# BENEFITS DERIVED FROM THE DATA MINING APPROACH TO PROJECT MANAGEMENT

Castejón, M.[(p)]; González, A.; Martínez, F.; Escribano, R.

## Abstract

This paper reports the recent experiences gained at applying data mining techniques to projects performance data. The projects considered are those collected by the International Software Benchmarking Standards Group (ISBSG). These are software development projects but the approaches described can be extended to other areas seamlessly although these methodologies are particularly suitable for this arena. From previous experiences that prove feasible this approach, we show how data mining techniques can provide unique insights on a based on facts approach.

*Keywords: Project management, data mining, ISBSG.*

## Resumen

Este artículo da cuenta de las experiencias recientes al aplicar las técnicas de minería de datos a los históricos de desarrollo de los proyectos. Los proyectos considerados en este estudio son los recogidos por el grupo ISBSG. Si bien estos son proyectos de desarrollo de software y las propuestas presentadas son particularmente adecuadas para las condiciones dadas, puede ser perfectamente aplicable a cualquier tipo de proyecto. A partir de experiencias previas que han demostrado la viabilidad de este enfoque, mostramos cómo las técnicas de minería de datos pueden revelar información única a partir de un enfoque basado en hechos.

*Palabras clave: Gestión de proyectos, minería de datos, ISBSG.*

## 1. Introduction

This work is a consequence of the interest of the EDMANS group in applying data warehousing and mining technique to project management (visit http://apiur.unirioja.es/edmans/ for more details).

The EDMANS research team (Engineering Data Mining And Numerical Simulations) comprises engineers and scientists specialized in different fields from engineering (design of machines, manufacturing process, structural analysis, environmental management, computer science, quality, statistics, automatization, modeling, etc.). This group develops its activity mainly supported by research projects funded by the European Commission and National Research Authorities. The main goals of the projects currently in progress are, amongst others:

- Optimization and modeling of industrial and environmental processes by means of Data Mining and Multivariate Statistics techniques. Usually, the target is to obtain hidden knowledge as models from historical records of those processes. Mainly, the uses are the quality improvement of the products, reduction of faults and fostering the process control. Normally, it is necessary first to redesign the data gathering system as well as to spend some efforts by data pre-processing.

- Advanced design coupled to numerical simulation of mechanical, CFD and thermal problems related to structural systems, machine components and environmental processes. Usually the target is to reduce cost and to improve the final quality of products.
- Application of data warehousing and data mining techniques in project management in order to build management support systems to help improving decision making processes.

Our interest in applying data warehousing and data mining techniques to project management is focus in providing useful tools, such as decision support systems to aid project professionals in their management daily tasks.

IPMA competence element 1.09 “Project Structures” and competence element 1.17 “Information and documentation” acknowledge the need of management on the different information flows associated with the project and the potential provided by data warehousing and mining tools in order to extract hidden knowledge from databases supporting the increasingly more commonly used information systems.

Data mining tools focus on retrieving underlying substantial information, thus reducing the amount of information but improving its quality. Nevertheless, project management is a field where practioners are not frequently accustomed to using these tools. This is mainly caused for their recent advent and yet shallow penetration amongst project management professionals. Nevertheless, the international community is paying more and more attention to the opportunities provided by these techniques. We describe, in the following lines the most relevant works following a data driven approach to managerial processes, a based on acts approach that coherent with ISO9000 standards principles.

The sixties witnessed the appearance of parameter estimation models applied to variables as critical in software projects management as their cost itself. From then, these estimates have been applied for multiple purposes and goals:

- Budgeting
- Risk analysis
- Planning and monitoring
- Strategic analysis

Since their birth, the interest in a precise estimation of the most critical parameters of the projects has been an active field of research. During the following years, a number of models and techniques were proponed. They can be coarsely classified as:

- Parametric models
- Experience based techniques
- Learning oriented techniques
- Dynamics based models
- Regression models
- Mixed Bayesian models
- Hybrid models

The increasingly growing computational power of commodity computers currently enable us to approach the estimation problems from a new point of view, that provided by data warehousing and mining tools. This new science comprises a vast set of tools and techniques. Their main purpose is to reveal hidden knowledge in also increasingly larger databases.

## 2. Recent works

Mendes and Lokan (2008) [1] analyzed the difference observed amongst other previous works by Jeffery et al. (Using public domain metrics to estimate software development effort. Proceedings Metrics'01, London, pp 16–27, 2001; S1) to compare the effort prediction accuracy between cross- and single-company effort models; Mendes et al. (A replicated comparison of cross-company and within-company effort estimation models using the ISBSG Database, in Proceedings of Metrics'05, Como, 2005; S2); Lokan and Mendes (Cross-company and single-company effort models using the ISBSG Database: a further replicated study, Proceedings of the ISESE'06, pp 75–84, 2006; S3). Their results corroborated those from S1, suggesting that differences in the results obtained by S2 were likely caused by legitimate differences in data set patterns. By applying the experimental procedure of S2 to the data set used in S1 (study S3), and the experimental procedure of S1 to the data set used in S2 (study S4), they investigate the effect of all the variations between S1 and S2. Their results for S4 support those of S3, suggesting that differences in data preparation and analysis procedures did not affect the outcome of the analysis. Thus, the different results of S1 and S2 can most likely be seen as caused by some fundamental difference in the data sets themselves.

Ruchi Shukla and Arun Kumar Misra (2008) [2] focused on software maintenance as an essential component of software developement. They developed a Neural Network (NN) based effort estimator using Matlab. A feed forward back-propagation NN employing Bayesian regularization training was selected and trained for one dataset. Various categories of software maintenance cost drivers and their effect on maintenance effort were analyzed using different combinations of number of hidden layers and hidden neurons etc. Their results successfully modeled the maintenance effort.

Huang et al. (2008) [3] showed their interest in precision in estimating the required software development effort as a critical factor in the success of software project management. Their work aimed to explore the effects of accuracies of the software effort estimation models established from the clustered data by using the International Software Benchmarking Standards Group (ISBSG) repository. The ordinary least square (OLS) regression method was adopted to establish a respective effort estimation model in each cluster of datasets. The results obtained by their empirical experiment results showed that the estimation accuracies did not not reveal significant differences within the respective dataset clustered by each software effort driver. It also demonstrated that software effort estimation models from the clustered data presented almost similar accuracy results compared to models from the entire data without clustering.

Wei Xia et al. (2008) [4] discussed the concepts of calibrating Function Point, whose aims were to estimate a more accurate software size that fits for specific software application, to reflect software industry trend, and to improve the cost estimation of software projects. A FP calibration model called Neuro-Fuzzy Function Point Calibration Model (NFFPCM) that integrated the learning ability from neural network and the ability to capture human knowledge from fuzzy logic was proposed. The empirical validation using International Software Benchmarking Standards Group (ISBSG) data repository showed a 22% accuracy improvement of mean magnitude relative error (MMRE) in software effort estimation after calibration.

Aroba et al. (2008) [5] reported a parametric software cost estimation model to estimate the effort and time required to develop a software product. Their solution proposed a segmented model based on fuzzy clusters of the project space. The use of fuzzy clustering allowed obtaining different mathematical models for each cluster and also the items of a project database to contribute to more than one cluster, while preserving constant time execution of the estimation process. Their results in an evaluation of a concrete model using the ISBSG project database yielded better figures of adjustment than its crisp counterparts and encourage further work in this area..

Qin Liu et al. (2008) [6], in ‘‘Evaluation of Preliminary Data Analysis Framework in Software Cost Estimation Based on ISBSG R9 Data,’’ showed how to lower prediction errors through a preliminary analysis of raw data. The accuracy of predictions is the greatest challenge when trying to perform project cost estimation. By applying their method to historical data sets, the authors were able to significantly improve the accuracy of predictions.

Bourque et al. (2007) [7] with an approach based on the empyrical análisis of the database compiled by the International Software Benchmarking Standards Group (ISBSG), studied different models related with the duration estimation in software engineering projects. These models reflected the behavior both of the data set in general and of subsets formed considering their capacity as well: personal computers, workstations and mainframes. Projects were also classified depending on whether the project required less than 400 hundred people or not. Considering these restriction they evaluated the usefulness of adding an additional independent variable, the maximum number of resources, and the possibility of obtaining duration models directly from the project size as measured in functional points as well.

Cuadrado-Gallego et al. (2007) [8] studied the possibility of obtaining predictive models related to the necessary effort to develop software project. Their approach was based on multiple parametric models generated by unsupervised data classification. Using this classification of the ISBSG data set, the proceed pursuing to improve the precision of traditional parametric models. In their paper, they describe the inputs, techniques, tools and results obtained. Their results show the benefits of their approach. an extension of existing models that provides significant improvements while not increasing the complexity of the estimation process.

Cuadrado-Gallego y Sicilia (2007) [9] extended previous works (Cuadrado-Gallego et al., 2007) [8] by proposing an algorithm that spawns segmented parametric estimation models oriented to the estimation of the effort required in software projects. Their proposal tackles the problem of obtaining a parametric model from the historical data recorded in databases of real projects variables. These, as in the ISBSG case, comprise highly heterogeneous data sets from projects of sundry sizes, different process and different requirements, which makes difficult to capture such richness with only one parametric model, thus usually leading to a poor fit of the model. Segmented parametric model take advantage of the regression technique to derive models with local context validity, as many as the number of adequate partitions to describe the nature of the considered data. Their approach proposes a generic algorithm that provides candidate models by an automatically calibrated model based on the data set which supports it, also considering an empirical evaluation of the data that uses the well-known EM (expectation—maximization) algorithm altogether with conventional parametric models.

Abran et al. (2007) analyzed the behavior of a black box tool in the estimation of software projects parameters. Their work shows the importance that other techniques have traditionally enjoined in building support systems to aid project managers in their decision making processes. These tools have in common the lack of documentation on their performance. They face this problem using the International Software Benchmarking Standards Groups databases. Their analysis is presented in three stages: first they apply some preprocessing to the data set, as usual; second, they obtain different estimation models directly from the preprocessed data set, both considering and removing outliers;

third, and finally, they evaluate the most widely spread commercial software estimation tools. In all, the observed behavior of these widely spread tools is quite poor, which emphasizes the need for correctly documenting the applicability ranges of these estimation techniques, an effort that developers must make in order to succeed.

## 3. Conclusions

A brief revision of the state of the art has been performed on the proposals appeared in 2007 and early 2008 in the field of applying data warehousing and data mining technique to analyze databases related to software engineering projects. In particular, we focused on those that use the data sets compiled by the International Software Benchmarking Standards Group.

A high interest and intense activity can be seen in this particular field. Among the main benefits derived from the use of this approach, the improvements observed in the effort and cost estimation must be highlighted. Naturally, these are the parameters more carefully studied by the scientific community.

## References

[1] Emilia Mendes and Chris Lokan. "Replicating studies on cross- vs single-company effort models using the ISBSG Database". Empirical Software Engineering, Vol. 13, núm. 1, pp. 3-37, 2008

[2] Ruchi Shukla and Motilal Nehru. "Estimating software maintenance effort: a neural network approach". Proceedings of the 1st conference on India software engineering conference. Hyderabad, India. pp. 107-112. 2008.

[3] Sun-Jen Huang, Nan-Hsing Chiu and Yu-Jen Liu. "A comparative evaluation on the accuracies of software effort estimates from clustered data".Information and Software Technology, Article in Press (doi:10.1016/j.infsof.2008.02.005), 2008.

[4] Wei Xia, Luiz Fernando Capretz, Danny Ho and Faheem Ahmed. "A new calibration for Function Point complexity weights". Journal of Systems and Software, Article in Press (doi:10.1016/j.infsof.2007.07.004 ). 2008

[5] Javier Aroba, Juan J. Cuadrado-Gallego, Miguel-Ángel Sicilia, Isabel Ramos and Elena García-Barriocanal. "Segmented software cost estimation models based on fuzzy clustering". Journal of Systems and Software, Article in Press (doi:10.1016/j.jss.2008.01.016). 2008.

[6] Qin Liu, Wen Zhong Qin, Robert Mintram and Margaret Ross. "Evaluation of preliminary data analysis framework in software cost estimation based on ISBSG R9 Data". Software Quality Journal, article in press (doi:10.1007/s11219-007-9041-4). 2008

[7] Bourque P., Oligny S., Abran S., Fournier B. “Developing Project Duration Models in Software Engineering”, Journal of Computer Science and Technology, Vol. 22, núm. 3, pp. 348-357. 2007.

[8] Cuadrado-Gallego, J.J., Rodríguez, D., Sicilia, M.A., Garre Rubio, M., García Crespo, A. “Software project effort estimation based on multiple parametric models generated through data clustering”, Journal of Computer Science and Technology, Vol. 22, núm. 3. pp. 371-378. 2007.

[9] Cuadrado-Gallego, J.J., Sicilia, M.A. “An algorithm for the generation of segmented parametric software estimation models and its empirical evaluation”, Computing and Informatics, Vol. 26, núm. 1, pp. 1-15. 2007

[10] Abran, A., Ndiaye, I., Bourque, P. “Evaluation of a black-box estimation tool: A case study”, Software Process Improvement and Practice, Vol. 12, núm. 2, pp. 199-218. 2007

## Acknowledgements

The authors gratefully acknowledge support from the Dirección General de Investigación, Ministerio de Educación y Ciencia de España through grants DPI2006-03060, DPI2006-14784, DPI-2006-02454 and DPI2007-61090; from the European Union through project RFS-PR-06035. Finally, the authors also thank the Gobierno Autonomo de La Rioja for its support through the 3rd Plan Riojano de I+D+i.

## Correspondencia (Para más información contacte con):

Manuel Castejón Limas
Área de Proyectos de Ingeniería.
Escuela de Ingenierías Industrial e Informática
Campus de Vegazana s.n., 24071 León (España)
Phone: +34 987 291 000 – ext. 5382
E-mail: mcasl@unileon.es