Available online at www.sciencedirect.com
SCIENCE DIRECT®
The Journal of Systems and Software 79 (2006) 353–361
The Journal of Systems and Software
www.elsevier.com/locate/jss

An empirical study of process-related attributes in segmented software cost-estimation relationships

Juan J. Cuadrado-Gallego [a], Miguel-Ángel Sicilia [a,*], Miguel Garre [a], Daniel Rodríguez [b]

[a] Computer Science Department, Polytechnic School, University of Alcalá. Ctra. Barcelona km. 33.6, 28871 Alcalá de Henares, Madrid, Spain
[b] Computer Science Department, University of Reading, Reading RG6 6AY, UK

Received 15 February 2005; received in revised form 23 April 2005; accepted 23 April 2005
Available online 1 July 2005

Abstract

Parametric software effort estimation models consisting on a single mathematical relationship suffer from poor adjustment and predictive characteristics in cases in which the historical database considered contains data coming from projects of a heterogeneous nature. The segmentation of the input domain according to clusters obtained from the database of historical projects serves as a tool for more realistic models that use several local estimation relationships. Nonetheless, it may be hypothesized that using clustering algorithms without previous consideration of the influence of well-known project attributes misses the opportunity to obtain more realistic segments. In this paper, we describe the results of an empirical study using the ISBSG-8 database and the EM clustering algorithm that studies the influence of the consideration of two process-related attributes as drivers of the clustering process: the use of engineering methodologies and the use of CASE tools. The results provide evidence that such consideration conditions significantly the final model obtained, even though the resulting predictive quality is of a similar magnitude.

Keywords: Parametric software effort estimation; Clustering algorithms; Software cost drivers; EM algorithm

1. Introduction

The *Parametric Estimating Handbook* (PEH) (PEI, 1999) defines parametric estimation as "a technique employing one or more cost estimating relationships (CERs) and associated mathematical relationships and logic". These techniques are nowadays widely used to measure and/or estimate the cost associated with software development (Boehm et al., 2000a). CERs are mathematical devices that obtain numerical estimates from main *cost drivers* that are known to affect the effort or time spent in development. According to the PEH, these drivers are the controllable system design or planning characteristics that have a predominant effect on system cost. Parametrics uses the few important parameters that have the most significant cost impact on the software being estimated. Nonetheless, even though the final CERs should use only the most significant parameters, it is often also useful to consider other parameters as a foundation for the logics of deriving the mathematical relationships from empirical data. The notion of "cost realism" as described in the PEH clearly points out to this dimension of reasonable and justified usage of data.

One important aspect of the process of deriving models from databases is that of the heterogeneity of data. Heteroscedasticity (non-uniform variance) is known to be a problem affecting data sets that combine data from heterogeneous sources (Stensrud et al., 2002). When

* Corresponding author. Tel.: +34 916 249 104; fax: +34 916 249 103.
E-mail addresses: jjcg@uah.es (J.J. Cuadrado-Gallego), msicilia@uah.es (M.-Á. Sicilia), miguel.garre@uah.es (M. Garre), d.rodriguez-garcia@rdg.ac.uk (D. Rodríguez).

0164-1212/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2005.04.040