A Systematic Mapping Study on Software Reliability Growth Models that Consider Evolution

L. Alhazzaa\(^1\) and A. Amschler Andrews\(^2\)

\(^1\)Computer Science Department, Al-Imam Muhammad Bin Saud Islamic University, Riyadh, Saudi Arabia
\(^2\)Department of Computer Science, University of Denver, Denver, CO, USA

Abstract—Software Reliability Growth Models (SRGMs) are used to assess and predict reliability of a software system. Many of these models are effective in predicting future failures unless the software evolves. Our objective is to find and compare solutions in literature for reliability prediction for evolving systems. We conduct this mapping study by collecting and analyzing publications in the field to identify available solutions, gaps and possible future work. We selected 54 papers from 1986 to 2018 and summarized trends with respect to year, research organizations, publication venues and academia versus industry. Papers are classified according to solution extent, proposed method, and research type.

Keywords: Software Reliability, evolution, change-point, failure prediction, systematic mapping study, legacy systems.

1. Introduction

Software reliability is one of the key aspects when evaluating software quality. Software Reliability Growth Models (SRGM) are models used to describe the behavior of software in terms of cumulative number of failures over time, in order to estimate and predict the reliability of that software product. The use of SRGM has been effective in many systems but the predictions are less accurate in legacy systems that undergo periodic changes due to corrective, perfective or adaptive maintenance and enhancements. Legacy systems can be vital to an organization.

Use of SRGMs started in 1972 with the work of Jelinski and Moranda [1]. Some of the more common models are the Musa model [2], delayed S-shaped model [3], and the Yamada exponential model [4]. Many studies and empirical studies were performed to apply those models, either using analytical methods or curve fit methods. Stringfellow and Andrews in [5] presented an empirical method for selecting SRGMs to make release decisions. They provided some of the few studies which use a curve-fit approach to evaluate reliability models. Curve-fit methods perform a regression and evaluate applicability of a set of candidate models with little or no assumptions about operational profile, defect fixes, etc.

Our goal is to study emerging solutions that deal with change and evolution in legacy systems. Our objective is to map the body of research in the area and compare solutions. We explore publications in industry and academia, find research affiliations and identify contributions through different venues. We then classify the papers according to solution extent, proposed method and research type, and finally discuss focus and gaps.

The remainder of the paper is laid out as follows: we begin by first describing the research method in Section 2, followed by the study classification scheme in Section 3. This scheme is applied to the selected papers to present the actual mapping of the field in Section 4. Section 5 covers threats to validity. Conclusive remarks are in Section 6.

2. Research method

Following the main guidelines in [6], we propose our systematic mapping going through the following four steps: Definition of research questions, identification of search string and source selection, study selection criteria and data mapping. Each step is explained in greater detail in the following subsections.

2.1 Definition of research questions

To meet our objective, we ask the following research questions:

RQ1: What are the publication trends for reliability models used during change and evolution?
- RQ1.1: What is the annual number of publications?
- RQ1.2: What are the main venues of publication?
- RQ1.3: Which publications are affiliated with academia versus industry? What institutes are the most active in the field?

RQ2: What are the existing research gaps related to SRGMs that can handle evolution and changes in failure prediction?
- RQ2.1: What solutions have been proposed?
- RQ2.2: What methods are used to apply those solutions?
- RQ2.3: What types of research are conducted in the area?

By looking first into research trends we can explore emerging and abandoned trends, the progression of research activity during a specific time span or through a specific research group. The second set of questions covers the most important solutions and approaches presented along
with the gaps and underrepresented methods. In addition, we analyze their contributions and their research method.

2.2 Identification of search string and source selection

In order to find reliable peer reviewed articles we selected articles from the following sources: IEEE, ACM Digital Library, Elsevier, Springer and Wiley. We found evolution and change to be referred to using different phrases such as change-point, multi-up gradation, or multiple change-points. The search string was derived as follows:

\[(\text{Failure prediction OR failure estimation}) \text{ AND (software reliability models OR SRGM OR software reliability growth models}) \text{ AND (evolution OR change OR change-point OR multiple change-points OR multi up-gradations)}\]

After performing the search on Google Scholar, Web of Knowledge and the search engines of the libraries indicated above, papers were selected according to the study selection process explained in the following section.

2.3 Study selection criteria

To select relevant studies we defined inclusion and exclusion criteria as shown in Table 1. The study selection followed the steps demonstrated in Fig. 1. The total number of papers retrieved from the search is 1,786 papers from 1986 to 2018. After reading titles and using the snowballing technique to find further papers, the remaining number of papers was 434. Next, we read the abstracts and kept 229 papers. The authors of this mapping study had to vote on which papers are relevant and which papers are excluded according to our inclusion/exclusion criteria. In addition, some subjects were found irrelevant to our objective. Since we were concerned with finding SRGM solutions for legacy systems, we found that reliability models that are concerned with imperfect debugging is not in our interest. There were 128 papers left for full text skimming. This left 54 relevant papers for our mapping study.

2.4 Data mapping

The 54 papers chosen covered the time period 1986 to 2018. The choice of the start date was based on the fact that most common SRGM’s were introduced between 1983-1986. We performed data mapping using three categorizations: Solution extent, proposed method, and research type.

Each relevant paper is reviewed by two reviewers and each reviewer suggests the proper categorization of the paper. If they both agree on the categorization of the paper, then the paper is assigned to the agreed upon category; if there is no agreement, additional review is performed by skimming the text. In some cases, skimming text resulted in the paper being declared not relevant, and therefore it was discarded.

3. Study classification scheme

The study classification is divided into two main categories. The first category is related to the publication trends which highlights the years of publications, the venues, active research organizations and affiliations. These trends answer the first set of research questions in Section 2.1. The second category is concerned more with the research contents and possible research gaps which are: Solution Extent, Proposed Method, and Research Type. Detailed descriptions are given below.

3.1 Solution extent

Solution Extent discusses what solution is proposed according to the type and quantity of change: single change-point, multiple change-points and multi up gradation. Single
change-point refers to the sudden change that happens to the cumulative failure rate as a result of code addition, deletion, or modification. This change affects the estimation of failures after that point. Multiple change-points provides a solution to a reliability model that is aimed at systems with multiple change-points in the cumulative failure curve. Multi up-gradation is dedicated towards systems that are regularly upgraded.

3.2 Proposed method

The proposed method refers to the method of selecting the proper reliability model for the proposed solution. They can be of two types: analytical and curve-fit. Analytical methods and which derive a solution analytically by providing assumptions for the proposed model or method and developing a model based on these assumptions. Alternatively curve-fit methods make little or no assumptions. They entirely rely on statistical curve-fit methods for one or more types of functions and select the most accurate one.

3.3 Research type

Each research paper has a specific contribution and each contributed solution could be evaluated. Specifying the research type provides more understanding of the maturity of the proposed work. If a method proposed is validated or evaluated empirically, this gives more confidence in the provided solution. Research type would fall under one of these two major categories: empirical or non-empirical. Empirical research provides direct or indirect observation of an experience. It demonstrates an application of a given method or model through a case study or an experiment. Non-empirical research on the other hand is more theoretical that provides information without empirical data. In this category a researcher relies on interpretation or observation to come to a conclusion.

For the empirical type of research, the work is evaluated according to some major research aspects based on empirical study evaluation criteria proposed by [7]. We ask the following questions:

- Are the objectives or hypothesis stated?
- Is the targeted application domain specified?
- Are the subject system specifications provided?
- Are there any details about how model parameter are set?
- Do they measure effectiveness of the solution?
- What is their baseline for comparison?
- Is there data analysis of the findings?
- Do they mention any threats to validity?
- How good is the quality of data used?

4. Analysis of the results

In this section we provide our synthesis of the relative studies we collected. In the following section we present the trends and the potential gaps among the 54 publications of the time period covering 1986-2018.

4.1 Publication trends between 1986 and 2018

![Fig. 2: Annual distribution of publications](image)

Publications trends are analyzed in terms of distribution over the span of years covered. The main venues of publications are then highlighted, as well as the contributions of academia and industry in the field of study. We then look into the most active research organizations either in academia or industry.

4.1.1 Distribution of publications

Publication distributions from 1986-2018 have increased in general. The first spike in terms of number of publications was in 2005, then the interest grew higher in 2010, 2011 and 2017, as shown in Fig. 2. The first increase in number of publication began on 2005, Huang C.Y. and a research group at National Tsing Hue University worked together to investigate several aspects of software reliability modeling with the existence of change-points ([8][9][10][11][12][13][14][15][16][17]). In 2010, the term multi up-gradation was introduced by Kapur et al.[18]. From that year on, Kapur as well as other members of their research group published a number of publications about reliability modeling in multi up-gradated systems. In 2017, an increased interest on the subject of reliability modeling for Open Source Systems (OSS) arose ([19][20][21][22]).

4.1.2 Main publication venues

The top five venues with publications related to the subject of interest are ranked in Table 3. Most publications are published in IEEE journals and conferences, 63%. Springer Verlag comes next with 24% of the publications. Elsevier has 9%, then ACM and Wiley each contributed to 2% of the publications.
4.1.3 Academia and industry representation

Both academia and industry are interested in finding solutions. 80% of research is proposed by academia alone. 13% is a product of industry, and 7% is a joint effort between academia and industry. Fig. 2 shows the distribution of contributed papers over the years. A total of 85 research organizations investigated reliability with change and evolution in 16 different countries. The research organizations with the highest number of publications are listed in Table 2.

4.2 Focus and potential gaps

The following sections demonstrate the gaps and focus of the existing body of work in terms of the solution extent discussed, proposed methods, and research type and quality.

4.2.1 Solution extent

Table 3: Top publications venues

<table>
<thead>
<tr>
<th>Rank</th>
<th>Type</th>
<th>Venue</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Journal</td>
<td>International Journal of System Assurance Engineering and Management</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>Journal</td>
<td>IEEE Transactions on Reliability</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Journal</td>
<td>IEEE Transactions on Software Engineering</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Journal</td>
<td>Journal of Systems and Software</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Conference</td>
<td>2013 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)</td>
<td>2</td>
</tr>
</tbody>
</table>

Collected articles fall into one of the three solution extents highlighted in Section 3.1: Single change-point, multiple change-points, multi up-gradation. We find that 43% of the publications focus on solutions for multiple change-points, since it provides more generic and applicable solutions than just single change-point solutions, which contribute 24% to the body of work. Multi up-gradation gained a great deal of interest and represents 33% of the papers in the field.

Some solutions incorporate an additional aspect or dimension to the model proposed, such as testing effort or code complexity. Some add an additional dimension to have two dimensional models such as using testing effort factor and fault reduction factor in Kapur et al.[33]. Table 4, classifies the research in the filed based on solution extent and additional modification to the model so it would consider contributing factors in order to better estimate models and predict future failures. More explanation regarding the proposed solutions are explained in Section 4.2.4.

4.2.2 Proposed method

In Section 3.2, We defined the two types of proposed methods, (1) analytical method and (2) curve-fit method. Analytical methods represent 94% of the research contribution, while curve-fit methods represent 6%. In the analytical methods presented, they would identify change-points and divide the failure curve into segments and then apply the proposed model with different parameters to fit each segment accordingly ([52] [41] [49]). These efforts are valuable in the study but they are very limited. They merely adjust parameters, but they do not consider the combined effect of the failure process in old and new systems.

4.2.3 Research type

According to our discussion in Section 3.3, research type is divided into either empirical or non-empirical research. 94% of the papers represent empirical research, and only 6% are non-empirical. We noticed that the empirical work we found are mostly case studies, and there was a notable lack of controlled experiments. This finding is to be expected, since one cannot expect to produce multiple versions of evolving software simultaneously. This would simply not be financially viable. We evaluated the 51 empirical studies, we evaluated their quality as discussed in Section 3.3, as shown in Table 5. The objective or the hypothesis of the case study is stated in all empirical studies. 44 of the studies explain the system specifications and parameter setting, which makes the case study easier

Table 4: Solution Extent

<table>
<thead>
<tr>
<th>Solution Extent</th>
<th>Single CP</th>
<th>Multiple CP</th>
<th>Multi UG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fault Severity</td>
<td>[29] [10] [22] [30]</td>
<td>[23]</td>
<td>[24] [25] [26] [27] [28]</td>
</tr>
<tr>
<td>Testing Effort</td>
<td>[17] [9] [31] [13]</td>
<td>[12] [15]</td>
<td>[32] [33] [34]</td>
</tr>
<tr>
<td>Environmental &amp; learning effects</td>
<td>[35] [36]</td>
<td></td>
<td>[26] [37]</td>
</tr>
<tr>
<td>TCF</td>
<td>[12] [15]</td>
<td></td>
<td>[33] [19]</td>
</tr>
<tr>
<td>FRF</td>
<td></td>
<td></td>
<td>[21]</td>
</tr>
<tr>
<td>Code complexity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing coverage</td>
<td>[30]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Queueing model</td>
<td>[14]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hazard Rate</td>
<td>[38] [39] [40]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted means</td>
<td>[16] [8] [11]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>[41] [42] [43]</td>
<td>[44] [45] [46]</td>
<td>[56] [57] [18] [58] [20] [59] [60] [61]</td>
</tr>
</tbody>
</table>
to understand. The target application domain is specified in only 28 studies. We may assume that the other studies are generic and work for a wide spectrum of software systems. Most studies, 47 of them, would measure the goodness-of-fit of a proposed model but very few measure how well a model can predict future failures, only 13 papers did so. In terms of data analysis, the majority of the available work presented a decent description of descriptive statistics, results of the case study, and actual significance of the results. Some papers had a minimal description. In many cases this occurred in a short version of the case study in a conference paper which was followed with a more detailed version in a journal paper. All papers provided different variations of descriptive statistics. 47 out of the 51 papers mentioned the statistical significance of the results. On the other hand, many of the case studies are lacking any mentioning of threats to validity, in fact out of the three studies, only one explicitly specifies the threats and how they overcome them. Finally, data in the studies available are of a decent quality, i.e. they all come from reliable sources (benchmarks) and no major problems are mentioned regarding the quality of the data collected. The only problem with some publications is that they use old data that may not reflect the accuracy of the model for current software technology. A study published in 2007 using data from a source published in 1980, may not provide realistic outcomes, since techniques evolve throughout the years causing changes in failure processes modeled. We found that 17 papers use relatively new data.

4.2.4 Focus

Looking at the progression of research over the years, we found that there is an increased interest in the subject for both academia and industry, but academia has the most contributions. We also find that some of the top organizations conduct cohesive research in specific subjects. This is made clear when looking into the progress of research timeline (Fig. 3). Early efforts focused on dividing datasets into smaller sets and performing estimates on a smaller scale. The French National Center for Scientific Research sponsored research focused on trend analysis to detect change, dividing the failure dataset at the point of the change then applying SRGM to each partition ([55][43][54]). Between 2005 and 2014, National Tsing Hua University in Taiwan provided a body of work by C.Y. Huang, and C.T. Lin. Their focus was on providing a unified theory for SRGM by adding weights of means to the model. They also incorporated testing effort into their models and a Testing Compression Factor (TCF). TCF provides a ratio of change in fault detection between the testing phase and the operational phase ([8][9][10][11][12][13][14][15][16][17]). Between 2008 and 2014, Tottori University in Japan, Inoue and Yamada, conducted research regarding hazard rate modeling, proposing 2-dimensional models using time and effort and other environmental factors ([39][29][42][45][40][38]). Meanwhile, efforts were made to consider fault severity, learning effects and other environmental factors in the SRGM ([44][23][53][41][60][59][35][36]).

From 2010 until 2018, a remarkable research effort was conducted by researchers at the University of Delhi and Amity University in India in addition to members from Islamic Azad University, Iran and Rutgers, The State University of New Jersey, NJ, USA. They defined the term of multi up-gradation, referring to change due to updates in software systems. They incorporated several factors in their models such as fault severity, testing effort, environmental functions. They also applied their methods to Open Source Systems (OSS) ([57][18][25][30][37][27][24][56][26][28][58][32][33][61][51][19][22][21]). Most of the efforts were focused to analytical solutions and enhancement on previously presented models. Many of the models were evaluated using Goodness-of-Fit in order to find the best fit model.

4.2.5 Gaps

Legacy systems are valuable systems that are costly to maintain but cost more to replace. Having a reliable model to predict the effect of change on a software system’s reliability is crucial. With available work focused on analytical methods, it is important to get a better understanding of how robust they are when specific assumptions they make are violated. For example, for a model that assumes perfect debugging, what degree of imperfect debugging will not cause inaccurate results? While some of this information exists for systems without change-points, engineering guidelines still have to be developed for evolving and legacy systems. We believe that more effort should be made to use curve-fitting methods for evolving systems as they do not require assumptions. In addition, Most of the work focus on measuring how fit is the proposed model. While predictive ability of a model briefly discussed. We would like to see how long far can a model predict into the future and how accurate these predictions are? How long can the model sustain and when do I need to use a different model or
update its parameters? Moreover, we find that there is a need of high quality empirical work that uses current data and meets all aspects of case study methodology provided by [62].

While academia has the major contributions in the field, an industrial point of view or more collaborative work between academia and industry would be a rich contribution. These collaborative efforts and recommendations will provide decision makers with better tools in how to make their systems evolve in a healthy and predictable manner. This is vital information for feasibility studies in bidding on projects and decision making in order to decide to improve a current software system for an organization or replace it.

We also found that simple single-change point or multiple change-point techniques are of a major interest, and discussed in a broad spectrum of organizations. Multi upgradation was also proposed by limited research groups in specific organization, in fact the term, multi up-gradation, is exclusively used by them. Overall, looking into the progression of the body of work, it shows that this field is gaining interest through the years and the quality of proposed work keeps improving.

5. Threats to Validity

The major threat is the possibility of missing important papers due to several restrictions explained in our inclusion and exclusion criteria in Section 2.3. Excluding papers that are written in languages other than English or articles that are not easily available (e.g. confidential to a company), put this study at risk of leaving out valuable papers. Also excluding papers in fields other than software engineering which may exclude the efforts performed in this area that could potentially be applied to software systems. Having papers from other fields would make it more difficult to categorize and focus our efforts in producing a clear mapping study.

The choice of keywords and search strings was made systematically, but if some keywords or terms were missing this would have prevented us from reaching some sources. Using an additional search technique, snowballing, gives further assurance that the maximum number of studies get recognized and included. There is a possibility also of excluding a valuable paper in the review by excluding a paper before skimming. This would be possible if the title or abstract was misleading or incomplete, or by voting to exclude the paper by the group of reviewers before fully skimming the paper. On the other hand, having more than one reviewer reduces the chance of discarding papers that are relevant.

6. Conclusion

As the demand increases on businesses to grow, legacy systems are expected to continue to evolve accordingly. Evolution, updates, and changes get more complex; so it is more likely to find a sudden increase in failure intensity.
The challenge then is to find a reliable method to predict system reliability and estimate future failures.

The main objective of this study is to obtain a holistic view of the existing studies in designing, validating and evaluating reliability models for evolving legacy software systems. Throughout the mapping study we identified 54 papers covering a spectrum of approaches of reliability models. These approaches are different in solution extent, techniques, and methods. In proposed methods, papers fell into one of the two main categories, analytical or curve-fit. We found that there is a high focus in the field of analytical methods in software reliability model with evolution, while curve-fit methods are limited although recent work has been published, ([49] [52][41]). Solution extent covered single-change point evolution, multiple change-point evolution or multi-up-gradation. There is an increase of interest in multi-up-gradation. The research in the field provides empirical work. A large proportion of case studies were of a high or a decent quality but some used low quality data, which affected the strength of their outcomes. We then conclude our work with some recommended areas for potential researchers to investigate: (1) Looking into more curve-fit solutions, (2) providing better quality empirical studies, (3) greater involvement of industry.

References


