.777, 0) X_1 + (60.572, 0) X_2 \tag{15}$$

$$S = 272.162 \tag{16}$$

Here, X_1 is the COSMIC-FFP size and X_2 is the project difficulty variable.

The comparison of the fuzziness of the two models about the equation (13) and (15) will provide information as to which model is a better fit to the data. The value of S or the system fuzziness for this model was found to be 272.162 while the fuzziness for that model, which only considers the size, was found to be 358.6. Therefore, the system fuzziness for the model, which considers the difficulty parameter additionally, is lower and hence the indefiniteness of the system is reduced. If the centre value of a fuzzy coefficient of an independent variable is 0, then it doesn't contribute to the model and hence the dependent variable doesn't depend on that particular variable. We also can find that project difficulty doesn't contribute to the effort estimate.

4. Preliminary Results

From the above analysis, two different fuzzy models, based on f-COCOMO and the fuzzy linear regression, are developed and optimized. The fuzzy set model is based on the extension principle while the fuzzy regression model depends on the premise that the error between the observed and estimated values is the fuzziness within the system.

The indices used to assess the performance used in the fuzzified regression models are the possibility values and the coverage of the actual value of effort in the fuzzy effort. Here, the possibility value is the degree to which the actual effort belongs to the fuzzy set that is the estimated effort. The coverage value is the average value of possibility for the data set. The variation of the coverage value is defined as $spread = (\beta - \alpha)/m$. The size input to this model is a fuzzy one. The spread here varies the fuzziness of the size input and hence we can get different values for coverage for the actual value of effort for different values of the input spread.

For the fuzzy linear regression model, the *h* value defines the level of the degree of belonging of the actual value of effort to the fuzzy effort. If the *h* value is given as 0.5, then the degree of belonging (or possibility) of the value of effort in the fuzzy effort estimate will be ⩾0.5 for all the values in the data set. From the Figure 3, it is obvious that an increase in *h* increases the output (i.e. effort) spread. An increase in the spread means an increase in the fuzzy width. The *h* value effectively varies directly with the fuzzy half width. Varying the *h* value in the fuzzy linear regression model and varying the spread coefficient in the fuzzy set model can effectively vary the fuzzy half width of the fuzzy output effort. A comparison of the relationship between coverage (i.e. average possibility value) and the output spread coefficients for the two models were performed. The variation in the coverage with *h* value and that with the input spread for the fuzzy set model is analogous to each other. A similar variation of coverage with the spread of the effort can be found.

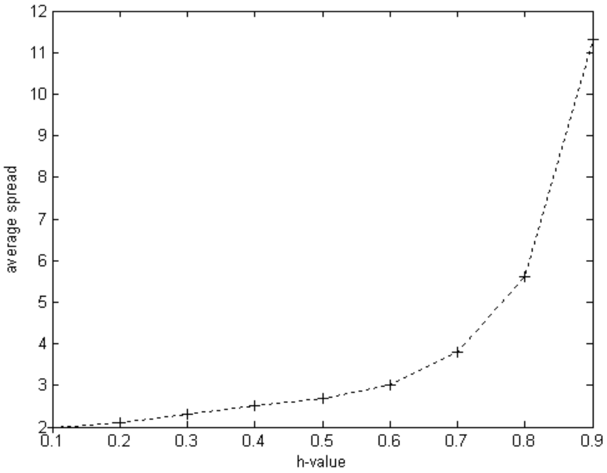

Figure 3: Spread vs. h-value for fuzzy regression model

From the above analysis, it can be inferred that the coverage values obtained from the fuzzy linear regression models are far more than those obtained from the fuzzy set model. The fuzzy linear regression model provides a certain level of control by providing variation in the degree of fit. This can be used to control the output spread and hence limit the uncertainty.

5. Conclusions

The size of the software is an important characteristic when estimating the required development effort. The size is often measured in terms of lines of code (i.e. LOC). Functional Size Measurement (FSM) is established as language-independent measures that are used to express the functionality of software, usually from the user's point of view. But Functional Size Measurement only measures a size of software; it does not consider other factors such as the difficulty of a software project etc.

In this paper, we have proposed the use of fuzzy sets and fuzzy regression to estimate the size and effort of a software project. A comparison was made between the two models. It was inferred that variation in the *h* value in the linear regression and the input (i.e. software size) spread value in the fuzzy set model leads to an increase in the fuzzy widths of the effort used as output.

6. References

[1] Cao Bingyuan, Applied Fuzzy Mathematics and System, Science Publishing House, October 2005

[2] Hu Qingbao, Fuzzy Theory Basis, Wu Han Publishing House, October 2004

[3] Musilek, P., Pedrycz, W., Succi, G., Reformat, M., Software Cost Estimation with Fuzzy Models, Applied Computing Review, Vol.8, No.2, 2000, pp.24-29

[4] Idri, A., Abran, A., Khoshgoftaar, T. M., Fuzzy Analogy: A New Approach for Software Cost Estimation. In: Dumke/Abran (eds.) Current Trends in Software Measurement, Shaker Publ., Germany, pp.127-142, 2001

[5] Souza Lima Jr., O., Farias, P.P.M, Belchior, A.D., Fuzzy Function Point Analysis. In: Proceedings of the 4th European Conference on Software Measurement and ICT Control, May 2001, Heidelberg, Germany, pp.161-172

[6] Abran, I.Silva, L.Primera, "Fields Studies using functional size measurement in building estimation models for software maintenance", Journal of Software Maintenance and Evolution, 2002(14)

[7] B.W.Boehm, R.W.Wolverton, "Software cost modeling: some lessons learned", Journal of System and Software, 1:195-201,1980

[8] H, Moskowitz and K.Kim, "On Assessing the H value in Fuzzy Linear Regression", Fuzzy Sets and Systems 58(1993) pp.303-327

[9] Alii Dri And Alain Abran, "COCOMO Cost Model Using Fuzzy Logic", 7th International Conference on Fuzzy Theory & Technology, Atlantic City, New Jersey, February 27-March 3, 2000

[10] Hsiao-Fan Wang, Ruey-Chyn Tsaur, "Insight of a Fuzzy Regression Model", Fuzzy Sets and Systems 112(2000), pp.355-369