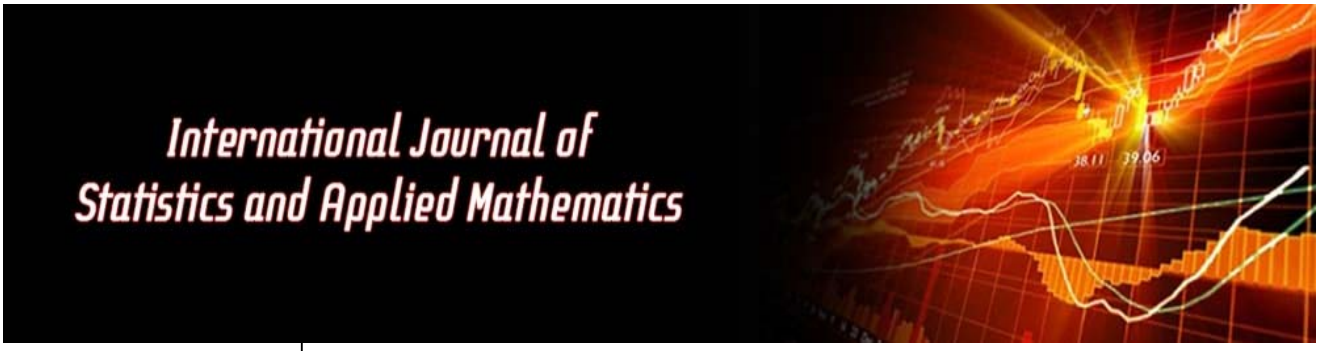
International Journal of Statistics and Applied Mathematics 2020; 5(4): 100-102



ISSN: 2456-1452
Maths 2020; 5(4): 100-102
© 2020 Stats & Maths
www.mathsjournal.com
Received: 16-05-2020
Accepted: 20-06-2020

R Jeromia Muthuraj
Assistant Professor,
Department of Statistics,
DD and CE, Manonmaniam Sundaranar University
Abishekapatti, Tirunelveli,
Tamil Nadu, India

A Mohamed Ashik
Assistant Professor in Statistics,
Department of Mathematics,
Merit Arts and Science College
Idaikal, Tirunelveli, Tamil Nadu, India

SM Karthik
Assistant Professor,
Department of Statistics,
DD and CE, Manonmaniam Sundaranar University
Abishekapatti, Tirunelveli,
Tamil Nadu, India

Corresponding Author:
R Jeromia Muthuraj
Assistant Professor,
Department of Statistics,
DD and CE, Manonmaniam Sundaranar University
Abishekapatti, Tirunelveli,
Tamil Nadu, India

Manipulating large-scale data in software reliability studies

R Jeromia Muthuraj, A Mohamed Ashik and SM Karthik

Abstract
Availability of information is abundant in this modern world due to the developments taking place in this ICT field. Drawing inference from such data requires the development of new methodologies. Dimensionality reduction with less amount of loss of information can be considered as one of the suitable solutions to this issue. This paper proposes a statistical methodology to reduce the dimension of the data with less amount of loss of information. The procedure was hierarchical Clustering method and extraction of principal components analysis. Applications of the methodology are illustrated with a data set.

Keywords: software reliability, principal component analysis, cluster analysis

1. Introduction
In a modern world, data might be represented in a spreadsheet, with one column representing each dimension. Dimensionality in statistics states that how many attributes a data set has. Naturally, healthcare records are notorious for having huge amounts of variables. High dimensional, that the number of dimensions is staggeringly high so calculations become extremely difficult. Dimensionality reduction is the process of reducing the dimensional of the feature data set with some acceptable level of loss of information. In the present world, data scientists observe large volumes of data, which is tough to process, information can be raw, unstructured, and high dimensional. Handling high dimensional data is hard, so the dimensionality reduction is essential to handling the data to analyze further. This paper proposes a methodology to reduce the dimensionality with less amount of loss of information in two different stages. Principal components analysis is used to reduce the dimensionality. The methodology is justified by the data which is collected to study the reliability of the software with Environmental factors. In software reliability studies, the reliability of the software studied through failure test data. Pham (2000) [6] and Patwa and Malviya (2014) [5] proposed Some environmental factors (EFs) to study the reliability of the software. Pham (2000) [6] proposed 32 EFs and Patwa and Malviya (2014) [5] used 26 EFs, from 26, 24 EFs are already used by Pham (2000) [6], Patwa and Malviya (2014) [5] introduced two EFs. Loganathan and Muthuraj (2016) [3] proposed that the 34 EFs are potential to study the reliability of the software.

2. Methodology
The Principal Components Analysis (PCA) is one of the classical methods of multivariate statistical data analysis. PCA is a dimensionality reduction technique, which is used to transform a high-dimensional dataset into a smaller-dimensional subspace. It is difficult to generate independent explanatory variables from these EFs. However, it is possible to extract uncorrelated explanatory variables from these EFs conducting principal components analysis (PCA). A methodology is used to reduce the volume of the data with less amount of loss of information with two stages. Cluster analysis subdivides a given set of individuals/objects into several disjoint subsets, called clusters, such that individuals /objects in the same cluster are more similar than individuals /objects in different clusters according to the same acceptable criterion of similarity. Clustering methods can be broadly categorized into two viz.,

hierarchical methods and non-hierarchical methods. Here 34 EFs has to be considered for the analysis, first, the factor is grouped into five clustered based on their similarities. Here it is proposed to perform principal component analysis in two stages. In the first stage, the principal components explaining 90% of the variation among the EFs in each cluster are extracted i.e., Intra-Cluster principal components extraction. Then, again principal components analysis is performed treating the intra-clusters principal components as variables. The Inter-Clusters principal components explaining 90% of the variation of the intra-cluster principal components are extracted and such principal components will be considered for further analysis.

3. Data Description

An investigation survey was conducted to studying the impact level of all the 34 EFs in view of the technical experts who are involved in developing and maintaining software. A questionnaire was designed to collect opinions of Managers, System Engineers, Software testers and Programmers. The respondents were given eight (8) options to express their opinion about the significance of each EF on determining software reliability. The opinions were scaled in 8-point scale as listed in Table 1.

Table 1: The opinion of Respondents and Score

S. No.	Opinion	Score
1	Extremely Significant	7
2	More Significant	6
3	Moderately Significant	5
4	Significant	4
5	May or May not be Significant	3
6	Less Significant	2
7	Moderately Insignificant	1
8	Not Significant	0

The questionnaire was sent to a randomly selected 85 persons who were working in 14 software development organizations. The respondents were communicated through e-mail attached with the questionnaire. The questionnaires were sent to the respondents along with Commercial Billing software for collecting their opinion on the significance of each EF with the aim of studying the reliability of this software. The total number of respondents is 85. Among them, 15% (13) of them were Managers, 29% (25) of them were System Engineers, another 18% (15) of them were testers and the remaining 38% (32) were Programmers. The information about the respondents is graphically presented in Fig. 1.

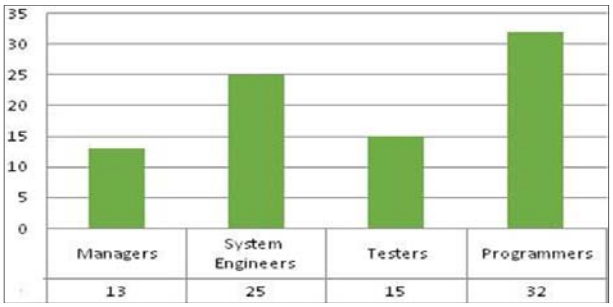


Fig 1: Data description

4. Results and Findings

4.1. Cluster Analysis*:* Opinions of the respondents on the significance of the EFs formed a new grouping of EFs. It may be noted that each cluster consists of EFs from different categories. It is interesting to note that the EFs in different phases of the software development process have similarity and the EFs in the same phase have dissimilarity. The agglomeration procedure is continued until all the EFs are put in a single cluster. The process of fusion is described in the dendrogram, which is displayed in Figure 2. A set of 6 clusters is selected and details of the cluster members are presented in Table 2.

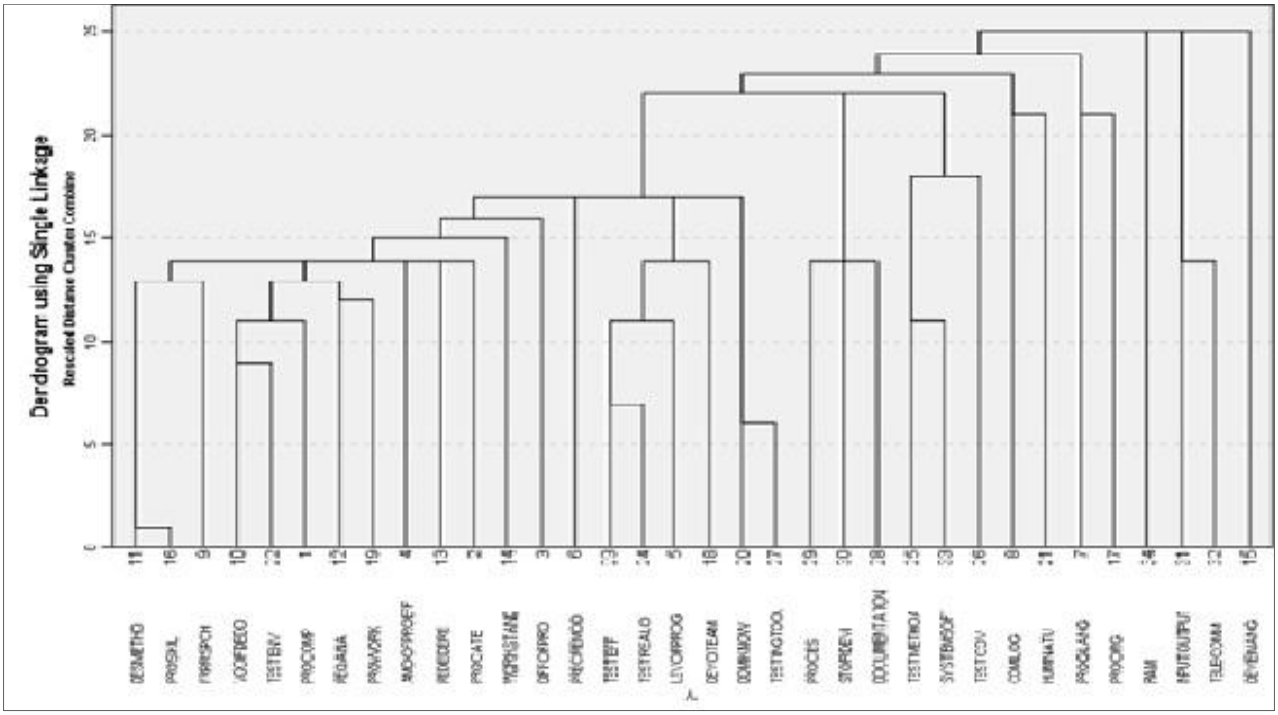


Fig 2: Dendrogram of cluster Formation

Table 2: Clusters of Environmental Factors

Cluster	No of EFS	Factors
Cluster 1	8	Design Methodology
		Programmer Skill
		Frequency of Program specification change
		Volume of Program design documents
		Testing Environment
		Program Complexity
		Requirements Analysis
		Program Workload (stress)
Cluster 2	6	Amount of Programming effort
		Relationship of detailed Design to Requirement
		Program Categories
		Work Standards
		Difficulty of Programming
		Percentage of Reused modules
Cluster 3	5	Testing Effort
		Testing Resource allocation
		Level of Programming technologies
		Development Team size
		Domain Knowledge
Cluster 4	6	Testing Tools
		Storage Devices
		Human Nature
		Documentation
		Testing Methodologies
		System Software
Cluster 5	4	Testing Coverage
		Complexity in Logic
		Processors
		Programming Language
Cluster 6	5	Programmer Organization
		Random Access Memory
		Input/ Output Devices
		Telecommunication Devices
		Development Management

4.2. Intra- Cluster Principal Components Extraction

It is proposed to perform principal component analysis in two stages. In the first stage, the principal components explaining 90% of the variation among the EFs in each cluster are extracted i.e., Intra-Cluster principal components extraction. Then, again principal components analysis is performed treating the intra-clusters principal components as variables.

Table 3: Number of Intra - Cluster Principal Components.

Cluster of EFs	No. of Principal Components
Cluster I	5
Cluster II	4
Cluster III	3
Cluster IV	4
Cluster V	3
Cluster VI	3
Total	22

4.3. Inter-Cluster Principal Components Extraction

It should be noted that the extracted principal components which are uncorrelated within the respective clusters only. Moreover, the principal components are a linear combination of the EFs in the respective groups. Under this circumstance, it should conduct PCA considering these 22 principal components as variables. A perusal of Table 4 the first 11 principal components explain about 91.732% of the total variance of the 22 intracluster principal components. Among them, the first two Inter Group principal components explain respectively 26.186% and 14.530% of the total variance. Also, the first six principal components explain, in total. 73.759% of the total variance. Inclusion of the principal components from 7 to 11 has added 17.973% of the total variance.

Table 4: Inter-Cluster Principal Components Extraction

Principal Components	Eigenvalues	% of Variance	Cumulative % of Variance
1	5.761	26.186	26.186
2	3.197	14.53	40.715
3	2.36	10.727	51.443
4	2.221	10.093	61.536
5	1.464	6.653	68.189
6	1.225	5.57	73.759
7	0.999	4.539	78.299
8	0.902	4.101	82.4
9	0.868	3.945	86.345
10	0.609	2.77	89.115
11	0.576	2.616	91.732
12	0.513	2.333	94.065
13	0.352	1.599	95.664
14	0.296	1.345	97.009
15	0.227	1.034	98.043
16	0.167	0.758	98.8
17	0.117	0.53	99.33
18	0.068	0.308	99.638
19	0.045	0.203	99.841
20	0.021	0.095	99.936
21	0.01	0.047	99.983
22	0.004	0.017	100

5. Conclusion

This study considered 34 potential environmental factors which are important to study the reliability of the software. It is proposed that, while handling large-scale data, the proposed methodology is appropriate for further analysis with less amount of loss of information and with less dimensions.

6. References

1. Garson GD. Cluster Analysis, Statistical Associate Publishers USA, 2012.
2. Jelinski Z, Moranda P. In Statistical Computer Performance Evaluation, in W. Freiberger, Ed. New York: Academic Press, 1972, 465-484.
3. Loganathan A, Muthuraj RJ Importance of Environmental Factors affecting Software Reliability, Global and Stochastic Analysis. 2016; 2:101-105.
4. Loganathan A, Muthuraj RJ. A new methodology for data reduction in software reliability studies, Communications in Statistics: Case Studies, Data Analysis and Applications. 2017; 4:119-125.
5. Patwa S, Malviya K. A Survey on Factors affecting Testing Techniques in Object-Oriented Software, International Journal of Applied Research on Information Technology and Computing. 2014; 5:78-85.
6. Pham H.: System Software Reliability, Springer Series in Reliability Engineering, 2000.
7. Zhang X, Pham H. An analysis of factors affecting software reliability. Journal of Systems and Software. 2000; 50(1):43-56.
8. Zhang X, Shin MY, Pham H. Exploratory analysis of environmental factors for enhancing the software reliability assessment, Journal of Systems and Software. 2001; 57:73-78.
9. Zhang X, Zhu MY, Pham H. A comparison analysis of environmental factors affecting software reliability, Journal of Systems and Software. 2015; 109:150-160.