Are size measures better than expert judgment?
An industrial case study on requirements volatility

Annabella Loconsole
Department of Computing Science
Umea University
SE-90187 Umea, Sweden
bella@cs.umu.se

Jürgen Börstler
Department of Computing Science
Umea University
SE-90187 Umea, Sweden
jubo@cs.umu.se

Abstract

Expert judgment is a common estimation approach in industry. However, there is very little research on the accuracy of expert judgment outside the area of effort estimation. In this paper, we present an industrial case study investigating subjective and objective measures of requirements volatility. Data was collected in retrospect for all use cases of a medium-size software project. In addition, we determined subjective volatility by interviewing developers and managers of the project. Our data analysis show that structural measures perform better than expert judgment in estimating the total number of changes to use case based requirements. These results confirm results from a previous case study by the authors and suggest that project managers should not rely on expert judgment alone for decision making.

1 Introduction

Software project success is influenced by many factors; one of them is requirements stability [26] (stability is an antonym of volatility). Software measurement can help us in providing guidance to the requirements management activities by quantifying the amount of changes and predicting the risk of frequent changes to requirements. Requirements and other entities in software engineering can be measured by observing their structural properties (like size, complexity, or functionality), or by expert judgment.

Expert judgment is a common estimation approach in industry [9, 12] because it is often quicker to collect subjective measures compared to the objective ones [26]. However, as stated by Briand et al. [4], an interesting question that has not been investigated in depth to date, is whether structural measures can perform as well as or better than experts in predicting quality attributes. With the aim of answering this question, Tomaszewski et al. show, in an industrial study on fault prediction, that structural measures are more accurate than expert estimations [25]. Turning towards the requirements volatility field, few empirical studies on structural measures and expert judgment are available. Zowghi and Nurmuliani [30] have performed a study on the impact of requirements volatility on software project and collected data on perceptions of requirements volatility. Ambriola and Gervasi have performed an empirical validation of structural measures of requirements stability [1].

In the present study, we want to add to this research by investigating subjective and objective measures of requirements volatility. This study1, is a replication of the study described in [17], where we found that measures of requirements size (number of lines, words, actors, and use cases) were good assessors of number of changes, while expert judgments were not.

In the present study (case study B), we examined requirements size measures and subjective volatility and correlated them to objective measures of volatility (see table 1). We analyzed data in retrospect for 21 use cases (UC) for a medium size software project and interviewed the stakeholders of the project about what they recalled of requirements volatility of the project we analyzed.

The remaining part of the paper proceeds as follows: section 2 describes the goals, hypotheses, and data collected in the present correlational study. The data analysis is described in section 3. Threats to validity are discussed in section 4. In section 5, we discuss the differences and similarities of the two studies. Conclusions are presented in section 6. The form distributed to the subjects and the collected subjective data is included in the appendix.

1For simplicity we will call our earlier case study “case study A” and the present one “case study B”.
2 Empirical study description

We followed the process suggested by Wohlin et al. [27] and Kitchenham et al. [13]. All materials of the study is available online to allow for possible replications [18].

2.1. Goals of the study

The goal of the present study is to investigate the ability of requirements measures to estimate the volatility of requirements. The goal can also be seen as an empirical validation of requirements structural measures and subjective volatility as measures of requirements volatility. This goal is also meant to confirm or falsify the results of our previous case study A [17].

2.2. Context selection

The present study was performed in retrospect at BAE Systems Hägglunds AB, Sweden. The company produces automotive systems with embedded software and is ISO9001 and ISO14001 certified. The project chosen for the study (host project) had as goal to construct the information system that runs on vehicles constructed by the company. We analyzed and collected data from the use case-based requirements specifications². This software system comprises 26 use cases and other non functional requirements described in two documents. At the time of the analysis the software system had been in operation for approximately 24 months. The Rational Unified Process (RUP) was used in the project. About fifteen people had worked on the software development project, including two project managers.

According to the terminology used in the studied project, a use case description should contain: overview, revision history, references, description, state-diagram, normal flow, alternative flows, special requirements, start conditions, end conditions, and extension points. The actors of the intended software system were described in a higher level requirements specification document called “use case summary”. The project started during summer 1999 and ended in 2003. The time delay between the project end and the case study was about two years.

The context selection of the case study can briefly be characterized as: 1) online, because it is performed in an industrial software development environment; 2) professional, i.e. non-student environment; 3) real, because it addresses a real problem; 4) specific, since it is focused on volatility measures.

The objects of the study were 26 use cases of the project at the company. Among these, four use cases were discarded, because the data available was incomplete. Another use case document was discarded because the file contained the description of two use cases and these were later split in two files. The final number of analyzed use cases was 21. Other documentation was used to gather data about the objects, in particular project plans, iteration plans, and test plans. The guidelines for filling in the forms distributed to the subjects of the study was included in an email sent to all participants (see appendix A). The measurement instruments used were email form, Microsoft Excel forms and Word for comparison, and Minitab (a software package for statistical analysis).

2.3. Variables selection

The entities analyzed in the project were requirements documents. Intuitively, the larger the document the more changes there are. Therefore, we believe that the size of requirements is an influential factor affecting volatility. The measures are described in table 1. These are quite intuitive, except for NREVISION which is described in section 2.3.2. The measurement rules, applied when collecting data, are described in [19].

2.3.1 State variables

The independent or state variables were the measures of size of UC, and subjective volatility. As suggested by Fenton and Pfleeger [7], size can be seen as composed of length, functionality, and complexity. In our case, NLINE and NWORD are measures of length, NACTOR and NSCENARIO are measures of complexity. The measures NLINE and NWORD are very similar to each other and collinear [19]. However, both were included in the study in order to explore all possible size measures and see which one is best correlated to volatility.

In our previous case study A, we collected data for the measure “number of use cases per requirement document” (NUC), which is a measure of functionality. This measure was not considered in the present study because the analyzed requirements documents were composed of one use case per document. One possible measure of volatility could also be “number of dependencies between use cases”, which is related to the internal attribute “complexity”. In our case it was not possible to collect data for this measure because there were no associations between use cases.

Subjective volatility was measured by subject ratings, which were collected manually by an e-mailed form (see appendix A). In order to determine the subjective volatility, we contacted 44 stakeholders at the company. Fifteen of them were software developers, the others had different roles like market responsible, project managers, quality assurance responsible, system engineers. Customers and end-users were not interviewed for confidentiality reasons.

²We are aware of the fact that some researchers do not consider use cases to be requirements [8, 14, 23, 28].
2.3.2 Response variables

In the available literature [3, 6, 10, 17, 21, 22, 24], volatility is treated as a quantitative measure. Likewise, we define requirements volatility as the amount of changes to a requirements document over time and we measure it as the sum of the change densities of a requirements document in time.

\[
\text{VOLATILITY} = \sum_{i=1}^{N\text{REVISION}} \frac{N\text{CHANGE}_i}{N\text{WORD}_i}.
\]

Our definition of requirements volatility is a function of: number of changes (NCHANGE), time measured in number of revisions (NREVISION), and size of the requirements document measured in NWORD. NCHANGE is a count of changed words per each revision, therefore, NWORD was chosen to calculate the change density (having the same unit of measurement). NREVISION expresses the time factor and is used to calculate VOLATILITY. Revisions are done to any files when changes are performed, but also to validate the UCs. A revision can include several changes or no changes at all. Usually, a large amount of revisions corresponds to a large amount of changes.

There is one difference between our operational definition of volatility and the ones available in the literature. We look at volatility document by document, instead of treating all requirements as one set. Our measure of volatility allows us to rate the volatility of each use case using a single number instead of a trend. This makes it possible to distinguish requirements documents with high levels of volatility from those with low levels.

For this study we choose to collect data for two response variables: VOLATILITY as defined above and total number of changes to a use case document in time (total NCHANGE). Total number of changes, which was a response variable in case study A, is useful in order to analyse volatility trends. VOLATILITY instead is a measure independent of the size of the UC documents and can be used to compare the volatility of use cases among each other.

Our dependent variables have been determined by comparing versions of requirements documents by means of a tool and counting the changes from one version of a document to the next.

Table 1. State and response variables

<table>
<thead>
<tr>
<th>Measures</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>State Variables</td>
<td>Subjective volatility: Subjective volatility is a good assessor of require-</td>
</tr>
<tr>
<td></td>
<td>ments volatility, i.e., there is significant correlation between the measures</td>
</tr>
<tr>
<td></td>
<td>of change to UC (total NCHANGE and VOLATILITY) and the rating of volatility</td>
</tr>
<tr>
<td></td>
<td>of the UCs made by the subjects.</td>
</tr>
<tr>
<td></td>
<td>H2: The structural measures of size of requirements (NACTOR, NSCENARIO, NL-</td>
</tr>
<tr>
<td></td>
<td>LINE, and NWORD) are good assessors of volatility.</td>
</tr>
<tr>
<td>Response Variables</td>
<td>Measures used to calculate VOLATILITY. Revisions are done to any files when</td>
</tr>
<tr>
<td></td>
<td>changes are performed, but also to validate the UCs. A revision can include</td>
</tr>
<tr>
<td></td>
<td>several changes or no changes at all. Usually, a large amount of revisions</td>
</tr>
<tr>
<td></td>
<td>corresponds to a large amount of changes.</td>
</tr>
</tbody>
</table>

Table 2. Data collected on use cases

<table>
<thead>
<tr>
<th>UC</th>
<th>NREVISION</th>
<th>Total NCHANGE</th>
<th>NSCENARIO</th>
<th>NACTOR</th>
<th>NLINE</th>
<th>NWORD</th>
<th>Subj. VOLATILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC1</td>
<td>4</td>
<td>2566</td>
<td>1</td>
<td>6</td>
<td>136</td>
<td>1145</td>
<td>0.21</td>
</tr>
<tr>
<td>UC2</td>
<td>2</td>
<td>1775</td>
<td>4</td>
<td>4</td>
<td>200</td>
<td>1428</td>
<td>0.36</td>
</tr>
<tr>
<td>UC3</td>
<td>2</td>
<td>923</td>
<td>3</td>
<td>4</td>
<td>146</td>
<td>813</td>
<td>0.21</td>
</tr>
<tr>
<td>UC4</td>
<td>11</td>
<td>3214</td>
<td>2</td>
<td>4</td>
<td>409</td>
<td>2451</td>
<td>0.07</td>
</tr>
<tr>
<td>UC5</td>
<td>5</td>
<td>4680</td>
<td>0</td>
<td>4</td>
<td>248</td>
<td>1168</td>
<td>0.00</td>
</tr>
<tr>
<td>UC6</td>
<td>5</td>
<td>2148</td>
<td>1</td>
<td>4</td>
<td>181</td>
<td>1412</td>
<td>0.07</td>
</tr>
<tr>
<td>UC7</td>
<td>2</td>
<td>1590</td>
<td>4</td>
<td>2</td>
<td>168</td>
<td>980</td>
<td>0.50</td>
</tr>
<tr>
<td>UC8</td>
<td>3</td>
<td>475</td>
<td>2</td>
<td>4</td>
<td>110</td>
<td>647</td>
<td>0.43</td>
</tr>
<tr>
<td>UC9</td>
<td>2</td>
<td>215</td>
<td>1</td>
<td>4</td>
<td>38</td>
<td>187</td>
<td>0.00</td>
</tr>
<tr>
<td>UC10</td>
<td>2</td>
<td>627</td>
<td>1</td>
<td>4</td>
<td>80</td>
<td>390</td>
<td>0.00</td>
</tr>
<tr>
<td>UC11</td>
<td>3</td>
<td>346</td>
<td>1</td>
<td>4</td>
<td>85</td>
<td>335</td>
<td>0.00</td>
</tr>
<tr>
<td>UC12</td>
<td>2</td>
<td>1596</td>
<td>6</td>
<td>2</td>
<td>240</td>
<td>1501</td>
<td>0.21</td>
</tr>
<tr>
<td>UC13</td>
<td>2</td>
<td>487</td>
<td>4</td>
<td>4</td>
<td>57</td>
<td>312</td>
<td>0.07</td>
</tr>
<tr>
<td>UC14</td>
<td>3</td>
<td>1208</td>
<td>4</td>
<td>4</td>
<td>131</td>
<td>877</td>
<td>0.07</td>
</tr>
<tr>
<td>UC15</td>
<td>2</td>
<td>406</td>
<td>2</td>
<td>2</td>
<td>50</td>
<td>288</td>
<td>0.36</td>
</tr>
<tr>
<td>UC16</td>
<td>2</td>
<td>627</td>
<td>1</td>
<td>2</td>
<td>101</td>
<td>547</td>
<td>0.14</td>
</tr>
<tr>
<td>UC17</td>
<td>2</td>
<td>485</td>
<td>2</td>
<td>2</td>
<td>86</td>
<td>561</td>
<td>0.14</td>
</tr>
<tr>
<td>UC18</td>
<td>2</td>
<td>610</td>
<td>3</td>
<td>2</td>
<td>101</td>
<td>565</td>
<td>0.29</td>
</tr>
<tr>
<td>UC19</td>
<td>3</td>
<td>527</td>
<td>3</td>
<td>2</td>
<td>108</td>
<td>481</td>
<td>0.21</td>
</tr>
<tr>
<td>UC20</td>
<td>2</td>
<td>325</td>
<td>0</td>
<td>1</td>
<td>32</td>
<td>72</td>
<td>0.07</td>
</tr>
<tr>
<td>UC21</td>
<td>2</td>
<td>319</td>
<td>0</td>
<td>1</td>
<td>31</td>
<td>72</td>
<td>0.07</td>
</tr>
</tbody>
</table>

2.4. Hypotheses formulation

Our hypotheses were the following:

\( H_1 \): Subjective volatility is a good assessor of requirements volatility, i.e., there is significant correlation between the measures of change to UC (total NCHANGE and VOLATILITY) and the rating of volatility of the UCs made by the subjects.

\( H_2 \): The structural measures of size of requirements (NACTOR, NSCENARIO, NLINE, and NWORD) are good assessors of volatility.

2.5. Operation

The preparation of the subjects was made by explaining the definition of requirements volatility, describing the forms and showing an example of how the form should be filled in (see appendix A). Each subject worked alone (in fact, the subjects did not know who was participating in the study) and could use unlimited time. They rated the relative volatility defined as trends in changes to UCs in the RUP phases inception, elaboration, construction, and transition. The rate was on a scale of three linguistic labels (low, medium, and high) plus a “I do not know” for those who have only been partially involved in the project or do not remember. We asked also for an estimation of a total volatility.
for each use case over the phases, and a total volatility for each phase for all the use cases. This allowed us to verify whether the subjective ratings were consistent.

Creating a form with four linguistic labels allowed us to fit the form on a single page that was easy and quick to fill in, in order to encourage the subjects to participate in the study. We asked the subjects not to read the documentation of the UCs in order to avoid letting the objective volatility affect the subjective volatility. This documentation contains some information (like the number of revisions, the description and the date of changes to the UC) which may hint the subjects about the volatility of the UC.

We collected data for the measures in table 1 starting from the first available revision of the UCs which were dated July 2001 (three months after the beginning of the project) and distributed the forms shown in appendix A. The data collection was semi-automatic, carried out by the authors by studying the historical documentation of the project (see table 5). From the first available revisions of requirements documents, all files were analyzed following the rules described in [19].

The study did not affect the development project because it was done in retrospect. We collected data by reading historical project documentation. Forty-four stakeholders were contacted at the company for the subjective data, and we received answers from eight of them. One of the answers was discarded because it was incomplete. The other seven answers were valid and were selected.

3 Results

To verify the two hypotheses we checked the distributions of data for normality. The data distributions were not normal for the variables NSCENARIO and NWORD, while the other distributions were normal. Therefore, we used non-parametric statistics and applied the Spearman correlation coefficient. In order to judge the significance of our correlations, we consider the Spearman cutoff of 0.435 as a reference point (for a sample of size 21 and a level of significance $\alpha = 0.05$) [46].

3.1. Descriptive statistics

Table 3 shows the descriptive statistics for the data collected. The columns SE Mean, and SDer state respectively mean standard error, and standard deviation. Please note that the data collected for the measures NACTOR, NSCENARIO, NLINE, and NWORD are not summed in time in contrast to what is done for total NCHANGE. The values of the measures (except total NCHANGE) were taken from the last revision of the requirement document analyzed.

As can be observed in table 3, except for NACTOR, the means of all the measures are larger than the medians, which might be due to outliers. Outliers affect the mean more than the median, therefore, it is preferred the median to the mean to express central tendency [29]. The measures with lowest variance are subjective volatility and VOLATILITY while total NCHANGE has the largest mean and standard deviation. Compared to NCHANGE, NLINE, and NWORD, the other measures have relatively low means, standard deviations, and variances. The low variance of NACTOR is expected, since UCs should be kept small according to common guidelines [11, 15].

All our measures have more than five non-zero data points. There is therefore sufficient variance in all the measures to proceed with the analysis [5].

3.2. Analysis of the first hypothesis

Because the rating of volatility made by the subjects yielded ordinal data, we applied a simple transformation (weighted average) to obtain ranked data. For the current analysis, we averaged the answers to one value for each UC. Each of the dependent variables were correlated separately with the transformed rating of volatility made by the subjects. Because all values of the Spearman correlation coefficient are below the cutoff (see table 4, subjective volatility), we can conclude that there is no significant correlation between the data collected for our measures and the subjective rating of volatility. For this hypothesis, the results of case study A were confirmed.

3.3. Analysis of the second hypothesis

The dependent variables (total NCHANGE and VOLATILITY) were correlated separately with the measures of size of UC. As can be observed in table 4, the measure total NCHANGE is highly correlated with NLINE, NWORD, and NACTOR. The same results were obtained in [17]. For the moment, we do not claim any causality (i.e. that larger use cases cause a higher number of changes). Further studies are needed to check the causality.

3 We counted the number of occurrences of each answer and multiplied the number obtained by 0 (low), 0.5 (medium), or 1 (high). Finally we summed the result and divided by 7 (the max number of subjects).
4 Validity evaluation

4.1. Threats to conclusion validity

One issue that could affect conclusion validity is the size of the sample data (21 UCs and seven subjects). If the sample size is small, non-parametric tests can lead to wrong conclusions. Therefore, we consider our results as preliminary. Data for NLINES, NWORDS, and NCHANGES have been calculated using a computerized tool and are therefore reliable. Data for NREVISION, NACTOR, and NSCENARIO have been collected by reading the documentation and is therefore subjective. However, we have defined measurement rules to keep the judgment as objective as possible [19]. The participants of the study formed a heterogeneous group with respect to their professional expertise and background. We could have chosen a homogenous group with the disadvantages of decreasing the number of subjects and affecting external validity. However, if the group is very heterogeneous the variations due to the differences among the subjects is bigger than the variations due to the treatments, which actually is a threat in our case.

4.2. Threats to construct validity

The subjective volatility is based on subjects’ memory. The subjects chosen were all involved in the host project as developers and managers. In order to obtain reliable subjects’ rating, we applied the alternative form reliability [20]. The subjects were asked to rate volatility in different ways: a) rating the volatility of each use case by RUP phase; b) rating the volatility of each use case across all phases; c) rating the volatility of the all set of use cases by RUP phase. After a transformation of the ordinal data in rank data, we checked the correlations between a), b) and c), obtaining values above 0.9 (which is a high correlation). The construct validity of the measures is ensured, since we theoretically validated them [16].

4.3. Threats to internal validity

Differences among subjects. Because the group of subjects was heterogeneous, error variance due to differences among subjects is reduced. Our subjects had different background, but it was not necessary to have previous knowledge about requirements engineering to be able to fill in the forms distributed. Therefore this threat was considered small.

Knowledge of the domain. All UCs belonged to the same universe of discourse, and the knowledge of the project’s domain did not affect internal validity.

Accuracy of collected data. Changes to requirements were not well documented. They were determined by comparing several versions of files. Rules of measurements were defined and can be found in [19]. However, there is a risk that the way we defined changes to requirements can be different from the subjects view of change.

Accuracy of responses. One factor affecting the reliability of the answers can be the time that had passed since the end of the project. The best time to collect reliable data and reduce the recall bias is debatable. Considering that the average length of a typical software development project at the company is about three years, two year of delay in a project context is a fairly long time. However, the subjects work on very few projects in parallel. Other factors affecting the subjective measures can be personal problems and mood. The developers might not have read the definition of volatility carefully and answered the form randomly. This threat is small, because we asked to rate the volatility in three different ways and there was consistency between the answers. The developers might even have read the documentation related to the UCs, affecting the subjective volatility by studying the objective data. However, this threat is small because it takes time to read the documentation of all the UCs and we believe that the developers were not willing to spend that time. This would have been a threat in case of correlation between the subjective and objective volatility.

Motivation of subjects. In order to get as many answers as possible, we distributed cinema tickets to the subjects who filled in the forms. However, we received only 8 answers with completed forms out of 44 contacted stakeholders. Seven received forms were valid, and four of them were filled in by the subjects most heavily involved in the project.

Other factors. To fill in the form required less than 30 minutes, therefore fatigue effects were not a relevant factor. Plagiarism could be checked easily, while influence among the subjects could not be controlled for and we could only trust the answers received.

4.4. Threats to external validity

Our results are based on two industrial projects performed in the same company, and therefore have local validity. However, the projects analyzed were quite different from each other (the systems developed, the developers, the size); furthermore, the materials used and the project were
real, and the subjects were professionals, therefore, we are quite confident that the results can be generalised. One threat could be the relatively small size of the project.

The number of subjects who answered is small, this can also be a threat. This is a general problem in empirical software engineering where it is typically difficult to identify subjects available for experiments [2].

5 Comparison of the two studies

Many studies are isolated and it is hard to understand how widely applicable the results are. One way to increase the applicability of empirical results is to replicate the study. However, even when replications are run, it is sometimes hard to understand the commonalities and the differences so that general conclusions can be drawn [2]. Therefore, in this section we will describe the differences among the two studies and draw some conclusions.

Replications of experiments can be classified in 3 major groups: replications that do not vary any research hypothesis, the ones that vary research hypothesis, and the replication that extend the theory [2]. Our case study is an internal replication (because it is performed in the same company), which falls in the second group. However, it is not a strict replication, because we vary the hypotheses, variables, and instrumentations in order to compensate for threats found in case study A. The problem is that the attempt to compensate for threats to validity may lead to changes such that the replication becomes less strict and therefore difficult to generalise.

A summary of the differences between the two studies can be seen in table 5. Differences between the studies concerning the state and response variables have been described in section 2.3.

The stakeholders contacted in both studies were employees at the same company and had different roles (from software developers to project managers to hardware developers). The material we distributed to the subjects is similar. In case study A we delivered a form where we described the definitions of volatility and relative volatility. In case study B we added the definitions of requirement and change. In both studies the form to be filled in was divided by the RUP phases. Furthermore, in case study B the subjects could choose among low, medium, high, and “I do not know”, while in the previous study there were only three choices (low, medium, high). These modifications were performed in order to minimise the threats to validity. The material was given in an email form in both studies.

The measurement rules for counting changes were different. In case study A we classified changes in minor, moderate, and major. In case study B we counted only the changed words. These differences affected the choice of the response variable in case study B. Differences in the context of the study and in the datasets are described in [19].

From an analysis of the data collected for the state and response variables common to the two studies, we obtain the following results: 1. Number of lines, words, and actors show high correlation to total number of changes in both studies. 2. Subjective volatility did not show any correlation to number of changes in both studies. However, we cannot consider these results as definitive. Further empirical studies are needed in order to confirm or falsify these results.

6 Conclusions

In this paper, we have described a retrospective case study on requirements volatility performed at BAE Systems Hägglunds AB, Sweden. We collected data for four structural measures of size and measures of change to UCs. We furthermore interviewed project stakeholders about what they recalled of requirements volatility. This study is an internal replication of case study A described in [17].

Summarizing the results, we found that the measures of length of UCs (number of lines and words) and the complexity measure number of actors are highly correlated with the measure total number of changes to UCs. The complexity measure number of scenarios did not show significant correlation. These results confirm the results of case study A, where we found that all requirements size measures correlated significantly with the total number of changes [17]. The results of another correlational study performed by the authors are also confirmed [19]. The results of all three studies imply that the measures of length of use cases are empirically validated in the specific environment and can be considered good estimators of total number of changes. With this result we scientifically prove that larger UCs are more volatile than smaller ones. The correlation seems to be linear [19], indicating that however we decompose a UC into smaller modules, the total number of changes will remain the same. This indicates that the notion of use cases

<table>
<thead>
<tr>
<th>Table 5. Differences between the two case studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Independent variables</strong></td>
</tr>
<tr>
<td>Number of developers</td>
</tr>
<tr>
<td>Project duration</td>
</tr>
<tr>
<td>Use case documents (UCD)</td>
</tr>
<tr>
<td>Average UCD size in kB (Medium)</td>
</tr>
<tr>
<td>Total number of req. documents</td>
</tr>
<tr>
<td>Supplementary req. documents</td>
</tr>
<tr>
<td>Total number of use cases</td>
</tr>
<tr>
<td>Reliability of subjective volatility</td>
</tr>
<tr>
<td>Inducement</td>
</tr>
<tr>
<td>Stakeholders contacted</td>
</tr>
<tr>
<td>Received valid answers</td>
</tr>
<tr>
<td>Dependent variables</td>
</tr>
<tr>
<td>Supplementary materials</td>
</tr>
<tr>
<td>Inducement</td>
</tr>
<tr>
<td>Stakeholders contacted</td>
</tr>
<tr>
<td>Received valid answers</td>
</tr>
<tr>
<td>Dependent variables</td>
</tr>
<tr>
<td>Supplementary materials</td>
</tr>
</tbody>
</table>

...
automatically forces developers to modularize requirements in such a way that requirements volatility can be easily predicted.

No structural size measure showed significant correlation to volatility. However, although not significant, the length measures are positively correlated to volatility in opposite to subjective estimations and complexity measures. The subjective volatility did not show any significant correlation to total number of changes (in both studies) and volatility (in the present study). Once again, the perceptions did not match our measures. There are many possible explanations for our results and further investigations are needed. The estimations of the subjects are reliable (see section 4.2) but different from our operational definition of volatility. This can mean that our operational definition is not measuring volatility as perceived by the subjects. It is also possible that the subjects did not reliably recall the evolution of all UCMs or that their view has been affected by late changes. Further factors might have contributed to subjective volatility, like the actual impacts of changes, priorities of use cases, functional details in general, or changes to design decisions. The perception of volatility in the industry is an open question and need to be investigated further, by performing a qualitative study. In any case, our results indicate that decision makers may take a high risk if basing decisions solely on subjective requirements volatility. Therefore, we suggest project managers at this company to complement their use of subjective estimation with simple structural measures such as ours in order to minimize this risk.

Although our approach currently uses use case based requirements, the defined measures are quite general and can therefore be used for all use case documents written in textual form. Furthermore, the measures NLINE and NWORD can be applied to any kind of requirement written in textual form. Our structural measures are simple, effective, easily interpreted, and completely automated.

Our results are based on two industrial projects performed in the same company, and therefore have local validity. However, the projects analyzed were quite different from each other (the systems developed, the developers, the size), therefore, we are quite confident that the results can be generalised. Further replications of this study in other contexts would further contribute in building a body of knowledge in the field of requirements volatility.

### A Instruments

**DEFINITIONS:**
- Requirement = a functionality (usually a use case) or a quality of the software system
- Change= any modification to the requirements (even small modifications like one word changed)

### Table 6. Subjective data collected on 21 use cases

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Inception</th>
<th>Elaboration</th>
<th>Construction</th>
<th>Transition</th>
<th>All phases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirement 1</td>
<td>H</td>
<td>M</td>
<td>L</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>Requirement 2</td>
<td>M</td>
<td>M</td>
<td>L</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>Requirement 3</td>
<td>L</td>
<td>L</td>
<td>M</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>In general for all requirements</td>
<td>D</td>
<td>M</td>
<td>L</td>
<td>H</td>
<td>D</td>
</tr>
</tbody>
</table>

- Volatility= the amount of changes to a requirement in time
- Relative volatility = the volatility of a requirement compared to other requirements.

For each requirement and RUP (rational unified process) phase, please indicate its relative volatility in the project by placing a letter (L, M, H, D) in each box.

L = low
M = medium
H = high
D = I do not know

More concretely, you should compare the amount of changes to each requirement to the amount of changes in average. See the example in table 7, where:
- inception = correspond to planning and initial requirements specification phase
- elaboration = more detailed requirements specification and initial design
- construction = detailed design of the system, coding and testing phases
- transition = deployment, transferring the system to the customer.

Thus, the meaning of the example table given above is the following: in the inception phase, the amount of changes to requirement 1 is high in relation to requirement 2 and it is very high in
relation to requirement 3. In the elaboration phase, the amount of changes to requirements 1 and 2 is higher in relation to requirement 3. The volatility of requirement 2 in the construction phase was higher than requirements 1 and 3. And so on.

References


