Modeling Uncertainty when Estimating IT Projects Costs

Michel Winter, Isabelle Mirbel, Pierre Crescenzo

LABORATOIRE

INFORMATIQUE, SIGNAUX ET SYSTÈMES
DE SOPHIA ANTIPOLIS
UMR7271

Modeling Uncertainty when Estimating IT Projects Costs

Michel Winter, Isabelle Mirbel, Pierre Crescenzo

Équipe MODALIS - WIMMICS

Rapport de Recherche
ISRN I3S/RR-2014-03-FR

Mars 2014

Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis (I3S) - UMR7271 - UNS CNRS
2000, route des Lucioles - Les Algorithmes - bât. Euclide B 06900 Sophia Antipolis - France
http://www.i3s.unice.fr

Modeling Uncertainty when Estimating IT Projects Costs

Michel Winter [1], Isabelle Mirbel [2], Pierre Crescenzo[3],

Équipe MODALIS - WIMMICS

ISRN I3S/RR-2014-03-FR
Mars 2014 - 4 pages

Abstract: In the current economic context, optimizing projects' cost is an obligation for a company to remain competitive in its market. Introducing statistical uncertainty in cost estimation is a good way to tackle the risk of going too far while minimizing the project budget: it allows the company to determine the best possible trade-off between estimated cost and acceptable risk. In this paper, we present new statistical estimators derived from the way IT companies estimate the projects' costs. In the current practice, the software to develop is progressively divided into smaller pieces until it becomes easy to estimate the associated development workload and the workloads of the usual additionnal activities (documentation, test, project management,...) are deduced from the development workload by applying ratios. Finally, the total cost is derived from the resulting workload by applying a daily rate. This way, the overall workload cannot be calculated nor estimated analytically. We thus propose to use Monte-Carlo simulations on PERT and dependency graphs to obtain the cost distribution of the project.

Key-words: IT Project ; Cost Estimation ; Monte-Carlo Method ; PERT

[1] Univ. Nice Sophia Antipolis, 06900 Sophia Antipolis, France – michel.winter@unice.fr
[2] Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France – WIMMICS (Inria Sophia Antipolis / Laboratoire I3S) - isabelle.mirbel@unice.fr
[3] Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France – MODALIS - pierre.crescenzo@unice.fr

Modeling Uncertainty when Estimating IT Projects Costs

Michel Winter
MIAGE,
Univ. Nice Sophia Antipolis,
1645 rte des Lucioles, 06900
Sophia Antipolis, France
michel.winter@unice.fr

Isabelle Mirbel
Univ. Nice Sophia Antipolis,
CNRS, I3S, UMR 7271, 06900
Sophia Antipolis, France
isabelle.mirbel@unice.fr

Pierre Crescenzo
Univ. Nice Sophia Antipolis,
CNRS, I3S, UMR 7271, 06900
Sophia Antipolis, France
pierre.crescenzo@unice.fr

ABSTRACT

In the current economic context, optimizing projects' cost is an obligation for a company to remain competitive in its market.

Introducing statistical uncertainty in cost estimation is a good way to tackle the risk of going too far while minimizing the project budget: it allows the company to determine the best possible trade-off between estimated cost and acceptable risk.

In this paper, we present new statistical estimators derived from the way IT companies estimate the projects' costs. In the current practice, the software to develop is progressively divided into smaller pieces until it becomes easy to estimate the associated development workload and the workloads of the usual additionnal activities (documentation, test, project management,...) are deduced from the development workload by applying ratios. Finally, the total cost is derived from the resulting workload by applying a daily rate.

This way, the overall workload cannot be calculated nor estimated analytically. We thus propose to use Monte-Carlo simulations on PERT and dependency graphs to obtain the cost distribution of the project.

1. INTRODUCTION

When developping software, cost estimation can be reduced to effort estimation since the project cost is derived by applying a daily rate.

The effort can be estimated in two ways: following an approach relying on expert judgment or an approach based on parametric models. Approaches based on expert judgment rely on a work breakdown structure; each task is sized based on an expert intuition and it is possible to model uncertainty by associating probabilistic distribution to each task estimate.

Concerning approaches relying on parametric models, many different types are available [2]. Even if model-based effort estimation processes may rely very much on expert judgment-based input, propagating uncertainty within the model becomes complicated.

In this paper we consider an hybrid approach, combining pure expert judgement for development task workload estimation and simple model-based estimators for other activities. Uncertainty are introduced wherever expert inputs are required and propagated to the global project effort estimation.

The paper is organized as follows: in section 2, we introduce a running example to illustrate our approach. Section 3 is dedicated to related work. In Section 4, we discuss the model we propose to propagate uncertainty into cost models. The way our model may be used on a concrete example is described in section 5. Finally, in section 6 we conclude the paper.

2. RUNNING EXAMPLE

We explain our approach along with a running example which is presented in the following.

A simple software has to be developped. An expert analysed the software requirements and identified two development tasks named Dvt_1 and Dvt_2. He estimated the associated development effort as follows:

- Dvt_1 should costs 6 ± 1 *man.days (md)*,
- Dvt_2 should also costs between 5 to $7md$ but it is more likely to be closer to 5 rather than 7.

In addition to the development activity, the project manager establishes that for such kind of project the test activity, Tst, represents 50% of the development effort $Dvt_1 + Dvt_2$, and the reporting activities, *Rpt*, represents $1md$ every week, during the whole project duration. If we consider that the two development tasks can be performed simultaneously, the sequence diagram modeling the project is shown in figure 1. Since we have an uncertainty on Dvt_1 and Dvt_2 duration,

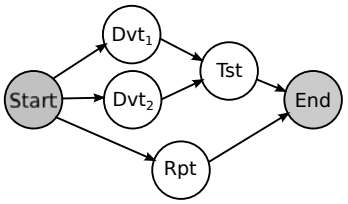

Figure 1: Task sequencing graph

and since activities Tst and *Rpt* depend on these development tasks, we need an approach to determine what is the best value to consider for the global effort and what is the resulting uncertainty.

3. RELATED WORK

3.1 Tasks and distributions

In project management, the classic way to introduce uncertainty in project cost-or-duration computation consists in considering each estimation as a random variable associated to a beta distribution [4]. The beta distribution is defined as follows:

$$f(x) = \begin{cases} a + (b-a)\left[\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,\mathrm{d}u}\right] & \forall x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

with: a, b the minimum and maximum of the distribution,
α, β defining the support of the function shape parameters

Modeling task Dvt_1 would lead to the following values: $a = 5$, $b = 7$, $\alpha = \beta = 2$. The associated distribution is illustrated in figure 2. Selecting $\alpha = \beta$ leads to a symmetrical

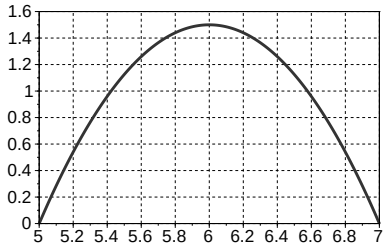

Figure 2: Beta distribution with $a = 5$, $b = 7$ and $\alpha = \beta = 2$

distribution centered around 6. It is also possible to model Dvt_2 for witch the workload is likely to be closer to 5 by setting β to 3 rather than 2, as shown in figure 3. Considering

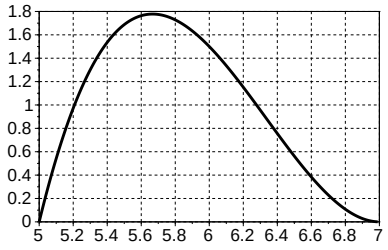

Figure 3: Beta distribution with $a = 5$, $b = 7$, $\alpha = 2$ and $\beta = 3$

a bigger project, each task i of the project can be modeled in the same way as a random variable X_i described by a beta distribution with parameters a_i, b_i, α_i and β_i. The workload of the whole project is defined as the sum of all these random variables. The workload can then be determined in two ways:

- using the Central Limit Theorem or
- applying Monte-Carlo simulations.

3.2 The Central Limit Theorem

The Central Limit Theorem (CLT) [11] states that the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed.
If we consider the random variable W associated to the project workload, defined as the sum of X_i random variables, we have:

$$W = \sum_{i=1..N} X_i \tag{1}$$

The CLT then states that W converges to a normal distribution whose mean $E(W)$ and variance $Var(W)$ are approximated to:

$$E(W) = \sum_{i=1..N} E(X_i) \tag{2}$$

$$= \sum_{i=1..N} a_i + \sum_{i=1..N} \frac{(b_i - a_i)\alpha_i}{\alpha_i + \beta_i} \tag{3}$$

and:

$$Var(W) = \sum_{i=1..N} Var\,(X_i) \tag{4}$$

$$= \sum_{i=1..N} \frac{(b_i - a_i)\alpha_i\beta_i}{(\alpha_i + \beta_i)^2(\alpha_i + \beta_i + 1)} \tag{5}$$

CLT is an efficient tool to manage uncertainty within the workload estimation as it is defined in (1), but it requires strict conditions [9]:

- the final random variable is a sum of initial ones,
- initial random variables must be identically distributed, with finite mean and variance,
- initial random variables are independent,
- the sum of random variables tends to be normally distributed as the number of these variables increases. In practice, the approximation is considered as acceptable for a sum of 30 random variables, even 50 if the distribution of random variables is asymmetric.

In the context described earlier, the second condition is not strictly fulfilled since different values for α and β lead to different distributions. An improvement to the CLT, named Lindberg's condition, states that the convergence to a normal distribution is also guaranteed if the second condition is omitted [8]. However, the first and third conditions are not fulfilled when considering our running example. Test and reporting activities are not terms that just have to be summed with development tasks and they are clearly correlated to the other tasks. CLT cannot be used.

3.3 Monte-Carlo experiments

The Monte-Carlo method is an alternative method allowing to compute complex combination of random variables. It relies on repeated random (statistical) sampling of elementary random variables. Simulations are run many times over in order to calculate the resulting probability heuristically. The algorithm is simple; the only complexity consists in generating samples according to associated distributions. It generally involves transforming a uniform random number in some way. Two methods are used [3]:

- the inversion method, that consists in integrating up to an area greater than or equal to the random number;
- the acceptance-rejection method, involves choosing x and y values according to a uniform distribution (the standard random function) and testing whether the distribution function of x is greater than the y value. If it is, the x value is accepted. Otherwise, the x value is rejected and the algorithm tries again.

This approach is a way to compute a numerical approximation which error can be lowered by increasing the number of generated samples: if we denote R the number of generated sampling sets, the error is reduced by a factor of $1/\sqrt{R}$ [7].

Since no condition is required unlike in the CLT approach, Monte-Carlo is clearly an approach that we can use to propagate uncertainty from expert inputs to the global effort. In order to achieve this, we are now going to elaborate the analytical model of the effort in accordance with the way it is expressed within the running example.

4. A MODEL COMBINING HETEROGENEOUS TASK ESTIMATORS

As explained previously, development tasks are modeled using beta distributions: the developer estimates what is the minimum and maximum time he would need to implement the associated feature (parameters a and b). Then he tunes the shape parameters α and β in order to reflect his feeling: will it be more likely close to the maximum b or to the minimum a.
Let's note X_i the associated random variables with $X_i \in \mathfrak{T}_\beta$, where $\mathfrak{T}_\beta$ corresponds to the set of tasks that are modeled using beta distributions.
When the effort associated to an activity is deduced by applying a ratio to a set of other tasks workloads, the corresponding random variable is modeled as follows:

$$X_j = f(a_j, b_j, \alpha_j, \beta_j) \sum_{X_l \in \mathfrak{T}} \delta_{j,l} X_l \tag{6}$$

where:
- $\delta_{j,l}$ is a boolean value that defines the set of random variables that are summed. $\delta_{4,8} = 1$ means that X_8 is used in the calculation of X_4.
- $\mathfrak{T}$ is the set containing all the tasks of the project.

Again, function $f()$ represents the beta distribution; it allows to model the fact that the ratio may also be uncertain. Considering our running example, let's name X_1,X_2 and X_3 the random variables associated to tasks Dvt_1, Dvt_2 and Tst respectively. We can write:

$$X_3 = f(0.4, 0.6, 2, 2)\,(X_1 + X_2) \tag{7}$$

$$= f(0.4, 0.6, 2, 2) \sum_{X_l \in \mathfrak{T}_\beta} \delta_{3,l} X_l \tag{8}$$

with $\delta_{3,1} = 1$ and $\delta_{3,2} = 1$. We also decided to add an uncertainty to the ratio of 50%; so we replace it by a beta distribution taking its values between 0.4 and 0.5.
Finally, for task *Rpt* we need to introduce another type of estimator: a ratio is applied to the duration of a set of tasks. It requires that the scheduling of the different tasks has already been elaborated. This schedule is defined as a directed graph $G_{sched} = (\mathfrak{T}, E)$ where $\mathfrak{T}$ is the set of node; each node represents a task of the project and E is a set of edges, i.e. pairs of $\mathfrak{T}$. The workload of each task is associated to each node. Finding the duration of a set of activities then consists in computing the longest path between the related start and ending nodes. Let's note $Dur(G_{sched}, s, e)$ the function that computes the duration between the two nodes s and e within the graph G_{sched}, we have:

$$X_k = f(a_k, b_k, \alpha_k, \beta_k) Dur(G_{sched}, s_k, e_k) \tag{9}$$

In our example, graph G_{sched} is illustrated in figure 1. If we note X_4 the random variable associated to task *Rpt*, we would write:

$$X_4 = f(0.15, 0.25, 2, 2) Dur(G_{sched}, Start, End) \tag{10}$$

Again, we added an uncertainty to the ratio of 20%; these 20% represent the effort of $1md$ per week estimated in our running example.

Our major contribution consists in the resulting global workload W that is synthesized below:

$$W = \sum_{X_m \in \mathfrak{T}} X_m \tag{11}$$

$$= \sum_{X_i \in \mathfrak{T}_\beta} f(a_i, b_i, \alpha_i, \beta_i) + \sum_{X_j \in \mathfrak{T}_{ratio}} \left[(a_j, b_j, \alpha_j, \beta_j) \sum_{X_m \in \mathfrak{T}} \delta_{j,m} X_m \right] + \sum_{X_k \in \mathfrak{T}_{dur}} [f(a_k, b_k, \alpha_k, \beta_k) Dur(G_{sched}, s_k, e_k)] \tag{12}$$

Where $\mathfrak{T}_{ratio}$ and $\mathfrak{T}_{dur}$ are the sets of random variables that are calculated according to model (6) and (9) respectively.

5. APPLYING MONTE-CARLO SIMULATIONS

The random variables are clearly no more independent and the project workload is now more complex than a sum a initial random variables. As a result, it is not possible to calculate the project workload using the CLT. The Monte-Carlo method has to be used. However, the project cost as expressed in (12) raises two difficulties described hereafter.

5.1 Tasks dependency

Random variables from $\mathfrak{T}_{ratio}$ and $\mathfrak{T}_{dur}$ define a dependency graph $G_{dep} = (\mathfrak{T}, E_2)$, where the arcs within E_2 are defined by the adjacency matrix Δ, gathering $\delta_{i,j}$ as defined in (6). The associated dependency graph is illustrated in figure 4.

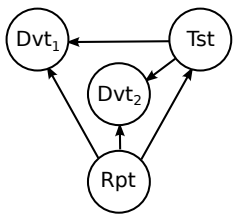

Figure 4: Dependency graph

To be able to compute W, it is necessary to derive an evaluation order that respects the given dependencies. Cycles within G_{dep} would lead to a situation in which no valid evaluation order exists, because none of the objects in the cycle may be evaluated first.
If it does not have any circular dependencies, the graph is a directed acyclic graph, and an evaluation order may be found by topological sorting. There are several algorithms that performs topological ordering; we used the one described in [6]. Considering our running example, the ordering $S = \{X_1, X_2, X_3, X_4\}$ is valid.

5.2 Tasks duration

For random variables from $\mathfrak{T}_{dur}$, we need to compute the duration of a group of tasks. In graph theory, it consists in computing the longest path between two nodes wihtin G_{sched}. Several algorithms answer this need but they all require edge-weighted graphs while we defined a vertex-weighted graph.
Therefore, G_{sched} has to be converted into its edge-weighted dual G'_{sched}. In the dual graph, the edges represent the activities, and the vertices represent the start or the finish of an activitiy. For this reason, the dual graph is also called an event-node graph.
The easiest way to tranform G_{sched} into G'_{sched} is to replace each node v in the original graph with two nodes, v_{in} and v_{out}, connected by a single edge with weight equal to the original vertex weight. Then the original edges (u, v) are replaced with edges from u_{out} to v_{in} of zero weight. Finally, when an original node v has only one predecessor, v_{in} is removed. Considering our example, the resulting dual graph is shown in figure 5.

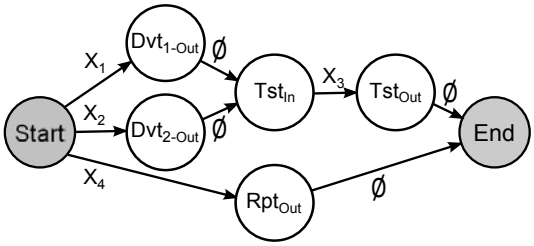

Figure 5: Scheduling dual graph

Since it is a directed acyclic graph, a longest path in the graph G'_{sched} corresponds to a shortest path in a graph $-G'_{sched}$ derived from G'_{sched} by changing every workload to its negation [10]. Longest paths in G'_{sched} can be found in linear time by applying a linear time algorithm for shortest paths in $-G'_{sched}$ such as [1, 5].

5.3 Running example

A Monte-Carlo simulation of 10000 turns has been performed on our running example; the repartition of the effort is given in figure 6. Using this distribution, we can easily tackle the uncertainty of our project. Depending on the way we want to manage the associated risk, two kind of questions can help the project manager in his decision making:

- What is the likehood that the project duration is for instance less than 19 m.d? In our example, the answer is **4%**

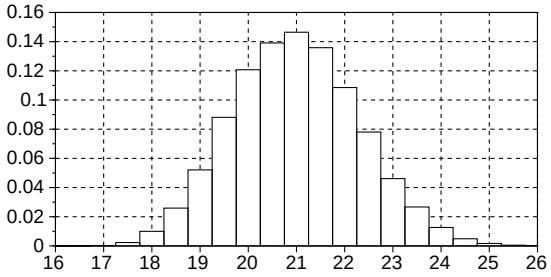

Figure 6: Distribution of the effort

- What is the maximal workload of the project, with a likehood of 90%? The answer is **23 m.d**

6. CONCLUSION

In this paper we proposed new statistical estimators for expert judgement-based effort estimation processes.
Because the global workload is a complex combination of random variables, it is not possible to calculate an analytical expression of its distribution. The propagation of uncertainty within the global workload model thus requires a numerical approach such as the Monte-Carlo simulation.
Even if the mean of the global workload could be easily determined without the help of Monte-Carlo simulation, our approach also provides an approximation of the workload distribution. This distribution allows the project manager to estimate more accurately the project cost and to better decide the maximum risk he can afford.

7. REFERENCES

[1] R. Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16:87–90, 1958.

[2] L. C. Briand and I. Wieczorek. Resource estimation in software engineering. *In J. J. Marcinak, Encyclopedia of software engineering*, pages 1160–1196, 2002.

[3] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[4] P. M. Institute. *A Guide to the Project Management Body of Knowledge: PMBOK Guide*. Newtown Square, Pennsylvania, 2013.

[5] L. R. F. Jr. *Network Flow Theory*. RAND Corporation, 1956.

[6] A. B. Kahn. Topological sorting of large networks. *Communications of the ACM*, 11(5):558–562, 1962.

[7] E. Koehler, E. Brown, and S. J.-P. A. Haneuse. On the assessment of monte carlo error in simulation-based statistical analyses. *Am Stat*, 63(2):155–162, 2009.

[8] J. W. Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 1(15):211–225, 1922.

[9] J. Rice. *Mathematical Statistics and Data Analysis (Second ed.)*. Duxbury Press, 1995.

[10] R. Sedgewick and K. D. Wayne. *Algorithms (4th ed.)*. Addison-Wesley Professional, 2011.

[11] G. W. Snedecor and W. G. Cochran. *Statistical Methods, Eighth Edition*. Iowa State University Press, 1989.