The impact of COTS components on software quality in IT Industry: A Survey and its analysis

N. Gnanasankaran
MCA Department
KSR College of Arts & Science
Tiruchengode – 637 215, India
Sankarn.iisc@gmail.com

K. Iyakutti
Department of Physics & Nanotechnology
SRM University, Kattankulathur, Chennai – 603 203, India
iyakutti@gmail.com

K. Alagarsamy
Department of Computer Applications
Madurai Kamaraj University, Madurai – 625 021, India
alagarsamy.mku@yahoo.co.in

S. Natarajan
School of Physics
Madurai Kamaraj University, Madurai – 625 021, India
s_natarajan50@yahoo.com

Abstract: ***Commercial Off-The-Shelf (COTS) components are widely used in many software industries and also in scientific computing. This paper first considers the definition of the term COTS and then tries to find how it typically been used in IT industry. A survey has been undertaken to study about the usage of COTS in IT industry. This paper deals with the details about the analysis of the data using statistical calculations and the results obtained.***

Key words: COTS, Survey, Statistical Calculations.

I. INTRODUCTION

Commercial-Off-The–Shelf(COTS) components are defined as “components which are bought from third party vendors and integrated into the system”. However, a more detailed and expanded view of COTS components should be taken. A COTS component could be as small as a routine that computes the square root of a number or as large as a credit card validation software. The important thing is that a COTS component already exists and was created by people outside the software development organization that will actually use it. A COTS component can therefore be defined as, “any software component that already exists, that was created by people outside the organization that will be using it, and that was purchased from a third party vendor”[1].

Almost without exception, every software-related endeavour will utilize a significant percentage of COTS software components. The application of COTS components in crystallographic software was evaluated by us earlier[2]. In another contribution, the case of a molecular modeling software, viz., GROMACS was taken up as a case study[3-4]. Another case study dealt with a software(SCILAB) used for mathematical computations[5]. The impact of COTS components on software quality in IT Industry was studied based on a survey. The details of this survey and its analysis are furnished in this paper.

II. INTRODUCTION TO THE SURVEY

It was planned to study the impact of COTS components on software quality. To study the awareness and usage of COTS by various experts, it was decided to conduct a survey. The survey consisted of distributing a questionnaire dealing with the different aspects on the various usages of COTS components in building software. In order to design a questionnaire, it was decided to prepare a set of questions and send to select experts for their opinion. This pilot study(described below) is necessary to get a first-hand view on this survey and analysis of the collected data.

III. PILOT SURVEY UNDERTAKEN REGARDING COTS

The questionnaire contained a list of questions for which the respondents have to answer with a "Yes" or "No". The questions covered the various aspects of the usage of COTS software, viz., (1) awareness about COTS, (2) evaluation and selection of COTS, (3) vendor selection and (4) COTS integration. The total number of questions framed were thirty. The questionnaire used for the above purpose is available elsewhere[6].

The mode chosen for conducting the above pilot survey was to send e-mails to several persons working in IT firms, The responses were received from 18 persons working in the software field. These responses helped us to understand the nature of various types of questions so that a full set of questionnaire could be designed.

IV.I DEVELOPMENT OF A NEW QUESTIONNAIRE AND FINAL SURVEY UNDERTAKEN REGARDING COTS

The items in the new questionnaire had more options to select from choices such as strongly agree, agree, unable to decide, disagree, strongly disagree (a five-point scale). In order to collect responses from software professionals a survey website known as "SurveyCrest"(www.surveycrest.com) was used. Although there are several sites to help in carrying out a survey, many of them prescribe restrictions on the number of statements in the questionnaire and the number of respondents to be contacted. In some cases, they charge the user heavily. After a careful scrutiny of these websites for service in "surveying", the website "SurveyCrest" was chosen. In this site, they do not prescribe a limit for the number of questions/statements to be asked and they also allow a maximum of 1000 respondents. Making use of this website, the final survey was conducted. The details regarding this survey, analysis of data and results obtained are furnished in this paper.

IV.II SURVEY PROCESS USING THE WEBSITE "SURVEYCREST"

The respondents for this study were selected randomly from various databases of software professionals. The total number of persons to whom e-mails were sent was 377. Out of this, 86 persons participated in the survey. Some of the details regarding the statements used in the survey, the positive/negative ranking given, the topics covered by the statements and the places of the respondents are furnished in the next section.

IV.III DETAILS ABOUT THE STATEMENTS USED IN THE QUESTIONNAIRE FOR THE SURVEY

1. The Questionnaire contained 20 statements. The number of respondents were 86.
2. As the items numbered as 1,2,3,4,9,10,11,12,13 and 18 measured the options on positive ranking, the marks 5,4,3,2, and 1 were assigned, respectively, for strongly agree, agree, unable to decide, disagree and strongly disagree.
3. On the other hand, the items numbered as 5,6,7,8,14.15.16.17,19 and 20 measured the options on negative ranking. Hence the marks given to these options were in the reverse order, viz., 1,2,3,4, and 5 for the options strongly agree, agree, unable to decide, disagree and strongly disagree, respectively.
4. The statement numbers 1 to 10 covered the topic: COTS: Awareness and vendor selection and 11 to 20 covered the topic: COTS**:** Evaluation, selection and integration, respectively.
5. The statements were further categorized based on the topic, as shown in Table 1.
6. Places of work of the respondents are given in Table 2.
7. The data also included a sample case and an unknown case, in addition to the data collected from the 86 respondents.

TABLE I. CATEGORIZATION OF THE STATEMENTS BASED ON THE TOPIC CHOSEN

Topic	Notation	Statements
Awareness about COTS	AWARE	1 to 8
Vendor selection	VEN	9 and 10
Evaluation of COTS	EVAL	11 and 12
Selection of COTS	SEL	13 to 17
Integration of COTS	INT	18 to 20

TABLE II. PLACES OF WORK OF THE RESPONDENTS

Places of the Respondents	Number
Chennai	29
Madurai	17
Bangalore	16
Tirunelveli	1
Idukki(Kerala)	1
Pune	1
Delhi	1
Amritsar	1
USA	4
UK	4
Australia	3
Singapore	2
Belgium	2
Canada	1
Malaysia	1

V. STATISTICAL CALCULATIONS CARRIED OUT USING THE SOFTWARE SPSS (SPSS.inc) and R(www.r-project.org)

Originally the data obtained from "SurveyCrest" were stored in an Excel format which is a matrix consisting of 86 rows and 20 columns. Here, 86 corresponds to the number of respondents in the survey. Marks were awarded to the answers ticked by the respondents to the twenty statements, based on the criteria discussed in the earlier section. These marks obtained by the respondents were stored in a matrix of size 86 x 20. Two more columns were created to accommodate the information about the total marks scored by the respondents (for the twenty statements) and their place of work.

V.I CLEANING OF DATA

It is imperative that the data must be scanned for abnormal or inconsistent observations before carrying out any statistical analysis. This is also called "Cleaning of data". As a first step, one has to look at the correlations of (86) scores of each item with the total marks (obtained by totaling the scores for each respondent on 20 items). This will reveal how far the selected items are consistent with the overall performance. The Pearson coefficient of correlations were calculated and it was found that the questions 11 and 12 had poor correlation coefficients with the total. Hence, it was decided to drop these items from further considerations.

As a second step, one has to look for abnormal or extreme cases that have inconsistent values compared to most of the respondents. For this the box-whisker plot was drawn(Fig. 1).

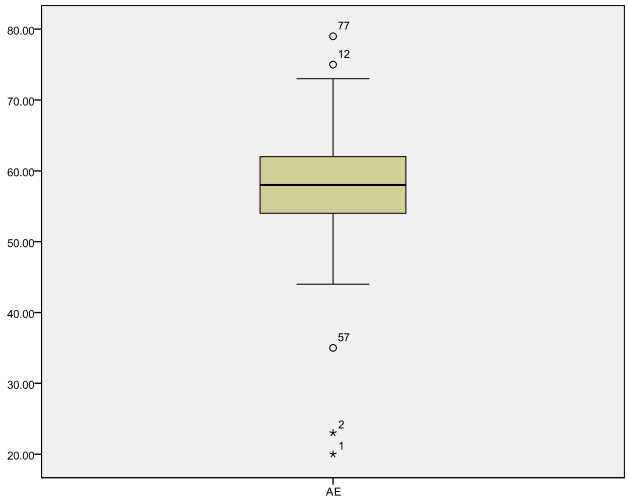

FIGURE. 1 THE BOX-WHISKER PLOT

It was observed that the cases 12 and 77 had high scores and they were not consistent with the others. As such they can be regarded as outliers; in the same way case 57 is an outlier on the lower side. The cases 1 and 2 are extreme ones. Hence the cases 1, 2, 12, 57and 77 were dropped from the data base.

Further calculations showed that the two groups of questions namely Aware and Eval are highly correlated with the total score AE.(See Table 3). Therefore, these two groups are consistent with the total.

TABLE III TABLE SHOWING THE RESULTS OF THE PEARSON CORRELATION COEFFICIENTS FOR THE TWO GROUPS OF STATEMENTS(AWARE AND EVAL)

	Correlations	Aware	Eval	AE
Aware	Pearson Correlation	1	.641	.922
	Sig. (2-tailed)		.000	.000
	N	86	86	86
Eval	Pearson Correlation	.641	1	.888
	Sig. (2-tailed)	.000		.000
	N	86	86	86
AE	Pearson Correlation	.922	.888	1
	Sig. (2-tailed)	.000	.000	
	N	86	86	86

After eliminating cases 1, 2, 12, 57 and 77 and items I11 and I12, the resultant data base was used for carrying out further analysis. The box plot for the 81 score of total AE was found to be symmetrical and all the scores were found to lie within 95% of the total area (Fig. 2). The descriptive statistics for the scores obtained by the 81 respondents is given in Table 4.

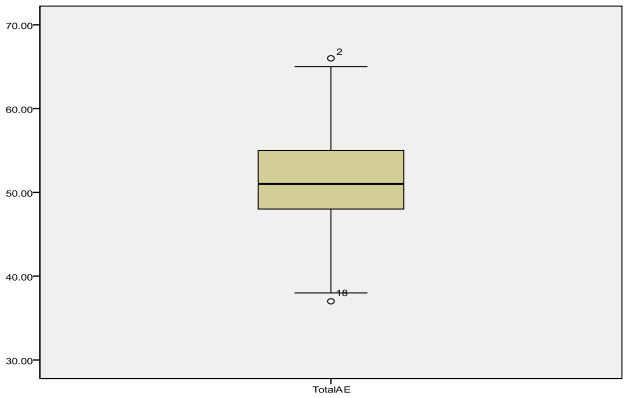

FIGURE. 2 THE BOX-WHISKER PLOT

TABLE IV THE DESCRIPTIVE STATISTICS FOR THE SCORES OBTAINED BY THE 81 RESPONDENTS

	N	Min	Max	Mean	S D
TotalAE	81	37.00	66.00	51.506	6.2372
Valid N (listwise)	81			2	3

Now it was found that all the data lie within 99% of the total data. The Gaussian curve enveloping the data is shown in Fig. 3.

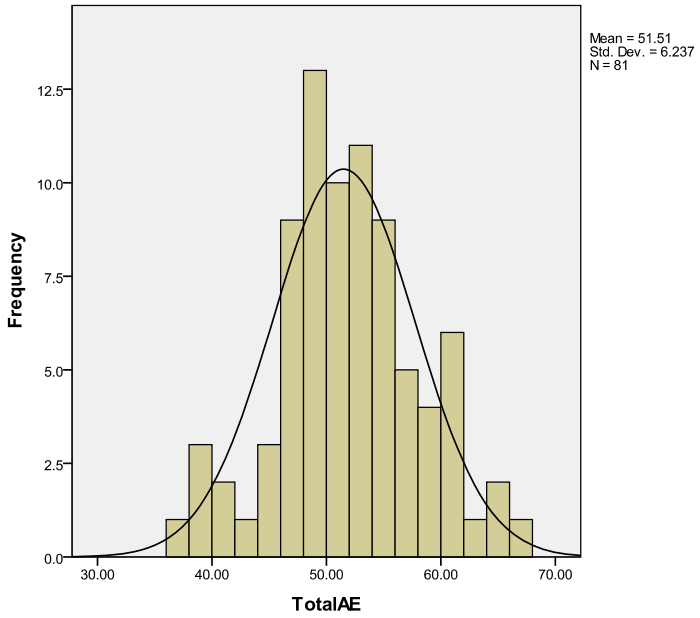

FIGURE. 3 THE GAUSSIAN CURVE ENVELOPING THE HISTOGRAM

The descriptive statistics for the total scores obtained on the two topics, viz., awareness and evaluation were calculated. The scores obtained by the respondents were divided into three groups as low, medium and high and the corresponding cross-tabulation was also obtained.

It was observed that of the 81 respondents (**3+33+8**)*100/81% cases are consistent on score in both attributes. (7+8)*100/81% have low score in AwareG but high score in EvalG. (12+8)*100/81% cases have low score in EvalG but high in AwareG.

Next, the expected counts (scores) and adjusted counts were computed and the cross tabulation made(Table 5). Next, the Chi-square test was performed. and the results are given in Table 6.

TABLE V THE CROSS TABULATION BETWEEN EXPECTED COUNTS(SCORES) AND ADJUSTED COUNTS

			Eval G			Total
			Low	Medium	High	
Aware G	Low	Count	3	7	1	11
		Exp Count	2.2	6.5	2.3	11.0
		Adjusted Residual	.7	.3	-1.0	
	Medium	Count	12	33	8	53
		Exp Count	10.5	31.4	11.1	53.0
		Adjusted Residual	.9	.8	-1.8	
	High	Count	1	8	8	17
		Exp Count	3.4	10.1	3.6	17.0
		Adjusted Residual	-1.6	-1.2	3.0	
	Total	Count	16	48	17	81
		Exp Count	16.0	48.0	17.0	81.0

TABLE VI RESULTS OF THE CHI-SQUARE TEST

	Value	Df	Asymp Sig(2 Sided)
Pearson Chi-Square	9.862[a]	4	.043
Likelihood Ratio	9.334	4	.053
Linear-by-Linear Association	7.012	1	.008
N of Valid Cases	81		

4 cells (44.4%) have expected count less than 5. The minimum expected count is 2.17.

Then, the descriptive statistics was obtained using the scores of the respondents in the four separate sections of the questionnaire, viz., AWARE, VEN, SEL and INT groups(Refer Table 1). This is given in Table 7. Then, regrouping based on the scores obtained was carried out. resulting in the information given in the Tables 8(a - d).

TABLE VII THE DESCRIPTIVE STATISTICS FOR THE TOTAL SCORES OBTAINED IN THE FOUR TOPICS, VIZ., AWARENESS, VENDOR SELECTION, SELECTION OF COTS AND INTEGRATION OF COTS

DESCRIPTIVE STATISTICS

	N	Min	Max	Mean	S D
AWARE	81	15	32	23.99	3.558
VEN	81	3	10	6.90	1.758
SEL	81	6	19	12.00	2.660
INT	81	4	13	8.62	1.875
Valid N (listwise)	81				

TABLE VIII (A - D). THE FREQUENCY TABLE FOR THE FOUR ITEMS ON REGROUPING, BASED ON THE SCORES

A) AWARE

		Freq	%	V%	C%
Valid	High	20	24.7	24.7	24.7
	Low	22	27.2	27.2	51.9
	Medium	39	48.1	48.1	100
	Total	81	100	100	

B) VEL

		Freq	%	V%	C%
Valid	High	49	60.5	60.5	60.5
	Low	32	39.5	39.5	100
	Total	81	100	100	

C) SEL

		Freq	%	V%	C%
Valid	High	24	29.6	29.6	29.6
	Low	33	40.7	40.7	70.4
	Medium	24	29.6	29.6	100
	Total	81	100	100	

D) INT

		Freq	%	V%	C%
Valid	High	20	24.7	24.7	24.7
	Low	22	27.2	27.2	51.9
	Medium	24	48.1	48.1	100
	Total	81	100	100	

In the next step, the total marks obtained by the respondents for all the statements put together was divided into four categories (Low: scores up to 48; Medium: scores between 49 and 51; High: scores between 52 to 55; Supreme: scores from 56 to the highest) and the frequency calculated.

Next, it was felt useful to carry out the statistics based on the regions(their work places) of the respondents. The four regions considered are as given below:

Region 1: Madurai (including Tirunelveli)

Region 2: Chennai

Region 3: Bangalore

Region 4: All the other cities (including the respondents from abroad)

The cross tabulations and the results of the chi-square test are furnished in Tables 9 and 10, respectively. It was found that the total mark groups and region_ new are NOT independent. Thus we conclude that there is a regional disparity in the total response score.

TABLES IX THE CROSS TABULATIONS BASED ON THE FOUR REGIONS OF THE RESPONDENTS

Region_New * Total Group Cross tabulation

Count	Total Group				Total
	High	Low	Medium	Supreme	
Region_ New	4	6	2	4	16
	10	17	9	7	43
	5	1	4	8	18
	1	0	3	0	4
Total	20	24	18	19	81

TABLE X THE RESULTS OF THE CHI-SQUARE TEST BASED ON FOUR REGIONS OF THE RESPONDENTS

	Value	df	Asymp sig(2 sided)
Pearson Chi square	17.576a	9	.040
Likelihood Ratio	18.929	9	.026
N of valid cases	81		

The above table gives the 'expected' counts and adjusted residuals

Z = (observed count – expected count)/Standard error

The following interesting observations were made from the above table:

1. Nearly for all the cells (region, Totalgroup), the observed count does not differ from expected count.

For eg., (region_new = 2, totalgroup = High),the expected count is 10.6, while the observed count is 10. But for the cell (Region_new=4, Totalgroup=Medium), the expected count is 0.9, i.e., nearly 1, but the observed count is 3. To test whether the difference is significant or not, the following criterion is used:

If -1.96<Z <1.96, the difference between observed count and expected count is NOT significant; otherwise significant at 95% confidence.

Similarly one may use -2.56 and 2.56 for 99% confidence.

2. It was already noted that the region_new and the total groups are not independent. Though, almost all cells do not have significant difference between observed and expected counts, it is to be pointed out that the cells at (3, L), (3, S) and (4, M) have significant differences between observed and expected counts.

That is, respondents from Bangalore get Low marks which is lower than the expected counts; and also get supreme (very high) marks which is higher than what is expected for this cell.

3. For Madurai and Chennai regions, the responses are homogeneous and are consistent.

VI. COMPUTATIONS OF CLASSIFICATION AND REGRESSION TREE (CART)

Classification and Regression Trees (CART), a recursive partitioning method, builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). CART uses historical (past or prior) data to construct the so-called decision trees. Decision trees are then used to classify new data.

CART methodology was developed in the 1980s by Breiman et al.[7]. Decision trees are represented by a set of questions which splits the learning sample into smaller and smaller parts.

CART asks only yes/no questions. A possible question could be:

- Is age of candidate greater than 50? or
- Is sex of the participant male?

CART algorithm will search for all possible variables and all possible values in order to find the best split - the question that splits the data into two parts with maximum homogeneity.

We all make decisions, intuitively, by considering various possibilities in the available actions. The past knowledge helps us to fix some rules which is used to take decision. This may lead to some errors (also called misclassification errors). CART uses systematic procedure and statistical measures to minimize these errors. The data on hand was analyzed to find out the possible decisions as to high and low scorers among the respondents.

The computations of Classification and Regression Tree (CART) were carried out with the following variables:

ScoreGroup (Low scorer=0, High scorer=1)

Predictor variables; aw1_8, ven9_10, sel13_17 and Int18_20.

The decision tree obtained is shown in Fig.4

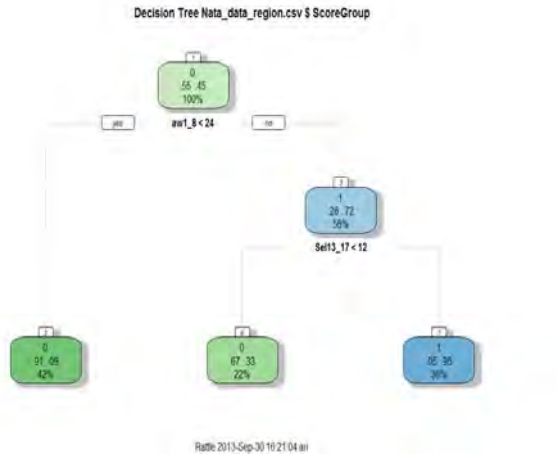

FIGURE.4 THE DECISION TREE OBTAINED FOR THE FOUR REGIONS AND THE TOTAL SCORES OBTAINED BY THE RESPONDENTS*****

***** aw_8 : AWARE group, Sel13 : SEL group

EXPLANATION FOR THE CART DIAGRAM

The top most entry represents the root. Nearly 55 % and 45 % of candidates scored less than and greater than 23.5, respectively. Of the later group, 28 % and 72 % candidates have scored less than and greater than 11.5, respectively. Thus, of those having aw_8 score above 24 and score in Sel13 greater than 12, 95 % of participants are high scorers.

Hence the association rules for finding high scorers can be stated as follows:

The association rules are

Tree as rules:

Rule number: 7 [ScoreGroup=1 cover=20 (36%) prob=0.95]

aw1_8>=23.5

Sel13_17>=11.5

Rule number: 6 [ScoreGroup=0 cover=12 (22%) prob=0.33]

aw1_8>=23.5

Sel13_17< 11.5

Rule number: 2 [ScoreGroup=0 cover=23 (42%) prob=0.09]

aw1_8< 23.5

VII. GENERAL CONCLUSIONS

From the computations carried out and presented above, several conclusions have been obtained as given below;

1. The SW professionals of South India seem to be fairly knowledgeable about COTS and their applications.
2. There appears to be regional disparities in the awareness and also usage of COTS SW among the professionals at different places, viz., Madurai, Chennai, Bangalore and other places.
3. The results of this analysis only complement the results obtained in the earlier Pilot survey.

VIII. ACKNOWLEDGEMENTS

One of the authors (NG) thanks the Management and the Principal of the K.S.R. College of Arts and Science, for the encouragements. Authors are thankful to Prof. G. Arivarignan, CSIR Emeritus Scientist, Madurai Kamaraj University for his assistance in carrying out the statistical calculations using R((www.**r**-project.org).

REFERENCES

[1] L.Brownsword, T.Oberndorf, C.Sledge. Developing New Processes for COTS Based Systems, IEEE Software, 17(4), 48 –55(2000).
[2] N. Gnanasankaran, S. Natarajan, K. Alagarsamy and K. Iyakutti, Application of COTS Components: A Software Package for Crystallography, British Journal of Mathematics & Computer Science, 3(2):99-107,(2013).
[3] N. Gnanasankaran, S.Natarajan, K. Alagarsamy and K. Iyakutti, A case study of the application of COTS components in a molecular dynamics software, Paper presented at the Fourth International Conference on Electronics Computer Technology (ICECT 2012), Kanyakumari, India, 6-8 April, 2012.
[4] N. Gnanasankaran, S.Natarajan, K. Alagarsamy and K. Iyakutti, A case study of the application of COTS components in a molecular dynamics software, Lecture Notes on Software Engineering, 1(2), 141-143(2013).
[5] N. Gnanasankaran, S. Natarajan, K. Alagarsamy and K. Iyakutti, Commercial Off-The-Shelf(COTS) Components in Software Engineering: The Software Package SCILAB, Int. J. Computer Technology & Applications, 4(1), 68-71(2013).
[6] N. Gnanasankaran, Ph.D. Thesis, Madurai Kamaraj University, 2013.
[7] Breiman L, Friedman J, Olshen R, and Stone C, Classification and Regression Trees: An Introduction, Chapman & Hall/CRC, 1984.

- www.ibm.com/software/analytics/spss
- www.r-project.org
- www.surveycrest.com