CHAPTER 3

INVESTIGATION OF MODULE SIZE DISTRIBUTION
IMPACT ON DEFECT DENSITY

Dependence on the software products is increasing day by day since a few decades. This trend forces the software developers to pay more attention to the software product quality. Software developers are using some quality attributes to measure the readiness of the product before its formal release. Defect density is the external attribute that identifies the possible improvement in the product before its release. Defect density is a parameter used by the developer to measure the reliability of the product before its formal release. In most of the previous studies, this parameter is positively related to the size of the product. Once the product reaches in the modification phase to rectify the defect discovered after testing, the size of the product will vary due to adding or deleting code in the product. Hence, the establishment of a relationship between software size and defect density is the need of research. In this chapter, a model has been proposed that gives a fair relationship between varying sizes of the module with defect density. The model has been used to identify the optimal size of the module with minimum defect density. Well known three available data sets have been used to analyze increasing as well as the decreasing trend in defect density with respect to the module size. The analysis result inferred that the effective module size distribution leads to significance optimization in defect density.

3.1 INTRODUCTION

In the recent era of software development, the development of secure and bug-free software is a major challenge for software developers. Software developers are trying to minimize the number of defects in the software product although it is next to impossible task. The software developers are trying to define the quality of software products in terms of some specific measurable attributes which are in the interest of the user. In view of software product quality, reliability is highly desirous criteria for
assessment before the launching of product in the market. External attributes of the product are used to assess the reliability of the product under development. Defect density is one of the external attributes that represents the different quality aspects of the product. The typical definition of defect density is “a ratio between numbers of known defects with the size of the product in thousand line of code” [Westfall, (2012)]. This measure gives an indication of the product readiness in terms of reliability. The defect density is considered by most of the developers as an indicator of product reliability whenever the product is about to release [Zhu and Faller, (2013)]. However, this external measure is used by the developers to prioritize the possible improvement in different modules of the software product. Whenever subsequent version/release of any product is developing, this parameter is also used to compare the quality with the previous version/release of the product [Westfall, (2012)].

A larger number of researchers is indicating the dependency of the defect density measure on the size of the product. In addition to the defect density, size is also an important measure in software engineering. The principle objective of most of the software organizations is to follow modular programming approach in their software development. It is also widely believed that the modular approach to software development decreases the design complexity. It is already a well-known phenomenon that size of a module is positively related to the defect proneness of the module. Some corrective actions will be taken after identifying any error/bugs in the software product. Meanwhile, few lines of code in the defective modules either added or deleted. This process of modification in the module changes the size of the module. Hence, the defect density of product will vary according to the size of the software modules. This varying size of modules affects the reliability of software product due to changes in the defect density. However, most of the studies indicating that the defect proneness is increasing with the size of the module; few studies showed the opposite trend of defect density with the module size. Hence, there is a need to research about the optimal module size with minimum defect density.

There are some essential parameters those are affecting defect density of the product; for example, size, defects, coupling, etc. These parameters provide the ideas about how to identify the possible ways of minimizing the defect insertion rate. Hence, the analysis of these factors is required to identify the possible technique for improvement in product reliability. The initial defect density estimated with the help of these analyses can further
use for the planning of testing. Further, the effort to the testing activities can be resourcefully distributed with the help of initial defect density measure. Two contemporary models use multiple parameters for defect density estimation [Munson and Khoshgoftaar, (1996); Farr, (1996)]. The performance of these models depends on the number of these factors. More number of factors gives the best performance in the defect density estimation [Takahashi and Kamayachi, (1985)]. ROBUST [Malaiya and Denton, (1997); Li and Malaiya, (1995)] and RADC [Munson and Khoshgoftaar, (1996); Farr, (1996)] are the models which are using the multiplicativity of factors for improving the performance of their estimation.

Previous software reliability models use the code complexity, developer’s experience, and the process of development [Bhattacharjee et al. (2011)]. These factors only influence the performance of models when there are frequent changes in the software requirement [Malaiya and Denton, (1999)]. On the basis of previous studies, it can be inferred that the defects are randomly distributed in the different modules of the software systems. Hence, this nature of defects shows higher dependency on the size of the module.

An effective distribution of modules sizes throughout the software development activities is the need for research in the interest of software developers. This distribution of module sizes helps to minimize the defect density of the product, and consequently, the reliability of the product can be improved. There are several methods that have been proposed to show the impact of software size on the defect density. Further, the prediction of defect density using the size of the module is the most desirable research topic for software quality improvement.

In the view of defect density prediction, defect prediction in the module is a very important activity. More number of studies are discussed only the prediction of faulty modules that will not serve the purpose to estimate the defect density. There are some methods those are used for prediction of the number of defects during the testing activity of development [Staron and Meding, (2008)].

In this chapter, a model has been formulated that can be used for prediction of defect density with variable size of modules. The model has been used to estimate the optimal size of the module for minimum defect density. This model is formulated with the
inspiration of Ferdinand (1974) and the classification of defects according to the Malaiya and Denton (2000). The study of Ferdinand (1974) emphasized on the relation between numbers of defects and number of the code segment. This study states that the number of defects is proportional to a power of \( n \). Further, in a study of Malaiya and Denton (2000), they categorized the defect in two classes. Firstly, the defect may arise due to the integration between modules. Secondly, the defects those may occur due to improper implementation of the modules. The model proposed in this chapter formulated with the consideration of these two phenomena.

### 3.2 RELATED WORK

The size of the module plays an important role in the software design. Software design with the larger size of the module, increase the design complexity. Hence, decomposing the larger modules into smaller modules makes some less complex design as compared to the larger size modules. The software design for the small module size is effortless as compared to larger module size. Therefore, the design of the software may be improved by minimizing the size of the module.

It is a well-known phenomenon that the complexity of the module is positively related to the size of the module. Further, a more complex module is more error prone compared to the less complex module. Research on the software complexity and errors [Basili and Perricone, (1984)] contradict this phenomenon. Basili and Perricone (1984) study states that the larger modules are less defective as compared to the smaller one. The major limitation of this study is that they used FORTRAN modules with size less than 200 lines. According to the analysis results of the study, it can easily be observed that this opposite trend is true even for the complex module. The dataset used in this study contain a large number of smaller modules, and there were too few modules of higher size. The most believable fact which supports this opposite behavior in this study is that a large number of interface errors were equally distributed in all modules. The other reason behind this trend is that the few larger modules were developed with more care, whereas less attention was given to a large number of smaller modules by the developers.
A similar approach discussed by Alsmadi and Najadat (2011), in which similar classification has been done from a large number of modules those are randomly selected from the tested data sets. These types of classifications can be further used to estimate the defect density due to the faulty modules.

Weyuker et al. (2008) formulated a methodology to assess the effectiveness of the software development, software maintenance, and software validation techniques [Weyuker et al. (2008)]. The method suggested through this study attempts to recognize the defect properties with different possible values. Further, these values used to describe the faults in different areas. Another study has been carried out by the same authors to illustrate the usability of fault data from successive releases of the industrial product [Weyuker et al. (2010)]. Also, the authors explained the relationship of fault proneness with the size of the module.

In a study of Ada software [Withrow, (1990)], most of the modules are larger in size compared with the Basili and Perricone’s (1984) study, which examined the modules written in FORTRAN Language. The analysis result of this study shows a decreasing trend in defect density up to a certain size of the module and after that the defect density increases with module size. The analysis has been performed in all sizes of modules where most of the modules are larger than 200 lines. The major difference between the study of Basili and Perricone (1984) and Withrow (1990) is the measurement of the size in two different languages. The earlier one only counted the executable lines and did not consider the common data statements, while in later, both declaration and body lines have been counted. Both the studies did not consider the commented lines in their analysis. The analysis result of this study justifies the hypothesis given by Banker and Kemerer (1989) where they have reconciled two opposing trends about the existence of economies or diseconomies of scale in software development.

Hatton (1997) indicated that the larger and smaller size software modules are less reliable compared to the average size module. The size of the module near about 200-400 lines of code has been considered as an average size. Modules beyond this size have been considered less reliable or more error prone. The analysis has been performed on the Withrow’s (1990) dataset. The analysis results in this study showing logarithmic growth in the number of defects for software product of the size up to 200 lines of code.
In another study of Rosenberg (1997), the fair relationship has been discussed between module size and defect density. The author’s compound their study with threefold uses of source line of code; (1) the formatting of source line of code does not affect the counting, (2) the source line of code is correlated with other metrics of software those are affecting the quality of software, and (3) source line of code is inversely related to the software defect density. This third observation of Rosenberg (1997) has been used by Malaya and Denton (2000) into the formulation of their model.

Malaiya and Denton (2000) proposed a model that describes how the module size distribution can control the defect density in the software system. They have considered the inverse relationship of module size with its defect density to formulate the model. This study performs the analysis on three different datasets i.e. Withrow (1990) data for Ada modules, Basili and Perricone’s (1984) data of Fortran modules, and Columbus assembly data are given by Hatton (1997). The analysis has been performed to identify how effectively the module size distribution can lead to minimizing the overall defect density. The analysis results in this study support the argument of Withrow’s (1990) and Hatton (1997); larger and smaller modules are more error prone compared to the middle size modules. The model considers two types of defect i.e. integration fault and instruction fault. The result provides the interpretation for both decreasing and increasing trends in defect density with the size of the module. The decreasing trend has been seen up to a certain size after that the increasing trend has been observed with the higher size of modules. The conclusion of this study drowns in these two regions that were supporting the previous studies about the relationship between defect density and module size.

### 3.3 FORMULATION OF MODEL

In this chapter, a model has been proposed which was inspired by Malaiya and Denton (2000) study. For classifying the faults in the model, two mechanisms have been considered. In today’s scenario, most of the software development organizations are developing the software products by following a modular approach. These modules itself developed by using the set of instructions. This approach of software product
development has been incorporated by Malaiya and Delton (2000) to formulate their model. In addition to the types of faults and module sizes, the model uses some parameters to estimate the defect density. The limitation of this model is that the model does not describe the approach to control the defect density with variable size of the modules. The model is restricted to only three parameters. It seems to modify the model by adding some more parameters that minimize the defect density with varying size of modules. This chapter introduces a new model that is an improved version of the model proposed by Malaiya and Delton (2000). In this proposed model, two new parameters have been added for the consideration of the statement giving by Ferdinand (1974); the number of defects is proportional to a power of \( n \) to the size of software. Further, the model has been analyzed to minimize the defect density with varying size of modules. The model has been formulated by considering two categories of faults where the first category considered as an integration faults and the second category considered as implementation faults. The integration faults may arise due to improper way of value transfer between modules or incorrect parameters passing between modules. This category of faults is related to more than one module. Hence, these can also be termed as module-related faults. The occurrence of the second category of faults may be due to improper implementation of instructions in the modules. This second category of faults is only related to the single module in which the faulty instruction is implemented. Hence, these can be termed as implementation faults. In the proposed model defect density due to the first category of faults will be denoted by \( D_m \) and due to the second category of faults, the defect density will be denoted by \( D_i \). The overall defect density in the system by adding both categories of faults will be denoted by \( D \).

### 3.3.1 Integration Fault

These types of faults are generated due to the parameters passing amongst the modules of the software product. Here the faults are related to more than one module. The occurrence of these faults may infer due to global data declaration as well as postulate by one module to the other modules in the product. In this proposed model, it has been considered that these types of faults are uniformly distributed among all modules. According to the
previous studies discussed in section 3.2, defects are proportionally decreased with increase in module size.

Defect density $D_m$ for module-related faults with size $s$ given by

$$D_m(s) \propto \frac{1}{s^l},$$

$$D_m(s) = \frac{\alpha}{s^l}, \quad \alpha \propto \quad \text{................................................................. (3.1)}$$

Where, $l$ any non-zero integer that is added to the model to reduce the defect density, and $\alpha$ is a suitable parameter. It has been assumed in this model that the size of any module should not be less than one. Hence, the minimum value of $s$ in equation (3.1) is one and $\alpha$ is a suitable parameter for proportionality constant. This relationship of defect density with module size presented in equation (3.1) support the model proposed by Shen et al. (1985).

### 3.3.2 Implementation Fault

The reason of occurring of this category of faults may be due to improper implementation of module instructions. Hence, these faults can be considered as implementation faults of instructions. These types of faults only relate to the module in which the instruction is improperly implemented. It can be understood that the larger modules have more number of instructions hence increase the probability of more faults compared to the smaller modules. To estimate the defect density with respect to these faults, here two components have been considered which are related to the probability of incorrect instructions. The first component assumed as constant that is $\beta$ and another component depended on the interaction among instructions in the same module and represented as $\gamma$. Hence, this second component is proportional to the module size $s^n$. So that the defect density $D_i$ due to instruction-related defects can be expressed as

$$D_i = \beta + \gamma s^n, \quad \text{................................................................. (3.2)}$$
Where $\beta$ and $\gamma$ are the parameters that affect the insertion of second category of defects and $n$ is any non-zero parameter added to reduce the defect density. From equation (3.1) and (3.2), the total defect density can be expressed as

$$D(s) = D_m(s) + D_i(s) = \frac{\alpha}{s^\beta} + \beta + \gamma s^n$$ .................................................... (3.3)

The equation (3.3) represents the sum of the defect density that occurs due to both categories of faults. In the above equation (3.3), the defect density $D(s)$ is considered as a dependent variable, whereas the size of module $s$ is considered as an independent variable.

It is understandable from the equation (3.3) that the total defect density can be optimized with respect to module size $s$. According to the mathematical analysis, maxima and minima concept can be applied to calculate the maximum and a minimum value of defect density of the equation (3.3) under some assumption. These assumptions have been considered that the equation (3.3) is well defined and differentiable in its domain.

Now for finding the minimum value of defect density, differentiating the equation (3.3) with respect to $s$,

It gives,

$$\frac{d}{ds} D(s) = \frac{-1 \alpha}{s^{\beta+1}} + n \gamma s^{n-1}$$ ................................................................. (3.4)

For maxima and minima equate the equation (3.4) to zero, put $\frac{d}{ds} D(s) = 0$,

$$\frac{-1 \alpha}{s^{\beta+1}} + n \gamma s^{n-1} = 0$$ ................................................................. (3.5)

$$s = \left(\frac{\alpha}{n \gamma} \right)^{\frac{1}{\beta+1}}$$ ................................................................. (3.6)
Equation (3.6) gives the minimum value of size $s$. Further, to calculate the minimum defect density with respect to this minimum size of modules, the second derivative of the equation (3.4) has been taken that gives

$$\frac{d^2}{ds^2} D(s) = \frac{l(l+1)\alpha}{s^{l+2}} + n(n-1)\gamma s^{n-2} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots
• First Case

First the value of \( l = 1 \) and \( n = 1 \), has been taken that it gives \( D_m(s) = \frac{\alpha}{s} \) and \( D_f(s) = \beta + \gamma s \), and

Total defect density is given by

\[
D(s) = D_m(s) + D_f(s) = \frac{\alpha}{s} + \beta + \gamma s \\
\tag{3.10}
\]

Differentiating to the equation (3.10) with respect to \( s \), it gives

\[
\frac{-\alpha}{s^2} + \gamma = 0 \\
\tag{3.11}
\]

Hence, the module size \( s_{\text{min}} \) for minimum defect density is given by

\[
s_{\text{min}} = \sqrt[\gamma]{\alpha} \\
\tag{3.12}
\]

And the minimum defect density is given by

\[
D_{\text{min}} = \left(2\sqrt{\alpha\gamma} + \beta\right) \\
\tag{3.13}
\]

• Second Case

If the value of \( l = 2 \) and \( n = 2 \), taken then it gives \( D_m(s) = \frac{\alpha}{s^2} \) and \( D_f(s) = \beta + \gamma s^2 \), and

Total defect density is given by

\[
D(s) = D_m(s) + D_f(s) = \frac{\alpha}{s^2} + \beta + \gamma s^2 \\
\tag{3.14}
\]
Hence, the module size $s_{\text{min}}$ for minimum defect density is given by

$$s_{\text{min}} = \left(\frac{\alpha}{\gamma}\right)^{1/4}$$ ... (3.15)

And the minimum defect density is given by

$$D_{\text{min}}(s) = \frac{\alpha}{s^2} + \gamma \left(\frac{\alpha}{\gamma}\right)^{1/2} + \beta$$ ... (3.16)

- **Third Case**

If the value of $l=2$ and $n=1$ taken then it gives $D_m(s) = \frac{\alpha}{s^2}$ and $D_f(s) = \beta + \gamma s$, and

Total defect density is given by

$$D(s) = D_m(s) + D_f(s)$$

$$= \frac{\alpha}{s^2} + \beta + \gamma s$$

$$\frac{d}{ds} D(s) = 0 \Rightarrow -\frac{2\alpha}{s^3} + \gamma = 0 \Rightarrow -\frac{2\alpha}{s^3} = -\gamma \Rightarrow s^3 = \frac{2\alpha}{\gamma}$$

and under some suitable constraints over $\alpha, \beta$ and $\gamma$.

Hence, the module size $s_{\text{min}}$ for minimum defect density is given by

$$s_{\text{min}} = \left(\frac{2\alpha}{\gamma}\right)^{1/3}$$ ... (3.17)

And the minimum defect density is given by

$$D_{\text{min}}(s) = \frac{\alpha}{(2\alpha)^{1/3}} + \beta + \gamma \left(\frac{2\alpha}{\gamma}\right)^{1/3}$$ ... (3.18)
Fourth Case

If the value of \( l = 1 \) and \( n = 2 \) taken then it gives \( D_m(s) = \frac{\alpha}{s^l} \) and \( D_l(s) = \beta + \gamma s^2 \), and

Total defect density is given by

\[
D(s) = D_m(s) + D_l(s) = \frac{\alpha}{s^l} + \beta + \gamma s^2
\]

\[
\frac{d}{ds} D(s) = 0 \Rightarrow -\frac{\alpha}{s^2} + 2\gamma s = 0, \Rightarrow s^3 = \left(\frac{\alpha}{2\gamma}\right) \Rightarrow s = \left(\frac{\alpha}{2\gamma}\right)^{1/3}
\]

and under some suitable constraints over \( \alpha, \beta \) and \( \gamma \).

Hence, the module size \( s_{\text{min}} \) for minimum defect density is given by

\[
s_{\text{min}} = \left(\frac{\alpha}{2\gamma}\right)^{1/3} \quad \text{.................................................................................................................. (3.19)}
\]

And the minimum defect density is given by

\[
D_{\text{min}}(s) = \frac{\alpha}{\left(\frac{\alpha}{2\gamma}\right)^{1/3}} + \beta + \gamma \left(\frac{\alpha}{2\gamma}\right)^{2/3} \quad \text{.................................................................................................................. (3.20)}
\]

By restricting the values of constant \( \alpha, \beta \) and \( \gamma \) under some special values, the values of defect density given by this model can be classified in two different trends. The first trend of defect density indicates the decreasing value by increasing the size of the module. On the other hand, after a certain value (optimum value) of module size, the defect density will increase with the size of the module. These two different trends in the defect density with the size of modules support the previous analysis of Malaiya and Delton (2000).
3.4 ANALYSIS OF PROPOSED MODEL

In the previous section of this chapter, a model has been developed which was indicating the relationship between defect density and module size. This model needs to be analyzed with the different available data sets. This section elaborates about the data collected for analysis of the model and calculation of the different parameters used in the model. Further, the calculation of the defect density with the parameter values, and finally the optimal size of the module for minimum defect density have been discussed.

3.4.1 Data Collection

The model proposed in this chapter is an improved model of Malaiya and Delton (2000), where two additional parameters have been added to minimize the defect density. These two new parameters are restricted to positive integers, although more values of these can be further used to analyze the model in future. In this chapter, the model has been analyzed with two discrete values of these two new parameters i.e. \( l = 1, n = 1 \) and \( l = 2, n = 2 \). These two values have been considered as a special case (i) and case (ii) respectively. Since the proposed model is the improved version of the previous model [Malaiya and Delton, (2000)], the same data set has been used to perform the analysis. In addition to the two data sets used by the previous model [Malaiya and Delton, (2000)], a new data set is also used to validate the model with the current data set. The first data set has been taken from the data made available by Basili and Perricone (1984). The limitation of this dataset is that it only comprises small sizes of modules. There are too few modules, larger than 200 lines of code in this data set. This dataset is presented in Table 3.1. Another data set has been taken from the data given by Withrow (1990); the previous model also uses the data set [Malaiya and Delton, (2000)]. In this data set, the sizes of modules are larger than compared to the first data set. This data set has been presented in Table 3.2. In addition to these two data sets, the third data set has been collected from the 23 closed source software project [PDR-1, (2013)]. The information about these closed source software that has been collected includes some modules, size, and defect density. The data set has been presented in Table 3.3.
3.4.2 Parameter Calculation

This section of the chapter describes the methodologies used to estimate all the parameter values used in the proposed model. The analysis of the proposed model has been performed on three data sets as mentioned in the previous section of this chapter. Additional new parameters are restricted to the positive integer while other parameter values for $\alpha, \beta, and \gamma$ have been estimated using the method of least squares also known as a curve fitting method.

As mentioned in the previous section only two special cases have been considered to analyze the proposed model, i.e. case (i) and case (ii). Parameters calculation for others values of $l$ and $n$ (for example, greater than 2) increase the complexity of calculations and provide the values of parameters in large frition of decimal point. Hence, the parameter values have been estimated for first two special cases. In the first data set, there are too few modules with size larger than 200 lines of code. Analysis result with this data set is not showing an increasing trend of defect density with the increasing size of modules while the same data is expressing the decreasing trend in defect density with increasing size of modules. Here the parameter $\gamma$ plays a very little role to establish the relationship of defect density with the size of the modules. Hence, only the one trend is shown in this data, the optimal module size of the module and minimum defect density for this data set can’t be defined. So these two values in Table 3.1 and Table 3.2 kept blank.

All estimated parameter values and the optimal size of the module for which the defect density will be minimum are mentioned in Table 3.1 and Table 3.2 for a case (i) and case (ii) respectively. The optimal size of the module and minimum defect density for case (i) has been estimated using equation (3.12) and equation (3.13) respectively.
Table 3.1: Parameter Values with special case (i)

<table>
<thead>
<tr>
<th>Data</th>
<th>Parameter Values</th>
<th>Optimal size of module</th>
<th>Minimum defect density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \alpha )</td>
<td>( \beta )</td>
<td>( \gamma )</td>
</tr>
<tr>
<td>Basili</td>
<td>-6.136</td>
<td>18.384</td>
<td>-0.0489</td>
</tr>
<tr>
<td>Withrow</td>
<td>36.41</td>
<td>4.13</td>
<td>0.0009</td>
</tr>
<tr>
<td>PDR</td>
<td>704.71</td>
<td>0.125</td>
<td>0.00005</td>
</tr>
</tbody>
</table>

Table 3.2: Parameter Values with special case (ii)

<table>
<thead>
<tr>
<th>Data</th>
<th>Parameter Values</th>
<th>Optimal size of module</th>
<th>Minimum defect density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \alpha )</td>
<td>( \beta )</td>
<td>( \gamma )</td>
</tr>
<tr>
<td>Basili</td>
<td>6960.4</td>
<td>13.6</td>
<td>-0.0001</td>
</tr>
<tr>
<td>Withrow</td>
<td>2253.3</td>
<td>4.544</td>
<td>0.000000154</td>
</tr>
<tr>
<td>PDR</td>
<td>98450</td>
<td>0.5499</td>
<td>7E-10</td>
</tr>
</tbody>
</table>

Similarly, for the case (ii), these two values have been estimated using equation (3.15) and equation (3.16) respectively.
3.4.3 Defect Density Calculation

In this section, calculated defect density values have been presented using the parameter values mentioned in Table 3.1 and Table 3.2. After putting these parameters values and module size in the proposed model, the defect densities for each module size has been calculated with respect to the case (i) and case (ii).

Table 3.3: Basili data [Basili and Perricone, (1984)]

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Size (LOC)</th>
<th>Modules</th>
<th>DD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>1</td>
<td>1-50</td>
<td>258</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>51-100</td>
<td>70</td>
<td>12.6</td>
</tr>
<tr>
<td>3</td>
<td>101-150</td>
<td>26</td>
<td>12.4</td>
</tr>
<tr>
<td>4</td>
<td>151-200</td>
<td>13</td>
<td>7.6</td>
</tr>
<tr>
<td>5</td>
<td>&gt;200</td>
<td>3</td>
<td>6.4</td>
</tr>
</tbody>
</table>

These calculated values of defect density for case (i) and case (ii) have been compared with actual observed values from the data set itself. First, the defect density has been estimated for the Basili and Perricon (1984) data set. This comparison of estimated values for both the cases and actual observed values have been listed in Table 3.3, and comparison using the graph has been shown in Figure 3.1.
It can be easily understood from the graph shown in Figure 3.1 that only one trend is showing by the estimated defect density as well as actual observed values of this data set. Decreasing rate in defect density with increase size of modules can be seen from the graph. More linearity in defect density with varying module size has been observed in the proposed model compared to the previous model [Malaiya and Delton, (2000)]. The sizes of modules in case (ii) are varying in a geometric progression that gives the minimum defect density for this type of module size distribution.

As mentioned in section 3.4.1, the data set of Basili and Perricone’s (1984) consist small size of modules only. That stimulated only a single trend in the proposed model analysis. This relationship of defect density with module size, using this dataset, has been supported by previous studies [Banker and Kemerer, (1989); Rosenberg, (1997)]. Another dataset proposed by Withrow (1990) has been considered in analyzing the both trends in the model. The inclusion of larger as well as smaller size modules inspired to use this data set for analysis. The estimated defect densities using this data in case (i) and case (ii) with the actual observed defect density have been mentioned in Table 3.4. The comparison of both cases and observed values of defect density can be seen from the graph shown in Figure 3.2.
Table 3.4: Withrow data [Withrow, (1990)]

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Size (LOC)</th>
<th>Modules</th>
<th>DD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>1</td>
<td>1-62</td>
<td>93</td>
<td>5.4</td>
</tr>
<tr>
<td>2</td>
<td>63-97</td>
<td>39</td>
<td>4.9</td>
</tr>
<tr>
<td>3</td>
<td>98-154</td>
<td>52</td>
<td>3.4</td>
</tr>
<tr>
<td>4</td>
<td>155-250</td>
<td>53</td>
<td>1.8</td>
</tr>
<tr>
<td>5</td>
<td>251-397</td>
<td>46</td>
<td>5.2</td>
</tr>
<tr>
<td>6</td>
<td>398-625</td>
<td>31</td>
<td>5.6</td>
</tr>
<tr>
<td>7</td>
<td>626-949</td>
<td>22</td>
<td>6.8</td>
</tr>
<tr>
<td>8</td>
<td>950-5160</td>
<td>26</td>
<td>8.3</td>
</tr>
</tbody>
</table>

The analysis results with this data set given in Table 3.4 shows both the trends in defect density with respect to the increasing size of modules. Initially, the defect density is decreasing with the increasing size of modules. At the optimal size of the module the minimum defect density has been observed. After an optimal value of module size the rising trend in density has been observed in the case (i) and case (ii). This relationship of defect density with the variable size of modules support the both the trend proposed in the previous model [Malaiya and Delton, (2000)].
The growing trend in defect density, with respect to the larger modules, could be due to improper testing of the large size of modules compared to the smaller modules. It may be possible that testing of the large size of modules requires more test cases, but the testing team only create fewer test cases. This improper testing with limited test cases may lead to uncovering some defects from the modules. These undiscovered defects in the larger modules lead to higher defect density. Hence, the higher defect density rate has been observed in large size modules in comparison to the smaller modules.

The first data set comprises the FORTRAN modules, whereas the second data set comprises the Ada modules. These two traditional data sets have been used by the previous model [Malaiya and Delton, (2000)]. In this chapter, other data set has been used to perform the analysis. This third data set has been collected from the PDR [PDR-1, (2013)]. This data set comprises modules programmed in Java.

The estimated defect densities using this data in case (i) and case (ii) with the actual observed defect density have been mentioned in Table 3.5. The comparison of both cases and observed values of defect density can be seen from the graph shown in Figure 3.3.
The analysis result of this data set also shows both the trends in defect density with respect to the increasing size of modules similar to the results with Withrow data set.

**TABLE 3.5: PDR Data [PDR-1, (2013)]**

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Size (LOC)</th>
<th>Modules</th>
<th>DD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>1.</td>
<td>1-132</td>
<td>30197</td>
<td>5.6</td>
</tr>
<tr>
<td>2.</td>
<td>133-465</td>
<td>5209</td>
<td>1.39</td>
</tr>
<tr>
<td>3.</td>
<td>466-1263</td>
<td>1065</td>
<td>0.77</td>
</tr>
<tr>
<td>4.</td>
<td>1264-2927</td>
<td>222</td>
<td>0.66</td>
</tr>
<tr>
<td>5.</td>
<td>2928-9878</td>
<td>40</td>
<td>0.61</td>
</tr>
</tbody>
</table>

**Figure 3.3:** Defect Density Variation of PDR (2013) data with case (i) and case (ii)
Initially, from the Figure 3.3, the defect density is decreasing with the increasing size of modules. At the optimal size of the module the minimum defect density has been observed. After this point, the rising trend in density has been observed in the case (i) and case (ii). This relationship of defect density with the variable size of modules support the both the trend proposed in the previous model [Malaiya and Delton, (2000)].

The analysis result of proposed model using these data sets established the relationship between defect density and module size. According to the graph shown in Figure 3.3, it can be inferred that the better reduction in defect density is given by case (ii) comparison to the case (ii). Small variations in the defect density have been seen because in this data set most of the modules are larger in size. Similar to the previous data set, it is understood that once the module size is distributed on a geometric progression, the rate of defect density reduction can be increased significantly.

### 3.4.4 Validation of Model

The previous section presented the estimated values of defect density using the proposed model for the case (i) and case (ii). All the estimated values for the proposed model indicate the observation how the defect density will be affected by varying the size of the modules. This observation needs to be statistically fit the measured value of the proposed model. Measures of goodness of fit normally review the inconsistency between observed values and the values expected under the model in question.

In this section, all the statistics to measure of goodness of fit have been calculated using the SPSS statistical analysis tool [SPSS, (2013)]. In this goodness of fit measure, three statistics have been calculated to find the fitness of the proposed model. The Sum of Squares Due to Error (SSE)\(^1\) statistics gives the measurement of the deviation of response values from the fit to response values. The value of this statistical measure closer to 0 indicates that the model has a smaller random error component, and the fit will be more useful for prediction. Another statistics R-Square\(^2\) (\(R^2\)) measures that how successful the model is explaining the variation of data to the fitted values. The values of this measure near to 1 indicates that most of the proportion of the variance in explaining by the model

---

1 See APPENDIX for more details

2 See APPENDIX for more details
in question. The value of this measure lies between 0 and 1. For example, an R$^2$ value 0.79 indicating that the model can predict the 79% of the variation in the dependent variable by the independent variable. The third measure, p-value$^3$, used to decide whether the null hypothesis of the study in question can be accepted or rejected. In detail, the value of this measure gives the probability of accepting or rejecting the null hypothesis. The lowest p-value (less than 0.05) indicates that the null hypothesis has to be rejected. Alternatively, the lower value of this parameter indicates that the predictor or independent variables can be added to the model. In this study, the null hypothesis has been considered as “there is no goodness of fit between the observed value of defect density and fitted values of defect density.” The results of this statistical measurement for the goodness of fit, for all three data sets used in this chapter with respect to special case (i) and case (ii), have been presented in Table 3.6.

**Table 3.6: Goodness of fit measure statistics**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Case</th>
<th>Measure Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MSE</td>
</tr>
<tr>
<td>Basilli</td>
<td>case (i)</td>
<td>0.238</td>
</tr>
<tr>
<td></td>
<td>case (ii)</td>
<td>0.227</td>
</tr>
<tr>
<td>Withrow</td>
<td>case (i)</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>case (ii)</td>
<td>0.029</td>
</tr>
<tr>
<td>PDR</td>
<td>case (i)</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>case (ii)</td>
<td>0.992</td>
</tr>
</tbody>
</table>

It is indicated from the Table 3.6 that all the three data sets give more goodness of fit measure for case (ii) compared to the case (i) and actual observed value.

---

$^3$ See APPENDIX for more details
3.5 PERFORMANCE EVALUATION

The previous model discussed the relationship between defect density with the module size [Malaiya and Delton, (2000)]. In this study, an effective distribution of modules according to the size indicates the minimization of defect density. The model proposed by Malaiya and Delton (200) is only restricted to the smaller size of modules, whereas in the model proposed in this chapter analyzed with the large data sets that comprise the broader range of modules. Two new parameters are added to the proposed model to give more effective size distribution for minimum defect density comparison to the previous model [Malaiya and Delton, (2000)]. According to the analysis results of the proposed approach with all three data sets (two data set as used by Malaiya and Perricone’s and data set from PDR ); improvement in reduction rate of defect density has been observed compared to the previous model [Malaiya and Delton, (2000)]. The analysis result for Basili and Perricone’s (1984) data sets, case (ii) gives closer results to the actual observed values compared to the case (i). Similarly, for Withrow and PDR data sets, the result gives improvised reduction in defect density for optimal module size. This impact of varying size of modules on the defect density can be observed from Figure 3.1, Figure 3.2, and Figure 3.3. For all three data sets, case (ii) gives better results compared to other cases. Although, in this proposed model, the parameters are restricted to the integer value up to 2, more improvement in the defect density reduction can be obtained with other values of these parameters.

3.6 SUMMARY

In this chapter, a model has been proposed that established the relationship between defect density and module sizes of the software product. This relationship would be used to make the prediction about the defect proneness of modules, and this information can be utilized to refractor the size distribution of the module. This redistribution of module size according to the proposed model gives an improvement in reduction of defect density. Hence, the quality in terms of reliability of the product can be improved significantly. The analysis of this proposed model shows that the defect density is decreasing in the smaller size of modules, after an optimal size of the module; the defect density is growing with
the size of the modules. In this approach, the module size distribution in geometric progression has been suggested to minimize the defect density. The proposed model gives effective module size distribution using two new additional parameters comparisons to previous contemporary model [Malaiya and Delton, (2000)]. The model has been analyzed with three data sets; two traditional data sets as used by the previous model [Malaiya and Delton, (2000)] and third publicly available data set has been used from the Promise Data Repository. Two special cases have been analyzed where the additional parameters are restricted to positive integer values; the result shows optimization in defect density reduction comparison to the actual values of defect density. It is suggested through this model that the effective distribution of modules with respect to the optimal module size gives a better reduction in defect density. The module size may be redistributed as; the smaller modules may combine to form a module in which size is near to optimal module, and too large modules can be broken into modules whose sizes are near to optimal module. Further, the proposed model can be implemented using more variables with varying values to achieve a higher reduction in defect density. Consequently, this type of consideration about the parameters needs to take care of tradeoffs between the desired results and product complexity.