Distributed and Colocated Projects: a Comparison

Alessandro Bianchi*, Danilo Caivano*, Filippo Lanubile*, Francesco Rago°, Giuseppe Visaggio*
**Dipartimento di Informatica, Università di Bari – Via Orabona, 4 – 70126 – Bari – Italy*
°Italy Solution Center, EDS Italia, Viale Edison, Lo Uttaro, 81100 - Caserta - Italy
{bianchi, caivano, lanubile, visaggio}@di.uniba.it, francesco.rago@eds.com

Abstract

The aim of this work is to point out through experimentation some of the problems that arise with distributed software development, such as the need for new techniques and methods for managing projects and processes, so as to achieve better assignment of activities among the various working groups, together with efficient communication among the members of each team. The paper analyzes the post mortem data on two projects, one conducted at a single site and the other at several different sites concurrently. The findings show that effort estimates are less accurate if the project is a small one, while increasing the number of staff members increases the risk of defects and hence rework; this generates a greater gap between expected and real staff requirements. The study also confirms that in distributed processes there is a greater need for communication among the working members than in colocated processes.

1. Introduction

The new forms of competition and cooperation that have arisen in software engineering as a result of the globalization process have an impact on the whole software process. Software development and maintenance have thus become processes distributed over various geographical sites and involve increasing numbers of staff with different cultural backgrounds. It has been pointed out in [CA01] that at present, 50 different nations are collaborating in different ways in software development.

However, global software development has a number of disadvantages, such as the need to use ad hoc methods for managing larger, and geographically distant working groups [Coc00], as well as knowledge sharing tools [NFK97, SY99], while there are new overheads involved in the problems of staff communication interchanges [ED01]. Herbsleb and Moitra have identified in [HM01] a set of issues connected with global software development. These consist of:

- *strategic issues*, concerning the decisions for subdividing the tasks among the different sites, so as to be able to work as independently as possible while maintaining efficient communication among sites;
- *cultural issues*, that arise when the staff come from different cultural backgrounds;
- *inadequate communication*, caused by the fact that geographical distribution of the staff over several sites increases the costs of official communications among team members and limits the possibility of carrying on the informal interchanges that traditionally helped to share experiences and foster cooperation to attain the targets;
- *knowledge management*, that is more difficult in a distributed environment as information sharing may be slow and occur in a non uniform manner, thus limiting the opportunities for reuse;
- *project and process management issues*, having to do with all the problems of synchronization of the work at the various different sites;
- *technical issues*, that have an impact on the communication network linking the various sites.

This work presents an analysis of data acquired in industrial projects, aiming to assess the impact of project and process management and the need for communication on the results of distributed processes. A post-mortem analysis is made of the data acquired in two projects carried out in EDS Italia: one involving several different geographical sites and the other a single site belonging to the same company. Data are analyzed to show the impact of distributed projects over scheduling of the activities, their subdivision and the degree of synchronization achieved, as described in [PV98]. Furthermore, communication among team members is analyzed analogously to [PSV94], as also we look for evidence of higher cost overheads in distributed processes due to practical limitations of this communication.

The paper is organized as follows: section 2 presents the projects and the metrics used in the analysis; section 3 illustrates the main lessons learnt from the investigation; section 4 draws some conclusions.

2. Case Study Setting

2.1 Characterization of the Projects

The first project, that will hereafter be indicated as the *distributed project*, was conducted over 3 different geographical sites of EDS-Italia, and involved resources that operated as a single team. It was a large project requiring a high number of human resources to carry out massive, non routine maintenance of a large software system to solve the Y2K problem. The software system considered was subdivided into *functional areas* (FA), each consisting of a *work-packet* (WP). There were 100 WP, and the maintenance effort had to deal with 65 of them. The size of each WP is expressed by the number of items, i.e. programs, library elements or JCL procedures, included. Each WP included a variable number of items ranging from a minimum of 6 to a maximum of 7506, making up a total of 25044 items, i.e. an average of 385.29 items per WP involved in the maintenance effort.

The second project, indicated as the *colocated project*, was conducted at a single site and consisted of corrective maintenance of the software system of a large services company. The goal was to remove all parts of the software

system, which caused an unexpected behavior. The software to be maintained consisted of 4 *subsystems*, each including a variable number of *subprojects*[1]. Each of the latter involved 111 or 112 items, for a total of 6672 items. The maintenance operations involved 58 subprojects.

2.2 Data Collection

To carry out the post-mortem analysis of the data acquired in these projects, the work packets and subprojects covering all the phases of the working cycle were taken into account. The following measures were collected:

- *estimated duration and actual duration* of the projects required to complete the WPs or subprojects, expressed as working days;
- *size* of the WPs or subprojects, expressed as number of items;
- *reliability metrics* of the WPs, i.e. number of *faults* and *failures*, number of requests for change made during the development projects or changes made on the system after it became operative (hereafter simply indicated as number of *changes*) and total number of problems of any nature that arose during the observation period (simply indicated as number of *issues*);
- *effort* involved to complete the WPs or subprojects, expressed as working days/ person;
- *staff size*, i.e. number of people who took part in executing the WPs or subprojects;
- *number of reports* produced to describe the work progress;
- *number of messages*, i.e. number of information exchanges among the various working groups;
- *number of meetings* officially held among the members working on the WPs or subprojects.

These observed metrics gave rise to the following calculated metrics:

- *mean of estimated duration* of the WPs belonging to a FA (hereafter abbreviated as *MED*), expressed as working days, calculated as $MED = \frac{\sum_{i=1}^{n} ED_i}{n}$, where ED_i is the value of the estimated duration of the i-th WP in the FA and n is the number of WP in that FA;
- *FA size normalized over the number of included WP* (*FANS*), expressed as items, calculated by $FANS = \frac{\sum_{i=1}^{n} S_i}{n}$, where S_i is the value of the actual size of the i-th WP in the FA and n is the number of WP in that FA;
- *discrepancy between estimated and actual duration* of the i-th WP (DIS_D_i), expressed as a percentage, calculated by $DIS\_D_i = \frac{AD_i - ED_i}{RD_i}$, where AD_i and ED_i are the estimated and the actual durations of the i-th WP, respectively;
- *discrepancy between estimated and actual staff* (DIS_S_i) for the i-th WP or subproject, expressed as the number of people and subprojects, calculated by $DIS\_S_i = ES_i - AS_i$, where ES_i and AS_i are the estimated and the actual staff of the i-th WP or subproject, respectively.

3. Data Analysis

3.1 Estimated vs Actual Duration

The first analysis made on the distributed project data assessed the ability of management to estimate the duration. The differences between the estimated and the actual durations were compared. Figure 3.1 shows the gap in percentage between the two values for each WP.

It can be seen that for the first WPs there was a tendency to overestimate the required time; this was followed by a chunk of WPs whose time estimates largely correspond to the actual duration. Finally, in the last part of the project, there was a tendency to underestimate the time required to conclude all the activities pertaining to a WP, with only a few exceptions, such as WP G.081.

This discordance between estimated and actual duration of the WPs is due to two main factors: a) a low rate of distribution of the WPs among the different sites, b) poor attention paid to the estimates of small WPs in comparison with the others belonging to the same FA, so that these were estimated to take much less than the average duration of the WPs belonging to that particular FA.

To investigate the effect of the distribution of the WPs among the sites, we consider them as components of each FA, rather than of the entire system. Table 3.1 demonstrates the correspondence between WPs and FAs and shows the percentage of WPs belonging to each FA performed at a single site.

This subdivision shows that the WPs whose times tended to be overestimated belonged to FA P01, and the underestimated WPs to FA P03, and, to a lesser extent, to FA P04. Finally, the chunk of WPs with a good time estimate belonged to FA P02. As to the percentage of WPs performed at a single site (third column in table 3.1), there is a clear

[1] Note that the subsystems and the subprojects in the colocated project correspond to the functional areas and work packets in the distributed project.

correlation between the true correspondence of the estimate and the distribution of the WPs over several sites. Data in the table show a greater gap between estimated and actual duration for FAs whose WPs were performed at a single site.

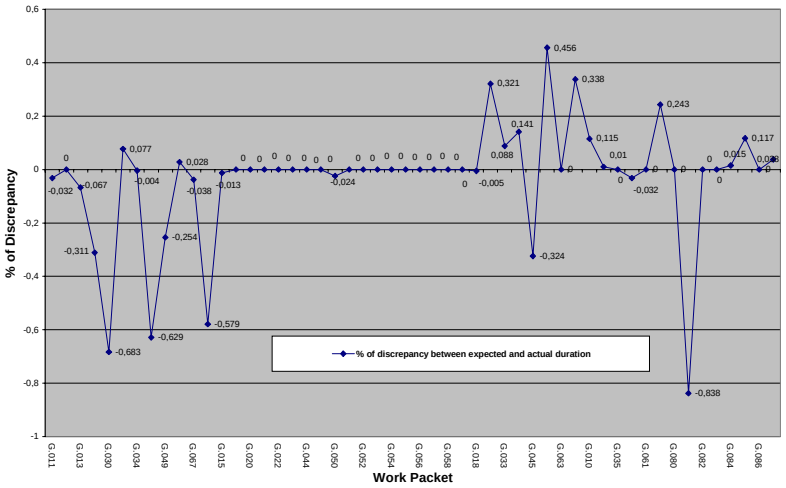


Figure 3.1. Comparison of the percentage discrepancy between estimated and actual duration of the work packets in the distributed project

To investigate the effect of the size of the WPs on the time estimates, table 3.2 shows the data on: MED, FANS, estimated duration of the WPs, DIS_D, and WP size for all those WPs with a greater than 20% gap between estimated and actual duration.

FA	Work Packet	% of WPs performed at a single site
P01	G.011, G.012, G.013, G.014, G.030, G.031, G.034, G.038, G.049, G.060, G.067,G.068	66,1%
P02	G.015, G.017, G.020, G.021, G.022, G.043, G.044, G.046, G.050, G.051, G.052, G.053, G.054, G.055, G.056, G.057, G.058, G.059	60,7%
P03	G.018, G.028, G.033, G.040, G.045, G.062, G.063, G.087	73,8%
P04	G.010, G.016, G.035, G.036, G.061, G.071, G.080, G.081, G.082, G.083, G.084, G.085, G.086, G.506	76,4%

Table 3.1. FAs to which each WP belongs, and percentage of WPs performed at a single site.

Analysis of these data shows a greater gap for WPs with a lower estimate than the mean for that FA, except in the case of WP G.028. Thus, we can conclude that less accurate estimates were made for smaller WPs, whose duration was estimated to be much lower than the mean for WPs belonging to the same FA.

FA	WP	MED	FANS	Estimated WP duration	DIS_D	WP size
P01	G.014	232,7 days	586,7 items	148 days	-31%	81 items
	G.030			175 days	-68%	579 items
	G.038			202 days	-63%	60 items
	G.049			79 days	-25%	18 items
	G.068			90 days	-56%	12 items
P03	G.028	106,4 days	365,5 items	125 days	32%	233 items
	G.045			94 days	-32%	56 items
	G.062			31 days	46%	158 items
	G.087			47 days	34%	165 items
P04	G.071	170,3 days	666,3 items	128 days	24%	31 items
	G.081			125 days	83%	39 items

Table 3.2. Data on WPs with a greater than 20% gap between estimated and actual duration.

3.2 Estimated vs Actual Staff

The second analysis assessed the ability to estimate the amount of staff needed to perform the WPs in the distributed project and the subprojects of the colocated project. Figure 3.2 shows the differences between estimated and actual staff in the distributed project (straight line) and the colocated project (dashed line).

Firstly, the figure shows that there was a greater difficulty in estimating the staff for the distributed than for the colocated project. There is also a more regular distribution of the gaps between estimated and actual values in the colocated project, whereas in the distributed project, there are evident peaks. The highest gaps in the distributed project can be attributed to the variable duration of the WPs in this project, already examined in the section on Duration.

Secondly, the WPs in the distributed project required a higher staff size than the subprojects of the colocated project: the average staff size for the distributed project was just under 20 people per WP (compared with an estimate of just over 20), whereas the colocated project required an average staff size of 15.5 people (compared with an estimated 17.6).

Finally, it can be observed in figure 3.2 that only overestimates were made for the colocated project whereas there are also some underestimates for the distributed project.

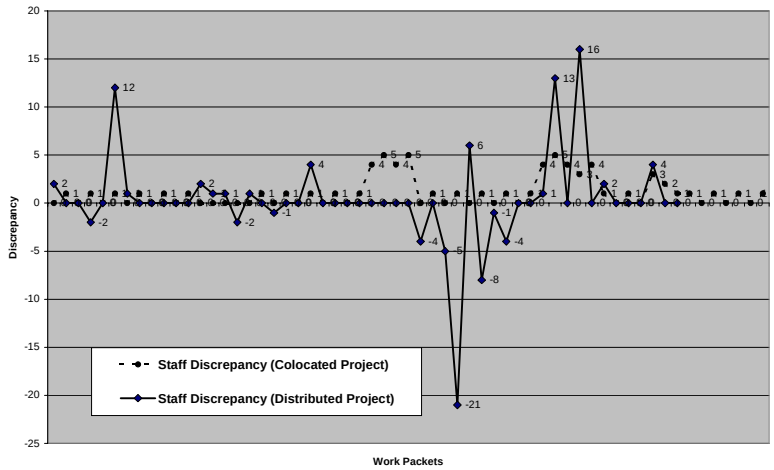


Figure 3.2. Comparison of the overall discrepancy between estimated and actual staff required in the two projects

To identify the causes of these differences, in table 3.3 we show the size and reliability measures for the WPs in the distributed project with a difference of more than 6 people between estimated and actual requirements.

WP	DIS_S	WP size	Faults	Failures	Changes	Issues
G.031	12 people	4096 items	12	5	5	7
G.036	13 people	764 items	5	4	6	15
G.040	-21 people	407 items	0	0	2	6
G.045	6 people	56 items	0	0	1	1
G.062	-8 people	158 items	9	0	1	2
G.071	16 people	31 items	0	0	1	2
Mean of all WPs in project	3 people	918.67 items	1,3	0,4	1,4	2,3

Table 3.3. Data on WPs with highest gaps between estimated and actual staff required.

There are large gaps for large WPs, like G.031, but also for small ones, like G.045 and G.071, so the size of the WP has no impact on the discordance. This is caused, instead, by the faults that occur during execution of the WPs. In fact, in the distributed project, each of the WPs shown has a significantly higher value for at least one of the fault metrics than the mean (last row in table). We can conclude that the higher the number of people estimated for a WP, the higher the number of faults. This requires more reworking and therefore causes a greater gap between estimated and actual values.

3.3 Communication

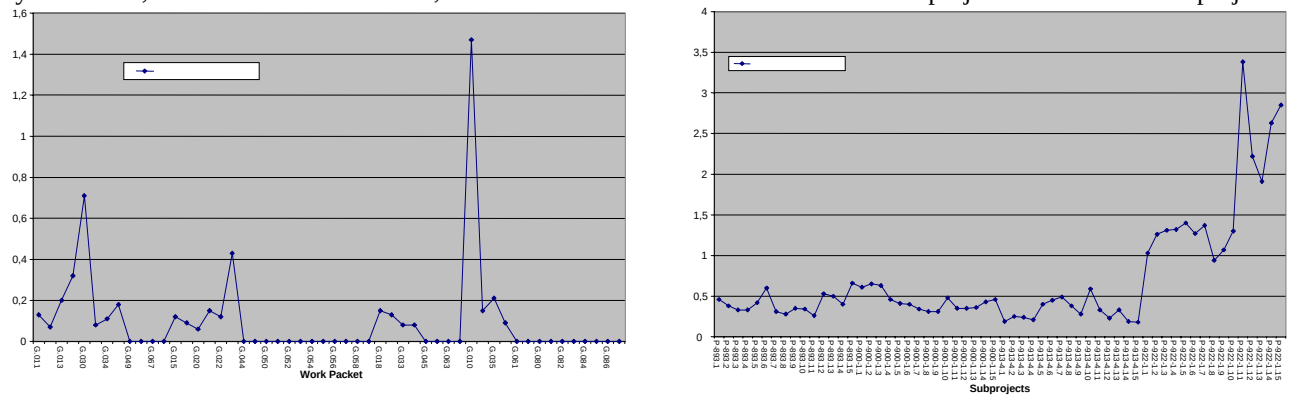
The third aspect investigated was related to the communication among team members. Table 3.4 shows the measures of communication for each project, expressed both as an absolute value and (in brackets) as a normalized value with respect to the effort required to carry out the relative FA or subsystem.

The number of reports produced for each subsystem of the colocated project is constant, whereas it varies for the distributed project. When the value is normalized over the effort, the distributed project is seen to require more reports. The same can be said of meetings. Instead, the number of messages shows the opposite trend, as more messages were exchanged for the colocated than for the distributed project. If we consider only the number of reports and the number of meetings, data confirm that adequate communication between staff members working on the distributed project requires greater effort and hence entails greater cost overheads than for the colocated project.

Distributed Project	N° Reports	N° Msg	N° Meetings	Effort
AF1	229 (3.88)	185 (3.14)	73 (1.24)	59 person/days
AF2	206 (3.12)	106 (1.61)	137 (2.08)	66 person/days
AF3	105 (2.19)	65 (1.36)	39 (0.82)	47.8 person/days
AF4	122 (3.99)	118 (3.86)	99 (3.24)	30.6 person/days
Colocated Project				
SS1	30 (0.10)	700 (2.40)	15 (0.05)	291.4 person/days
SS2	30 (0.08)	924 (2.41)	14 (0.04)	383.5 person/days
SS3	30 (0.15)	533 (2.63)	15 (0.07)	203 person/days
SS4	30 (0.03)	2909 (2.44)	15 (0.01)	1191.1 person/days

Table 3.4. Metrics for communication, expressed both as an absolute value and normalized over the effort

To investigate the opposite tendency for messages, figures 3.3.a and 3.3.b show the distribution of messages exchanged by each team, normalized over the effort, for the WPs of the distributed and for the subprojects of the colocated project.



a) **b)**

Figure 3.3. Distribution of the number of messages exchanged by the team, normalized to the effort, for the WPs in the distributed project (a) and for the subprojects of the colocated project (b)

These graphs show that in the distributed project, messages are not always exchanged for each WP; in fact, in 30 of 52 cases, (over 50%), the metric value is zero. Moreover, the values for the single WPs and subprojects shows that there is a more intense exchange of messages in the distributed project. This means that in the distributed project, the exchange of messages is affected by the characteristics of the WP and, despite the generally greater intensity, does not contribute to increase the organizational costs, because it is offset by the zero costs for more than 50% of cases.

4. Conclusions

This post mortem analysis of the data on two different types of projects confirms some of the global software development problems that have still to be solved. It demonstrates the poor management of distributed projects, especially as regards estimation of the duration of work packets and of the number of staff required to execute them. The gap between estimated and actual values is largely due to the fact that those making the estimates have not gained sufficient experience in distributed environments and therefore define the baselines according to their experience with colocated projects. In short, the fact of being able to perform activities in larger work groups causes the management to underestimate the problems inherent to distributed processes. Moreover, the estimates are more accurate for larger WPs. This can be explained by the fact that for larger WPs with longer estimated times, the problems of inefficiency due to the distributed work are smoothed over time. It can also be observed that increased staff results in an increased number of faults and hence reworks, generating a larger gap between estimated and actual staff needed. Analysis of the duration provided a further indirect confirmation of the conclusions of various authors (e.g. [HMFG01], [ED01], [NFK97]) on the delay introduced by distributed work with respect to work carried out in a single site.

Analysis of data on communication within the team show that distributed work requires more reports and meetings than work carried out in a single site, because the informal discussions that enable information sharing among the various participants are lacking. This causes an increase in cost overheads for distributed projects.

In conclusion, new techniques and methods for software engineering need to be studied to meet the new requirements created by global software development. Simply transposing traditional technologies does not enable the best exploitation of the potential of distributed work projects and indeed, causes these to appear less efficient than centralized work projects.

5. References

[CA01] E. Carmel, R. Agarwal, “Tactical Approaches for alleviating Distance in Global Software Development”, *IEEE Software*, Mar-Apr 2001, pp. 22-29.

[Coc00] A. Cockburn, “Selecting a Project’s Methodology”, *IEEE Software*, July-August 2000, pp.64-71.

[ED01] C. Ebert, P. De Neve, “Surviving Global Software Development”, *IEEE Software*, Mar-Apr 2001, pp.62-69

[HMFG01] J.D. Herbsleb, A. Mockus, T.A. Finholt, R.E. Grinter, “An Empirical Study of Global Software Development: Distance and Speed”, *Proc. Intl. Conf. on Software Engineering*, 2001, pp. 81-90.

[HM01] J.D. Herbsleb, D. Moitra, “Global Software Development”, *IEEE Software*, Mar-Apr 2001, pp. 16-20.

[NFK97] K. Nakamura, et al., “Distributed and Concurrent Development Environment via Sharing Design Information”, *Proc. of the 21st Intl. Computer Software and Applications Conference*, 1997.

[PSV94] D.E. Perry, N.A. Staudenmayer, L.G. Votta, “People, Organization and Process Improvement”, *IEEE Software*, Jul-Aug 1994, pp.36-45.

[PV98] D.E. Perry, L.G. Votta, “Parallel Changes in Large Scale Software Development: An Observational Case Study”, *Proc. Intl. Conf. on Software Engineering*, 1998, pp. 251-260.

[SY99] J. Suzuki, Y. Yamamoto, “Leveraging Distributed Software Development”, *Computer*, Sep 1999, pp.59-65.