*Proceedings of the 17th IEEE International Conference Automated Software Engineering (ASE 2002)*, IEEE Computer Society Press, September 23-27, 2002, Edinburgh, Scotland.

# On CASE tool usage at Nokia

Alessandro Maccari, Claudio Riva
*Nokia Research Center*
*P. O. Box 407 – FIN 00045*
*Nokia Group, Finland*
*Tel.: +358 7180 08000 Fax: +358 7180 36308*
*{alessandro.maccari | claudio.riva}@nokia.com*

Francesco Maccari
*Università degli Studi di Siena,*
*Facoltà di Economia R. M. Goodwin*
*Piazza S. Francesco 53100 Siena, Italy*
*maccari2@unisi.it*

## Abstract

*We present the results of a research work targeted to understanding CASE tools usage in Nokia. By means of a survey questionnaire, we collected data aimed to identify what features are most useful and best implemented in current CASE tools according to senior developers and managers. With the aid of both descriptive and inferential statistical data analysis methods, we found out that the features that are rated most useful belong to the graphical editing, version management and document generation categories. The statistical methods we use allow us to extend the results to the whole population with a certain degree of confidence.*

*The analysis of the data seems to give the indication that there is a general level of dissatisfaction on the quality of currently available CASE tools. Also, there is evidence that some of the most advanced features (reverse engineering, code generation) are not deemed as useful as others.*

*Further research should focus on extending the survey to other types of industries, and attempt generalization of the results. This may constitute precious feedback for the software tools industry in order to develop products that correspond more to industry needs.*

## 1. Introduction

CASE (Computer Aided Software Engineering) tools are programs that automate certain parts of software development cycle. The term "CASE tool", however, defines products that are rather different from each other, and implement different subsets of a very vast amount of features. CASE tools may support the automation of the following generic software development activities: modelling (of requirements, design, architectures, etc.), implementation, maintenance (including evolution), documentation, configuration management, process management, quality, testing. Therefore, it is not really clear what functionality a CASE tool should provide.

It is not clear what consequences using a CASE tool bears. Many people seem to believe it automatically implies automating the software lifecycle. This, in turn, is assumed to have positive outcome on the final product. However, such belief is largely unjustified (as shown by Fenton, 1994 and Glass, 1999). In general, research work on CASE tools usage is insufficient (with a few exceptions: Orlikowski, 1993, Post, 1998, Iivari, 1998), and almost inexistent in the telecom software domain.

One exception is our previous study (Maccari, 2000), where we illustrated the outcome of a survey we carried out within Nokia Mobile Phones (the mobile handset software development unit of Nokia). The survey addressed the following two research questions.

- Among all the features that are offered by the various CASE tools, which are reputed most useful?
- In the CASE tools that are currently used in Nokia Mobile Phones, how satisfactorily are such features implemented?

The survey was conducted on a small population sample, but nonetheless showed interesting results. Simple descriptive analysis of the data indicated that CASE tools are mainly used as drawing, documentation and repository tools. Features involving automation of the software development activities were not regarded as very useful by the interviewees. This already contradicted the (rather widespread) assumption that CASE usage and software engineering automation are related.

Here we apply descriptive statistics to the same data more thoroughly, performing two kinds of inferential tests on the data set: non-parametric (sign) test, and cluster analysis. Descriptive statistics allow to understand the analyzed sample, and formulate founded hypotheses on the population based on the observed data. Inferential statistics, instead, provide the means to extend results to the whole population.

The results seem to be rather coherent with our previous conclusions, that are thus strengthened by this work. In particular, we identified 10 features (out of a total 33) that are rated "extremely well implemented" by the interviewees.

The conclusions seem to indicate that, relatively to our environment, software engineering automation is still fairly immature. What is currently available in terms of tools and methods does not seem to correspond to what the people judge as valuable. We present a research roadmap for this controversial and important subject.

## 2. Research background

The research background for the first study has already been described in our previous publication (Maccari, 2000). We will limit the discussion to the background behind the analysis hereby described.

### 2.1. Motivations

The motivation for extending the analysis presented in our previous paper is twofold. First, the descriptive analysis of the data thereby presented was, in a way, incomplete: we used the descriptive method since it allows a better understanding of the respondent sample; however, it does not provide a means to extend the conclusion to the whole population. Second, some of the issues that were raised by that research work needed further investigation. The survey was structured as a multiple-choice questionnaire, a form of data collection that is known to be inflexible and excessively summarizing. Having to do with a fairly large respondent set, we decided to sacrifice flexibility for speed and cost. In an industrial environment, such choices are often inevitable.

### 2.2. Objectives

We hereby pursue the following research objectives:

- understanding the usefulness of the various CASE features as rated by Nokia's mobile handset software developers and managers;
- understanding how the developers rate the implementation of such features in existing CASE tools.

The more general research question that we address can be phrased as follows: “how far is the current offer of CASE tools from what is really required in practice?”. The paper attempts an answer to this question limited to our company’s environment. We advocate further research to validate the generalization of the results we obtained.

## 3. Validity threats

We will consider two different types of validity: internal, which refers to the characteristics of this research work, and external, which concerns the generalization of results to different contexts.

### 3.1. Internal validity

The object population consists of all software developers and managers in Nokia's mobile handset development unit, Nokia Mobile Phones (NMP).

The collected data has been organized in two data sets, containing, respectively, the answers to the "ideal CASE tool" evaluation and to the "used CASE tools" evaluation. We will refer to them as to the "ideal data set" and "used data set".

Each evaluated feature corresponds to a variable containing the given score. Therefore, the object of the analysis is the set of variables, which is the same in the "ideal CASE tool" and "used CASE tools" data sets.

Three of the internal validity threats that we identify originate from the *choice of the population sample*.

The first one is the *non-representativeness of the chosen sample*. The list of all population members was not available in the beginning of the survey. Therefore, we could not determine the inclusion probability of every sample unit. The sample is formed by 49 software developers and managers that were thought to have experience in CASE tools. We cannot formulate the hypothesis that the sample is representative of the population.

The second validity threat concerns the relatively high *non-response rate*. We sent the questionnaire by e-mail: 14 people answered evaluating the "ideal CASE tool", while only 12 of them evaluated the "used CASE tools". The occurred non-response rates have been 70.8% for the "ideal CASE tool" evaluation and 75% for the "used CASE tools" evaluation.

Such response rates are expectable for mail questionnaires (Edwards, 1972). In any case, they are too high to consider imputing the responses, i.e. filling them with the use of *ad hoc* methods.

The third threat to validity is the *low number of observations* per each variable. For this reason, it has not been possible to analyse the five evaluated CASE tools separately. We received seven answers for Rational Rose 98 [RationalRose], two for Object Time [ObjecTime] and one each for Prosa [Prosa], Rhapsody [Rhapsody] and Qualiware QLM [QLM]. This prevented us from drawing conclusions about the quality of any CASE tool. The purpose of the analysis on the "used CASE tools" data set is therefore to investigate a sort of "average" quality of those five.

Two other internal validity threats are caused by *the structure of the data collection method (questionnaire)*, especially as concerns the use of scores.

The first is the *qualitative domain for scores*. The acceptable scores are the integer numbers from one to five. These, in practice, correspond to qualitative scores. Therefore, scores should not be processed as numbers, even though they can be ordered.

The fifth validity threat is *the subjectivity in the interpretation of scores:* a score of, say, 3 may not have the same meaning for different respondents: some people may think of 3 as a poor score, while others may place it

just as an average score. We are forced to ignore it when processing data.

### 3.2. External validity

Generalization of the results of this research to different contexts is subject to a number of validity threats.

We claim that the validity of the conclusions can be accepted only with reference to our object population. Our sampling method (choice of a subset of Nokia Mobile Phones chief developers and architects that were known to the authors of the work and other people to be working with tools) does not imply representativeness of the sample. Unfortunately, in very large companies such as Nokia, it is virtually impossible to reach all the members of the target population (even mailing lists are not accurate, since it is not possible to send a mail to all employees). Therefore, we cannot eliminate this validity threat, but should take it into account when attempting to generalize the results.

Extension of our results to the whole population sample (Nokia Mobile Phones developers and managers) or, even more, to other contexts (e.g. different companies) is particularly risky. To be correctly based, any further study about the same subject that is performed in a different environment should not assume any of our results as universally valid. Instead, it should try to repeat the same steps and generate new results that are (hopefully) valid for other contexts. When a reasonably large amount of research has been performed, the results that show consistency may be considered universally valid with reasonable certainty.

## 4. Data analysis

The "ideal data set" is composed of 14 observations, while the "used data set" has only 12 observations. Both data sets contain 33 variables.

In order to pursue our research goals, we have carried out the following statistical analyses:

- analysis of the "ideal data set" with the purpose of identifying the most useful CASE tool features;
- analysis of the "used data set" with the purpose of synthesizing the CASE tool evaluations given by the interviewees;
- comparison between the results of the two analyses in order to check whether the used CASE tools provide a satisfactory implementation of the most useful features.

The basic idea is to perform two different levels of analysis on the two data sets: descriptive analysis has the purpose of getting some indications about the empirical shape of the variables; inferential analysis aims to divide the evaluated features into groups on the basis of either their importance (or quality of their implementation) as deduced from the inferential process.

The choice of suitable methods for these analyses has been substantially influenced by the presence of a reduced number of observations per variable.

### 4.1. Descriptive analysis

Generally, the purpose of descriptive analysis is twofold: first, having a first rough idea of the characteristics of the data; second, gathering some useful indication for hypothesis formulation.

The use of any inferential method (see section 4.2) must be coherent with the initial hypotheses. This aspect is particularly relevant in our case, since we are not experienced enough in this field of application to be able to make any *a priori* assumption.

#### *4.1.1 Methods*

Histograms seem to be the most suitable instrument to achieve the specified aim, as they show whether variables tend to assume high, low or medium scores. They also give an idea about their distribution, symmetry, kurtosis and concentration around a central value.

In addition to histograms, we need a position indicator, i.e. a value that summarizes the information contained in the collected sample for each variable. For this purpose, we decided to use the median score.

We did not use the mode mainly because some of the responses are not uni-modal: in these cases, there is no consensus, so the mode is not appropriate. On the other hand, the mean is not suitable for variables with qualitative domain and, moreover, extreme scores too easily influence it. Finally, we cannot be sure of the existence of the real mean, because of the lack of any assumption on the distribution of the variables.

In order to enrich the information provided by the median scores, we will show other simple statistics. In particular, quartiles, maximum score and minimum score give an idea about the concentration of the data. The number of observations gives an indication about the degree of reliability of results. A useful reference for basic statistics is Hoel, 1966.

The main reason why we have performed a descriptive analysis is taking a decision about plausible hypotheses to make. Besides, we will use it also to obtain a first classification of CASE tool features as regards both evaluations.

Therefore, we will put in evidence all the features that possess the following properties:

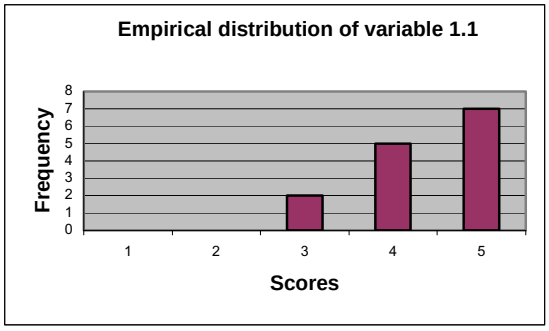
§ the median score is higher than or equal to 3.5;
§ the first quartile is higher than or equal to 3;
§ the third quartile is higher than or equal to 4.

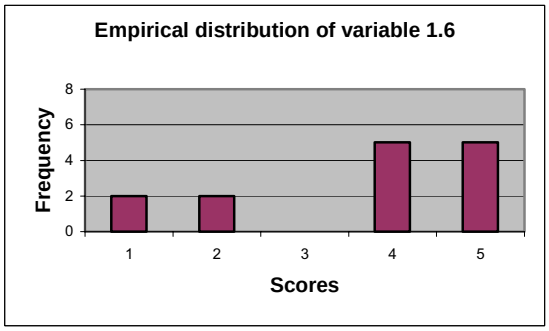
Variables that fulfil all the three conditions have been considered indicative of high importance of features for the first data set and of satisfactory implementation for the second one.

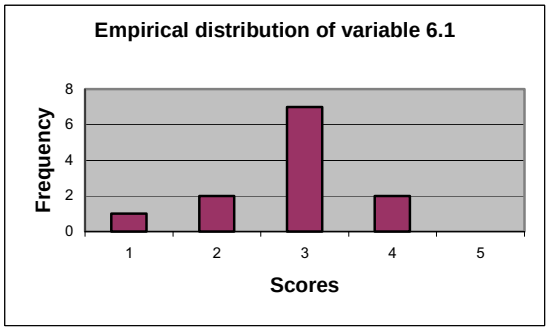
### *4.1.2 Application*

Due to lack of space, only some histograms that relate to the "ideal data set" are presented in Figure 1. However, they show some interesting recurrent characteristics of the gathered responses: high proportion of high scores (variable 1.1), multi-modal response (variable 1.6), concentration around medium scores (variable 6.1) and no significant concentration of the scores (variable 8.6).

As regards the choice of initial hypotheses, we do not feel like assuming to know the real distribution of the variables; therefore, the most suitable inferential analysis methods will be distribution-free. The hypothesis of symmetry of the variables does not look to be plausible either, since some scores do not tend to concentrate around a central value, as we can see, for instance, from the histogram of variable 1.1.



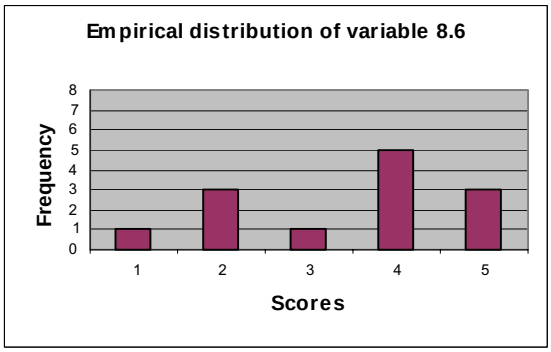




**Figure 1. Empirical distribution of some interesting variables.**

Considering the lack of assumption together with the discreet and qualitative nature of the variables, our priority in the inferential analysis will definitely be to look for methods that have a very wide range of applications and whose effectiveness is not bound to any strong restrictions.

From the observation of histograms, it is also clear that most of the assigned scores to the ideal CASE tool features are quite high. This means that we will discriminate among features that are considered somewhat useful by the respondents.

We will not show any histograms for the "used data set", because they put in evidence the same characteristics as for the "ideal data set" variables. The only difference is a general lower level of the scores. This constitutes some evidence of defective implementation in the used CASE tools.

Tables A and B in Appendix A contain some descriptive statistics of the sampled variables in the two data sets. The rows in bold indicate, in both tables, those features whose scores meet the conditions specified in section 4.1.1. A careful examination of the tables confirms the idea that the respondents assigned considerably higher scores to the usefulness of features than to their quality of implementation in the used CASE tools.

## 4.2. Inferential analysis

Inferential analysis allows us to divide the evaluated features in different categories, on the basis of assigned scores. This can be done for both the "ideal data set" and the "used data set".

The characteristics of our case of study which mostly bind the choice of inferential methods are the low number of observations per each variable, the qualitative nature of the scores and the absence of hypotheses on the distribution of variables (see section 3).

We believe that the two inferential methods that best meet the above conditions are nonparametric tests of hypothesis and cluster analysis. Nonparametric tests are distribution free; therefore they need no assumptions on the shape of the variables distribution. Cluster analysis methods are essentially applied when no *a priori* assumption can be made, since they just divide the collected data in the best possible way, based on the distances between the objects.

### *4.2.1 Nonparametric tests – Methods*

We test the following hypothesis system on the median of each variable in the two data sets:

$$\begin{cases} H_0 : M_i = 3 \\ H_1 : M_i > 3 \end{cases}$$

Where $M_i$ is the real median of the *i*th variable, while $H_0$ and $H_1$ are, respectively, the *null* and the *alternative* hypotheses.

We choose the threshold value 3 as a separator between high and low scores.

This test can be interpreted as follows: if the empirical evidence gives indications towards the refusal of $H_0$ (null hypothesis) in favour of $H_1$ (alternative hypothesis), we can conclude that the interviewees consider the feature very useful, or well implemented in the used CASE tools.

We initially consider two non-parametric statistics: the Sign and the Wilcoxon statistics.

Both work well with a reduced number of observations, since their distributions have been exactly calculated for more than one observation. They also adapt rather well to discreet variables, even though they were introduced essentially for continuous ones.

In fact, their application to discreet variables causes the loss of all ties observations, i.e. those observations that are equal to the value specified in the null hypothesis.

We favour the use of the Sign test (like in Siegel and Castellan, 1988) since, unlike the Wilcoxon test; it does not assume the symmetry of the object variable, which is not very plausible in our case (see section 3). For each variable, the key output is represented by the observed significance of the null hypothesis: the lower the observed significance is, the less compatible the null hypothesis is with the sample data (see Lehmann, 1997).

We classify the features, for both data sets, according to different levels of observed significance. We also show results of the execution of Sign test to the "ideal data set" and to the "used data set" (see 4.2.3).

We apply the Sign test to each data set independently (instead of performing a matched pairs test on the differences between the correspondent variables). We do this to obtain an independent classification of features for each evaluation.

### *4.2.2 Cluster analysis – Methods*

Cluster analysis is used to organize collections of data into meaningful structures. It can be applied to both cases and variables. Although the term "cluster analysis" encompasses different algorithms, the most suitable in our case seems to be "k-means clustering". It allows us to pre-determine the final number of clusters and pursues the most significant classification (see Mac Queen, 1967).

Our scale is formed of five scores. The optimal choice is to split the features in three groups. The resulting clusters should be sufficiently distinguished from each other, and small enough to provide a reasonable selection of features.

While other clustering methods (e.g. hierarchical methods) do not need any *a priori* assumptions, k-means method includes ANOVA significance testing on the differences among the resulting groups. ANOVA is based on the hypothesis of normal distribution (Scheffé, 1967). However, this does not represent a strong bind to its application: ANOVA test is performed with Snedecor F statistic, and therefore is remarkably robust with non-normal distributions (see Lindman 1992). Finally, despite the fact that assumption violation affects the result of the test, it does not affect the output of the algorithm.

Instead, the large number of item non-responses may threaten the effectiveness of k-means clustering.

No detailed information about non-respondents was available; therefore we cannot apply any refined imputation methods.

We get around this problem by substituting the missing data with the mean values of the correspondent variable. This may be a rough method, but it is the only one that was easily supported by the tool we used.

### *4.2.3 Nonparametric tests - Application*

For both data sets, we form three groups according to the level of observed significance: respectively, lower than 0.01, between 0.01 and 0.05, higher than 0.05.

The first group contains the most useful features for the "ideal data set", and the best implemented features for the "used data set". The results of this method are shown in Tables C and D.

### *4.2.4 Cluster analysis – Application*

ANOVA testing results indicate low plausibility for the hypothesis of no difference between the resulting groups. Therefore, it is possible to interpret each cluster on the basis of scores level.

This way, we can identify the cluster containing extremely useful features (ideal data set) and the cluster containing very well implemented features (used data set). The results are very similar to those obtained with the

Sign test application. We will illustrate them in the following section.

### 4.3. Commonality and difference

We now try to combine the results in order to draw some conclusions. From the two evaluations, we locate the features whose scores are high according both to the Sign test and the cluster analysis.

To our surprise and relief, the two methods lead to very similar results as regards the subdivision of features. Therefore, we refer only to Sign test results, shown in tables C and D in the Appendix.

We compare the two tables, in order to check how satisfied CASE users are with the features that they have rated "extremely useful". Table E shows the classification of the extremely useful features set according to their rated quality of implementation in the used CASE tools. All the features that were rated "useful" (table C) seem to have been also rated "unsatisfactorily implemented".

### 4.4. Final remarks

The one presented here is a pilot study, aimed to: a) giving general indications regarding the opinions of NMP managers and software developers about CASE tools; b) giving rise to further research, especially as regards analysis methods that may be used in similar cases.

Expectedly, sending questionnaires by e-mail has not been very effective. Direct interviews, although difficult to lead and hard to interpret, would probably reduce non-response rates and allow more detailed investigations on single CASE tools. Obtaining an evaluation for single CASE tools is the most important of our research target. There is still a lot of work to do in this direction.

## 5. Conclusion and research roadmap

This analysis presented here confirms that, in spite of the improvements that have been reached in the past few years, the area of computer-aided software engineering is still immature. The available CASE tools are actually fairly different from one another. Very often the features they provide match only partially those that are required by the users. This phenomenon emerges pretty well from our statistical analyses, but, due to the validity threats explained in section, has little potential for generalization.

We advocate further research work in the following areas.

- Some controversial points should be studied more: in particular, the reasons why certain features are rated "extremely useful" but "not well implemented" should be investigated. This can provide feedback to tool vendors for improvement of their products.
- The impact of CASE tools usage on software development should be studied. Does tools usage really make software developers more productive? is software developed using CASE tools better? less buggy? easier to maintain? answers to these questions are needed.
- A thorough evaluation of single tools is needed. Each tool has its strong and weak points. What are they? we have found very little in the literature that helps us answer this question (see e.g. Post, 1998 or Hendrickson, 1999).
- Research should be performed in organizations different than Nokia. The research framework we present here can be reused for this. The community should dedicate more effort to this issue.

At Nokia we commit to investigating the matter further. As an example, The Eureka Σ!2023 Programme, ITEA project ip00004 (CAFÈ) project is currently looking into issues relating to the application of tools to product family engineering. We plan to address some of the research questions raised by this study. Other research should be performed, especially in the context of different organizations.

## 6. References


B. Edwards, *Statistics for business students,* 1st Edition, Collins, 1972.

N. Fenton, S. Lawrence Pfleger, R. L. Glass, *Science and substance: a challenge to software engineers,* IEEE Software, July 1994, pp. 86-95.

R. L. Glass, *The realities of software technology – Payoffs,* Communications of the ACM, February 1999.

E. Hendrickson, *Evaluating CASE tools,* Software Testing & Quality Engineering, January/February 1999, p. 38-42.

P. G. Hoel, *Elementary statistics*, in Wiley series in probability and mathematical statistics, John Wiley and Sons, 1966.

J. Iivari, J. Maansaari, *The usage of systems development: why are we stuck to old practices?,* Information and Software Technology 40 (1998), p. 501-510, Elsevier Science, 1998.

E. L. Lehmann, *Significance level and power*, in O. F. Hamouda and J. C. R. Rowley, *Foundations of probability, econometrics and economic games*, p. 29-38, Cheltenam: Elgar, 1997.

H. R. Lindman, *Analysis of variance in experimental design,* in Springer text in statistics, Springer, 1992.

A. Maccari, C. Riva, *Empirical evaluation of CASE tool usage at Nokia,* Fourth International Conference on Empirical Assessment & Evaluation in Software Engineering, Keele, UK, April 2000, accepted for publication on the International Journal of Empirical

Software Engineering, Kluwer academic Publishers, December 2000.

J. Mac Queen, *Some methods for classification and analysis of multivariate observations*, in L. M. Le Cam and J. Neyman, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. I: Statistics, p. 281-297, University of California Press, Berkeley and Los Angeles, 1967.

[ObjecTime]: see http://www.objectime.com/

W. J. Orlikowski, *CASE tools as organizational change investigating incremental and radical changes in systems development,* Management Information Systems Quarterly, vol. 17 n. 3, September 1993.

G. Post, A. Kagan, R. T. Keim, *A comparative evaluation of CASE tools,* The Journal of Systems and Software 44 (1998), p. 87-96, Elsevier Science, 1998.

[Prosa]: see http://www.insoft.fi/

[QLM]: see http://www.qualiware.com/

[RationalRose]: see http://www.rational.com/

[Rhapsody]: see http://www.ilogix.com/

H. Scheffé, *The analysis of variance*, in Wiley Publications in statistics, John Wiley and Sons, 1967.

S. Siegel, N. J. Castellan, *Nonparametric statistics for the behavioral sciences*, McGraw-Hill, 1988.

M. Wood, J. Daly, J. Miller, M. Roper, *Multi-method research: an empirical investigation of object-oriented technology,* The Journal of Systems and Software 48 (1999), 13-26.

## Appendix A – Tables

| *Variable* | *No. Obs.* | *Median score* | *Min score* | *Max score* | *First quartile* | *Third quartile* |
|---|---|---|---|---|---|---|
| **1.1** | **14** | **4.5** | **3** | **5** | **4** | **5** |
| **1.2** | **14** | **4** | **3** | **5** | **3.5** | **5** |
| **1.3** | **14** | **5** | **2** | **5** | **4** | **5** |
| **1.4** | **14** | **3.5** | **2** | **5** | **3** | **4** |
| **1.5** | **14** | **5** | **3** | **5** | **4** | **5** |
| 1.6 | 14 | 4 | 1 | 5 | 2 | 5 |
| 1.7 | 11 | 4 | 1 | 5 | 2.5 | 4.5 |
| **1.8** | **14** | **5** | **4** | **5** | **5** | **5** |
| **1.9** | **14** | **3.5** | **1** | **5** | **3** | **5** |
| **2.1** | **14** | **4** | **1** | **5** | **3** | **5** |
| 2.2 | 14 | 4 | 2 | 5 | 2.5 | 5 |
| **3.1** | **11** | **5** | **3** | **5** | **3.5** | **5** |
| 3.2 | 14 | 3 | 1 | 5 | 2 | 3 |
| 3.3 | 14 | 2.5 | 1 | 5 | 2 | 3 |
| **4.1** | **14** | **5** | **3** | **5** | **4** | **5** |
| **4.2** | **14** | **5** | **4** | **5** | **4** | **5** |
| **4.3** | **14** | **5** | **3** | **5** | **4** | **5** |
| 4.4 | 14 | 4 | 1 | 5 | 2 | 4.5 |
| 4.5 | 14 | 3 | 2 | 5 | 3 | 3 |
| **5.1** | **12** | **4** | **2** | **5** | **3** | **4** |
| **5.2** | **13** | **5** | **2** | **5** | **4** | **5** |
| 6.1 | 12 | 3 | 1 | 4 | 2 | 3 |
| 6.2 | 13 | 3 | 1 | 4 | 2 | 3 |
| 6.3 | 13 | 3 | 1 | 5 | 3 | 4 |
| 7.1 | 12 | 3 | 1 | 4 | 3 | 4 |
| 7.2 | 12 | 3 | 1 | 4 | 3 | 3 |
| 8.1 | 13 | 4 | 1 | 5 | 2 | 4.5 |
| **8.2** | **14** | **3.5** | **2** | **5** | **3** | **4.5** |
| **8.3** | **12** | **3.5** | **2** | **5** | **3** | **4** |
| 8.4 | 13 | 3 | 2 | 5 | 2.5 | 4 |
| **8.5** | **13** | **4** | **2** | **5** | **3** | **4.5** |
| 8.6 | 13 | 4 | 1 | 5 | 2 | 4 |
| 8.7 | 12 | 4 | 1 | 5 | 2 | 4 |

**Table A – Descriptive statistics of the Ideal CASE tools.**

| *Variable* | *No. Obs.* | *Median score* | *Min score* | *Max score* | *First quartile* | *Third quartile* |
|---|---|---|---|---|---|---|
| **1.1** | **14** | **4.5** | **3** | **5** | **4** | **5** |
| **1.2** | **14** | **4** | **3** | **5** | **3.5** | **5** |
| **1.3** | **14** | **5** | **2** | **5** | **4** | **5** |
| **1.4** | **14** | **3.5** | **2** | **5** | **3** | **4** |
| **1.5** | **14** | **5** | **3** | **5** | **4** | **5** |
| 1.6 | 14 | 4 | 1 | 5 | 2 | 5 |
| 1.7 | 11 | 4 | 1 | 5 | 2.5 | 4.5 |
| **1.8** | **14** | **5** | **4** | **5** | **5** | **5** |
| **1.9** | **14** | **3.5** | **1** | **5** | **3** | **5** |
| **2.1** | **14** | **4** | **1** | **5** | **3** | **5** |
| 2.2 | 14 | 4 | 2 | 5 | 2.5 | 5 |
| **3.1** | **11** | **5** | **3** | **5** | **3.5** | **5** |
| 3.2 | 14 | 3 | 1 | 5 | 2 | 3 |
| 3.3 | 14 | 2.5 | 1 | 5 | 2 | 3 |
| **4.1** | **14** | **5** | **3** | **5** | **4** | **5** |
| **4.2** | **14** | **5** | **4** | **5** | **4** | **5** |
| **4.3** | **14** | **5** | **3** | **5** | **4** | **5** |
| 4.4 | 14 | 4 | 1 | 5 | 2 | 4.5 |
| 4.5 | 14 | 3 | 2 | 5 | 3 | 3 |
| **5.1** | **12** | **4** | **2** | **5** | **3** | **4** |
| **5.2** | **13** | **5** | **2** | **5** | **4** | **5** |
| 6.1 | 12 | 3 | 1 | 4 | 2 | 3 |
| 6.2 | 13 | 3 | 1 | 4 | 2 | 3 |
| 6.3 | 13 | 3 | 1 | 5 | 3 | 4 |
| 7.1 | 12 | 3 | 1 | 4 | 3 | 4 |
| 7.2 | 12 | 3 | 1 | 4 | 3 | 3 |
| 8.1 | 13 | 4 | 1 | 5 | 2 | 4.5 |
| **8.2** | **14** | **3.5** | **2** | **5** | **3** | **4.5** |
| **8.3** | **12** | **3.5** | **2** | **5** | **3** | **4** |
| 8.4 | 13 | 3 | 2 | 5 | 2.5 | 4 |
| **8.5** | **13** | **4** | **2** | **5** | **3** | **4.5** |
| 8.6 | 13 | 4 | 1 | 5 | 2 | 4 |
| 8.7 | 12 | 4 | 1 | 5 | 2 | 4 |

**Table B – Descriptive statistics of the Used CASE tools.**

*The rows in bold contain the variables whose median score is higher than or equal to 3.5 and the first quartile is higher than or equal to 3 and the third quartile is higher than or equal to 4.*

**Table C – Classification of ideal CASE tool features based on the results of Sign test**

*Extremely useful features*

| *Variable* | *Feature* | *Sign test observed significance* |
|---|---|---|
| 1.1 | **Support for standard UML notation** | 0.0002 |
| 1.2 | **Be able to edit all the UML diagrams** | 0.0005 |
| 1.3 | **Perform diagram analysis (e.g. consistency checks)** | 0.0017 |
| 1.5 | **Support design specification** | 0.0002 |
| 1.8 | **Be intuitive and easy to use** | 0.0001 |
| 3.1 | **Utilise a repository** | 0.0020 |
| 4.1 | **Allow easy editing of text notes inside diagrams** | 0.0001 |
| 4.2 | **Allow easy editing of graphical data (diagrams)** | 0.0001 |
| 4.3 | **Automatically generate well structured documents from models** | 0.0001 |
| 5.2 | **Manage versioning** | 0.0032 |

*Useful features*

| *Variable* | *Feature* | *Sign test observed significance* |
|---|---|---|
| 1.4 | **Support requirements specification methods** | 0.0352 |
| 2.1 | **Generate correct, well structured code** | 0.0193 |
| 8.5 | **Runtime analysis** | 0.0195 |

*Other features*

| *Variable* | *Feature* | *Sign test observed significance* |
|---|---|---|
| 1.6 | **Help performing simulation** | 0.0898 |
| 1.7 | **Help building prototypes** | 0.0898 |
| 1.9 | **Allow concurrent editing of the same model** | 0.0898 |
| 2.2 | **Help in the debugging phase** | 0.0730 |
| 3.2 | **Be able to read and analyse existing code** | 0.9922 |
| 3.3 | **Be able to obtain models of existing, non-modelled code** | 0.9453 |
| 4.4 | **Support hypertext navigation in the model** | 0.1938 |
| 4.5 | **Support free form attachments in the model** | 0.6875 |
| 5.1 | **Help tracking modification within the model** | 0.0547 |
| 6.1 | **Track project deliverables in the model** | 0.8125 |
| 6.2 | **Analyse and report on project status** | 0.8906 |
| 6.3 | **Support process (lifecycle) management** | 0.0625 |
| 7.1 | **Help in managing quality parameters** | 0.1094 |
| 7.2 | **Provide support for risk management** | 0.6875 |
| 8.1 | **Automatic testing** | 0.1938 |
| 8.2 | **Module testing** | 0.0898 |
| 8.3 | **Regression testing** | 0.1445 |
| 8.4 | **Integration testing** | 0.2539 |

| | | |
|---|---|---|
| 8.6 | **Analyse test coverage** | 0.1938 |
| 8.7 | **Support automatic test result verification** | 0.1719 |

**Table C – Classification of used CASE tool features based on the results of Sign test**

*Very well implemented features*

| *Variable* | *Feature* | *Sign test observed significance* |
|---|---|---|
| 4.2 | **Allow easy editing of graphical data (diagrams)** | 0.0078 |

*Well-implemented features*

| *Variable* | *Feature* | *Sign test observed significance* |
|---|---|---|
| 1.1 | **Support for standard UML notation** | 0.0193 |
| 1.3 | **Perform diagram analysis (e.g. consistency checks)** | 0.0352 |
| 4.1 | **Allow easy editing of text notes inside diagrams** | 0.0352 |

*Other features*

| *Variable* | *Feature* | *Sign test observed significance* |
|---|---|---|
| 1.2 | **Be able to edit all the UML diagrams** | 0.0730 |
| 1.4 | **Support requirements specification methods** | 0.8906 |
| 1.5 | **Support design specification** | 0.1445 |
| 1.6 | **Help performing simulation** | 0.9102 |
| 1.7 | **Help building prototypes** | 0.9453 |
| 1.8 | **Be intuitive and easy to use** | 0.2266 |
| 1.9 | **Allow concurrent editing of the same model** | 0.9648 |
| 2.1 | **Generate correct, well structured code** | 0.3633 |
| 2.2 | **Help in the debugging phase** | 0.5000 |
| 3.1 | **Utilise a repository** | 0.3438 |
| 3.2 | **Be able to read and analyse existing code** | 0.9961 |
| 3.3 | **Be able to obtain models of existing, non-modelled code** | 1.0000 |
| 4.3 | **Automatically generate well structured documents from models** | 0.7461 |
| 4.4 | **Support hypertext navigation in the model** | 0.7734 |
| 4.5 | **Support free form attachments in the model** | 0.8906 |
| 5.1 | **Help tracking modification within the model** | 0.9922 |
| 5.2 | **Manage versioning** | 0.9922 |
| 6.1 | **Track project deliverables in the model** | 1.0000 |
| 6.2 | **Analyse and report on project status** | 1.0000 |
| 6.3 | **Support process (lifecycle) management** | 0.9688 |
| 7.1 | **Help in managing quality parameters** | 1.0000 |
| 7.2 | **Provide support for risk management** | 1.0000 |
| 8.1 | **Automatic testing** | 0.9688 |

| | | |
|---|---|---|
| 8.2 | **Module testing** | 0.7734 |
| 8.3 | **Regression testing** | 0.9688 |
| 8.4 | **Integration testing** | 0.9688 |
| 8.5 | **Runtime analysis** | 0.8906 |
| 8.6 | **Analyse test coverage** | 1.0000 |
| 8.7 | **Support automatic test result verification** | 0.9688 |

**Table M – Classification of very useful features according to their implementation in the evaluated CASE tools**

*Very well implemented features*

| Feature | Ideal CASE tool evaluation | | Used CASE tools evaluation | |
|---|---|---|---|---|
| | Median score | Sign test observed significance | Median score | Sign test observed significance |
| **Allow easy editing of graphical data (diagrams)** | 5.0 | 0.0001 | 4.0 | 0.0078 |

*Well-implemented features*

| Feature | Ideal CASE tool evaluation | | Used CASE tools evaluation | |
|---|---|---|---|---|
| | Median score | Sign test observed significance | Median score | Sign test observed significance |
| **Support for standard UML notation** | 4.5 | 0.0002 | 4.0 | 0.0193 |
| **Perform diagrams analysis (e.g. consistency checks)** | 5.0 | 0.0017 | 4.0 | 0.0352 |
| **Allow easy editing of text notes inside diagrams** | 5.0 | 0.0001 | 4.0 | 0.0352 |

*Not well-implemented features*

| Feature | Ideal CASE tool evaluation | | Used CASE tools evaluation | |
|---|---|---|---|---|
| | Median score | Sign test observed significance | Median score | Sign test observed significance |
| **Be able to edit all the UML diagrams** | 4.0 | 0.0005 | 4.0 | 0.0730 |
| **Support design specification** | 5.0 | 0.0002 | 4.0 | 0.1445 |
| **Be intuitive and easy to use** | 5.0 | 0.0001 | 3.0 | 0.2266 |
| **Utilise a repository** | 5.0 | 0.0020 | 3.5 | 0.3438 |
| **Automatically generate well structured documents from models** | 5.0 | 0.0001 | 2.5 | 0.7461 |
| **Manage versioning** | 5.0 | 0.0032 | 1.5 | 0.9922 |