NIKOLAOS A. MITTAS

CURRICULUM VITAE

KAVALA 2012

PERSONAL INFORMATION

Name / Surname	**:**	Nikolaos Mittas
Father's name	**:**	Athanasios
Date / Place of birth	**:**	15 May 1981, Kavala, Greece
Address	**:**	Anagnostaki 13, Kavala
Telephone	**:**	2510-441732 (home), 2510-225871 (office)
Mobile Phone	**:**	6932879555
e-mail	**:**	nmittas@csd.auth.gr

STUDIES

- **2003: Bachelor** degree in Mathematics, University of Crete, Greece.
- **2005: Msc** in Information Systems, Department of Informatics, Aristotle University of Thessaloniki (AUTh), Greece.
- **2009: PhD** diploma thesis in Statistics (Statistical and Computational Methods for Development, Improvement and Comparison of Software Cost Estimation Models), Department of Informatics, Aristotle University of Thessaloniki (AUTh), Greece.
- **2011: Postdoctoral** research in Project Management-Software Cost Estimation, Department of Electrical Engineering and Information Technologies, Cyprus University of Technology.

ACADEMIC EXPERIENCE

- Visiting Assistant Professor, Technological Educational Institute (TEI) of Kavala. (2004-today).
- Visiting Lecturer, Department of Informatics, Aristotle University of Thessaloniki (AUTh), Greece (2009- today).

TEACHING EXPERIENCE

- **Research Methodology and Data Analysis** (Post-graduate course, Department of Informatics, Aristotle University of Thessaloniki).
- **Advanced Mathematics (Calculus I and II)** (Department of Petroleum Technology and Natural Gas, Technological Educational Institute of Kavala).

- **Computational Mathematics (Numerical Analysis, Fourier Series, Fourier and Laplace Transformation)** (Department of Electrical Engineering, Technological Educational Institute of Kavala).
- **Probabilities and Statistics** (Department of Electrical Engineering, Technological Educational Institute of Kavala).
- **Introduction to Corporate Statistics** (Department of Business Administration, Technological Educational Institute of Kavala).
- **Statistical Data Analysis** (Department of Accounting, Technological Educational Institute of Kavala).
- **Information Technology** (General Department of Physical Sciences, Technological Educational Institute of Kavala).

FUNDED PROJECTS

- Participation in 13 funded R&D Greek or European projects. Indicatively:
  - Software quality observatory for open source software SQO-OSS. (IST/EU).
  - Free/libre/open source software metrics and benchmarking study. European Commission - Information Society and Media Directorate General.

PUBLICATIONS

PhD Dissertation

Mittas N. (2009). Statistical and Computational Methods for Development, Improvement and Comparison of Software Cost Estimation Models. *Department of Mathematics, Aristotle University (In Greek).*

Journal Publications

J1. **MITTAS** N., ATHANASIADES M., ANGELIS L. (2008). Improving Analogy – Based Software Cost Estimation by a Resampling Method. *Information and Software Technology (Elsevier), 50, 221–230.*

J2. **MITTAS** N., ANGELIS L. (2008). Comparing Cost Prediction Models by Resampling Techniques. *Journal of Systems and Software (Elsevier). Special Issue on Software Process and Product Measurement, 81, 616–632.*

J3. **MITTAS** N., ANGELIS L. (2010). Visual Comparison of Software Cost Estimation Models by Regression Error Characteristic Analysis. *Journal of Systems and Software (Elsevier), 83, 621-637.*

J4. **MITTAS** N., ANGELIS L. (2010). LSEbA: Least Squares Regression and Estimation by Analogy in a Semi-Parametric Model for Software Cost Estimation. *Empirical Software Engineering (Springer), 15 (5), 523-555*.

J5. **MITTAS** N., ANGELIS L. (2011). A Permutation Test based on Regression Error Characteristic Curves for Software Cost Estimation Models. *Empirical Software Engineering (Springer)*. ***Special Issue*** *on Repeatable Results in Effort Estimation, 17* (*1-2*), 34-61.

J6. **MITTAS N.** (2012). Evaluating the Performances of Software Cost Estimation Models through Prediction Intervals. *Journal of Engineering Science and Technology Review* (accepted for publication).

International Conferences

C1. BIBI M., **MITTAS** N., ANGELIS L., STAMELOS I., MENDES E. (2007). Comparing Cross- vs. Within-Company Effort Estimation Models Using Interval Estimates. *IWSM-Mensura 2007 (International Conference on Software Process and Product Measurement), Palma de Mallorca, Spain, 5-8 November 2007.*

C2. **MITTAS** N., ANGELIS L. (2008). Partial Regression Error Characteristic Curves for the Comparison of Software Cost Prediction Models, *Workshop on Artificial Intelligence, Techniques in Software Engineering (AISEW 2008), July 2008, Patras, Greece.*

C3. **MITTAS** N., ANGELIS L. (2008). Comparing Software Cost Prediction Models by a Visualization Tool. *34th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA), September 2008, Parma, Italy.*

C4. **MITTAS** N., ANGELIS L. (2008). Combining Regression and Estimation by Analogy in a Semi-parametric Model for Software Cost Estimation. *International Symposium on Empirical Software Engineering and Measurement ESEM'08, October 9–10, 2008, Kaiserslautern, Germany.*

C5. **MITTAS** N., ANGELIS L. (2009). Bootstrap Confidence Intervals for Regression Error Characteristic Curves Evaluating the Prediction Error of Software Cost Estimation Models. *2nd Artificial Intelligence Techniques in Software Engineering Workshop (AISEW2009), at the 5th IFIP Conference on Artificial Intelligence Applications and Innovations, April, Thessaloniki, Greece. Proceedings online, http://ceur-ws.org/Vol-475, pp. 221-230.*

C6. **MITTAS** N., ANGELIS L. (2009). Bootstrap Prediction Intervals for a Semi-Parametric Software Cost Estimation Model. *35th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2009). August 2009, Patras, Greece, Proceedings published by IEEE, pp. 293-299.*

C7. **MITTAS** N., ARGYROPOYLOY V. ANGELIS L. (2010). Modeling the Relationship between Software Effort and Size Using Deming Regression. *6th International Conference on Predictive Models in Software Engineering. September 12-13, 2010, Timisoara, Romania.*

C8. KOSTI M., **MITTAS** N., ANGELIS L. (2010). DD-EbA: An algorithm for determining the number of neighbors in cost estimation by analogy using distance distributions. *Workshop on Artificial Intelligence, Techniques in Software Engineering (AISEW 2010), September 2010, Ayia Napa, Cyprus.*

C9. HAMDAN K., ANGELIS L. **MITTAS** N. (2010). Estimating learning by analogy: A case study in the UAE univerisity. *International Conference of Education, Research and Innovation (ICERI 2010). November 15-17, 2010, Madrid, Spain.*

C10. **MITTAS** N. (2011) Evaluating the Performances of Software Cost Estimation Models through Prediction Intervals. *International Conference on Econophysics, June 2-3, 2011, Kavala, Greece.*

Book Chapters

B1. ANGELIS L., SENTAS P., **MITTAS** N., CHATZIPETROU P. (2010). Methods for Statistical and Visual Comparison of Imputation Methods for Missing Data in Software Cost Estimation. *In Modern Software Engineering Concepts and Practices: Advanced Approaches, Idea Group Inc. Publishing.*

Greek Conferences/Publications

G1. **MITTAS** N., L. ANGELIS (2006). Confidence intervals and hypothesis tests in the comparison of software cost estimation methods. *Proceedings of the 19th Panhellenic Conference in Statistics, Kastoria (In Greek).*

G2. DIMOKAS N., **MITTAS** N., NANOPOULOS A., ANGELIS L. (2008). A Prototype System for Educational Data Warehousing and Mining. *12th Pan-Hellenic Conference on Informatics PCI 2008 August 2008, Samos, Proceedings published by IEEE.*

G3. **MITTAS** N., L. ANGELIS (2009). Graphical comparison of software cost estimation models with regression error characteristic analysis. *22nd Panhellenic Conference in Statistics, Chania, Greece (In Greek).*

G4. **MITTAS** N., FLOROU G., POLYCHRONIDOU P. (2009). Examination of Duration of Studies for the Accounting Department of the Technological Institution of Kavala through Survival Analysis.Διάρκεια Φοίτησης στο Τμήμα Λογιστικής ΤΕΙ Καβάλας μέσω της Ανάλυσης Επιβίωσης. *22nd Panhellenic Conference in Statistics, Chania, Greece (In Greek).*

G5. POLYCHRONIDOU P., **MITTAS** N., FLOROU G. (2009). Factors that Affect the Duration of Studies of the Accounting Department of the Technological Institution of Kavala. *5th Pan-Hellenic Data Analysis Conference with International Participitation, September 2009, Rethimno*.

G6. SALTAS B., KONGUETSOF A., KALAMBAKAS A., POLYCHRONIDOU P., **MITTAS** N., TSIANTOS B. (2010). Evaluation of the Skills for the 1st year Students in Mathematics for the Department of Petroleum Technology and Natural Gas of Technological Educational Institute of Kavala, *23rd Panhellenic Conference in Statistics, Veria, Greece (In Greek)*.

Impact Factor of Journals:

Journal	**IF (2009)**	**5-Year IF**
Journal of Systems and Software (Elsevier)	**1.227**	**1.290**
Information and Software Technology (Elsevier)	**1.507**	**1.426**
Empirical Software Engineering (Springer)	**1.776**	**1.783**

Acceptance ratio of Conferences (Indicatively):

Conference	**Ποσοστό**
2nd ACM-IEEE international symposium on Empirical software engineering and measurement (ESEM 2008)	**28%**
35th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2009)	**39%**
12th Panhellenic Conference on Informatics 2008	**49%**
34th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2008)	**56%**

REVIEWER IN JOURNALS

- Journal of Systems and Software (Elsevier)
- IEEE Transactions on Software Engineering (IEEE).

ACADEMIC PROGRAM COMMITTEE

- 2nd IEEE International Conference on Computer and Communication Technology (ICCCT - 2011), 15-17 September 2011 in Allahabad, India.

- Academic Program Committee on 12th EANN / 7th AIAI Joint Conferences, 15 - 18 September 2011, Corfu, Greece Engineering Applications of Neural Networks / Artificial Intelligence Applications and Innovations.
- Academic Program Committee on CISE 2012: 2nd International Workshop on Computational Intelligence in Software Engineering September 27-30, 2012 Halkidiki, Greece.

CITATIONS FROM OTHER RESEARCHERS
(Last Update – December 2011)

Citations for J1

J1-1. (2008) LI Q., WANG Q., YANG Y., LI M. Reducing biases in individual software effort estimations: a combining approach. *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement, 223-232.*

J1-2. (2008) KAMEI Y., KEUNG J., MONDEN A., MATSUMATO K. An over-sampling method for analogy-based software effort estimation. *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement, 312-314*

J1-3. (2008) LI Y.F., XIE M., GOH T.N. Optimization of feature weights and number of neighbors for analogy based cost estimation in software project management. *IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2008, art. no. 4738130, pp. 1542-1546*

J1-4. (2008) LI Y.F., XIE M., GOH T.N. A study of analogy based sampling for interval based cost estimation for software project management *Proceedings of the 4th IEEE International Conference on Management of Innovation and Technology, ICMIT, art. no. 4654377, pp. 281-286*

J1-5. (2008) LI Y.F., XIE M., GOH T.N. A bayesian inference approach for probabilistic analogy based software maintenance effort estimation *Proceedings of the 14th IEEE Pacific Rim International Symposium on Dependable Computing, PRDC 2008, art. no. 4725294, pp. 176-183*

J1-6. (2009) AZZEH M., NEAGU D., COWLING P. Software effort estimation based on weighted fuzzy grey relational analysis. *Proceedings (ACM) of the 5th International Conference on Predictor Models in Software Engineering table of contents Vancouver, British Columbia, Canada.*

J1-7. (2009) AZZEH M., NEAGU D., COWLING P. Fuzzy grey relational analysis for software effort estimation. *Empirical Software Engineering, 15(1), 60-90.*

J1-8. (2009) ZHANG J.G. Validity verifying method of software project management. *Proceedings of the International Conference on Management and Service Science, MASS 2009.*

J1-9. (2010) BO J., XIAOJUN Z., LIFENG X. Back propagation neural network based product cost estimation at an early design stage of passenger vehicles. *International Journal of Industrial and Systems Engineering, 5 (2), 190 – 211*.

J1-10. (2010) LI Y.F. Improvement and implementation of analogy based method for software project cost estimation. *Doctoral Thesis. Department of Industrial and Systems Engineering, Wuhan National University of Singapore.*

J1-11. (2011) WEN J., LI S., LIN Z, HU Z., HUANQ C. Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology*.

J1-12. (2012) NAGPAL G., UDDIN M, KAUR A. A Hybrid Technique using Grey Relational Analysis and Regression for Software Effort Estimation using Feature Selection. *International Journal of Software Computing and Engineering* (*IJSCE*) *1 (6), January 2012, pp. 20-27*.

Citations for J2

J2-1. (2009) KITCHENHAM B., MENDES E. Why comparative effort prediction studies may be invalid. *Proceedings of the 5th International Conference on Predictor Models in Software Engineering table of contents, Vancouver, British Columbia, Canada.*

J2-2. (2009) VIJAY J.F., MANOHARAN C. Initial Hybrid Method for Analyzing Software Estimation, Benchmarking and Risk Assessment Using Design of Software. *Journal of Computer Science 5 (10): 717-724.*

J2-3. (2010) ARSHID A., SALMAN Q., SYED SHAH M., JALIL A., MUHAMMAD TARIQPERVA., SARFARAZ A. Software Cost Estimation through Entity Relationship Model. *Journal of American Science, 6 (11), 47-51*.

J2-4. (2010) KHAN K. The evaluation of well-known effort estimation models based on predictive accuracy indicators. *Master Thesis. Number: MSE- 2010:03, School of Computing Blekinge Institute of Technology, Sweden.*

J2-5. (2010) PAHARIYAA J., RAVIA, V., CARRA M., VASUA M. Computational Intelligence Hybrids Applied to Software Cost Estimation. *International Journal of Computer Information Systems and Industrial Management Applications 2, 104-112*.

J2-6. (2012) FALESSI D., CANTONE G., CANFORA G. Empirical Principles and an Industrial Case Study in Retrieving Equivalent Requirements via Natural Language Processing Techniques. IEEE Transactions on Software Engineering.

J2-7. (2012) NAGPAL G., UDDIN M, KAUR A. A Hybrid Technique using Grey Relational Analysis and Regression for Software Effort Estimation using Feature Selection. *International Journal of Software Computing and Engineering* (*IJSCE*) *1 (6), January 2012, pp. 20-27*.

Citations for J3

J3-1. (2010) FERNADEZ-DIEGO M., MARTINEZ-GOMEZ M. TORRABLA-MARTINEZ J.M. Sensitivity of results to different data quality meta-data criteria in the sample selection of projects from the ISBSG dataset. *Proceedings of the 6th International Conference on Predictive Models in Software Engineering. September 12-13, 2010, Timisoara, Romania.*

J3-2. (2010) PAHARIYAA J., RAVIA, V., CARRA M., VASUA M. Computational Intelligence Hybrids Applied to Software Cost Estimation. *International Journal of Computer Information Systems and Industrial Management Applications 2, 104-112*.

Citations for J4

J4-1. (2011) S. GUPTA, G. SIKKA, H. VERMA. Recent methods for software effort estimation by analogy. *ACM SIGSOFT Software Engineering Notes, 36 (4), 1-5.*

J4-2. (2011) WEN J., LI S., LIN Z, HU Z., HUANQ C. Systematic literature review of machine learning based software development effort estimation models. Information and Software Technology.

J4-3. (2012) NAGPAL G., UDDIN M, KAUR A. A Hybrid Technique using Grey Relational Analysis and Regression for Software Effort Estimation using Feature Selection. *International Journal of Software Computing and Engineering* (*IJSCE*) *1 (6), January 2012, pp. 20-27*.

Citations for J5

J5-1. (2012) MENZIES T., SHEPPERD M. Special issue on repeatable results in software engineering prediction. *Empirical Software Engineering (Springer)*. ***Special Issue*** *on Repeatable Results in Effort Estimation.*

Citations for C4

C4-1. (2009) SEO Y.-S., YOON K.-A., BAE, D.-H. Improving the accuracy of software effort estimation based on multiple least square regression models by estimation error-based data partitioning. *Proceedings - Asia-Pacific Software Engineering Conference, APSEC*: pp. *3-10*.

C4-2. (2009) ASSUNÇÃO S.C. Caracterização da prática da gestão de projectos de desenvolvimento de software - Perspectiva dos especialistas de prestadores de serviços. DISSERTAÇÃO DE MESTRADO EM TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO. UNIVERSIDADE DE TRÁS-OS-MONTES E ALTO DOURO.

C4-3. (2010) KHAN K. The evaluation of well-known effort estimation models based on predictive accuracy indicators. *Master Thesis. Number: MSE-2010:03, School of Computing Blekinge Institute of Technology, Sweden.*

Citations for C6

C6-1. (2010) KHAN K. The evaluation of well-known effort estimation models based on predictive accuracy indicators. *Master Thesis. Number: MSE-2010:03, School of Computing Blekinge Institute of Technology, Sweden.*

C6-2. (2011) KLAS M., TRENDOWICZ A., ISHIGAI Y., NAKAO H. Handling Estimation Uncertainty with Bootstrapping: Empirical Evaluation in the Context of Hybrid Prediction Methods. *Proceedings of the 5th International Symposium on Empirical Software Engineering and Measurement (ESEM 2011), Sept. 22-23, Banff, Canada.*

Citations for C7

C7-1. (2010) PAPATHEOCHAROUS E., ANDREOU A. Size-based software cost modelling with artificial neural networks and genetic algorithms. *Artificial Neural Networks – Application, InTech*, pp. 167-188, April 2011.

Citations for C8

C8-1. (2011) S. GUPTA, G. SIKKA, H. VERMA. Recent methods for software effort estimation by analogy. *ACM SIGSOFT Software Engineering Notes, 36 (4), 1-5.*

Citations for G2

G2-1. (2011) AREVALILLO-HERAEZ M., MORENO-CLARI P., CERVERON-LLEO V. Educational knowledge generation from administrative data. *Journal of Educational Technology Research and Development (Springer).*

G2-2. (2010) WARNARS S. Tata Kelola Database Perguruan Tinggi Yang Optimal Dengan Data Warehouse. *TELKOMNIKA Indonesian Journal of Electrical Engineering, 8 (1), pp. 25 – 34.*

G2-3. (2011) Y. Xuejian, L. Xueqing. A Multidimensional Data Analysis System Based on MDA for Educational Data Warehousing. *Proceedings of the 6th International Conference on Computer Science & Education (ICCSE 2011), August 3-5, SuperStar Virgo, Singapore, pp. 88-94.*

NIKOLAOS A. MITTAS

ABSTRACTS OF PUBLICATIONS

KAVALA 2011

PhD Dissertation

MITTAS N. (2009). Statistical and Computational Methods for Development, Improvement and Comparison of Software Cost Estimation Models. *Department of Mathematics, Aristotle University*.

The plethora of Software Cost Estimation models proposed in the literature reveals that the prediction of the cost for a new software project is a vital task affecting the well-balanced management of the development process. The overestimation of a project may lead to the canceling and loss of a contract, whereas the underestimation may affect the earnings of the development organization. Hence, there is an ongoing research in the SCE area attempting to build prediction models that provide accurate estimates of the cost.

The present dissertation deals with the introduction of statistical and computational methods for the comparison, improvement and development of Software Cost Estimation models. More specifically, the contribution of the dissertation focuses on the following subjects.

Chapter 3 deals with the comparison procedure of alternative prediction models. Since there are a lot of models that can be fitted to certain data, a crucial issue is the selection of the most efficient prediction model. Most often this selection is based on comparisons of various accuracy measures that are functions of the model's errors. However, the usual practice is to consider as the most accurate prediction model the one providing the best accuracy measure without testing if this superiority is in fact statistically significant. This policy can lead to unstable and erroneous conclusions since a small change in the data is able to turn over the best model selection. On the other hand, the accuracy measures used in practice are statistics with unknown probability distributions, making the testing of any hypothesis, by the traditional parametric methods, problematic. In this Chapter, the use of statistical simulation tools is proposed in order to test the significance of the difference between the accuracy of two prediction methods. The statistical simulation procedures involve permutation tests and bootstrap techniques for the construction of confidence intervals for the difference of measures. These techniques repeat the data analysis a large number of times on replicated datasets, all drawn by resampling from the original observed set of data. The resampling techniques can be used on their own in carrying out a hypothesis test without worrying about the distribution of the variables or they can also be utilized with the traditional procedures in order to reinforce their results.

Chapter 4 also deals with the comparison procedure of alternative prediction models, whereas the research interest focuses on the graphical investigation of the performances. More precisely, we introduce the Regression Error Characteristic analysis, a powerful visualization tool with interesting geometrical properties, in order to validate and compare different prediction models easily, by a simple inspection of a graph. The proposed formal framework covers different aspects of the estimation process such as the calibration of the prediction methodology, the assessment of the applicability of the estimation method to a specific dataset, the identification of factors affecting the error, the investigation of errors on certain ranges of the actual cost and the examination of the distribution of the cost for certain errors. The experimentation portrays the benefits and the significant information obtained by this analysis.

Chapter 5 studies a well-known question in Software Cost Estimation area that is whether there are differences in the predictive accuracy of various software cost estimation techniques when effort estimates are based on datasets with completed projects from a single company (within-company predictions) or from different companies (cross-company predictions). The question is examined on models that produce either point estimates accompanied by prediction intervals or interval estimates with a derived point estimate. Several known measures of prediction errors were used for the comparison of point estimates while for the comparison of the prediction intervals a new measure was defined - a generalization of the Hit-rate accuracy measure. This new measure was designed to take into account (a) whether the actual cost value fell into an interval, (b) the similarity of intervals under comparison and (c)

the width of the intervals. The experimentation reveals that the use of within-company data generally improved the results, especially regarding the intervals.

Chapter 6 deals with a well-known technique that is Estimation by Analogy. The popularity of the method is due to its straightforwardness and its intuitively appealing interpretation of the whole procedure which mimics the human instinctive decision-making by comparing with similar cases. However, in spite of the simplicity in application, the theoretical study of the method is quite complicated. In this Chapter, we exploit the relation of the method to the nearest neighbor non-parametric regression in order to suggest a resampling procedure, known as iterated bagging, for reducing the prediction error. The improving effect of iterated bagging is validated using both artificial and real datasets from the literature, obtaining very promising results.

Chapter 7 deals with the possibility of aggregating the parametric Least Squares regression and the non-parametric Estimation by Analogy in a semi-parametric model, trying to incorporate, in a systematic way, the linear and non-linear information obtained from the same dataset. Least Squares regression is called parametric, since it assumes that there is a functional form between a dependent and a set of independent variables, fully described by a finite set of parameters. However, a pre-selected parametric model for all independent variables is often a too strong and too restricted assumption, especially for datasets with categorical variables. Such a model often fails to fit unexpected effects of the attributes. On the other hand, a non-parametric approach, such as Estimation by Analogy, can offer a flexible procedure in explaining unknown and complicated relationships because it is based on the concept of similarity which can be calculated even for categorical data. However, the inclusion of all independent variables in the procedure of computing similarities is also a strong assumption which is able to mask a strong parametric relation, failing to take advantage of valuable information. The aforementioned considerations led us to believe that a combination of methods appears to be a more realistic approach for the SCE datasets which usually contain a portion of variables (for example the size) that is parametrically correlated with the cost variable and another portion that has a significant impact on cost, but in an undefined and non-linear form. Experimentation on representative datasets verifies the benefits of the proposed model in terms of accuracy, bias and spread of the predictions.

Finally, Chapter 8 concludes this dissertation and presents extensions and directions for future work.

Journal Publications

[J1] MITTAS N, ATHANASIADES M., ANGELIS L. (2008). Improving Analogy – Based Software Cost Estimation by a Resampling Method. ***Information and Software Technology (Elsevier), 50, 221–230.***

Estimation by analogy (EbA) is a well-known technique for software cost estimation. The popularity of the method is due to its straightforwardness and its intuitively appealing interpretation. However, in spite of the simplicity in application, the theoretical study of EbA is quite complicated. In this paper, we exploit the relation of EbA method to the nearest neighbor non-parametric regression in order to suggest a resampling procedure, known as iterated bagging, for reducing the prediction error. The improving effect of iterated bagging on EbA is validated using both artificial and real datasets from the literature, obtaining very promising results.

[J2] MITTAS N., ANGELIS L. (2008). Comparing Cost Prediction Models by Resampling Techniques. ***Journal of Systems and Software (Elsevier), Special Issue on Software Process and Product Measurement, 81, 616–632.***

The accurate software cost prediction is a research topic that has attracted much of the interest of the software engineering community during the latest decades. A large part of the research efforts involves the development of statistical models based on historical data. Since there are a lot of models that can be

fitted to certain data, a crucial issue is the selection of the most efficient prediction model. Most often this selection is based on comparisons of various accuracy measures that are functions of the model's relative errors. However, the usual practice is to consider as the most accurate prediction model the one providing the best accuracy measure without testing if this superiority is in fact statistically significant. This policy can lead to unstable and erroneous conclusions since a small change in the data is able to turn over the best model selection. On the other hand, the accuracy measures used in practice are statistics with unknown probability distributions, making the testing of any hypothesis, by the traditional parametric methods, problematic. In this paper, the use of statistical simulation tools is proposed in order to test the significance of the difference between the accuracy of two prediction methods: regression and estimation by analogy. The statistical simulation procedures involve permutation tests and bootstrap techniques for the construction of confidence intervals for the difference of measures. Four known datasets are used for experimentation in order to validate the results and make comparisons between the simulation methods and the traditional parametric and non-parametric procedures.

[J3] MITTAS N., ANGELIS L. (2010). Visual Comparison of Software Cost Estimation Models by Regression Error Characteristic Analysis. ***Journal of Systems and Software*** **(*Elsevier*), *83, 621-637*.**

The well-balanced management of a software project is a critical task accomplished at the early stages of the development process. Due to this requirement, a wide variety of prediction methods has been introduced in order to identify the best strategy for software cost estimation. The selection of the best technique is usually based on measures of error whereas in more recent studies researchers use formal statistical procedures. The former approach can lead to unstable and erroneous results due to the existence

of outlying points whereas the latter cannot be easily presented to non-experts and has to be carried out by an expert with statistical background. In this paper, we introduce the regression error characteristic (REC) analysis, a powerful visualization tool with interesting geometrical properties, in order to validate and compare different prediction models easily, by a simple inspection of a graph. Moreover, we propose a formal framework covering different aspects of the estimation process such as the calibration of the prediction methodology, the identification of factors that affect the error, the investigation of errors on certain ranges of the actual cost and the examination of the distribution of the cost for certain errors. Application of REC analysis to the ISBSG10 dataset for comparing estimation by analogy and linear regression illustrates the benefits and the significant information obtained.

[J4] MITTAS N., ANGELIS L. (2010). LSEbA: Least Squares Regression and Estimation by Analogy in a Semi-Parametric Model for Software Cost Estimation. ***Empirical Software Engineering (Springer)*****, (accepted for publication).**

The importance of Software Cost Estimation at the early stages of the development life cycle is clearly portrayed by the utilization of several models and methods, appeared so far in the literature. The researchers' interest has been focused on two well known techniques, namely the parametric Regression Analysis and the non-parametric Estimation by Analogy. Despite the several comparison studies, there seems to be a discrepancy in choosing the best prediction technique between them. In this paper, we introduce a semi-parametric technique, called LSEbA that achieves to combine the aforementioned methods retaining the advantages of both approaches. Furthermore, the proposed method is consistent with the mixed nature of Software Cost Estimation data and takes advantage of the whole pure information of the dataset even if there is a large amount of missing values. The paper analytically illustrates the process of building such a model and presents the experimentation on three representative datasets verifying the benefits of the proposed model in terms of accuracy, bias and spread. Comparisons of LSEbA with linear regression, estimation by analogy and a combination of them, based on the average

of their outcomes are made through accuracy metrics, statistical tests and a graphical tool, the Regression Error Characteristic curves.

[J5] MITTAS N., ANGELIS L. (2011). A Permutation Test based on Regression Error Characteristic Curves for Software Cost Estimation Models. Empirical Software Engineering (Springer). ***Special Issue on Repeatable Results in Effort Estimation, 17 (1-2), 34-61*****.**

Background Regression Error Characteristic (REC) curves provide a visualization tool, able to characterize graphically the prediction power of alternative predictive models. Due to the benefits of using such a visualization description of the whole distribution of error, REC analysis was recently introduced in software cost estimation to aid the decision of choosing the most appropriate cost estimation model during the management of a forthcoming project.
Aims Although significant information can be retrieved from a readable graph, REC curves are not able to assess whether the divergences between the alternative error functions can constitute evidence for a statistically significant difference.
Method In this paper, we propose a graphical procedure that utilizes (a) the process of repetitive permutations and (b) and the maximum vertical deviation between two comparative Regression Error Characteristic curves in order to conduct a hypothesis test for assessing the statistical significance of error functions.
Results In our case studies, the data used come from software projects and the models compared are cost prediction models. The results clearly showed that the proposed statistical test is necessary in order to assess the significance of the superiority of a predictionmodel, since it provides an objective criterion for the distances between the REC curves. Moreover, the procedure can be easily applied to any dataset where the objective is the prediction of a response variable of interest and the comparison of alternative prediction techniques in order to select the best strategy.

[J6] MITTAS N. (2012). Evaluating the Performances of Software Cost Estimation Models through Prediction Intervals. ***Journal of Engineering Science and Technology Review*** **(accepted for publication).**

The task of predicting accurately the cost required for the completion of a new software project is a challenging issue in the Software Cost Estimation area, since it is closely related with the activities of project management and the wise decision-making of organizations in order to bid, plan and budget a forthcoming system. However, the accurate prediction of the cost is often obtained with great uncertainty and for this reason there has been noted a lack of convergence in experimental studies. The main reason for the discrepancy can be derived from the inherent characteristic of prediction methodologies, since they produce point estimates without taking into account the risk covering the whole process. In this study, we propose a statistical framework, so as to focus on the construction of Prediction Intervals which provide an “optimistic” and a “pessimistic” guess for the true magnitude of the cost. The proposed framework that incorporates different accuracy indicators, formal hypothesis testing and graphical inspection of the predictive performance is applied on a dataset with real software projects.

International Conferences

[C1] BIBI M., MITTAS N., ANGELIS L., STAMELOS I., MENDES E. (2007). Comparing Cross-vs. Within-Company Effort Estimation Models Using Interval Estimates. ***IWSM-Mensura 2007 (International Conference on Software Process and Product Measurement), Palma de Mallorca, Spain, 5-8 November 2007*****.**

This paper investigates whether effort predictions for projects from a single company that were obtained using a cross-company (CC) training set can be as accurate as effort predictions obtained using a within-company (WC) training set. We employed five different cost estimation techniques, two providing point estimates (estimation by analogy and stepwise regression) and three providing predefined interval estimates (ordinal regression, classification and regression trees and Bayesian networks). For the development and evaluation of both cross and within company models ISBSG release 9 was utilized. Our results showed no significant differences between CC and WC-based predictions, for all the cost estimation techniques, after comparing the medians of the absolute errors. Other accuracy metrics were also considered, providing in general similar results.

[C2] MITTAS N., ANGELIS L. (2008). Partial Regression Error Characteristic Curves for the Comparison of Software Cost Prediction Models, *Workshop on Artificial Intelligence, Techniques in Software Engineering (AISEW 2008), July 2008, Patras, Greece*.

The task of predicting accurately the cost required for the completion of a new software development project is a critical issue in the Software Cost Estimation area. The introduction of a variety of models and methods for this purpose verifies the in-creased interest of the researchers, whereas an abundance of stud-ies has been carried out in order to select the "best" prediction technique. More recent studies make use of statistical comparisons in order to signify their results. However, statistical tests have to be carried out by an expert with mathematical background, whereas the interpretation of the results is not trivial. In this paper, we in-troduce the utilization of Partial Regression Characteristic curves, a visualization technique originated from machine learning and data mining and inspired by Receiver Operating Characteristic (ROC) analysis. The tool can be applied to any cost estimation situation in order to study the behavior of several comparative statistical or artificial intelligence methods on certain ranges of actual cost val-ues and decide which of the prediction methods gives the "best" results in the range under consideration.

[C3] MITTAS N., ANGELIS L. (2008). Comparing Software Cost Prediction Models by a Visualization Tool. *34th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA), September 2008, Parma, Italy*.

A crucial issue in the Software Cost Estimation area that has attracted the interest of software project managers is the selection of the best prediction method for estimating the cost of a project. Most of the prediction techniques estimate the cost from historical data. The selection of the best model is based on accuracy measures that are functions of the predictive error, whereas the significance of the differences can be evaluated through statistical procedures. However, statistical tests cannot be applied easily by non-experts while there are difficulties in the interpretation of their results. The purpose of this paper is to introduce the utilization of a visualization tool, the Regression Error Characteristic curves in order to compare different prediction models easily, by a simple inspection of a graph. Moreover, these curves are adjusted to accuracy measures appeared in Software Cost Estimation literature and the experimentation is based on two well-known datasets.

[C4] MITTAS N., ANGELIS L. (2008). Combining Regression and Estimation by Analogy in a Semi-parametric Model for Software Cost Estimation. *International Symposium on Empirical Software Engineering and Measurement ESEM'08, October 9–10, 2008, Kaiserslautern, Germany*.

Software Cost Estimation is the task of predicting the effort or productivity required to complete a software project. Two of the most known techniques appeared in literature so far are Regression Analysis and Estimation by Analogy. The results of the empirical studies show the lack of convergence in choosing

the best prediction technique between the parametric Regression Analysis and the non-parametric Estimation by Analogy models. In this paper, we introduce the use of a semi-parametric model that achieves to incorporate some parametric information into a non-parametric model combining in this way regression and analogy. Furthermore, we demonstrate the procedure of building such a model on two well-known datasets and we present the comparative results based on the predictive accuracy of the new technique using several accuracy measures. We also perform statistical tests on the residuals in order to assess the improvement in the predictions attained through the new semi-parametric model in comparison to the accuracy of Regression Analysis and Estimation by Analogy when applied separately. Our results show that the semi-parametric model provides more accurate predictions than each one of the parametric and non-parametric approaches.

[C5] MITTAS N., L. ANGELIS (2009). Bootstrap Confidence Intervals for Regression Error Characteristic Curves Evaluating the Prediction Error of Software Cost Estimation Models. ***2nd Artificial Intelligence Techniques in Software Engineering Workshop (AISEW2009), at the 5th IFIP Conference on Artificial Intelligence Applications and Innovations, April, Thessaloniki, Greece. Proceedings online, http://ceur-ws.org/Vol-475, pp. 221-230.***

The importance of Software Cost Estimation at the early stages of the development life cycle is clearly portrayed by the utilization of several algorithmic and artificial intelligence models and methods, appeared so far in the literature. Despite the several comparison studies, there seems to be a discrepancy in choosing the best prediction technique between them. Additionally, the large variation of accuracy measures used in the comparison procedure constitutes an inhibitory factor which complicates the decision-making. In this paper, we further extend the utilization of Regression Error Characteristic analysis, a powerful visualization tool with interesting geometrical properties in order to obtain Confidence Intervals for the entire distribution of error functions. As there are certain limitations due to the small-sized and heavily skewed datasets and error functions, we utilize a simulation technique, namely the bootstrap method in order to evaluate the standard error and bias of the accuracy measures, whereas bootstrap confidence intervals are constructed for the Regression Error Characteristic curves. The tool can be applied to any cost estimation situation in order to study the behavior of comparative statistical or artificial intelligence methods and test the significance of difference between models.

[C6] MITTAS N., ANGELIS L. (2009). Bootstrap Prediction Intervals for a Semi-Parametric Software Cost Estimation Model. ***35th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2009). August 2009, Patras, Greece, Proceedings published by IEEE, pp. 293-299.***

The vital task of accurate Software Cost Estimation predictions remains a challenging problem attracting the interest of researchers and practitioners. Although Least Squares (LS) regression and Estimation by Analogy (EbA) are two of the most widely applied methods, there seems to be a discrepancy in choosing the best prediction technique. In this paper, we further extend our previous work on the utilization of a semi-parametric model, called LSEbA that achieves to combine the abovementioned methods. More precisely, we present a method of constructing prediction intervals by the bootstrap resampling technique. The prediction intervals obtained for LSEbA are compared with those of LS and EbA separately, with the aid of a new methodology that takes into account not only the ability of comparative intervals to capture the actual cost, but also their similarity and their width.

[C7] MITTAS N., ARGYROPOYLOY V. ANGELIS L. (2010). Modeling the Relationship between Software Effort and Size Using Deming Regression. ***To be presented at 6th International Conference on Predictive Models in Software Engineering. September 12-13, 2010, Timisoara, Romania.***

Background: The relation between software effort and size has been modeled in literature as exponential, in the sense that the natural logarithm of effort is expressed as a linear function of the logarithm of size. The common approach to estimate the parameters of the linear model is ordinary least squares regression which has been extensively applied to various datasets. The least squares estimation takes into account only the error arising from the dependent variable (effort), while the measurement of independent variable (size) is considered free of errors.
Aims: The basis of the study is that in practice the assumption of measuring the size without error is hardly true, since the size of a software project depends on the precision of the tool of measurement and often by the subjectivity of the rater. Moreover, the sizes of projects comprising a dataset have been measured by different measurement tools and this adds another source of variability in the independent variable.
Method: In this paper, we consider a regression technique, known as Deming regression, which takes into account the error in measurement of the independent variable, the size. Deming regression is applied to four publically available datasets in order to model the linear relationship between effort and size and to
compare it with ordinary least squares.
Results: Accuracy measures of fitting (MAE, MdAE, MMRE, MdMRE, pred25) are improved by the Deming regression. Comparison of Absolute Errors (AE) by the Wilcoxon test shows significant difference at <0.001 level of significance.
Conclusions: Deming regression is appropriate for datasets where the size is subject to measurement error. However some assumptions on the variances of the measurement errors are arbitrary and need to be studied. Further work is needed for using the Deming regression for effort prediction.

[C8] KOSTI M., MITTAS N., ANGELIS L. (2010). DD-EbA: An algorithm for determining the number of neighbors in cost estimation by analogy using distance distributions. ***To be presented at Workshop on Artificial Intelligence, Techniques in Software Engineering (AISEW 2010), September 2010, Ayia Napa, Cyprus*****.**

Case Based Reasoning and particularly Estimation by Analogy, has been used in a number of problem-solving areas, such as cost estimation. Conventional methods, despite the lack of a sound criterion for choosing nearest projects, were based on estimation using a fixed and predetermined number of neighbors from the entire set of historical instances. This approach puts boundaries to the estimation ability of such algorithms, for they do not take into consideration that every project under estimation is unique and requires different handling. The notion of distributions of distances together with a distance metric for distributions help us to adapt the proposed method (we call it DD-EbA) each time to a specific case that is to be estimated without loosing in prediction power or computational cost. The results of this paper show that the proposed technique achieves the above idea in a very efficient way.

[C9] HAMDAN K., ANGELIS L. MITTAS N. (2010). Estimating learning by analogy: A case study in the UAE univerisity. ***International Conference of Education, Research and Innovation (ICERI 2010), November 15-17, 2010, Madrid, Spain*****.**

This work is concerned with how to measure and, most importantly, how to predict the “amount” of learning in “problem-based” learning programs. Due to difficulties in dealing quantitatively with educational aspects like learning, which are rather abstract when it comes to measuring, this is not a typical problem which can be dealt with easily traditional statistical analysis. The ordinal nature of data collected from surveys requires suitable estimation methods. In this paper we use a method called “estimation by analogy” (EbA), known from its application to software engineering problems, for estimating the amount of learning in a class which is based on historical data containing information from previous learning measurements. The approach is simple since it does not require any mathematical

background and assumptions; it is intuitively appealing since it is based on the idea of "finding the most similar cases" and, most importantly, it gives very good results. The method is applied on a dataset from 56 Math/IT classes collected at the end of the fall 2005 from the enrolment of UGRU, United Arab Emirates University.

[C10] MITTAS N. (2011) Evaluating the Performances of Software Cost Estimation Models through Prediction Intervals. ***International Conference on Econophysics, June 2-3, 2011, Kavala, Greece*****.**

The task of predicting accurately the cost required for the completion of a new software project is a challenging issue in the Software Cost Estimation area, since it is closely related with the activities of project management and the wise decision-making of organizations in order to bid, plan and budget a forthcoming system. However, the accurate prediction of the cost is often obtained with great uncertainty and for this reason there has been noted a lack of convergence in experimental studies. The main reason for the discrepancy can be derived from the inherent characteristic of prediction methodologies, since they produce point estimates without taking into account the risk covering the whole process. In this study, we propose a statistical framework, so as to focus on the construction of Prediction Intervals which provide an "optimistic" and a "pessimistic" guess for the true magnitude of the cost. The proposed framework that incorporates different accuracy indicators, formal hypothesis testing and graphical inspection of the predictive performance is applied on a dataset with real software projects.

Book Chapters

[B1] ANGELIS L., SENTAS P., MITTAS N., CHATZIPETROU P (2010). Methods for Statistical and Visual Comparison of Imputation Methods for Missing Data in Software Cost Estimation. ***In Modern Software Engineering Concepts and Practices: Advanced Approaches, Idea Group Inc. Publishing.***

Software Cost Estimation is a critical phase in the development of a software project and over the years has become an emerging research area. A common problem in building software cost models is that the available datasets contain projects with lots of missing categorical data. The purpose of this chapter is to show how a combination of modern statistical and computational techniques can be used to compare the effect of missing data techniques on the accuracy of cost estimation. Specifically, a recently proposed missing data technique, the multinomial logistic regression, is evaluated and compared with four older methods: listwise deletion, mean imputation, expectation maximization and regression imputation with respect to their effect on the prediction accuracy of a least squares regression cost model. The evaluation is based on various expressions of the prediction error and the comparisons are conducted using statistical tests, resampling techniques and a visualization tool, the regression error characteristic curves.

Greek Conferences/Publications

[G1] MITTAS N., L. ANGELIS (2006). Confidence intervals and hypothesis tests in the comparison of software cost estimation methods. ***Proceedings of the 19th Panhellenic Conference in Statistics, Kastoria (In Greek)*****.**

The delivery of qualitative software in predetermined time limits, budget and according to predefined schedules is one of the most important objectives of the organizations around the world. The accurate software cost estimation is a topic that has occupied the researchers' interest the last decades. A crucial issue in software development is the selection of the prediction model. Usually, this selection is based on the most commonly used accuracy measures. The previous policy of determining the most accurate

prediction model is not been validated by hypothesis tests and could lead to unstable and erroneous conclusions. In this paper, we present two statistical simulation tools in order to test the significance of the difference between two comparative models, namely the bootstrap and permutation tests. Three bootstrap techniques are used for the construction of accurate confidence intervals for the difference of means, whereas permutation tests are utilized to test whether the most widely known accuracy measures are significantly different for the comparative models.

[G2] DIMOKAS N., MITTAS N., NANOPOULOS A., ANGELIS L. (2008). A Prototype System for Educational Data Warehousing and Mining. ***12th Pan-Hellenic Conference on Informatics PCI 2008 August 2008, Samos, Proceedings published by IEEE.***

Universities are encountering growing demands by legislators and communities who are clamoring for valuable information about student achievement and university system accountability. Not only are universities required to measure annual progress for every single student, but government (through ministries for education) aid is directly linked to these results. The department of Informatics of Aristotle university of Thessaloniki has developed a data warehouse solution that assists the analysis of educational data. In this paper we present the design and development of the proposed data warehouse solution, which facilitates better and more thorough analysis of department's data. The proposed system constitutes an integrated platform for a thorough analysis of department's past data. Analysis of data could be achieved with OLAP operations. Moreover, we propose a thorough statistical analysis with an array of data mining techniques, that are appropriate for the examined tasks.

[G3] MITTAS N., L. ANGELIS (2009). Graphical comparison of software cost estimation models with regression error characteristic analysis. ***22nd Panhellenic Conference in Statistics, Chania, Greece (In Greek).***

Universities are encountering growing demands by legislators and communities who are clamoring for valuable information about student achievement and university system accountability. Not only are universities required to measure annual progress for every single student, but government (through ministries for education) aid is directly linked to these results. The department of Informatics of Aristotle university of Thessaloniki has developed a data warehouse solution that assists the analysis of educational data. In this paper we present the design and development of the proposed data warehouse solution, which facilitates better and more thorough analysis of department's data. The proposed system constitutes an integrated platform for a thorough analysis of department's past data. Analysis of data could be achieved with OLAP operations. Moreover, we propose a thorough statistical analysis with an array of data mining techniques, that are appropriate for the examined tasks.

[G4] MITTAS N., FLOROU G., POLYCHRONIDOU P. (2009). Examination of Duration of Studies for the Accounting Department of the Technological Institution of Kavala through Survival Analysis. ***22nd Panhellenic Conference in Statistics, Chania, Greece (In Greek).***

In this paper we present a statistical analysis that concerns the Department of Accounting of the Technological Educational Institution of Kavala. Except of the primary descriptive statistical analysis, we use several techniques of multivariate analysis in order to detect correlations between the factors that affect the duration of studies and the degree class of Bachelor of the students. We took into account all the available information, even that of the students that continue their studies, so we use Survival Analysis, a statistical methodology that is well-known and it is used especially in Biostatistics.

[G5] POLYCHRONIDOU P., MITTAS N., FLOROU G. (2009). Factors that Affect the Duration of Studies of the Accounting Department of the Technological Institution of Kavala. *5th Pan-Hellenic Data Analysis Conference with International Participitation, September 2009, Rethimno*.

Institutions of Greece's higher education have to inform and update the scientific society and the relevant Ministries about their students' progress among their overall image. In this paper, we analyze statistically the duration of studies at the Accountancy Department of Kavala Institute of Technology. We describe the demographic characteristics of all of our Department's students through the years. We use several techniques of the multivariate analysis, in order to detect correlations among the factors that affect the duration of studies and the degree class of Bachelor.

[G6] SALTAS V., KALAMPAKAS A., KOGEUTSOF A., POLYCHRONIDOY P., MITTAS N., TSIANTOS V. (2010). Evaluation of the Mathematical Skills for First-year Students of the Petroleum and Natural Gas Department*. 23th Panhellenic Conference in Statistics, Veria, Greece*. (*In greek*)

In this study, we evaluate the results of the Mathematical skills for the first-year students of the Petroleum and Natural Gas Department of the Technological Educational Institution of Kavala. The study is portioned into three stages of evaluation that are related with the students, educational system and professors. Initially, we collect and analyze the performances of the first-year students and more particularly, we study the knowledge and capabilities that students have acquired.

For this purpose, the study is based on a questionnaire with 20 questions of multiple choices that concern the mathematical capabilities of students, whereas the students are clustered into three basic categories in accordance to their high-school education. The statistical analysis of the questionnaires will provide some interesting information and knowledge acquisition concerning the educational system and the mathematical skills of the Petroleum and Natural Gas Department

[G7] TSIANTOS V., KAPENIS K., PATSILIAS G., MITTAS N., CHATZIFOTIOU S. (2011). The Usage of Video on the Education of Mathematics. 2th Panhellenic Conference

ΤΣΙΑΝΤΟΣ Β., ΚΑΠΕΝΗΣ Κ., ΠΑΤΣΙΛΙΑΣ Γ., ΜΗΤΤΑΣ Ν., ΧΑΤΖΗΦΩΤΙΟΥ Σ. (2011). Η χρήση του Βίντεο στη Διδασκαλία των Μαθηματικών: Μία Πιλοτική Εφαρμογή στο Μάθημα «Μαθηματικά ΙΙ» του Τμήματος Βιομηχανικής Πληροφορικής του ΤΕΙ Καβάλας. *Πρακτικά 2ου Πανελλήνιου Συνέδριου Ένταξη των ΤΠΕ στην Εκπαιδευτική Διαδικασία, Μάρτιος 2011, Πάτρα*.