Research Contribution

Precise Formulation and Applicability of a Software Reliability Growth Model Based on Hyper-Geometric Distribution

RAYMOND JACOBY*† and YOSHIHIRO TOHMA*

In this paper, the Hyper-Geometric Distribution is used to estimate the number of faults in a program at the beginning of the test-and-debug phase. The Hyper-Geometric Distribution Growth Model (HGD Model) is well suited to estimating the growth curves of the observed accumulated number of detected faults. The advantage of the model is its applicability to all kinds of observed data. Application of a single model makes it possible to calculate exponential growth curves, as well as S-shaped growth curves.

First, the HGD Model is precisely formulated. Next, the exact relationship of the model to the NHPP Goel-Okumoto Growth Model and the Delayed S-shaped Growth Model is shown. Assumption of an appropriate value of $w(i)$, the sensitivity factor of the proposed model, will establish the S-shaped HGD Growth Model. The introduction of a variable fault detection rate significantly increases the goodness of fit of the estimated growth curve to the growth curve of actually observed faults.

Various examples of the applicability of our model to actually observed data demonstrate the characteristics of the HGD Model.

1. Introduction

In a software development process, a program is designed, coded, tested-and-debugged, and finally released. Test-and-debug is mostly carried out by a testing team independent of the software programmers. Usually, the test-and-debug team detects numerous faults, whatever the programmers' confidence in the quality of their product.

At the beginning of the test-and-debug phase, nobody knows exactly how many faults are still resident in the software product. Furthermore, the testing team members are unable to guarantee when the program being tested can be released. Therefore, it is necessary to rely on the estimate of the number of initial software faults. Such estimates are of major interest to guarantee high reliability and quality assessments of a program after the test-and-debug phase.

Various estimation models have been proposed, based on the Gompertz curve, the logistic curve, and Non-Homogeneous Poisson Process (NHPP). In a previous paper [11], we presented our model based on the hyper-geometric distribution. In the model, we distinguish between the manifestation of faults and the detection of faults in the process for estimating the number of initial faults ($E[m]$) of a given program at the test-and-debug phase. One of the key concepts is $w(i)$, the sensitivity factor in our model, which is a measure that represents how many faults manifest themselves as errors upon the application of a test instance i.

A second main characteristic is the applicability of our model to various kinds of observed data. Exponential growth as well as S-shaped growth can be estimated, depending on the parameter values of the model. Since in the initial stages of the test-and-debug phase nobody knows the shape of the growth curve of observed detected faults, various growth models must be applied to the actually observed data. A single model should be applicable to various kinds of data set, since it allows such multiple model applications to be discarded.

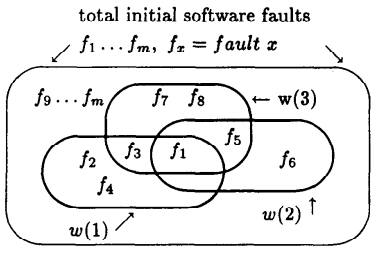
The first aim of this paper is to formulate our model precisely. In the next section, we give the exact mutual relationship of our model to the Goel-Okumoto NHPP Growth Model and the Delayed S-shaped Growth Model. We then establish the S-shaped Hyper-Geometric Distribution Growth Model with its variable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

2. Basic Concept and Precise Formulation of the HGD Model

2.1 Basic Concept

A program has been developed and debugged. When

*Department of Computer Science, Tokyo Institute of Technology, Meguro-ku, Ookayama, 2-12-1, Tokyo 152, Japan.
†Currently with TOSHIBA Corporation, Systems & Software Engineering Laboratory, Saiwai-ku, Kawasaki-shi, 210, Japan.



$w(1) = \{f_1, f_2, f_3, f_4\}$, $|w(1)| = 4$
$w(2) = \{f_1, f_5, f_6\}$, $|w(2)| = 3$
$w(3) = \{f_1, f_5, f_3, f_7, f_8\}$, $|w(3)| = 5$
$w(4) = \ldots$, $|w(i)|$ =number of faults sensed by $t(i)$

Fig. 1 Manifestation of $w(i)$ Faults.

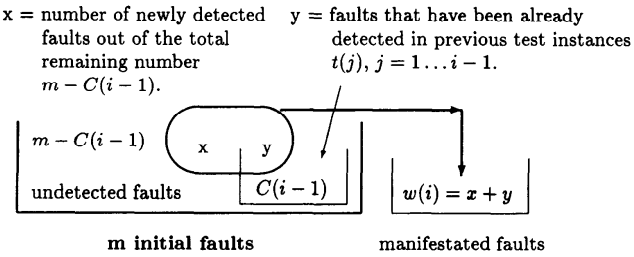


Fig. 2 Basic Idea of the HGD Model.

the programmers are confident that there are no more faults left, the product will be passed to the testing team for more rigorous test-and-debug. Usually, more faults are found by the testing team. In our model, we argue for the detection and removal (fixing) of faults at this test-and-debug stage. At the beginning of the test-and-debug stage m faults are in the program being tested. With the application of *test instances* (such as test runs), these faults are detected. We call a "set" of test instances a test.

Bearing in mind the distinction between the manifestation and detection of faults at the application of a test instance, we propose the Hyper-Geometric Distribution Model as a means of estimating the number of initial faults [11]. We make the following assumptions for the Hyper-Geometric Distribution Model:

1. Faults that manifest themselves upon the application of a test instance $t(i)$ will be removed (fixed) before the next test instance $t(i+1)$ is applied.
2. During the removal of detected faults, no new faults will be inserted. Thus, the reliability of the program will grow along with the progress in the test-and-debug phase. (This assumption is subject to criticism, but we are presently optimistic that it can be removed in future.)
3. A random set of $w(i)$ faults are sensed by test instance $t(i)$ out of m initial faults. They may or may not have been detected in previous test instances.

$w(i)$ is defined as the **sensitivity factor**. It is a measure that represents how many faults manifest themselves as errors upon the application of test instance $t(i)$. For the first test instance $t(1)$, the number of detected (and removed) faults is of course $w(1)$. When $t(2)$ is applied, the number of **newly** detected faults is not necessarily $w(2)$, because some of these $w(2)$ faults may already have been detected and removed by $t(1)$, and so on for all $t(i)$, $i=3 \ldots n$. In the example shown in Fig. 1, some of the $w(2)$ faults have been detected in $w(1)$, and the number of newly detected faults for $w(2)=2$.

In test instance $t(i)$, let $C(i-1)$ be the cumulative number of faults newly detected by $t(1)$, $t(2)$, . . . , $t(i-1)$, and let $N(i)$ be the number of faults newly detected by $t(i)$. $C(0)$ is defined to be 0. Some of the $w(i)$ faults sensed by $t(i)$ may already have been counted in $C(i-1)$; the rest of the $w(i)$ faults are newly detected. This basic idea is depicted in Fig. 2.

2.2 Precise Formulation

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

$$Prob(x \mid m, C(i-1), w(i)) = \frac{\binom{m-C(i-1)}{x}\binom{C(i-1)}{w(i)-x}}{\binom{m}{w(i)}} \tag{1}$$

where $0 \le x \le U_x$, $U_x = \min\{w(i), m-C(i-1)\}$.

The probabilistic distribution of Eq. (1) is called a hyper-geometric distribution. The expected value of $N(i)$, denoted by $\overline{N(i)}$, is [1]

$$\overline{N(i)} = \{m - C(i-1)\}\frac{w(i)}{m} \tag{2}$$

and by definition, the cumulative number of newly detected faults is given as

$$C(i-1) = \sum_{k=1}^{i-1} N(k) \tag{3}$$

From now on, the estimate for a parameter P will be denoted by $E[P]$. Thus, we can use an estimate for $C(i-1)$ such as

$$E[C(i-1)] = \sum_{k=1}^{i-1} \overline{N(k)} \tag{4}$$

and substitute it for $C(i-1)$ in Eq. (2). That is,

$$\begin{aligned} E[C(i)] &= E[C(i-1)] + \{m - E[C(i-1)]\}\frac{w(i)}{m} \\ &= E[C(i-1)]\left\{1 - \frac{w(i)}{m}\right\} + w(i) \end{aligned} \tag{5}$$

$E[C(i)]$ in Eq. (5) can be solved in a non-recursive form as follows. For example let us take the first two members of Eq. (5), $E[C(1)]$ and $E[C(2)]$. By defini-

tion, $E[C(1)]=w(1)$. Therefore

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

Thus, the equation for $E[C(2)]$ can be changed as follows:

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

Hence, we obtain the following theorem:

Theorem: Eq. (5) can be rewritten as the following non-recursive equation:

$$\boxed{E[C(i)]=m*\left[1-\prod_{j=1}^{i}\left(1-\frac{w(j)}{m}\right)\right] \quad \forall i=1\ldots n,} \tag{6}$$

with $E[C(0)]=0$.

Proof: by induction.

For $i=1$:

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

Assume the theorem holds for $i=n-1$. Then,

for $i=n$:

$$E[C(i)]=E[C(i-1)]*\left(1-\frac{w(i)}{m}\right)+w(i)$$
$$=m*\left[1-\prod_{j=1}^{i-1}\left(1-\frac{w(j)}{m}\right)\right]*\left(1-\frac{w(i)}{m}\right)+w(i)$$
$$=m*\left[\left(1-\frac{w(i)}{m}\right)-\prod_{j=1}^{i}\left(1-\frac{w(j)}{m}\right)\right]+w(i)$$
$$=m*\left[1-\prod_{j=1}^{i}\left(1-\frac{w(j)}{m}\right)\right] \quad \text{Q.E.D.}$$

Eq. (6) can be also be rewritten as an exponential function. The product within Eq. (6) can be resolved by applying the exponential function $x=e^{\ln x}$ to

$$\prod_{j=1}^{i}\left(1-\frac{w(j)}{m}\right)=e^{\ln\left(\prod_{j=1}^{i}\left(1-\frac{w(j)}{m}\right)\right)}$$

$$\prod_{j=1}^{i}\left(1-\frac{w(j)}{m}\right)=e^{\sum_{j=1}^{i}\ln\left(1-\frac{w(j)}{m}\right)}$$

and therefore, Eq. (6) can be written as

$$\boxed{E[C(i)]=m*\left[1-e^{\sum_{j=1}^{i}\ln\left(1-\frac{w(j)}{m}\right)}\right] \quad \forall i=1\ldots n,} \tag{7}$$

with $E[C(0)]=0$.

2.3 Approximated Expression for the HGD Model

Let us consider the following difference equation [12]:

$$E[C(i)]-E[C(i-1)]=\Delta E[C(i-1)]$$
$$=w(i)\left\{1-\frac{1}{m}E[C(i-1)]\right\}. \tag{8}$$

Changing this difference equation into a differential equation of a continuous function of i, and ignoring the difference between $i-1$ and i, we get

$$\frac{\partial E[C(i)]}{\partial i}=w(i)\left\{1-\frac{1}{m}E[C(i)]\right\} \tag{9}$$

$E[C(i)]$ can then be expressed as the following general function:

$$E[C(i)]=m*\left(1-e^{\left(-\frac{1}{m}\int_0^i w(x)\,dx\right)}\right) \tag{10}$$

Eq. (10) is a similar expression to the one given for hardware reliability models [7].

2.4 The Sensitivity Factor $w(i)$ of the HGD Model

In several previous papers [5, 6, 12], the sensitivity factor $w(i)$ was related to information available from data sets. Therefore, the following function was applied:

$$w(i)=X(i)*(a*i+b), \textit{ with}$$
$$X(i)=\{\textit{number of tester}(i) \text{ or } \textit{computer time}(i) \text{ or } \textit{test items}(i) \text{ or } 1.0\} \tag{11}$$

where i represents the ith test instance. This function takes account of the linear change in the ease of testing as the testing progresses.

2.5 Parameters and Evaluation of Optimal Parameter Values

The sensitivity factor $w(i)$ and the total number of initial faults m are unknown parameters to be estimated, ($E[a]$, $E[b]$, $E[m]$). Their values are tentatively calculated by using the full scan over a possible range of values. (Our research is directed toward a new methodology for determining the parameter values analytically.) Thus, $E[C(i)]$ for $i=1, 2, \ldots$ in Eq. (6) and Eq. (7) is obtained by comparison with the actually observed $C(i)$, using an evaluation function that measures the minimal distance between the observed growth curve and the estimated growth curve in test instance i, as follows:

$$EF1=\frac{1}{n}\sum_{i=1}^{n}|C(i)-E[C(i)]| \tag{12}$$

The 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 (MLH) method for parameter value estimation [9, 13]. With the MLH method, normalized data are used together with the Kolmogorov-Smirnov Goodness of Fit Test to calculate analytically the optimal parameter values. The normalization process implies that the last estimated cumulative number faults is identical to the last observed cumulative number of faults. This condition is unrealistic in a real test-and-debug and estimation environment. Furthermore, the normalized data changes the original difference between the observed and estimated values of the cumulative number of faults.

Therefore, to preserve the original difference between $C(i)$ and $E[C(i)]$ in test instance i and to respect the actual conditions of a test-and-debug environment, we apply EF1 to determine the optimal parameter values.

3. Mutual Relationship to Other Models

This section shows the exact mathematical relationship of the HGD Model to other models, especially NHPP models. We prove 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 Growth Model and the HGD Model, but the estimated results for $E[m]$ are not the same for both models. This difference is explained by the assumption of a variable fault detection rate $p(i)$.

3.1 Mutual Relationship of the HGDM and the NHPP Goel-Okumoto Model

The Goel-Okumoto NHPP exponential growth model has the following mean value function of a nonhomogeneous Poisson process:

$$\hat{m}(i)=E[m]*(1-e^{-\phi i}) \quad \forall i=1 \ldots n \tag{13}$$

This equation can be compared to Eq. (7), assuming $E[m]$ to have the same value in both models. The following relationship between the two models can be easily derived, with $w(j)=constant$,

$$w(j)=b \quad constant \quad \forall j;$$

$$e^{\sum_{i=1}^{i} ln\left(1-\frac{b}{E[m]}\right)}=e^{i*ln\left(1-\frac{b}{E[m]}\right)}=e^{-\phi i}$$

$$\boxed{\phi=-ln\left(1-\frac{b}{E[m]}\right) \quad \forall i=1 \ldots n} \tag{14}$$

Table 1 Results of Estimations for the Data of [3].

Model	$E[m]$	Parameter Values	EF1-value	Kolm.-Smirnov Value
HGDM, EF1	141.37	$b_1=16.503$	4.293922	0.121346
NHPP, EF1	141.37	$\phi=0.124131$	4.293922	0.121346
[3]	142.32	$\phi=0.1246$	4.557864	0.122594

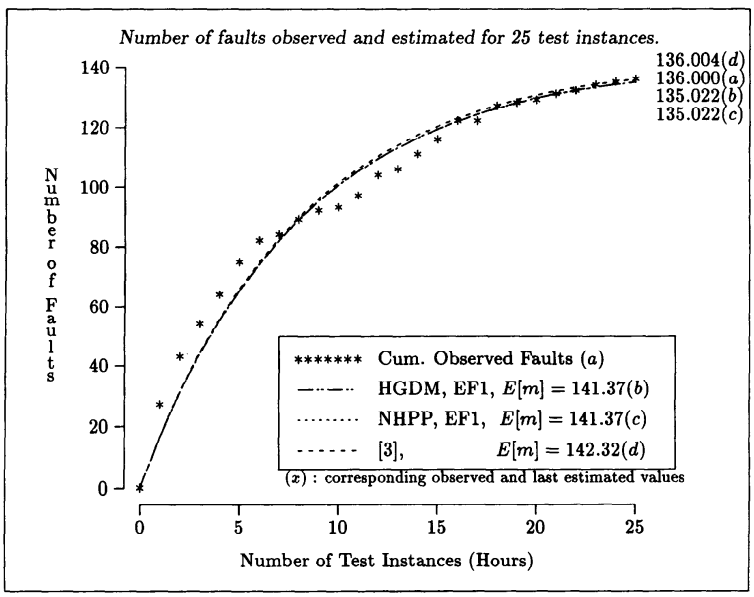


Fig. 3 Comparative Estimation Results for Data in Goel [3].

Resolving Eq. (14) for b, we have

$$\boxed{b=E[m]*[1-e^{-\phi}] \quad \forall i=1 \ldots n} \tag{15}$$

Thus, the exact mathematical relationship between the two models is given by Eq. (14) and Eq. (15).

Example 1.

The results of estimations for the data in Goel [3] are given in Table 1. The estimated growth curves are shown in Fig. 3. As the mathematical proof indicates, the estimated growth curves EF1 for both HGDM and NHPP as well as the estimates for $E[m]$ of both models coincide. The difference between Goel's results [3] and ours stems from the different approach to parameter value determination, which we discussed in Section 2.

3.2 Mutual Relationship of the HGD Model to the 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

$$\hat{g}(i)=E[m]*(1-(1+\rho i)*e^{-\rho i}) \quad \forall i=1 \ldots n \tag{16}$$

Let us define $f(\rho, i)$ so that $f(\rho, i)=(1+\rho*i)*e^{-\rho i}$. Taking Eq. (6) into account, and using the same $E[m]$ for both models, we can state that

$$\text{for } i\text{: } \prod_{j=1}^{i}\left(1-\frac{w(j)}{E[m]}\right)=f(\rho, i)$$

$$\text{for } i-1\text{: } \prod_{j=1}^{i-1}\left(1-\frac{w(j)}{E[m]}\right)=f(\rho, i-1)$$

Dividing these functions,

$$\frac{\prod_{j=1}^{i}\left(1-\frac{w(j)}{E[m]}\right)}{\prod_{j=1}^{i-1}\left(1-\frac{w(j)}{E[m]}\right)}=\left(1-\frac{w(j)}{E[m]}\right)=\frac{f(\rho, i)}{f(\rho, i-1)} \tag{17}$$

we have

$$w(i)=E[m]*\left[1-\frac{f(\rho, i)}{f(\rho, i-1)}\right] \quad \text{with } f(\rho, 0)=1$$

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

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

$$\boxed{w(i)=E[m]*\left[1-\left(1+\frac{\rho}{1+\rho*(i-1)}\right)*e^{-\rho}\right] \quad \forall i=1 \ldots n} \tag{18}$$

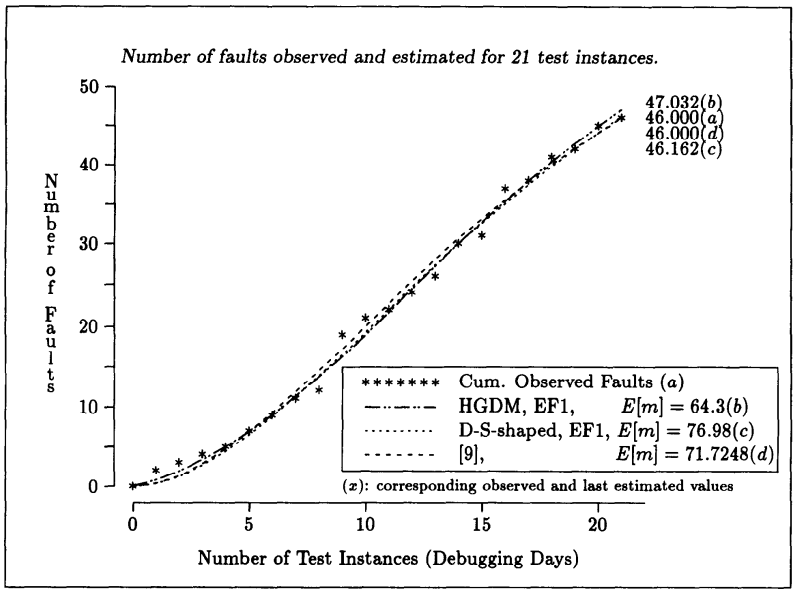


Fig. 4 Comparative Estimation Results for Data in Ohba [9].

Table 2 Results of Estimations for the Data in Ohba [9].

Model	$E[m]$	Parameter Values	EF1-value	Kolm.-Smirnov Value
HGDM, EF1	64.3	$w(i)=0.304*i+0.53$	0.837071	0.094860
S-Shaped, EF1	76.98	$\rho=0.096240$	0.988674	0.099198
[9]	71.7248	$\rho=0.103967$	1.042750	0.114289

For Eq. (18) two interpretations of the parameter value are possible. The equality only holds in the following two cases:

1. For $i=1$, ρ can be calculated numerically, and from the value of ρ all other $w(i)$ can be calculated for $i \geq 2$. The constant ρ defines a set of $w(i)=ai+b$ values that are not the same at each i. Therefore the estimates for $E[a]$ and $E[b]$ are not constant in test instance 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, ρ is not constant. In this case we consider a variable fault detection rate, as discussed in the next section, on the S-shaped HGD Model.

Because of these two cases for Eq. (18), the respective estimated values of $E[m]$ for the HGD Model and the Delayed S-shaped Growth Model are different, although the equality in Eq. (18) holds.

Example 2.

For the data set in Ohba [9], 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 the actually observed growth curve better than does that of the Delayed S-shaped growth model (D-S-shaped, EF1). In fact, it fits the observed growth curve even better than the growth curve obtained in Ohba [9] by the method of normalized data.

4. The S-shaped HGD Model and Variable Fault Detection Rate

Example 1 showed that the HGD Model can estimate an exponential, actually observed growth curve. Example 2 gave the estimated results of the HGD Model for an S-shaped growth curve. Depending on the choice of the function for the sensitivity factor $w(i)$ and its parameter values, the HGD Model is well suited to all kind of data set. In this section, we will study the S-shaped behavior of the HGD Model more precisely.

4.1 S-shaped HGD Model and Variable Fault Detection Rate

In Eq. (18) we established the relationship of the HGD Model to the Delayed S-shaped growth model. When $w(i)=a*i+b$ and $E[a]$ and $E[b]$ are the optimal estimated parameter values in all test instances i, ρ is not a constant value for all i. With the assumption of the same value $E[m]$ for both models, the equality of Eqs. (6) and (16) is given by the following relationship:

Table 3 Comparison of Results for Variable Fault Detection Rate.

i	$\prod_{j=1}^{i}\left(1-\frac{aj+b}{m}\right)$	$(1+\rho(i)*i)*e^{-r(i)*i}$	$\rho(i)$	$\rho=0.11699$
1	0.987030	0.987029	0.17041	0.993668
2	0.969561	0.969562	0.13483	0.976544
...	...	...	...	...
6	0.860002	0.859996	0.10901	0.843520
7	0.824452	0.824464	0.10802	0.801972
8	0.786473	0.786470	0.10771	0.759316
9	0.746526	0.746521	0.10786	0.716303
10	0.705078	0.705075	0.10835	0.673533
...	...	...	...	...
40	0.013577	0.013577	0.16116	0.058045
40	0.010898	0.010899	0.16348	0.052722

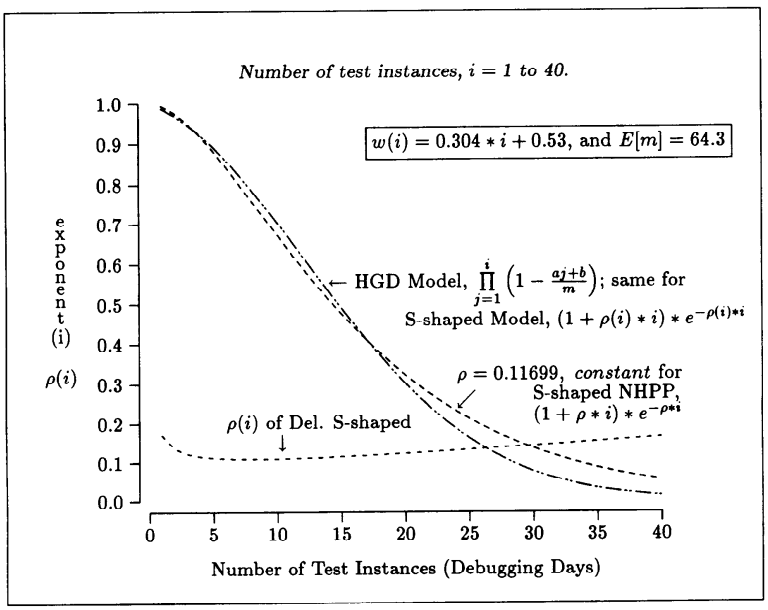


Fig. 5 Variable Fault Detection Rate for $\rho(i)$.

$$\prod_{j=1}^{i}\left(1-\frac{aj+b}{m}\right)=(1+\rho(i)*i)*e^{-\rho(i)*i} \quad \forall i=1 \ldots n \tag{19}$$

This equality holds only when ρ in Eq. (16) is replaced by the variable $\rho(i)$. In Eq. (19), $\rho(i)$ does not have 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 in Example 2, $E[m]=64.3$, $E[a]=0.304$, and $E[b]=0.53$, are used in Eq. (19). In Table 3, the numerical values of the variable fault detection rates $\rho(i)$ that satisfy the equality in test instance i in Eq. (19) are given. The Delayed S-shaped growth curve in Ohba [9] can calculate the same growth curve as obtained by the S-shaped HGD Model only for the variable $\rho(i)$.

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 shown together with the failure intensity function for the Delayed S-shaped growth model where $E[m]=64.3$ and the 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 ρ cannot calculate the better-fitting growth curve of the HGD Model.

4.2 S-shaped HGD Model and Goodness of Fit

The S-shaped HGD model significantly increases the goodness of fit of the estimated growth curve to the real observed growth curve. Table 4 shows the optimal estimated parameter values as well as the goodness of fit values for the optimal estimated growth curves represented in Example 2. The optimal constant fault detection rate ρ for an estimated initial number of faults $E[m]=64.3$ (the same value as estimated by the HGD Model) is also listed for the Delayed S-shaped growth model. The S-shaped HGD Model evaluates the best-fitting estimated growth curve. With a constant fault detection rate ρ, the Delayed S-shaped model is incapable of calculating the better-fitting growth curve of the S-shaped HGD Model.

4.3 Conclusion for S-shaped HGD Model

The HGD Model with $w(i)=a*i+b$, $a>0$ and $b\geq 0$, has a variable fault detection rate $\rho(i)$, whereas the Delayed S-shaped Growth Model does not. The usage of $\rho(i)$ 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 of software reliability modeling, as realized by the HGD Model, can be seen as a great improvement on previous growth models that consider only constant fault detection rates. A variable fault detection rate also seems more realistic in an actual test-and-debug environment.

5. Applicability to Different Kinds of Data

In this section, three data sets of real observed data will be analyzed. The values of the parameters estimated by the HGD Model are compared with those estimated by other models.

Table 4 Comparison of Goodness of Fit.

Model	$E[m]$	Parameter Values	EF1-value	Kolm.-Smirnov Value
HGD Model, EF1	64.3	$w(i)=0.304*i+0.53$	0.837071	0.094860
D-S-shaped, EF1	76.98	$\rho=0.09624$	0.988674	0.099198
D-S-shaped, EF1; with $E[m]=64.3$ assumed	64.3	$\rho=0.116990$	1.285424	0.135735
[9]	71.7248	$\rho=0.103967$	1.042750	0.114289

Table 5 Results of Estimations for 111 Test Instances.

Model	$E[m]$	Parameter Values	EF1-value	Kolm.-Smirnov Value
HGDM, EF1	475.9	$w(i)=0.663*i+3.74$	10.208244	0.125941
Del. S-shaped	483.04	$\rho=0.068653$	12.491662	0.133920
G-O NHPP	497.29	$\phi=0.030796$	25.641797	0.195207
HGDM, EF1 [5]	484	$w(i)=testworker(i)*$ $(0.082*i+1.36)$	9.825453	0.111240
Del. S-shaped, EF1	481.8	$\rho=0.071568$	12.369654	0.153361
G-O NHPP, EF1	527.8	$\phi=0.02646$	23.150693	0.175228

5.1 Example 3, Monitoring and Real-Time Control Software

A data set presented in [11] is used. The data were collected during the test-and-debug of a monitoring and real-time control software package consisting of about 200 modules, each about 1000 lines of code written in FORTRAN. The test conditions are not known. The only given facts are the number of test workers involved in the test and the number of newly detected faults on a day-by-day basis for 111 debugging days.

In our paper [5], we used $w(i)$ as the following "ease of test" function:

$$w(i) = testworker(i) * (a * i + b)$$

For the estimations here, however, we define $w(i)$ as a linear function of $w(i) = a * i + b$. Table 5 shows the results of estimations for the different parameters as well as the goodness of fit values obtained by EF1 and the Kolmogorov-Smirnov Goodness of Fit Test. The estimated growth curves are plotted in Fig. 6.

The Goel-Okumoto NHPP Model is inappropriate for this data, 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 best fit the actually observed data. This better fit of the estimated growth curve to the actually observed growth

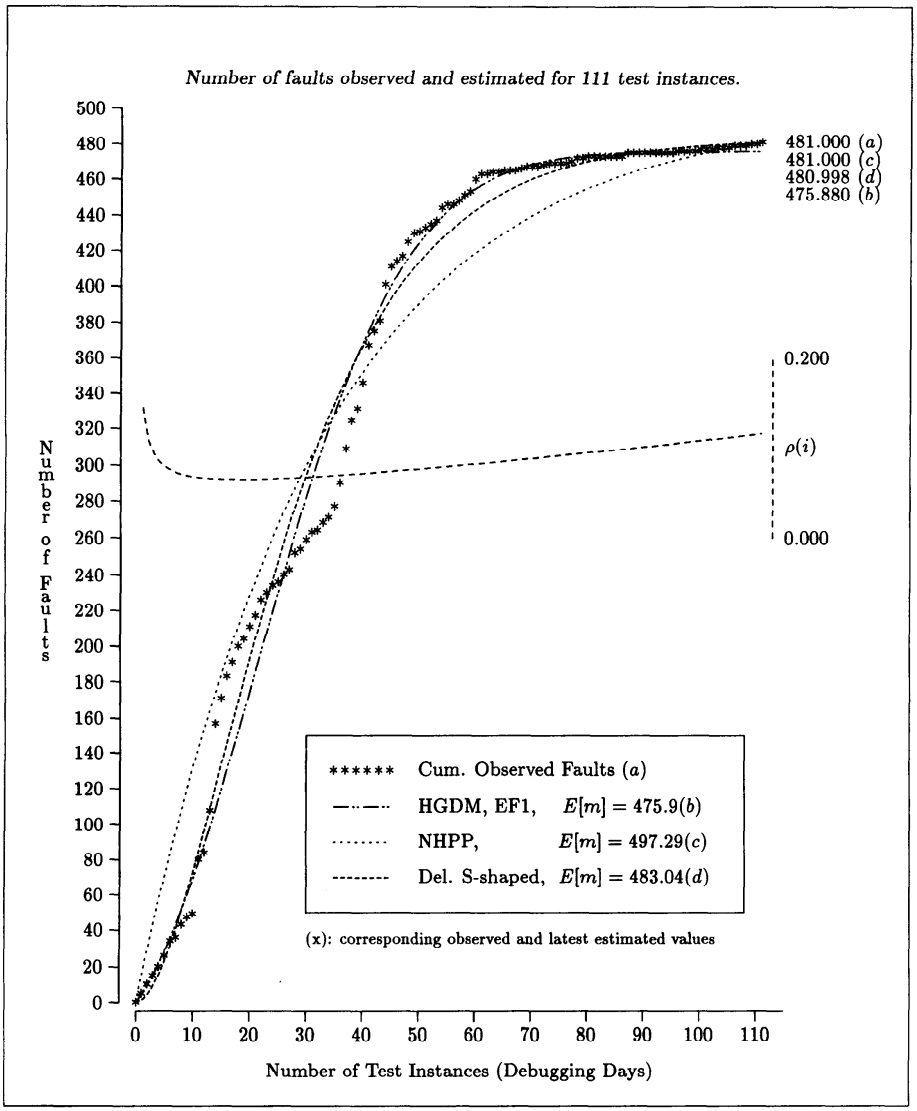


Fig. 6 Estimated Growth Curves for Monitoring and Real-Time Control Software.

curve is realized by the assumption of the variable fault detection rate $\rho(i)$ in our model. The application of NHPP models in combination with the evaluation function EF1 in Eq. (12) also gives better-fitting estimated growth curves than those obtained by the maximum likelihood method.

5.2 Example 4, Railway Interlocking System

The data analyzed here was given in our previous paper [11]. It is the bug report from a program consisting of about 14.5 KLOC of ASSEMBLER language for a railway interlocking system. Only the number of

Table 6 Results of Estimations for 199 Test Instances.

Model	$E[m]$	Parameter Values	EF1-value	Kolm.-Smirnov
HGDM, EF1	65.6	$w(i)=0.0069*i+0.0000$	2.037733	0.128216
Del. S-shaped	76.0672	$\rho=0.01282$	2.517038	0.127030
Del. S-shaped, EF1	97.0	$\rho=0.01034$	2.114081	0.108003

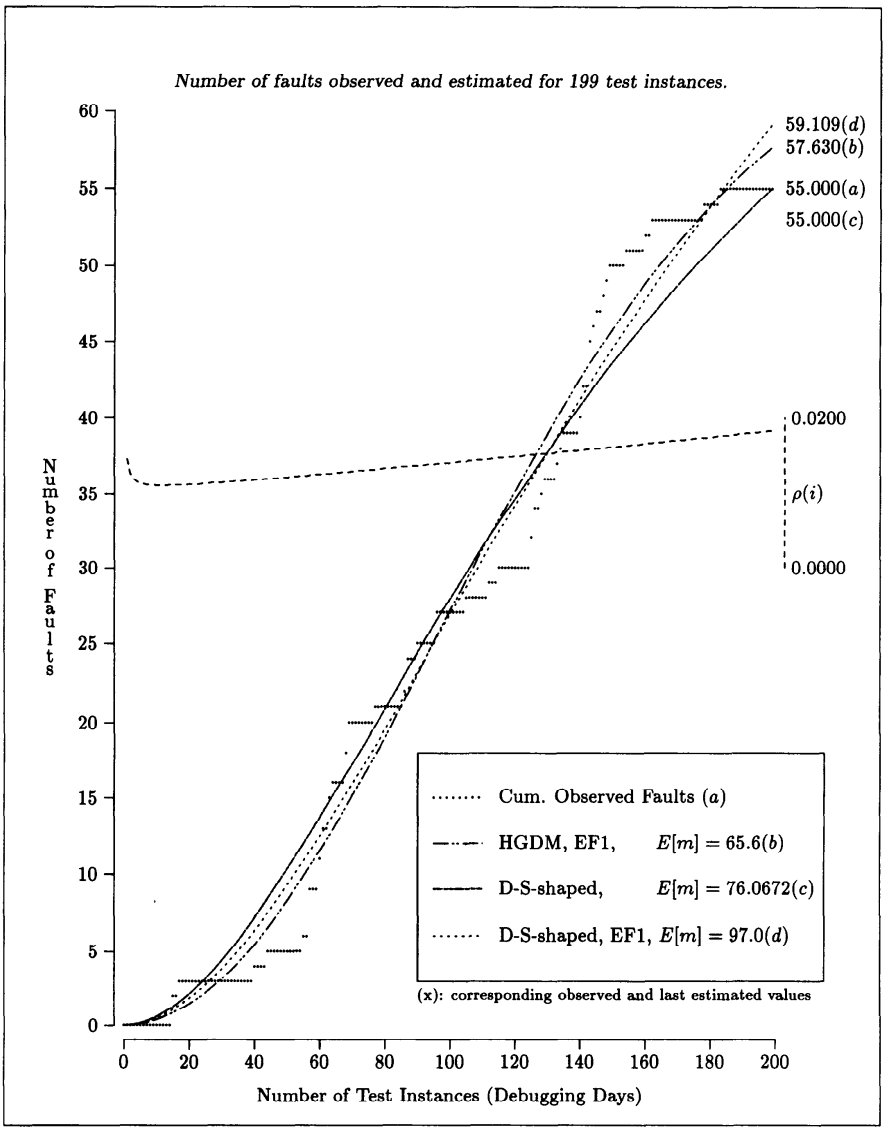


Fig. 7 Estimated Growth Curves for Railway Interlocking System Software.

newly detected faults was reported. Therefore, the results for the HGD Model, with $w(i)=a*i+b$, are compared with those for the Delayed S-shaped growth model.

The numerical results for the estimation are given in Table 6 and the respective estimated growth curves, as well as the variable fault detection $\rho(i)$-curve, are plotted in Fig. 7.

The estimated growth curve of the HGD Model (HGDM, EF1) best fits the actually observed data, owing to the variable fault detection rate. In the final stage of estimation, this estimated growth curve bends earlier than the other estimated growth curves and therefore calculates $E[m]=65.6$ initial faults, whereas the Delayed S-shaped growth model estimates 76.0672 faults by maximum likelihood method. The growth curves estimated by *EF*1 fit these data much better than those obtained by the maximum likelihood method.

5.3 Example 5, PL/I Application Program Test Data

The following are test data for a PL/I database application program [9]. The program is of about 1.317 KLOC. The data reported for a period of 19 weeks. The execution times and the number of faults detected per week are reported. The total number of observed failures is 358. The results of estimations for the HGD

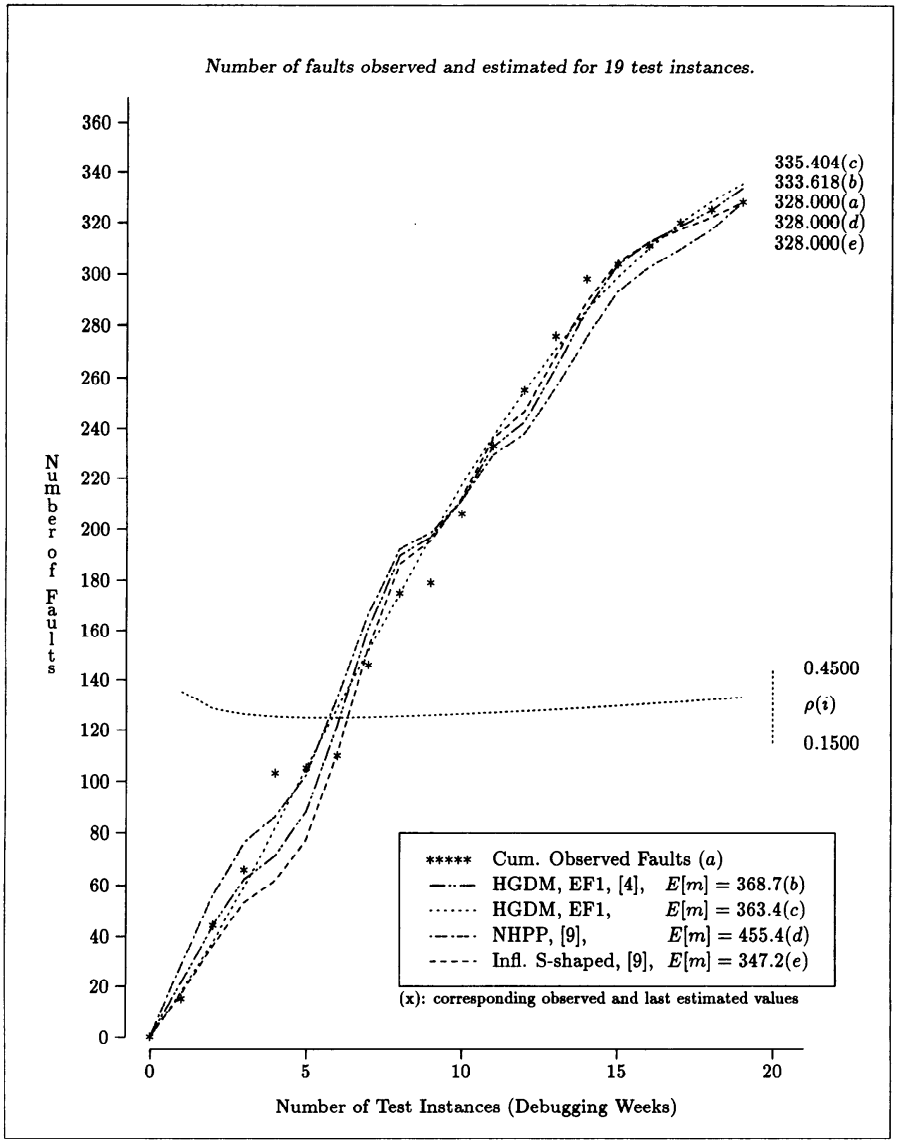


Fig. 8 Estimated Growth Curves for PL/I Database Software.

Table 7 Results of Estimations for 19 Test Instances (Weeks).

Model	$E[m]$	Parameter Values	EF1-value	Kolm.-Smirnov Value
HGDM, EF1, [4]	368.7	$w(i)$=exec.-time(i)* (0.86*i+7.97)	8.887544	0.156170
HGDM, EF1	363.4	$w(i)=3.07*i+14.72$	6.793139	0.127944
Exponential G-O, [9]	455.4	$\phi=0.0267368$	12.672613	0.174075
Inflection S-shaped, [9]	347.2	$\phi=0.0935493$, $r=0.2$	8.898889	0.131950
Exponential G-O, EF1	859.6	$\phi=0.0274$	9.818134	0.122025
Delayed S-shaped, EF1	385.2	$\rho=0.1845$	9.640437	0.116258
Totally Observed	358			

Model are compared with those obtained by Ohba [9].

In Table 7, the numerical results of estimations are given. For a comparison of the goodness of fit, the numerical results of estimation for the Delayed S-shaped model and the NHPP Goel-Okumoto model, in combination with the evaluation function, Eq. (11), are also given.

The estimated growth curve of the HGD Model and those of the Inflection S-shaped growth model and the NHPP Goel-Okumoto model are shown in Fig. 8. Here, to keep the representation of data consistent, the estimated growth curves in Ohba [9] for the inflection growth model and the NHPP Goel-Okumoto growth model are plotted for test instances i rather than for cumulative execution times t.

For these data also, the growth curves estimated by the HGD Model better fit the actually observed data. Furthermore, the estimates for $E[m]$ of the HGDM are the closest to the actually observed total number of 358 faults detected.

From these data, it is difficult to judge which of the NHPP models estimates a growth curve closest to the actually observed data. With the application of the HGDM, we do not need to worry about the probable shape of the observed growth curve. This is very useful, because it allows the HGDM to make estimations for all kinds of observed growth curves. The parameters of the model are responsible for a better fitting to the actually observed data.

6. Conclusion

In this paper, we have presented the basic concepts of the Hyper-Geometric Distribution Growth Model for the estimation of the number of faults at the beginning of the test-and-debug phase. The exact formulation of the model is established. To evaluate the optimal parameter values of our model, we use some distance function. This approach is different from the idea of applying the maximum likelihood method to estimate the best-fitting growth curves.

The exact formulation of the HGD Model makes it easy to compare with other models. We presented the relationship of the HGD Model to the Goel-Okumoto NHPP Growth Model and the Delayed S-shaped Growth Model. The Goel-Okumoto NHPP model can be regarded a special case of the HGD Model. The relationship to the Delayed S-shaped growth model is given, but the estimated growth curves obtained by the HGD Model correspond more closely to the actually observed data than those obtained by the Delayed S-shaped model. We introduced the concept of a variable fault detection rate. This variable fault detection rate increases the goodness of fit for growth curves estimated by our model.

A very important property of the HGD Model is that it can be applied to various kinds of data, as we have shown in this paper. A single model can be used to estimate exponential growth as well as S-shaped growth. This means, that using this overall model, we do not need to worry about which model is to be applied to which actually observed data. This is very advantageous in reliability estimations based on reliability growth curves.

Future research will attempt to eliminate some of the unrealistic assumptions common to many software reliability growth models. One major point of interest is the introduction of new faults during the fault elimination process. We are also conducting research on a more general "ease of test" function for $w(i)$ that does not have a linear property, and on the establishment of some analytical method for determining the parameter values of our model.

References

1. FRY, T. C. Probability and its Engineering Uses (2nd Ed.), pp. 205, Van Nostrand Co. Ltd., 1965.

2. GOEL, A. L. and OKUMOTO, K. Time-Dependent Error-Detection Rate Model for Software Reliability and Other Performance Measures, *IEEE Transactions on Reliability*, R-**28**, 3 (August 1979), 206–211.

3. GOEL, A. L. Software Reliability Models: Assumptions, Limitations, and Applicability, *IEEE Trans. Softw. Eng.*, SE-**11**, 12 (December 1985).

4. JACOBY, R. and TOHMA, Y. Notes on the Application of the Hyper-Geometric Distribution to Estimate the Total Number of Faults Initially Resident in a Software Under Test, *IECE Technical Report*, FTS**88**-9 (May 1988), 51–58.

5. JACOBY, R. and TOHMA, Y. Hyper-Geometric Distribution Estimation Model. From Precise Formulation to Mutual Relationship With Other Models. The S-shaped HDG Model, 10th Software Reliability Symposium, Osaka Garden Palace (October 1989), 24–25.

6. MURATA, Y. and TOHMA, Y. A Model for Estimating the Number of Software Faults Considering the Progress of Test, *IECE Technical Report*, FTS*86*-31 (February 1987).
7. MUSA, J. D. et al. Software Reliability. Measurement, Prediction, Application, McGraw-Hill International Edition, 1988.
8. OHBA, M., YAMADA, S., TAKEDA, K. and OSAKI, S. S-shaped Software Growth Curve: How good is it?, Proc. COMPSAC 1982, Chicago (1982), 38-44.
9. OHBA, M. Software Reliability Analysis Models, IBM Journal of Research and Development, **28**, 4 (July 1984), 428-443.
10. TOHMA, Y., NAGASE, S. and MURATA, Y. Structural Approach to Software Reliability Growth Model based on the Hyper-Geometric Distribution, *IECE Technical Report*, FTS86-23 (November 1986), 47-55.
11. TOHMA, Y., TOKUNAGA, K., NAGASE, S. and MURATA, Y. Structural Approach to the Estimation of the Number of Residual Software Faults Based on the Hyper-Geometric Distribution, *IEEE Trans. Softw. Eng.*, **15**, 3 (March 1989), 345-355.
12. TOHMA, Y., JACOBY, R., MURATA, Y. and YAMAMOTO, M. Hyper-Geometric Distribution Model to Estimate the Number of Residual Software Faults, COMPSAC89, Orlando, Florida (September 1989), 610-617.
13. YAMADA, S. Software Reliability Estimation Technique, Software Reliability Growth Model Introduction, HBJ Integrated Libraries, No. 42 (May 1989) (in Japanese).

(Received February 2, 1990)