Isabella Wieczorek: **Improved Software Cost Estimation - A Robust and Interpretable Modeling Method and a Comprehensive Investigation**

Promotion: Universität Kaiserslautern
Erstgutachter: Prof. Dr. H. Dieter Rombach (Universität Kaiserslautern)
Zweitgutachter: Dr. Lionel Briand (Carleton University, Kanada)
Datum der Prüfung: 11.06.2001
Veröffentlichung: Stuttgart: Fraunhofer IRB Verlag, 2001 (PhD Theses in Experimental Software Engineering 7). Zugl.: Kaiserslautern, Univ., Diss., 2001; ISBN 3-8167-6033-3

Kurzfassung:

Delivering software on time, within budget, and to an agreed level of quality is crucial for many organizations. To ensure their business reputation and competitiveness organizations have recognized the importance of software cost estimation. However, it is not well understood in practice how to obtain and interpret software cost predictions. This is because the estimation of cost for software projects leads to many practical measurement and modeling difficulties.

These difficulties are not satisfactory addressed by common estimation methods. Classical statistical methods have limited model building capabilities regarding software engineering data. Other methods produce models with low interpretability and limitations in capturing uncertainty. This restricts the acceptability and usefulness of estimation models.

To overcome these limitations, there is a need to derive methods that make less assumptions about the underlying data sets and whose models are easy to interpret. This thesis proposes an enhancement of the method Optimized Set Reduction (OSR) to solve estimation problems. This method is based on machine learning and robust statistics. OSR exploits the rigor of statistical modeling and still generates easy-to-interpret, rule-based models.

To evaluate the benefits of OSR, this thesis describes a subjective assessment and a comprehensive accuracy evaluation. OSR and commonly used cost estimation methods are applied to large software cost data sets from different application domains. The data allow for a comparison of estimation methods across organizations, as well as within one company.

Local, company-specific data are widely believed to provide a better basis for accurate estimates. On the other hand, multi-organizational databases provide an opportunity for fast data accumulation and shared information benefits. Therefore, this thesis also trades off the potential advantage and drawbacks of using local data as compared to multi-organizational databases.

The results from the comparative evaluation suggest that OSR performs relatively accurate and is more robust to outlying predictions than other methods considered. In the context of missing data, OSR optimally utilizes the available data and performs better than comparable methods.

The results investigating the feasibility of multi-organizational data sets suggest that to really benefit from collecting company-specific cost data, one should focus on collecting the important factors in an organization 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.

Beside prediction accuracy, a broader range of issues should be considered when selecting an estimation method. OSR may be favored as it has the potential in providing easy-to-interpret predictive models. These allow for a better understanding of the factors impacting software costs and the uncertainties associated with software projects.