In-virtuo Experimental Studies: An Approach Based on Genetic Algorithms

Rogério Eduardo Garcia[*1], **Jeffrey Carver**[1],
Maria Cristina Ferreira de Oliveira[2], **José Carlos Maldonado**[2]

[1]Department of Computer Science and Engineering
Mississippi State University
Butler Hall, Room 214/329, Box 9637
Mississippi State, MS 39762
regarcia@icmc.usp.br,carver@cse.msstate.edu

[2]Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Av. Trabalhador São-Carlense, 400, C.P. 668
CEP 13560-970 São Carlos – SP, Brazil
{*cristina,jcmaldon*}*@icmc.usp.br*

Abstract. *The cost of running empirical studies is high. Although needed to establish a broad knowledge base, it is not always feasible to conduct as many experiments as desired. To address these issues, in this paper we propose a process for conducting in-virtuo experiments. In the approach proposed, data from a real experiment is used to seed a genetic algorithm that then creates simulated data. In this paper we discuss generating simulated subjects based on the experience profiles of real subjects. This type of study, with its increased data set, allows researchers to examine the effects of various parameters on the results of the study. The major goal of in-virtuo studies is to uncover patterns that may not be evident in smaller data sets and then pose new hypotheses to be tested in real experiments. A benefit of in-virtuo experiments is that they do not incur the same costs as running additional in-vivo or in-vitro experiments.*

Resumo. *O custo para conduzir estudos empiricos é alto. Embora seja mandatório para estabelecer um corpo de conhecimento amplo, nem sempre é possível conduzir tantos experimentos quanto desejado. Neste artigo é proposto um processo para a condução de experimentos in-virtuo, em que dados reais são submetidos a um algoritmo genético para criar dados simulados. Como estudo de caso, apresentamos a geração de experiências de participantes baseado no perfil de participantes reais. Esse tipo de estudo permite que pesquisadores examinem os efeitos de vários parâmetros nos resultados de seus estudos. O principal objetivo de estudos in-virtuo, nesta abordagem, é permitir a descoberta de de padrões que podem não ser evidente em conjunto de dados pequenos, e com isso gerar novas hipóteses para serem verificadas em experimentos reais. Um benefío de experimentos in-virtuo é o custo reduzido, comparado aos experimentos in-vivo ou -in-vitro.*

Keywords: Empirical Software Engineering, In-Virtuo Experiments, Simulation, Genetic Algorithms.

[*]Rogerio Eduardo Garcia is a visiting student from the University of São Paulo at São Carlos

1. Introduction

Empirical Software Engineering attempts to measure the performance of models, methods and techniques for software development. These measurements provide a base of knowledge to aid developers in their selection of suitable methods, techniques and tools. It has been pointed out that a usable body of knowledge cannot be built from isolated studies, because results from any single experiment may not be representative of an entire population [Miller 2005]. Differences in system domains, subjects' motivation and previous experience, and the effect of professional settings and cultural backgrounds are known to affect subject performance [Shull et al. 2002], and typically cannot be isolated and controlled in a single experimental design.

In-vitro experimental studies such as those conducted in the scope of the Readers Project [Maldonado et al. 2005, Shull et al. 2003] demand considerable planning and long execution times. Travassos and Barros [Travassos and de Oliveira Barros 2003] argue that the traditional classification of SE experimental studies as either *in-vivo* or *in-vitro* should be augmented to include *in-virtuo* and *in-silico*. Also, they argue that, although *in-vivo* and *in-vitro* experiments are certainly necessary to build a body of knowledge, *in-virtuo* and *in-silico* experiments can help to increase productivity in empirical software engineering and reduce risks and costs. These types of studies would allow researchers to more easily vary the study parameters and observe the effect on the variables of interest. Based on the effects observed, researchers can then generate new hypotheses and new experiments. This scenario sets the motivation for this paper: we contribute towards improving experimentation in SE through *in-virtuo* experiments. For this improvement, we propose an approach for experimental design taking into account the needs of *in-virtuo* experiments. This approach addresses the initial step that eventually leads to simulation of experimental results, as proposed in Garcia et al. [Garcia et al. 2005].

This paper is organized as follows: Section 2. contains a brief background on the Taxonomy and Empirical Software Engineering Process, which is followed to establish our approach. In Section 3. we present the approach to define the model for simulating *in-virtuo* experiments, considering the knowledge acquired by previous *in-vitro* studies. Finally, Section 4. contains a discussion on the results and perspectives for further work.

2. Background

Extending software engineering studies beyond *in-vivo* and *in-vitro* studies to *in-virtuo* – studies in which real subjects interact with a computer model of reality – and *in-silico* – studies in which both the subjects and the "environment" are modeled on the computer – poses some interesting possibilities for study replication. First, an experiment may be replicated *in-vitro* – with or without variations – but actually requiring its physical execution. Second, *in-virtuo* replications may be conducted if environmental conditions can be simulated to map the experimental design onto a model that simulates results. Therefore, to conduct *in-virtuo* studies, behavior models must be defined. These models must represent consistent and coherent behavior in a defined context.

In-virtuo studies create additional needs in the experimentation process. The five-step experimental process by Wohlin et al. [Wohlin et al. 1999] was used as starting point. The first step – *Definition* – consists of clearly establishing the object of study, intention, effect studied, perspective and context. The foundation for the experiment must be stated in the *Planning* activity: a) the determination of the context, including the participants and the environment; b) the formal statement of the hypotheses (null hypothesis and alternative hypotheses); c) the definition of dependent and independent variables, including their scales and values; d) the identification of subjects; e) the design of the experiment, and the preparation of the instrumentation; f) validity considerations (internal, external, construct and conclusion). The *Operation* activity consists of three steps: preparation, execution and data validation. *Analysis & Interpretation* is concerned with understanding the data collected, reducing the data set, if possible, and hypothesis testing. *Packaging* activity is concerned with the Lab Package creation, documentation of results, and instructions for experimental replication.

3. Generating *In-Virtuo* Experiments

To define the model required for an *in-virtuo* experiment, knowledge from *in-vitro* experiments and parameters for the model are required. The activities to gather the knowledge and the parameters have to be added to the traditional experimentation process. To define the *in-virtuo* experiment (*Definition* activity), we need to consider two different situations. First, when using only one *in-vitro* experiment as input, Wohlin's definition tasks are already done. Therefore, a new question must be considered: how can the data set be expanded? The analyses and results from the original experiment are used to define the new objective within the constraints of the original experimental design. Secondly, using two or more *in-vitro* experiments as input requires some method for combination. After a separate analysis for each experiment, analyses of data sets combined must be conducted focused on identifying possible conflicts among the studies, such as context. In both cases, one study or more than one study, it is important to observe the dependent and the independents variables. The result of analyzing previous experiments defines the requirements and constraints for the model to be built. Some questions may arise at this point, such as "how does one combine data from multiple contexts?". Based on the analyses conducted, a new *in-virtuo* experiment can be defined, using Wohlin's task set . The experimenter must determine which characteristics of the *in-vitro* experiment will be explored in the *in-virtuo* experiment.

Focusing on the variables, the next activity (*Planning*) can be generalized as follows. To simulate new *in-virtuo* experiments we need to establish model parameters for the simulation. The first step in obtaining the parameters is to select independent variables from from previous (*in-vitro*) experiments. These independent variables will be used to map the dependent variable. For example, using the Readers Project experiments, independent variables could be the subject profile, made up of several variables that capture experiences in specific activities. Conversely, *Document* could not be used as an independent variable, because there is not enough information to specify a model. More information about the document is needed to support model specification, such as number of pages, defects seeded and their criticality.

The independent variables and the approach used to create the sampling model influence the sampling strategy. We have proposed an approach based on Genetic Algorithms (GA) to generate the scenarios. The idea is to create "virtual" values for independent variables based on genetic mutation of real values. In this approach we are using a heuristic based on genetic operators to mutate real values without considering if there is an improvement on fitness function. Such consideration allows introducing variations on real values, but not focused on obtaining the best values for independent variables. Combining Visual Data Analysis and Statistical Data Analysis allows researchers to define the metrics and how to obtain them. Statistical measures can support the choice of independent variable to be used in the model, and support the reduction of dimensionality to be considered. This analysis is useful to consider meaningful variables to define the model. To validate the distribution of the generated values, we have used measures of correlation among the chosen variables. These measures help guarantee that the virtual values could exist in real world (*in-vitro*) experiments. The model for sampling independent variables can be used to generate the scenarios to be explored. For example, it is possible to simulate a scenario using only experienced subjects, using a subset of original data set as source. When using this type of subset, we must determine the coverage for that subset. The *coverage* is a measure describing the representativeness of the generated data. If there are 100 source records and only 10 are used to support the generation, i.e. selected for mutation, it means that the model is not as broad as it could be. The same rationale can be applied to the generated data to understand the correlation between the generated and source data. Also, the knowledge needed to create the model is obtained by analyzing the entire and the generated population.

To validate the model for planning the *in-virtuo* experiment different strategies can be used depending on the available source data. One example is to use a subset of the data for planning the experiment and another subset for validation. The validation can be done using the same data analysis process used to extract measures and acquire knowledge. A second example is to use data sets from multiple replications. In the case of multiple data sets, it is important to note that meta-analysis must first be applied to combine data.

In the case of the single data set source, other metrics can be extracted for validation: e.g. the percentage of profiles selected from the original data set and the percentage of generated profiles that match original profiles . The first metric indicates how supported by real values the generated sample is: for example, if a sample is generated using only outliers in *in-vitro* experiments, then this metric indicates low support. The second metric indicates how related the generated values are to the real values: in this case, the ideal is that the simulated data is not very close or very far away from the original data. Very close would indicate that the simulated data is simply a copy of the original data. Very far would mean that dependent variables could not be safely simulated. A strategy to evaluate the similarity between real and simulated values must consider the measure chosen to support the mutation. Initial studies show that variation between 10% and 20% is enough to obtain data with new information keeping the main characteristics.

After data analysis of the planning data set (independent variables), *in-virtuo* experiments can be run. Although the simulated data set is not real – considering that it did not come from real subjects – it is possible to use it to enlarge the data set. This larger data set allows investigations that can be useful for generating new hypothesis. Both by generating new hypothesis and by conducting new *in-vivo* or *in-vitro* replications that aggregate meaningful data, it is possible to increase the body of knowledge.

3.1. Generating a new *In-Virtuo* Experiment

We have used data from two experiments conducted under the Readers Project to evaluate the proposed process. Therefore, the first step was to analyze data from the two experiments to evaluate the Lab Package. We needed to obtain information about constraints on independent variables, hypotheses already considered and establish the objective for the new study: to explore variations on subject profile. The next step was a cycle of analyses, both visual and statistical, to define a set of parameters for the model for generation of subjects. For example, Figure 1 presents the normal distribution obtained for experience as Developer, based on the entire source data set.

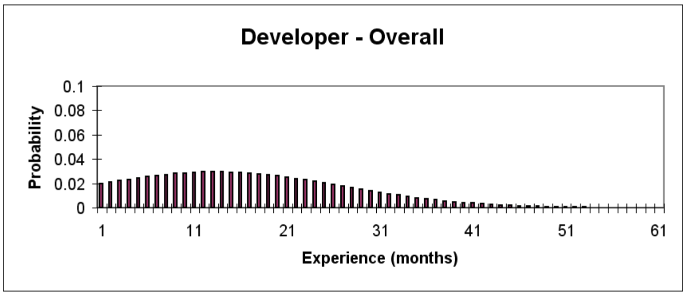

Figure 1. Normal Distribution to Experience as Developer

Next, a model based on GA was defined to generate subject profiles. The model takes a random real subject as seed and generates another profile through mutation, using the normal distribution as a parameter. The profiles generated were analyzed again to validate the strategy adopted. Comparing the correlations between items in the original profiles and those in the generated profiles, it was possible to conclude that the model obtained is coherent. For example, the correlation obtained between the *Experience as Analyst* and *Experience Using A Requirements Document* were 0.81 and 0.91, for the original and generated data sets respectively. The correlation between *Experience as Tester* and *Experience as Analyst* were 0.84 and 0.77, also for original and generated data sets respectively. Despite the variation, the correlations are similar. Comparing raw data separately, we observed that the model obtained is good enough: although there is a tendency to reduce the variation in the values, because of the measures used. The values obtained show that the GA approach introduced a jitter into the values of each variable.

4. Conclusions

A new and richer body of data can be generated by mapping the experimental design and results into a suitable GA model. This data allows researchers to speculate on the effects of the parameters on the results without the cost of running additional *in-vitro* experiments. The universe of *in-virtuo* experiments generated will also support the design of new *in-vitro* studies, i.e., we may investigate if the *in-virtuo* experiments would really occur in practice.

The generation of *in-virtuo* studies, based on results from *in-vitro* studies, opens up a new area for investigation. *In-vivo* and *in-vitro* replications require a lot of effort and a high cost because "real" subjects are involved. *In-virtuo* simulation allows us to cost-effectively explore different "what-if" scenarios, which can provide valuable insights for planning new studies. Another drawback of *in-vivo* and *in-vitro* replications is their typically small sample size. Small sample size with respect to the expected effect leads to weak evidence. By increasing the size of the data set, data analysis possibilities increase. Results become amenable to approaches such as data mining, visualization and visual data mining. These analysis techniques help uncover trends that can be explored in future *in-vivo* or *in-vitro* studies.

This paper illustrates a potential contribution of GA based *in-virtuo* experiment simulations. We illustrate the possibility of expanding the universe of data collected by simulating feasible data based on trends observed in a series of "real" experimental studies. With this approach it is possible to explore variations of the experimental conditions that may be hard to achieve in practice. For example, simulating experiments using only highly experienced subjects. We have validated the model using different data sets from real replications. Data obtained from a new *in-vitro* study might be used in the future to calibrate the simulation model, validating the simulated results obtained. As further work, analyses and meta-analyses – combining data from different scenarios – may be conducted on *in-silico* data sets to gain further insights on data trends and possible experimental variations. Finally a complete simulation, based on the GA approach, will be defined including defect list generation for "virtual" subjects.

Acknowledgments

The authors acknowledge the financial support of FAPESP – the State of São Paulo Research Funding Agency – Grant 01/13993-0.

References

[Garcia et al. 2005] Garcia, R. E., de Oliveira, M. C. F., and Maldonado, J. C. (2005). Genetic algorithms to support software engineering experimentation. (Accepted to International Symposium on Empirical Software Engineering – 2005).

[Maldonado et al. 2005] Maldonado, J. C., Carver, J., Shull, F., Fabbri, S., Dória, E., Martimiano, L., Mendonça, M., and Basili, V. (2005). Perspective-Based Reading: a Replicated Experiment Focused on Individual Reviewer Effectiveness. *Empirical Software Engineering: An International Journal - (EMSE)*. Accepted for publication in 7/5/05.

[Miller 2005] Miller, J. (2005). Replicating software engineering experiments: a poisoned chalice or the holy grail. *Information & Software Technology*, 47(4):233–244.

[Shull et al. 2002] Shull, F., Basili, V., Carver, J., Maldonado, J. C., Travassos, G. H., Mendonça, M., and Fabbri, S. (2002). Replicating software engineering experiments: Addressing the tacit knowledge problem. In IEEE Computer Society, editor, *2002 International Symposium on Empirical Software Engineering*, pages 7–16, Nara, Japan.

[Shull et al. 2003] Shull, F., Carver, J., Travassos, G. H., Maldonado, J. C., Conradi, R., and Basili, V. R. (2003). *Empirical Software Engineering*, chapter Replicated Studies: Building a Body of Knowledge about Software Reading Techniques, pages 39–84. River Edge, NJ.

[Travassos and de Oliveira Barros 2003] Travassos, G. H. and de Oliveira Barros, M. (2003). Contributions of in virtuo and in silico experiments for the future of empirical studies in software engineering. In Verlag, F. I., editor, *The Future of Empirical Studies in Software Engineering - 2ND International Workshop*, volume 2, pages 117–130, Roman Castles, Italy. ISBN 3-8167-6418-5.

[Wohlin et al. 1999] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (1999). *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers, Boston (USA).