Statistical Approach for Predicting Factors of Mood Method for Object Oriented

Firas Jassim1, Fawzi Altaani2

1 Management Information Systems Department, Irbid National University, Irbid, Jordan
2 Management Information Systems Department, Irbid National University, Irbid, Jordan

Abstract
Object oriented design is becoming more popular in software development and object oriented design metrics which is an essential part of software environment. The main goal in this paper is to predict factors of MOOD method for OO using a statistical approach. Therefore, linear regression model is used to find the relationship between factors of MOOD method and their influences on OO software measurements. Fortunately, through this process a prediction could be made for the line of code (LOC), number of classes (NOC), number of methods (NOM), and number of attributes (NOA). These measurements permit designers to access the software early in process, making changes that will reduce complexity and improve the continuing capability of the design.

Keywords: Software engineering, Software metric, Object Oriented, MOOD.

1. Introduction
Software metrics are most often proposed as the measurement tools of choice in empirical studies in software engineering, and the field of software metrics is the most often discussed from the perspective referred to as measurement theory. Software Metrics can be defined by measuring quality or characteristic of a software objects in any complex software project. Object oriented approach is capable of classifying the problem in terms of objects and provide many benefits like reliability, reusability, decomposition of problem into easily understood object and aiding of future modifications [2]. Nowadays, a quality engineer can choose from a large number of object–oriented metrics. The question posed is not the lack of metrics but the selection of those metrics which meet the specific needs of each software project. A quality engineer has to face the problem of selecting the appropriate set of metrics for his software measurements. A number of object–oriented metrics exploit the knowledge gained from metrics used in structured programming and adjust such measurements so as to satisfy the needs of object–oriented programming. On the other hand, other object–oriented metrics have been developed specifically for object–oriented programming and it would be pointless to apply them to structured programming [6]. Recently, many companies have started to introduce object-oriented (OO) technologies into their software development process. Many researchers have proposed several metrics suitable for measuring the size and the complexity of OO software. Some of them are in terms of Function Point (FP), others are in the terms of Lines of Code (LOC). Traditional metrics such as (FP) are unsatisfactory for predicting software size. On the other hand, LOC are quit satisfactory because it can be used to measure the software size [1, 7].

2. MOOD Method
The MOOD (Metrics for Object-Oriented Design) method is a collection of metrics which is used to evaluate the main abstraction of OO [4], such as inheritance, encapsulation, coupling, and information hiding or polymorphism and finally how to reuse that, together, for the increase in software quality. MOOD includes the following metrics [3, 5, 6, 13]:

- Method Hiding Factor (MHF)
- Attribute Hiding Factor (AHF)
- Method Inheritance Factor (MIF)
- Attribute Inheritance Factor (AIF)
- Coupling Factor (CF)
- Polymorphism Factor (PF)
These metrics are intended to present the presence or the absence of a certain property or attribute. Mathematically speaking, it can be viewed as probabilities ranging from 0 (total absence) to 1 (total presence).

Objects are an encapsulation of information that is relative to some entity. The class can be viewed as an abstract data type (ADT), which includes two types of features: methods and attributes, where the number of defined methods in a class \( C_i \) is given as:

\[
M_d(C_i) = M_v(C_i) + M_h(C_i)
\]

\( M_d \) (represents defined methods), \( M_v \) (represents visible methods), and \( M_h \) (represents hidden methods).

Then we define the Method Hiding Factor (MHF), as follows:

\[
MHF = \frac{\sum_{i=1}^{TC} M_h(C_i)}{\sum_{i=1}^{TC} M_d(C_i)}
\]

\( TC = \text{Total Classes} \)

Conversely, the number of attributes defined in class \( C_i \) (using the same manner above) is given by:

\[
A_d(C_i) = A_v(C_i) + A_h(C_i)
\]

\[
AHF = \frac{\sum_{i=1}^{TC} A_h(C_i)}{\sum_{i=1}^{TC} A_d(C_i)}
\]

And all other factors are calculated using similar mathematical formulas. So, \( MIF \) and \( AIF \) can be defined through equations (5) and (6), as:

\[
MIF = \frac{\sum_{i=1}^{TC} M(C_i)}{\sum_{i=1}^{TC} M_d(C_i)}
\]

\[
AIF = \frac{\sum_{i=1}^{TC} A_i(C_i)}{\sum_{i=1}^{TC} A_d(C_i)}
\]

PF is defined as the ratio of the actual number of possible different polymorphic situations for class \( C_i \) to the maximum number of possible distinct polymorphic situations for class \( C_i \), and can be defined as:

\[
PF = \frac{\sum_{i=1}^{TC} M_o(C_i)}{\sum_{i=1}^{TC} [M_n \times DC(C_i)]}
\]

where \( M_o \) represents overridden methods, \( M_n \) for new methods, and DC for descendants methods.

Polymorphism arises from inheritance and [10] suggest that in some cases overriding methods could contribute to reduce complexity and therefore to make the system more understandable and easier to maintain. While, [14] have shown that this metric is a valid measure within the context of the theoretical framework.

Finally, CF is defined as the ratio of the maximum possible number of couplings in the system to the actual number of couplings not imputable to inheritance.

\[
CF = \frac{\sum_{i=1}^{TC} \sum_{j=1}^{TC} is\_client(C_i,C_j)}{TC^2 - TC}
\]

where:

\[
TC^2 - TC = \text{maximum number of coupling in a system with TC classes.}
\]

\[
is\_client(C_i,C_j) = \begin{cases} 1 & \text{iff } C_i \Rightarrow C_j \land C_i \neq C_j \\ 0 & \text{otherwise} \end{cases}
\]

Coupling Factor (CF) has a very high positive correlation with all quality measures [11]. Therefore, as coupling among classes increases, the defect density and normalized rework is also expected to increase. This result shows that coupling in software systems has a strong negative impact on software quality and then should be avoided during design. In fact, many authors have noted that it is desirable that classes communicate with as few others as possible.
because coupling relations increase complexity, reduce encapsulation and reuse.

3. Estimation of Factors

MOOD method used widely to measure many target OO programs and many studies have compare it with other methods. Mainly, our focus will be on line of code (LOC), number of classes (NOC), number of methods (NOM), and number of attributes (NOA), so to reach this; we have collect our data from 33 systems [9, 10, 12, 14] to be suitable for normal distribution curve. Results obtained using SPSS package.

According to table (1), we can plot the relation between LOC (in the x-axis), and NOC, NOM, and NOA (in the y-axis), as shown in fig.1.

<table>
<thead>
<tr>
<th></th>
<th>NOL</th>
<th>NOC</th>
<th>NOM</th>
<th>NOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15837</td>
<td>65</td>
<td>1446</td>
<td>537</td>
</tr>
<tr>
<td>2</td>
<td>23570</td>
<td>57</td>
<td>1535</td>
<td>876</td>
</tr>
<tr>
<td>3</td>
<td>47016</td>
<td>91</td>
<td>2141</td>
<td>1178</td>
</tr>
<tr>
<td>4</td>
<td>23154</td>
<td>51</td>
<td>1420</td>
<td>538</td>
</tr>
<tr>
<td>5</td>
<td>20747</td>
<td>154</td>
<td>2814</td>
<td>1113</td>
</tr>
<tr>
<td>6</td>
<td>44930</td>
<td>92</td>
<td>2224</td>
<td>1132</td>
</tr>
<tr>
<td>7</td>
<td>28582</td>
<td>71</td>
<td>1978</td>
<td>839</td>
</tr>
<tr>
<td>8</td>
<td>19254</td>
<td>69</td>
<td>1815</td>
<td>675</td>
</tr>
<tr>
<td>9</td>
<td>20085</td>
<td>74</td>
<td>1876</td>
<td>700</td>
</tr>
<tr>
<td>10</td>
<td>57086</td>
<td>140</td>
<td>322</td>
<td>81</td>
</tr>
<tr>
<td>11</td>
<td>92231</td>
<td>201</td>
<td>481</td>
<td>124</td>
</tr>
<tr>
<td>12</td>
<td>167541</td>
<td>355</td>
<td>735</td>
<td>204</td>
</tr>
<tr>
<td>13</td>
<td>261260</td>
<td>562</td>
<td>1193</td>
<td>297</td>
</tr>
<tr>
<td>14</td>
<td>838128</td>
<td>1966</td>
<td>3227</td>
<td>611</td>
</tr>
<tr>
<td>15</td>
<td>2062982</td>
<td>5107</td>
<td>6735</td>
<td>2297</td>
</tr>
<tr>
<td>16</td>
<td>2129555</td>
<td>5035</td>
<td>7292</td>
<td>2294</td>
</tr>
<tr>
<td>17</td>
<td>1948354</td>
<td>4566</td>
<td>5975</td>
<td>2095</td>
</tr>
<tr>
<td>18</td>
<td>64492</td>
<td>222</td>
<td>210</td>
<td>81</td>
</tr>
<tr>
<td>19</td>
<td>70514</td>
<td>243</td>
<td>229</td>
<td>88</td>
</tr>
<tr>
<td>20</td>
<td>113919</td>
<td>349</td>
<td>325</td>
<td>132</td>
</tr>
<tr>
<td>21</td>
<td>177356</td>
<td>565</td>
<td>516</td>
<td>185</td>
</tr>
<tr>
<td>22</td>
<td>6593</td>
<td>324</td>
<td>1310</td>
<td>60</td>
</tr>
<tr>
<td>23</td>
<td>1023</td>
<td>25</td>
<td>103</td>
<td>220</td>
</tr>
<tr>
<td>24</td>
<td>1729</td>
<td>20</td>
<td>134</td>
<td>185</td>
</tr>
<tr>
<td>25</td>
<td>50000</td>
<td>46</td>
<td>2025</td>
<td>510</td>
</tr>
<tr>
<td>26</td>
<td>300000</td>
<td>1000</td>
<td>11000</td>
<td>10960</td>
</tr>
<tr>
<td>27</td>
<td>500000</td>
<td>1617</td>
<td>37191</td>
<td>17141</td>
</tr>
<tr>
<td>28</td>
<td>9189</td>
<td>339</td>
<td>1993</td>
<td>4022</td>
</tr>
<tr>
<td>29</td>
<td>7102</td>
<td>45</td>
<td>711</td>
<td>482</td>
</tr>
<tr>
<td>30</td>
<td>830</td>
<td>10</td>
<td>175</td>
<td>89</td>
</tr>
<tr>
<td>31</td>
<td>1602</td>
<td>26</td>
<td>180</td>
<td>247</td>
</tr>
<tr>
<td>32</td>
<td>3451</td>
<td>18</td>
<td>170</td>
<td>145</td>
</tr>
<tr>
<td>33</td>
<td>549</td>
<td>15</td>
<td>33</td>
<td>172</td>
</tr>
</tbody>
</table>

Total N = 33

Table 1: Product metrics from 33 commercial samples

Now, by implementing log transform to avoid large number scale we can plot the data again as fig. 2.

The main contribution in this article is to use statistics, especially regression; to predict number of classes needed for the software, also number of attributes and methods needed. Hence, linear regression model is used to find the relationship between factors and their influences on OO software measurements. Fortunately, through this process we can predict the suitable number of LOC, classes (objects), methods, and attributes we need to satisfy the software metrics using MOOD.
4. Regression Analysis

Actually, we can use linear regression model to predict the LOC, NOC, NOM, and NOA needed. Statistically speaking, in order to investigate the correlations and relationships between the object-oriented metrics and software quality we conducted a correlation and a multiple linear regression analysis. The mathematical formula for the model is as follows:

\[ \text{LOC} = \beta_0 + \beta_1 \text{NOC} + \beta_2 \text{NOM} + \beta_3 \text{NOA} \] (10)

\[ \text{NOC} = \beta_0 + \beta_1 \text{LOC} + \beta_2 \text{NOM} + \beta_3 \text{NOA} \] (11)

\[ \text{NOM} = \beta_0 + \beta_1 \text{LOC} + \beta_2 \text{NOC} + \beta_3 \text{NOA} \] (12)

\[ \text{NOA} = \beta_0 + \beta_1 \text{LOC} + \beta_2 \text{NOC} + \beta_3 \text{NOM} \] (13)

Each time we have used one variable as an independent variable while the others as the dependent variables. To reach the fact that, each one of these variables responsible for the efficiency of the MOOD method. The regression analysis shows the values of the coefficients of the model \((\beta_0, \beta_1, \beta_2, \text{and} \ \beta_3)\).

The independent variable in an experiment is the variable that is systematically manipulated by the investigator. In most experiments, the investigator is interested in determining the effect that one variable; has one or more effect on the other variables. On the other hand, the dependent variable in an experiment is the variable that the investigator measures to determine the effect of the independent variable.

First, we consider LOC as the dependent variable and the other factors as the independent variables, equation 10, table (2) shows the value of \((\beta_0, \beta_1, \beta_2, \text{and} \ \beta_3)\), and the significances \((p\text{-value})\).

Since the value of significant \((p\text{-value})\) is less than 0.05. This means that LOC mainly affect the other factor according to table (4), which shows the ANOVA (ANalysis Of VAriance).

Second, we consider NOC as the dependent variable and the other factors as the independent variables, table (5) shows the value of \((\beta_0, \beta_1, \beta_2, \text{and} \ \beta_3)\), and the significances \((p\text{-value})\).

Also, if we want to use the values of the coefficients above, we may re-write the regression line as:

\[ \text{NOC} = 24.439 + 0.002 \text{LOC} - 0.006 \text{NOM} + 0.037 \text{NOA} \]

Therefore, if we want to predict the value of NOC we can substitute the given values of NOC, NOM, and NOA in the above formula and get an estimated (predicted) value for NOC. Also, from the values of \(p\text{-value}\) we can see that the values of \(\beta_1 \text{ and} \ \beta_3\) only are less than 0.05, so we can conclude that LOC are mainly affected by NOC and NOA.

Also, from the values of \(p\text{-value}\) we can see that the values of \(\beta_1 \text{ and} \ \beta_3\) only are less than 0.05, so we can conclude that LOC are mainly affected by NOC and NOA.

There is some statistical measures used to measure the goodness of fit and it is an indicator of how well the model fits the data. The higher the value of R square, the more accurate the model is. These values can be seen in table (3).

Table 2: Results of \(\beta_0, \beta_1, \beta_2, \text{and} \ \beta_3\) when LOC is the dependent variable

<table>
<thead>
<tr>
<th>Regression coefficients</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_0)</td>
<td>9458.918</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>421.994</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>3.025</td>
</tr>
<tr>
<td>(\beta_3)</td>
<td>-16.009</td>
</tr>
</tbody>
</table>

So, if we want to use the values of the coefficients above, we may re-write the regression line as:

\[ \text{LOC} = -9458.918 + 421.994 \text{NOC} + 3.025 \text{NOM} - 16.009 \text{NOA} \]

Therefore, if we want to predict the value of LOC we can substitute the given values of NOC, NOM, and NOA in the above formula and get an estimated (predicted) value for LOC.

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.
conclude that NOC are mainly affected by LOC and NOA. On the other hand, NOM does not affect LOC too much. As previously mentioned the values of R square and the ANOVA table are shown in tables 6 & 7.

Table 6: The value of R square and adjusted R square for the regression model when NOC is the dependent variable

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.998</td>
<td>.997</td>
<td>.996</td>
<td>87.55</td>
</tr>
</tbody>
</table>

Table 7: ANOVA results for NOC as the dependent variable

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>64118680</td>
<td>3</td>
<td>2 1372893.45</td>
<td>2788.435</td>
<td>.000</td>
</tr>
<tr>
<td>Residual</td>
<td>222280.2</td>
<td>29</td>
<td>7 664.035</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>64340961</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similarly, we can do the same thing for NOM and NOA, but we mainly focused on the LOC and NOC because of their main role in MOOD method [8].

5. Conclusions

A simple and easy technique has been constructed to use statistics for predicting the values of MOOD factors, in the same manner one can use this technique to estimate other factors rather than LOC, NOC, NOM, and NOA, which can be used to evaluate software quality. Additionally, using linear regression model can be extended to non-linear model and multivariate analysis to add more complicated model to give more accurate estimation for these factors and also use another statistical estimation approaches such as maximum Likelihood Estimator (MLE) to give better estimation than regression model, and to be standards for MOOD method and to give more accurate measurements for object-oriented metrics.

Acknowledgments

Author Firas A. Jassim pays his regards to Mrs. Hind Emad Qassim for giving her moral support and help to carry out this research work.

References


Firas Jassim received the BS degree in mathematics and computer applications from Al-Nahrain University, Baghdad, Iraq in 1997, and the MS degree in mathematics and computer applications from Al-Nahrain University, Baghdad, Iraq in 1999 and the PhD degree in computer information systems from the University of Banking and Financial Sciences, Amman, Jordan in 2012. His research interests are Image processing, image compression, image enhancement, image interpolation and simulation.  

Fawzi Altaani received the BS degree in public administration from Al-Yarmouk University, Irbid, Jordan 1990, and Higher Diploma in health service administration from university of Jordan, 1991 and the MS degree in health administration from Red Sea University, Soudan in 2004 and the PhD degree in management information systems from the University of Banking and Financial Sciences, Amman, Jordan in 2010. His research interests are management information system and public administration, and Image processing.