Title of DT:	Effective Parametric Model for System Engineering Project Estimation
Student:	Ho Le Thi Kim Nhung
Supervisor:	Assoc. Prof. Ing. Zdenka Prokopová, CSc.
Author of assessment:	doc. Ing. Petr Čermák, Ph.D.

Relevance of the chosen dissertation topic

Relevance of the selected topic - a comprehensive approach to improving estimation accuracy and minimizing project risks in the early stages of software development is very high. The topic is very actual and the student makes a significant contribution to improving estimation accuracy methods to minimize the risks associated with project development.

Definition and fulfilment of the dissertation objectives Aims/objectives:

The aim is to extend the accuracy of existing estimates and minimise project risks in the early stages of software development. The objectives of the dissertation are given in subsection 1.3. In my opinion, the student fulfill the stated objectives, which consist in a new method of optimizing correction factors and its evaluation in 4 namely Ex1, ..., Ex4 experiments in chapter 5 and 6.

Achieved results of the dissertation and evaluation of the work in terms of contribution of new knowledge, contribution to practice or development of the scientific field

The new contribution is the extension of the ExOCF method and procedure for better risk estimation of correction factors (technical and environmental). There is also interesting comparison OCP and UCF methods combined with machine learning methods. The proposed methods and procedures was subsequently verified using different error indicators (SSE, PRED, MAE, MdMRE, MBRE, MIBRE and RMSE). It was shown that the proposed method better optimizes, i.e., minimizes the risk of estimating correction factors.

In practical use, the application is offered for better estimation of complexity and size of software projects in engineering practice in the workflow/process of software design and implementation.

Since the student has to deal with two parts of model identification - structure identification and parametric identification, he is close to identification using TSK FNN, I have a question in the assessment "Defense questions".

Evaluation of the formal aspects of the thesis, including characteristics of the selection and use of sources

The dissertation is written in good English, the pictures are of good quality. Adequate selection of study sources. Variables should be written in italics.

Fulfilling the conditions of independent creative scientific work and publishing

The core of this dissertation is supported by publications, namely the number of :
1 publication in journal as lead author;

2 publications in journal as next author (less than fourth in order);
4 publication in proceedings as lead author;
9 publications in proceedings 1 as next author (less than fourth in order);

The core of the thesis is very well supported by publications.

Defence questions

Would it be beneficial to use the TSK fuzzy neural network to create models related to the subject of the dissertation?
Which methods and algorithms do you implement yourself?
Which toolbox and library do you use as part of the implementation in your experiments?

Final decision
Ho Le Thi Kim Nhung systematically deals with estimation methods complexity and size of software projects. The dissertation demonstrates the ability of scientific work and experience in the field.

Given that
1. relevance of the topic;
2. the quality of the dissertation in terms of content and formality;
3. Publication activity that supports the core of the dissertation;
i recommend Miss Ho Le Thi Kim Nhung's dissertation for defense. I further recommend that upon successful defense, the degree of Ph.D. in Engineering Informatics be awarded.

Opava, 11.11.2022

doc. Ing. Petr Čermák, Ph.D.

Tomas Bata University in Zlín
Faculty of Applied Informatics

Review of the dissertation

Author: MSc. Ho Le Thi Kim Nhung
Title: Effective Parametric Model for System Engineering Project Estimation
Supervisor: Assoc prof. Ing. Zdenka Prokopová, CSc.
Consulting supervisor: Assoc prof. Ing. Radek Šilhavý, Ph.D.

Topicality, content and the structure of the qualifying document

The PhD student has selected the topic of current importance. The title is related to the software development effort estimation and is very important in practice for managers who can reasonably plan the scheduling and budgeting. Nevertheless, the submitted thesis has an unusual format. The goals of the thesis are missing entirely. Therefore, it is very complicated to evaluate the work. It is impossible to evaluate what is missing, whether the aims were achieved or even achieved successfully.
The thesis contains 80 pages of text, figures and tables (tables are on 20 pages out of 80) divided into 9 chapters: Introduction, Theoretical framework, Current state of the issues dealt with, The proposed methods, Research Methodology, Results and Discussion, Threat of validity, Contributions of the thesis to science and practice and Conclusions.
The Introduction section is divided into 4 subchapters: Motivation, Problem statement, Research contribution and Organization of the thesis. In the text of the Introduction, there is the fusion of what is essential, what has been achieved, and how the student contributed to the science. In my opinion, chapter 1.3. is suitable more for the discussion or conclusion at the end of the thesis, not at the beginning. Even though it is mentioned what has been achieved, there is no goal, and one does not know what the aim to be achieved was and whether it was successful or not. Most of the thesis is written in "we" form, which does not clearly highlight what has been achieved by the student herself.

Since the goals are missing, the structure of the thesis does not fulfil the requirements for the dissertation thesis.

Technical quality of the thesis

The dissertation is written in English, and the used language is on a standard level. The images are at a standard printing level. The appropriateness of the selection and amount of literature used is sufficient and in line with the requirements of the dissertation. A list of abbreviations could be ordered in alphabetical form.

Selected methods of the processing

The topic is very complex, and the data from projects are usually missing. The student, hopefully (since again written in "we" form), has proposed 4 methods to improve the accuracy. The student

has suggested 4 experiments on 4 datasets available for this kind of task. But it is not clearly stated why these experiments were suggested in this form and what they should prove. For instance, the experiment 3 is even hard to follow; it contains full of abbreviations in tables 5.6 and 5.7 related to used machine learning techniques but for an unfamiliar reader is unclear.
Although the thesis contains a lot of tables of results, the summary of what has been achieved and the discussion of the results are missing. The comparisons with other researchers are missing too.
I appreciate the section Threat of validity. But I consider the Conclusions section insufficient.

Aiming of the thesis and new knowledge which brings

Based on the chapter – Contribution to the science and practice – it can be stated that the thesis proposed new ensemble techniques which have improved the estimation accuracy for the effort estimation of a project. I appreciate such an approach as a contribution to the science since it improves the accuracy of available datasets. And this technique can be helpful for the manager to estimate the necessary budget and schedule... which is crucial in the negotiation between a customer requiring the software and a supplier of the software.

Remarks and questions

Remarks:

1. It is hard for an unfamiliar reader to follow the text full of abbreviations even, especially if they are close to each other. I would recommend repeating the full version behind the "abbreviation" from time to time.
2. It is very impractical to look for defined hypotheses in previous sections since the result section contains, for instance, only – "we accept the alternative hypothesis H1". I would recommend repeating what the hypothesis was and whether it was or not accepted. And maybe more comments and discussion.

Questions:

1. Since most of the text of the thesis is written in "we" form, please, specify what your contribution was.
2. It is unclear why grid search was used to optimize the configuration settings for those particular methods in experiment 3. Please, justify.
3. Have you compared your results/approaches with other researchers? If not used by other researchers, have you tried on their "data"? It is essential not only to compare different versions of your proposed technique but also to fit it into the context of the current state-of-the-art techniques.

Review conclusion

I am convinced that the topic is significant and worth doing research on. MSc. Ho Le Thi Kim Nhung does not prove standard ability of creative and inventive scientific work since planning, designing and risen of the research questions and goals are important. Even though some results were achieved and I am convinced that the estimation accuracies were improved, the dissertation needs major revision. The goals, and a discussion of whether the goals were fulfilled or not, must be implemented together with the comparison with other state of the art techniques in the field.

Because of the given reasons,

I do not recommend the candidate to be awarded with the Ph.D. degree.

In Zlín, 4th November 2022

Assoc. prof. Ing. Zuzana Komínková Oplatková, Ph.D.
m.p.

Department of informatics and artificial intelligence
Faculty of the Applied Informatics
Tomas Bata University in Zlín
Nad Stráněmi 4511
760 05 Zlín

Review of the Ph.D. thesis
„Effective Parametric Model for System Engineering Project Estimation"
by Ho Le Thi Kim Nhung

Doctoral student Ho Le Thi Kim Nhung's dissertation deals with estimating the complexity of software project solutions using Use Case Points (UCP), complementing the concurrent dissertation of doctoral student Vo Van Hai, focusing on function point analysis.

Common to both works is the diversity of software projects due to their different purposes, they can involve relatively simple operations in information systems (adding, updating and deleting), and on the other hand systems for demanding scientific and technical calculations (e.g. finite element method and solving nonlinear differential equations), therefore finding a general method for evaluating software complexity is a non-trivial problem. As the work of PhD student Ho Le Thi Kim Nhung is oriented towards the early stages of software development, her goal is to estimate the functions it will provide.

These estimates have their economic significance because underestimating the complexity of development can cause significant cost increases due to failure to meet software delivery deadlines, just as unnecessarily high estimates lead to a waste of staff time and capacity that could be used for other tasks.

As the idea of the project's complexity is burdened with a considerable degree of uncertainty at an early stage of its development, the issue under investigation is very challenging and provides an opportunity to find new approaches.

The ability to use expert knowledge is limited by the difficulty of obtaining relevant data. Moreover, the information can only be described verbally and the problem of transforming it into a form that would allow machine processing arises.

As the author points out, a mechanical calculation expressed as a ratio of work to scope (productivity factor) from completed projects is misleading because, in general, each project is unique.

The aim of the work was to identify the software complexity factors that significantly affect the accuracy of the estimates and to propose a new formula for the calculation of the correction factors. In addition, to propose a suitable model or models for determining the size of the software.

It can be concluded that the focus of the thesis is topical, the objectives of the thesis are challenging with a clearly formulated own contribution, and the thesis is **dissertable** if they are met.

In Chapter 2, the author discusses in detail the UCP method, the diagram representation, the parameters (here called *actors*), their weights and scalarizing aggregations into a complex use case, then added 13 correction factors, and finally the method of determining the software complexity, expressed in person-hours. The next section then followed up by expanding on statistical evaluations and machine learning techniques such as a multilayer neural network with backpropagation algorithm adjusting synapse weights, decision trees, and other methods.

Chapter 3 is a survey of works in the world literature, describing both algorithmic and non-algorithmic estimation methods and models of statistical evaluation and machine learning. It also refers to systems freely available from websites and introduces their user environment and the way they work.

In chapter 4, the author proceeds to propose her own method OCF (Optimization Correction Factors), which is an improvement of the UCP method and should contribute to the refinement of the estimates. Her contribution is a new relationship for the calculation of the correction factor (hence the name of the method) and a multiple linear regression following it in order to minimize the prediction error. This section presents a number of mathematical equations to the LASSO (Least Absolute Shrinkage and Selection Operator) regression to determine the correction coefficient. With the interpretation of multiple linear regression, the author moves on to matrix equations, but the readability is reduced by the fact that the notation of matrices and "ordinary" variables in the typography is not distinguished by the style.

In subsection 4.3, the stacking ensemble model of correction factor optimization is introduced, where the name is somewhat confusing because stacking does not mean a data structure with a LIFO approach.

The author evaluates the quality of the regression model using coefficients of determination R^2 ("R squared").

The remainder of Chapter 5 is devoted to the experiments and the results obtained, illustrated by graphs and tables.

A detailed evaluation of the results and a discussion of the improvements over competing methods is provided in Chapter 6 and in verbal form in Chapters 7 and 8.

In the conclusion, the author suggests possible further research directions that may inspire other PhD students.

The work yields new results and the stated objectives have been met.

In terms of content and graphics, the work is at a very good level, as is the language level of the written text in English.

Formal comments:

- The author does not follow the typographic rules of writing symbols at all, she writes them all in a normal style.

Questions to the dissertant:

1. In several cases, you report that your OCP method improves the estimate by a certain number of percentages (e.g., 53.6 % on page 28) compared to the UCP method. But how is this value determined when estimates are applied to the early stages of software development? Is the calculation applied to historical cases of early stage projects and then compared to the reality of completed projects?
2. The problem defined by equation (4.1) on page 25 is nonlinear. What method was used to solve it?

Conclusion:

PhD student Ho Le Thi Kim Nhung's dissertation demonstrated a broad overview of optimization techniques and their creative application to estimating the complexity of software projects, including the design of original approaches.

The PhD student has applied the results in 16 publications in international forums - 3 in journals with impact factor and 13 conference papers - with citations. The results can be built upon in further research by the author or her followers. Therefore,

I recommend

Ho Le Thi Kim Nhung's Ph.D. thesis to be accepted by the Committee to be presented and defended in the Engineering Informatics study branch

Brno, October 31, 2022

Prof. RNDr. Ing. Miloš Šeda, Ph.D.
Institute of Automation and Computer Science
Faculty of Mechanical Engineering
Brno University of Technology