Relatório Técnico

First Steps in Mining Resource-Constrained Project-Scheduling Stochastic Models for New Valuable Insights in Project Management

A. J. Alencar
G. G. Rodrigues
E. A. Schmitz
A. L. Ferreira
S. J. Mecena

NCE - 03/06

Núcleo de Computação Eletrônica

Universidade Federal do Rio de Janeiro

First Steps in Mining Resource-Constrained Project-Scheduling Stochastic Models for New Valuable Insights in Project Management

Antonio Juarez Alencar[1], Gelson Guedes Rodrigues[1], Eber Assis Schmitz[2], Armando Leite Ferreira[3] and Sérgio José Mecena[4]

[1]Institute of Mathematics, Federal University of Rio de Janeiro,
P.O. Box 68530, 21941-590 - Rio de Janeiro - RJ, Brazil

[2]Electronic Computer Center, Federal University of Rio de Janeiro,
P.O. Box 2324, 20001-970 - Rio de Janeiro - RJ, Brazil

[3]The COPPEAD School of Business, Federal University of Rio de Janeiro,
P.O. Box 68514, 21945-970 - Rio de Janeiro - RJ, Brazil

[4]Production Engineering Department, Fluminense Federal University,
Rua Passo da Pátria nº 156, 24210-240 - Niterói - RJ, Brazil

juarezalencar@br.inter.net, gelsongr@gmail.com,
eber@nce.ufrj.br, armando@coppead.ufrj.br, smecena@globo.com

Abstract. *In this work we show how classification trees, a family of non-parametric statistical methods, can be used together with RCPS (Resource Constrained Project Scheduling) stochastic modeling and simulation to provide project managers with better insights into the project they run. Such insights make it easier for managers to anticipate changes in planning that favor projects to be delivered on time, within budget, and in compliance with available cash flow and the requirements they were set to satisfy. Also, we discuss the implications of these insights for both the management of complex projects and the construction of effective business strategies.*

Resumo. *Neste artigo descrevemos como árvores de classificação, uma família de métodos estatísticos não-paramétricos, pode ser utilizada juntamente com modelagem e simulação estocástica para o escalonamento de projetos com recursos escassos (RPCS) com vistas a fornecer aos gerentes um entendimento mais profundo dos projetos sobre sua responsabilidade. Este entendimento facilita a implementação de mudanças no planejamento das atividades, favorecendo a entrega de projetos dentro do prazo, de acordo com o orçamento e fluxo de caixa, e em sintonia com os requisitos que se comprometeram a satisfazer. Em adição discutimos as implicações das técnicas aqui apresentadas para a gerência de projetos de grande complexidade e para a construção de estratégias de negócio.*

1. Introduction

Over the last few decades the environment in which organizations do business has changed considerably with enormous consequences for project management. The business paradigms that prevailed during the industrial revolution are giving way to new ones dictated by the information age, knowledge age and technology revolution which we are currently experiencing [Jones et al. 2002]. For example, the existence of rigid production lines of impersonalized tangible products that characterized the industrial revolution are being successfully challenged by increasing demand for highly customized products and services, decentralized work force and intangible products, opening new vistas for industrial and business development [Hosni and Khalil 2004]. The cell phone communication, computer network, TV on demand and distance learning industries are just a few examples of such vistas [Wagner 2005].

The increasing competition for market space in many industries worldwide has become an important societal force of this time and age, putting extreme pressure on organizations to make their complex, customized outputs available as quickly as possible. Nowadays, response must come faster, decisions must be made sooner and results must occur more quickly. Hence, time-to-market has become a critical success factor in several areas of business [Meredith and Mantel 2002]. On top of that, for companies to be competitive, they have to reduce their costs and focus on satisfying their customers [Frame 2002]. If you do not deliver the products and services your customers value, for the price they are willing to pay, someone else will.

Market pressure and the growing demand for new and more complex products and services have required organizations to practice creative destruction on a non-stop basis by replacing old ways of doing business in order to create new ways [Hamill 2000, Arora et al. 2001]. However, this process of destruction and construction requires all sort of different specialized knowledge, e.g. marketing, emerging technologies, finance, logistic, business strategy, etc. Therefore, although invisible, knowledge has emerged as one of the most strategic assets for organizations [Kakabadse et al. 2001].

All of this has forced organizations worldwide to increase both the quantity and complexity of projects they run. To remain competitive organizations have to innovate constantly. The introduction of every innovation in products and services requires the execution of one or, more frequently, several projets. As the pace of innovation increases, so does the number of projects. On the other hand, the demand for new and more complex products calls for the execution of more complex projects; requiring the coordination of efforts from multidisciplinary teams with advanced technical and business skills, the establishment of strategic alliances with external partners, the outsourcing of project activities, and the use of recently developed technology.

Nevertheless, short time-to-market and fierce competition for market space require projects to be executed faster, with tighter budgets and less margin for errors. Undoubtedly such constrains tend to put projets managers under enormous pressure to produce results. When the pressure to deliver becomes too great, common sense often goes out the window, crucial steps in the project development are ignored, the end result is shoddy, and the rework required to repair the damage is far more expensive than if it were caught in the planning stages [Info-Tech Research Group 2003].

Effective project management in a lean resource environment requires good planning and timely information, allowing problems to be anticipated and dealt with before the worse happens[Wright et al. 1999]. With the view of antecipating problems that frequently prevent projects from finishing successfully, over the years, managers have resorted to planning methods such as Gantt Charts, CPM (Critical Path Method), PERT (Performance Evaluation and Review Technique) and, more recently, to RCPS (Resource Constrained Project Scheduling) stochastic modeling and simulation [Lewis 2000, Kerzner 2003].

Although the stochastic modeling and simulation of RCPS problems has proved to be a powerful tools for project planning in constrained environments, the construction and analysis of RCPS activity networks require considerable experience in quantitative model building and advanced knowledge of both mathematics and statistics.

This work shows how classification trees, a family of non-parametric statistical methods, can be used together with RCPS stochastic modeling and simulation to provide valuable insight into project planning in constrained environments, making it easier for managers to foresee changes in planning that favor projects to be delivered on time, within budget and in compliance with available cash flow and the requirements they were set to satisfy. Also, we discuss the implication of such insights for both the project management of complex projets and the construction of effective business strategies.

2. Conceptual Framework

One of the main goals of project scheduling is to produce a detailed plan of project related activities with the view of allowing managers to deal with a large variety of problems before the worse happens. Among these problems one may find lack of specialized or experienced labor, incompatible start and finish dates among activities, insufficient funding, activities being executed in sequence when they should be executed in parallel, unavailability of equipment, etc. However, the most frequent problem managers use project scheduling to solve is the minimization of makespan[1], other common possibilities include minimization of cost, maximization of financial results, maximization of quality measures, etc. [Zhu et al. 2005].

2.1. Resource-Constrained Project Scheduling

Project schedules that are subjected to precedence and resource constraints are called "Resource-Constrained Project Scheduling" in the literature, or RCPS for short. Not surprisingly, the vast majority of projects in the real world face RCPS problems. Hence, the great amount of attention that the academic world and industry have been paying to efforts in dealing efficiently with RCPS problems.

Over the years numerous methods have been proposed to solve RCPS problems, such as: implicit enumeration, branch-and-bound procedure, schedule generation schemes (SGS), X-pass, etc. A comprehensive review of these methods is found in [Demeulemeester and Herroelen 2002, Kolisch and Hartmann 2006]. Despite the differences these methods may have, they can all be classified in just two categories: exact methods and heuristic methods [Castillo and Muñoz 2004].

Exact methods are used to find the precise maximum or minimum value of a variable under consideration. For example, the minimum project duration or its maximum financial return. However, these methods have considerable limitations that become evident when projects have a great number of activities or complex resource constraints.

[1] Total project duration.

Usually, in these circumstances a solution cannot be found within reasonable computing time. Heuristic methods, on the other hand, use a schedule priority criteria to provide an approximate solution to RCPS problems. They are particularly useful when the use of exact methods are not feasible.

In the heuristic methods every activity is initially a potential candidate to be scheduled. However, activities can only be scheduled if all its precedent activities have been completed and the resources required for its execution are fully available. Unfortunately, whenever activities share resources, it may imply that they cannot be executed concurrently. The use of a heuristic is then necessary to decide when each activity should receive its required resources and be executed as a result. When no real candidate activities are available, the clock advances until one of the activities in progress is completed. At that point in time resources are freed and the process is repeated, i.e. candidate activities are identified, resources are checked for availability and activities are scheduled. The total duration of a project is the time required for the completion of all its activities.

The following are examples of heuristics frequently used to deal with RCPS problems: shortest processing time (SPT), most immediate successors (MIS), late start time (LST), first come first served (FCFS). A thorough account of these heuristics is provided by [Goerlich and Olaguibel 1989]. As a mathematical problem RCPS belongs to a category where, in general, an optimal solution cannot be found in a linear or polynomial time frame, i.e. the NP-hard category [Blazewicz et al. 1983].

2.2. Classification Trees

The techniques described in this article to analyze RCPS stochastic models make extensive use of classifications trees, a class of statistical inferential methods conceived by Morgan and Sonquist in the 1960's [Morgan and Sonquist 1963] and later perfected by others, such as: [Kass 1980], [Breiman et al. 1984], [Quinlan 1992], [Loh and Vanichestakul 1988] and [Loh and Shih 1997].

As an inferential method, classification trees seek to explain the behavior of a target variable by combining different values of a given set of predictive variables. To achieve this goal, the values of the predictive variables are successively combined in such a way that an n-dimensional space is portioned into increasingly more homogenous sets of values with respect to the target variable. The way the portioning is done allows the final result to be presented as a flow-chart whose format resembles a tree, so the name of the family of methods. Furthermore, the information presented in the flow-chart may be easily translated into a set of rules that indicate the likelihood of occurrence of values in the domain of the target variable for different circumstances.

Classification trees are non-parametric methods, i.e. no restrictions are imposed on the distribution of values of predictive and target variables. In addition, these variables are allowed to hold all sort of relationships among themselves. All of this makes it easier to use classification trees to solve problems in the real word, where the distribution of values of variables and the relationships that they hold among themselves are frequently unknown. Therefore, not only classification trees are a family of robust methods, but the way the results are presented makes understanding easier and facilitates their dissemination among all interested parties.

Moreover, classification trees have been widely reported as presenting satisfactory results even in the presence of noise and when little data is available; making it a very attractive class of methods to be used in all sorts of different situations [Witten and Frank 2005].

In formal terms, given a set of observations $\mathcal{O} = \{o_1, o_2, \cdots, o_m\}$, where every $o_{j \in [1..m]} \in \mathcal{O}$ is a tuple $(x_1, x_2, \ldots, x_n)$ and each tuple component $x_{i \in [1..n]}$ may take value in a different domain, then let $v_{i \in [1..n], j \in [1..m]}$ be the value of the tuple component x_i of o_j. In these circumstances, a variable $w_{i \in [1..n]}$ in $\mathcal{O}$ is an undefined value of the set $\{v_{i,1}, v_{i,2}, \ldots, v_{i,m}\}$. Moreover, let w_n be the target variable, whose behavior one hopes to explain by combining the values of $w_1, w_2, \cdots, w_{n-1}$. Also let w_n take value in a finite set J.

In the classification tree paradigm the set of variables in $\mathcal{O}$ is exhaustively examined for relations of the form:

$$\begin{array}{ll} x_i \leq r, & \text{if } x_i \text{ is a continuous variable} \\ \text{or} & \\ x_i \in \{k_1, k_2, k_3, \cdots, k_l\}, & \text{if } x_i \text{ is a discrete variable.} \end{array}$$

where r and $k_{j \in [1..l]}$ belong the domain of x_i. Each relation $R_{x_i \leq r}$ or $R_{x_i \in \{k_1, \cdots, k_l\}}$ that is found allows $\mathcal{O}$ to be partitioned into two sets, i.e. $\mathcal{O}_R$ and $\mathcal{O}_{\neg R}$, such that $\mathcal{O}_R \cap \mathcal{O}_{\neg R} = \{\}$ and $\mathcal{O}_R \cup \mathcal{O}_{\neg R} = \mathcal{O}$. In this case, $\mathcal{O}_R$ contains the observations in $\mathcal{O}$ for which R holds and $\mathcal{O}_{\neg R}$ the observations for which the negation of R holds.

A metric M is then used to evaluate how diverse the data in $\mathcal{O}$, $\mathcal{O}_R$ and $\mathcal{O}_{\neg R}$ are regarding the elements in J. The difference between $M(\mathcal{O})$ and the weighted average of $M(\mathcal{O}_R)$ and $M(\mathcal{O}_{\neg R})$ indicates the contribution of relation R for the reduction in the diversity of the elements in J as a result of partitioning $\mathcal{O}$ into $\mathcal{O}_R$ and $\mathcal{O}_{\neg R}$. Because one wants diversity to be reduced as fast as possible, the relation R that provides the biggest reduction in diversity is initially used to partition $\mathcal{O}$. The process is then successively reapplied to $\mathcal{O}_R$ and $\mathcal{O}_{\neg R}$, until no gain in the reduction of diversity is achieved.

In the classification tree paradigm, when J has only two elements, one of the most frequently used diversity metric is the Gini diversity index. The index was initially proposed by Corrado Gini [Gini 1939] and later adapted by [Breiman et al. 1984] for the development of classification methods. In formal terms, for a given set of observations $\mathcal{O}$, Gini is calculated as

$$I(\mathcal{O}) = 1 - S, \tag{1}$$

where $S = \sum_{i=1}^{|J|} P(j_i|\mathcal{O})^2$ for $j_i \in J$, and $P(j_i|\mathcal{O})$ is the probability of occurrence of objects j_i in $\mathcal{O}$. In these circumstances the reduction in diversity provided by a relation R is given by

$$\Delta I(\mathcal{O}) = I(\mathcal{O}) - (I(\mathcal{O}_R) \times p_R + I(\mathcal{O}_{\neg R}) \times p_{\neg R}) \tag{2}$$

where p_R and $p_{\neg R}$ are respectively the proportion of elements in O_R and $O_{\neg R}$.

3. Mining RCPS Stochastic Models with Classification Trees

According to Seneca (4 BC-AD 65), the Roman philosopher:

> "Rules make the learner's path long, while examples make it short and successful".

As a result, the mining process presented in this article is introduced with the help of a real-world inspired example.

Suppose that a chain of furniture stores decides to create a catalog to promote some of the products they sell to a large group of prospects at a special price. The proper undertaking of this task requires that eight interdependent activities are efficiently executed within a time frame, i.e.

1. *Product Selection* - that chooses the products that will be advertised in the catalog;
2. *Prospect Selection* - that identifies the prospects to whom the catalogs are going to be mailed;
3. *Pricing* - that establishes the promotional price of every product to be advertised in the catalog;
4. *Catalog Design* - where the graphic and textual aspect of the catalog, and accompanying advertising material are conceived and put together;
5. *Label Printing* - where labels with prospects' names and addresses are printed and organized;
6. *Stock Control* - that makes sure that the products advertised in the catalog will be available for shipping when they are ordered ;
7. *Catalog Printing* - where the actual print of the catalog is done;
8. *Catalog Labeling & Mailing* - which labels the catalogs with prospects' names and addresses and sends them to their intended destinations over the mail.

Tough competition in the furniture business has brought down profit margins over the years, as a result the chain of furniture stores operates with a skinny employee structure. Therefore, only two people have been selected to work on the catalog project. They are going to be named *Mimi* and *Ed*. Table 1 shows the human resources required by each of the project's activities.

Table 1. Resources required by each of the project's activities.

Activity		Resource Required
Label	**Description**	
PdS	Product Selection	Mimi and Ed
PsS	Prospect Selection	Mimi
P	Pricing	Mimi and Ed
CD	Catalog Design	Ed
LP	Label Printing	Mimi or Ed
SC	Stock Control	Mimi or Ed
CP	Catalog Printing	Mimi or Ed
CLM	Catalog Labeling & Mailing	Mimi and Ed

Figure 1 presents the project's network of activities. In that figure *Product Selection* is the first activity to be executed and *Catalog Labeling and Mailing* the last. Moreover, an arrow connecting two activities such as *Product Selection* ⟶ *Pricing* indicates that the latter may only start when the former has been completed and all the necessary resources are available.

As the project has to be completed within a time frame, it is crucial that management is made aware of its expected makespan in respect to the current resource constraints

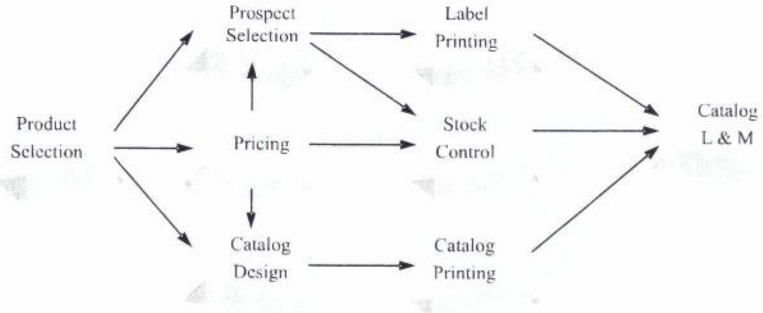

Figure 1. The project's network of activities.

and the duration of the different project activities. However, because these activities have not been executed yet, their duration can only be estimated. In this case, with the support of other experienced project managers and a database of previously executed projects, a three point estimate has been established for each activity, indicating their minimum, most likely and maximum expected duration. Table 2 presents these figures.

Table 2. Minimum, maximum and most likely duration of project activities.

Activity	Activity Duration		
	Minimum	Most Likely	Maximum
PdS	2	3	6
PsS	3	6	8
P	4	6	7
CD	1	8	10
LP	7	9	11
SC	5	6	7
CP	4	5	8
CLM	1	2	4

Considering that the exact duration of each activity is unknown and that the available resources may be insufficient to ensure that all activities are executed when their precedents are completed, a stochastic simulation model has been built to analyze the project makespan. In this model, the duration of each activity is described by a triangular probability density function, one of the most widely used functions to describe activity duration [Chung 2003].

Subsequently the stochastic model was subjected to a simulation process, where resources were granted to the first activity to place a request for them. During simulation the duration of each activity was recorded in the form of a percentage of their respective time range. Also, the activity waiting time for resource allocation was recorded together with the outcome of each simulated scenario, indicating whether the project finished within the allowed time. Table 3 presents the figures collected during simulation. Table 4 describes the meaning of the columns presented in Table 3. See [Chung 2003] for an introduction to simulation.

For example, as indicated by variables Scn and CD, in the first simulated scenario the execution of *Catalog Design* consumed 25.4% of its time range, i.e., $1+0.254 \times (10-1) = 3.3$ time units. Moreover, the variable L_{CD} shows that the necessary resources for its execution were being used by other activities at the time they were requested and that 4.1 time units passed until they were released. It should also be noted that only two activities had their waiting time for resource allocation registered. This is due to the fact that no other activities had to wait for the allocation of resources they needed.

Figure 2 shows the project makespan classification tree built with the help of the SPSS software (www.spss.com), using the data generated during the simulation process and the CRT option, an enhanced variation of [Breiman et al. 1984] approach to classification. In that figure the box labeled *Node 0* is the root of the classification tree. It contains the total number of observations available for analysis, i.e. 5,000 in this case. Also, it displays the number and proportion of the different scenarios in which the catalog project finished and did not finish on time.

For example, initially, in 60.7% of the generated scenarios the project finished on time, whilst in 39.3% it did not. It should be noted that the relation that is used to partition the initial set of scenarios is $L_{LP} \leq 4.8$, and that, as a consequence, *Node 1* contains the 4,110 observations for which this relation holds and *Node 2* the remaining 890.

Because of the incremental way in which the tree is constructed, all relations that hold for the observations in a node also hold for the the observations in its descendants nodes. For example, in the scenarios that are part of *Node 3* the label-printing waiting time for resource allocation is smaller than or equal to 4.8 time units ($L_{LP} \leq 4.8$) and

Table 3. Simulation results connecting the duration of project activities and waiting time for resource allocation with the project's outcome.

Scn	*PdS*	*PsS*	*P*	*CD*	*LP*	*SC*	*CP*	*CLM*	L_{CD}	L_{LP}	Rst
1	44.4	88.3	1.2	25.4	24.5	36.6	36.8	10.9	4.1	0.0	In
2	80.6	13.6	70.4	67.6	14.7	16.1	24.1	76.9	0.0	3.5	In
3	34.8	78.7	0.9	13.8	32.9	66.1	49.3	21.2	4.1	0.0	In
4	40.7	84.2	80.5	29.2	64.5	20.7	84.2	81.6	0.0	4.4	Out
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	38.5	92.2	27.3	84.6	45.2	49.1	85.8	31.3	0.0	4.6	Out

Table 4. Meaning of the variables whose values were collected during the simulation process.

Variable	Meaning
Scn	Scenario identifier.
$PdS, PsS, \cdots, CLM$	Duration of activities $PdS, PsS, \cdots, CLM$ expressed as a percentage of their respective time range.
L_{CD} and L_{LP}	Waiting time for resource allocation of activities CD and LP.
Rst	The outcome of a given scenario, where In indicates that the project finished within the allowed time and Out indicates otherwise.

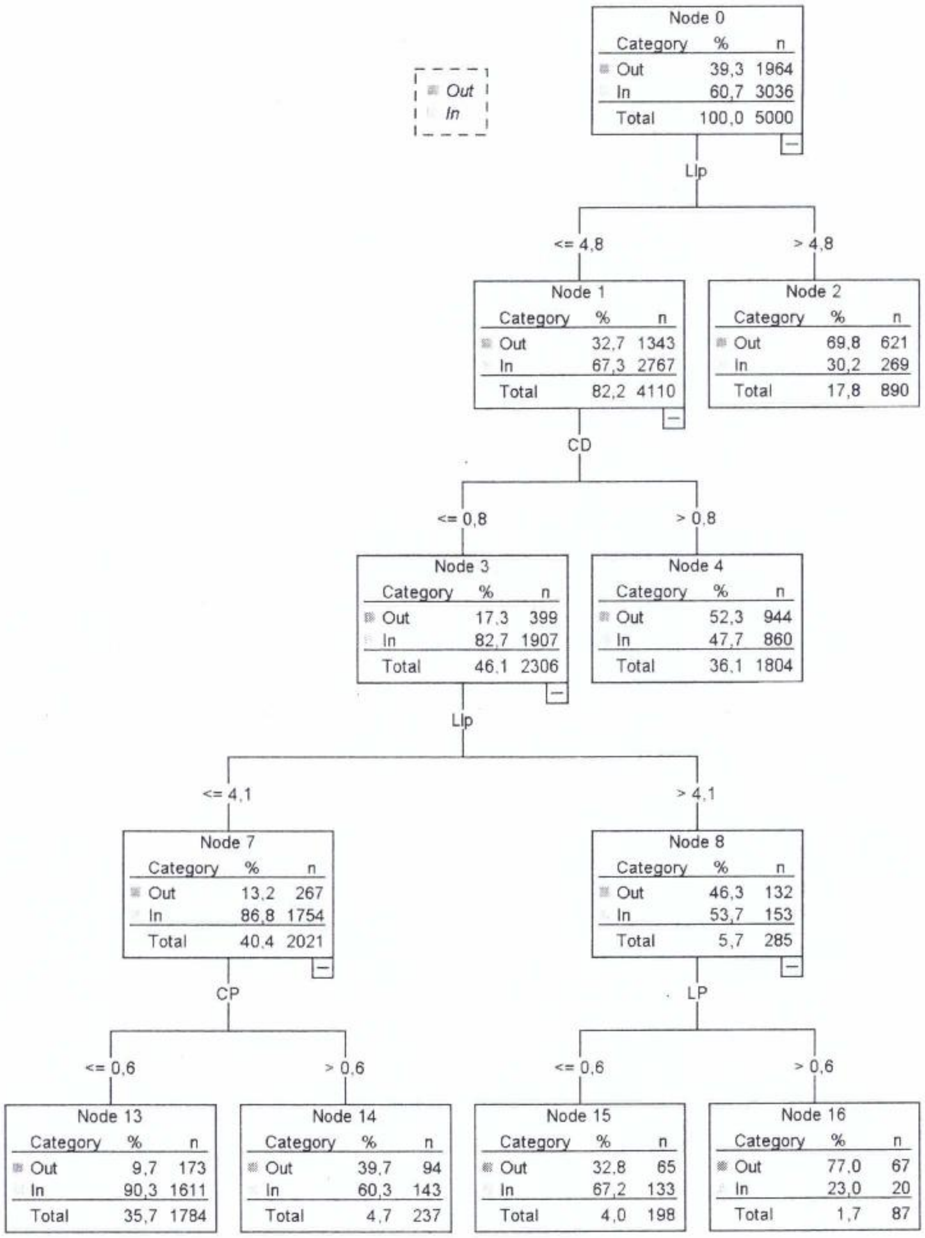

Figure 2. The project's makespan classification tree.

the time spent in the design of the catalog is smaller than or equal to 80% of its estimated time range ($CD \leq 0.8$). Table 5 shows all the relations that hold for the leaves of the tree presented in Figure 2, i.e. for the nodes that have no descendants, together with the proportion of In's and Out's.

Table 5. Rules indicating in which circumstances the catalog project is more likely to be finished on time.

Node	Rule	Outcome	
		In	Out
2	$L_{LP} > 4.8$	30.2%	69.8%
4	$L_{LP} \leq 4.8$ and $CD > 0.8$	47.7%	52.3%
13	$L_{LP} \leq 4.1$ and $CD \leq 0.8$ and $CP \leq 0.6$	90.3%	9.7%
14	$L_{LP} \leq 4.1$ and $CD \leq 0.8$ and $CP > 0.6$	60.3%	39.7%
15	$L_{LP} > 4.1$ and $L_{LP} \leq 4.8$ and $CD \leq 0.8$ and $LP \leq 0.6$	67.2%	32.8%
16	$L_{LP} > 4.1$ and $L_{LP} \leq 4.8$ and $CD \leq 0.8$ and $LP > 0.6$	23.0%	77.0%

It should be noted that according to the information displayed in Figure 2, if the project manager can ensure that the label-printing waiting time for resource allocation is smaller than or equal to 4.8 time units ($L_{LP} \leq 4.8$), then the chances of the project finishing on time goes from 60.7% to 67.3%. However, if they can also ensure that the time spent on the catalog design is smaller than or equal to 80% of its time range ($CD \leq 0.8$), then the chances of the project being delivered on time increases even further to 82.7%.

In addition, if the label-printing waiting time for resource allocation is shortened to at most 4.1 time units ($L_P \leq 4.1$) and the catalog printing is carried out in less than or equal to 60% of its estimated time range ($CP \leq 0.6$), then the chances of the project being finished as planned goes to 90.3%.

4. Discussion

At the outset of this article we undertook to successfully mine RCPS stochastic models with the view of making it easier for managers to anticipate tactical changes in project planning. Bellow we answer some key questions about the implications of the insights provided by classification trees for both project management and the construction of effective business strategies.

4.1. What are the necessary steps leading to the successful mining of RCPS stochastic models?

To successfully mine RCPS stochastic models with the view of helping managers to anticipate changes in project planning that increase the chances of finishing the project on time one may take the following steps:

1. Build a stochastic model that properly represents the project's activities, their interdependency relations and the resource constrains to which they are subjected;
2. Run a simulation process that, for each scenario, collects the project outcome indicating whether it finished on time or not. In these circumstances a scenario is composed of the duration of all project activities as percentage of its time range, together with their respective waiting time for resource allocation;

3. Use the data collected during the simulation process to construct a classification tree that has the project outcome as its target variable;
4. Examine the tree carefully looking for ways to improve the chances of finishing the project on time;
5. Make the necessary adjustments in the project planning of activities;
6. If any adjustments are made, then the stochastic model ought to reflect these changes, the simulation process should be rerun and the classification tree rebuilt.

4.2. How do users of RCPS stochastic models benefit from classification trees?

One of the main benefits of building an RCPS stochastic model is the insight it can potentially provide on the dependency relations that exists among the duration of project activities, the waiting time for their resource allocation and the project outcome. The understanding of these relations allows the identification of activities that most strongly influence the outcome of a project, prompting changes in planning that increase the chances of projects being delivered on time.

However, even the construction and analysis of small RCPS stochastic models require considerable knowledge of mathematics and statistics, including the use of probability density functions for both discrete and continuous variables, sampling, randomly generated scenarios, measures of central tendency and spread, estimation, etc. As a result, many project managers are unable to enjoy the benefits of using this truly powerful tool, because they just lack the necessary technical skills.

In addition, the two statistical methods that are most frequently employed to identify cause and effect relationships among variables in RCPS stochastic models are linear correlation and regression. The proper use of these methods to analyze the relationship that exists among the duration of project activities, waiting time for resource allocation and project outcome requires these relations to be linear, which most often is not the case.

Moreover, linear regression requires predictive variables to be independent, i.e. the value that one of these variables may take may not depend upon the value that other predictive variables take. Unfortunately, there is a natural tendency that the duration of project activities depend upon each other. For example, in many industries time spent on planning and designing tends to strongly influence time spent on testing and delivery. Therefore, situations in which the proper use of linear regression is welcome do not come by as frequently as one may hope for.

Finally, both correlation and linear regression do not deal easily with categorical variables. Linear regression requires these variables to be transformed into a set of independent variables. Linear correlation requires variables to be of ratio scale, while rank order correlation requires them to take value in an ordinal set. Furthermore, linear regression requires the target variable to be continuous. See [McClave and Sincich 2005] for an introduction to regression and correlation, and [Kurpius and Stafford 2005] for a discussion about measurement scales.

Although, classification trees cannot make the construction of RCPS stochastic models any easier, they do help with the analysis of the simulation process, particularly with the extraction of information that favors tactical changes in project planning. The use of classification trees does not require predictive variables to be independent nor to hold a linear relation with the target variable. Also, they can take value in any kind of set, despite its scale of measurement being nominal, ordinal, scalar or ratio. All of this not only greatly facilitates the analysis of RCPS stochastic models, but also makes it faster and less prone to errors.

4.3. Are there any additional benefits for users of RCPS stochastic models?

While correlation uses a number that varies within well defined bounds, typically from -1 to 1, to indicate the intensity of a relation that holds between two variables, linear regression uses a polynomial equation to describe the relationship that holds between a set of predictive variables and a target variable. Although the interpretation of these results are not particularly difficult, they do require some knowledge of mathematics and statistics.

On the other hand, classification trees yield simple logical expressions that describe the dependency relation that hold among predictive variables and a target variable. Such expressions are easy to read and understand, even by those with a restricted knowledge of mathematics and statistics. Moreover, they do indicate, clearly and incrementally, the most critical values that predictive variables with discrimination power may take, i.e., values that most strongly influence a project's outcome. This makes it easier for project managers and their teams to identify what needs to be done and when it needs to be done, before the worst happens.

4.4. Can the techniques described in this article be used to analyze other aspects of project management besides makespan?

Although all the examples that have been used throughout this article have been concerned with project makespan, there is no reason why classification trees cannot be successfully combined with stochastic modeling and simulation to analyze different aspects of project management, such as cost, financing, budget and cash flow compliance, operational risk, human resource hiring and allocation, quality conformity, etc.

Each of these aspects is more easily and effectively analyzed with the right choice of predictive and target variables. For example, for project cost analysis it is usual to choose the estimated cost of each activity as predictive variables and the project total cost as the target variable. However, because project cost may also be influenced by many other factors, as for example the existence of penalties for late delivery, loss of qualified and experienced labor, sudden change in the interest and exchange rates, lack of proper resources, equipment malfunction etc., it is not uncommon that project cost analysis is improved by the inclusion of a multitude of other variables related to both financial and non-financial aspects of project management.

Whatever the situation, the right combination of predictive and target variables is bound to yield rules that help managers to keep their project on track, making tactical changes in planning where and when they are shown to be necessary.

4.5. Can the combination of classification trees and RPCS stochastic modeling and simulation be used as a negotiation tool?

Despite all the efforts that project managers may place on their tasks, projects are full of uncertainties that make planning a difficult endeavor. Not only are projects concerned with interrelated events and activities that have not happened yet, but also once these activities have been completed, they will not be executed in the future exactly the same way as in the past. Therefore, reliable activity cost and time span estimation are not easily achieved. Also, it is not unusual that projects will have people working together for the first time; not to mention that many projects are influenced by a variety of economic, social and political factors such as interest rate, flow of imports and exports, change in consumer behavior, proximity to the electoral period, introduction of new technology, etc.

Hence, projects are very unlikely to run as planned, requiring frequent adjustments in the course of time. The more complex and lengthy the project, the more adjustments it is likely to require. While some projects many not need adjustments involving over budget investments and deadline extensions, others go the other way around.

Because the rules generated by classification trees are conjunctions of logical expressions that are easy to read and understand, and can be presented to non-technical personnel, they may be used as an awareness tool to make senior management more conscious of critical aspects of a project; in particular of those aspects where further investment is necessary. In this sense, the incremental way in which the rules are presented makes it easier for project managers to indicate where the extra funding should go, as well as the consequent increase in the likelihood of finishing the project on time.

For example, in the catalog project if funding is made available to ensure compliance with $L_{LP} \leq 4.8$ and $CD \leq 0.8$, the chances of having the project finished on time go from 60.4% to 82.7%. However, if additional funding is made available to also ensure that $L_{LP} \leq 4.1$ and $CP \leq 0.6$, then the chances of finishing the project on time increase to 90.3%. See *Node 13* in Figure 2. Therefore, not only classification trees make the analysis of RCPS stochastic model easier, but they can also be used as a negotiation tool with stakeholders to ensure that a project receives the resources it needs in a timely manner.

4.6. What are the implications of the insights provided by classification trees to project management and business strategy?

In the extremely competitive world that has emerged as a consequence of the advent of the Internet and the globalization of the world economy, organizations in many lines of businesses have found themselves under enormous pressure to innovate, improving constantly the quality of the products and services they market. Because innovation depends upon the execution of projects, the number and complexity of projects that organizations run annually have increased considerably.

A combination of tools such as classification trees and RCPS stochastic modeling and simulation that helps to increase the chances of projects being delivered on time, within budget and according to the quality aspects that were agreed upon, is certain to favor more effective project management with positive impact on the number of projects that are successfully completed and, as a result, on the competitive aspect of business.

5. Conclusion

In this article we have demonstrated the viability of successfully combining classification trees with RCPS stochastic modeling and simulation to provide managers with the means to anticipate changes in project planning that favor projects being finished successfully. This combination of mathematical tools may be used to analyze different aspects of project management including project makespan and cost, with many advantages of more classical methods.

Classification trees are not difficult to build as predictive variables are not required to have any particular distribution of values or hold any kind of relationship among themselves. Moreover, the rules generated by classification trees are easy to read, understand and communicate to all interested parties, helping to avoid mis-communication among team members and also with the identification of most needed adjustments in project planning.

Furthermore, easy to understand rules favor the involvement of senior management with critical aspects of project planning and execution, making it easier for project managers to ensure that the necessary investments are made where and when they are most needed. All of this makes classification trees a very attractive tool to be used in combination with RCPS stochastic modeling and simulation to support the management of complex projects in the real world.

References

Arora, A., Fosfuri, A., and Gambardella, A. (2001). *Markets for Technology: The Economics of Innovation and Corporate Strategy*. MIT Press.

Blazewicz, J., Lenstra, J., and Kan, A. R. (1983). Scheduling subject to resource constraints: Classification and complexity. *Discrete Applied Mathematics*, 5:11–24.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC Press.

Castillo, A. L. and Muñoz, D. F. (2004). A decision support system to schedule operations in water heater manufacturing. In Hart, S. T. . G., editor, *IIE Annual Conference and Exhibition*, Houston, Texas. Institute of Industrial Engineers.

Chung, C. A. (2003). *Simulation Modeling Handbook: A Practical Approach*. CRC Press.

Demeulemeester, E. L. and Herroelen, W. S. (2002). *Project Scheduling: A Research Handbook*. International Series in Operations Research & Management Science. Springer-Verlag, 1st edition.

Frame, J. D. (2002). *The New Project Management: Tools for an Age of Rapid Change, Complexity, and Other Business Realities*. Jossey Bass Business and Management Series. Jossey-Bass, 2nd edition.

Gini, C. (1939). *Memorie di Metodologica Statistica*, volume I, chapter Variabilità e concentrazione, pages 359–408. Dott. A. Giuffrè, Milano, Italy. Text written in Italian.

Goerlich, J. M. T. and Olaguibel, R. A.-V. (1989). Heuristics algorithms for resource-contrained project scheduling: A review and empirical analysis.

Hamill, L. (2000). The introduction of new technology into the household. *Personal Technologies*, (4):1–16.

Hosni, Y. A. and Khalil, T. (2004). *Management of Technology - Internet Economy: Opportunities and Challenges for Developed and Developing Regions of the World*. Elsevier Science.

Jones, A., Kovacich, G. L., and Luzwick, P. G. (2002). *Global Information Warfare: How Businesses, Governments, and Others Achieve Objectives and Attain Competitive Advantages*. CRC Press, 1^{st} edition.

Kakabadse, N. K., Kouzmin, A., and Kakabadse, A. (2001). From tacit knowledge to knowledge management: leveraging invisible assets. *Knowledge and Process Management*, 8(3):137–154.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:119–127.

Kerzner, H. (2003). *Project Management: A Systems Approach to Planning, Scheduling, and Controlling*. John Wiley & Sons, 8^{th} edition.

Kolisch, R. and Hartmann, S. (2006). Experimental investigation of heuristics for resource- constrained project scheduling: An update. *European Journal of Operational Research*. To appear.

Kurpius, S. E. R. and Stafford, M. E. (2005). *Testing and Measurement : A User-Friendly Guide*. SAGE Publications.

Lewis, J. P. (2000). *Project Planning, Scheduling & Control*. McGraw-Hill, 3^{rd} edition.

Loh, W. Y. and Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7:815–840.

Loh, W. Y. and Vanichestakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association*, 83:715–728.

Info-Tech Research Group (2003). *Effective Project Management: Tools, Templates & Best Practices*. Info-Tech Research Group. Technical Report.

McClave, J. T. and Sincich, T. (2005). *Statistics*. Prentice Hall, 10^{th} edition.

Meredith, J. R. and Mantel, S. J. (2002). *Project Management: A Managerial Approach*. John Wiley & Sons.

Morgan, J. N. and Sonquist, J. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415–434.

Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, New York.

Wagner, E. D. (2005). Enabling mobile learning. *EDUCAUSE Review*, 40(3):40–53.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.

Wright, W. F., Smith, R., Jesser, R., and Stupeck, M. (1999). Information technology, process reengineering and performance measurement: A balanced scorecard analysis of Compaq Computer Corporation. *Communications for The Association of Information Systems*, 1(8):1–61.

Zhu, G., Bard, J. F., and Yu, G. (2005). Disruption management for resource-constrained project scheduling. *Journal of the Operational Research Society*, (56):365–381.

Núcleo de Computação Eletrônica
Universidade Federal do Rio de Janeiro
Cx. Postal: 2324 - CEP: 20001-970 - Rio de Janeiro - RJ
Tel.: PABX (021) 598-3212/598-3130 - FAX: (021) 270-8554