International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 03, March-2015

Survey Based Analysis of Effect of Code Clones on Software Quality

Shahid Ahmad Wani
Research Scholar,
Dept. of MMICT&BM
Maharishi Markandeshwar University
Mullana Amballa, Haryana, India

Shilpa Dang
Assistant Professor,
Dept. of MMICT&BM
Maharishi Markandeshwar University
Mullana Amballa, Haryana, India

***Abstract*—Code clones are similar code portions. Cloning is a process of duplicating code segments by copy–paste activities that is a common activity in software development. It is believed that the presence of code clone is one of the factors that have a great impact on software quality attributes. In literature many techniques have been proposed to detect and eliminate code clones on this basis. Various research efforts are being performed to reduce somber problems caused by code clones. This paper presents the study of the effect of code clones on software quality. In this paper an industrial study is presented to understand impact of code clones on a software system from software developer's point of view. This study involves a questionnaire survey and collects enough data about the reasons behind the cloning activity and the impact of code clones on a software system. The results of the study show that clones have a harmful effect on the system. This study also suggests that maintenance is the mostly effected software quality attribute.**

Keywords—*Code Clones, Abstract Syntax Tree (AST), Program Dependence Graph (PDG).*

I. INTRODUCTION

Code clones are similar or identical code portions in software programs. Generally, code clones are introduced by copy–paste programming activity in software development [1]. Such programming practices are common, very easy and can reduce programming effort as well as time as they reuse on hand code rather than rewriting related code from scratch. But this can cause somber problems in long run to a software system. It is said that code clones have a negative impact on software development and quality. For example clones may increase bug occurrence, if an expression of duplicate code is changed for fixing bugs or adding new features, its correspondents must be changed simultaneously, if the corresponding duplicate code is not changed, new bugs are introduced to them. It is agreed that clones exist in software systems and they must be detected to maintain, manage or remove them. It is believed that code clones have a negative impact on software quality attributes. One of the attributes is maintenance. Software maintenance is the most expensive phase in the entire software development process. It is reported that a lot of money is spent in the maintenance of existing systems of a software development company [2].

The renowned researchers have put efforts for improving and resolving the problems caused by code clones. Many techniques have been proposed to detect, manage and remove code clones [3]. The various clone detection techniques Abstract Syntax Tree (AST) [4], Program Dependence Graph (PDG) [5], code metrics [6] and program tokens [7] [8] provide an automated assistance to identity code clones in source code. Visualization and query-based techniques [9] [10] have been proposed to manage and inspect detected code clones in large software systems. It has also been proposed that code clones can also be removed through refactoring [11]. Many clone detection tools have been developed which use these techniques for automatic detection of clones. All these tools have possible advantages and disadvantages and more hybrid approaches are needed to overcome the limitations of these tools [12].

In spite of the great success of these clone detection and refactoring techniques, little work has been done in understanding why and in which situations developers introduce clones into a software system. Clones have a huge impact on software quality attributes. But whether the clones in a software system are harmful or not is still an open question. In this study, it is tried to find out the answers to these questions. This work involves a survey through a questionnaire presented to the people working in software industry. The experience of professional people working in software industry with clones how they define a clone and impact of clones over software quality attributes is introduced in this paper. There are some studies [1] [11] [13] [14] on finding intentions and reasons behind code cloning practices. However, these studies are based on the personal experience of researchers with no support from industrial studies. These study focus on the introduction of clones into the system, with little study concerning impact of clones on software quality attributes. There are still many research questions that remain unanswered. This research work focuses on the following questions related to code clones:

1. What is a code clone?
2. Why and how often developers perform cloning activities?
3. Whether clones are harmful or not to a software system?
4. Which software quality attribute is mostly effected due to clones?

5. Effect of clones on performance, complexity, scalability and maintenance of a software system.

To find the answers of the above questions a questionnaire is given to the industry professionals. Based on their answers in the questionnaire better results are found. In this study enough data is collected from industrial perspective, to answer the above discussed questions regarding code cloning practices in software development.

II. CATEGORICAL VARIABLES FREQUENCY TABLE

In order to understand the code cloning practice in software industry, a survey is conducted through a questionnaire. The questionnaire is sent to software professionals working in industry to understand their perception about code cloning. The survey consisted of 34 pre-defined multiple choice questions, which were designed to understand the participant's view about clone definition, why they copy-paste code, impact of clones on software quality and which of the software quality attribute is mostly effected by code clones. Table 1 shows a summary of survey questions. The responses of the survey are analyzed using SPSS-19. In total 40 engineers participated in the survey. The majority of the participants have experience of 6-10 years. 70% of them were developers, 20% of them were test engineers, 7.5% of them were maintenance engineers and 2.5% of them were design engineers. The participants had different experience in software industry with C, C++, C# and Java programming languages. The response frequency and percentage for each categorical variable is shown in table 2.

TABLE 1: SUMMARY OF SURVEY QUESTIONS

S. No.	**Question**
1.	How do you define a "code clone"?
2.	How often do you perform copy–paste or cloning activities?
3.	What according to you are the reasons for cloning?
4.	How do you agree that clones have a negative impact on software quality?
5.	Which of the following software attribute is mostly effected by code clones?
6.	How do you agree that clones reduce the performance of the software system?
7.	How do you agree that clones make complex the maintenance of a software system?
8.	How often do clones have negative impact on performance of software?
9.	What percentage of negative impact do clones have on the performance of software system?
10.	How do you agree that clones increase the size of a software program?
11.	How do you agree difficult to change code is reason for code cloning?
12.	How do you agree risk avoidance is reason for code cloning?
13.	How do you agree that clones increase the size of software program?
14.	How do you agree that clones effect maintenance effort and cost?
15.	What percentage of negative impact do clones have on the maintenance of software system?

TABLE 2: FREQUENCIES AND PERCENTAGE OF CATEGORICAL VARIABLES

Participant Experience									
1 years– 5 years		6 years – 10 years		11 years – 15 years		16 years – 20 years		More than 20 years	
Freq.	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*
18	45.0%	21	52.5%	1	2.5%	0	.0%	0	.0%
Work Area									
Development		Test		Design		Maintenance		Other	
Freq.	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*
28	70%	8	20%	3	7.5%	1	2.5%	0	.0%
Language Familiar									
C		C++		C#		Java		Other	
Freq.	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*
0	.0%	5	12.5%	1	2.5%	33	82.5%	1	2.5%

III. SURVEY BASED ANALYSIS OF CODE CLONES

Code cloning is an active research area since 1980s. In literature large amount of research on software code clone has been carried out that mainly focus automatic techniques to identify, detect, manage and eliminate code clones. In spite of the promising results of these clone detection techniques, code clone still remains a big challenge to quality of a software system. Most of the studies available have examined open-source systems and focused on code analysis and did not consider reasons and intentions for the introduction of code clones. Furthermore, these studies do not analyze information from developer and industrial perspectives. In this concern this study conducts a survey based analysis to understand views of software developer regarding the effect of code clones on software quality. This research work is based on five questions as discussed in section 1. The results of the industry professionals are analyzed based on these questions.

A. What is a Code Clone?

Existing literature says that there is still vagueness in the definition of clones. This study finds the opinion of software developers about the definition of clones. It is found that developers have different views about the definition of code clones. 40% of participant define clone as duplicate code, 10% define clone as copy of original code, 37.5% define clone as copy-pasted code and 12.5% define it as similar code. Table 3 shows the frequencies and percentages for definition of code clones. It can be seen that majority of the developers called clone as duplicate code and copy-paste code. It is found that developers have different opinions regarding the definition of code clones and there is still an ambiguity in the definition. Fig. 1 shows the response for the definition of code clones.

B. Why and how often developers performs cloning activities?

The other objective of survey is to understand the reasons for cloning practices in software industry. The responses of developers suggest that there are different reasons for code cloning practice. It is found that risk avoidance is mostly the motivation of introducing clones. The developers try to avoid the risk of making changes to the existing system as they are afraid that this may crack the system. The other reason for introduction of code clones is time limitation, the pressure of submitting a project in time force a developer to reuse the exiting code that ultimately leads to the clone. The other three reasons of code cloning are difficult to change existing code, unaware of harmfulness of clones and skill limitation. By the industry developers finds it difficult to change the existing code so they keep on reusing the code, they copy-paste code in the initial phase of project development as they are unaware of its harmful consequences. Due to less knowledge and limited skills of developers people try to find the available code for problem that leads to the introduction of code clones in a system. Table 4 summarizes response percentage for the reasons of code clones. 47.5% of developer's claim that risk avoidance as the reason for cloning, 27.5% strongly agree and 60% agree for time limitation as a reason, 55% state unawareness of harmfulness of clones as the reason of cloning. The satisfaction level of survey participants for the reasons of code clones is shown in Fig. 2. Out of all five reasons risk avoidance and skill limitation are the two most important reasons.

This study also finds that how often does software developers copy-paste. In table 5 frequencies of response of how often developers perform cloning activity is shown. The responses of participants suggest that cloning is a frequent activity from developer's point of view. 17.5% of developer's state that they often copy-paste code, 60% state that they sometimes copy-paste code, 20% of developer's say that they rarely do so and none of the participants said that they never copy-paste code. Fig. 3 depicts percentage of how often developers perform cloning.

TABLE 3: DEFINITION OF CODE CLONES

Definition	**Freq.**	**%age**
Duplicate Code	16	40%
Copy of original code	4	10%
Copy-pasted code	15	37.5%
Similar code	5	12.5%
Other	0	0.0%

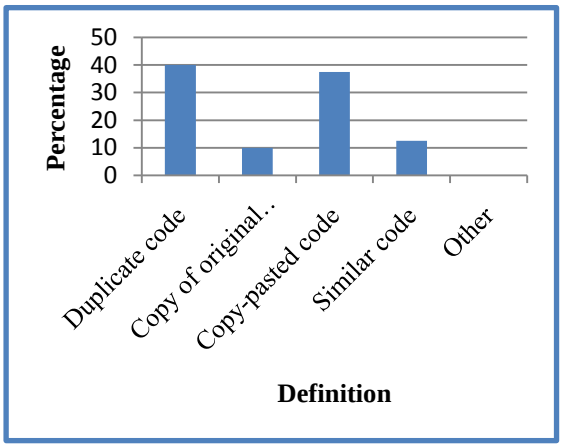


Fig. 1: Definition of code clones.

TABLE 4: REASON OF CODE CLONING

Reason	Strongly Agree		Agree		Can't Say		Disagree		Strongly Disagree	
	Freq.	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*
Difficult to change existing code	4	10.0%	18	45.0%	6	15.0%	11	27.5%	1	2.5%
Time Limitation	11	27.5%	24	60.0%	1	2.5%	4	10.0%	0	0.0%
Risk Avoidance	19	47.5%	19	47.5%	0	0.0%	2	5.0%	0	0.0%
Unaware of harmfulness of clones	0	0.0%	22	55.0%	4	10.0%	10	25.0%	4	10.0%
Skill Limitation	15	37.5%	15	37.5%	8	20.0%	0	0.0%	2	5.0%

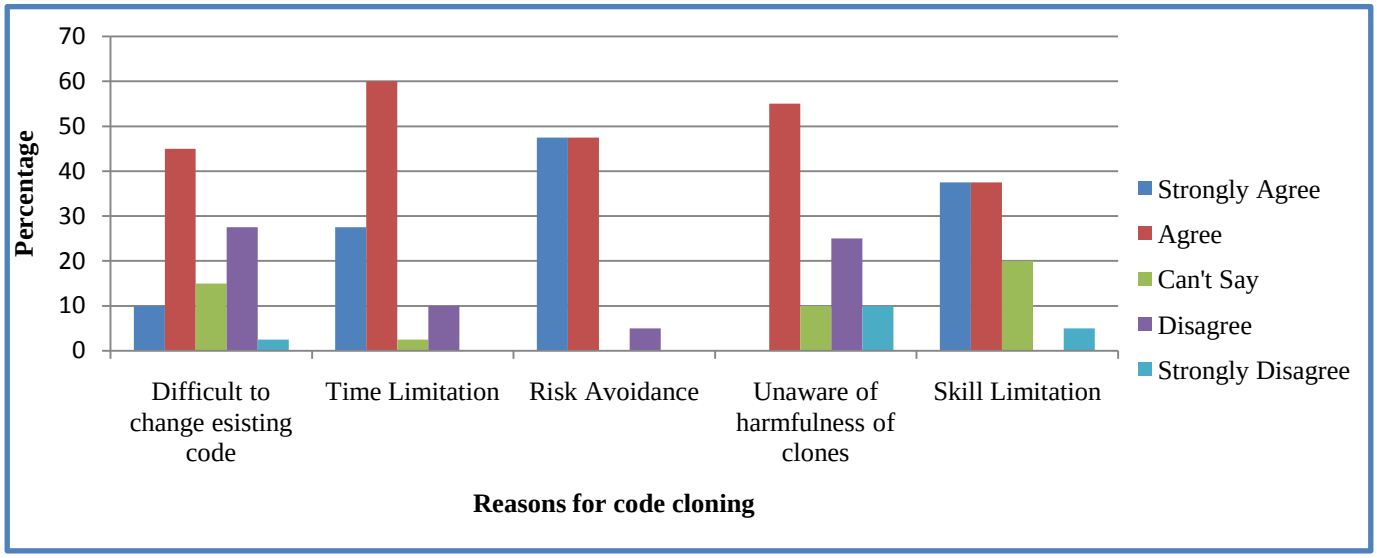


Fig. 2: Reasons of code cloning.

TABLE 5: HOW OFTEN DEVELOPERS PERFORM CLONING

How often	**Freq.**	**%age**
Often	7	17.5%
Sometimes	24	60%
Rarely	8	20%
Never	0	0.0%
Depends on Situation	1	2.5%

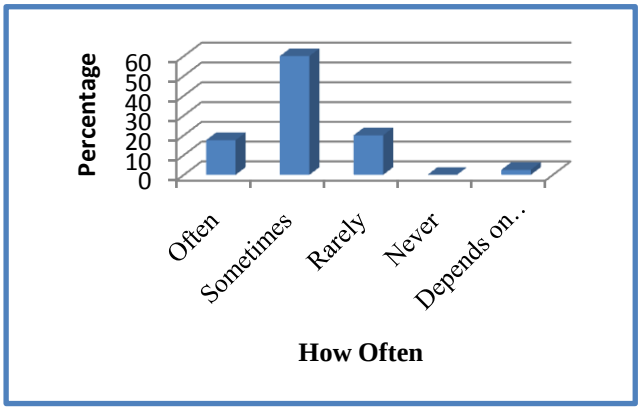


Fig. 3: How often developers perform cloning.

C. *Whether Clones are Harmful or Not?*

In order to answer the question whether code clones are harmful are not, certain questions were provided in the questionnaire concerned with the impact of clones on the software. This study finds the level of satisfaction of the participants for the options provided with the questions and it is checked by providing the scale from strongly agree to strongly disagree. Table 6 shows the responses of whether clones are harmful or not. It can be noticed that 27.5% of software developer's strongly agree that clones have negative impact on software quality, 67.5% of the participant's state that they agree for the statement while as 5% participants claim that they can't say anything about the statement but none of the participant state that they disagree with the statement. The response of the participants shows that clones are harmful for a system. In view of finding how clones are harmful to a software system the questionnaire contained related questions such as mostly effected software quality attribute, impact of clones on performance of a software system, is scalability effected by clones, does clones increase complexity of software system and impact of code clones on the maintenance of the software system.

In this study it is also found that how often clones have negative effect on a software system. Table 6 also shows how often clones have harmful effect on software system. 50% responses suggest clones sometimes negatively affect a software system, 32.5% participants claim for frequent negative impact of clones, 10% suggest seldom while as 7.5% state that it depends on the number of clones present in a system. It can be noted that clones often have a negative impact on a software system. One common statement among the developers is that they all are agree that clones are harmful for the software system. But, it depends on the type of system, environment and situations. Fig. 4 shows whether clones are harmful or not and how often.

TABLE 6: WHETHER CLONES ARE HARMFUL OR NOT AND HOW OFTEN

Clones have harmful effect on system									
Strongly Agree		Agree		Can't Say		Disagree		Strongly Disagree	
Freq.	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*
11	27.5%	27	67.5%	2	5.0%	0	.0%	0	.0%
How often clones have harmful effect on system?									
Often		Sometimes		Seldom		Never		Depends on number of clones	
Freq.	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*
13	32.5%	20	50.0%	4	10.0%	0	0.0%	3	7.5%

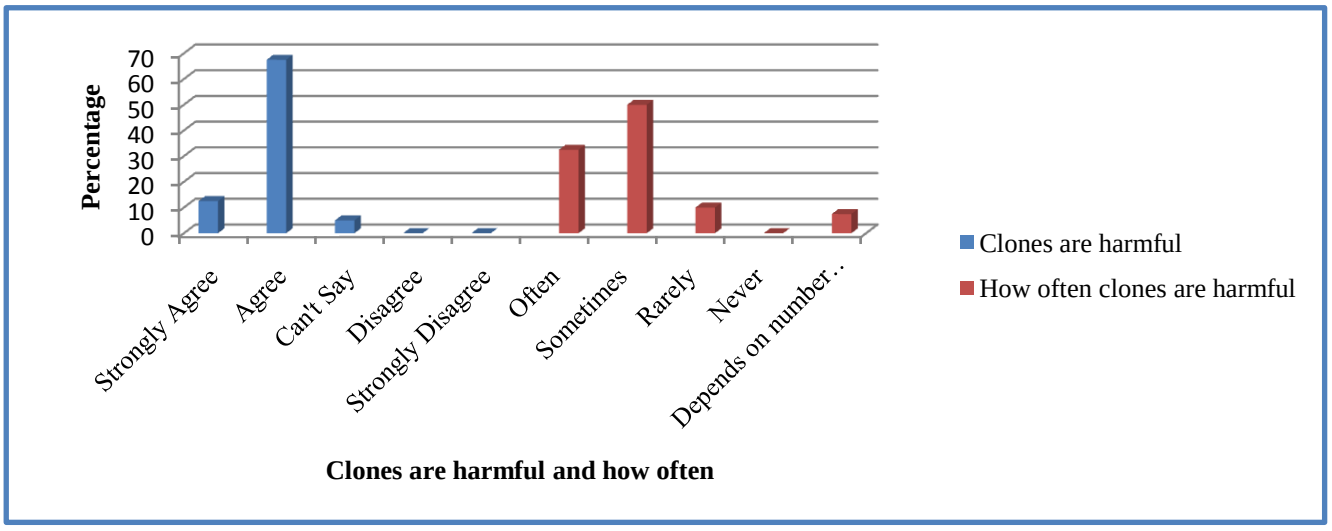


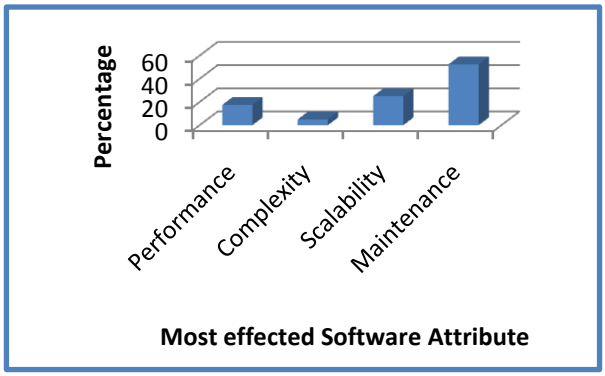
Fig. 4: Whether clones are harmful or not and how often.

D. Which of the software quality attribute is mostly effected by clones?

Although the researchers have found that clones have a negative impact on maintenance, there are other software attributes that are effected by clones. In order to find out which of the software quality attribute is mostly effected from developer's point of view a few related question were provided in the questionnaire. Based on existing and previous study of code clones four software quality attributes are chosen that are effected by clones. . The four chosen software quality attributes that are most significantly related with cloning activity are performance, scalability, complexity and maintenance. Table 7 shows the frequency of mostly effected software attribute. It can be seen that the professionals working in industry state that maintenance of software system is mostly affect by software system and claimed by 52.5% participants. 17.5% claim that performance is affected mostly, 5% state complexity while as 25% participants claim that scalability is mostly effected by code clones. Fig. 5 shows percentage of mostly effected software attribute. Out of the four software quality attributes it is found that maintenance and scalability are the two most significant attributes effected by code clones. Thus it is suggested that necessary measures should be taken at the inception of a project to prevent from the serious problems caused by code clones.

TABLE 7: MOSTLY EFFECTED SOFTWARE ATTRIBUTE

Software Attribute	Freq.	%age
Performance	7	17.5%
Complexity	2	5%
Scalability	10	25%
Maintenance	21	52.5%



Mostly Fig. 5: Mostly effected software attribute.

E. Effect of clones on performance, complexity, scalability and maintenance of a software system

This section presents the impact of code clones on performance, complexity, scalability and maintenance software quality attributes. The impact ratio of clones on quality attributes is also presented in this section. Table 8 shows response frequency and percentage of effect of clones on performance, complexity, scalability and maintenance. Although maintenance is effected by code clones mostly but the other software attributes are also affected. 72.5% participant agree with the statement that clones reduce the performance of the software system, 12.5% participant strongly agree while as only 5% of people disagree with the statement. In response of the statement "clones increase complexity" it is found that 50% of people disagree with the statement and only 32% agree. Thus it can be concluded that clones also have negative impact on complexity. It is also found that clones have a great impact on the scalability of system. In response of the statement "clones increase size" of software program, it is identified that 55% of people strongly agree and 40% are agree and just 5% disagree with the statement. The response of the statement "clones have negative impact on maintenance" suggest that clones have huge harmful affect on maintenance as 37.5% strongly agree and 62.5 % agree with the statement. Fig. 6 shows effect of clones on performance, complexity, scalability and maintenance.

Table 8 shows that majority of the participants agree with the statements. This implies that clones have a negative impact on the quality attributes. Table 9 shows how often clones have a negative impact on software attributes. The responses suggest that the complexity is the only attribute that is seldom affected by code clones as stated by 52.5% of participants. 37.5% of participants state often and 50% say that performance is sometimes effected by code clones. Only 2.5% state that performance is never effected by clones. For scalability 50% state often and 47.5% state sometimes as effected by clones but none of the participants says never. 65% participants' claim that maintenance is often effected and 35% claim that it is effected sometimes. The response percentage of how often clones effect quality attributes is shown in Fig. 7.

TABLE 8: EFFECT OF CLONES ON PERFORMANCE, COMPLEXITY, SCALABILITY AND MAINTENANCE.

Statement	Strongly Agree		Agree		Can't Say		Disagree		Strongly Disagree	
	Freq.	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*
Clones reduce performance of system	5	12.5%	29	72.5%	3	7.5%	2	5.0%	1	2.5%
Clones increase complexity of system	4	10.0%	13	32.5%	2	5.0%	20	50.0%	1	2.5%
Clones increase size of software program	22	55.0%	16	40.0%	0	0.0%	2	5.0%	0	0.0%
Clones have negative impact on maintenance	15	37.5%	25	62.5%	0	0.0%	0	0.0%	0	0.0%

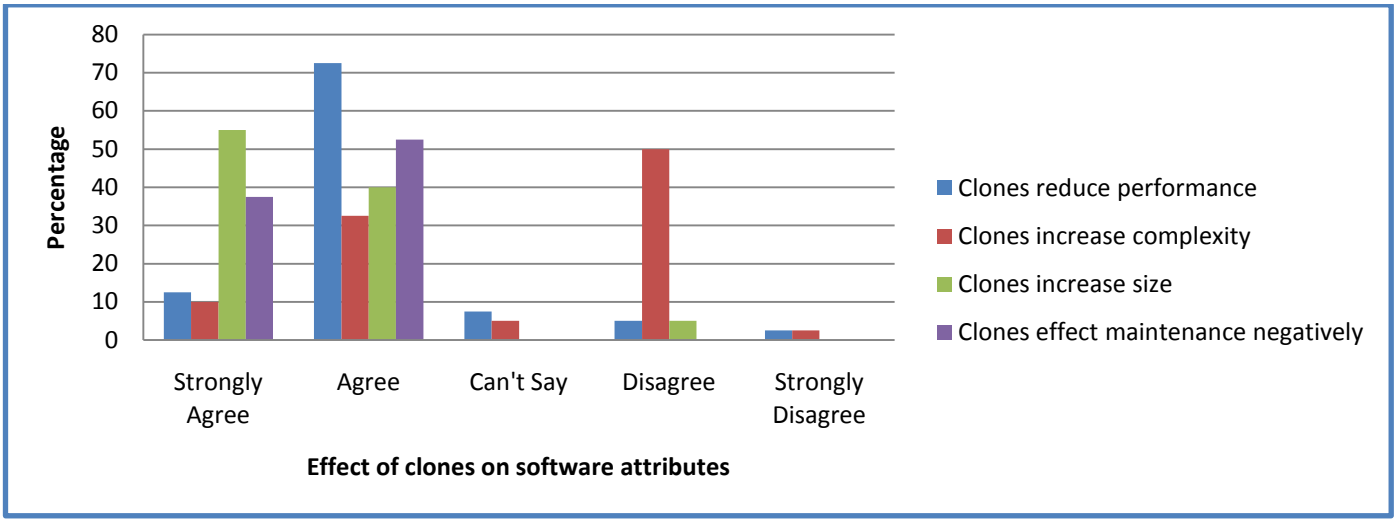


Fig. 6: Effect of clones on Performance, complexity, scalability and maintenance.

TABLE 9: HOW OFTEN CLONES HAVE NEGATIV IMPACT ON SOFTWARE ATTRIBUTES.

Software Attribute	Often		Sometimes		Seldom		Never		Can't Say	
	Freq.	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*
Performance	15	37.5%	20	50.0%	4	10.0%	1	2.5%	0	0.0%
Complexity	4	10.0%	14	35.0%	21	52.5%	0	0.0%	1	2.5%
Scalability	20	50.0%	19	47.5%	1	2.5%	0	0.0%	0	0.0%
Maintenance	26	65.0%	14	35.0%	0	0.0%	0	0.0%	0	0.0%

Finally how much amount of percentage each software attribute is effected due to clones is evaluated. Table 10 shows how much impact does clones have on software quality attributes. It is found that clones drastically increase the lines of code as the responses suggest that 41-60% increase is found in the size of program due to clones. The responses claim that 21-40% negative effect is found on performance and maintenance and 1-20% on complexity of system. A greater part of the participants claim that clones have 21-40% of harmful impact on performance,

scalability and maintenance attributes. 5% of the participants claim that clones have 61-80% impact on scalability and 2.5% state for the same percentage of effect on maintenance attribute of software system. The responses of the survey suggests that complexity is the only attribute that is least effected by code clones. Fig. 8 shows how much percentage of software attributes are effected by clones.

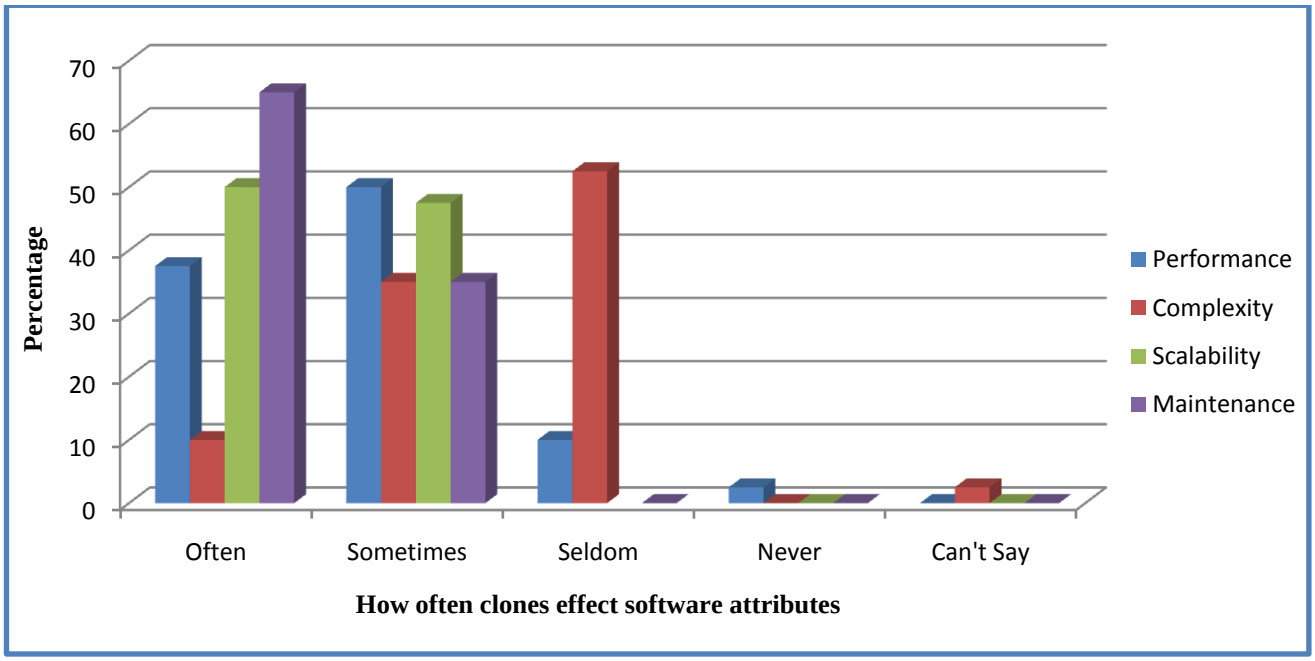


Fig. 7: How often clones effect quality attributes.

TABLE 10: HOW MUCH IMPACT DOES CLONES HAVE ON SOFTWARE QUALITY ATTRIBUTES

Software Attribute	1-20%		21-40%		41-60%		61-80%		81-100%	
	Freq.	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*	*Freq.*	*%age*
Performance	9	22.5%	17	42.5%	14	35.0%	0	0.0%	0	0.0%
Complexity	21	52.5%	15	37.5%	4	10.0%	0	0.0%	0	0.0%
Scalability	1	2.5%	15	37.5%	22	55.0%	2	5.0%	0	0.0%
Maintenance	1	2.5%	30	75.0%	8	20.0%	1	2.5%	0	0.0%

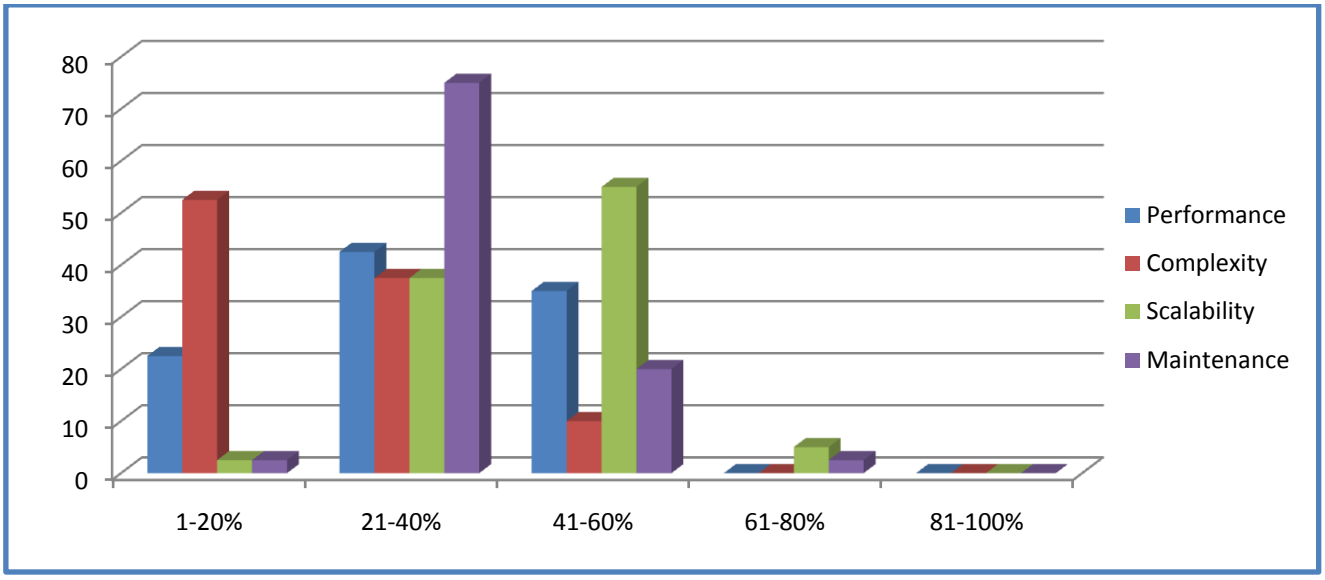


Fig. 8: How much percentage of software attributes are effected by clones.

Since the results found in this study are the questionnaire based responses of the professionals, therefore it may be a threat to the validity of the results. Also some participants might have hide real thoughts due to personal and organizational reasons, participants may have misunderstood certain questions. The analysis of study may also be biased because of the incomplete knowledge about the background of the participants.

IV. CONCLUSION

This study presented a survey based analysis of effect of code clones on software quality. The survey was conducted in software industry using a questionnaire to find out the answers of some significant research questions. The questionnaire was developed based on the previous study and the existing literature of code clones. The response of the survey is analyzed using the SPSS-19 software package. A good amount of empirical data was collected and based on the analysis of data results were presented. It is found that definition of clones is ambiguous. Risk avoidance was found to be the most significant reason of code cloning. This work suggests that clones have a harmful effect on software quality attribute.

REFERENCES

[1] C. K. Roy and J. R. Cordy, "A survey on software clone detection research", Technical Report No. 2007-541, School of Computing, Queen's University, Canada, 2007.
[2] Yoshiki Higo, Yasushi Ueda, Toshihro Kamiya, Shinji Kusumoto, Katsuro Inoue, "On software maintenance process improvement based on code clone cnalysis".
[3] S. Bellon, R. Koschke, G. Antoniol, J. Krinke, and E. Merlo, "Comparison and evaluation of clone detection tools," *IEEE Transactions on Software Engineering*, vol. 32, no. 10, pp. 804–818, 2007.
[4] I. Baxter, A. Yahin, L. Moura, and M. S. Anna, "Clone detection using abstract syntax trees", pp. 368-377, ICSM 1998.
[5] J. Krinke. "Identifying similar code with program dependence graphs", pp. 301-309, WCRE 2001.
[6] J. Mayrand, C. Leblanc, E. Merlo, "Experiment on the automatic detection of function clones in a software system using metrics", pp. 244-253, ICSM 1996.
[7] T. Kamiya, S. Kusumoto, and K. Inoue, "CCFinder: A multilinguistic token-based code clone detection system for large scale source code", IEEE Trans. on Soft. Eng., Vol. 28(7), pp. 654- 670, 2002.
[8] B. Baker, "On finding duplication and near-duplication in large software systems", pp. 86-95, WCRE 1995.
[9] M. Rieger, S. Ducasse and M. Lanza, "Insights into system wide code duplication", pp. 100-109, WCRE 2004.
[10] Y. Zhang, H. A. Basit, S. Jarzabek, D. Anh, and M. Low, "Query-based filtering and graphical view generation for clone analysis", pp. 376-385, ICSM 2008. [11] F. V. Rysselberghe and S. Demeyer, "Evaluating clone detection techniques from a refactoring perspective", pp. 336-339, ASE 2004.
[11] F. V. Rysselberghe and S. Demeyer, "Evaluating clone detection techniques from a refactoring perspective", pp. 336-339, ASE 2004.
[12] Shahid Ahmad Wani and Shilpa Dang, "A comparative study of clone detection tools", IJARCSMS, Vol. 3(1), pp. 37-41, 2015.
[13] C. Kapser and M. W. Godfrey. "Clones considered harmful" considered harmful", pp. 19-28, WCRE 2006.
[14] M. Kim, L. Bergman, T. Lau, and D. Notkin, "An ethnographic study of copy and paste programming practices in OOPL", pp. 83-92, ISESE 2004.