Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 9 (2013) 705 – 711

Procedia Technology

CENTERIS 2013 - Conference on ENTERprise Information Systems / PRojMAN 2013 - International Conference on Project MANagement / HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies

Advance data mining for Monte Carlo simulation in project management

Sergio Sebastián Rodríguez *

Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain

Abstract

It is well known the potential and the convenience of using Monte Carlo (MC) simulation to forecast projects' execution results in other to define and adjust its roadmap. MC simulation has great capabilities for project management solutions, and enough margins to be extended in terms of evolving the simulation model and offering better data mining of the results generated, in order to facilitate Project Managers (PM) labor. In this paper is expose some of the functionalities that are being studied in a wider research based on exploring advance capabilities of MC simulators for project management. For this purpose, a practical example is studied to show explicitly the benefits of the development.

Keywords: Monte Carlo, Simulation, Data mining, Scenarios, Context, Time, Cost, Critical Path

1. Introduction

In the past few years there has been a boom not only in the interest awakened by the MC method [1][2], but also in the various business solutions developed by various companies. There is no doubt about the possibilities that MC is able to offer to PM, but even though there are some horizons to be explored in the scope of the

* Corresponding author. Tel.:+34 626 384 572.
E-mail address: sergio.sebastian.rodriguez@gmail.com

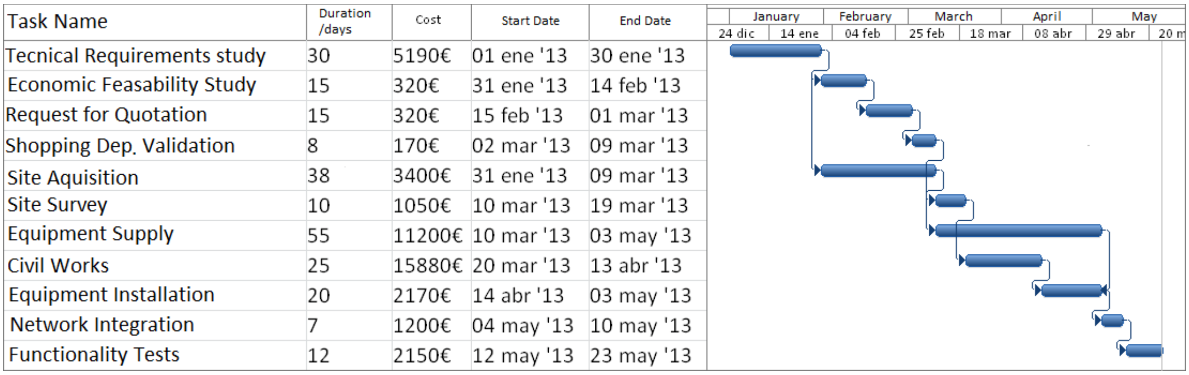
doi:10.1016/j.protcy.2013.12.078

simulation model inputs, such as deeper integration of costs and their interaction with the whole model, and the data mining of the generated results [3].

Regarding the second aspect studied, the data mining, in the framework of the ongoing research (*Exploration of advance capabilities of MC simulation for project management*) it has been observed that depending on the nature of the project simulated, the results given by the simulator should be processed and represent in a specific way. The reason for this is that although one of the key characteristics of MC simulation is its ability to consider the whole spectrum of possible outcomes, this can become one of its main weaknesses, as it can hide important and specific results due to overlapping the information in the global results metrics or graphs [4]. Does this mean that up to now PM have not been able to do useful simulations of their work? No, but this shortage of the simulation platforms offered in the software solutions imply to rerun different simulations modifying inputs in order to consider specific information needed. This problem is partially solved as this solutions typically allow to compare the information generated in various simulations, but this is far from the proper path, because this assumes that all the successive simulations run by the PM to find out particular information of relevant conditions (from now on scenarios) will be correctly proposed, and that the comparison or addition of the information obtained would lead to a correct interpretation [5]. To avoid this, efforts should be made to adapt the data mining of the results, defining when and how specific results processes must be considered. In the present paper we are going to expose the need of this functionality in order to achieve robust project simulation software, and the motivation of its development.

2. Exploring the needs of projects for the MC data mining

In first place, it will be exposed how adapting the data mining of the MC simulator can help the PM to obtain richer information for a better project planning. All the project's information, such as the schedule, tasks durations and costs, conditional and probabilistic events, and time and cost risks and constraints are simulated using a fully developed simulator for the research, based in MATLAB^{TM} (its matrix calculus potential is ideal for the task if the simulation model is done with this specific element arrangement [6]), which allows to integrate additional modules of any kind (inputs or data mining) to evolve the tool. The basis of its difference compared to present simulation platforms is that its implementation of modules and processes is completely orientated to generate and treat separately the data. This has been the main issue of its development, as robust simulators should have a wide variety of inputs, and possible interactions between them, that complicates the task of having the data disaggregated through all the simulation process.



Task Name	Duration /days	Cost	Start Date	End Date
Tecnical Requirements study	30	5190€	01 ene '13	30 ene '13
Economic Feasability Study	15	320€	31 ene '13	14 feb '13
Request for Quotation	15	320€	15 feb '13	01 mar '13
Shopping Dep. Validation	8	170€	02 mar '13	09 mar '13
Site Aquisition	38	3400€	31 ene '13	09 mar '13
Site Survey	10	1050€	10 mar '13	19 mar '13
Equipment Supply	55	11200€	10 mar '13	03 may '13
Civil Works	25	15880€	20 mar '13	13 abr '13
Equipment Installation	20	2170€	14 abr '13	03 may '13
Network Integration	7	1200€	04 may '13	10 may '13
Functionality Tests	12	2150€	12 may '13	23 may '13

Figure 1. Gantt chart containing the deterministic time and cost information [MS Project ^{TM}]

It is going to be studied an example based in a real project, the deployment of a radio base station of a telecommunications company. In the next figure we represent the project's tasks schedule, along with the cost of each task (according to the time in/dependant resources of each), driven by explicit time and cost risk factors [7] (based on the risk driver method) to be modeled as close to reality as possible.

The results obtained for the project's total duration and cost are shown below. These results show the most basic information given by the simulator, as it also specifies output data such as the probability density and cumulative distribution functions (PDF and CDF), cash flows, sensibility analysis, critical indexes, correlations and more; but they are not strictly needed for the purpose.

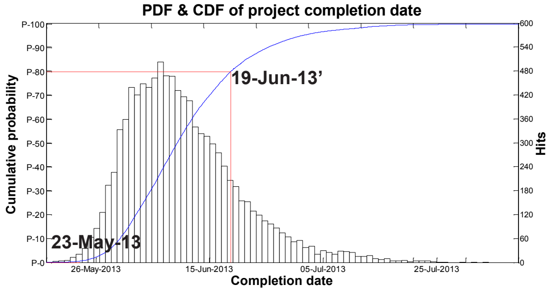


Figure 2. Results of the completion date

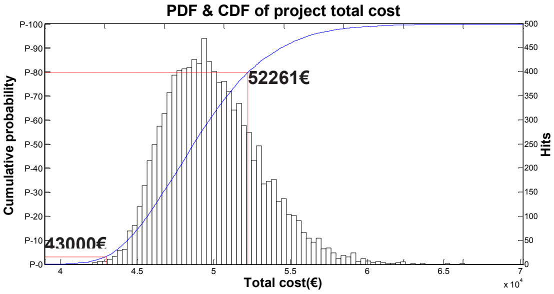


Figure 3. Results of the total cost

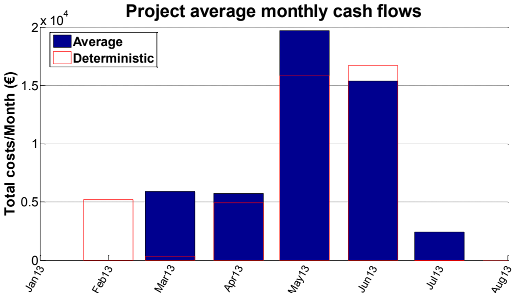


Figure 4. Results of the average cash flow per month

As shown above, figures 2 and 3 emphasize two metrics, the deterministic value and the P_{80} confidence level, which shows us how important is to consider uncertainty in our predictions. Apparently, we should not doubt about the quality of the results, considering that the simulation data supplied to the simulator has enough accuracy. But how much a PM should rely on this forecast? Even if we consider a conservative confidence level for the results such as P_{80}, we may be incurring not only in an underestimation of the risks, but also in a misunderstanding of what possible situations can be encountered in the project execution [8]. Where is this misunderstanding coming from, if a MC simulator with a robust model and accurate input information is all what a PM could desire for forecasting? The reason for this is that every result must be put into context.

Following this reasoning it is understandable that MC simulation contexts would be as wide as the proper project feasible roadmaps. This means that not necessarily always a PM should put in doubt the results given by the MC simulations, as long as it is not needed. What kind of context should require from the PM a deeper analysis of the given results? The desired answer for this would be a clearly defined set of conditions, but as said before, this is far away from being a bounded criterion, but there are some conditions that should make PM demand a further analysis of the information generated by the MC simulation [9].

To identify these situations, first we need to consider the possible simulation inputs that include the simulation model for a project. Typically these inputs are the tasks durations and costs PDF, schedule, duration and cost risks, conditional and probabilistic events, and time and cost constraints, whose existence or randomness could create the commented conditions. The inputs that could lead to require a deeper results analysis are: parallel paths in the schedule layout, that can cause uncertainty in the tasks finalization sequence or critical paths; schedule branching triggered by conditional events that manifest time and cost constraints, which could mean the execution of contingency plan; and probabilistic events, that model project events such as the success or fail of an acceptance test. This inputs, if they exist and are considered in the model when running a MC simulation of the project to be studied, can hide critical situations in global results like those shown in figures 2, 3 and 4; while they should be presented and analyzed specifically.

To illustrate this circumstance, consider a project that has a time constraint that would require a delay for the start of a particular task if its predecessor exceeds a determined date. Consider also that the delay required is considerable compared to the total project duration. The next figure shows the possible results:

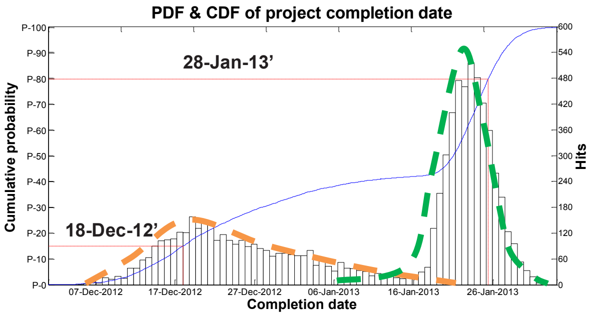


Figure 5. Project with two scenarios easily to distinguish in the PDF of the global completion date results

As shown above, the PDF of the project's total duration exhibits two well differentiated modes, which should be interpreted as the two possible scenarios able to come about in the project. These scenarios are highlighted with an orange contour line for the one that represents no delay in the execution, and the green contour line which represents the one where the time constraint forces to delay the beginning of a particular task, postponing the completion of the project. Even if the scenarios where not highlighted, there would probably not be any problems interpreting the results, as the key input that creates two differentiated scenarios is told to be considerable with respect to the project's total duration, which translate into an clear difference between the possible modes for the completion date, and therefore not requiring excessive data mining to interpret correctly the results given. How often these situations occur in a project, where critical circumstances are clearly interpretable within the whole bunch of results that MC simulation is able to offer? Indeed very occasionally, and inversely proportional to the complexity of the project simulated. This means that in order to develop a useful MC project simulator is vital to consider enough functionality that will alert and allow analyzing the global results to a deeper level whenever it is found to be necessary [10].

This functionality has been developed for the research simulator, and in the following block it is going to be tested to see how it can make PM reach a higher level of understanding in the required simulations.

3. Putting into practice the extended MC data mining

Once we have laid the foundations, the project exposed in figure 1 is going to be resumed to see how this extended functionality of the simulator can help the PM. As it was told above, considering the context of the project studied, the PM may not rely completely in the results given by the global results. As it can be seen, the schedule of the project in figure 1 contains elements like parallel paths, and its layout can take us to the need of disaggregate the global results into specific scenarios. The existence of parallel paths in the schedule logic does not necessarily mean that we should make special emphasis on every possible critical path, as it is not always significant which path would delay the merge point milestone, but as our simulated project model contains a large amount of parameters that affect the model elements dynamically, we should ensure that this do not lead to significant circumstances to consider in our project forecast.

The schedule layout in figure 1 can have four possible critical paths, and therefore the data mining of the simulation results is going to give specific data for each one, so that the PM can guarantee that the project planning can be adjusted to the occurrence of any of those possible critical paths, if needed.

The following figure will show each scenario's time, cost, and cash flow results:

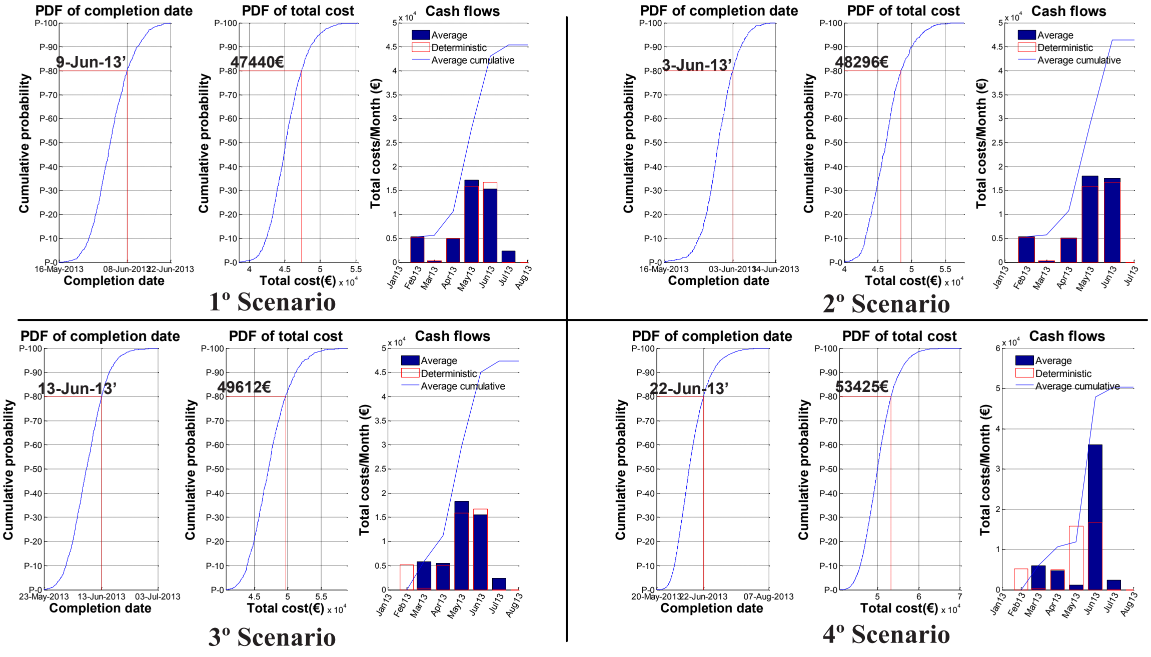


Figure 6. Results (date of completion, total cost and cash flow) obtained for the four scenarios of the project simulated

According to critical paths, the first scenario in figure 6 represents the one that has in its tasks the *Economic feasibility study*, the *Request for quotation (RFQ)*, and the *Shopping department validation* as critical for the first parallel path (or merge point of the schedule), and the task of *Equipment supply* being critical for the second parallel path. The second scenario has the same critical path for the first merge point, but has the tasks of *Site Survey*, *Civil works* and *Equipment installation* being critical for the second merge point. The third and fourth scenarios have the task of *Site acquisition* as critical for the first merge point. The third scenario has the same critical path for the second merge point as the first scenario, while the fourth shares the critical path of the second merge point with the second scenario described.

At first sight, it can be seen that although there are a few similar results for the cost metric, there is a reasonable difference between the four possible scenarios present in the project if we take into account all the metrics (time, cost and cash flow) that completely define them.

The results show that the first scenario raises a distribution of the cash flow, and therefore of the execution of the tasks, different from the global results of figure 4, as well as determining an earlier completion date, and a reduction of the project's budget of around a 10%. The results for the second scenario show also a smaller budget needed (not as much as the first), but particularly throws the earliest completion date of all the possible ones, with more than two weeks of difference with respect to the global results; and a cash flow arranged similarly as the deterministic estimate. The third scenario determines a smaller budget as the previous two, with a later estimate of the completion date (but earlier compared to the global results), but agrees with the global cash flow shown in figure 4. Finally, the fourth scenario would be the one that would alert a PM when analyzing these results, mainly because all the metrics raised are the least desirable compared to the other scenarios, but also to the global results, which means that only relying on figures 2 and 3 results could lead to an underestimation of the risks and uncertainties. Besides a larger budget needed, it is important to take a look to the foreseeable cash flow, which describes an considerable expenditure in the month of June, something that should be taken into account from the corporative perspective, as this kind of situations must be considered in the financial forecast in order to have enough capital available for all the current projects taking place, as well as other concerns such as fiscal issues or any kind of planned investments.

Shall we expect such a difference in our project scenarios when there are situations like various possibilities for the critical path? Not necessarily, the reason for this notorious difference between the scenarios remains in the inputs that define the project. It is possible for us to reckon quite easily the date of completion having the information of the PDF of each task and the schedule logic, but, what we don't take into account is that the dynamism of the MC simulation lead to the interaction of various model parameters that, in a specific critical path can drive the project's results away from the global estimates. Indeed, that is what exactly occurred for the fourth scenario, due to its nature the critical path that it describes forces certain parameters to be driven to values that result in such results. In this case, the longer duration of the tasks of *Site Survey*, *Civil works* and *Equipment installation* implicitly mean that the existing risk of *Requirements variation* is especially high and active, and as it jeopardize a wide range of tasks of the project, it will increase all of those tasks durations, and therefore a wide range of time dependant resources, raising the total cost and duration of the project.

After the previous analysis, the PM has very precise information of the possible situations that may occur, designing specific planning strategies like contingency plans that can make the difference to achieve the project success. It should be noted that these results shown are just a part of the information that the simulator developed offers, and should be complemented with the sensibility analysis, correlation results, risk prioritization algorithms, critical indexes or resource allocation, among others; all of them specific for each scenario. In the example, a PM would probably desire to adjust the roadmap as much as possible to the second scenario, which represents one of the best time-cost metrics. It is important to keep in mind that just making efforts for executing a project through a determinate critical path would be far from ensuring a proper planning; this tool offers just a useful hint for such a complicated labor like project management is.

4. Conclusions

As it has been stated, although project management MC simulators have a great potential for estimating projects forecast, they should be complemented with additional functionality in order to obtain the maximum advantage from them. The examples exposed show that global results may incur not only in underestimation, but also in a misunderstanding of the forthcoming events, as most times these are results made up of various scenarios. Also, the PM should have the enough criterion to decide when project inputs lead to situations that would be useful to analyze specifically, selecting the appropriate data mining parameters in the simulator. As an extended part of the research, is being developed a full automated decision process that delivers these specific results whenever required, based in inputs and simulation results, define by customizable thresholds.

From the point of view of the software development, the simulator modules and internal processes must be orientated to treat separately the information demanded by the PMs. This is a delicate issue, as the existence of any king of input, and the possible interactions between them, should be considered explicitly in the simulator data model, processing, and results evaluation modules.

Acknowledgements

The author wishes to express appreciation to the Universidad Autónoma de Madrid (U.A.M.) and to the project management team of the backbone network department of Vodafone ^{TM} Spain.

References

[1] Project Management Institute, 2009. “A guide to the Project Management Body of Knowledge: (PMBOK guide)”, Project Management Institute.
[2] Reuven Y. Rubinstein, 2008. “Simulation and the Monte Carlo method”, Wiley-Interscience.
[3] Dragan Z. Milosevic, 2003. “Project management toolbox: tools and techniques for the practicing project manager” John Wiley & Sons.
[4] Johnathan Mun, 2006. “Modeling risk: applying Monte Carlo simulation, real options analysis, forecasting, and optimization techniques”, John Wiley & Sons.
[5] Vose, D., 2008. “Risk analysis: a quantitative guide”, Wiley.
[6] Huu Tue Huynh, 2008. “Stochastic simulation and applications in finance with MATLAB programs”, John Wiley & Sons.
[7] Hullet, D., 2011. “Integrated Cost-Schedule Risk Analysis”.
[8] Laurent Condamin, 2006. “Risk quantification: management, diagnosis and hedging”, John Wiley.
[9] Yacov Y. Haimes, 2009. “Risk modeling, assessment, and management”, Wiley.
[10] Stephen Grey, 1995. “Practical risk assessment for project management”, Wiley.