Regression Analisys of Segmented Parametric Software Cost Estimation Models Using Recursive Clustering Tool

M. Garre[1], M.A. Sicilia[1], J.J. Cuadrado[1], and M. Charro[2]

[1] University of Alcalá, Ctra. Barcelona km 33.6 - 28871,
Alcalá de Henares, Madrid, Spain
`{miguel.garre, jjcg, msicilia}@uah.es`
[2] Dept. Applied Sciences, Torrejón Air Base - Technical School, 28850 - Torrejón de
Ardoz, Madrid, Spain

Abstract. Parametric software effort estimation models rely on the availability of historical project databases from which estimation models are derived. In the case of large project databases with data coming from heterogeneous sources, a single mathematical model cannot properly capture the diverse nature of the projects under consideration. Clustering algorithms can be used to segment the project database, obtaining several segmented models. In this paper, a new tool is presented, Recursive Clustering Tool, which implements the EM algorithm to cluster the projects, and allows use different regression curves to fit the different segmented models. This different approaches will be compared to each other and with respect to the parametric model that is not segmented. The results allows conclude that depending on the arrangement and characteristics of the given clusters, one regression approach or another must be used, and in general, the segmented model improve the unsegmented one.

Keywords: Software Engineering, Effort estimation, Segmented parametric model, Recursive Clustering Tool (RCT), Clustering, EM algorithm.

1 Introduction

Segmented parametric software estimation models has been used against not segmented, in a variety of works. Concretely J.J. Cuadrado [1] show the influence of cost drivers, CASET (Use of Case Tools) and METHO (Use of a Methodology), on estimation effort using segmented parametric models. In the work of M. Garre [2], a recursive process is described to obtain a set of segments that shows more homogeneous characteristics than previously. Finally the same author, compares the two models (segmented and not segmented) in [3]. These works are based in the use of the EM algorithm [7,5,4], which has been integrated into a software tool: "Recursive Clustering Tool" (RCT).

Segmented parametric estimation models have arisen in contrast to not segmented, due to the use of heterogeneous historical projects databases on the estimation process. For example the projects database ISBSG (International