Empirical Validation of Complexity and Extensibility Metrics for Software Product Line Architectures

Edson A. Oliveira Junior, José C. Maldonado
Computing Systems Department
University of São Paulo (USP)
São Carlos-SP, Brazil
Email: {edsonjr, jcmaldon}@icmc.usp.br

Itana M. S. Gimenes
Informatics Department
State University of Maringá (UEM)
Maringá-PR, Brazil
Email: itana@din.uem.br

Abstract—The software product line (PL) architecture (PLA) is one of the most important PL core assets as it is the abstraction of the products that can be generated, and it represents similarities and variabilities of a PL. Its quality attributes analysis and evaluation can serve as a basis for analyzing the managerial and economical values of a PL. We proposed metrics for PLA complexity and extensibility quality attributes. This paper is concerned with the empirical validation of such metrics. As a result of the experimental work we can conclude that the metrics are relevant indicators of complexity and extensibility of PLA by presenting their correlation analysis.

Keywords—complexity; empirical validation; extensibility; metrics; product line architecture; software product line.

I. INTRODUCTION

In the last decades effective methodologies to evaluate software architectures, such as ATAM (Architecture Trade-off Analysis Method) and SAAM (Software Architecture Analysis Method), were proposed and consolidated by both industrial and academic segments [6]. Such a consolidation is corroborated by the analysis of the number of published research papers and technical reports providing important examples of how to carry out a software architecture evaluation based on quality attributes. Thus, these methodologies are essential for evaluating single-product architectures.

However, in recent years, the software product line (PL) [8] engineering has emerged as a promising reusability approach, which brings out some important benefits, such as increases the reusability of its core assets, while decreases the time to market. One of the most important assets of a PL is its architecture (PLA). The PLA plays a central role at the development of products from a PL.

The evaluation of a PLA must be supported by a set of metrics [7]. Such metrics must both evidence the quality of PL and serve as a basis to analyze the managerial and economical value of a PL [2]. The PLA must exploit the common (similarities) and variable (variabilities) aspects of a PL. The variability impact analysis on the PL development can determine the aggregated value of a PL for an organization. Metrics for a PLA are applied to a set of assets from which variants can be generated rather than one specific product. Thus, it is necessary to define specific PLA metrics to provide effective indicators with regard to the overall PL development and evolution.

We proposed two metrics for PLA quality attributes [10]: one for complexity (CompPLA) and one for extensibility (ExtensPLA). These metrics were defined to provide an indicator of how complex and extensible is a PL by measuring derived PL products. Complexity is measured based on McCabe’s Cyclomatic Complexity [9] while extensibility is measured based on the relation between abstract classes and methods over concrete classes and methods. Class extensibility is calculated taking into account the number of abstract methods divided by the total number of methods (concrete plus abstract) of a class. Thus, component extensibility is the sum of the extensibility of all classes that form a component. The metrics are presented as follows:

CompPLA: is the sum of the complexity of each component of a PLA. This metric is represented as:

\[
\text{CompPLA(PLA)} = \sum_{i=1}^{nCpt} \text{CompV arComponent}(Cpt_i),
\]

where:

- \( \text{CompV arComponent} \) is the complexity of a PLA component. It is the sum of the complexity of each variability in class
- \( Cpt_i \) is the \( i^{th} \) component of a PLA
- \( nCpt \) is the number of variable components of a PLA

ExtensPLA: is the sum of the extensibility of each component of a PLA. This metric is represented as:

\[
\text{ExtensPLA(PLA)} = \sum_{i=1}^{nCpt} \text{ExtensV arComponent}(Cpt_i),
\]

where:

- \( \text{ExtensV arComponent} \) is the extensibility of a PLA component. It is the sum of the extensibility of each variability in class
- \( Cpt_i \) is the \( i^{th} \) component of a PLA
- \( nCpt \) is the number of variable components of a PLA

Variabilities are related to PLA class and/or components. Each variability is related to variation points and/or variants.
that realize it. A variation point or variant might be a PLA class or component. The complexity or extensibility of a variability can be calculated based on the complexity or extensibility of each variation point or variant. Thus, \textit{CompVarComponent} is the sum of the complexity of all variation points and/or variants related to a specific variability in a PLA component. \textit{ExtensVarComponent} is the sum of the extensibility of all variation points and/or variants related to a specific variability in a PLA component.

Both theoretical and empirical validations [4] are necessary to validate a set of metrics. Theoretical validation is concerned with demonstrating that a metric is measuring the concept it is purporting to measure. The first requirement for theoretical validation is that either the analyst has an intuitive understanding of the concept that is being measured and/or that the software engineering community has a consensual intuitive understanding of the concept. Theoretical validation of the complexity and extensibility metrics have been done in [10].

This paper is concerned with the empirical validation of the proposed metrics for PLA complexity and extensibility quality attributes. The validation aims at correlating the metrics \textit{CompPLA} and \textit{ExtensPLA} with subject’s complexity and extensibility rating, respectively, when generating PLA configurations. A PLA configuration represents a derived PL product with variabilities resolved.

This paper is organized as follows: Section II presents how the experimental study was planned and carried out to validate the complexity and extensibility metrics; Section III discusses the results obtained in this study; and Section IV provides the conclusions and directions for future work.

II. EXPERIMENTAL STUDY

In this section we describe the experiment we have carried out to empirically validate the proposed metrics as indicators of PLA complexity and extensibility.

We have followed the suggestions provided by Wohlin et al. [15] and Perry et al. [11] on how to perform controlled experiments with minor changes. We are concerned with the following main activities: definition, planning, operation, analysis and interpretation, validity evaluation, and presentation and package.

A. Definition

Based on the Goal-Question-Metric (GQM) template [1], the goal of the experiment is presented as follows:

**Analyze** collected metrics from UML models and source code

**For the purpose of** validating

**With respect to** the capability to be used as PLA complexity and extensibility indicators

**From the point of view of** software product line architects

In the context of graduate students and lecturers of the Software Engineering area at the University of Waterloo (UWaterloo), University of São Paulo (ICMC-USP), and State University of Maringá (UEM)

B. Planning

1) **Context Selection:** The experiment was carried out in an academic environment as it composed of a group of graduate students and a lecturer of the Software Engineering area.

2) **Selection of Subjects:** The six subjects were: one Master student and one PhD candidate from ICMC-USP, one lecturer from UEM, one Master student and two PhD candidates from UWaterloo. They have at least minimal experience in the design of product lines and variabilities using UML.

3) **Variable Selection:** The independent variables were the class and component complexity and extensibility of a PLA. The dependent variables were the complexity and extensibility of each product generated from the PLA.

4) **Instrumentation:** The objects were: a document describing the Arcade Game Maker (AGM) PL [12]; AGM UML class and component models, a traceability model from classes to components; and a resolution model containing the variabilities to be resolved at class level. In addition, each subject received a copy of the experiment’s Terms of Agreement and a questionnaire to answer. The independent variables were measured by the proposed metrics. The dependent variables were measured according to the subjects ratings of complexity and extensibility.

5) **Hypothesis Formulation:** The following hypothesis were tested in this study:

- **Null Hypothesis (H0):** There is no significant correlation between the PLA complexity metric (\textit{CompPLA}) and the subject’s complexity rating for each PLA configuration, and neither between the PLA extensibility metric (\textit{ExtensPLA}) and the subject’s extensibility rating for each PLA configuration;

- **Alternative Hypothesis (H1):** There is a significant correlation between the PLA complexity metric (\textit{CompPLA}) and the subject’s complexity rating for each PLA configuration, but there is no significant correlation between the PLA extensibility metric (\textit{ExtensPLA}) and the subject’s extensibility rating for each PLA configuration;

- **Alternative Hypothesis (H2):** There is a significant correlation between the PLA extensibility metric (\textit{ExtensPLA}) and the subject’s extensibility rating for each PLA configuration, but there is no significant correlation between the PLA complexity metric (\textit{CompPLA}) and the subjects complexity rating for each PLA configuration; and

- **Alternative Hypothesis (H3):** There is a significant correlation between both the PLA complexity metric (\textit{CompPLA}) and the subjects complexity rating, and the...
PLA extensibility metric (ExtensPLA) and the subjects extensibility rating.

6) Experiment Design: It was chosen a within-subject design experiment. It means that all the tasks had to be solved by each of the subjects.

C. Operation

1) Preparation: When the experiment was carried out, all of the subjects had graduated in the Software Engineering area, in which they have learned how to design at least object-oriented (OO) class diagrams using UML. In addition, all of the subjects had at least basic experience in applying PL and variability concepts to OO systems designed using UML. Moreover, subjects were given a training session with respect to the AGM PL before the experiment took place. However, the subjects were unaware about the hypothesis formulated.

The material prepared to the subjects consisted of:
• the class diagram representing the core asset of the AGM PL;
• the AGM component diagram, representing its logical architecture;
• an AGM traceability model from classes to components;
• the description of the AGM PL;
• the SMartyProfile, which is a UML metamodel, thus the subjects can understand how the variabilities are represented in class and component diagrams;
• a variability resolution model, which the subjects could resolve the variabilities to generate AGM configurations; and
• a test (questionnaire) describing complexity and extensibility concepts, which the subjects had to rate the associated complexity and extensibility of each generated AGM configuration based on the following linguistic labels (Table I).

We selected five linguistic labels as we considered they are significant to cover all the possible categories of our variables: complexity and extensibility. The selection of an odd and prime number of labels was based on some guidelines provided by [3], as such a number brings out balance in order to obtain better results.

2) Execution: The six subjects were given the material described in Preparation (Section II-C1). It was required to each subject to generate five AGM configurations. It was done by following instructions on how to resolve the AGM variability resolution model, and how to rate the complexity and extensibility associated to the configurations generated from the subjects view point. All the tasks were performed by each subject alone, with no time limit to solve them. All the data was collected including the subjects rating obtained from the responses of the questionnaire and the metric values were calculated manually.

3) Data Validation: The tasks realized by the subjects were collected. Once they were complete, we verified whether the subjects have at least depth knowledge in designing class diagrams for OO systems and basic knowledge in PL and variability management using UML. These facts are corroborated by analyzing the questionnaires responses. Thus, we consider the subjects subjective evaluation reliable.

D. Analysis and Interpretation

We summarized the collected data by calculating the metrics CompPLA and ExtensPLA for the thirty AGM configurations generated by the six subjects, as well as verifying the complexity and extensibility rating of such configurations. Table II presents the observed values for CompPLA and ExtensPLA metrics from the generated AGM configurations.

1) Descriptive Statistics and Frequency Distribution: Figure 1 presents the descriptive statistics and frequency distribution of the CompPLA observed values (Table II).

Figure 2 presents the descriptive statistics and frequency distribution of the ExtensPLA observed values (Table II).

We can observe that in Figure 2 there are fifteen configurations with value 0.61 for extensibility. There are three main abstract classes in the AGM PL. Each of these classes belongs to a different game that can be produced from the AGM PL. Most of the generated configurations has exactly one game. Thus, for the AGM PL, products with exactly one game have at least value 0.61 for extensibility.

2) Normality Tests: We can clearly observe that the CompPLA and the ExtensPLA observed values distributions (Figure 1 and Figure 2) are non-normal. In spite of it, Shapiro-Wilk and Kolmogorov-Smirnov normality tests were conducted to make sure of it.

The following hypothesis were proposed for both normality tests with regard to the CompPLA metric:
• Null Hypothesis (H₀): the CompPLA observed values distribution is normal, i.e., the significance value (p) is greater than 0.05 (p > 0.05); and
• Alternative Hypothesis (H₁): the CompPLA observed values distribution is non-normal, i.e., the significance value (p) is less or equal to 0.05 (p ≤ 0.05).

Taking into account a sample size (N) of 30, with mean (µ) 0.6545, standard deviation (σ) 0.1842, and median (x) 0.5895, the CompPLA metric obtained a significance value:
• p < 0.01 (0.01 < 0.05) for the Kolmogorov-Smirnov test;
• p = 0.0118 (0.0118 < 0.05) for the Shapiro-Wilk test.
Table II

OBSERVED VALUES FOR COMPPLA AND EXTENSPLA METRICS FROM THE GENERATED CONFIGURATIONS.

<table>
<thead>
<tr>
<th>Configuration #</th>
<th>CompPLA</th>
<th>ExtensPLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.51</td>
<td>0.61</td>
</tr>
<tr>
<td>2</td>
<td>0.56</td>
<td>0.61</td>
</tr>
<tr>
<td>3</td>
<td>0.51</td>
<td>0.81</td>
</tr>
<tr>
<td>4</td>
<td>0.83</td>
<td>0.80</td>
</tr>
<tr>
<td>5</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>0.50</td>
<td>0.61</td>
</tr>
<tr>
<td>7</td>
<td>0.53</td>
<td>0.61</td>
</tr>
<tr>
<td>8</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>0.47</td>
<td>0.61</td>
</tr>
<tr>
<td>10</td>
<td>0.57</td>
<td>0.80</td>
</tr>
<tr>
<td>11</td>
<td>1.00</td>
<td>0.61</td>
</tr>
<tr>
<td>12</td>
<td>0.61</td>
<td>0.80</td>
</tr>
<tr>
<td>13</td>
<td>0.48</td>
<td>0.61</td>
</tr>
<tr>
<td>14</td>
<td>0.69</td>
<td>0.61</td>
</tr>
<tr>
<td>15</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>16</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>17</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td>18</td>
<td>0.82</td>
<td>0.80</td>
</tr>
<tr>
<td>19</td>
<td>0.52</td>
<td>0.61</td>
</tr>
<tr>
<td>20</td>
<td>0.82</td>
<td>0.80</td>
</tr>
<tr>
<td>21</td>
<td>0.49</td>
<td>0.61</td>
</tr>
<tr>
<td>22</td>
<td>0.47</td>
<td>0.61</td>
</tr>
<tr>
<td>23</td>
<td>0.53</td>
<td>0.61</td>
</tr>
<tr>
<td>24</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>25</td>
<td>0.62</td>
<td>0.80</td>
</tr>
<tr>
<td>26</td>
<td>0.47</td>
<td>0.61</td>
</tr>
<tr>
<td>27</td>
<td>0.53</td>
<td>0.61</td>
</tr>
<tr>
<td>28</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>29</td>
<td>0.40</td>
<td>0.61</td>
</tr>
<tr>
<td>30</td>
<td>0.78</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Thus, there are evidences, for both normality tests, that the null hypothesis ($H_0$) must be rejected at a significance level of 5%. Then, we cannot consider the CompPLA observed values distribution normal and, consequently, a non-parametric statistic method must be used to analyze the data.

We also performed both normality tests to the ExtensPLA observed values distribution (Figure 2), to which we pro-
posed the following hypothesis:

- **Null Hypothesis** ($H_0$): the ExtensPLA observed values distribution is normal, i.e., the significance value ($p$) is greater than 0.05 ($p > 0.05$); and
- **Alternative Hypothesis** ($H_1$): the ExtensPLA observed values distribution is non-normal, i.e., the significance value ($p$) is less or equal to 0.05 ($p \leq 0.05$).

Taking into account a sample size ($N$) of 30, with mean ($\mu$) 0.7390, standard deviation ($\sigma$) 0.1487, and median ($\tilde{x}$) 0.7060, the ExtensPLA metric obtained a significance value:

- $p < 0.01$ ($0.01 < 0.05$) for the Kolmogorov-Smirnov tese;
- $p = 0.00001$ ($0.00001 < 0.05$) for the Shapiro-Wilk tese.

Thus, there are evidences, for both normality tests, that the null hypothesis ($H_0$) must be rejected at a significance level of 5%. Then, we cannot consider the ExtensPLA observed values distribution normal, specially for Shapiro-Wilk significance, which is way far to be normal, and, consequently, a non-parametric statistic method must be used to analyze the data.

3) **Spearman’s Rank Correlation**: As both CompPLA and ExtensPLA distributions are non-normal, we applied the non-parametric Spearman’s Correlation method [14] to support the interpretation of the data. This method allows to establish whether there is a correlation between two sets of data.

The Spearman’s Rank Correlation ($\rho$) is calculated following these steps:

1) create a table from the two sets of data to be correlated;
2) rank the two data sets. Ranking is achieved by giving the ranking “1” to the biggest number in a column, “2” to the second biggest value and so on. The smallest value in the column will get the highest ranking. This should be done for both sets of measurements. Tied scores are given the mean (average) rank;
3) find the difference in the ranks (d): this is the difference between the ranks of the two values on each row of the table;
4) square the differences (d^2);
5) sum the squared differences (\sum d^2); and
6) calculate the coefficient ($\rho$) using Equation (3). The answer will always be between 1.0 and -1.0 as shown in Figure 3.

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2, \text{ where } n \text{ is the sample size (N)} \quad (3)$$

We performed three correlations:

**Corr.1 CompPLA and the subjects complexity rating,** which shows that the understanding of complexity by the subjects corroborates to the CompPLA metric, establishing how to measure complexity in PLA;

**Corr.2 ExtensPLA and the subjects extensibility rating,** which shows that the understanding of extensibility by the subjects corroborates to the ExtensPLA metric, establishing how to measure extensibility in PLA; and

**Corr.3 CompPLA and ExtensPLA,** which shows whether there is a correlation between complexity and extensibility in terms of PLA.

Table III presents the Spearman’s ranking correlation for Corr.1. The Spearman $\rho$ coefficient (Equation 3) for Corr.1 is calculated as follows:

$$\rho_{\text{Corr.1}} = 1 - \frac{6}{30(360^2-30)} \times 293.5 = 1 - \frac{6}{26970} \times 293.5 = 1.0 - 0.07 = 0.93$$

Thus, according to Figure 3, there is a strong positive correlation ($\rho_{\text{Corr.1}} = 0.93$) between the metric CompPLA and the subjects complexity rating.

Table IV presents the Spearman’s ranking correlation for Corr.2. The Spearman $\rho$ coefficient (Equation 3) for Corr.2 is calculated as follows:

$$\rho_{\text{Corr.2}} = 1 - \frac{6}{30(360^2-30)} \times 841.5 = 1 - \frac{6}{26970} \times 841.5 = 1.0 - 0.19 = 0.81$$

Thus, according to Figure 3, there is a strong positive correlation ($\rho_{\text{Corr.2}} = 0.81$) between the metric ExtensPLA and the subjects extensibility rating.

Table V presents the Spearman’s ranking correlation for Corr.3. The Spearman $\rho$ coefficient (Equation 3) for Corr.3 is calculated as follows:

$$\rho_{\text{Corr.3}} = 1 - \frac{6}{30(360^2-30)} \times 2139 = 1 - \frac{6}{26970} \times 2139 = 1.0 - 0.48 = 0.52$$

Thus, according to Figure 3, there is a strong positive correlation ($\rho_{\text{Corr.3}} = 0.52$) between the metrics CompPLA and ExtensPLA.

Based on correlations Corr.1 and Corr.2, we have evidences to reject the null hypothesis $H_0$ of the study, and accept the alternative hypothesis $H_1$ (Section II-B5), which states that both complexity and extensibility metrics are significantly correlated to complexity and extensibility of PLA.

Corr.3 demonstrates that there is a strong correlation between the complexity and extensibility quality attributes. It means that when performing quality attribute trade-off analysis, complexity and extensibility must be carefully analyzed in order to prioritize one of them.

Linear regressions were used to demonstrate how the CompPLA and ExtensPLA metrics are related to each other, as well as to express the equation gathered from such correlations.

4) **Linear Regression**: We applied the linear regression to the three correlations in order to mathematically show the equation for each correlation done.

Figure 4 shows the linear regression graph for the CompPLA metric and the subjects complexity rating.
Such a correlation can be expressed in terms of the following equation:

\[
y = 4.7213x + 0.6433, \text{ where}
\]

\[x \text{ is the CompPLA value}
\]

\[y \text{ is the subjects complexity rating}
\]

Based on Equation (4), obtained from Corr.1 linear regression, we proposed a scale to measure the complexity of a PLA configuration, presented in Figure 5. To measure the complexity of a given PLA configuration just replace its CompPLA value with the “\(x\)” in Equation 4. Then, compare the obtained result (the \(y\) value) to the complexity scale (Figure 5) and find how complex such a PLA configuration
Table V
SPEARMAN’S CORRELATION FOR CORR. 3: COMPPLA AND EXTENSPLA.

<table>
<thead>
<tr>
<th>Config. #</th>
<th>CompPLA</th>
<th>ra</th>
<th>ExtensPLA</th>
<th>rb</th>
<th>d</th>
<th>d^2</th>
<th>Config. #</th>
<th>CompPLA</th>
<th>ra</th>
<th>ExtensPLA</th>
<th>rb</th>
<th>d</th>
<th>d^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.51</td>
<td>22.5</td>
<td>0.61</td>
<td>23.0</td>
<td>6.5</td>
<td>0.25</td>
<td>16</td>
<td>0.98</td>
<td>2.0</td>
<td>1.00</td>
<td>3.0</td>
<td>1.0</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.56</td>
<td>16.0</td>
<td>0.61</td>
<td>23.0</td>
<td>7.0</td>
<td>0.70</td>
<td>17</td>
<td>0.77</td>
<td>11.0</td>
<td>0.80</td>
<td>11.0</td>
<td>1.0</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0.51</td>
<td>22.5</td>
<td>0.81</td>
<td>6.0</td>
<td>16.5</td>
<td>272.25</td>
<td>18</td>
<td>0.62</td>
<td>7.5</td>
<td>0.80</td>
<td>11.0</td>
<td>3.5</td>
<td>12.25</td>
</tr>
<tr>
<td>4</td>
<td>0.83</td>
<td>8.0</td>
<td>0.80</td>
<td>11.0</td>
<td>5.0</td>
<td>25.00</td>
<td>19</td>
<td>0.92</td>
<td>20.5</td>
<td>0.61</td>
<td>23.0</td>
<td>2.5</td>
<td>6.25</td>
</tr>
<tr>
<td>5</td>
<td>0.91</td>
<td>4.0</td>
<td>0.61</td>
<td>23.0</td>
<td>19.0</td>
<td>361.00</td>
<td>20</td>
<td>0.82</td>
<td>7.5</td>
<td>0.80</td>
<td>11.0</td>
<td>3.5</td>
<td>12.25</td>
</tr>
<tr>
<td>6</td>
<td>0.50</td>
<td>24.0</td>
<td>1.00</td>
<td>3.0</td>
<td>21.0</td>
<td>441.00</td>
<td>21</td>
<td>0.49</td>
<td>25.0</td>
<td>0.61</td>
<td>23.0</td>
<td>2.0</td>
<td>4.00</td>
</tr>
<tr>
<td>7</td>
<td>0.47</td>
<td>27.5</td>
<td>0.61</td>
<td>23.0</td>
<td>4.5</td>
<td>20.25</td>
<td>22</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>3.0</td>
<td>2.0</td>
<td>4.00</td>
</tr>
<tr>
<td>8</td>
<td>0.53</td>
<td>18.0</td>
<td>0.61</td>
<td>23.0</td>
<td>5.0</td>
<td>25.00</td>
<td>23</td>
<td>0.52</td>
<td>20.5</td>
<td>0.61</td>
<td>23.0</td>
<td>2.5</td>
<td>6.25</td>
</tr>
<tr>
<td>9</td>
<td>0.67</td>
<td>14.0</td>
<td>0.80</td>
<td>11.0</td>
<td>3.0</td>
<td>9.00</td>
<td>24</td>
<td>0.42</td>
<td>29.0</td>
<td>0.61</td>
<td>23.0</td>
<td>6.0</td>
<td>36.00</td>
</tr>
<tr>
<td>10</td>
<td>0.90</td>
<td>6.0</td>
<td>1.00</td>
<td>3.0</td>
<td>2.0</td>
<td>4.00</td>
<td>25</td>
<td>0.62</td>
<td>10.5</td>
<td>0.60</td>
<td>11.0</td>
<td>4.0</td>
<td>16.00</td>
</tr>
<tr>
<td>11</td>
<td>0.53</td>
<td>18.0</td>
<td>1.00</td>
<td>3.0</td>
<td>15.0</td>
<td>225.00</td>
<td>26</td>
<td>0.47</td>
<td>27.5</td>
<td>0.61</td>
<td>23.0</td>
<td>4.5</td>
<td>20.25</td>
</tr>
<tr>
<td>12</td>
<td>0.97</td>
<td>3.0</td>
<td>0.61</td>
<td>23.0</td>
<td>20.0</td>
<td>400.00</td>
<td>27</td>
<td>0.53</td>
<td>18.0</td>
<td>0.61</td>
<td>23.0</td>
<td>5.0</td>
<td>25.00</td>
</tr>
<tr>
<td>13</td>
<td>0.48</td>
<td>26.0</td>
<td>0.61</td>
<td>23.0</td>
<td>3.0</td>
<td>9.00</td>
<td>28</td>
<td>0.70</td>
<td>12.0</td>
<td>0.80</td>
<td>11.0</td>
<td>1.0</td>
<td>1.00</td>
</tr>
<tr>
<td>14</td>
<td>0.89</td>
<td>13.0</td>
<td>0.81</td>
<td>23.0</td>
<td>19.0</td>
<td>100.00</td>
<td>29</td>
<td>0.40</td>
<td>30.0</td>
<td>0.61</td>
<td>23.0</td>
<td>7.0</td>
<td>49.00</td>
</tr>
<tr>
<td>15</td>
<td>0.74</td>
<td>11.0</td>
<td>0.80</td>
<td>11.0</td>
<td>0.0</td>
<td>0.00</td>
<td>30</td>
<td>0.78</td>
<td>9.0</td>
<td>0.80</td>
<td>11.0</td>
<td>2.0</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Figure 4. CompPLA x Subjects Complexity Rating Linear Regression.

is. For instance, let’s take a CompPLA value of a hypothetical configuration, which is 0.64. Resolving Equation 4, we find $y = 3.66$. Then, we can state that such a configuration has high complexity based on the proposed complexity scale (Figure 5).

Figure 6 shows the linear regression graph for the ExtensPLA metric and the subjects extensibility rating.

Such a correlation can be expressed in terms of the following equation:

$$\text{ExtensPLA} = (0.3934 \times \text{CompPLA}) + 0.4817 \quad (6)$$

Correlation from Figure 8 can be expressed in terms of the following equation:

$$\text{CompPLA} = (0.604 \times \text{ExtensPLA}) + 0.2081 \quad (7)$$

E. Validity Evaluation

In this section we discuss the empirical study’s threats to validity and how we tried to minimize them.

1) Threats to Conclusion Validity: The conclusion validity defines the extent to which conclusions are statistically valid. The only issue that we take into account is the sample size (N=30), which can be increased during prospective replications of this study in order to reach normality of the observed values.

\[ y = 4.7213x + 0.6433 \]

\[ R^2 = 0.9189 \]
2) Threats to Construct Validity: The construct validity is the degree to which independent and dependent variables are measured by their appropriated instruments. Our dependent variables are complexity and extensibility. We proposed subjective metrics for them, as linguistic labels, collected based on the subjects rating. As the subjects have enough experience in modeling OO systems using at least class diagrams, we take their ratings as significant. The construct
validity of the metrics used for the independent variables is guaranteed by some insights carried out on a previous study of metrics for PLA [10].

3) Threats to Internal Validity: The internal validity is the degree to which conclusions can be drawn about cause-effect of independent variables on the dependent variables. We dealt with the following issues:

- **Differences among subjects.** As we dealt with a small sample, variations in the subject skills were reduced by applying the within-subject task design. Thus, subjects experiences had approximately the same degree with regard to UML modeling, and PL and variabilities basic concepts.

- **Accuracy of subject responses.** Complexity and extensibility were rated by each subject. As they have medium experience in UML modeling, and PL and variabilities concepts, we considered their responses valid.

- **Fatigue effects.** On average the experiment lasted for 69 minutes, thus fatigue was considered not very relevant. Also, the variability resolution model contributed to reduce such effects.

- **Measuring PLA and Configurations.** As PLA can be analyzed based on its products (configurations), measuring derived configurations provide a means to analyze PLA quality attributes by allowing the performing of trade-off analysis to prioritize such attributes. Thus, we consider valid the application of the metrics to PLA configurations to rate the overall PLA complexity and extensibility.

- **Other important factors.** Influence among subjects could not really be controlled. Subjects did the experiment under supervision of a human observer. We believe that this issue did not affect the study validity.

4) Threats to External Validity: Based on the greater the external validity, the more the results of an empirical study can be generalized to actual software engineering practice, two threats of validity have been identified, which are:

- **Instrumentation.** We tried to use representative class and component diagrams of real cases. However, the PL used in the experiment is non-commercial, and some assumptions can be made on this issue. Thus, more empirical studies taking a “real PL” from software organizations must be done.

- **Subjects.** Obtaining well-qualified subjects was difficult, thus we used lecturers and advanced students from the Software Engineering academia. More experiments with practitioners and professionals must be carried out allowing us to generalize the study results.

### F. Presentation and Package

The documents related to this study is available at Edson’s web site [http://edsonjr.pro.br/expAGM01](http://edsonjr.pro.br/expAGM01) as we agree with the importance of the diffusion of the experimental data to external replications of the current experiment [5] and knowledge-sharing [13].

### III. Discussion of Results

Obtained results of the study lead us to conclude that the metrics CompPLA and ExtensPLA are relevant indicators of PLA complexity and extensibility, respectively. Although the correlation between CompPLA and ExtensPLA almost had a limit value, i.e. 0.52, we can conclude that they are strongly correlated as expressed by equations 6 and 7. Several more experiments must be carried out, as well as more PLA configurations must be both derived and incorporated to enhance the conclusions. In addition, we need to apply our metrics to a commercial PL in order to reduce external threats to the study validity and for gathering real evidences that these metrics can be used as complexity and extensibility indicators.
IV. Conclusion

Current literature claims the need of metrics to allow PL architects empirically analyze the potential of a PLA, as well as PL managers analyze the aggregated managerial and economical values of a PL throughout its products.

Performing empirical validation of metrics is essential to demonstrate their practical usefulness. The proposed metrics for complexity (CompPLA) and extensibility (ExtensPLA) PLA quality attributes were empirically validated based on their application to a set of 30 products generated by experiment subjects from the Arcade Game Maker (AGM) PL. The observed metric values were submitted to normality tests which proved their non-normality. Then, Spearman’s rank correlation was used to demonstrate the metrics correlations, which are:

• CompPLA has a strong positive correlation with the subjects complexity rating;
• ExtensPLA has a strong positive correlation with the subjects extensibility rating;
• CompPLA and ExtensPLA have a strong positive correlation with each other.

Linear regression was used to come up with equations that express the correlation between the metrics and the subjects rating, as well as the metrics themselves.

Although we have used a non-commercial PL to conduct our experiments, we had evidences that our proposed metrics can be used as relevant indicators of complexity and extensibility of a PLA based on its derived products.

We are currently proposing some changes on various issues to improve our experiments with metrics, which are:

• increase the derived configurations sample size, which is important to stay closer to real projects and to generalize the results;
• conduct experiments in a more controlled environment;
• deal with real data from commercial PL obtained from industrial environments; and
• recruit more subjects from the Software Engineering area, both from academic and industrial environments.

Acknowledgements

The authors would like to thank CAPES-Brazil for funding Edson’s visiting term (Feb-Dec/2009) at University of Waterloo, Canada, where the experiment was partially conducted, as well as Prof. Márcio Barros from UNIRIO, Marcilio Mendonça, Rafael Lotufo, Thiago Baritolomei, Eduardo Barrenechea, Prof. Ladan Tahvildari, Mehdi Amoui, and Maziejar Salehie from UWaterloo, Prof. Thelma Colanzi from UEM, and Marco A. Graciotto and Paulo G. Queiroz from ICMC-USP, for supporting this study.

References