The Hyper-Geometric Distribution Software Reliability Growth Model (HGDM):
Precise Formulation and Applicability.

Raymond Jacoby
Yoshihiro Tohma
Department of Computer Science, Tokyo Institute of Technology
Ookayama 2-12-1, Meguro-Ku, Tokyo 152
Email: {jacobym, tohma]@cs.titech.ac.jp/relay.cs.net

Abstract

In this paper, the Hyper-Geometric Distribution is applied to make estimations for the number of initial faults residual in a software at the beginning of the test-and-debug phase. The Hyper-Geometric Distribution Growth Model (HGDM) is well suitable to make estimations on the observed growth curves of the accumulated number of detected faults. The advantage of our model is the applicability to all kind of observed data. By application of a single model, exponential growth curves, as well as S-shaped growth curves can be estimated.

We first present the precise formulation of the HGDM. Secondly the exact relationship of our model to the NHPP Goel-Okumoto Growth Model and the Delayed S-shaped Growth Model will be shown. By assuming w(i), the sensitivity factor of our model appropriately, we will establish the S-shaped HGD Growth Model. With the introduction of a variable fault detection rate, the goodness of fit of the estimated growth curve to the growth curve of real observed faults is increased significantly.

Different examples of applicability of our model to real observed data show the characteristics of the HGD Model.

Index Terms - hyper-geometric distribution, software reliability growth model, software test and debugging, initial software faults estimation, S-shaped estimation model.

1 Introduction

In a software development process, a program is designed, coded, test-and-debugged, and finally put into operation. Test-and-debugging is mostly carried out by some test team, independent from the programmers of the piece of software. Usually, the test-and-debug team detects many more faults, in spite of the programmers' confidence of the quality of their product.

At the beginning of the test-and-debug phase, nobody knows definitely how many faults are still resident in the software product. Also, test workers are unable to certify when the program under test can be released into service. Therefore, we have to rely on the estimation of the number of initial residual software faults. Such estimations are of major interest to guarantee high reliability and quality assessments of a program after the test-and-debug phase.

Different estimation models have been proposed, based on the Gompertz curve, the logistic curve, and Non-Homogeneous Poisson Process (NHPP). In [10], our model based on the hyper-geometric distribution was presented and in [11] an approximated expression for the mean value function of the model was given. We distinguish between the manifestation of faults and the detection of faults in the estimation process for the number of initial faults \( E[m] \) of a given program under test. \( w(i) \), the sensitivity factor in our model which is a measure to represent how many faults manifest themselves as errors upon the application of a test instance \( i \), is one of the key concepts.

A second main characteristic is the applicability of our model to various kind of observed data. Exponential growth, as well as S-shaped growth can be estimated. Since at beginning stages of the test-and-debug phase nobody knows about the shape of the growth curve of the observed detected faults, various growth models must be applied to real observed data. In order to discard such multiple model applications, the applicability of a single model to various kind of data sets is favorable.

The first aim of this paper is to show the precise formulation of our model. In a next section, the exact mutual relationship of our model to the Goel-Okumoto NHPP Growth Model and the Delayed S-shaped Growth Model is given. We then establish the S-shaped Hyper-Geometric Distribution Growth Model with its variable fault detection rate. Finally, the results of applicability of our model to different sets of real observed test-and-debug data will be compared to the results obtained by other growth models.

2 Basic Concept and Precise Formulation of the HGDM

2.1 Basic Concept

With our model, we argue the detection and removal (fixing) of faults at the test-and-debug stage. At the beginning of the test-and-debug stage \( m \) initial faults are in the program under test. With the application of test instances (i.e. test runs, ...), these faults are detected. We call a "set" of test instances the test.

Considering the distinction between the manifestation and detection of faults at the application of a test instance, we proposed the Hyper-Geometric Distribution Model for the estimation of the number of initial faults. The basic assumptions and concepts have been made clear in [10] and in [11].

![Figure 1: Manifestation of w(i) Faults](image)

\( w(i) \) is defined as the sensitivity factor. It is a measure to represent how many faults manifest themselves as errors upon the application of test instance \( i \). For the first test instance \( t(1) \), the number of detected (and removed) faults is, of course, \( w(1) \). On the application of \( t(2) \), the number of newly detected faults is not necessarily \( w(2) \), because some of these \( w(2) \) faults may have been detected and removed by \( t(1) \). And so on for all
2.2 Precise Formulation

In this model the probability that \( N(i) = x \) is given by

\[
Prob(x|m, C(i - 1), u(i)) = \frac{m - C(i - 1)}{m} \cdot \left( \frac{m - C(i - 1)}{m} \right)^x \cdot \left( 1 - \frac{m - C(i - 1)}{m} \right)
\]

where \( 0 \leq x \leq C_i, C_i = \min\{u(i), m - C(i - 1)\} \).

The probabilistic distribution of Eq.(1) is called hyper-

\[E[C(i)] = E[C(i - 1)] \left( 1 - \frac{u(i)}{m} \right) + w(i) \quad \text{(2)}\]

Eq.(2) is a recursive equation. The solution of \( E[C(i)] \) in Eq.(2) into a non-recursive form can be calculated. For definition, \( E[C(1)] = w(1) \), and therefore

\[
E[C(2)] = w(1) \cdot \left( 1 - \frac{w(2)}{m} \right) + w(2) = \frac{m - w(1)}{m} \cdot \left( 1 - \frac{w(2)}{m} \right) + \frac{m - w(2)}{m} + m - m = \frac{m - w(1)}{m} \cdot \left( 1 - \frac{w(2)}{m} \right) + \frac{m - w(2)}{m} \cdot \left( 1 - \frac{w(1)}{m} \right)
\]

\[
E[C(i)] = m \cdot \left[ 1 - \prod_{j=1}^{i-1} \left( 1 - \frac{w(j)}{m} \right) \right], \quad \forall i = 1 \ldots n, \quad \text{(3)}
\]

with \( E[C(0)] = 0 \). Proof for \( \text{Eq}(3) \) by induction.

Eq.(3) can also be changed into an exponential function of

\[
E[C(i)] = m \cdot \left( 1 - \frac{\sum_{j=1}^{i} \ln \left( 1 - \frac{w(j)}{m} \right) }{m} \right), \quad \forall i = 1 \ldots n, \quad \text{(4)}
\]

2.3 \( w(i) \), the sensitivity factor of the HGD Model

In [4], [5], sensitivity factor \( w(i) \) was related to information available from data sets. Therefore, the following "ease of test" functions have been applied.

\[
w(i) = X(i) \cdot (a \cdot i + b), \quad \text{with} \quad X(i) = \{ \text{number of tester(s) or computer times}(i) \}
\]

where \( i \) represents the \( i \)th test instance. Such function for \( w(i) \) takes into consideration the linear change of the ease of test along the progress in test. For this paper, in order to compare our model precisely to other models, we use \( w(i) \) as

\[
w(i) = a \cdot i + b \quad \forall i = 1 \ldots n,
\]

2.4 Parameters and Evaluation of Optimal Parameter Values

\( w(i) \), the sensitivity factor of this model, and \( m \), the total number of initial faults are unknown parameters to be estimated. \( E[a], E[b], E[m] \). Their values are calculated tentatively using the full scan over possible range of values. (Our research is guided to a new methodology to determine the parameter values analytically.) Thus, \( E[C(i)] \) for \( i = 1, 2, \ldots \) of Eq.(3) and Eq.(4) is obtained by making the comparison with actually observed \( C(i) \), using an evaluation function that measures the minimal distance between the observed growth curve and the estimated growth curve at test instances \( i \), as the following

\[
EF1 = \frac{1}{n} \sum_{i=1}^{n} \left| E[C(i)] - \hat{C}(i) \right|
\]

Those values \( E[m], E[a] \) and \( E[b] \) that minimize this \( EF1 \) function are taken as the optimal parameter values for the estimation of the number of initial faults.

This method is different from the maximum likelihood method (MLM) of parameter value estimation. With the MLM, normalized data is used together with the Kolmogorov-Smirnov Goodness of Fit Test to calculate analytically the optimal parameter values. Such normalization process implies that the last estimated number of cumulated faults is identical to the last observed number of cumulated faults. Such a condition is unrealistic in a real test-and-debug and estimation environment. Furthermore, through normalized data the original difference between observed values and estimated values of the cumulated number of faults is changed.

Therefore, in order to keep the original difference between \( C(i) \) and \( E[C(i)] \) at test instances \( i \) and in order to respect the real conditions of a test-and-debug environment, we apply \( EF1 \) for determination of the optimal parameter values.

3 Precise Mutual Relationship to Other Models

In this section, the exact mathematical relationship of the HGD Model to other models, especially NHPP models, is shown. It is proved that the Goel-Okumoto NHPP Model can be represented exactly by the HGD Model. It is also possible to define the relationship between the Delayed S-shaped Model and the HGDM, but the estimates for \( E[m] \) are not the same for both models. The reason for such different \( E[m] \) estimates is explained by the assumption of a variable fault detection rate \( p(i) \).

3.1 Mutual Relationship of HGDM and NHPP Goel-Okumoto Model

The Goel-Okumoto NHPP exponential model has the following mean value function of a nonhomogeneous Poisson process
This equation can be compared to Eq.(4), considering $E[m]$ of same value in both models. The following relationship between the two models can be easily derived, with $w(i) = b = \text{constant}.$

$$\phi = \ln \left(1 - \frac{b}{E[m]}\right), \forall i = 1 \ldots n$$

Resolving Eq.(8) for $b,$ we have

$$b = E[m] \cdot \left[1 - e^{-\phi}\right], \forall i = 1 \ldots n$$

Thus, the exact mathematical relationship between the two models is given by Eq.(8) and Eq.(9).

**Example 1**

Table 1 gives the results of estimations for the data in [2]. The estimated growth curves are shown in Fig. 3. As supported by the mathematical proof, both estimated growth curves for HGDM and NHPP, EF1 as well as the estimates for $E[m]$ of both models coincide. The existing difference in result of [2] and our model is founded on the different approach to parameter value determination as discussed in Section 2.4.

<table>
<thead>
<tr>
<th>Model</th>
<th>$E_[m]$</th>
<th>Parameter Values</th>
<th>EF1 - value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGDM, EF1</td>
<td>141.37</td>
<td>$b_1 = 16.503$</td>
<td>4.293922</td>
</tr>
<tr>
<td>NHPP, EF1</td>
<td>141.37</td>
<td>$\phi = 0.1246$</td>
<td>4.557564</td>
</tr>
</tbody>
</table>

**Figure 3:** Comparative Estimation Results for Data of [3]

### 3.2 Mutual Relationship of HGD Model and Delayed S-Shaped Growth Model

In this section, the mathematical relationship of the HGD Growth Model to the Delayed S-shaped Growth Model is established. The mean value function of the Delayed S-shaped Growth Model is given by

$$\hat{h}(i) = E[m] \cdot \left(1 - e^{-\rho i}\right), \forall i = 1 \ldots n$$

Let us define $f(p, i)$ such as $f(p, i) = (1 + \rho \cdot i) \cdot e^{-\rho i}.$ Taking Eq.(3) into account, and having the same $E[m]$ for both models, the following holds.

$$\prod_{i=1}^{i} \left(1 - \frac{w(i)}{E[m]}\right) = f(p, i) = (1 + \rho \cdot i) \cdot e^{-\rho i}$$

Assuming the same for $(i-1),$ and dividing as $f(p, i)/f(p, i-1),$ we have

$$w(i) = E[m] \cdot \left[1 - \frac{f(p, i)}{f(p, i-1)}\right], \text{with } f(p, 0) = 1 \text{ and } i \geq 1$$

$$\quad = E[m] \cdot \left[1 - \frac{(1 + \rho \cdot (i-1) + \rho) \cdot e^{\rho i}}{(1 + \rho \cdot (i-1))} \right]$$

Thus, the exact mathematical relationship between the two models is given by Eq.(8) and Eq.(9).

For Eq.(12) two cases of parameter value interpretation are possible. The equality only holds in the following two situations:

1. For $i = 1,$ $\rho$ can be calculated numerically and from such a value $\rho$ all other $w(i),$ for $i \geq 2,$ can be calculated. Such a constant $\rho$ defines a set of $w(i) = a_i + b$ values that are not the same at each $i.$ Therefore the estimates for $E[a]$ and $E[b]$ are not constant at test instances $i.$

2. $w(i)$ is not a constant for all test instances $i.$ If the optimal $E[a]$ and $E[b]$ are constant values for all $i,$ $\rho$ is not constant. In this case we consider a variable fault detection rate, as discussed in the next Section, S-shaped HGD Model.

These two situations are the reason that the respective estimated $E[m]$ for the HGD Model and the Delayed S-shaped Growth Model are different although the equality in Eq.(12) holds.

**Example 2**

For a data set given in [8], the estimated numerical results are given in Table 2 and the estimated growth curves are shown in Fig. 4.

It can be seen from the EF1-values of Table 2 that the estimated growth curve of the HGD Model fits better the real observed growth curve than the estimated one by the Delayed S-shaped growth model ( Del. S-shaped, EF1). Furthermore, the estimated growth curve of the HGD Model fits even much better than the growth rate obtained in [8] by the method of normalized data.

<table>
<thead>
<tr>
<th>Model</th>
<th>$E[m]$</th>
<th>Parameter Values</th>
<th>EF1 - Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGDM, EF1</td>
<td>64.3</td>
<td>$w(i) = 0.304 \cdot i + 0.53$</td>
<td>0.837071</td>
</tr>
<tr>
<td>Del. S-, EF1</td>
<td>76.38</td>
<td>$\rho = 0.0052040$</td>
<td>0.989637</td>
</tr>
<tr>
<td>[2]</td>
<td>71.248</td>
<td>$\rho = 0.103307$</td>
<td>1.042750</td>
</tr>
</tbody>
</table>

**Table 2:** Results of Estimations for the Data of [9]
4.1 S-shaped HGD Model and Variable Fault Detection Rate

In Eq.(12) we have established the relationship of the HGD Model to the Delayed S-shaped growth model. In the case where \( w(t) = a + b \) and \( E[a] \) and \( E[b] \) are the optimal estimated parameter values at all test instances \( i \), \( \rho \) is not a constant value for all \( i \). With the assumption of the same value \( E[i] \) for both models, the equality of equations Eq.(3) and Eq.(10) is given by the following relationship

\[
\prod_{i=1}^{n} (1 - \frac{a_j + b_j}{m}) = (1 + \rho(i) \ast i) \ast e^{-\rho(i) \ast i}, \forall i = 1 \ldots n \tag{13}
\]

This equality holds only in the case in which \( \rho \) of Eq.(10) is replaced by the not constant \( \rho_i \). In Eq.(13), \( \rho(i) \) is not a constant value for all \( i \) and we can therefore introduce \( \rho(i) \), the variable fault detection rate.

Example of variable fault detection rate \( \rho(i) \)

In order to represent the variable \( \rho(i) \), the optimal estimated parameter values of Example 2, \( E[m] = 64.3, E[a] = 0.304 \), and \( E[b] = 0.53 \) are used with Eq.(13). In Table 3, the numerical values of the variable fault detection rates \( \rho(i) \) that satisfy the equality at test instances \( i \) in Eq.(13) are given. Only in the case of a variable \( \rho(i) \), the Delayed S-shaped growth curve of [8] can calculate the same growth curve as obtained by the S-shaped HGD Model.

In Fig. 5, the variable fault detection rate \( \rho(i) \) is represented for 40 test instances. The failure intensity function for the optimal parameter values of the HGD model is depicted together with the failure intensity function for the Delayed S-shaped growth model with \( E[m] = 64.3 \) with its estimated optimal constant fault detection rate \( \rho \) = 0.11699. It can be seen that the Delayed S-shaped growth model with its constant fault detection rate \( \rho \) cannot calculate the better fitting growth curve of the HGD Model.
4.3 Conclusion for S-shaped HGD Model

Comparing the HGD Model with \( w(i) = a + i + b \), \( a > 0 \) and \( b \geq 0 \), to the Delayed S-shaped Growth Model, a variable fault detection rate \( \rho(i) \) is realized. The usage of such variable fault detection rate \( \rho(i) \) has a positive impact on the goodness of fit of the estimated growth curves to the real observed growth curve. Therefore, the introduction of a variable fault detection rate into the theory on software reliability modeling, as realized by the HGD Model, can be seen as a great improvement from previous growth models that consider only constant fault detection rates. Also, a variable fault detection rate seems more realistic in a real test-and-debug environment.

5 Applicability to Various Kind of Data

In the following, four data sets of real observed test-and-debug data will be analyzed. The HGD Model is applied to the estimate the number of initial faults and the results of estimations are compared to those obtained by other models.

5.1 Example 3, Software Project X

For a software project X the test-and-debug data was reported. The data lists day-by-day, for 109 debug days, the number of faults newly detected and the number of testers involved in a test.

In Table 5, the results of estimations are given. The estimated growth curves are plotted in Fig. 6.

For this data the Good-Okamoto NHPP Model is inappropriate, as can be seen from the estimate for \( E[a] \) of the HGD Model and the goodness of fit values. The estimated growth curves obtained by the HGD Model are best fitting the real observed data. This better fit of the estimated growth curve to the real observed growth curve is realized by the assumption of the variable fault detection rate \( \rho(i) \) in our model, as depicted in Fig. 6.

![Figure 6: Estimated Growth Curves for Software Project X](image)

Also, the application of NHPP models in combination with the evaluation function \( EF1 \), Eq.(6), gives better fitting estimated growth curves than the curves obtained by maximum likelihood method, as shown by the numerical results in Table 5.

![Table 5: Results of Estimations for 109 Test Instances](image)

5.2 Example 4, Railway Interlocking System

The data analyzed here was given in [10]. It is the bug report about a software consisting of about 14.5 kilolines of ASSEMBLER language for a railway interlocking system. For this data, only the number of newly detected faults was reported. Table 6 gives the numerical results of the estimations and the respective estimated growth curves, as well as the variable fault detection rate \( \rho(i) \)-curve are plotted in Fig. 7.

The estimated growth curve of the HGD Model (HGD, \( EF1 \)) fits the best the real observed data due to the variable fault detection rate with an estimated \( E[a] = 65.6 \).

![Figure 7: Esti. Growth Curves for Railway Interl. Sys. Sw.](image)

![Table 6: Results of Estimations for 199 Test Instances](image)

5.3 Ex. 5, PL/1 Application Program Test Data

The data of Table 7 is the test data of a PL/1 database application program [8]. The size of the software is about 1.317 kilolines of code. The data consists of 19 weeks of reported detected bugs. The execution times and the number of detected faults per week are reported. The total number of observed failures is 358.

Since the observations are recorded on a week-by-week basis and by the elapsed execution times, we show the applicability of the HGD Model to this data for the following two methods: Applicability Based on Time of Observation (Weeks), and Applicability Based on Elapsed Execution Times (4). The results of estimations will show that the HGD Model is well applicable in both cases.
5.3.1 HGD Model and Time of Observation (t)

In [11] we discussed the application of our model to this data with \( w(t) = \text{exec.time}(t) \cdot (a \cdot t + b) \). Here, we use \( w(t) = a \cdot t + b \) as "ease of test" function. In Table 8, the numerical results of estimations based on observation times (Weeks) are given. For reason of comparing the goodness of fit, the numerical results of estimation for the Delayed S-shaped model and the NHPP Goel-Okamoto, in combination with the evaluation function \( EF_1 \) of Eq.(6) are also listed.

The estimated growth curves of the HGD Model and those of the Delayed S-shaped growth model and the NHPP Goel-Okamoto model are depicted in Fig. 8.

![Figure 8: Estimated Growth Curves for PL/I Database](image)

The HGD Model estimates the better fitting growth curves to real observed data in each case, with realistic estimates for \( E[m] \).

5.3.2 HGD Model and Elapsed Execution Times (t)

Eq.(3) gives the exact formulation of our model for test instances \( t(i) \). If we want to apply our model to the cumulative execution times of Table 7, we need to change the notion of test instance. We regard one test instance to be the smallest time slice of the recorded data. In this example here, the smallest time slice \( t_s = 0.01 \) execution times, and therefore we can use 47.65 test instances. Eq.(3) is used as follows

\[
E[C(t_s)] = E[m] \cdot \left[ 1 - \prod_{j=t_s}^{t_n} \left( 1 - \frac{w(j)}{m} \right) \right]
\]

\( t_1 \leq t_2 \leq \ldots \leq t_n \) and with \( E[C(0)] = 0 \).

In Table 9, the numerical results of estimations based on elapsed execution times are given. The estimated growth curves of the HGD Model and those of the Delayed S-shaped growth model, the NHPP Goel-Okamoto model, and the Inflection S-shaped growth model are depicted in Fig. 9.

![Figure 9: Estimated Growth Curves for PL/I Database Based on Elapsed Execution Times](image)

The HGD Model estimates the best fitting growth curve. The estimate for \( E[m] \) is even very close to the real observed number of initial software faults. The goodness of fit for the NHPP models is increased by using \( EF_1 \) instead of the Maximum Likelihood Method.

5.3.3 Conclusion

This example shows the applicability of our model to test instances based on time observations and based on elapsed execution time. With Eq.(14) our model is well applicable to make estimations, considering the elapsed time to fault detection.

One more observation is relevant. For this data it is difficult to judge which one of the NHPP models estimates the better fitting curve.
growth curve to the real observed data. With the application of the HGDM we do not need to worry about the probable shape of the observed growth curve. The parameters of the model are responsible for a smooth fitting to the real observed data.

5.4 Example 6, Software Project Y, [12]
The data analyzed here was given in [11]. For this data, the number of newly detected faults and the number of test items was reported. We argued the segmentation of this data in [11]. In this paper here, with no segmentation of the data set, we make estimations with \( w(i) = a * i + b \) for 46 test days.

![Figure 10: Estimated Growth Curves for Software Project Y](image)

The estimated growth curve of the HGDM (HGDM, EF1) and the Goel-Okumoto growth model, with EF1, are the same, and fit the best the real observed data.

6 Conclusion
In this paper the exact formulation of the Hyper-Geometric Distribution model as software reliability growth model is established. This exact formulation enables an easy comparison to other models. We presented the relationship of the HGDM to the Goel-Okumoto NHPP Growth Model and the Delayed S-shaped growth model. The relationship to the Delayed S-shaped growth model can be established mathematically. But, as shown in various examples, the estimated growth curves obtained by the HGDM are better fitting the real observed data than those obtained by the Delayed S-shaped model. We introduce the concept of a variable fault detection rate. Due to this variable fault detection rate, the goodness of fit is increased for growth curves estimated by our model.

A very important property of the HGDM is that it can be applied to various kind of data. With one single model, exponential growth as well as S-shaped growth can be estimated. This means, using this overall model, we do not need to worry about which model is to be applied to which real observed data. This is a very favorable situation for the reliability estimations based on reliability growth curves.

We are proceeding our research activities to a more general "ease of test" function \( w(i) \) and to define an analytical method for the parameter value determination. Furthermore, we try to apply our model for estimating software availability during the operational phase.

Acknowledgment. This work was supported by the Ministry of Education of Japan under Grant of Scientific Research. We also express our thanks to companies which kindly provided us with real test-and-debug data.

References