Empirical Software Engineering, 7, 177–180, 2002.

Dissertation

Improved Software Cost Estimation – A Robust and Interpretable Modelling Method and a Comprehensive Empirical Investigation

ISABELLA WIECZOREK* Isabella.Wieczorek@iese.fhg.de
Fraunhofer Institute for Experimental Software Engineering, Sauerwiesen 6, 67661 Kaiserslautern, Germany

Keywords: Software cost estimation, estimation methods, model evaluation, multi-organisational software cost databases.

Delivering software on time, within budget, and to an agreed level of quality is crucial for the business reputation and competitiveness of many organisations. Accurate estimates are essential for a better project planning, tracking, and control and pave the way for successful product delivery. This calls for support in software cost estimation, since many software companies are working within tight schedules and finish their projects behind schedule and budget, if they finish them at all. This is because the estimation of cost for software projects leads to many practical measurement and modelling difficulties.

These difficulties are not satisfactory addressed by common estimation methods leading to their restricted acceptability. A number of cost estimation methods evolved from a variety of fields such as statistics, machine learning, and knowledge acquisition. Though many estimation methods have been developed, appropriately using such methods is not easy and may not always lead to accurate estimates. Many statistical, data-driven methods produce models that are still difficult to interpret. However, the estimated values are meaningless without an understanding of their accuracy and sensitivity to risks associated with the project. Beside the interpretability of their models, estimation methods have to deal with several constraints prevalent in software engineering data, such as uncertainty in predictions, or problems associated with missing data. Appropriately dealing with these constraints has an impact on prediction accuracy. Existing commonly used estimation methods are limited as they only address some of these issues and thus provide potential for improvements (Briand and Wieczorek, 2002).

* Ph.D. thesis submitted and accepted at the University of Kaiserslautern. Advisors: Prof. Dr. Dieter Rombach, Prof. Dr. Lionel Briand.

To overcome these limitations, this research proposes an enhancement of the method Optimised Set Reduction (OSR) to solve estimation problems. The goal is to provide an estimation method that overcomes the main drawbacks of other methods and that produces models with an accuracy comparable to other commonly used models. The OSR method is originally defined to solve classification problems and is based on machine learning and robust statistics (Briand et al., 1992). It exploits the rigor of statistical modelling and still generates easy-to-interpret, rule-based models. OSR determines which subset of a given database provides the best characterisation of an object to be assessed. The algorithm automatically generates a collection of logical expressions that characterise trends observable in the data set. The main distinction to other modelling methods is that OSR builds an individual model for each object to be predicted, rather than one general model based on the underlying data set. This research extends the OSR algorithm (called OSR^+) to solve cost estimation problems preserving those characteristics. To evaluate the benefits of OSR^+, this thesis describes a subjective assessment and a comprehensive empirical accuracy evaluation of commonly used estimation methods (Wieczorek, 2001).

The accuracy evaluation of OSR^+ aims for external validity of the results. Thus, the conditions under which the validation is performed are essential. Many different studies comparing software cost estimation methods have been performed in the past. But despite intense research only few conclusions can be drawn from existing results. This is because many studies were restricted to small and medium-sized data sets. Studies tend to be incomplete considering only small subsets of estimation methods for their evaluation. Only few studies were replicated and even if the same data set was used across different studies, the results were not always comparable because of different experimental designs. Studies were difficult to replicate, because they were not reported in a form that allows for the comparison of results. As in any other experimental field, replication is key to establishing the validity and the identification of consistent trends of results.

To overcome the drawbacks of many previous studies, OSR^+ was consistently compared with a cross-section of common cost modelling methods using two large software cost data sets from different application domains. The estimation methods are selected according to several criteria, such as utility in software engineering, suitable input requirements, or automation. Beside OSR^+, the methods that fulfilled the criteria were Ordinary least-squares Regression, robust Regression, stepwise Analysis of Variance, Regression Trees, and Analogy-based estimation.

The two data sets used were the European Space Agency (ESA[1]) and the Laturi database. Both are the result of a rigorous data quality assurance process and both are administered by professional institutions, an International Business School (INSEAD[2]) and the Software Technology Transfer Finland (STTF[3]), respectively. The ESA data are mainly from space and military applications, whereas the Laturi data consist of business application. Both data sets are fairly large, multi-organisational and from different European countries. This makes these databases very suitable to assess cost estimation methods both within and across companies.

The results from the comparative evaluation suggest that OSR^+ outperforms other non-parametric methods (Regression Trees, Analogy) and produces results comparable to the ones generated by parametric methods (Analysis of Variance, robust and least-squares Regression). OSR^+ is more robust to outlying predictions than all other methods considered. In the context of missing data, OSR^+ optimally utilises the available data and performs better than comparable methods. Thus, in situations where interacting with domain experts is crucial for the interpretation of a cost model, OSR^+ may be considered as a good and better alternative to parametric and non-parametric estimation methods, respectively.

Local, company-specific data are widely believed to provide a better basis for accurate estimates. This is because they allow for a number of advantages, such as a better control of the data collection process, or the possibility to define tailored cost drivers. However, practitioners are often faced with the lack of explicit data collected from past projects, because data collection is an expensive, time-consuming investment for individual companies. On the other hand, multi-organisational databases provide an opportunity for fast data accumulation. Unfortunately, collecting consistent data may turn out to be difficult and trends may differ significantly across organisations. Therefore, this thesis also trades off the potential advantage and drawbacks of using local data as compared to multi-company databases.

The results suggest that to really benefit from collecting company-specific cost data, one should focus on collecting the important factors in an organisation through a tailored measurement program. In case few local data are available, common data repositories across companies within homogenous application domains could be considered as a beneficial alternative.

In general, the appropriate selection of a method is context dependent and other criteria beside the predictive accuracy may justify the usage of a particular method. Criteria such as complexity, interpretability, and practical issues are rarely evaluated in previous studies. Therefore, this work also subjectively evaluates OSR^+ and the considered methods with respect to these criteria. A subjective evaluation framework was defined and applied using criteria relevant to make decisions regarding the adoption of an estimation method.

This research has highlighted and addressed three primary issues eluding previous investigations in cost estimation. Firstly, the limitations of currently used data-driven estimation methods. Secondly, the difficulty to appropriately select cost estimation methods. Thirdly, the benefits and drawbacks for cost estimation from collecting data from multiple organisations. As a potential solution for the first issue, the OSR method was enhanced with alternative algorithms (called OSR^+) suited for predicting continuous project attributes, such as development effort. This results in a generalisation of OSR to solve classification and regression problems. Addressing the second and the third issue, a comprehensive and repeatable approach to evaluate OSR^+ and other commonly used cost estimation methods has been provided. This comparison includes a subjective assessment as well as an objective empirical evaluation.

Notes

1. ESA Bulletin. The ESA Initiative for Software Productivity Benchmarking and Effort Estimation. No 87, August 1996; also available at http://www.esapub.esrin.esa.it/bulletin/bullet87/greves87.htm
2. INSEAD. http://www.insead.fr
3. STTF. http://www.sttf.fi/index.html

References

Briand, L.C., Basili, V., and Thomas, W., 1992. A pattern recognition approach for software engineering data analysis. *IEEE Transactions on Software Engineering*, 18(11): 932–942.

Briand, L.C., and Wieczorek, I. Software Resource Estimation. To be published in 2002: *Encyclopaedia of Software Engineering*; a previous version is published as ISERN-Technical Report, ISERN-00-05.

Wieczorek I. 2001 Improved Software Cost Estimation. A Robust and Interpretable Modelling Method and a Comprehensive Empirical Investigation. Ph.D. Theses in Experimental Software Engineering, vol. 7. Fraunhofer IRB Verlag. ISBN-3-8167-6033-3.

Isabella Wieczorek received the degree Diplom-Informatikerin (M.S.) in computer science 1994 from the University of Koblenz, Germany. Currently, she is a competence manager at the Fraunhofer Institute for Experimental Software Engineering (IESE) in the Quality and Process Engineering department. She has been working for several years in a diversity of industrial and research projects in the area of software cost and quality measurement. Her main interests are software measurement, empirical model building, and the application, evaluation, and development of quantitative methods to improve software project performance. She received the 'Best Dissertation Award 2000/2001' from the University of Kaiserslautern.