ABSTRACT
The definition of imprecise estimates is one of the crucial problems of software industry. In this context, this work presents an approach for prediction of stated periods of projects based on simulation, using statistical methods. The simulation is used to provide more confidence to the estimates defined by the project manager, so that they can reflect the organization's reality. Based on the estimates provided by the simulator, statistical methods are used to predict an interval, to which the end of the project, in practice, will have chances to belong to. A case study was performed in the end of the research attesting, statistically, the potential of the approach defined in this work.

Categories and Subject Descriptors
D.2.9 [Software Engineering]: Management – Time estimation.
I.6.8 [Simulation and Modelling]: Types of Simulation – Discrete event.

General Terms
Management, Experimentation, Verification.

Keywords
simulation, prediction, project management, statistical methods.
composed by probable finish dates for the project being estimated. This interval is provided through a statistical analysis.

To attest the approach defined in this work, a case study was performed in a small software company. According to the results obtained at the end of the case study, it could be verified that all pilot projects, executed in practice, finished in dates that belonged to the interval estimated by the methodology, suggesting its estimation capacity.

2. PROJECT PLANNING

2.1 Problem Description

Jones [4] emphasizes that imprecise stated periods and costs estimates besides the absence of an appropriate management process is one of the main causes of software projects failures. When they don’t result on total failures, the out of date delivery of projects causes customer dissatisfaction, besides damages related to contract fines.

One way to solve this problem is to give conditions so that the organization can predict, precisely, the chances of a future project to be executed with success, considering the stated periods, costs and quality defined. Through the perception of project execution, the stated periods can be negotiated with more confidence, reducing the risk of delays.

In this context, it is interesting the estimates are defined according to the organization’s current reality. For this, the simulation [3] is a useful tool that allows foreseeing the execution of the project, estimated by the manager, according to the current reality of physical and human resources in the organization [11].

Another important point to consider during the definition of project estimates is the dynamism present in a software development environment. The occurrence of events, related to risks of projects, can cause delays on the execution of critical activities [7]. In this way, it’s necessary that this possible delay is considered in the prediction, so that its results can reflect the reality of such environments.

A possible approach is to analyze, statistically, the history of projects executed by the organization, trying to identify its “profile” related to delays of projects. Such delays represent the organization’s natural dynamism, and can be used as a complement to the simulation process. The use of statistical methods is defined in quality models like CMMI [12]. In one of its levels, the model suggests the adoption of a statistical methodology that can be used to measure software process non-conformances. In the context of this work, the use of statistical methods aims to provide more predictability to the future projects in the organization.

2.2 Related Works

The works [10] e [11] aim to help improving project planning, presenting approaches to foreseeing processes execution behavior using simulation. The work [11] presents a simulation model based on cooperative intelligent agents that execute tasks, according to their perception of the environment, skills and affinity with the others agents. The work [10] presents an environment for process studying. The simulation component is able to identify problems related to process models previously defined in the environment. In this case, the component also uses intelligent agents to simulate the processes execution, but the agents’ features such as skills and affinities are not considered, in a significant way.

The works [1] e [2] present approaches for training and learning in the context of project planning and management area. Both works are based on simulation games in which the player needs to plan and conduct the project with success to win the game.

The work presented in this article places itself in the project planning area, using simulation to validate the project chronogram estimated by the project manager, verifying its execution according to the reality of the development team. However, this work differs from others by the use of simulation for prediction of stated periods of projects, aiming to help improving the prediction process in software development organizations.

3. AN APPROACH FOR PREDICTION OF PROJECTS BASED ON SIMULATION

This work has as its main contribution an approach for prediction of stated periods of software projects, using simulation and statistical analysis. The Figure 1 details the proposed approach.

According to the Figure 1, the project chronogram, estimated by the manager, besides the information about the development team, such as availability, skills and affinities are used as parameters for the simulation tool, called ProSimulator. In the end of simulation, the simulator provides a set of metrics, including the project finish date, according to the simulation of the tasks scheduling based on the features of the development team. In parallel time, the history of projects already executed by the organization is analyzed statistically, providing a statistical distribution that represents the organization profile related to delays of projects. The finish date, provided through the simulation and the statistical distribution are used to predict an interval, to which, in practice, the end of the project will have possibilities to belong to. The next sections detail each one of the steps of simulation, statistical analysis and prediction.

Figure 1. Prediction of projects based on simulation.

3.1 ProSimulator: Simulation of the Project Chronogram

ProSimulator is a discrete-event simulation tool, developed in the context of the environment ImPProS [6]. In this environment, the project manager defines a set of activities to be executed in the project and estimates a medium stated period for them based on the roles responsible for their execution, such as systems analyst,
software engineer, among others. So, the execution of the estimated chronogram is simulated by the tool, considering the information about the development team.

In the ProSimulator, the events are represented by the chronogram activities, disposed in a queue, ordered by estimated start date, in way that the simulation clock is always updated to the next start date related to the activity to be simulated. The simulation process ends when the activities queue is empty.

During the simulation, the activities are dequeued and allocated to the members of the development team, represented by simulation agents. The allocation decision and the execution of activities are both influenced by the agents’ skills and affinities levels related to them. In this way, the agent to be allocated to an activity will be the one that is: available in its estimated date (because of the possible parallelism); has the highest level of skills for its execution; and has the highest affinity average with the other agents already allocated to it. According to the allocated agents and their skills and affinities levels, the time spent in the execution of the activity can be equal, higher or lower than the duration time, estimated by the manager.

The calculation of the simulated time of an activity is defined by the equation 1, which considers the agents’ skills and affinities levels as heuristic information. Such data are known as critical influence factors during the development of the project [11] and, because of that, were considered in this work through statistical data related to the members’ experience in other projects already executed by the organization. Such information must reflect the reality of the development team, available for the project, being that the main purpose of the simulator in the context of the prediction approach proposed in this work.

\[
\text{time}_{\text{activity}} = \text{estimated time} \times \text{factor}(\text{skill}(x)) \times \text{factor} \left( \sum_{y \in B} \text{affinities}(x,y) \right) \quad (1)
\]

Where:
- \( x \in B, y \in B; B > 0; x \neq y \)
- \( B \): set of agents allocated to execute the activity.

### Figure 2. ProSimulator: an example of a simulated chronogram.

The Figure 2 shows an example of the results provided by the ProSimulator, after the simulation of the chronogram estimated by the project manager. The red bars represent the estimates defined by the manager, and the blue ones represent the simulated chronogram, according to the reality of the development team. Besides that, the tool provides textual reports that present the failures detected during the simulation process, such as: extrapolation of stated periods and costs of the project; verification of idle and overload agents; verification of activities that extrapolate the estimated costs and stated periods in a significant way, among others.

### 3.2 Statistical Analysis

In parallel time, a sample of projects already executed by the organization must be defined and statistically analyzed, aiming to identify the organization’s profile related to delays. The sample of projects must be chosen according to criteria, such as: choosing more recent projects, so that the profile identified is close to the current reality; choosing projects with different size and complexity, so that the sample can be representative and non-tendentious, among others.

After choosing the sample, the variable to be analyzed (delay) is measured to each element in the sample. That is, for each project in the sample, it is defined the percentage of delay considering the gap between the estimates defined by the manager and the project finish date, according to its execution in practice, registered in the organization’s database. After that, the average (\( \mu \)) and standard deviation (\( \sigma \)) of the sample are calculated, based on the percentages of delay related to each project, previously measured. These data are used to identify the probabilistic distribution that represents the sample’s profile related to delays of projects.

In the context of this work, it’s essential that the percentages of delay, defined for the projects in the sample, can represent, in a significant way, the normal curve. In this way, the statistical analysis can be conducted with more confidence, based on a concept widely studied in the literature [13]. In this case, after defining and analyzing the sample of projects, it is necessary to verify its normality, so that the data can be used in the prediction. For this, it must be applied a normality test. In this work, it was applied the Shapiro-Wilk test [9], that attests, with 95% of confidence, the normality of a certain distribution. It’s important to emphasize that this work is defined exclusively for a set of projects that represents the normal curve. In the case of non-normal distributions, the approach defined in this work needs to be changed.

### 3.3 Prediction of the Termination Interval

If the sample that identifies the delay in projects execution represents the normal curve, it’s possible to define an interval, attempting to predict the real termination of the future project being estimated. The prediction is performed in the following way: the project finish date, indicated by the ProSimulator, is located on the curve that represents the normal distribution, as showed in Figure 3. After that, it’s added to the date provided by the simulator, the average percentage of delay, converted in days, determining, in this way, a new finish date, considering the organization’s delay profile. This new date will be the center of the normal curve, as showed in Figure 3.

Studies indicate that 95% of the elements in a normal distribution are situated in the interval defined as the product of standard deviation and constant 1.96 (1.96\( \sigma \)) [13]. So, after defining the new project finish date, according to the
organization’s delay profile, the interval is determined around this new finish date, displacing to the right and to the left the product of standard deviation (related to the sample of projects) and the constant 1.96, converted in days. In this way, there are 95% of chances the project, in practice, finishes in a date belonged to the estimated interval.

Figure 3. Prediction of the termination interval.

It’s important to emphasize that the approach for prediction defined in this work is adaptable to any project to be estimated, because it is based on the percentage of delay and not the number of days that a project delayed. In this way, both interval size and shape of the normal curve will vary according to the project size being estimated.

4. RESULTS

To verify the potential of the approach for prediction defined in this work, a case study was performed in a small software company called SWFactory, with, approximately, 20 employees. The next sections detail the case study performed in the company.

4.1 Choice of Projects

Initially, the company’s history of projects was analyzed, aiming to select the projects to be used in the experiment to compose the sample and the pilot projects test base. In the end, 14 projects were chosen to compose the sample and other six were chosen to be pilot projects. The choice of projects to compose the sample was based on the criteria defined in the section 3.2.

4.2 Statistical Analysis of the Sample of Projects

After the choice of projects to compose the sample, a statistical analysis was performed, calculating the percentage of delay related to each project. The average and standard deviation of the sample were also calculated (29.05% and 57.36%, respectively). After that, the Shapiro-Wilk test (section 3.2) was applied to the sample of projects, obtaining a p-value = 0.4924, attesting, in this way, the normality of the distribution related to the sample (a p-value equal or higher than 0.05 guarantees, with 95% of confidence, the normality of the distribution [9]). The test was performed using the statistical software R.

4.3 Simulation of Pilot Projects

Parallel to the statistical analysis, the pilot projects were simulated. The pilot projects’ chronograms, estimated by the company’s project manager, were inserted into the ProSimulator. Statistical information about skills and affinities related to the members of the development team was also inserted into the simulation tool. Such information was collected in the company’s history base that contained information about the developers’ behavior, during the execution of past projects.

4.4 Prediction of Pilot Projects Termination Intervals

Based on the pilot projects finish dates, provided through the simulation, and the delay average calculated through the statistical analysis, the pilot projects termination intervals could be statistically estimated (Table 1). For comparison, the pilot projects finish dates, occurred in practice, were collected in the organization’s history.

Table 1. Prediction of pilot projects termination intervals

<table>
<thead>
<tr>
<th>Pilot Proj. Id</th>
<th>Finish date: simulation + average delay (in days)</th>
<th>μ * 1.96</th>
<th>Termination Interval (x₀)</th>
<th>Termination Interval (x₁)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>04/24/2006</td>
<td>12</td>
<td>04/12/2006</td>
<td>05/06/2006</td>
</tr>
<tr>
<td>2</td>
<td>07/30/2005</td>
<td>12</td>
<td>07/18/2005</td>
<td>08/11/2005</td>
</tr>
<tr>
<td>3</td>
<td>10/04/2005</td>
<td>12</td>
<td>09/22/2005</td>
<td>10/16/2005</td>
</tr>
<tr>
<td>5</td>
<td>11/26/2005</td>
<td>10</td>
<td>11/16/2005</td>
<td>12/06/2005</td>
</tr>
</tbody>
</table>

Trough the data presented in Table 2, it could be verified the potential of the approach for prediction defined in this work. Among the six pilot projects tested, all of their finish dates, occurred in practice, belonged to the correspondent estimated interval. The results were obtained considering the level of confidence 95%, considered reasonable for statistical tests.

Table 2. Termination interval estimated versus projects finish dates in practice.

<table>
<thead>
<tr>
<th>Pilot Proj. Id</th>
<th>Termination Interval (x₀)</th>
<th>Termination Interval (x₁)</th>
<th>Finish date (in practice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>04/12/2006</td>
<td>05/06/2006</td>
<td>04/25/2006</td>
</tr>
<tr>
<td>2</td>
<td>07/18/2005</td>
<td>08/11/2005</td>
<td>08/08/2005</td>
</tr>
<tr>
<td>3</td>
<td>09/22/2005</td>
<td>10/16/2005</td>
<td>10/06/2005</td>
</tr>
<tr>
<td>4</td>
<td>09/24/2006</td>
<td>10/22/2006</td>
<td>10/05/2006</td>
</tr>
<tr>
<td>5</td>
<td>11/16/2005</td>
<td>12/06/2005</td>
<td>12/06/2005</td>
</tr>
</tbody>
</table>

It can be seen that the decrease of the level of confidence increases the restriction of the estimated interval and, consequently, it’s less probable that, in practice, the project will end in a date belonged to the interval. For a prediction considering a lower level of confidence (68.23%), also tested in this work, among the six projects analyzed, three of them had their finish dates belonged to the estimated intervals.

5. FINAL REMARKS

The work presented in this article defines an approach for prediction of stated periods of projects, based on simulation, using
statistical methods. The aim of this work was to provide more confidence to the estimates defined by the project manager, in a context that estimates are defined based on roles, not considering the reality of organizational resources.

Using the ProSimulator, the prediction of the termination interval could be based on estimates closer to the organization’s reality, considering the personal features of the development team, and not only the medium time estimates defined by the manager.

It’s important to emphasize that the potential of the approach defined in this work depends, directly, on the information about the organization, used in the prediction. Such information must be updated, aiming to reflect the current organization’s reality related to its both human resources and delay profile. In the first case, it’s suggested the automatic capture of these data through knowledge discovery tools, installed in development environments, so that the organization’s reality can be reflected in such data (the mechanism called “knowledge discovery”, proposed in [5]). In the second case, the history of projects, in which the sample is based on, must be always updated with recent projects. In this way, the prediction results will have more chances to be close to the future projects results, observed in practice.

As future works of this research it’s proposed: the automation of the statistical analysis in ProSimulator (currently, the analysis is performed using electronic spread sheets); new case studies using the approach for prediction, aiming to calibrate it, according to the detected problems.

6. ACKNOWLEDGMENTS
Thankfulness to CNPq (National Council for Scientific and Technological Development) by the financial support in this project.

7. REFERENCES