COCOMO II Local Calibration Using Function Points

Mauricio Aguiar
TI Metricas
mauricio@metricas.com.br

Abstract

COCOMO II is an updated version of the COCOMO software cost estimation model published by Barry Boehm in 1981. COCOMO II was originally calibrated to 161 project data points from several sources. Even though the original calibration can be used by a variety of organizations, software cost estimation models generally perform better with local calibration. This is particularly important for organizations whose project profiles differ from those included in the original COCOMO II calibration. Local calibration involves many challenges, the first being to obtain a set of completed projects that can be used in the calibration process. Because it is not feasible to get a random sample the usual practice is to target typical projects from the organization's most critical business area. Once the sample is defined several project attributes have to be collected, such as project name, description, development platform, language, software size, scale factors, effort multipliers, actual effort, and actual schedule. In Brazil, most organizations that implement software measurement use function points as the sole measure of size. This makes function points the natural measure of size when collecting project data in this setting. This paper discusses the challenges, difficulties, and lessons learned in calibrating COCOMO II models for 5 Brazilian organizations of the government and private sectors. Calibration results and recommendations are presented and discussed.

1. INTRODUCTION

1.1. The COCOMO Model

COCOMO (an acronym for COnstructive COst MOdel) is a widely known software engineering cost model created by Barry Boehm circa 1979 and fully described in [1]. COCOMO II is an update of the original model to account for changes in the way organizations build software. It is aligned with modern software practices such as iterative and incremental development processes, UML-based analysis and design, and risk-driven life-cycles [2]. COCOMO II supports three estimation models: Application Composition, Early Design, and Post-Architecture. The Application Composition model is used on projects that use I-CASE tools for rapid application development, the Early Design model is used when there is only enough project information for a rough estimate, and the Post-Architecture model is used when most of the life-cycle architecture has been defined [8]. COCOMO II uses source lines of code (SLOC) and function points as its major size measures. Because the number of delivered source instructions is the primary COCOMO cost driver [1], function points have to be converted to SLOC before being used as input to the model [2]. COCOMO II was originally calibrated to 83 data points using a multiple regression approach for a prediction level PRED(.30) = 52%, meaning the model yielded results within 30 percent of the actuals 52% of the time [8]. Prediction level is a measure of estimation accuracy [6]. When the model was calibrated for each of the major sources of data, its prediction level increased to PRED(.30) = 64%. Some years later, a new model was calibrated using a Bayesian approach. The prediction level of this model was PRED(.30) = 75%, meaning the model yielded results within 30 percent of the actuals 75% of the time. After stratifying the model into sets based on the eighteen major sources of project data results improved to PRED(.30) = 80% [2]. Local calibration improves prediction accuracy because of inconsistencies in the interpretation of model definitions and parameters across organizations, as well as differences in software processes and life-cycles. It is generally recommended that organizations collect historical data in order to calibrate software cost models for local conditions [7],[2].

1.2. Scope of the Study

This paper describes the calibration of COCOMO II estimation models for 5 Brazilian organizations of the government and private sectors. In Brazil most organizations that implement software measurement use function points as the only measure of size, therefore all projects addressed in this study were measured in function points. Because participating organizations were all interested in obtaining estimates very early in the life-cycle, the COCOMO II Early Design Model was used. In addition to discussing the challenges, difficulties, and lessons learned in the calibration process, a specific goal was to investigate the use of function points as a size measure with COCOMO II. It is relevant to note that all calibrations were done in a learning environment – most participants (except consultants) were doing this kind of work for the first time.

2. MODEL CALIBRATION FRAMEWORK

Four out of the five organizations involved in this study started a COCOMO II calibration project with the intention of establishing a software project estimation process. The remaining organization needed to estimate effort and schedule for a software project before launching a Request for Proposals. All five COCOMO II models were calibrated using the following framework:

Data Collection

- Study the environment and establish project categories
- Select a target project category
- Select projects to be measured
- Determine actuals – effort and schedule
- Measure selected projects in function points
- Determine COCOMO II scale factors and effort multipliers for each project

Model Calibration

- Calibrate the COCOMO II model using the CALICO [12] software

Analysis

- Assess calibration and analyze results

Each step is described in the following paragraphs.

2.1. Study the Environment and Establish Project Categories

The general goal of this step is to acquire knowledge on the organization and its software development process. The specific goal is to establish a set of project categories in order to target those with the highest business priority for the organization. This is usually accomplished in interviews with project managers and technical personnel. The categories identified should be ranked in accordance with their business relevance for the organization.

2.2. Select Target Category

After project categories have been determined and the organization's objectives and priorities clarified, a single project category should be selected as target for the model calibration effort. One important consideration when selecting a category is that completed projects should be available for that category, as well as people who understand the corresponding applications and can answer questions on their functionality and characteristics.

2.3. Select Projects to Be Measured

A set of at least 10 projects of the selected category should be chosen for measurement. Ideally more projects should be selected, however sometimes one will have to do with less, as shown in the case studies that follow. It is quite common that after a detailed inspection some projects will be found ineligible for a variety of reasons: use of a different technology, problems with the team, radical change of scope, or any other condition that might make the project an outlier, or indicate it should be placed in a different category.

2.4. Determine Actuals – Effort and Schedule

In order to calibrate a COCOMO II model actual effort and schedule will have to be determined for each selected project. This will usually require the analysis of time cards and project schedules, as well as setting up interviews with project participants. COCOMO II was originally designed and calibrated to produce estimates for the Elaboration and Construction life-cycle phases [2], as defined in Unified Process terminology, or their Waterfall counterparts. Nevertheless, it is possible to calibrate a particular model to produce estimates for the full life-cycle, from Inception to Transition. In the following calibration projects, project categories were defined in such a way as to allow estimation of the Inception phase as a constant percentage of the total project. A similar assumption was made for the Transition phase, excluding full-scale installation and training, so a COCOMO II model could be properly calibrated to estimate effort for the full life-cycle.

2.5. Measure Selected Projects in Function Points

Unadjusted function points were used in the calibration process, with a uniform SLOC to FP factor equal to 100. Multiplication by 100 was performed primarily to avoid problems with software tools that are better prepared to deal with SLOC. Multiplying FPs by 100 makes FP values comparable to SLOC in order of magnitude and does not affect the estimates provided the same factor is used when producing estimates. To minize counting effort, in most cases function points were estimated using the NESMA "estimated function point counting" technique [10].

2.6. Determine Scale Factors and Effort Multipliers

The COCOMO II Early Design Model requires rating 5 scale factors, and 7 effort multipliers. Five out of the seven Early Design effort multipliers are defined as combinations of the Post-Architecture effort multipliers (the remaining ones are RUSE and SCHED – see [2] for details). In order to make effort multiplier calculations easier the team used a spreadsheet that combined Post-Architecture drivers to obtain the Early Design drivers. Whenever a weighted average was required in [2] a simple average was used. The USC COCOMO II software was used to calculate the Effort Adjustment Factor – EAF.

2.7. Calibrate Model

Calibration can be done using a spreadsheet, or specialized software such as the free COCOMO II tool from USC [11], or CALICO, a free tool from SoftStar Systems [12]. For consistency only CALICO was used for calibration. For better results the team always calibrated both the constant and the exponent in the effort equation, even for as many as 6 projects only.

2.8. Assess Calibration and Analyze Results

The calibrated COCOMO II model was used to obtain estimates for the same projects used in the calibration. In an ideal setting an out-of-sample validation approach would be used [2]. However, the small number of projects available did not allow for that kind of procedure. Results were assessed based on the MRE and PRED(.30) values obtained. Projects with a high percent error were further investigated for measurement errors and/or problems with driver ratings.

3. STUDY RESULTS

In this section each case study is described and discussed. For brevity only effort estimation is addressed, even though schedule was also estimated in each case. Confidentiality issues prevented the disclosure of some data. Participating organizations were labeled A, B, C, D, and E, three of them being government organizations and two of the private sector. Two organizations were financial institutions, one was a service organization, one was an IT organization, and one was a manufacture. The following figures depict size in function points (only once size is depicted in SLOC), and effort in Person-Months.

3.1. Case Study: Organization A

The goal of Organization A was to estimate effort and schedule for one project before launching an RFP. All measurements were done by consultants. The client's team provided the required information.

3.1.1. Data Collection

Because the goal was to estimate effort and schedule for a single project, only one project category was defined. The organization had prepared a detailed description of the target project, including high-level use cases and some architectural details. Based on that information the consulting team was able to create a project profile. The team then searched for completed projects with the defined profile. Basically they were interested in projects that had used the same language, platform, and development process as planned for the target project. The organization was able to find 8 completed projects that satisfied the criteria. However, closer inspection revealed that 2 of those projects had actually been programmed in a language other than the target language, so eventually only 6 projects were used in the calibration. In the absence of reliable documentation, actual effort and schedule were determined in interviews with former project participants. Each project was measured using the IFPUG 4.1.1 function point counting technique [14]. The organization had skilled resources and software to count lines of code, so the team was able to obtain SLOC measures using the checklist available in [2]. Scale factors and effort multipliers were determined for each project. All scale factors were kept at the nominal level. Figure 1 depicts the Effort Adjustment Factor (EAF) for each project.

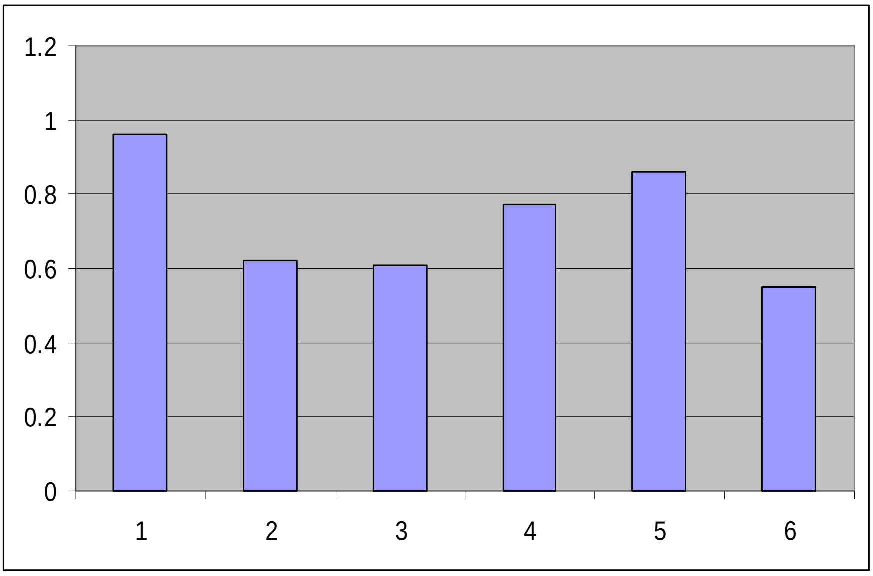

Figure 1 – EAF (Mean = 0.73; Std. Dev. = 0.16)

3.1.2. Model Calibration

Two COCOMO II Models were calibrated for this organization, one based on SLOC and another on function points. Figures 2 and 3 depict actual and estimated values for SLOC and FP.

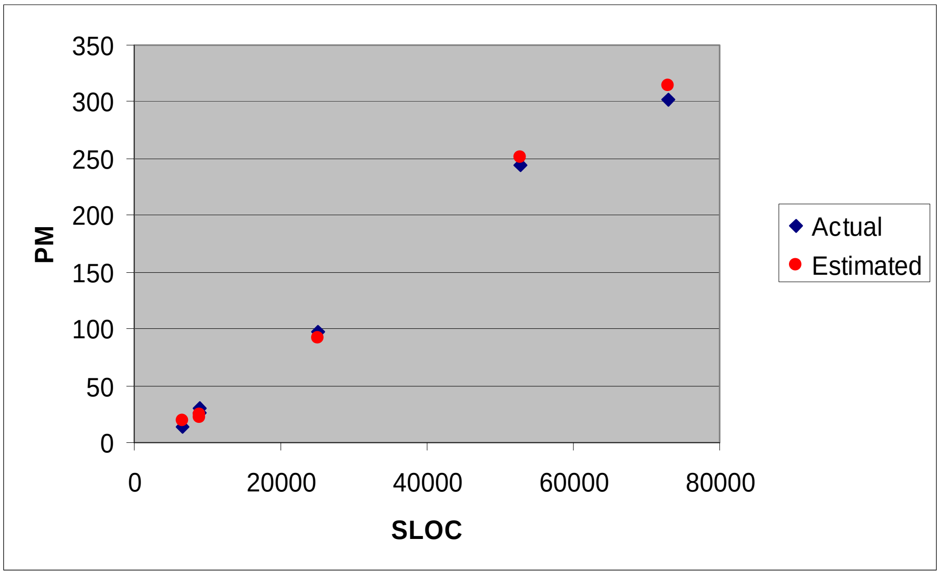

Figure 2 – Calibration Results for SLOC
MRE = 11.68% – PRED(.30) = 83%
A = 4.25, B = 0.85

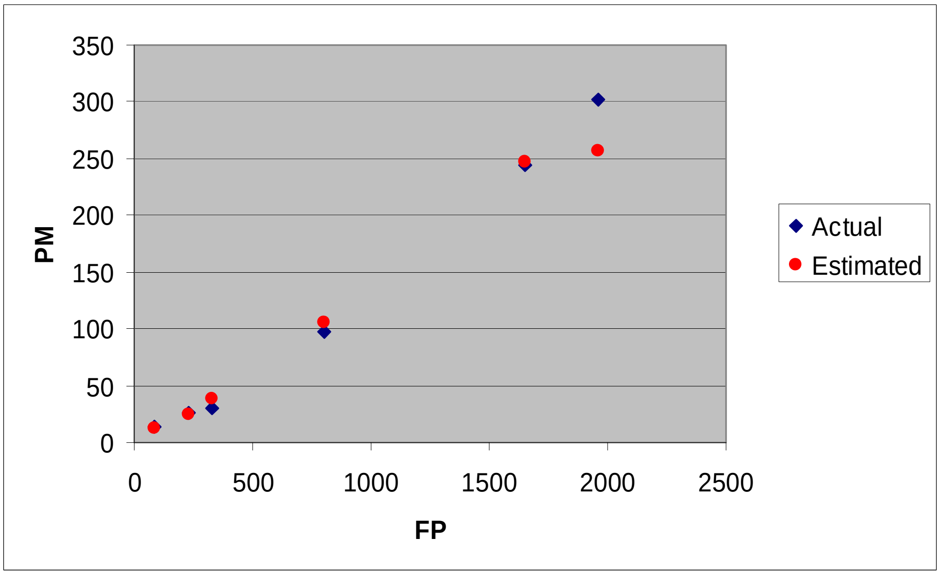

Figure 3 – Calibration Results for Function Points
MRE = 11.38% – PRED(.30) = 100%
A = 2.96, B= 0.68

3.1.3 Analysis

The two models yielded comparable results. The MRE is practically the same and the difference in PRED(.30) could be explained by looking at a single project – its percent error was 31.6% in the SLOC calibration, slightly above the PRED(.30) limit. Comparing the PRED(.30) values found with the COCOMO II.2000 calibration where PRED(.30) was 80% [2] led the team to assess the calibration obtained as good. Examination of other PRED values and of the equation coefficients made the team confident that both models obtained were reasonable, even though only 6 projects had been used.

3.2. Case Study: Organization B

The goal of Organization B was to implement a COCOMO II estimation process. Function point sizing and cost driver ratings were done by the client's team supported by consultants, as part of a training program.

3.2.1. Data Collection

This organization preferred to select completed projects based on availability. All projects were small, with actual schedule ranging from 2 to 4 months. Only 6 projects were available so those were selected. Actual effort and schedule determination was based on interviews with former project participants. Each project was measured using the NESMA technique [10] as part of a function point counting training program. Scale factors and effort multipliers were determined for each project. All scale factors were found to be at the nominal level. Figure 4 depicts the Effort Adjustment Factor (EAF) for each project.

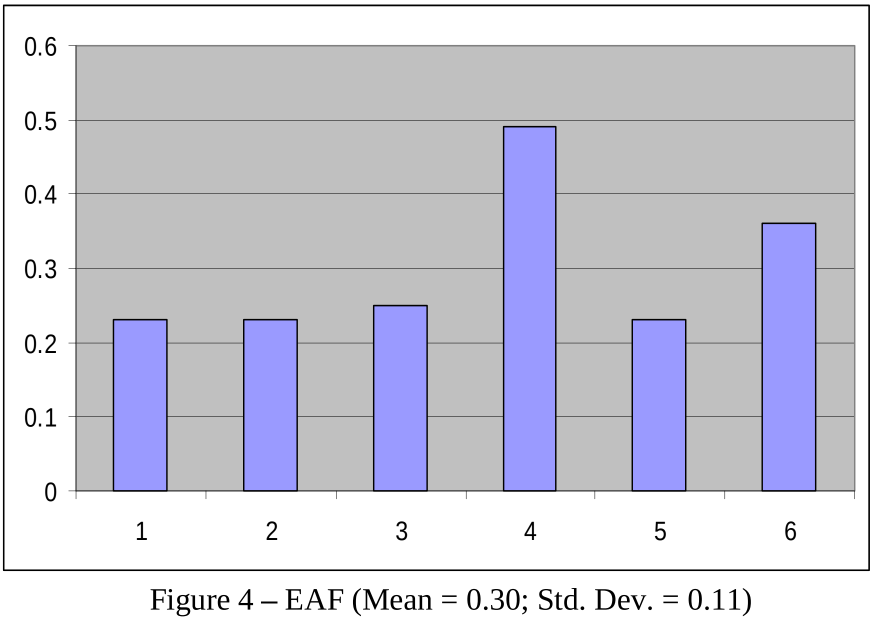

Figure 4 – EAF (Mean = 0.30; Std. Dev. = 0.11)

3.2.2. Model Calibration

Figure 5 depicts actual and estimated values for the calibrated model.

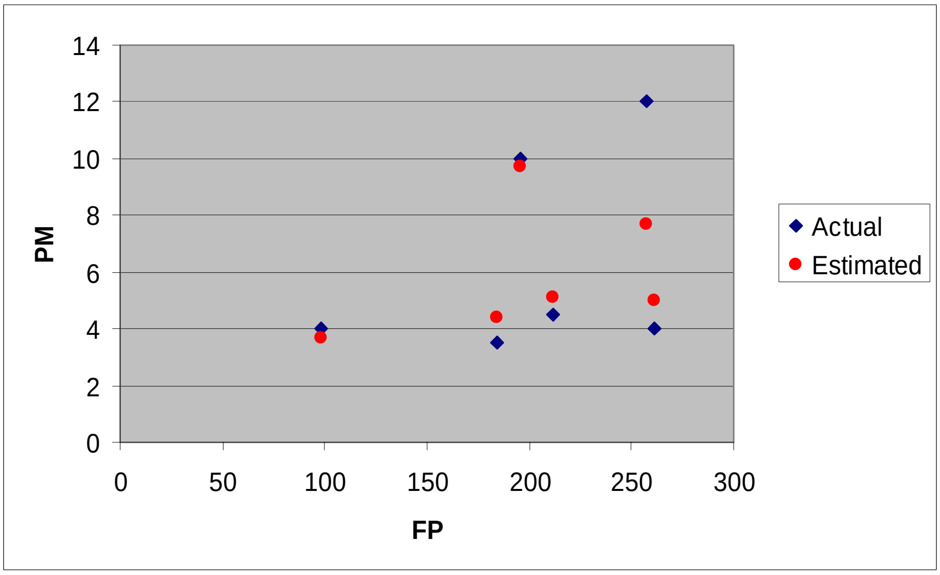

Figure 5 – Calibration Results
MRE = 18.50% – PRED(.30) = 83%
A = 7.76, B= 0.12

3.2.3. Analysis

Calibration results were considered satisfactory. One caveat is that only 6 projects were used. The organization was advised to improve the model with more projects before using it as a basis for doing business.

3.3. Case Study: Organization C

The goal of Organization C was to implement a COCOMO II estimation process. Function point sizing and cost driver ratings were done by the client's team supported by consultants.

3.3.1. Data Collection

This organization selected a specific platform for the COCOMO II model calibration effort. A total of 16 projects were collected, all from the same project category. Actual effort and schedule determination was exclusively based on interviews with former project participants. Projects were measured using the NESMA technique [10]. Most participants had been previously trained in FP counting. Scale factors and effort multipliers were determined for each project. All scale factors were found to be at the nominal level. Figure 6 depicts the Effort Adjustment Factor (EAF) for each project.

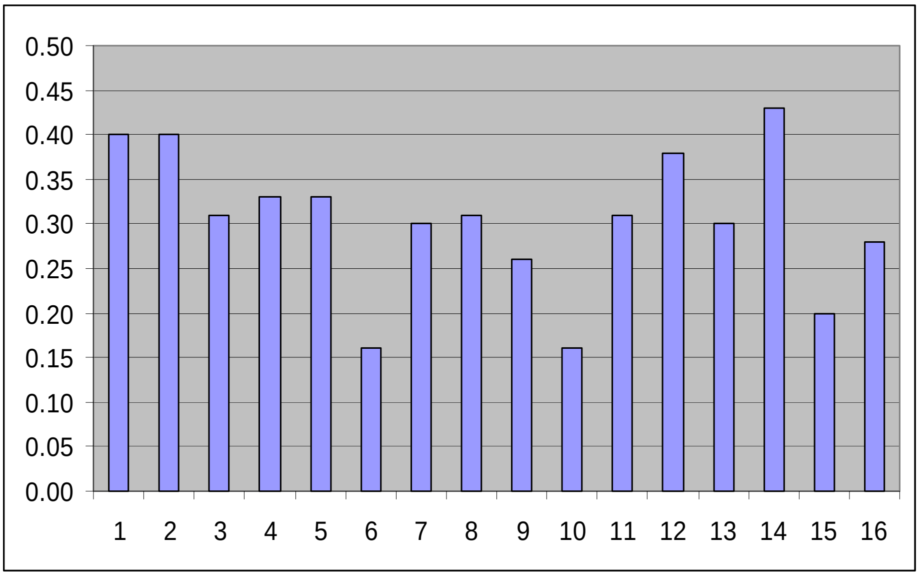

Figure 6 – EAF (Mean = 0.30; Std. Dev. = 0.08)

3.3.2. Model Calibration

Figure 7 depicts actual and estimated values for the calibrated model.

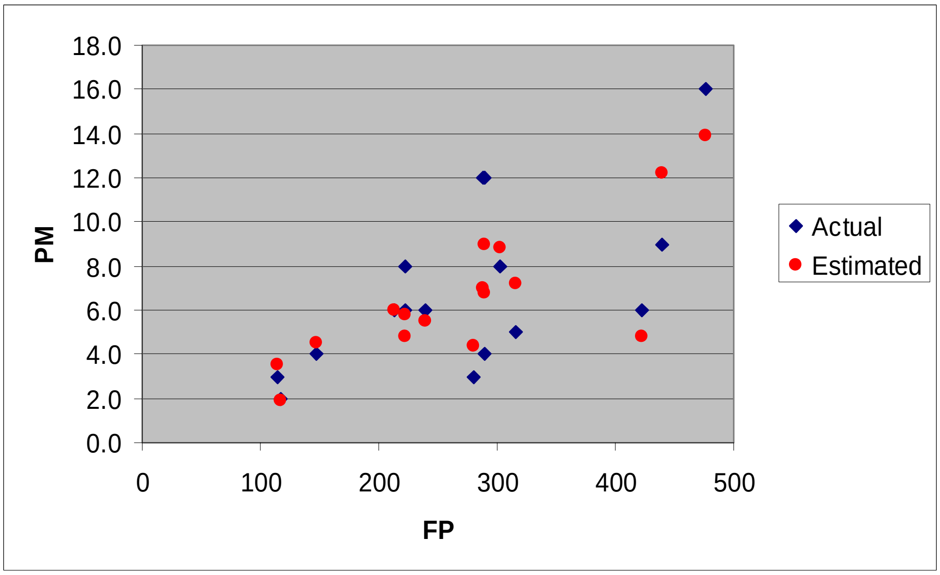

Figure 7 – Calibration Results
MRE = 29.52% – PRED(.30) = 56%
A = 2.00, B= 0.54

3.3.3. Analysis

The scatter diagram shows a large variation in effort around 300 FP. For those projects effort varied from 3 to 12 Person-Months for approximately the same size. A similar situation occurs in the vicinity of 450 FP. Size in function points and COCOMO II cost drivers alone were unable to explain this variation. Further analysis will be necessary in order to refine the model.

3.4. Case Study: Organization D

Organization D was interested in implementing a COCOMO II estimation process. Function point sizing and cost driver ratings were done by the client's team supported by consultants.

3.4.1. Data Collection

This organization was able to find 8 projects available for the calibration effort. Most projects came from the same platform and used the same technology. Actual effort and schedule determination was based on interviews with former project participants. Each project was measured using the NESMA [10] technique. Some participants had been previously trained in FP counting. Scale factors and effort multipliers were determined for each project. All scale factors were kept at the nominal level. Figure 8 depicts the Effort Adjustment Factor (EAF) for each project.

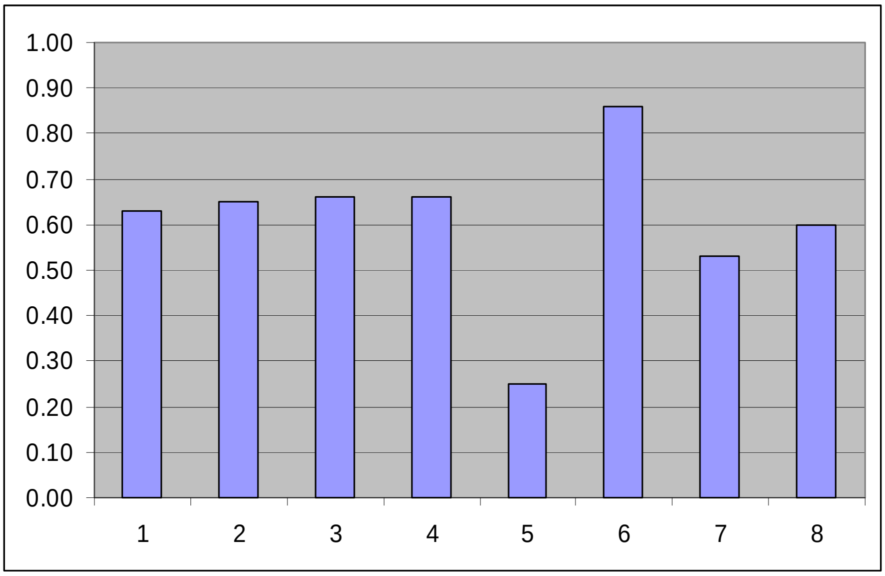

Figure 8 – EAF (Mean = 0.61; Std. Dev. = 0.17)

3.4.2. Model Calibration

Figure 9 depicts actual and estimated values for the calibrated model.

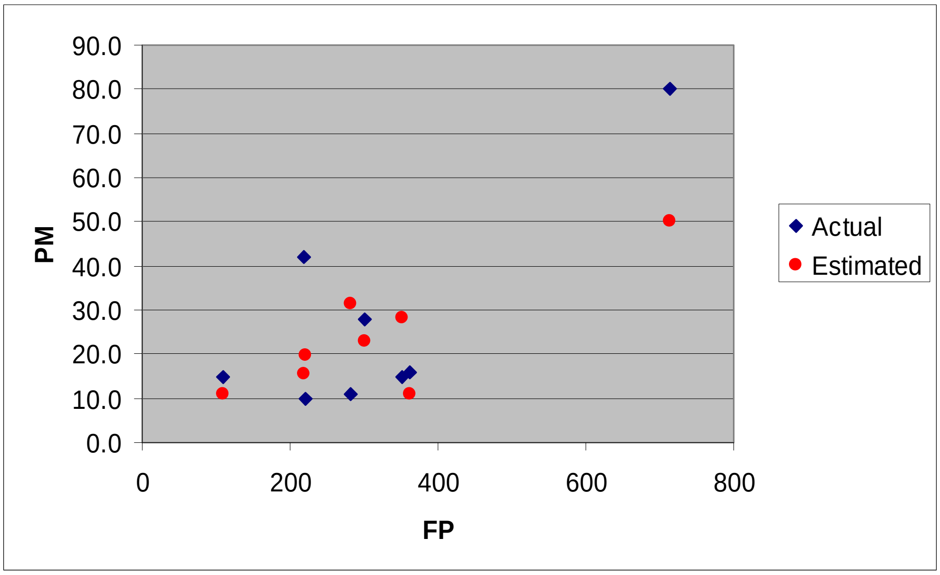

Figure 9 – Calibration Results
MRE = 68.24% – PRED(.30) = 25%
A = 2.54, B= 0.61

3.4.3. Analysis

A number of graphs were used to identify possible causes for the low value of PRED(.30). As an example, Figure 10 shows bar charts for Early Design effort multipliers RCPX, RUSE, PDIF, and PERS for the 8 selected projects. Light blue bars depict driver ratings, and dark red bars show percent error for each project. These charts and others were used to visually inspect driver ratings for potential errors.

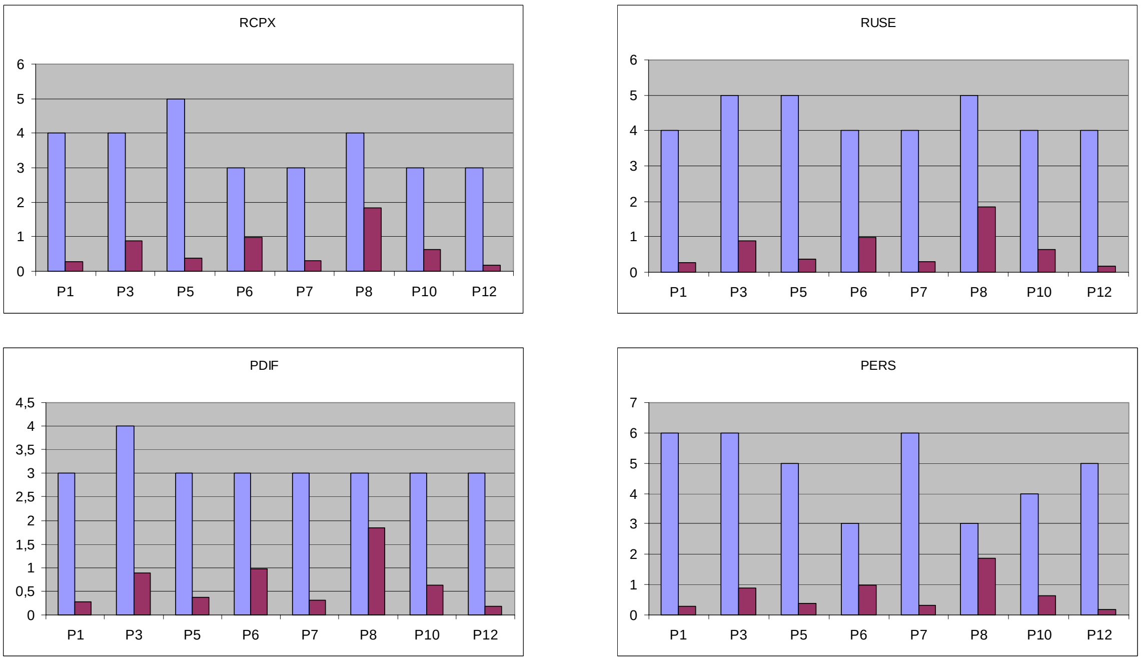

Figure 10 – Cost Driver Ratings and Percent Error per Project

Some potential causes for the low predictive level in this case were:
- some projects were interrupted and then resumed
- some projects had very small teams (like 1 individual working alone)
- inconsistent rating of the DATA effort multiplier
- some projects had the construction phase executed by a different organization
- some projects did not follow the organization's standard software process

The organization intended to investigate those possibilities, collect more data, and attempt another calibration.

3.5. Case Study: Organization E

Organization E had an estimation process and was interested in upgrading it to COCOMO II. Function point sizing and cost driver ratings were done by the client's team supported by consultants, as part of a training program.

3.5.1. Data Collection

This organization was able to find 7 completed projects available for the calibration effort. Because projects were chosen on an availability basis, they came from several platforms and technologies. Actual effort and schedule determination was based on existing documentation and few interviews with project managers. Each project was measured using the NESMA technique [10]. Scale factors and effort multipliers were determined for each project. All scale factors were kept at the nominal level. Figure 11 depicts the Effort Adjustment Factor (EAF) for each project.

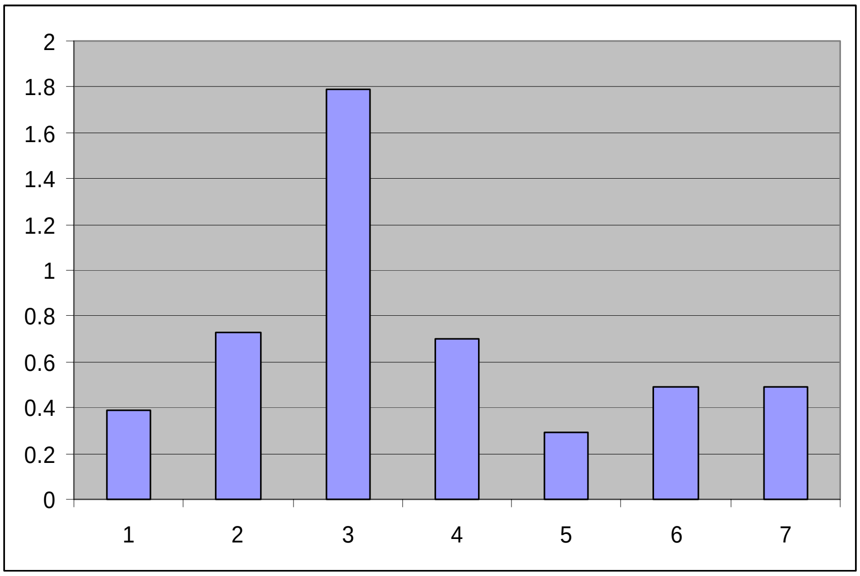

Figure 11 – EAF (Mean = 0.70; Std. Dev. = 0.51)

3.5.2. Model Calibration

Figure 12 depicts actual and estimated values for the calibrated model.

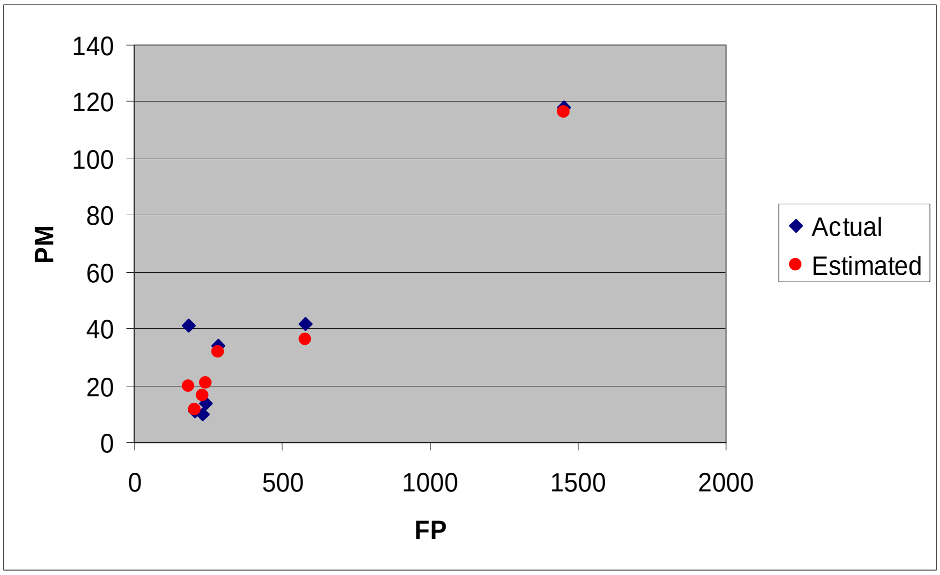

Figure 12 – Calibration Results
MRE = 27.42% – PRED(.30) = 57%
A = 19.92, B= 0.0473

3.4.3. Analysis

Also in this case, size in function points and COCOMO II cost drivers alone were unable to explain the variation. A single, large project may have strongly influenced the model. When compared to other calibrations, constant A assumed a very high value while B was very small. Further analysis will be required in order to refine the model.

4. CONCLUSIONS AND FUTURE WORK

4.1. Using Original COCOMO II Calibrations as a Baseline

COCOMO II original calibrations were used as a baseline to assess the calibrations obtained for the 5 organizations. The bar chart below (Fig. 13) allows one to compare each calibration in this study with both COCOMO II.1997 and COCOMO II.2000 calibrations. COCOMO II values displayed are for stratification, e.g., calibrating the model to each of the major sources of project data. Of the 5 organizations, 2 (A and B) were able to obtain PRED values comparable to COCOMO II.2000, 2 (C and E) obtained values between COCOMO II.2000 and COCOMO II.1997, and one (D) obtained a value below COCOMO II.1997.

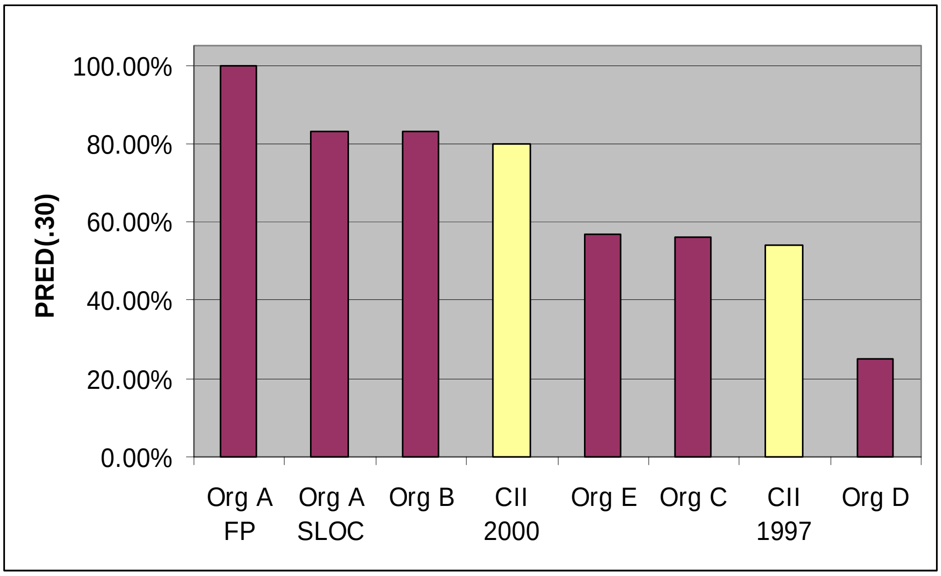

Figure 13 – Comparing Calibration Results

Calibrating two models for the same data using SLOC and function points was important to confirm to the client and to the team that comparable results could be obtained using either size measure with COCOMO II.

4.2 Calibration Difficulties, Lessons Learned, and Recommendations

One of the challenges in calibrating an estimation model is data collection. Obtaining a set of completed projects that can be measured is often difficult. Counting function points requires documentation and knowledgeable people who can help the counter interpret requirements. It is recommended that a simplified counting process such as NESMA's "estimated function point count" [10] be used in initial COCOMO II calibration projects. Initial calibration efforts will usually happen in a scenario of inaccurate information, so it may not be cost-effective to spend resources on exact function point counts. High-maturity organizations may require more accurate methods.

Another difficulty lies in obtaining actual values for effort and schedule. It seems that the preferred method for recording effort is staff-hours [3]. However, when documentation is not available it is usually easier to obtain effort in person-months (as used in the COCOMO II model) than in staff-hours. As a rule people will be able to remember who worked on a project, month-by-month. Staff-hours can then be derived from person-months as needed. Interviews can be more reliable than documentation, depending on the organization's policies. When interviewees are aware of data collection goals and sure that the information they provide will not be used against them they tend to reveal details not disclosed otherwise. Regular information channels sometimes lead employees to provide incorrect information. This kind of situation is extensively discussed in [9].

The determination of COCOMO II scale factors and effort multipliers requires special care. Because the rating of those drivers involves interpretation, consistency is very important in order to minimize subjective bias. Ideally, driver rating should be done by trained professionals. Even though some effort has been directed toward reducing driver rating subjectivity [4], [5], this is still a pending issue. One way of dealing with this problem is creating local driver rating criteria. For example, an organization may create formal, objective rules for rating the ACAP effort multiplier according to its HR policies. Another way of dealing with cost driver uncertainty is Monte Carlo simulation, whereby driver rating errors can be factored into the estimation process [13].

4.3. Future Work

Possibilities for future work include helping organizations to:
- add more projects to their baselines and recalibrate their COCOMO II models
- calibrate new models for other project categories
- create organization-specific cost driver rating scales
- explore new ways of grouping projects into categories for model building

4.4. Final Words

This paper has described some challenges, difficulties, and lessons learned in calibrating COCOMO II models for 5 Brazilian organizations. In addition, some results on the use of function points as a size measure with COCOMO II were provided.

5. REFERENCES

[1] B. Boehm, *Software Engineering Economics*. Prentice-Hall, 1981.
[2] B. Boehm, C. Abts, A.W. Brown, S. Chulani, B.K. Clark, E. Horowitz, R. Madachy, D. Reifer, and B. Steece, *Software Cost Estimation with COCOMO II*. Prentice-Hall, 2000.
[3] W.B. Goethert, E.K. Bailey, and M.B. Busby, *Software Effort & Schedule Measurement: A Framework for Counting Staff-hours and Recording Schedule Information* (CMU/SEI-92-TR-021). Software Engineering Institute, Carnegie-Mellon University, 1992.
[4] B.K. Clark, *COCOMO II Rating Scale Elaborations*. 18th International Forum on COCOMO and Software Cost Modeling, October 2003.
[5] B.K. Clark et al., *Elaboration Cost Drivers Workshop*. 18th International Forum on COCOMO and Software Cost Modeling, October 2003.
[6] R.D. Stutzke, *Estimating Software-Intensive Systems – Projects, Products, and Processes*. Addison-Wesley, 2005.
[7] C.F. Kemerer, *An Empirical Validation of Software Cost Estimation Models*. Communications of the ACM, May 1987.
[8] B.K. Clark, S. Devnani-Chulani, B. Boehm, *Calibrating the COCOMO II Post-Architecture Model*. 20th International Conference on Software Engineering, 1998.
[9] R.D. Austin, *Measuring and Managing Performance in Organizations*. Dorset House, 1996.
[10] NESMA, *Early FPA Counting*. Netherlands Software Metrics Users Association, http://www.nesma.nl/english/earlyfpa.htm, accessed in September, 2005.
[11] USC, *USC COCOMO II Version 2000.2 User Manual*. University of Southern California, 1999. Available in http://sunset.usc.edu/research/COCOMOII/cocomo_main.html#downloads, accessed in September, 2005.
[12] D. Ligett, *CALICO 7.0 Calibration Software*, SoftStar Systems, 2005. available in http://www.softstarsystems.com/calico.htm, accessed in September, 2005.

[13] P. McDonald, D. Strickland, and C. Wildman, *NOSTROMO: Using Monte Carlo Simulation to Model Uncertainty Risk in COCOMO II*. 18th International Forum on COCOMO and Software Cost Modeling, October 2003.
[14] IFPUG, *Function Point Counting Practices Manual Release 4.1.1*. International Function Point Users Group, 2000.