International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment***, 21-24 December, 2012.

ISBN**: 978-984-33-5876-9**

Forecasting Failure Number Using Warranty Claims in Multiplicative Composite Scale

BenojirAhammed and Md. MesbahulAlam

Department of statistics, University of Rajshahi, Rajshahi-6205, Bangladesh
E-mail: benojir_11@yahoo.com, mesbah_ru@yahoo.com

Abstract. Depending on the technological development, forecasting of failure number is main aspect of a manufacturing company to the plan of services; predict the approximate warranty cost and customer satisfactions. Warranty data analysis is crucial for predicting the quantity of product that will be returned under warranty. In this paper we use multiplicative composite model based on age and usage for forecasting the failure number. For this purpose, to make the incomplete data due to censored usage as complete several approaches are proposed. Also, few existing approaches are considered. Incorporating this information, the failure claims under two-dimensional warranty scheme are predicated and compared with age based forecasting through a brief simulation, which shows that the proposed method on multiplicative composite scale can fairly forecast the failure claims, even better than age based forecasting.

Keywords: Reliability, Warranty Policy, Month-in-Service (MIS), Censoring.

1 Introduction

Reliability is an important aspect of product quality. The reliability of product is the probability that the item will perform its intended function throughout a specified time period when the operated in a normal environment (Blischke and Murthy [1]). That is, reliability is quality over time (Condra [2]). In this modern age rapid advance in technology, development of high sophisticated products, intense global competition, and increasing customer expectation have put new pressure on manufactures to produce high quality products. Failure is directly related to the products and the manufactures main target is to reduce failure and the improvement of quality of product that the customer expected. Manufactures always try to fulfill the customer satisfaction and expectation. So manufactures follow a warranty policy. Warranty policy is a statement, in connection with sale of products, on the kind (e.g., repair/ replacement, part refund, free service etc.) and extend (length of period) of compensation offered by the manufacturer in the event of failure. Warranty policy helps the manufacturer to (i) impress upon the prospective customers about the superior product quality and reliability, when compared to that of the competitors, and (ii) carry out post sale service for better customer satisfaction.

Generally failure of a component or product are mainly depends on its age or usage or both age and usage. There are multiple failure time scales and each scale contributes partial failure information. The warranty policy can be developed on only age or usage based model (single variable model) and composite(both age and usage) based model (multiple variable model).

Jiang and Jardine [3] note that the failure prediction based on the composite scale model will be more accurate than that based on the model of single variable. This is because the information capacity carried by the composite variable is larger than that carried by any individual variable or measure. So, the composite scale model is expected to have better capacity in forecasting failure number than that of single variable model.

There are mainly two types of composite scale models, which are: (i) Linear models and (ii) Multiplicative models. Three major applications of the composite scale models as discussed in [3] are:

(i) To build a composite scale model when the variables are all time scales;

(ii) To build a composite covariate model when the variables are all covariates; and

(iii) To build a failure time model with covariates when the variables include time variables and covariates.

Meeker and Escobar [4] state that, “field data are vital resources. There will always be risk of failures with any product design. It is necessary to develop processes for the collection and appropriate use of field feedback to quickly discover and correct potential problems before they become widespread, thereby reducing overall risk”.

Forecasting the number of failure is the challenging work for manufactures. It can be done from warranty database and policy. Warranty database provide information on a product’s performance in real field condition. However it is very difficult and costly to track the performance in real field conditions. But it is only way to collect warranty claim

data. In reliability analysis, two dimensional warranty claims data are available and the two dimensional information are time-to-failures or ages and their usages. In previous, several works have been done only on age based information for forecasting the number of failures [4]. In this paper, we use multiplicative composite scale model for forecasting the number of failure for two dimensional (2-D) warranty data.

The rest of the paper is organized as follows. Section 2 discusses the warranty policy and model setting. Sections 3 describe the multiplicative composite model for two-dimensional warranty claims. Section 4 illustrates the proposed and existing approaches to composite modeling data analysis. Sections 5 describe essential data and parameter estimation procedure. Section 6 discusses a simulation study for different parameter set and finally, Section 7 concludes the paper with a brief summary.

2 Warranty Policy and Model Setting

Blischke and Murthy [5] give a detailed discussion of the various types of warranty policies. The policy can broadly be grouped into two types: one and two-dimensional. The one-dimensional policy is characterized by a time interval (e.g., 12 months) as warranty period or by usage interval (e.g., 1, 20,000 copies for photocopiers) as warranty period. The two-dimensional warranty policy ends after a specified time or amount of use, whichever comes first. A typical example of two-dimensional warranty policy for an automobile is as follows: the manufacturer agrees to repair or provide replacement for its failed parts free of cost for the maximum period of 10 years or maximum usage of 100000 km, whichever occurs first, from the starting time scale. When a warranted product fails within the warranty period, and the consumer makes legitimate claim to the manufacturer for repair or replacement of product, the claim is known as *warranty claim*. This paper considered warranty claims data with a two-dimensional scheme: *age warranty* and *usage warranty*. Different manufacturer offer different warranty schemes. Nowadays warranty periods are being longer. The manufacturer companies are paying a lot of money for warranty costs, which are increasing.

Warranty claim records, which are usually provided by maintenance or service department. When a warranty claim is made, failure related information is recorded in the warranty database. Two variables that are important in the field reliability studies are month-in-service (MIS) and usages at the time of warranty claim. The warranty claim databases can be represented as in Table 1.

Manufacturer can represent, where each cell corresponding to $t$month-in-service (MIS) and $l$month-to-failure (MTF) will have failure usages of r_t^l failed components or products or items. Thus with MIS t, out of N_t sold items there will be failure usages of r_t failed items, and these failure usages constitute the data set under interest.

Table 1: Information available from warranty database.

Sales Month	MIS, t	Sales amount	MTF					Failure number	Censored number
			1	…	l	…	T		
1	T	N_T	r_T^1	…	r_T^l	…	r_T^T	r_T	c_T
…	…	…	…	…	…	…		…	
$T-t+1$	t	N_t	r_t^1	…	r_t^l			r_t	c_t
…	…	…	…	…				…	
T	1	N_1	r_1^1					r_1	c_1
Total		**N**						**n**	**N − n**

The notation in the Table 1 is summarized as follows:

- t : months-in-service (MIS); $t = 1,2\ldots,T$
- T : maximum MIS where $T = \min(W, S, M)$
- W : warranty period in MIS
- S : total months of sales
- M : observation period in calendar time
- N_t : number of product or component with t MIS
- r_t^l : number of failures claimed at l MTF for the product or component with t MIS
- r_t : number of failures with t MIS, $r_t = \sum_{l=1}^{t} r_t^l, t = 1, \ldots, T$
- c_t : number of non-failures with t MIS, $c_t = N_t - r_t$
- n : total number of failures
- N : total number of product or component in the field

Let X be a lifetime variable under study measured by actual usage, and let Y_t be a censoring random variable representing the total usage accumulated by an item at t MIS, which is independent of X. The following competing risks model is often used for censoring problems in which the observations $(U_{ti}, \delta_{ti}), i = 1, \dots, N_t; t = = 1, \dots, T$, have been obtained:

$$U_{ti} = \min\{X_i, Y_{ti}\}, \text{ and}$$

$$\delta_{ti} = \begin{cases} 1 \; if \; X_i \leq Y_{ti} \\ 0 \; if \; X_i > Y_{ti} \end{cases}$$

for the i-th usages $(i = 1, \dots, N_t; t = 1, \dots, T)$. Several studies have been done on the censoring problem [18].

3 Multiplicative Composite Model for Two-Dimensional Warranty Claims

Multiplicative composite scale modeling is used to combine several scales or variables into a single scale or variable. The multiplicative composite scale is expected to have better failure prediction capability than individual scales [6-11]. The traditional method of building a composite model is to make the transformed data in the composite scale have a minimum coefficient of variation (CV). The smaller it is, the more accurate the prediction of failure time it will be [3]. When products are sold with two-dimensional warranties, the warranty claims data are two dimensional. Usually the first dimension refers to the age of the products or items at the time of failure and the second to usage [16]. Now we consider, T be the age and U be the usage. The two dimensional warranty provides convergence over a region $R = [0, A) \times [0, B)$, for example, $A = 18$ months and $B = 150{,}000$ km. Thus the failures are covered under warranty only if $(T, U) \in R$. Composite scale model involves forming a new variable V that is the combination of usage U and age T. Here we follow the approach suggested by Gertsbakh and Kordonsky (called the G-K approach [12]), in which the composite scale model is a multiplicative combination of the form

$$V = T^a \cdot U^b \tag{1}$$

The parameters a and b can be determined by minimizing the sample variance of V subject to the sample mean of V being equal to the maximum of sample means of $T$and U [3]. In mathematically,

Minimize sample variance of, $V = T^a \cdot U^b$,

subject to the condition, $\mu_v = \mu_0$.

The choice for μ_0 may be: $\mu_0 = max(\mu_T, \mu_U)$ where μ_v is the sample mean of V.

Here, the two-dimensional problem is effectively reduced to a one-dimensional problem.

4 Approaches to Approximate Censored Usage

The two-dimensional warranty data are effectively reduced to one-dimensional data by treating usage as a function of age or age as a function of usage. In this section, we assume that the usage rate for a customer is constant over the warranty period but varies across the customer population. Modeling of failures under warranty is then done using 1-D models by conditioning on the usage rate [13-14]. We assume that usage rate is different for each different customer. Let (T_{ij}, U_{ij}) denote the age and usage at the time of the j-th warranty claim for customer i.The usage rate for customer i with a single failure is

$$z_i = \frac{U_{ij}}{T_{ij}}, j = 1, i = 1,2, \dots \dots \dots, I_1 \tag{2}$$

The underlying model formulations for the several approaches we look at failures and censors. Subsequent failures data are available by their age but not their usage. So we involve the different approaches to modeling 2-dimensinal failures. Approach 1 and Approach 2 already exist and discussed in [14]. Further three approaches are newly proposed, called Approach 3, Approach 4 and Approach 5. Now the approaches are given bellow:

Approach-1: In this approach we compute average sample usage rate from the warranty claims data, which can be treated as a censoring observation for each age (e.g., day, week, month etc.) by multiplying usage by sequential age (e.g., week, month etc.). Here usage rate is computed by dividing the accumulated usage by its corresponding age.

Approach-2: In this approach we compute median sample usage rate from the claims data, which can be treated as a censoring observation for each age by multiplying usage by sequential age.

Approach-3: In this approach we compute maximum sample usage rate form the claims data corresponding to each age, that can be treated as a censoring observation for each age. If there is no maximum sample use rate corresponding to each age, in that place we use average or/and median sample use rate.

Approach-4: In this approach we compute average sample usage rate form the claims data corresponding to each age that can be treated as a censoring observation for each age. If there is no average sample use rate corresponding to each age, in that place we use average or/and median or/and maximum sample use rate.

Approach-5: In this approach we compute median sample usage rate form claims data corresponding to each age that can be treated as a censoring observation for each age. If there is no median sample use rate corresponding to each age, in that place we use average or/and median or/and maximum sample use rate.

5 Essential Data and Parameter Estimation

The analysis may be done at the component or product level. For simplicity, let *N* denote the total number of items in the data set, including original purchases, and let *n* be the total number of failures. Observed values are denoted x_i,andy_i. Renumber the data so that (x_i, y_i) correspond to failures for $i = 1, 2 \dots n$, and to age and usage of non-failed items for $i = n + 1 \dots N$. Age data for all items are easily obtained from the sales data (at an aggregated level, if necessary). Usage data, however, are rarely available for un-failed items.

Assuming that data have been aggregated at some level, with *k* = number of periods for which the following data are available for $j = 1, 2, \dots, k$

N_j = sales in period *j*
A_j = number of valid claims on the N_j units sold in period *j*
M_j = number of items sold in period *j* for which no claims were made = $N_j - A_j$
T_j = age at censoring for items sold in period *j*
K = total number of valid claims = $\sum A_j$
U_j = usage at censoring for items sold in period *j*

If usage is not known, it may be estimated as indicated in the previous Section, that can be used forecasting. Let

$$R_j = (x, y): x \in [0, U_j), y \in [0, T_j) \tag{3}$$

Valid warranty claims resulting from item failures must be such that $(x_i, y_i) \in R_j$. $R_j$may be viewed as service ages and usages for un-failed items sold in period j. Data for analysis are failure data, namely those for which $(x_i, y_i) \in R_j$, andcensored data, i.e., unobserved random variables (X_j, Y_j) which are not in theregion $R_j$for any j.

Let T_j and U_j denote the values of censored age and usage for items sold in period j. The corresponding censored values of V_j, denoted $\tilde{V}$, are calculated as

$$\tilde{V} = T_j^a \cdot U_j^b, j = 1,2, \dots \dots, k \tag{4}$$

Failure data on the *v*-scale are

$$v_i = x_i^a \cdot y_i^b, i = 1,2, \dots \dots, n \tag{5}$$

The value of a and b are determined by a procedure, which discuss in Section 3.

Now we select a distribution F and use the one-dimensional data V and $\tilde{V}$associated with the values of a and b to calculate MLEs of the parameters of F basedon the following likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(v_i; \theta) \prod_{j=1}^{K} [1 - F(\tilde{V}_j; \theta)]^{M_j} \tag{6}$$

In the G-K approach, the Weibull distribution is usually used for this purpose. We also use the Weibull distribution in our purpose. We consider the distribution function and density function of the chosen distribution are $F(v; \theta)$ and $f(v; \theta)$, respectively.

MLEs of the elements of the parameter θ are obtained by maximizing $logL(\theta)$given in (6) for the selected values of $a$and b, using Newton Raphson iteration method for solution [15-16]. The asymptotic normality of the MLEs may be used to obtain confidence intervals (CI) for the estimated parameters. This is done by forming the matrix as inverting the matrix, and substituting the estimated values of unknown parameters into the result. This provides estimated variances and covariance of the estimators and the confidence intervals (CI) are obtained by use of the standard normal distribution.

6 Simulation Study

Our interest in this study is to forecast the number of failure using the multiplicative composite scale model in warranty claim data with different approach and compares the composite model with age based model. In this paper we perform a simulation study. The whole processes were repeated 1000 times and we use average values from 1000 repetitions. To perform the simulation study data were generated assuming that the lifetime variable $X \sim Weibull\ (\beta_0, \eta_0)$ and the

censoring variable $Y_t \sim LN\ (\mu + log\ t, \sigma^2)$for three set of true parameter values [17-18]. In previous several works have been done for considering the lifetime variable and censoring variable [18]. Based on that information we also use these distributions for generating the warranty claims data. In this paper we consider about three types of parameter settings: increasing failure rate (IFR), constant failure rate (CFR), and decreasing failure rate (DFR). The true parameter settings considered for the simulation studies are given in the Table 2.

Table 2: The three parameter sets of the simulation study.

Setting	Parameters				Type
	β_0	η_0	μ	σ	
1	2.00	30000	6.50	0.70	IFR
2	1.00	95000	6.50	0.70	CFR
3	0.70	90000	5.50	0.70	DFR

6.1 Forecasting the failure number for multiplicative composite scale model:

In this paper we want to forecast the failure number of company product over the next year. The warranty period that the company offers is 12 months. Hence 9 month failure data are available that are generated using the parameter (Table 2) and we want to forecast the failure numbers in next 3 months that completes 12 months warranty. The products are sold 2,000 units per month and we also consider that the company will have sales of 2000 units per month over the next year. We consider , the setup for the warranty folio given that we have sales data for 10 months stating in month-1(M-1), return data for 9 months starting in month-2 (M-2), and we also want to include future sales for next 12 months. We generate warranty claims (failures) data to month-in-service and we transform this data to month-to-failure and fit a distribution. We have chosen to fit a 2-parameter Weibull distribution using MLE as the parameter estimation method due to the large number of suspended data points in the data set. Once we have obtained a failure distribution, we can use the concept of conditional reliability in order to calculate the probability of failure for the reaming units after each sales period then make forecasting of future failure. The equation of the conditional probability of failure is:

$$Q(\tau|T) = 1 - R(\tau|T) = 1 - \frac{R(T+\tau)}{R(T)} \tag{7}$$

where$Q(\tau|T)$ is the unreliability of a unit for the next τ months given that it has already operated successfully for T months, $R(\tau|T)$ is the reliability of a unit for the next τ months given that it has already operated successfully for T months and $R(.)$ is the reliability function.

Let, out of Z units that are sold in Month-1(M-1), B units are still out in the field as of the end of Month 10 (M-10). That is calculated by subtracting the total number of returns of the M-1 shipment from the number of sales for that month. Then the forecasted number of failure can be calculated as $NF = B \cdot Q(\tau|T)$.

For the IFR type data (parameter set 1, Table 2) and different approaches to approximate censored usage as discussed in Section 4, the average MLE of the parameters of Weibull distribution under multiplicative composite scale model is presented in the Table 3.

Table 3: The average MLE of the Weibull parameters of composite model from 1000 repetition.

Approaches	MLE of the parameters		Standard Error		90% Confidence Interval			
					$\hat{\beta}$		$\hat{\eta}$	
	$\hat{\beta}$	$\hat{\eta}$	S. E. ($\hat{\beta}$)	S. E. ($\hat{\eta}$)	Lower	Upper	Lower	Upper
Approach 1	1.8435	52.7610	0.0620	4.2519	1.7418	1.9452	45.7876	59.7336
Approach 2	1.8967	49.2982	0.0635	3.6221	1.7926	2.0008	43.3580	55.2384
Approach 3	2.2755	35.4427	0.1977	5.0685	1.9512	2.5998	27.1304	43.7549
Approach 4	2.7978	25.5851	0.4575	6.5407	2.0475	3.5482	14.8583	36.3118
Approach 5	2.8734	24.6350	0.4931	6.6387	2.0648	3.6820	13.7472	35.5221

Table 4: Forecasted number of failures by MIS for different approaches with IFR type data.

Approaches	**Number of failure in MIS**											
	M-1	**M-2**	**M-3**	**M-4**	**M-5**	**M-6**	**M-7**	**M-8**	**M-9**	**M-10**	**M-11**	**M-12**
Actual	3	11	23	40	61	79	110	139	170	201	234	268
Approach 1	1	4	9	18	32	49	71	94	122	153	184	218
Approach 2	1	4	9	18	32	50	71	96	124	155	188	222
Approach 3	1	3	7	16	31	51	75	103	134	169	205	243
Approach 4	0	1	4	13	30	52	80	114	153	196	239	287
Approach 5	0	1	4	13	30	53	82	116	156	200	246	295
Age	1	6	16	34	58	82	113	154	190	229	277	317

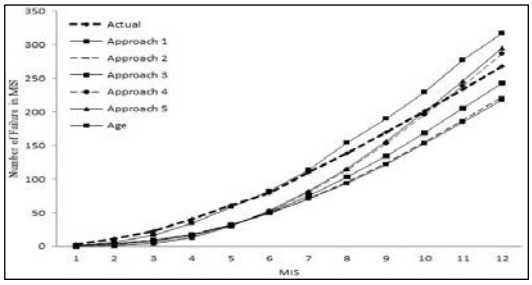


Figure 1: Comparison of different approaches in forecasting future failure claims (IFR type data).

Table 5: Forecasted failure number in the next 12 months for IFR type data with Approach 4.

Sales month	MIS	Sales amount	No of failure	M-10	M-11	M-12	M-13	M-14	M-15	M-16	M-17	M-18	M-19	M-20	M-21
M-1	9	2000	170	33	39	45									
M-2	8	2000	139	28	33	39	45								
M-3	7	2000	111	23	28	34	39	45							
M-4	6	2000	84	18	23	28	34	40	45						
M-5	5	2000	60	13	18	23	28	34	40	46					
M-6	4	2000	40	9	13	18	23	28	34	40	46				
M-7	3	2000	23	6	9	14	18	23	29	34	40	46			
M-8	2	2000	11	3	6	10	14	18	23	29	34	40	46		
M-9	1	2000	3	1	3	6	10	14	18	23	29	35	41	47	
M-10				0	1	3	6	10	14	18	23	29	35	41	47
M-11					0	1	3	6	10	14	18	23	29	35	41
M-12						0	1	3	6	10	14	18	23	29	35
M-13							0	1	3	6	10	14	18	23	29
M-14								0	1	3	6	10	14	18	23
M-15									0	1	3	6	10	14	18
M-16										0	1	3	6	10	14
M-17											0	1	3	6	10
M-18												0	1	3	6
M-19													0	1	3
M-20														0	1
M-21															0
Total				134	173	221	221	222	223	224	224	225	226	227	227

Using 9 month data we forecast the number of failures for the next 3 months for choosing an appropriate approach. The forecasting value of 12 MIS is given in the Table 4 and also their plot in Figure 1.

From Table 4 and Figure 1, it is observed that the Approach 4 performs better than other approaches. It is also better than age based forecasting because the number of failure of Approach 4 in MIS 12 is closely near to the actual failure in MIS 12. Now the forecasting of failure number in the next 12 months using the Approach 4 is presented in the Table 5. The 90% confidence bound of the number of failure for Approach 4 along with actual failure are shown in Figure 2. Finally, Figure 3 shows the expected failures for each month of next 12 months that are still out in the field along with the 90% upper and lower confidence bounds.

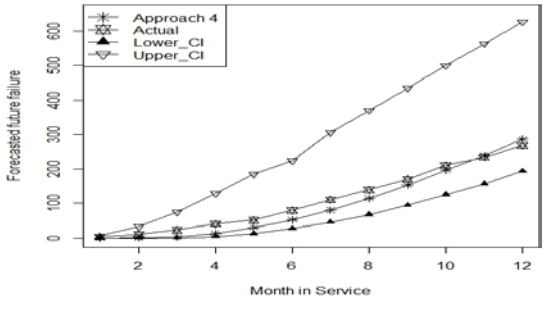


Figure 2: 90% Confidence bound of the number of failure by MIS (IFR type data).

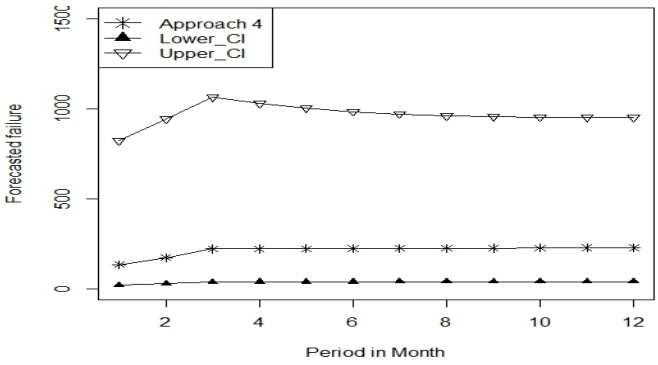


Figure 3: 90% Confidence bound of the expected failures for each month of next 12 months (IFR type data).

Table 6: Forecasted number of failures by MIS for different approaches with CFR type data.

Approaches	Number of failure in MIS											
	M-1	M-2	M-3	M-4	M-5	M-6	M-7	M-8	M-9	M-10	M-11	M-12
Actual	18	36	53	69	86	105	120	134	152	169	185	196
Approach 1	7	21	41	77	110	140	168	195	222	248	272	296
Approach 2	8	22	40	71	99	124	148	171	192	215	235	254
Approach 3	11	25	41	65	87	107	127	146	164	182	199	216
Approach 4	13	25	36	52	68	84	100	117	133	149	164	179
Approach 5	14	25	36	50	66	82	98	114	130	145	160	176
Age	9	23	40	67	91	112	133	153	171	190	207	226

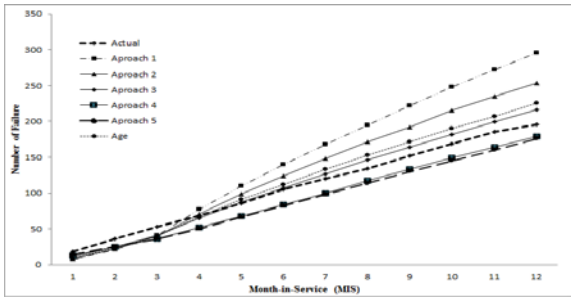


Figure 4: Comparison of different approaches in forecasted future failure claims (CFR type data).

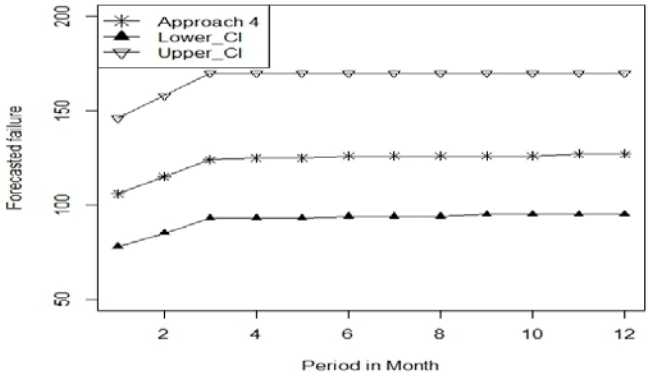


Figure 5: 90% Confidence bound of the expected failures for each month of next 12 months (CFR type data).

For the CFR type data (parameter set 2, Table 2), similar simulation result are obtained. The simulation study for CFR type data, the result indicates that the proposed Approach 4 performs better than other approaches. It is also better than age based forecasting. The specific results are shown in Table 6 and Figure 4. Finally, Figure 5 shows the

expected failures for each month of next 12 months that are still out in the field along with the 90% upper and lower confidence bounds.

Again the simulation study for DFR type data (parameter set 3, Table 2), the result indicates that the proposed Approach 5 performs better than other approaches and also Approach 4 performs similar to the Approach 5. Since the multiplicative composite scale model gives better forecast than single variable model (e.g., age based) so the specific results for DFR type data are shown in Table 7 and Figure 6. Also 90% Confidence bound of the expected failures for each month of next 12 months are present in the Figure 7.

Table 7: Forecasted number of failures by MIS for different approaches with DFR type data.

Approaches	Number of failure in MIS											
	M-1	M-2	M-3	M-4	M-5	M-6	M-7	M-8	M-9	M-10	M-11	M-12
Actual	37	62	78	91	105	120	132	146	168	175	182	189
Approach 1	4	31	101	257	442	650	867	1079	1274	1446	1591	1706
Approach 2	5	30	86	206	330	462	596	730	861	987	1105	1216
Approach 3	8	36	84	178	259	334	408	478	545	610	672	732
Approach 4	12	33	60	115	150	178	203	225	247	266	285	302
Approach 5	13	33	57	107	138	163	185	206	224	240	258	273

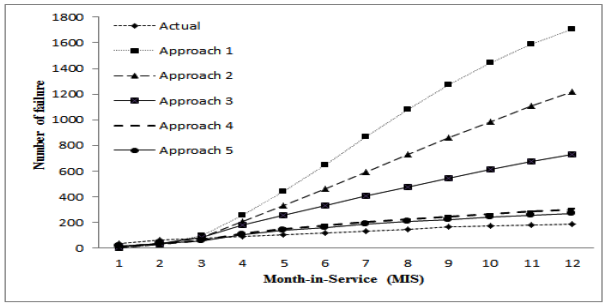


Figure 6: Comparison of different approaches in forecasting future failure claims (DFR type data).

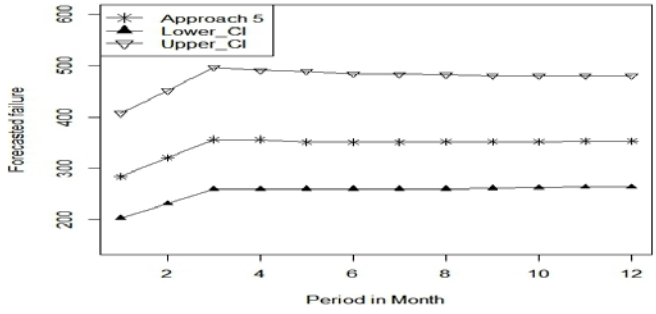


Figure 7: 90% Confidence bound of the expected failures for each month of next 12 months (DFR type data).

7 Conclusion

In this paper, the number of future failure for multiplicative composite scale model based on age and usage in the presence of censored observations has been examined. As warranty data is incomplete due to unavailability of censored usage, three approaches are proposed to approximate the censored usage along with two existing approaches. This information is used to make the incomplete data as complete. Weibull distribution is then fitted to the approximate complete data in multiplicative composite scale and is used to forecast the failure claims.

The primary goal of this research was to investigate whether composite scale is capable or not to forecast failure claims. If so, how it performs as compared to age based forecasting. For this purpose, a brief simulation was performed. The simulation results reveals that one can fairly predict future failure claims in two-dimensional warranty scheme in composite model. Further we can use these approaches in real data analysis.

This paper considered only single failure mode. In future we are interested to consider multiple failure modes in forecasting failure claims for warranted products. In such cases consideration of usage condition as a covariate would be interesting.

References

1. Blischke, W.R. and Murthy, D.N.P. (2003). *Case Studies in Reliability and Maintenance* (editors), New Jersey: John Wiley & Sons, Inc.
2. Condra, L.W. (1993). *Reliability Improvement with Design of Experiments*, 2nd Ed., New York: Marcel Dekker, Inc.
3. Jiang, R. and Jardine, A.K.S. (2006). Composite scale modeling in the presence of censored data, *Reliability Engineering and System Safety*,**91**(7), pp.756–764.
4. Meeker, W.Q. and Escobar, L.A. (1998). *Statistical Methods for Reliability Data*, New York: John Wiley & Sons, Inc.
5. Blischke, W.R. and Murthy, D.N.P. (1994). *Warranty Cost Analysis*, New York: Marcel Dekker, Inc.
6. Karim, M.R. and Suzuki, K. (2004). Analysis of warranty claim data: A literature review, *International Journal of Quality & Reliability Management*, **22**, pp. 667-686.
7. Oakes, D. (1995). Multiple time scales in survival analysis,*Lifetime Data Anal*, **1**(1), pp. 7-18.
8. Gertsbakh, I.B. and Kordonsky, K.B. (1998). Parallel time scales and two-dimensional manufacturer and individual customer warranties, *IIE Trans*, **30**(12): pp. 1181-9.
9. Duchesne, T. and Lawless, J. (2000). Alternative time scales and failure time models, *Lifetime Data Anal*, **6**(3), pp. 157-179.
10. Duchesne, T. and Lawless, J. (2002). Semi parametric inference methods for general time scale models, *Lifetime Data Anal*, 8(4), pp. 263-276.
11. Frickenstein, S.G. and Whitaker, L.R. (2003). Age replacement policies in two time scale, *Nav Res Log*, **50**(7), pp. 592-613.
12. Gertsbakh, I.B. and Kordonsky, K.B. (1993). Choice of the best time scale for system reliability-analysis, *Eur J Oper Res*,**65**(3), pp. 235-46.
13. Lawless, J.F., Hu, X.J. and Cao, J. (1995). Methods for the estimation of failure distributions and rates from automobile warranty data, *Lifetime Data Anal*, **1,** pp. 227–240
14. Blischke, W. R., Karim, M. R. and Murthy, D. N. P. (2011). *Warranty Data Collection and Analysis*,New York: Springer.
15. Murthy, D.N.P., Xie, M. and Jiang, R. (2004). *Weibull Models*, New York: Wiley.
16. Alam, M.M. and Suzuki, K. (2009).Lifetime estimation using only information from warranty database, *IEEE Transactions on Reliability*.
17. Rai,B. and Singh, N. (2006).Customer-rush near warranty expiration limit and nonparametric hazard rate estimation from the known mileage accumulation rates,*IEEE Transactions on Reliability*, **55**, pp. 480-489.
18. Alam, M.M., Suzuki, K. and Yamamoto, W. (2009). Estimation of lifetime parameters using warranty data consisting only failure information, *Journal of Japanese Society for Quality Control (JSQC)*, **39**(3), pp. 79-89.

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Application of Mixture Models for Analyzing Reliability Data: A Case Study

SabbaRuhi, S.M. Sayadat Amin, Tahsina Aziz and M. RezaulKarim

Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh

Emails: sabba.ruhi@gmail.com; sayadat_stat@yahoo.com; mrezakarim@yahoo.com

Abstract. Over the last two decades there has been a heightened interest in improving quality, productivity and reliability of manufactured products. Rapid advances in technology and constantly increasing demands of customers for sophisticated products have put new pressures on manufacturers to produce highly reliable products. There are situations where some components are produced over a period of timeusing different machines, collecting items from different vendors, etc.The physical characteristics and the reliability of such componentsmay be different, but it may be difficult todistinguish clearly. In such situations,mixtures of distributions are often used in the analysis of reliability data for these components.

In this study, a set of competitive two-fold mixture models (based on Weibull, Exponential, Normal and Lognormaldistributions) are applied to analyze a set of Aircraft windshield failure data. The data consist of both failure and censored lifetimes of the windshield. In the existing literature, there are many uses of mixture models for the complete data, but very limited literature available about it uses for the censored data case. Maximum likelihood estimation method is used to estimate the model parameters and the Akaike Information Criterion (AIC) and Anderson-Darling (AD) test statistic are applied to select the suitable models among a set of competitive models. It is found that the Weibull-Exponential and Normal-Lognormal mixture models fit good for the data. Various characteristics of the mixture models, such as the cumulative distribution function, reliability function, B10 life, mean time to failure, etc. are estimated to assess the reliability of the component.

Keywords. Reliability; Mixture models; Failure data; MLE

1 Introduction

Whenever customers purchase durable goods, they expect it to function properly at least for a reasonable period of timeunder usual operating conditions. That is, customers expect that the purchased products would be reliable and safe. So, the manufacturers bear the responsibility to inform their customers about the reliability of their products. Again for costly items, customers expect a minimum life time during which the item should work properly without any disturbance. Improving reliability of product is an important part of the larger overall picture of improving product quality. Therefore, in recent years many manufacturers have collected and analyzed field failure data to enhance the quality and reliability of their products and to improve customer satisfaction.

The mixture models have been extensively used in reliability theory and in many other areas. Murthy et al. (2004) mentioned a list of applications of mixture models in reliability theory. In this article we apply two-fold mixture models to analyze a set of Aircraft windshield failure data and to assess the reliability of the windshield.A number of standard lifetime distributions are considered as the distributions of the subpopulations for mixture models. Both the nonparametric (Probability paper plot and Kaplan-Meier estimate) and parametric (maximum likelihood method) estimation procedures are used for estimation purposes. The Akaike Information Criterion (AIC) and Anderson-Darling (AD) test are applied to select the suitable models for the data set. A variety of quantities of interest are estimated in investigating product reliability.

The remainder of the article is organized as follows: Section 2 describes a set of product failure data which will be analyzed in this paper. Section 3 discusses the mixture models. Section 4 explains the parameter estimation and model selection procedures of the lifetime models of the component. Finally, Section 5 concludes the article with additional implementation issues for further research.

2 Data Set

Field failure data are superior to laboratory test data in the sense that they contain valuable information on the performance of a product in actual usage conditions. There are many sources of collecting reliability- related data of a product. Warranty claim data are used as an important source of field failure data which can be collected economically and efficiently through repair service networks and therefore, a number of procedures have been

developed for collecting and analyzing warranty claim data (e.g. Karim and Suzuki, 2005; Karim et al., 2001; Lawless, 1998; Murthy and Djamaludin, 2002; Suzuki, 1985; Suzuki et al., 2001). The recent book written by Blischke et al. (2011) is an excellent reference on the collection and analysis of warranty data.

Table1: Windshield Failure Data

Failure Times				Service Times		
0.040	1.866	2.385	3.443	0.046	1.436	2.592
0.301	1.876	2.481	3.467	0.140	1.492	2.600
0.309	1.899	2.610	3.478	0.150	1.580	2.670
0.557	1.911	2.625	3.578	0.248	1.719	2.717
0.943	1.912	2.632	3.595	0.280	1.794	2.819
1.070	1.914	2.646	3.699	0.313	1.915	2.820
1.124	1.981	2.661	3.779	0.389	1.920	2.878
1.248	2.01	2.688	3.924	0.487	1.963	2.950
1.281	2.038	2.823	4.035	0.622	1.978	3.003
1.281	2.085	2.89	4.121	0.900	2.053	3.102
1.303	2.089	2.902	4.167	0.952	2.065	3.304
1.432	2.097	2.934	4.240	0.996	2.117	3.483
1.480	2.135	2.962	4.255	1.003	2.137	3.500
1.505	2.154	2.964	4.278	1.010	2.141	3.622
1.506	2.190	3.000	4.305	1.085	2.163	3.665
1.568	2.194	3.103	4.376	1.092	2.183	3.695
1.615	2.223	3.114	4.449	1.152	2.240	4.015
1.619	2.224	3.117	4.485	1.183	2.341	4.628
1.652	2.229	3.166	4.570	1.244	2.435	4.806
1.652	2.300	3.344	4.602	1.249	2.464	4.881
1.757	2.324	3.376	4.663	1.262	2.543	5.140

Aircraft Windshield Failure Data

As discussed in Murthy et al. (2004), the windshield on a large aircraft is a complex piece of equipment, comprised basically of several layers of material, including a very strong outer skin with a heated layer just beneath it, all laminated under high temperature and pressure. Data on all windshields are routinely collected and analyzed. At any specific point in time, these data will include failures to date of a particular model as well as service times of all items that have not failed. Data of this type are incomplete in that not all failure times have as yet been observed.

Failures of the Aircraft Windshield are not structural failures. Instead, they typically involve damage or delamination of the nonstructural outer ply or failure of the heating system. These failures do not result in damage to the aircraft but do result in replacement of the windshield. Data on failure and service times for a particular model windshield are given in Table-1 from Murthy, et al. (2004), originally given in Blischke and Murthy (2000). The data consist of 153 observations of which 88 are classified as failed windshields, and the remaining 65 are service time (censored time) of windshields that had not failed at the time of observation. The unit for measurement is 1000h.

3 Modeling

In the real world, problems arise in many different contexts. Problem solving is an activity that has a history as old as the human race. Models have played an important role in solving the problem. A variety class of statistical models have been developed and studied extensively in the analysis of the product failure data (Kalbfleisch and Prentice, 1980; Meeker and Escobar, 1998; Blischke and Murthy, 2000; Lawless, 2003; Murthy et al., 2004). The models that will be used to analyze the product failure data,given in Table 1, are discussed below.

Mixture Models : A general *n*-fold mixture model involves *n* subpopulations and is given by

$$G(t) = \sum_{i=1}^{n} p_i F_i(t) \quad , p_i > 0 , \sum_{i=1}^{n} p_i = 1 \quad , \tag{1}$$

where $F_i(t)$ is the CDF of the *i*-th sub-population and p_i is the mixing probability of the *i*-th sub-population. The density function is given by:

$$g(t) = \sum_{i=1}^{n} p_i f_i(t), \tag{2}$$

where $f_i(t)$ is the density function associated with $F_i(t)$.

The hazard function $h(t)$ is given by

$$h(t) = \frac{g(t)}{1-G(t)} = \sum_{i=1}^{n} w_i(t) h_i(t) \tag{3}$$

where $h_i(t)$ is the hazard function associated with subpopulation i , and

$$w_i(t) = \frac{p_i R_i(t)}{\sum_{i=1}^{n} p_i R_i(t)} \quad \sum_{i=1}^{n} w_i(t) = 1 \tag{4}$$

With $$R_i(t) = 1 - F_i(t), \quad 1 \le i \le n \,. \tag{5}$$

From (4), we see that the failure rate for the model is a weighted mean of the failure rate for the subpopulations with the weights varying with *t*.

Special Case: *Two-fold Mixture Model (n=2)*

The CDF of the two-fold mixture model is given by

$$G(t) = pF_1(t) + (1-p)F_2(t) \,. \tag{6}$$

For example, suppose, $f_1(t) \sim$ Weibull (α, β) and $f_2(t) \sim$exponential (σ) distribution. Hence, the distribution function for Weibull-Exponential mixture model from equation (6) is:

$$G(t) = 1 - p\exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] + p\exp\left[-\left(\frac{t}{\sigma}\right)\right] - \exp\left[-\left(\frac{t}{\sigma}\right)\right] \,. \tag{7}$$

The probability density function $g(t)$ is:

$$g(t) = p\left[\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}\right] + \frac{(1-p)}{\sigma}\exp\left[-\left(\frac{t}{\sigma}\right)\right] \,. \tag{8}$$

And the hazard function is:

$$h(t) = \frac{p\left[\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}\right] + \frac{(1-p)}{\sigma}\exp\left[-\left(\frac{t}{\sigma}\right)\right]}{p\exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] + (1-p)\exp\left[-\left(\frac{t}{\sigma}\right)\right]} \,. \tag{9}$$

Other two-fold mixture models can be derived by using different CDFs in (6) from different lifetime distributions, similarly.

4 Parameter Estimation

For censored data the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f(t_i)^{\delta_i} \left[R(t_i)\right]^{1-\delta_i}$$

where$\delta_i$is the failure-censoring indicator for t_i(taking on thevalue 1 for failed items and 0 for censored).Taking log on both sides we get,

$$\ln L = \sum_{i=1}^{n} \{\delta_i \ln[f(t_i)] + (1-\delta_i)\ln[R(t_i)]\} \,. \tag{10}$$

In the case of Weibull-Exponential mixture model putting the value of CDF and pdf of the model in equation (10), we obtain the log-likelihood function of Weibull-Exponential mixture model, which is:

$$\ln L = \sum_{i=1}^{n}\left[\delta_i \ln p\left[\frac{\beta}{\alpha}\left(\frac{t_i}{\alpha}\right)^{\beta-1}\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}\right] + \frac{(1-p)}{\theta}\exp\left[-\left(\frac{t_i}{\sigma}\right)\right]\right]$$
$$+\sum_{i=1}^{n}\left[(1-\delta_i)\ln\left\{p\exp\left[-\left(\frac{t_r}{\alpha}\right)^{\beta}\right] - p\exp\left[-\left(\frac{t_r}{\sigma}\right)\right] + \exp\left[-\left(\frac{t_r}{\sigma}\right)\right]\right\}\right]. \quad (11)$$

The maximum likelihood estimates of the parameters are obtained by solving the partial derivative equationsof (11) with respect to α, β, σ and *p*. But the estimating equations do not give any closed form solutions for the parameters. Therefore, we maximize the log likelihood numerically and obtain the MLE of the parameters. In this article, the "mle" function given in the R-package is used to maximize (11) numerically. It is very sensitive to initial values of parameters of these models.

Akaike Information Criterion (AIC)

Akaike's information criterion, developed by Hirotsugu Akaike under the name of "an information criterion" (AIC) in 1971, is a measure of the goodness of fit of an estimated statistical model. This is the most widely used criterion for selecting the best model for a given data set. The model with the lowest AIC being the best. In general case, the AIC is

$$AIC = 2k - 2\ln(L),$$

where *k* is the number of parameters in the model, and *L* is the maximized value of the likelihood function for the assumed model.

Anderson-Darling (AD) and Adjusted AD test statistics

The Anderson–Darling test is named after Theodore Wilbur Anderson and Donald A. Darling, who invented it in 1952. The Anderson-Darling test is based on the difference between the hypothesized CDF and empirical distribution function (EDF). The AD test statistic is given by:

$$A^2 = A_n{}^2 = \frac{-1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln\left\{F\left[t_{(i)}\right]\right\} + \ln\left\{1 - F\left[t_{(n-i+1)}\right]\right\}\right] - n.$$

And the Adjusted AD test statistic is given by $AD^* = \left(1 + 0.2/\sqrt{n}\right)A^2.$

Best distribution is the one with the lowest value of AD test statistic.

The value of -2log-likelihood, AIC, AD and the Adjusted AD test statistics of the seven mixture models are estimated for Windshield failure data. The results are displayed in Table-1.

Table1: Estimates of -2log-likelihood, AIC, AD and the Adjusted AD for seven mixture models

Mixture Models	-2 logL	AIC	AD	Adj AD
1. Weibull-Weibull	354.92	364.92	5.30	5.42
2. Weibull-Exponential	354.98	362.98	5.29	5.40
3. Weibull-Normal	353.32	363.32	4.99	5.09
4. Weibull-Lognormal	355.16	365.16	4.51	4.61
5. Normal-Exponential	359.12	367.12	5.82	5.94
6. Normal-Lognormal	355.44	365.44	4.50	4.60
7. Lognormal-Exponential	359.24	367.24	154.14	157.43

Here, the Weibull-Exponential mixture model contains the smallest AIC value and the Normal-Lognormal mixture model contains the smallest value of AD and Adjusted AD among all of the seven mixture models. Hence, we can say that, among these mixture models, Weibull-Exponential mixture model and Normal-Lognormal mixture model can be selected as the best two models for the data according to the value of AIC and AD test statistic, respectively for Windshield failure data.

The parameters of Weibull-Exponential, Normal-Lognormal and Weibull-Weibull mixture models, estimated by applying ML method are displayed in Table-2.

Table2:MLEs of the parameters

Mixture models	Parameters	MLE	Standard error
Weibull-Exponential	β1	2.6587	0.2569
	α1	3.4276	0.1374
	σ	4.6217	3.9682
	p	0.9855	0.0140
Normal-Lognormal	μ	0.2985	0.0940
	log μ	1.0514	0.0438
	σ	0.1809	0.0674
	log σ	0.4541	0.0349
	p	0.0256	0.0133
Weibull-Weibull	β1	1.1834	0.8007
	β2	2.6673	0.2538
	α1	0.2219	0.1647
	α2	3.4287	0.1367
	p	0.0149	0.0131

The probability distribution for a failure time T can be characterized by a cumulative distribution function, a probability density function, a survival/reliability function or a hazard function. We have estimated the CDF and $R(t)$ of Weibull-Weibull mixture model based on non-parametric and parametric approaches by using Kaplan–Meier (K-M) and maximum likelihood (ML) estimating methods, respectively. The CDF and $R(t)$ are also estimated by using WPP plot[estimates are taken from Murthy et al. (2004)]. Figure-1 represents the reliability function, to see either the WPP plot or the ML method gives the best result for the data set.

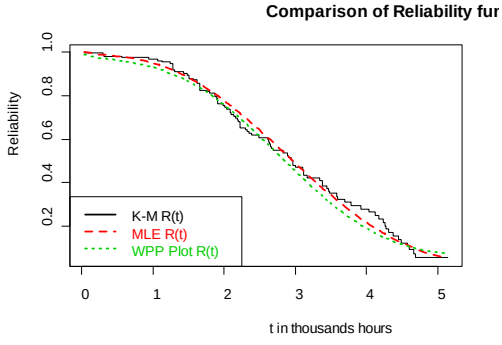


Figure1: Comparison of reliability functions of Weibull-Weibull mixture model based on K-M estimate, WPP plot and ML method for Windshield data

Figure 1 indicates that the reliability function obtained from the MLE is closer to the Kaplan-Meier estimate than that of the reliability function obtained from the WPP plot. So, we may say that, the maximum likelihood estimating procedure is much better than Weibull probability paper (WPP) plot procedure.

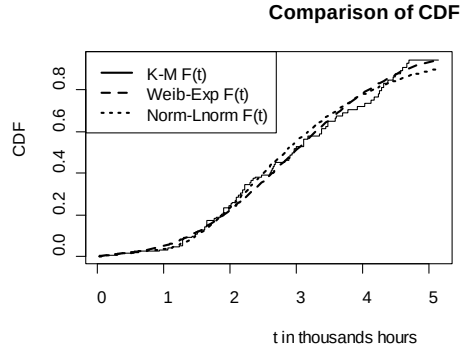


Figure2: Comparison of CDFs of Weibull-Exponential and Normal-Lognormal mixture models based on Kaplan-Meier and ML methods

Also the CDF of Weibull-Exponential and Normal-Lognormal mixture models, using K-M and ML estimating methods are estimated and the results are displayed in Figure-2 for a comparison. Figure 2 indicates that both of the CDFs based on Weibull-Exponential and Normal-Lognormal mixture models belong very closely to the nonparametric estimate of CDF. Hence we may consider both of the models for the data set.

We have also estimated the B10 life, median and B90 life of the CDF based on Kaplan-Meier estimate, ML method and WPP plot and displayed the result in Table-3.

Table3: Estimatesof B10 life, median and B90 life times (in thousand hours)

Estimates of	KM	ML	WPP
B10 life	1.432	1.432	1.262
B50 life or median	2.934	2.964	2.878
B90 life	4.570	4.663	4.694

From Table 3, we may conclude that, 10% of the total components fail at time 1.432 (in thousand hours) for K-M procedure, at time 1.432 for MLE method and at time 1.262 for WPP Plot method. 50% of the total components fail at time 2.934 for K-M procedure, at time 2.964 for MLE and at time 2.878 for WPP Plot. 90% of the total components fail at time 4.570 for K-M procedure, at time 4.663 for MLE and at time 4.694 for WPP Plot method. Hence we may say that the WPP plot under estimates the B10 life and the median life and over estimates the B90 life as compared tothe K-M and ML methods.

5 Conclusion

A set of seven competitive mixture models are applied to assess the reliability of Aircraft windshield based on failure and censored data. The conclusions on the analysis results are as follows:

- The Weibull-Exponential and Normal-Lognormal mixture models fit good for the Windshield failure data, based on the value of AIC and AD test statistic, respectively.
- Maximum likelihood estimate procedure gives much better fit than Weibull probability paper (WPP) plot procedure.
- For the data, the WPP plot method under estimate the B10 life and the median life and over estimates the B90 life as compared to K-M and ML methods.
- 10% of the total components fail at time 1.432 and 90% of the total components fail at time 4.663. These results would be useful for managerial implications for cost-benefit analysis, including improvement in reliability, reduction in warranty cost, and forecastingclaims rates and costs.
- This article analyzed a special type of product failure data. However, the proposed methods and models are also applicable to analyze lifetime data available in the fields, such as, Biostatistics, Medical science, Bio-informatics, etc.
- The article considered the first failure data of the product. If, there are repeated failures for any product, application of an approach of modeling repeated failures based on renewal function would be relevant.
- The Expectation-Maximization (EM) algorithm can be applied for estimating the parameters of the model and can be compared with the given method. Finally, further investigation on the properties of the methods and models by simulation study would be useful and the authors would like to show these results in the conference.

References

Blischke, W.R. and Murthy, D.N.P. (2000). *Reliability*, Wiley, New York.

Blischke, W.R., Karim, M.R. and Murthy, D.N.P. (2011). *Warranty data collection and analysis*, Springer.

Kalbfleisch J.D. and Prentice, R.L. (1980). The Statistical analysis of failure time data. John Wiley &Sons Inc., New York.

Karim, M.R. and Suzuki, K. (2005). Analysis of warranty claim data: a literature review, *International Journal of Quality & Reliability Management*, Vol. 22, No. 7, pp. 667-686.

Karim, M.R., Yamamoto, W. and Suzuki, K. (2001). Statistical analysis of marginal count failure data, *Lifetime Data Analysis*, Vol. 7, pp. 173-186.

Lawless, J.F. (1998). Statistical analysis of product warranty data, *International Statistical Review*, Vol. 66, pp. 41-60.

Lawless, J.F. (2003). Statistical methods for lifetime data. Wiley, New York.

Meeker, W.Q. and Escobar, L.A. (1998). Statistical methods for reliability data. Wiley, New York.

Murthy, D.N.P. and Djamaludin, I. (2002). New product warranty: a literature review, *International Journal of Production Economics*, Vol. 79, pp. 231-260.

Murthy, D.N.P., Xie, M., and Jiang, R., (2004). *Weibull Models*. Wiley, New York

Suzuki, K. (1985a). Non-parametric estimation of lifetime distribution from a record of failures and follow-ups, *Journal of the American Statistical Association*, Vol. 80, pp. 68-72.

Suzuki, K., Karim, M.R., and Wang, L. (2001). Statistical analysis of reliability warranty data, in N. Balakrishnan and C.R. Rao (Eds), *Handbook of Statistics:Advances in Reliability*, Elservier Science, Vol. 20, pp. 585-609.

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Competing Risk Model for Reliability Data Analysis

M. Rezaul Karim
Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh
Email: mrezakarim@yahoo.com

Abstract. The complexity of products has been increasing with technological advances. As a result, a product may fail in different ways or causes, which are commonly known as failure modes. Competing risk model is appropriate for modeling component failures with more than one failure modes. In this talk the competing risk model is applied for analyzing product reliability data with multiple failure modes. Maximum likelihood estimation method is used to estimate the model parameters and various characteristics of the modelto assess and predict the reliability of the product.

Keywords. Competing risk model; Exponential distribution; Weibull distribution; Reliability; MLE

1 Introduction

According to (ISO 8402, 1994), a product can be tangible (e.g. assemblies or processed materials) or intangible (e.g., knowledge or concepts), or a combination thereof. A product can be either intended (e.g., offering to customers) or unintended (e.g., pollutant or unwanted effects). Here we are concerned with tangible products, specifically manufactured goods.

The complexity of products has been increasing with technological advances. As a result, a product must be viewed as a system consisting of many elements and capable of decomposition into a hierarchy of levels, with the system at the top level and parts at the lowest level. There are many ways of describing this hierarchy. One such is the nine-level description shown in Table 1, based on a hierarchy given in Blischke and Murthy (2000) and Blischke, Karim and Murthy (2011).

Table 1: Multilevel decomposition of a product

Level	Characterization
0	System
1	Sub-system
3	Assembly
4	Sub-assembly
5	Module
6	Sub-module
7	Component
8	Part

The number of levels needed to describe a product from the system level down to the part level depends on the complexity of the product.

Many units, systems, subsystems, or components have more than one cause of failure. For example, (i) A capacitor can fail open or as a short, (ii) Any of many solder joints in a circuit board can fail, (iii) A semi conductor device can fail at a junction or at a lead, (iv) A device can fail because a manufacturing defect (infant mortality) or because of mechanical wear out, (v) For an automobile tire, tread can wear out or the tire may suffer a puncture, etc. The Competing risk model is appropriate for modeling component failures with more than one mode of failure. A failure mode is a description of a fault. It is sometimes referred to as fault mode.Failure modes are identified by studying the (performance) function.Assume a (replaceable) component or unit has K different ways it can fail. These are called failure modes and underlying each failure mode is a failure mechanism. Each mode is like a component in a series-system.

Improving reliability of product is an important part of the larger overall picture of improving product quality. Therefore, in recent years many manufacturers have collected and analyzed field failure data to enhance the quality and reliability of their products and to improve customer satisfaction. Here we apply the competing risk model to analyze product failure data and to assess and predict the reliability of the product.

The remainder of the article is organized as follows: Section 2 describes competing risk model formulation. Section 3 applies the competing risk model for analyzing a set of product failure data. Section 4 concludes the paper with additional implementation issues for further research.

2 Competing risk model formulation

A general *K*-fold competing risk model is given by

$$F(t) \equiv F(t;\theta) = 1 - \prod_{k=1}^{K} [1 - F_k(t;\theta_k)] \tag{1}$$

where $F_k(t) \equiv F_k(t;\theta_k)$ are the distribution functions of the *K* sub-populations with parameters $\theta_k, 1 \le k \le K$. Here $\theta \equiv \{\theta_k, 1 \le k \le K\}$ and we assume that $K \ge 2.$

This is called a "competing risk model" because it is applicable when an item (component or module) may fail by any one of *K* failure modes, i.e., it can fail due to any one of the *K* mutually exclusive causes in a set $\{C_1, C_2, \ldots, C_K\}$. The competing risk model has also been called the *compound model*, *series system model*, and *multi-risk model* in the reliability literature. Let T_k be a positive-valued continuous random variable denoting the time to failure if the item is exposed only to cause $C_k, 1 \le k \le K$. If the item is exposed to all *K* causes at the same time and the failure causes do not affect the probability of failure by any other mode, then the time to failure is the minimum of these *K* lifetimes, i.e., $T = \min\{T_1, T_2, \ldots, T_K\}$, which is also positive-valued, continuous random variable. Let $R(t)$, $h(t)$, and $H(t)$ denote the reliability, hazard, and cumulative hazard functions associated with $F(t)$, respectively, and let $R_k(t)$, $h_k(t)$, and $H_k(t)$ be the reliability function, hazard function and cumulative hazard function associated with the distribution function for $F_k(t)$, of the k^{th} failure mode, respectively. It is easily shown that

$$R(t) = \prod_{k=1}^{K} R_k(t) \tag{2}$$

$$H(t) = \sum_{k=1}^{K} H_k(t) \tag{3}$$

and

$$h(t) = \sum_{k=1}^{K} h_k(t) \tag{4}$$

Note that for independent failure modes, the reliability function for the item is the product of the reliability functions for individual failure modes and the hazard function for the item is the sum of the hazard functions. The density function of *T* is given by

$$f(t) = \sum_{k=1}^{K} \left\{ \prod_{\substack{j=1 \\ j \ne k}}^{K} [1 - F_j(t)] \right\} f_k(t) \tag{5}$$

which may be rewritten as

$$f(t) = R(t) \left\{ \sum_{k=1}^{K} \left[\frac{f_k(t)}{R_k(t)} \right] \right\} \tag{6}$$

Suppose that a component has *K* failure modes and that the failure modes are statistically independent. We look first at the general case in which the failure modes of some of the failed items are known and those of the remaining are

unknown. In addition, we assume that it is not possible to determine the failure modes (or causes of failure) for the censored (non-failed) items.

Two special cases of interest are as follows:

Case (i): The failure modes are known for all failed item.

Case (ii): The failure modes are unknown for all failed items.

Let n_1 be the number of failed units and n_2 the number of censored units. For the failed units, the post-mortem outcome is uncertain, that is, the failure modes for some units may not be known. Out of the n_1 failed items, let n_{1k} denote the number of items with failure mode $k, 1 \le k \le K$, and $n_{10} = n_1 - \sum_{k=1}^{K} n_{1k}$ the number of failures for which there is no information regarding the failure mode. Let t_{kj} denote the lifetime of the j^{th} item failing from failure mode k, and $\tilde{t}_i$ the i^{th} censoring time.

Note: For Case (i), $n_{10} = 0$, and for Case (ii) $n_{10} = n_1$.

For the general case, n_{1k} units out of n failed due to failure mode k, with failure times $\{t_{k1}, t_{k2}, \ldots, t_{kn_{1k}}\}$, and there are n_{10} units with failure times $\{t'_1, t'_2, \ldots, t'_{n_{10}}\}$ for which there is no information regarding the failure mode. In addition, there are $n_2 = n - \sum_{k=1}^{K} n_{1k} - n_{10}$ censored units, with censoring times $\{\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_{n_2}\}$. The likelihood function in the general case is given by

$$L(\theta) = \prod_{k=1}^{K}\left[\prod_{j=1}^{n_{1k}} f_k(t_{kj}) \prod_{\substack{l=1 \\ l\neq k}}^{K} R_l(t_{kj})\right] \times \prod_{k=1}^{K}\left[\prod_{j=1}^{n_{10}} f_k(t'_j) \prod_{\substack{l=1 \\ l\neq k}}^{K} R_l(t'_j)\right] \times \prod_{i=1}^{n_2}\prod_{k=1}^{K} R_k(\tilde{t}_i) \quad . \tag{7}$$

The MLEs of the parameters are obtained by maximizing the likelihood function (7). For most distributions the ML estimation method requires numerical maximization because of the lack of closed form solutions for the estimators.

The results for the two special cases are as follows:

Case (i): The expression for the likelihood function is given by (7) with the second term equal to unity, so that

$$L_1(\theta) = \prod_{k=1}^{K}\left[\prod_{j=1}^{n_{1k}} f_k(t_{kj}) \prod_{\substack{l=1 \\ l\neq k}}^{K} R_l(t_{kj})\right] \times \prod_{i=1}^{n_2}\prod_{k=1}^{K} R_k(\tilde{t}_i) \quad . \tag{8}$$

Case (ii): The expression for the likelihood function is given by (7) with the first term of equal to unity,

$$L_2(\theta) = \prod_{k=1}^{K}\left[\prod_{j=1}^{n_{10}} f_k(t'_j) \prod_{\substack{l=1 \\ l\neq k}}^{K} R_l(t'_j)\right] \times \prod_{i=1}^{n_2}\prod_{k=1}^{K} R_k(\tilde{t}_i) \quad . \tag{9}$$

The cause-specific (or failure mode-specific) hazard function for cause k can be written as

$$\tilde{h}_k(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t, C = k \mid T \ge t)}{\Delta t} = \frac{f_k(t)}{R(t)} \quad , \tag{10}$$

where $f_k(t)$ is the cause-specific PDF at time t that represents the unconditional probability of failure of an unit at time t from cause k, and $R(t)$ is the overall reliability function representing the probability of surviving from all causes up to time t. Relationship (10) implies that

$$f_k(t) = \tilde{h}_k(t)R(t) \quad (11)$$

Using (11) and (2), we can rewrite the likelihood functions (8) and (9), respectively as

$$L_1(\theta) = \prod_{k=1}^{K}\left[\prod_{j=1}^{n_{1k}} \tilde{h}_k(t_{kj})R(t_{kj})\right] \times \prod_{i=1}^{n_2} R(\tilde{t}_i) \quad (12)$$

and

$$L_2(\theta) = \prod_{k=1}^{K}\left[\prod_{j=1}^{n_{10}} \tilde{h}_k(t'_j)R(t'_j)\right] \times \prod_{i=1}^{n_2} R(\tilde{t}_i) \quad (13)$$

The MLEs of the parameters of the models are obtained by maximizing (8) or (12) for Case (i) and (9) or (13) for Case (ii). More details on the formulations and applications of mixture models can be found in Murthy, Xie, and Jiang(2004) and Blischke, Karim, and Murthy (2011).

3Examples

This section describes the following two examples.

3.1Exponential distribution

Suppose that $K = 2$, and the lifetimes of failure modes 1 and 2 independently follow exponential distributions with parameters λ_1 and λ_2, respectively. Time to failure is modeled by (1). We consider Case (i). The data consist of n units, with n_{11} units failing due to failure mode 1 with failure times $\{t_{11}, t_{12}, \ldots, t_{1n_{11}}\}$, n_{12} units failing due to failure mode 2 with failure times $\{t_{21}, t_{22}, \ldots, t_{2n_{12}}\}$, and $n_2 = n - n_{11} - n_{12}$ units censored, with censoring times $\{\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_{n_2}\}$.

In this case, from (2), we have $R(\tilde{t}) = R_1(\tilde{t})R_2(\tilde{t}) = \exp\left(-(\lambda_1 + \lambda_2)\tilde{t}\right)$ and using this in (12), the log-likelihood function becomes

$$\log L = n_{11}\log(\lambda_1) - (\lambda_1 + \lambda_2)\sum_{j=1}^{n_{11}} t_{1j} + n_{12}\log(\lambda_2) - (\lambda_1 + \lambda_2)\sum_{j=1}^{n_{12}} t_{2j} - (\lambda_1 + \lambda_2)\sum_{i=1}^{n_2} \tilde{t}_i \quad (14)$$

From this, the ML estimators of λ_1 and λ_2 are found to be

$$\hat{\lambda}_i = \frac{n_{1i}}{\sum_{j=1}^{n_{1j}} t_{1j} + \sum_{j=1}^{n_{2j}} t_{2j} + \sum_{i=1}^{n_2} \tilde{t}_i}, \; i = 1, 2 \quad (15)$$

It follows from (2) that the maximum likelihood estimate of the reliability function of the component is

$$\hat{R}(t) = \exp\left(-(\hat{\lambda}_1 + \hat{\lambda}_2)t\right), \; t \geq 0 \quad (16)$$

We consider an electronic component for which lifetimes follow an exponential distribution. The component exhibits a new mode of failure due to mounting problems. If incorrectly mounted, it can fail earlier, and this is also modeled by an exponential distribution. The parameters of the exponential distributions for failure modes 1 and 2 are $\lambda_1 = 0.0006$ and $\lambda_2 = 0.0004$ per day. From (16), the maximum likelihood estimate of the reliability function of the component is $\hat{R}(t) = \exp(-(0.0006+0.0004)t) = \exp(-0.001t), t \geq 0$.

Figure 1 displays a comparison of the estimated reliability functions for failure mode 1, failure mode 2 and combined failure modes 1 and 2 for $0 \leq t \leq 10000$ days.

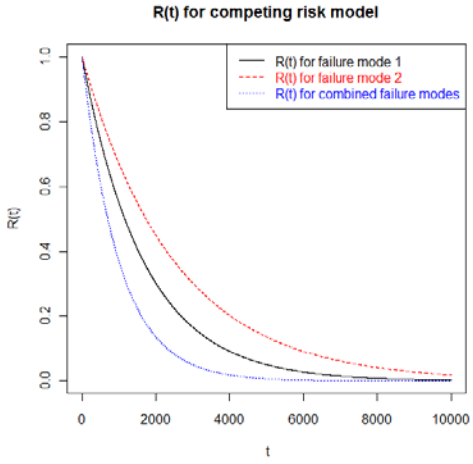


Figure 1: Comparison of ML estimates of reliability functions for competing risk model

This figure can be used to assess reliability of the component for given days. For example, the figure indicates the reliabilities of the component at age 2000 days are 0.30 for failure mode 1, 0.45 for failure mode 2 and 0.14 for the combined failure modes. Based on (16), the estimated MTTF of the component is found to be $\hat{\mu} = \int_0^\infty \hat{R}(t)dt$ = 1/($\hat{\lambda}_1 + \hat{\lambda}_2$) = 1000 days.

3.2 Device-G Data

Failure times and running times for a sample of devices from a field tracking study of a larger system are given in Meeker and Escobar (1998). 30 units were installed in typical service environments. Cause of failure information was determined for each unit that failed (lifetime in thousand cycles of use). Mode S failures were caused by failures on an electronic component due to electrical surge. These failures predominated early in life. Mode W failures, caused by normal product wear, began to appear after 100 thousand cycles of use. The purposes of the analyses are:

- Analyze the failure modes separately to investigate the effects of failure modes.
- How to improve product reliability – if one failure mode can be eliminated.
- Compare lifetime (with respect to the MLEs of parameters, MTTF, B10 life, median life, etc.) of the product with failure modes (competing risk model) and ignoring failure mode information.

When the failure modes S and W act independently, one can:

- Analyze the mode S failures only: In this case mode W failures are treated as right censored observations. This is the estimate of the failure-time distribution if mode W could be completely eliminated.
- Analysis of the mode W failures only: In this case mode S failures are treated as right censored observations. This is the estimate of the failure-time distribution if mode S could be completely eliminated.
- A combined analysis use the competing risk model assuming independence between mode S and mode W.

Out of 30 units, there are 8 censored units at censoring time 300 kilocycles. A preliminary analysis of failure modes are given in Table 2. It is an examination of failure mode frequency or relative frequency data to determine the most important failure modes that contribute to quality problems and to which quality improvement efforts should be directed.

Table 2: Frequencies and average lifetimes for failure modes S and W

Failure Mode	Frequency	Average Life (Failure only)
S	15	86.1
W	7	231.3

Table 2 indicates that failure mode S has considerably higher frequency and lower average lifetime (based on failure data only). Therefore, we may conclude that efforts should be concentrated on failure mode S to eliminate it or to reduce the risks associated with this failure mode. Figure 2 represents the Weibull probability plots for individual failure modes S and W with the MLEs of shape and scale parameters. This figure suggests that the Weibull distribution provides a good fit to both failure modes.

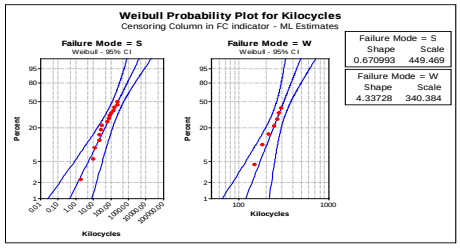


Figure 2: The Weibull probability plots for individual failure modes S and W

The maximum likelihood estimates of Weibull parameters with MTTFs for failure modes S and W are displayed in Table 3 and Table 4, respectively.

Table 3: Maximum likelihood estimates of Weibull parameters for failure mode S

Parameters and MTTF	Estimate	Standard Error	95.0% Normal CI	
			Lower	Upper
Shape	0.670993	0.157777	0.423221	1.06382
Scale	449.469	191.944	194.625	1038.01
Mean(MTTF)	593.462	342.422	191.539	1838.77

Table 4: Maximum likelihood estimates of Weibull parameters for failure mode W

Parameters and MTTF	Estimate	Standard Error	95.0% Normal CI	
			Lower	Upper
Shape	4.33728	1.45059	2.25183	8.35411
Scale	340.384	36.139	276.437	419.124
Mean(MTTF)	309.963	29.8906	256.582	374.45

Tables 3 and 4 indicate that for the failure mode W, the MLEs of shape parameter is much larger and the MTTF is smaller than that of the failure mode S. The estimates of MTTFs of Tables 3 and 4 indicate a contradiction to the conditional average lifetimes given in Table 2.

Figure 3 represents the Weibull probability plots for individual failure modes in the same scale. It suggests that the mode S failures predominated early in life whereas the mode W failures caused by normal product wear and began to appear after 100 thousand cycles of use.

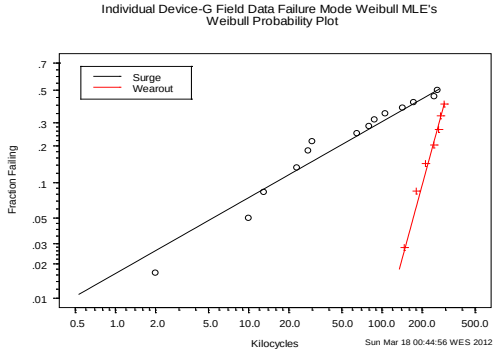


Figure 3: Weibull probability plots for individual failure modes in the same scale

Figure 4 shows the Weibull probability plot for competing risk model. This figure diverges rapidly after 200 thousand cycles.

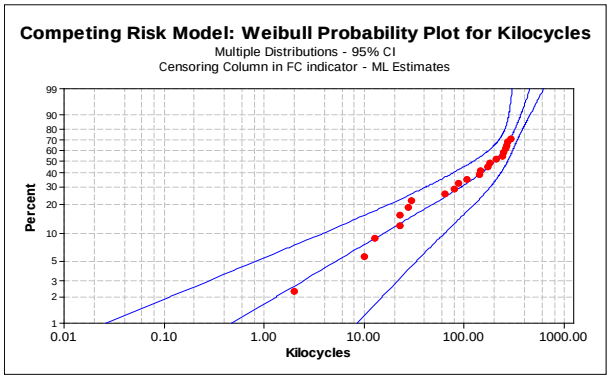


Figure 4: Weibull probability plot for competing risk model

The Weibull probability plot (ignoring failure mode information) is shown in Figure 5. Weibull analysis ignoring the failure mode information (Figure 5) shows evidence of a change in the slope of the plotted points, indicating a gradual shift from one failure mode to another.

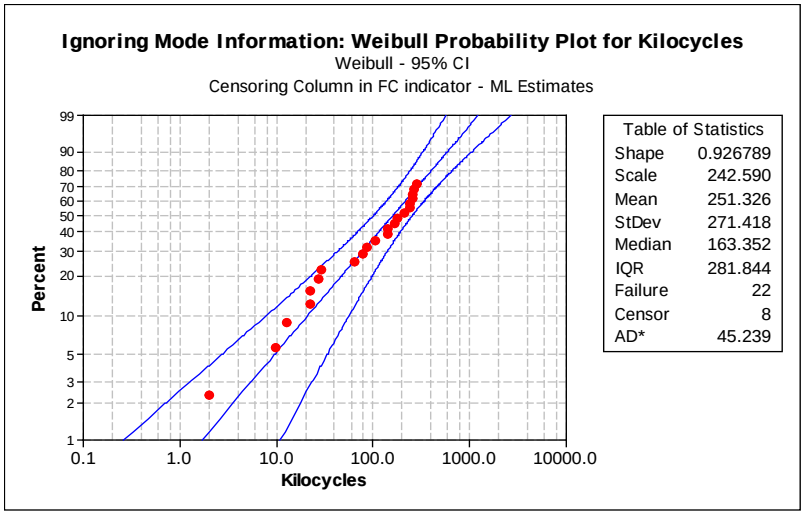


Figure 5: Weibull probability plot (ignoring failure mode information)

Maximum likelihood estimates of percentiles for both competing risk model and ignoring failure mode information are given in Table 5.From Table 5, we may conclude that, 10% of the total components fail at 15.71 kilocycles under competing risk model and at 21.4kilocycles under ignoring failure mode information. 50% of the total components fail at 203.06 kilocycles for competing risk model and at 163.35kilocycles for without failure mode information. Hence we may say that ignoring failure mode information over estimates the B10 life and B90 life and under estimates median life compared with the competing risk model.

Table 5: MLEs of percentiles for competing risk model and ignoring failure mode information

Percentile	**Competing Risk Model**			**Ignoring Mode Information**		
	Estimate	**95% L-CI**	**95% U-CI**	**Estimate**	**95% L-CI**	**95% U-CI**
5	5.37	0.85	33.78	9.84	2.81	34.44
10	15.71	3.86	63.63	21.4	7.97	57.43
50	203.06	124.25	273.72	163.35	102.47	260.4
90	369.4	280.7	455.89	596.63	334.03	1065.67

4 Conclusion

- The failure mode-wise frequencies and conditional mean lifetimes can be misleading to determine the most important failure modes that contribute to quality problems and to which quality improvement efforts should be directed.
- The failure mode or failure cause wise model with competing risk is better than combined model for assessing and predicting reliability of the product.
- This article analyzed the failure data based on Case (i), where the failure modes are known for all failed items. If, the failure modes are unknown for all failed items, application of the likelihood derived under Case (ii) would be relevant. However, it requires a complicated numerical maximization technique. The Expectation-Maximization (EM) algorithm might be applied in Case (ii). Further investigation on that case would be useful.

References

Blischke, W.R. and Murthy, D.N.P. (2000). *Reliability*, Wiley, New York.

Blischke, W.R., Karim, M.R. and Murthy, D.N.P. (2011). *Warranty data collection and analysis*, Springer.

ISO 8402 (1994) *Quality Vocabulary*. International Standards Organization, Geneva.

Meeker, Q. and Escobar, L.A. (1998): *Statistical Methods for Reliability Data*, John Wiley & Sons, Inc.

Murthy, D.N.P., Xie, M., and Jiang, R., (2004). *Weibull Models*. Wiley, New York.

Contributed Paper

- **Sample Survey**

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

In Sample and Out of Sample Forecasting Performance under Fat Tail and Skewed Distribution

Md. Mostafizur Rahman[2] Md. Monimul Huq[2] and M. Sayedur Rahman[2]

[1] Department of Statistics, Statistics and MathematicsSchool, YunnanUniversity of Finance and Economics, Kunming-650221, P.R. China. Email: mostafiz_bd21@yahoo.com

[2] Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh.

Abstract. The aim of this paper is to empirically investigate the in-sample and out-of-sample forecasting performance of GARCH, EGARCH and APARCH models under fat tail and skewed distributions. In our study we consider two fat tail distributions (Student-t and generalized error) and two skewed distributions (Skewed Student-t and Skewed GED) in case of Dhaka Stock Exchange (DSE) from the period January 02, 1999 to December 29, 2005 for a total of 1861 observations. First 1661 observations are taken for in-sample estimation and last 200 observations are taken for out-of-sample forecasting performance. Our empirical study suggests that in case of in-sample estimation, APARCH model with fat tail distribution gives better result than Skewed distribution and Gaussian distribution. Between these two fat tail distributions, Student-t gives better result. In case of out-of-sample forecasting performance we also found that APARCH model with Student-t distribution gives the lowest value of loss function. The superiority of out-of-sample volatility predictive performance of these models is verified by using SPA test of Hansen's (2005). These results also indicate that APARCH model with Student-t distribution is the most suitable model for both in-sample and out-of-sample forecasting performance in case of DhakaStock Exchange, Bangladesh.

Keywords: GARCH model, Fat tail distributions, Skewed distributions, Loss function, SPA test, DSE.

1.Introduction

It has been recognized that financial time series exhibits a changes in volatility over time that tend to be serially correlated. In particular, financial returns demonstrate volatility clustering, meaning that large changes tend to be followed by large changes and vice versa. The AutoRegressive Conditional Heteroscedasticity (ARCH) model introduced by Engle (1982) plays a milestone role for the studies of volatility because it can capture some of the empirical regularities of equity returns, such as volatility clustering and thick-tail ness. Subsequently, Bollerslev (1986) presented a Generalized ARCH (GARCH) model that is a parsimonious parameterization for the conditional variance. But GARCH model sometime fail to capture the thick tail property. This excess kurtosis has naturally led to the use of non normal distributions in to GARCH model. Bollerslev (1987), Baillie and Bollerslev (1989), Kaiser (1996) and Beine et al (2002) among others used Student-t distribution while Nelson (1991) and Kaiser (1996) suggested the Generalized Error Distribution (GED).

The existing literature has long recognized that the distribution of returns can be skewed. For instance, for some stock market indexes, returns are skewed toward the left, indicating that there are more negative than positive outlying observations. The intrinsically symmetric distribution such as normal, Student-t and generalized error distribution (GED) cannot cope with such skewness. As a result we have to use some skewed distributions such as Skewed Student-t and Skewed GED. Fernandez and Steel (1998) proposed Skewed Student-t distribution and Theodossiou (2000) proposed skewed generalized error distribution. Later, Lambert and Laurent (2000, 2001) used Skewed Student-t distribution in to GARCH framework. Peters (2001) examined the forecasting performance of GARCH, Exponential GARCH (EGARCH), Glosten-Jagannathan-Runkle (GJR) and Asymmetric Power ARCH (APARCH) models with three distributions (normal, Student-t and Skewed Student-t) for FTSE 100 and DAX 30 index and found that asymmetric GARCH models give better results than symmetric GARCH model and forecasting performance was not clear when using non normal densities.

Liu et al (2009) investigated the specification of return distribution influences the performance of volatility forecasting for two Chinese Stock Indexes using two GARCH models and found that GARCH model with skewed generalized error distribution give better results than GARCH model with normal distribution. Rahman et al (2008) examined a wide variety of popular volatility models with normal, Student-t and GED distributional assumption for Chittagong Stock Exchange (CSE) and found that Random Walk GARCH (RW-GARCH), Random Walk Threshold GARCH (RW-TGARCH), Random Walk Exponential GARCH (RW-EGARCH) and Random Walk Asymmetric

Power ARCH (RW-APARCH) models under Student-t distributional assumptions are suitable for CSE. Hasan et al (2004) investigated the return behavior of Dhaka Stock Exchange (DSE) within a GARCH-M framework. Rahman (2005) in his Ph.D. thesis discussed the different kinds of volatility modeling under normal and Student-t distributional assumptions for Dhaka Stock Exchange by non-parametric specification test. The earlier study indicate that only few researches have been conducted in case of DSE and none of the earlier studies estimate various GARCH models under fat tail and skewed distributions together.

On the other hand, accurate forecasting of volatility and correlations of financial asset returns is essential for optimal allocation and managing portfolio risk. For this reason to find out that accurate forecasting model for any particular stock market is always interesting for researchers because different model fit well for different stock markets. To make forecasting evaluation, we take great care in employing loss functions that are robust and lead to an optimal forecast. We use Hansen's (2005) Superior Predictive Ability (SPA) test to evaluate the forecasting performance across all volatility models. Therefore, the aim of this paper is to empirically examine the in-sample and out-of-sample forecasting performance of several GARCH-type models such as GARCH, EGARCH and APARCH model with Gaussian, Student-t, Generalized error, Skewed Student-t and Skewed GED densities by using different loss function and SPA test of Hansen (2005) in case of DSE. This study is important because this is the first time to compare the performance of different GARCH type models under fat tail and skewed distributional assumptions together by SPA test in case of DSE. Rest of the paper is organized as follows: Section 2 presents the methodology, Section 3 presents the forecasting method, Section 4 presents the empirical study and finally Section 5 present the conclusions.

2. Methodology

2.1 Models

2.1.1 Generalized ARCH (GARCH) model

Bollerslev (1986) introduced a conditional heteroscedasticity model that includes lags of the conditional variance as regressors in the model for the conditional variance in addition to the lags of the squared error term. The GARCH(p,q) model can be expressed as

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=i}^{p} \beta_j \sigma_{t-j}^2 \tag{1}$$

Using the lag or backshift operator L, the GARCH (p, q) model is

$$\sigma_t^2 = \alpha_0 + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2 \tag{2}$$

where $p \geq 0, q > 0; \alpha_0 > 0, \alpha_i \geq 0 (i = 1,2,..., q)$ and $\beta_j (j = 1,2,..., p)$. The term $\alpha(L) = \alpha_1 L + \alpha_2 L^2 + ... + \alpha_q L^q$ and $\beta(L) = \beta_1 L + \beta_2 L^2 + ... + \beta_p L^p$ based on equation (2), it is straightforward to show that the GARCH model is based on an infinite ARCH specification.

2.1.2 Exponential GARCH (EGARCH) Model

The EGARCH or Exponential GARCH model was proposed by Nelson (1991). The specification for conditional variance is:

$$\log(\sigma_t^2) = \alpha_0 + \sum_{j=1}^{q} \beta_j \log(\sigma_{t-j}^2) + \sum_{i=1}^{p} \alpha_i \left| \frac{\varepsilon_{t-i}}{\sigma_{t-i}} \right| + \sum_{k=1}^{r} \gamma_k \frac{\varepsilon_{t-k}}{\sigma_{t-k}} \tag{3}$$

The left hand side of equation (3) is the log of the conditional variance. This implies that the leverage effect is exponential rather than quadratic and that forecasts of the conditional variance are guaranteed to be non-negative. The presence of the leverage effects can be tested by the hypothesis $\gamma_i < 0$. The impact is asymmetric if $\gamma_i \neq 0$.

2.1.3 Asymmetric Power ARCH (APARCH) Model

Taylor (1986) and Schwert (1990) introduced the standard deviation of GARCH model, where the standard

deviation is modeled rather than the variance. Ding, Granger and Engle (1993) introduced the Asymmetric Power ARCH (APARCH) model. The APARCH (p, q) model can be expressed as:

$$\sigma_t^{\delta} = \alpha_0 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^{\delta} + \sum_{i=1}^{q} \alpha_i \left(|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i} \right)^{\delta} \tag{4}$$

where $\alpha_0 > 0$, $\delta \geq 0$, $\beta_j \geq 0 (j = 1,2,\ldots\ldots, p)$, $\alpha_i \geq 0$ and $-1 < \gamma_i < 1, (i = 1,2,\ldots, q)$.

This model is quite interesting since it couples the flexibility of a varying exponent with the asymmetry coefficient (to take the "leverage effect" into account).

2.2 Distributional Assumptions

To complete the basic ARCH specification, we require an assumption about the conditional distribution of the error term. Since it may be expected that excess kurtosis and skewness displayed by the residuals of conditional heteroscedasticity models will be reduced when an appropriate distributions is used. In our study we used Gaussian distribution, two fat tail distributions and two skewed distributions.

Most applications of the GARCH models used the Gaussian distributional assumption for the errors. The normal or Gaussian distribution is a symmetric distribution with density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/\sigma^2} \tag{5}$$

where μ is the mean value and σ^2 is the variance of the stochastic variable. The standard normal distribution is given by taking $\mu = 0$ and $\sigma^2 = 1$. The unconditional distribution of GARCH model with error term which is conditionally normal and leptokurtic. But this leptokurtosis is not enough to explain the leptokurtosis which is found in most of the financial data (Bollerslev, 1987). Therefore, one should take this into account and use conditionally leptokurtic distribution for the error.

One alternative possibility is Student-t distribution. The density function of Student-t is given by

$$f(x) = \frac{\Gamma\left((v+1)/2\right)}{\sqrt{v\pi}[v/2](1 + x^2/v)^{(v+1)/2}} \tag{6}$$

where v is the degrees of freedom (df) $v > 2$. If v tends to ∞ the Student-t distribution converges to a normal distribution.

Another generalization of the normal distribution is the Generalized Error distribution (GED). The GED is a symmetric distribution and platykurtic depending on the degree of freedom. The GED has the following density function

$$f(x) = \frac{v e^{\frac{1}{2}\left|\frac{x}{\lambda}\right|}}{\lambda 2^{(v+1)/v} \Gamma\left(1/v\right)} \tag{7}$$

where $\lambda = \left[\frac{2^{-2/v} \Gamma\left(1/v\right)}{\Gamma\left(3/v\right)} \right]^{1/2}$

It includes the normal distribution if the parameter v has the value 2. For df v <2 indicate fat tail distribution.

The main drawback of fat tail density is that it is symmetrical while financial time series can be skewed. Fernandez and Steel (1998) proposed an extension of the Student-t distribution by adding skewness parameter. Chen and Hong (2009) used the following form of Skewed Student-t distribution

$$f(x)=\begin{cases}\dfrac{AP}{\left[1+\dfrac{1}{\nu-2}\left(\dfrac{A\varepsilon+B}{1-\xi}\right)^2\right]^{(\nu+1)/2}} & if\ \varepsilon<-\dfrac{B}{A}\\ \dfrac{AP}{\left[1+\dfrac{1}{\nu-2}\left(\dfrac{A\varepsilon+B}{1+\xi}\right)^2\right]^{(\nu+1)/2}} & if\ \varepsilon\geq-\dfrac{B}{A}\end{cases} \tag{8}$$

where $0<\nu<\infty,\ -1<\xi<1,\ B=4\xi P\dfrac{\nu-2}{\nu-1},\ A=\sqrt{1+3\xi^2-B^2}$ and $P=\dfrac{\Gamma\left[(\nu+1)/2\right]}{[\pi(\nu-2)]^{1/2}\Gamma(\nu/2)}$

Another skewed distribution advocated by Theodossiou (2000) is skewed generalized error distribution. The density function of Skewed GED distribution is given by

$$f(x)=C.\exp\left(-\frac{\left|x_t-\delta\right|^{\nu}}{[1-sign(x_t-\delta)\xi]^{\nu}\theta^{\nu}}\right) \tag{9}$$

where $C=\nu\left(2\theta.\Gamma(1/\nu)\right)^{-1}$, $\theta=\Gamma(1/\nu)^{1/2}[(3/\nu)^{-1/2}.S(\xi)^{-1}]$, $\delta=2\xi A'S(\xi)^{-1}$, $S(\xi)=\sqrt{1+3\xi^2-4A^2\xi^2}$ and $A=\left[\Gamma(2/\nu)\Gamma(1/\nu)^{-1/2}\Gamma(3/\nu)^{-1/2}\right]$

where the shape parameter ν indicate height and fat tails and skewness parameter is ξ.

3. Forecasting Method

For evaluating the forecasting performance of each model we need to minimize the loss function. These loss functions are Mean Squared Error (MSE) and Mean Absolute Error (MAE), Heteroskedasticity-adjusted MSE and MAE, Logarithmic Loss (LL) function and QLIKE loss function. Patton (2011) analytically showed that the use of many loss functions can cause severe distortions when used with standard volatility proxies including squared returns, so he suggested two loss functions which give rise to optimal forecast and these are MSE and QLIKE. Hou and Suasdi (2012) used SPA method of Hansen (2005) with three loss function such as MSE, QLIKE and R^2_{vol} (total variation in the true volatility). So, at our study we use only two loss functions which are given below

$$MSE=\frac{1}{n}\sum_{t=1}^{n}(\sigma_t^2-\hat{\sigma}_t^2) \tag{10}$$

and the QLIKE which corresponds to the loss implied by a Gaussian likelihood is given by:

$$QLIKE=\frac{1}{n}\sum_{t=1}^{n}\left(\log\hat{\sigma}_t^2+\frac{\sigma_t^2}{\hat{\sigma}_t^2}\right) \tag{11}$$

Hansen (2005) proposed Superior Predictive Ability (SPA) test to find out the best performing forecasting models. So later we applied SPA test for our empirical study.

Hansen (2005) applied a Supremum over the standardized performances and tests the null hypothesis

$$H_0:\max_{k=1,2,\ldots m}\mu_k\leq 0 \quad \text{where } \mu_k=E(f_{k,t})$$

And the test statistics

$$T^S=\max\left\{\left(\max_k\frac{n^{1/2}\bar{f}_k}{\hat{\sigma}_k}\right),0\right\} \tag{12}$$

where $\hat{\sigma}_k$ is the standard deviation of $n^{1/2}\bar{f}_k$, then $\bar{f}_k = 1/n\sum_{t=1}^{n} f_{k,t}$ and $f_{k,t} = l_{k,t} - l_{0,t}$. Here $l_{0,t}$ is the value of the pre-specified loss function at time t for a benchmark model while $l_{k,t}$ is the value of the corresponding loss function for another competing model, n is the number of out of sample data. To reduce the influence of poor performing models while preserving the influence of the alternatives with $\mu_k = 0$, Hansen proposes the following consistent estimator for μ:

$$\hat{\mu}_k^c = \bar{f}_k 1_{\left\{n^{1/2}\hat{f}_k/\hat{\sigma}_k \geq -\sqrt{2\ln\ln n}\right\}}, k=1,...,m \tag{13}$$

where $1_{\{\cdot\}}$ is an indicator function. Hansen argues that the threshold rate $\sqrt{2\ln\ln n}$ ensures $\hat{\mu}_k^c$ is consistent estimator that effectively captures all alternative with $\mu_k = 0$, and thus leads to a consistent estimate of the null distribution, which improves the power of the test. The distribution of the test statistic under the null hypothesis can be approximated by the empirical distribution derived from the bootstrap resample based on the stationary bootstrap of Politis and Romano (1994).

$$U_{k,b,t}^* = f_{k,b,t}^* - h(\bar{f}_k) \text{ for } b = 1,...,B \text{ and } t = 1,...,n \tag{14}$$

where $h(\hat{f}_k) = \bar{f}_k 1_{\left\{n^{1/2}\hat{f}_k/\hat{\sigma}_k \geq -\sqrt{2\ln\ln n}\right\}}$. The p-value of the SPA test can be obtained by first calculating

$$T_b^{S*} = \max\left\{\left(\max_k \frac{n^{1/2}\bar{f}_k}{\hat{\sigma}_k}\right), 0\right\} \tag{15}$$

for each $b = 1,...,B$ and then the comparing T^S to the quantiles of T_b^{S*}.

$$P^S = \sum_{b=1}^{B} \frac{1_{\left\{T_b^{S*} > T^S\right\}}}{B} \tag{16}$$

4. Empirical Study

4.1 Data and In-Sample Estimation

In this section, we describe the data and in-sample estimation. The data we analyze is the daily closing price index for Dhaka Stock Exchange (DSE) in Bangladesh from January 02, 1999 toDecember 29, 2005 for a total of 1861 observations. Here we consider first1661 observations for in-sample estimation and last 200 observations for out-of-sample forecasting performance. The analysis is done by using EViews5.0 and MATLAB 7.0. The parameter estimation process we choice here MLE. The indices prices are transformed into their returns so that we obtain stationary series. The transformation is

$$r_t = 100 * [\ln(y_t) - \ln(y_{t-1})] \tag{17}$$

where y_t is the return index at time t. Daily returns of DSE are plotted at Figure 1.

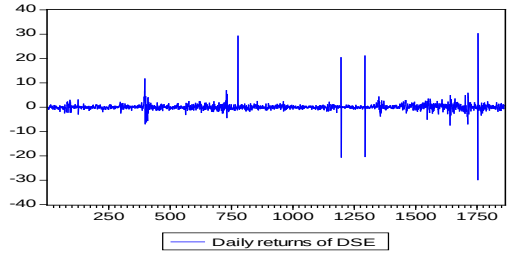


Figure 1: Daily returns of DSE

Some of the descriptive statistics of daily returns are displayed in Table-1. From Table-1 we found that mean return of the DSE is 0.0609. Volatility (measured as standard deviation) is 1.8553. The daily returns of DSE markets are leptokurtic in the sense that kurtosis exceeds positive three and the return series also display positive skewness. According to Jarque-Berra statistics normality is rejected for the return series. We found from Engle (1982) LM test which indicate the present of ARCH effect. Overall these results clearly support the rejection of the hypothesis that DSE time series of daily returns are time series with independent daily values. Parameter estimation results for different GARCH models such as GARCH, EGARCH and APARCH with Gaussian, Student-t and GED, Skewed Student-t and Skewed GED distributional assumptions are given at Table-2. In comparison the performance of GARCH models we use the simple mean equation: $y_t = \mu + \varepsilon_t$ for all models.

Table 1: Descriptive statistics of DSE

Sample size	Mean	Min.	Max.	SD	Skewness	Kurtosis	Jarque-Bera test	ARCH(2) test
1861	0.0609	-29.790	30.228	1.8553	2.2834	141.127	1480259 (0.000)	674.245

Table 2: Parameter estimation and their corresponding t statistics

Parameters		μ	α_0	α_i	β_j	γ_k (or γ_i)	υ	ξ	δ
GARCH	Gaussian	0.0605	5.1015	0.2094	-0.0269				
		(1.033)	(117.36)	(5.326)	(-14.29)				
	Student-t	0.0071	0.1591	0.6544	0.3038		3.7297		
		(0.494)	(7.610)	(8.278)	(10.134)		(19.638)		
	GED	0.0126	0.2792	0.7015	0.2160		0.2760		
		(1.363)	(8.583)	(7.180)	(5.807)		(63.339)		
	SK-t	0.0059	0.1311	0.7045	0.2714		3.116	-0.068	
		(0.421)	(6.641)	(8.978)	(9.224)		(17.836)	(-1.102)	
	SK-GED	0.0095	0.3142	0.6545	0.2455		0.2512	-0.541	
		(1.112)	(8.651)	(6.985)	(5.612)		(60.145)	(-1.112)	
EGARCH	Gaussian	0.0659	0.3092	0.4616	0.3965	-0.0555			
		(2.683)	(11.725)	(12.264)	(8.299)	(-2.278)			
	Student-t	0.0150	-0.2854	0.3015	0.7042	-0.0772	3.3509		
		(0.991)	(-10.09)	(12.274)	(27.736)	(-4.004)	(18.168)		
	GED	0.0163	-0.2670	0.1342	0.6715	-0.0220	0.7931		
		(1.527)	(-8.886)	(11.025)	(15.716)	(-3.428)	(60.348)		
	SK-t	0.0112	-0.3061	0.2554	0.6661	-0.0720	3.018	0.1741	
		(1.147)	(-9.548)	(10.145)	(25.321)	(-3.854)	(17.245)	(2.147)	
	SK-GED	0.0134	-0.2498	0.1570	0.6011	-0.0187	0.7354	0.0871	
		(1.508)	(-8.001)	(11.451)	(14.784)	(-2.965)	(57.146)	(2.114)	
APARCH	Gaussian	-0.012	0.8042	0.3009	-0.067	0.2027			0.0723
		(-2.44)	(193.34)	(82.746)	(-23.43)	(40.910)			(9.656)
	Student-t	0.0124	0.1595	0.4284	0.5422	-0.0122	3.6750		0.8419
		(3.189)	(7.216)	(10.817)	(14.491)	(-3.221)	(19.410)		(8.501)
	GED	0.0229	0.2629	0.4814	0.4389	-0.0063	0.7959		0.7764
		(7.460)	(7.897)	(7.772)	(7.829)	(-2.115)	(56.259)		(64.991)
	SK-t	0.0138	0.1441	0.3842	0.5562	-0.0141	3.2140	0.2125	0.8124
		(3.046)	(6.998)	(10.124)	(15.145)	(-3.012)	(18.264)	(2.847)	(7.485)
	SK-GED	0.0201	0.2500	0.5249	0.4012	-0.0054	0.6047	0.1421	0.6478
		(6.784)	(7.312)	(6.381)	(7.141)	(-2.874)	(52.569)	(2.664)	(57.211)

From Table-2 we found that most of the parameters are significant at 5% level of significance except the parameter μ for GARCH model with all distributional assumption and for EGARCH model with Student-t, GED, Skewed Student-t and Skewed GED distributional assumption. The t-statistics also suggest that the skewness parameters are insignificant in case of GARCH model. From this Table we also found that the sum of GARCH parameter estimates $\alpha_i + \beta_j$ is less than 1 which suggest that the volatility are limited and the data is stationary that's explain that the models are well fitted. Some model comparison criteria such as such as Box-Pierce statistic for both the residual and the squared residual, Akaike Information Criteria (AIC) and Log Likelihood value are given at Table-3.

Table 3: Model comparison criteria

	Model	$Q(20)$	$Q^2(20)$	AIC	Log-likelihood
Gaussian	GARCH	20.8040	7.9412	3.9589	-3677.804
	EGARCH	16.8660	6.2675	3.7024	-3437.240
	APARCH	28.5543	5.3582	3.7020	-3436.369
Student-t	GARCH	8.6162	3.2620	2.4594	-2281.202
	kEGARCH	11.3974	4.3191	2.5136	-2331.671
	APARCH	8.0667	3.2447	2.4443	-2266.289
GED	GARCH	9.2044	4.2736	2.5790	-2393.388
	EGARCH	13.1050	7.3160	2.6610	-2422.257
	APARCH	9.5411	4.2863	2.5659	-2379.356
Skewed-t	GARCH	10.5423	5.4712	2.6214	-2468.241
	EGARCH	13.8745	7.8901	2.7847	-2758.784
	APARCH	9.8654	6.1270	2.6042	-2390.421
Skewed-GED	GARCH	11.2483	5.6871	2.6632	-2510.326
	EGARCH	12.8792	8.1145	2.8567	-2762.341
	APARCH	11.2477	5.8791	2.6798	-2475.245

From Table-3 we found that all the models seem to do a good job in describing the dynamic of the first two moments of the series as shown by the Box-Pierce statistics for the residual and the squared residual which are all non-significant at 5% level. The Akaike Information Criteria (AIC) and log-likelihood values suggest that APARCH model give better results than GARCH and EGARCH models under all distributional assumption. Among these models EGARCH model showed the worst performance in case of DSE. Regarding the densities, fat tail densities such as Student-t and generalized error clearly outperform than the skewed and Gaussian densities. Among these densities Student-t density shows smallest AIC value for all the models than other distributions. The log likelihood value is strictly increasing in case of fat tail distributions and skewed distributions. But the log likelihood value under Student-t and GED distributions are higher than Skewed Student-t and Skewed GED distribution. Among all of these distributions Student-t distribution gives the highest log likelihood value. So, finally from Table-3 we found that in case of in sample estimation APARCH model with Student-t distribution gives better results than other models in case of DSE, Bangladesh.

4.2 Out-of-Sample Forecasting Performance

An out-of-sample test has the ability to control either possible over fitting or over parameterization problems and give more powerful framework to evaluate the performance of competing models. Most of the researchers are interested in having good volatility forecasts rather than good in-sample fit. In our study we use out-of-sample evaluation of one step ahead volatility forecast according to the loss function MSE and QLIKE. To better assess the forecasting performance of the various models we use the Superior Predictive Ability (SPA) test of Hansen (2005). The estimation results are given at Table-4.

Table 4: Forecasting evaluation criteria (MSE, QLIKE and SPA test)

	Model	MSE	QLIKE	SPA(p-value)	
				MSE	QLIKE
Gaussian	GARCH	1.2142	0.9945	0.021	0.121
	EGARCH	1.2188	0.9978	0.133	0.067
	APARCH	1.2101	0.9935	0.489	0.310
Student-t	GARCH	1.1972	0.9722	0.643	0.587
	EGARCH	1.1995	0.9756	0.342	0.442
	APARCH	1.1956	0.9707	0.906	0.967
GED	GARCH	1.2023	0.9794	0.012	0.124
	EGARCH	1.2079	0.9804	0.156	0.221
	APARCH	1.2012	0.9786	0.303	0.412
Skewed-t	GARCH	1.2046	0.9812	0.512	0.627
	EGARCH	1.2084	0.9834	0.210	0.313
	APARCH	1.2030	0.9821	0.687	0.597
Skewed-GED	GARCH	1.2059	0.9844	0.112	0.227
	EGARCH	1.2099	0.9868	0.274	0.124
	APARCH	1.2048	0.9834	0.671	0.354

From Table-4 we found that APARCH model with all distributional assumptions give the lowest value of MSE and QLIKE where EGARCH model with all distributional assumptions provides the poorest forecasting performance. The comparison among densities suggests that the model under fat tail and skewed densities give better results than Gaussian density and adding skewed distributions are not improving the forecasting performance than fat tail distributions. Among these distributions Student-t distribution gives the lowest MSE and QLIKE. In this Table we only reported the p-values of the SPA test under MSE and QLIKE loss function. Under the null hypothesis the base model is not outperformed by all of the other models, the higher p-values indicate the superiority of the forecasting performance. The p-value for the APARCH model with Student-t density are 0.906 and 0.967 for MSE and QLIKE loss function respectively which are virtually close to 1 suggesting that APARCH model with Student-t density presents the highest forecasting accuracy than other models in case of DSE.

5. Conclusions

In this paper we compared the in-sample and out-of-sample forecasting performance of several GARCH-type models such as GARCH, EGARCH and APARCH model with Gaussian, fat tail (Student-t and Generalized error distribution) and skewed (Skewed Student-t and Skewed-GED) densities in case of Dhaka Stock Exchange. Our empirical results show that noticeable improvements can be made when using asymmetric GARCH model with non normal distributional assumptions in case of in-sample estimation. Among these models, APARCH model with Student-t distribution fits the data well. In the case of out-of-sample forecasting performance we found that APARCH model with all distributional assumptions give the lowest value of MSE and QLIKE where EGARCH model with all marginal densities provides the poorest forecasting performance. The comparison among densities suggests that the model under the Student-t density gives better results than other densities. The superiority of out-of-sample volatility predictive performance of the APARCH model with Student-t distribution is verified by SPA test. Therefore, our study suggests that APARCH model with Student-t distributions is the most suitable model for both in-sample and out-of-sample forecasting performance in case of DhakaStock Exchange, Bangladesh.

References

1. Baillie, R. and Bollerslev, T.: 'The Message in Daily Exchange Rates: A Conditional-Variance Tale', *Journal of Business and Economic Statistics*, 7, 297–305, 1989.
2. Beine, M., Laurent, S. and Lecourt, C.: 'Accounting for Conditional Leptokurtosis and Closing Days Effects in FIGARCH Models of Daily Exchange Rates', *Applied Financial Economics*, Vol.12, 123-138, 2002.
3. Bollerslev, T.: 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics*, 31, 307–327, 1986.
4. Bollerslev, T.: 'A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return', *Review of Economics and Statistics*, 69, 542–547, 1978.
5. Chen, B. and Hong, Y.: 'A Unified Approach to Validating Univariate and Multivariate Conditional Distribution Models in Time Series', Seminar Paper of International Symposium on Recent Developments of Time Series Economietrics, Xiamen University, China, 2009.
6. Ding, Z., Granger, C. W. J. and Engle, R. F.: 'A Long Memory Property of Stock Market Returns and a New Model', *Journal of Empirical Finance*, 1, 83–106, 1993.
7. Engle, R.: 'Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation', *Econometrica*, 50, 987–1007, 1982.
8. Fernandez, C., and Steel, M.F.J.: 'On Bayesian Modelling of fat tails and skewness', *Journal of the American Statistical Association*, 93, 359-371, 1998.
9. Hansen, P.R.: 'A test for superior predictive ability', *Journal of Business and Economic Statistics*, 23, 365-380, 2005.
10. Hasan, M.K., Basher, S.A., and Islam, M.A.: 'Time Varying Volatility and Equity Returns in Bangladesh Stock Market', *Finance, 0310015, Economics working paper* Achieved at WUSTL, 2004.
11. Hou, A. and Suasdi, S.: 'A Nonparametric GARCH model of crude oil price return volatility',*Emergy Economics*, 34, Issue March, 618-635, 2012.
12. Kaiser, T.: 'One-Factor-GARCH Models for German Stocks -Estimation and Forecasting', *Working Paper*, Universiteit Tubingen, 1996.
13. Liu, H.C., Lee, Y.H. and Lee, M.C.: 'Forecasting China Stock Markets Volatility via GARCH Models under Skewed-GED distribution', *Journal of Money, Investment and Banking*, Issue 7, 5-15, 2009.
14. Lambert, P. and Laurent, S.: 'Modellingskewness Dynamics in series of Financial Data', *Discussion Paper*, Institute de Statistique, Louvain-la-Neuve, 2000.
15. Lambert, P. and Laurent, S.: 'Modelling Financial Time Series Using GARCH-Type Models and Skewed student Density', *Mimeo*, Universite de Liege. 2001.
16. Nelson, D.: 'Conditional heteroskedasticity in asset returns: a new approach', *Econometrica*, 59, 349–370, 1991.
17. Patton, A.J.: 'Volatility forecast comparison using imperfect volatility proxies', *Journal of Econometrics*, 160(1), 246-256, 2011.
18. Politis, D. N., and Romano, J.P.: 'The stationary bootstrap', *Journal of the American Statistical Association*, 89, 1303-1313, 1994.
19. Peters, J.P.: 'Estimating and forecasting volatility of Stock indices using asymmetric GARCH models and (Skewed) Student-t densities', Ecole d' Administration des Affairs, University of Liege, Belgium, 2001. http://www.unalmed.edu.co/~ndgirald/Archivos%20Lectura/Archivos%20curso%20Series%20II/jppeters.
20. Rahman, M.M., Zhu, J.P., and Rahman, M. S.: 'Impact study of volatility modeling of Bangladesh stock index using non-normal density', *Journal of Applied Statistics*, Vol. 35, No. 11, 1277-1292, 2008.

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Two Efficient Estimators for Small Areas

Dulal Chandra Roy[1] and Keya Rani Das[2]

[1]Professor, Department of Statistics, University of Rajshahi

[2]Lecturer, Department of Statistics, Bangabandhu Sheikh MujiburRahman Agriculture University, Gazipur

Abstract. In the event of small areas or domains, we have constructed two weighted estimators of the domain mean which perform better than the corresponding existing ones. We have also undertaken performance sensitivity in respect of the proposed estimators and worked out sensible values of the involved weights with a view to retaining superiority of the proposed estimators vis-a-vis the existing ones.

Key words: Small area (domain) estimation, Performance-sensitivity.

1 Introduction

Estimation at disaggregated levels, i.e., at the levels of small domains and areas has held out interest for a long time. But the relevant estimates are not usually obtainable for national sample surveys which generally provide estimates at the national level, and any attempt to extract estimates for small domains does not generally meet with success from the point of view of precision, validity and practicability. In recent years, the estimates for small domains have increasingly engaged the attention of sampling theorists and practitioners. A good collection of contributions to small domain estimation is available in Platek et al. (1987). The method of estimation for small domains may vary with the type of domains. For a classification of types of domains, we refer to Purcell and Kish (1979). In what follows, we have suggested two estimators as potentially efficient alternatives to the existing ones. One of the proposed estimators has been motivated and triggered by the thought of exploiting the available data in full as has been advocated by authors such as Sarndal et al. (1992).

2 Two Efficient Estimators

Suppose that a simple random sample s of size n is drawn from the population of size N which contains a small domain d of interest having size N_d. Let y be the study variable with the value y_i on unit i,i = 1,2,..., N. Further, let s_d be the sample of n_d units which are common to s and d.

If we denote the mean of the domain d by $\overline{Y}_d$, then its estimator that has often been considered in the literature is defined by

$$\bar{y}_d = \frac{1}{n_d}\sum_{i\in s_d} yi \qquad (1)$$

For this purpose, we refer the readers to Hedayat and Sinha (1991, pp. 342-45) and Sarndal et al. (1992, pp. 391-92). Let $\bar{y}$ be the sample mean based on s of size n.

If N_d is known in advance, the estimator $\bar{y}_d$ is asymptotically unbiased for $\bar{y}_d$ with large sample variance given by

$$V(\bar{y}_d) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{S^2}{\Phi_d} \qquad (2)$$

where $\Phi_d = \frac{N_d}{N}$, $S_d^2 = \frac{1}{N_d - 1}\sum_{i\in d}(y_{i-}\bar{Y}_d)^2$.

However, if N_d is not known, an adaptation of $\bar{y}_d$ available in the literature is

$$\bar{y}_d^* = \frac{\bar{y}^*}{\bar{x}} \qquad (3)$$

where $\bar{y}^* = \frac{1}{n}\sum_{i\in s} y_i^*$, $y_i^* = \begin{cases} y_i, \text{ if } i \in d \\ 0, \text{otherwise} \end{cases}$

and $\bar{x} = \frac{1}{n}\sum_{i\in s} x_i$, x being an auxiliary variable defined as

$x_i = \begin{cases} 1, & \text{if } i \in d \\ 0, & \text{otherwise} \end{cases}$

For a discussion of $\bar{y}_d^*$, we refer to Hedayat and Sinha (1991, pp. 345-46). The estimator $\bar{y}_d^*$ is identified as the ratio estimator which is unbiased, having the same variance (to terms of O (n^{-1})) as $\bar{y}_d$ given by (2).

We now propose the following two estimators:

$$\bar{y}_d' = \alpha\bar{y}_d + (1-\alpha)\bar{y} \tag{4}$$

and

$$\bar{y}^{*\prime} = \alpha^* \frac{\bar{y}^*}{\bar{x}} \tag{5}$$

where α and α^* are pre-assigned weights to be determined optimally. In proposing $\bar{y}^{*\prime}$, we have kept in mind the dictum that we must make the best of the data at hand.

2.1 Evaluation of the Performance of the First Estimator

The bias of $\bar{y}_d'$ is obtainable as

Bias $\left(\bar{y}_d'\right) = (1-\alpha)(\bar{Y} - \bar{Y}_d)$,

and can be reduced by suitable choice of α. However, we would, in practice, determine α with a view to minimizing the mean square error of $\bar{y}_d'$ given by

$$\text{MSE}\left(\bar{y}_d'\right) = \alpha^2(\bar{Y}_d - \bar{Y})^2(1+C^2) - 2\alpha[V(\bar{y}) - \text{Cov}(\bar{y}, \bar{y}_d) + (\bar{Y}_d - \bar{Y})^2] + V(\bar{y}) + (\bar{Y}_d - \bar{Y})^2 \tag{6}$$

where $V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2, S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$

$$V(\bar{y}_d) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{S_d^2}{\Phi_d} \qquad \text{Cov}(\bar{y}, \bar{y}_d) = \left(\frac{1}{n} - \frac{1}{N}\right)S_d^2 = \Phi_d V(\bar{y}_d)$$

and C = Coefficient of variation (C.V.) of $(\bar{y}_d - \bar{y})$.

The optimum value of α that minimizes (6) is found to be

$$\alpha_{opt} = \frac{A}{B} \tag{7}$$

where $A = \left(\frac{1}{n} - \frac{1}{N}\right)(S^2 - S_d^2) + (\bar{Y}_d - \bar{Y})^2$

and $B = \left(\frac{1}{n} - \frac{1}{N}\right)\left\{(S^2 - S_d^2) + S_d^2\left(\frac{1-\Phi_d}{\Phi_d}\right)\right\} + (\bar{Y}_d - \bar{Y})^2 = (\bar{Y}_d - \bar{Y})^2(1+C^2)$.

Using the optimum value of α, we have

$$\underset{\text{MSE}_{\text{opt}}}{(\bar{y}'_d)} = V(\bar{y}) + (\bar{Y}_d - \bar{Y})^2 - \frac{A^2}{B} \tag{8}$$

As regards the performance of the newly proposed estimator $\bar{y}'_d$, we can note that

$$\underset{\text{MSE}_{\text{opt}}}{(\bar{y}'_d)} - V(\bar{y}_d) = -\frac{(B - A)^2}{B} \leq 0,$$

which implies that $\bar{y}'_d$ is more efficient than $\bar{y}_d$.

However, an irritant that needs to be tackled lies in assuming knowledge of the optimum value of the weighting factor α which involves the population quantities. We now address ourselves to this problem.

Performance- Sensitivity of $\bar{y}'_d$

Let P_I denote the proportional inflation in the variance of $\bar{y}'_d$ resulting from the use of some α other than $\alpha_{opt,}$ i.e.,

$$P_I = \frac{\text{MSE}(\bar{y}'_d) - \text{MSE}_{\text{opt}}(\bar{y}'_d)}{\text{MSE}_{\text{opt}}(\bar{y}'_d)}$$

Then, defining

$$\nu = \frac{\alpha - \alpha_{\text{opt}}}{\alpha_{\text{opt}}} \Rightarrow \alpha = (1 - \nu)\,\alpha_{\text{opt}}$$

where ν is the proportional deviation in α_{opt}, we obtain

$$P_I = \nu^2 \left(\frac{\alpha_{\text{opt}}}{1 - \alpha_{\text{opt}}}\right)^2 \left(\frac{V(\bar{y}_d) - \text{MSE}_{\text{opt}}(\bar{y}'_d)}{\text{MSE}_{\text{opt}}(\bar{y}'_d)}\right) = \left(\frac{\alpha - \alpha_{\text{opt}}}{1 - \alpha_{\text{opt}}}\right)^2 G$$

where G is the gain in efficiency of $\bar{y}'_d$ (using α_{opt}) relative to $\bar{y}_d$.

The proposed estimator $\bar{y}'_d$ will continue to fare better than $\bar{y}_d$ even when some α other than α_{opt} is used provided the quantity P_I does not exceed G, i.e.,

$$\left|\frac{\alpha - \alpha_{\text{opt}}}{1 - \alpha_{\text{opt}}}\right| \leq 1. \tag{9}$$

From (9), it is clear that a good guess about α_{opt} can yield the range of values for α in order that P_I be smaller than G. For example, if an indication is available that $S^2 \geq S_d^2$ and hence $\alpha_{opt} \leq 1$ from (7), then α can be taken to be any value not exceeding 1. In essence, the fact that $P_I \leq G$ will always ensure superiority of $\bar{y}'_d$ (using α) relative to $\bar{y}_d$ because the desired gain in efficiency, say, G' in this case can be expressed as

$$G' = \frac{V(\bar{y}_d) - \text{MSE}(\bar{y}'_d)}{\text{MSE}(\bar{y}'_d)} = (G - P_I)\frac{\text{MSE}_{\text{opt}}(\bar{y}'_d)}{\text{MSE}(\bar{y}'_d)}$$

which will be non-negative if $G \geq P_I$. Thus, gain in precision would accrue via a sensibly chosen α that results in $G \geq P_I$.

2.2 Evaluation of the Performance of the Second Estimator

The bias and the mean square error of the second proposed estimator $\bar{y}_d^{*\prime}$, to terms of O (n^{-1}), are expressible as

$$\text{Bias } (\bar{y}_d^{*\prime}) = \bar{Y}_d\,(\alpha^* - 1)$$

$$\text{MSE } (\bar{y}_d^{*\prime}) = (\alpha^* - 1)^2\,\bar{Y}_d^2 + \alpha^{*2}\,V(\bar{y}_d^*) \qquad (10)$$

The optimum value of α^* that minimizes the expression (10) is given by

$$\alpha_{opt}^* = \frac{\bar{Y}_d^2}{\bar{Y}_d^2 + V(\bar{y}_d^*)} = \frac{1}{1+C_1^2}$$

where C_1, besides being the C.V. of $\bar{y}_d$ as explained earlier, is also the C.V. of $\bar{y}_d^*$ since both $\bar{y}_d^*$ and $\bar{y}_d$ have the same variance to terms of $O(n^{-1})$. Note that $0 < \alpha^*_{opt} \leq 1$. It would be apt to point out that α^*_{opt} being a function of a coefficient of variation only is obtainable in many practical situations. After some algebra, the optimum mean square error of $\bar{y}_d^{*\prime}$ can be worked out as

$$\text{MSE}_{opt}(\bar{y}_d^{*\prime}) = \frac{\bar{Y}_d^2\,V(\bar{y}_d^*)}{\bar{Y}_d^2 + V(\bar{y}_d^*)} = \frac{V(\bar{y}_d^*)}{1+C_1^2}$$

To appraise the performance of the estimator $\bar{y}_d^{*\prime}$, we find that

$$\text{MSE}_{opt}(\bar{y}_d^{*\prime}) - V(\bar{y}_d^*) = \frac{\{V(\bar{y}_d^*)\}^2}{V(\bar{y}_d^*) + \bar{Y}_d^2} \leq 0,$$

implying thereby that $\bar{y}_d^{*\prime}$ is more efficient than $\bar{y}_d^*$.

Performance- Sensitivity of $\bar{y}_d^{*\prime}$

Let P_I^* (similar to P_I of sub-section 2.1) symbolize the proportional increase in variance resulting from the use of some α^*- value other than α_{opt}^*, i.e.,

$$P_I^* = \frac{\text{MSE}\,(\bar{y}_d^{*\prime}) - \text{MSE}_{opt}\,(\bar{y}_d^{*\prime})}{\text{MSE}_{opt}\,(\bar{y}_d^{*\prime})}$$

Further, let ν^* be the proportional departure from α_{opt}^* owing to the use of α^*, i.e., notationally

$$\nu^* = \frac{\alpha^* - \alpha_{opt}^*}{\alpha_{opt}^*}$$

Then, after some algebra, we obtain

$$P_I^* = \left(\frac{\nu^*}{C_1^2}\right)^2 \frac{V(\bar{y}_d^*) - \text{MSE}_{opt}\,(\bar{y}_d^{*\prime})}{\text{MSE}_{opt}\,(\bar{y}_d^{*\prime})} = \left(\frac{\nu^*}{C_1^2}\right)^2 G^*$$

where G* is the gain in efficiency of $\bar{y}_d^{*\prime}$ (using α_{opt}^*) relative to $\bar{y}_d^*$. It can be noted that P_I^* will never exceed G^* if

$$\left|\frac{\nu^*}{C_1^2}\right| \leq 1 \Rightarrow -1 \leq \frac{\nu^*}{C_1^2} \leq 1$$

which, using $\alpha^{*}_{opt} = \frac{1}{1+C_1^2}$, yields

$$\frac{1-C_1^2}{1+C_1^2} \le \alpha^{*} \le 1. \qquad (11)$$

Thus, it is clear that a choice of α^{*} within the bounds given by (11) is enforceable in many sampling situations, thereby ensuring better performance of $\bar{y}_d^{*\prime}$ compared to $\bar{y}_d^{*}$. Further, the fact that $G^{*} \geq P_I^{*}$ will invariably ensure superiority of $\bar{y}_d^{*\prime}$ (using α^{*}) relative to $\bar{y}_d^{*}$ because the desired gain in efficiency, say, $G^{*\prime}$ which is expressible as

$$G^{*\prime} = \frac{V(\bar{y}_d^{*}) - MSE(\bar{y}_d^{*\prime})}{MSE(\bar{y}_d^{*\prime})}$$

$$(G_{*} - P_I^{*}) \frac{MSE_{opt}(\bar{y}_d^{*\prime})}{MSE(\bar{y}_d^{*\prime})}$$

is non-negative if $G^{*} \geq P_I^{*}$.

2.3 A Comparison of the Proposed Estimators

From the results of Sub-sections 2.1 and 2. 2, we have

$$MSE_{opt}(\bar{y}_d') - V(\bar{y}_d) = -\frac{(B-A)^2}{B}$$

$$MSE_{opt}(\bar{y}_d^{*\prime}) - V(\bar{y}_d^{*}) = -\frac{\{V(\bar{y}_d^{*})\}^2}{V(\bar{y}_d^{*}) + \bar{Y}_d^2},$$

And since $V(\bar{y}_d) = V(\bar{y}_d^{*}),$ we have

$$MSE_{opt}(\bar{y}_d') - MSE_{opt}(\bar{y}_d^{*\prime}) = -\frac{(B-A)^2}{B} + \frac{\{V(\bar{y}_d^{*})\}^2}{V(\bar{y}_d^{*}) + \bar{Y}_d^2}$$

$$= -\{V(y_d^{*})\}^2 \left[\frac{(1-\Phi_d)^2 \bar{Y}_d^2 (1+C_1^2) - (\bar{Y}_d - \bar{Y})^2 (1+C^2)}{\bar{Y}_d^2 (\bar{Y}_d - \bar{Y})^2 (1+C^2)(1+C_1^2)} \right] \qquad (12)$$

It follows from (12) that the estimator $\bar{y}_d'$ would perform better than $\bar{y}_d^{*\prime}$ if

$$(1-R)^2 \le (1-\Phi_d)^2 \frac{1+C_1^2}{1+C^2}$$

$$\Rightarrow 1-(1-\Phi_d)\left(\frac{1+C_1^2}{1+C^2}\right)^{1/2} \le R \le 1+(1-\Phi_d)\left(\frac{1+C_1^2}{1+C^2}\right)^{1/2} \qquad (13)$$

holds. When the domain mean overlaps with the population mean, the condition (13) stands satisfied. As a matter of fact, we can find the bounds on $R = \frac{\bar{Y}}{\bar{Y}_d}$ in many practical situations with view to guiding us with regard to a choice between $\bar{y}_d'$ and $\bar{y}_d^{*\prime}$.

3 Numerical Illustration

We have chosen two examples from Sarndal et al. (1992, pp. 414-415) to illustrate the findings of Section 2.

Example 1. The following data relate to the study variable y (number of conservative seats in Municipal Council) in respect of a Swedish population that comprises three regions:

Major Region d	N_d	$\sum_{i\in d} y_i$	$\sum_{i\in d} y_i^2$
1	47	349	3,375
2	50	437	4,159
3	45	194	1,000

For the purpose of application of the results of Section 2, we consider the region 1 as the domain d of our interest. The following quantities are obtained from the above population data:

N=142, $\overline{Y}$ =6.9014, $\overline{Y}_d$ = 7.4255, S^2 = 12.5576, S_d^2= 17.0324 and Φ_d =0.3310.

Now taking a sample of size 8, we can compute the following additional quantities:

V($\overline{y}$)= 1.4813, A = -0.2532, B = 3.8075, α_{opt}= -0.0665, α^*_{opt} =0.90, $V(\overline{y}_d) = V(\overline{y}^*_d)$ = 6.0698,

$MSE_{opt}(\overline{y}'_d)$=1.7391 and $MSE_{opt}(\overline{y}^{*\prime}_d)$= 5.4679.

Thus, the gains in efficiency of $\overline{y}'_d$ and $\overline{y}^{*\prime}_d$ relative to the $\overline{y}_d$ and $\overline{y}^*_d$ are found to be 249% and 11% respectively, meaning thereby that the proposed estimators perform better than the customary estimators available in the literature. We have prepared Tables 1, and 2 to present the impact of departures from $\alpha_{opt}$and α^*_{opt}.

Table 1: Performance- Sensitivity of $\overline{y}'_d$ and $\boldsymbol{\alpha_{opt}}$**= -0.0665**

α	P_I	G'
-1.10	2.3385	0.0454
-1.00	1.9078	0.2003
-0.75	1.0228	0.7254
-0.50	0.4114	1.4727
-0.10	0.0025	2.4815
0.10	0.0607	2.2904
0.50	0.7026	1.0499
0.75	1.4595	0.4190
0.95	2.2622	0.0699
1.00	2.4902	0.0

Table 2: Performance- sensitivity of $\overline{y}^{*\prime}_d$ and α^*_{opt} **= 0.90**

α^*	P_I^*	$G^{*\prime}$
0.81	0.908	0.0162
0.85	0.281	0.0789
0.90	0.000	0.1101
0.95	0.0281	0.0808
1.00	0.1101	0.0

The last column of the above Table 1 unfolds that the estimator $\overline{y}'_d$ retains its superiority for values of α ranging from- 1.10 to 1. In other words, despite substantial and significant departure from α_{opt} as reflected by a wide range of values of α, $\overline{y}'_d$ maintains its lead over $\overline{y}_d$.

Table 2 clearly shows that the estimator $\bar{y}_d^{*\prime}$ ensures better performance compared to $\bar{y}_d^{*}$ for the different values of α^* ranging from 0.81 to 1.

Example 2. The following data relate to the study variable y (real estate values according to 1984 assessment) in respect of a real Swedish population comprising three regions:

Major Region d	N_d	$\sum_{i\in d} y_i$	$\sum_{i\in d} y_i^2$
1	105	382,906	5,377,345,182
2	94	271,392	2,729,354,250
3	85	209,719	956,277,147

To demonstrate the application of the results of Section 2, we consider region 1 as the domain d of interest. The following quantities are obtained from the population data:

N = 284, $\bar{Y}$ =3077.4543, $\bar{Y}_d$ = 3646.7238, S^2 = 22520039.13, S_d^2 = 38278776.49 and Φ_d = 0.3697.
Considering a sample of size 15, we compute

$V(\bar{y})$ = 1422040.03, A = -671026.2152, B= 3449935.552, α_{opt}= -0.1945, α_{opt}^* = 0.6704,

$V(\bar{y}_d)$ = $V(\bar{y}_d^*)$ = 6538095.775, $MSE_{opt}(\bar{y}_d')$ = 1615590.52 and $MSE_{opt}(\bar{y}_d^{*\prime})$ = 4383164.896.

Thus, the gains in efficiency of $\bar{y}_d'$ and $\bar{y}_d^*$ relative to $\bar{y}_d$ and $\bar{y}_d^*$ are found to be 305% and 49% respectively, meaning thereby that the proposed estimators perform better than the customary estimators.

Again, for Example 2, we have prepared Tables 3 and 4 to present the impact of departures from $\alpha_{opt}$and α_{opt}^*.

Table 3: Performance- Sensitivity of $\bar{y}_d'$ and $\boldsymbol{\alpha_{opt}}$ **= - 0.1945**

α	P_1	G′
-1.35	2.8512	0.0501
-1.25	2.3790	0.1977
-1.00	1.3855	0.6964
-0.50	0.1993	2.3744
-0.10	0.0191	2.9711
0.10	0.1852	2.4144
0.50	1.0299	0.9936
0.75	1.9050	0.3931
0.95	2.7972	0.0658
1.00	3.0469	0.0

Table 4: Performance- Sensitivity of $\bar{y}_d^{*\prime}$ & α_{opt}^* **= 0.6704**

α^*	P_I^*	$G^{*\prime}$
0.35	0.4646	0.0185
0.40	0.3309	0.1208
0.50	0.1314	0.3184
0.60	0.1050	0.4589
0.70	0.0040	0.4857
0.80	0.0760	0.3863
0.90	0.2386	0.2043
0.95	0.3539	0.1018
1.00	0.4916	0.0

The last column of Table 3 clearly points to the superiority of the estimator $\bar{y}'_d$ over $\bar{y}_d$ even if we choose some α-value (rather than α_{opt}) in the range from- 1.35 to 1.00.

Table 4 alludes to the fact that, even if α^*_{opt} is not available, the lead of $\bar{y}^{*'}_d$ over $\bar{y}^*_d$ is maintained when $0.35 \leq \alpha^* \leq 1$, i.e., the proportional departure from α^*_{opt} is to the extent of about 50%.

References

1. Hedayat, A.S. and Sinha, B.K.(1991). *Design and Inference in Finite Population Sampling.* John Wiley & Sons, New York.
2. Platek, R., Rao, J.N.K., Sarndal, C.E. and Singh, M.P.(1987). Small Area Statistics:An International Symposium. John Wiley & Sons, New York.
3. Purcell, N.J. and Kish, L.(1979). Estimation for Small Domains. *Biometrics*, 35, 365-384).
4. Sarndal,C.E., Swensson, B. and Wretman, J.(1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Bayesian Sample Size Determination for Estimating Proportion

Farhana Sadia, Syed S. Hossain

Institute of Statistical Research and Training (ISRT), University of Dhaka, Bangladesh.

Abstract: Sample size is the number of observations used for calculating estimates of a given population and sample size determination(SSD) plays often an important role in planning a statistical study - and it is usually a difficult one. Sample size determination means the act of choosing the number of observations to include in a statistical sample. Specifically, we deal with the Bayesian approach to SSD which gives researchers the possibility of taking into account pre-experimental information and uncertainty on unknown parameters. In this context, the leading idea is to choose the minimal sample size using three different Bayesian approaches based on highest posterior density (HPD) intervals which are average coverage criterion(ACC),average length criterion(ALC) and worst outcome criterion(WOC). Here, necessary sample sizes by both non-Bayesian and Bayesian methods using real life data set are compared. We demonstrated that the developed Bayesian methods can be efficient and may require fewer subjects to satisfy the same length criterion.

Keywords: Average coverage criterion(ACC), Average length criterion (ALC), Worst outcome criterion (WOC).

1 Introduction

Statistical experiments are always better if they are carefully planned. For good planning of a statistical survey, sample size is a crucial component in statistical theory. For getting optimal sample size, sample size determination plays an important role in designing the statistical survey[8]. In case of sample size determination(SSD), a considerable number of criteria are available depending on the two types of inferential approach which are Frequentist and Bayesian approach.

Frequentist sample size determination methods depend directly on the unknown parameter of interest but Bayesian way do not depend on the guessed value of the true parameter; it depends on its prior distribution. Bayesian methods are based on the posterior distribution which combines the pre-experimental information of the parameter (prior distribution) and the experimental data that is likelihood. In case of Bayesian sample size determination, the marginal or prior-predictive distribution is used which is the mixture of the sampling distribution of the data and the prior distribution of the unknown parameters[4]. In this context, the minimal sample size is determined using three different Bayesian approaches based on highest posterior density (HPD) intervals which are average coverage criterion (ACC), average length criterion (ALC) and worst outcome criterion (WOC).

In sample size calculation, the literatures on Bayesian approaches have recently received much attention. Bayesian SSD for a single binomial parameter has been discussed by so many authors like Adcock(1988b, 1992, 1995), Pham-Gia(1995), Joseph et al.(1995a,b) and Pham-Gia and Turkan(1992). This paper presents several new results on Bayesian sample size determination for estimating binomial proportion for SRS, more complex survey and Bayesian sample size determination for proportion using different prior.

2 Bayesian Sample Sizes for Binomial Proportions

In case of binomial experiment, let θbe the binomial parameter to be estimated. In this case, the prior model is a beta distribution and the likelihood model is a binomial distribution. For binomial proportion, the three Bayesian sample size determination criteria are as follows.

2.1 Average Coverage Criterion (ACC)

In the case of a binomial parameter, the average coverage criterion finds the smallest n satisfying the following condition[2]

$$\sum_{x=0}^{n} \Pr\{\theta \in (a(x,n), a(x,n)+l)\}\, p(x,n) \geq 1-\alpha \qquad (1)$$

Where

$$\Pr\{\theta \in (a(x,n), a(x,n+l)\} \propto \int_{a(x,n)}^{a(x,n)+l} \theta^{x}(1-\theta)^{(n-x)} f(\theta) d\theta \qquad (2)$$

$l$is the user-specified length of the credible set, $f(\theta)$ is the prior distribution of $\theta$and $p(x,n)$ is the preposterior probability function of the data.

2.2 Average Length Criterion (ALC)

In the case of a binomial parameter, the average length criterion seeks the minimum n satisfying the following condition[2]

$$\sum_{x=0}^{n} l\acute{(}x,n)p(x,n) \leq l \qquad (3)$$

Where, $l\acute{(}x,n)$is the length corresponding to the HPD interval.

2.3 Worst Outcome Criterion (WOC)

The worst outcome criterion finds the smallest $n$satisfying the following condition

$$inf_{x\in\chi}\left\{\int_{a(x,n)}^{a(x,n)+l} f(\theta|x,n)d\theta\right\} \geq 1-\alpha \qquad (4)$$

Where both $l$and $\alpha$are fixed in advance.

3 Computation of Sample Size for Proportion with Simple Random Sampling

In case of simple random sampling, for determining an appropriate classical sample size n, pre-assigned values of $\alpha, d$and P have to be given and Bayesian sample size can be computed using the above three criteria. Now, a table of classical and Bayesian sample $\alpha = 0.05$is given below-

Table 1. Table of Classical and Bayesian Sample Size for SRS($\alpha = 0.05$)

Length	Classical Sample Size	Different Prior	Bayesian Sample Size		
			ACC	ALC	WOC
0.01	38416	(1,1)	27778	23747	38412
		(2,2)	31630	29990	38410
		(3,3)	33326	32576	38408
		(4,4)	34419	33981	38406
		(2,3)	31638	29956	38409
0.05	1537	(1,1)	1107	944	1534
		(2,2)	1262	1193	1532
		(3,3)	1327	1295	1530
		(4,4)	1368	1350	1528
		(2,3)	310	297	378
0.1	384	(1,1)	275	235	381
		(2,2)	312	298	379
		(3,3)	328	320	377
		(4,4)	336	332	375
		(2,3)	310	297	378
0.50	15	(1,1)	9	8	12
		(2,2)	8	8	10
		(3,3)	7	7	8
		(4,4)	6	6	6
		(2,3)	8	7	9

Table 1 compares the required sample sizes for classical and Bayesian criterion for proportion with simple random sampling assuming different prior distributions for $\alpha = 0.05$. The proportion that we have used in the above table is $\theta = 0.5$. From table 1, we can say that the three Bayesian criteria provides very different sample sizes and it is also worth observing from the table that $n_{ALC} \leq n_{ACC} \leq n_{WOC}$. In fact, all Bayesian criteria provides smaller sample sizes than the classical approach and from table 1, two main observation can be made. Firstly, Bayesian criteria ACC and ALC seem to lead very similar sample sizes whereas WOC criteria provides the largest sample sizes with no consistent ordering seen for the other two criteria. For example, in case of $l = 0.05$ and a prior about proportion θ,

$u = v = 3$, ACC and ALC yield the sample size of $n = 1327$ and $n = 1295$ which are somewhat similar but WOC yields a sample size of $n = 1530$ which is larger than the sample size of ACC and ALC. Secondly, As long as non-informative prior approaches to informative prior, the

sample size gradually increases. For example, in case of $l = 0.05$ and a prior about proportion θ, $u = v = 1$, ACC, ALC and WOC yield the sample size of $n = 1107$; $n = 944$ and $n = 1534$ but in case of the informative prior distribution $(u; v) = (2,2), (3,3), (4,4)$the sample sizes are gradually larger than the sample

size of the previous prior distribution.

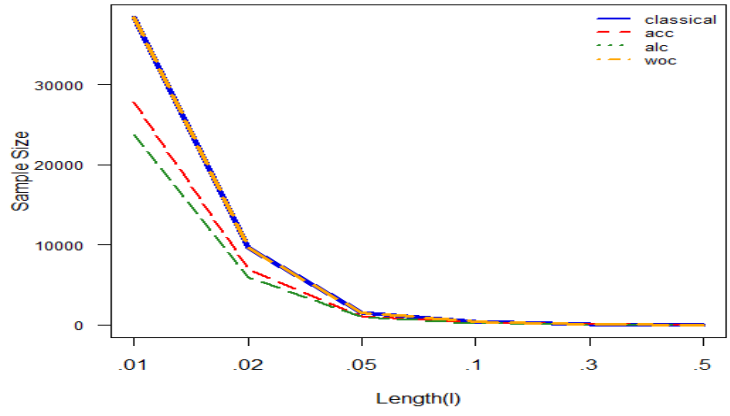


Figure 1: Classical and Bayesian sample size for different length for prior(1,1) .

The above figure represents the classical and Bayesian sample size for different length using non-informative prior (1,1) . The figure shows that as the desired length approaches to large length from small length, the sample size gradually decreases and this fact is true for both classical method and Bayesian method. This figure also shows that the all three Bayesian criteria give smaller sample size than the classical approach of sample size determination. Here, it is seen that the sample size suggested by the Bayesian criteria ACC and ALC are almost similar but the criteria WOC gives the sample size which is larger than the other two criteria ACC and ALC. It is also observing that the Bayesian criteria WOC provides almost same sample size like the classical approach. That means, from this figure we can easily say that

$$n_{ALC} \leq n_{ACC} \leq n_{WOC} \leq n_{Classical}$$

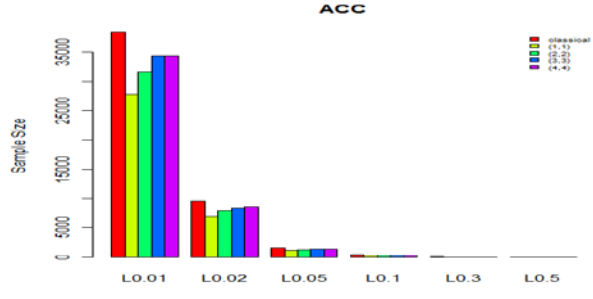


Figure 2: Classical and Bayesian ACC sample size for different prior and different length .

This figure represents the sample size of classical approach and the sample size of the Bayesian criteria ACC using different prior for different value of desired length of the HPD interval. From this figure, we can reveal that the Bayesian sample sizes of ACC criteria using different prior are smaller than the classical sample size. It is also important to note that as long as the non-informative prior approaches to informative prior, the sample size gradually increases but the sample size gradually decreases when we approaches to big length from small length of the HPD interval.

4 Computation of Sample Size for Proportion with Cluster Sampling

In case of cluster sampling, classical sample size for proportion can be determined by multiplying the classical sample size for SRS with the design effect. Usually, design effect is taken 1.5. Now, a table of classical and Bayesian sample size for $\alpha = 0.05$ is given below-

Table 2. Table of Classical and Bayesian Sample Size for Cluster Sampling($\alpha = 0.05$)

Length	Classical Sample Size	Different Prior	Bayesian Sample Size		
			ACC	ALC	WOC
0.01	57624	(1,1)	41651	35555	57618
		(2,2)	47445	44985	57615
		(3,3)	49989	48864	57612
		(4,4)	51629	50972	57609
		(2,3)	47457	44934	57614
0.05	2306	(1,1)	1654	1427	2301
		(2,2)	1893	1790	2298
		(3,3)	1991	1943	2295
		(4,4)	2052	2025	2292
		(2,3)	1889	1790	2297
0.1	576	(1,1)	413	353	572
		(2,2)	468	447	569
		(3,3)	492	480	566
		(4,4)	504	498	563
		(2,3)	465	446	567
0.50	23	(1,1)	14	12	18
		(2,2)	12	12	15
		(3,3)	11	11	12
		(4,4)	9	9	9
		(2,3)	12	11	14

The above table represents the comparison between the classical sample sizes and Bayesian sample size for proportion with cluster sampling assuming different prior distributions for $\alpha = 0.05$.. In this table, we have used the proportion $\theta = 0.5$ for determining the classical sample size. This table shows that the all three Bayesian criteria give very different sample sizes, that is, $n_{ALC} \leq n_{ACC} \leq n_{WOC}$. Actually the classical approach provides larger sample sizes than all three Bayesian criteria and it is also observing that Bayesian criteria ACC and ALC seem to lead very similar sample sizes whereas WOC criteria provides the largest sample sizes. For example, in case of $l = 0.05$ and a prior about proportion θ, $u = v = 1$, ACC and ALC yield the sample size of $n = 1654$ and $n = 1427$ which are somewhat similar but WOC yields a sample size of $n = 2301$ which is larger than the sample size of ACC and ALC. We can also conclude that as long as non-informative prior approaches to informative prior, the sample size gradually increases. For example, in case of $l = 0.1$ and a prior about proportion θ, $u = v = 1$, ACC and ALC yield the sample size of $n = 413$ and $n = 353$ and $n = 572$ but in case of the informative prior distribution $(u; v) = (2,2), (3,3), (4,4)$, the sample sizes are gradually larger than the sample size of the previous prior distribution.

5 Sample Size for Some Real Life Surveys in Bangladesh

5.1 Computation of Sample Size for Bangladesh Demographic and Health Survey (BDHS) 2004

The Bangladesh Demographic and Health Survey (BDHS) is a periodic survey conducted in Bangladesh to serve as a source of population and health data for policymakers, program managers, and the research community. The classical and Bayesian sample sizes according to some variables using BDHS 2004 data

set are given in the following table-

Table 3. Table of Classical and Bayesian Sample Size for Urban Women Using BDHS 2004($\alpha = 0.05$)

Variable	Classical Sample Size	Bayesian Sample Size		
		ACC	ALC	WOC
No Education	1732	1470	1268	3376
With Secondary Education or Higher	1416	1110	957	2577
Currently Married	45	107	93	219
Want No More Children	332	260	224	579
Mothers Received Tetanus Injection (last birth)	96	164	142	373
Mothers Received Medical Care at Delivery	1751	1610	1365	3740
Child Received Polio Vaccination	121	192	166	469

The above table shows the classical and Bayesian sample size for some variables for urban women using BDHS 2004 data set. In this table, the relative margin of error is used as 10%. This table describes that classical sample size is greater than Bayesian sample size of ACC and ALC criteria for some variables (no education, with secondary education or higher,want no more children, Mothers Received Medical Care at Delivery) and the classical sample size is smaller than the Bayesian sample size of all criteria for the remaining variables in the above table. This indicates that here sampling is not done properly. But it is important to note that the Bayesian WOC criteria provides the largest sample size among the other two criteria and the classical method of sample size determination because WOC criteria is a conservative criteria.

Table 4. Table of Classical and Bayesian Sample Size for Rural Women Using BDHS 2004($\alpha = 0.05$)

Variable	Classical Sample Size	Bayesian Sample Size		
		ACC	ALC	WOC
No Education	833	624	535	1347
With Secondary Education or Higher	138	1651	1407	3451
Currently Married	31	81	70	153
Want No More Children	249	193	166	392
Mothers Received Tetanus Injection (last birth)	129	174	148	374
Mothers Received Medical Care at Delivery	5285	11578	9907	23387
Child Received Polio Vaccination	127	156	134	364

The above table represents the classical and Bayesian sample size for some variables for rural women using BDHS 2004 data set. This table describes that classical sample size is greater than Bayesian sample size of ACC and ALC criteria for some variables (no education, with secondary education or higher, want no more children) and the classical sample size is smaller than the Bayesian sample size of all criteria for the remaining variables in the above table. This indicates that here sampling is not done properly.

6 Conclusion

In statistics, before collecting data, it is essential to determine the sample size requirements of a study. In fact, three types of criteria usually will need to be specified to determine the appropriate sample size which are –

6.1 Sample Size for Different Methods

For determining appropriate sample size, two types of methods are available -classical and Bayesian method. In classical method, sample sizes are calculated using some formulas and in Bayesian method sample sizes are calculated using three different criteria (ACC, ALC and WOC). Normally, Sample sizes in classical method are larger than the sample sizes of Bayesian method for all criteria. In case of sample size for classical and Bayesian criteria,

$$n_{ALC} \leq n_{ACC} \leq n_{WOC} \leq n_{Classical}$$

For real life surveys, this equation is not exactly satisfied because of lacking of conventional proportional sampling methods.

6.2 Sample Size for Different Margin of Error

The Margin of Error, sometimes called sampling error, is the range in which the true value of the population is estimated to be. We know, the smaller the margin of error, the bigger the sample. In case of classical method of sample size determination, sample sizes also reduces for large margin of error and this statement is also satisfied by Bayesian method of sample size determination. For example, in table 1, for $d = 0.005$, classical sample size = 38416 and Bayesian ACC sample size = 27778, for $d = 0.025$, classical sample size = 1537 and Bayesian ACC sample size = 1107 and for d = 0.05, classical sample size = 384 and Bayesian ACC sample size = 275.Therefore, these sample sizes satisfy our statement.

6.3 Sample Size for Different Prior

Prior means information gathered from the previous study, past experience or expert opinion. The estimated sample size is increased as long as we approach to the informative prior from non-informative prior. For example, in table 1, for $l = 0.1$, ,the estimated sample size for prior $(1,1) = 275$, for prior $(2,2) = 312$, prior $(3,3) = 328$, prior $(4,4) = 336$. That means, sample size is increased if the value of the prior is increased, that is, if we have more information about our estimator (proportion). Consequently, we can say that the proper use of prior information illustrates the power of the Bayesian method of sample size determination.

References

1. Pham-Gia T.: 'Sample Size Determination in Bayesian Statistics-A Commentary', Journal of the Royal Statistical Society, 44, 163-166, 1995.
2. Joseph L., M'Lan C. E. , Wolfson D.B. : 'Bayesian Sample Size Determination for Binomial Proportions', International Society for Bayesian Analysis, 3, Number 2, 269-296, 2008.
3. Dr. Syed S. Hossain, Basics of sample size determination , Institute of Statistical Research and Training University of Dhaka, Bangladesh, July 15,2007.
4. Joseph L., Wolfson D. B., Berger R. D. : 'Sample Size Calculations for Binomial Proportions via Highest Posterior density Intervals', Royal Statistical Society, 44, No. 2, 143-154, 1995.
5. Sahu L.K., Smith T. M. F. : 'A Bayesian method of sample size determination with practical applications', Royal Statistical Society, 169, 2, 235-253, 2006.
6. National Institute of Population Research and Training (NIPORT), Mitra and Associates, and Macro International. 2005. Bangladesh Demographic and Health Survey 2004. Dhaka, Bangladesh and Calverton, Maryland, USA: National Institute of Population Research and Training, Mitra and Associates, and Macro International, 2004.
7. National Institute of Population Research and Training (NIPORT), Mitra and Associates, and Macro International. Bangladesh Demographic and Health Survey 2007. Dhaka, Bangladesh and Calverton, Maryland, USA: National Institute of Population Research and Training, Mitra and Associates and Macro International, 2007.
8. Cochran W. G. : Sampling Techniques, John Wiley & Sons, New York.
9. Joseph L., Wolfson D. B., Berger R. D. : 'Some Comments on Bayesian Sample Size Determination', Royal Statistical Society, 44, 2, 161-171, 1995.
10. Naing L., Winn T., Rusli B. N. : 'Some Practical Guidelines for Effective Sample-Size Determination', Archives of Orofacial Sciences, 1, 9-14, 2006.
11. Chen M. H., Shao Q. M. : 'Monte Carlo Estimation of Bayesian Credible and HPD intervals', Journal of Computational and Graphical Statistics, 7, 1998.
12. Chadha V. : 'Sample size determination in health studies', National Tuberculosis Institute, Bangalore, NTI Bulletin , 42/3&4, 55 -62, 2006.

Contributed Paper

- **Signal Processing**

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Face Recognition System Using Principle Component Analysis and Genetic Algorithm

Md. Shahjahan Kabir
Department of Computer Science and Engineering
Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh , skabir.ruet@gmail.com

Abstract- Identifying an individual from his or her face is one of the most nonintrusive modalities in biometrics. However, it is also one of the most challenging ones. This thesis discusses why it is challenging and the factors that a practitioner can take advantage of in developing a practical face recognition system. PCA, a statistical technique to reduce the dimensionality and are used to extract features with the help of covariance analysis to generate Eigen components of the images. Genetic Algorithm represents intelligent exploitation of a random search within a defined search space to solve a problem.

Keywords-Feature Selection, Face Recognition, Principle Component Analysis, Genetic Algorithm.

1. Introduction

Feature extraction also plays an important role in law enforcement forensic investigation, security access control system, security monitoring Banking system, Face Recognition, intelligent robotic, safety alert system based on eye lid movement & it has various other applications. To better use the Face Recognition for this purpose, compression of data is mandatory. Images can be compressed as structural features such as contours and regions [2]. Images have been exploited to encode images at low bit rates. The method of Eigen faces uses Principal Components Analysis (PCA)toa low dimensional subspace (Eigen space). This subspace is defined by the principal components of the distribution of face images. Each face can be represented as a linear combination of the Eigen faces. Given an image, sub-images of different size are extracted at every image location. To classify images to any other face, its distance from the Eigen space vector is computed. Face recognition can typically be used for verification or identification. In verification an individual image is already enrolled in the reference database. In identification, an input image is matched with a biometric reference in the database. There are two outcomes: the person is not recognized or the person is recognized. Two recognition mistakes may occur: false reject (FR) which indicates a mistake that occur when the system reject a known person, false accept (FA) which indicates a mistake in accepting a claim when it is in fact false. These algorithms are two types: two dimensional (2D) approaches and three dimensional (3D) approaches. Mainly, the traditional 2D approaches are divided into six algorithms: eigenfaces (PCA), linear discriminant analysis (LDA), independent component analysis (ICA), support vector machine (SVM), Genetic Algorithm (GA) neural network and hidden Markov model (HMM) [1].

Genetic algorithm was developed by John Holland- University of Michigan to provide efficient techniques for optimization [7].Human face recognition is currently a very active research area with focus on way to perform robust and reliable biometric identification. Face Recognition means matching the original face from a set of face previously stored in the knowledge base**.** In earlier times, people have tried to understand which features help us to perform recognition tasks, such as identifying a person [6].In the genetic algorithm, the problem to be solved is represented by a list of parameters which can be used to drive an evaluation procedure, called chromosome or genomes.

The face of our primary focus of attention in social intercourse, playing a major role in conveying identity and emotion. Although the ability infer intelligence or character from facial appearance is suspect. Face Recognition System convey to provide the intelligence power to the computer so that it can recognize the faces to identify and differ from other faces through the heuristically learned faces. In the area of surveillance, secure trading terminals, credit verification and criminal identification, Close Circuit Television (CCT) control and user authentication uses the Face Recognition system, it draws considerable interest and attention from many researchers. To ensure legal voter identification and avoid corruption and duplicates voters as well as illegal votes, face attached voter ID card is implemented by the Face Recognition system.

2. Methodology

In PCA, the probe and gallery images must be the same size. Each image is treated as one vector. All images of the training set are stored in a single matrix and each row in the matrix represents an image. The average image has to be calculated and then subtracted from each original image. Then calculate the eigenvectors and eigenvalues of the covariance matrix. These eigenvectors are calledeigenfaces. The eigenfaces is the result of the reduction in dimensions which removes the unuseful information and decomposes the face structure into the uncorrelated

components(eigenfaces). Each image may be represented as a weighted sum of the eigenfaces. A probe image is then compared against the gallery by measuring the distance between their represent vectors. Distance can be measure using city block distance, Euclidian Distance and so on.
Block Diagram of Face Recognition System using Principle component analyses are:

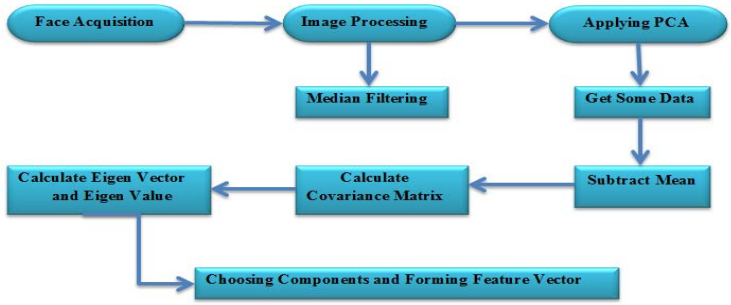


Fig.1. Block Diagram of PCA

A Genetic Algorithm is a problem solving method that uses genetics as its model of problem solving. It's a search technique to find approximate solutions to optimization and search problems. Fitness function is first derived from the objective function and used in successive genetic operations. Certain genetic operation requires that fitness function be non-negative, although certain operators don not have this requirement. A fitness function must be devised for each problem to be solved. Then, the genetic algorithm loops over an iteration process to make the population evolve. Each iteration consists of the following steps [5]: **Selection:** The first step consists in selecting individuals for reproduction. This selection is done randomly with a probability depending on the relative fitness of the individuals so that best ones are often chosen for reproduction than poor ones. **Reproduction:** In the second step, offspring are bred by the selected individuals. For generating new chromosomes, the algorithm can use both recombination and mutation. **Evaluation:** Then the fitness of the new chromosomes is evaluated. **Replacement:** During the last step, individuals from the old population are killed and replaced by the new ones. This cycle can be summarized as [8].

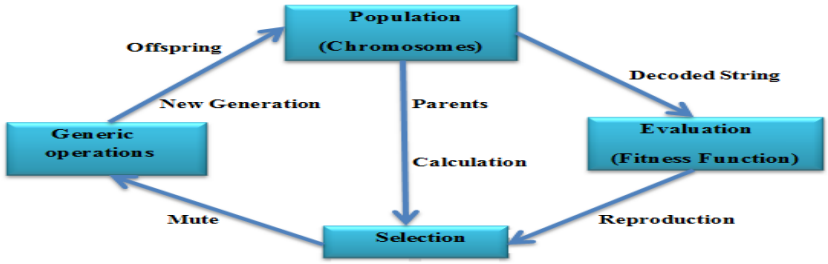


Fig.2. Cycle of Genetic Algorithm

The basic genetic algorithm is as following:

1. The algorithm begins by creating a random initial population.
2. The algorithm then creates a sequence of new populations. At each step, the algorithm uses the individuals in the current generation to create the next population. To create the new population, the algorithm performs the following steps:
    a. Scores each member of the current population by computing its fitness value.
    b. Scales the raw fitness scores to convert them into a more usable range of values.
    c. Selects members, called parents, based on their fitness.
    d. Some of the individuals in the current population that have lower fitness are chosen as*elite.* These elite individuals are passed to the next population.
    e. Produces children from the parents. Children are produced either by making random changes to a single parent—*mutation*—or by combining the vector entries of a pair of parents—*crossover.*
    f. Replaces the current population with the children to form the next generation.
3. The algorithm stops when one of the stopping criteria is met.

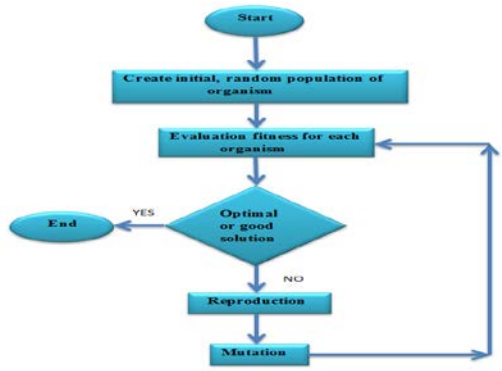


Fig.3: Flow chart of Genetic Algorithm

3. Approach

PCA: In this thesis, Japanese Female Facial Expression Standard Database has been used. A 2-D facial image can be represented as 1-D vector by concatenating each row (or column) into a long thin vector.*M* vectors of size *N* representing a set of sampled images. Obtain M training images it is very important that the images are centered. Represent each imageas a vector . The images are mean centered by subtracting the mean image from each image vector.

$$\Psi = 1/M \sum_{i=1}^{M} T_i \qquad C = N^2 * N^2$$

$$A = [\Phi_1, \Phi_2, \ldots\ldots \Phi_N] \qquad C = AA^T$$

Subtract the mean face from each face vector, to get a set of vectors,(1) $\Phi_i = T_i - \Psi$
Find the covariance matrix, wherebutAnd , C is huge.Findout $A = N^2 * M$
$L = A^T A = M * M$, can get V eigenvectorand then have to calculate the eigenvector of C using this equation,

$$U_i = AV_i \qquad (2)$$

$$W_j = U_j^T \Phi_i$$

Fig.4: Six Eigen faces

Weight calculation for each training image using this formula , at last calculated the individual weight vector for each training images

(3)
$$\Omega_i = \begin{bmatrix} w_1 \\ w_2 \\ w_N \end{bmatrix}$$

For test image also calculated , at the last step distance measured from test image to every training image. Using Euclidean or city blocks distance. Which distance is minimum that's are result

(4)
$$Eudist = \sqrt{(test - train)^2}$$

Test image Equivalent image

Fig. 5: Recognized image using PCA

GA: For the Genetic Algorithm first use the most frequent occur gray level in image that is shown in the bellow:

Fig. 6: Gray level Image

Calculate the frequency of each gray level. Then eliminate the gray values that have the probability less than 0.003. After the preprocessing the image, get the bellow figure for a sample image:

Fig.7: Preprocessed Image

At first take the input image and then take each database image. Assign the zero values for some rows and column of input image and database image. Find the deviation between both images. Zero values of 3 rows and column of the input image are assigned. Calculate the deviation values for five times. Each time the eliminated row and column of the database image increase 3 factors. 3 crossover factors and generation is 5 is considered. The resultant image is as following

Fig.8: The resultant Image after crossover between input image and the database image for five generation

Fitness values for each database and input images are find. If the total database image is 25 then we will find the 25 fitness values. Then take each fitness value for each database image and input image for each generation. At last add all fitness value. Sort the Fitness value. The minimum fitness value is selected. The corresponding index image is selected.

Test image Equivalent image

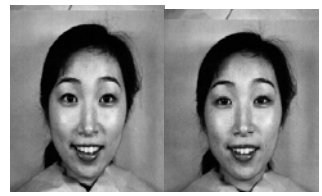


Fig.9: Recognized image using GA

4. Experimental Result And Analysis

Table 1: Efficiency calculation for PCA

No. of Face Image	Successfully Recognized Face Image	Unrecognized Face Image	Efficiency (%)
10	9	1	90%
20	18	2	90%
25	22	3	88%

Therefore the efficiency of Face Recognition System by using PCA is 89.33%.

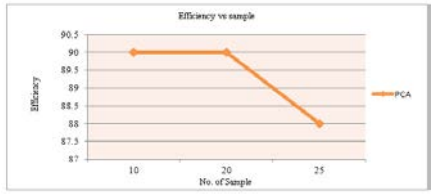


Fig.10: Efficiency versus No. of Sample graph using PCA

Table 2: Efficiency calculation for GA

No. of Face Image	Successfully Recognized Face Image	Unrecognized Face Image	Efficiency(%)
10	10	0	100%
20	19	1	95%
25	23	2	92%

Therefore the efficiency of Face Recognition System by using GA is 95.6%.

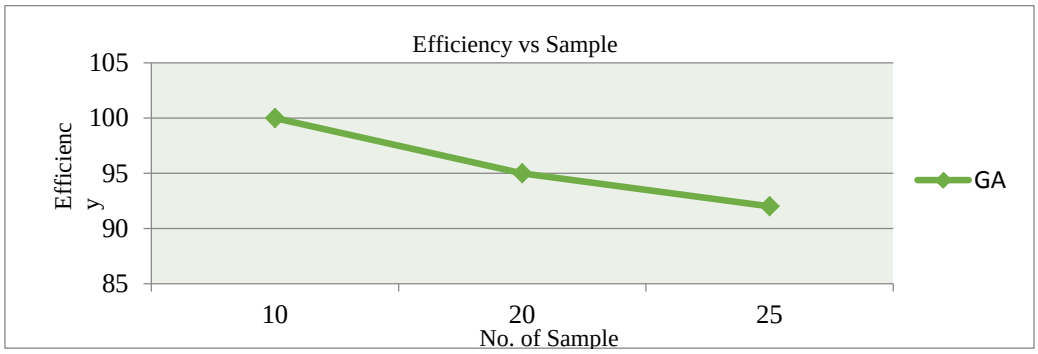


Fig.11: Efficiency versus No. of Sample graph using GA

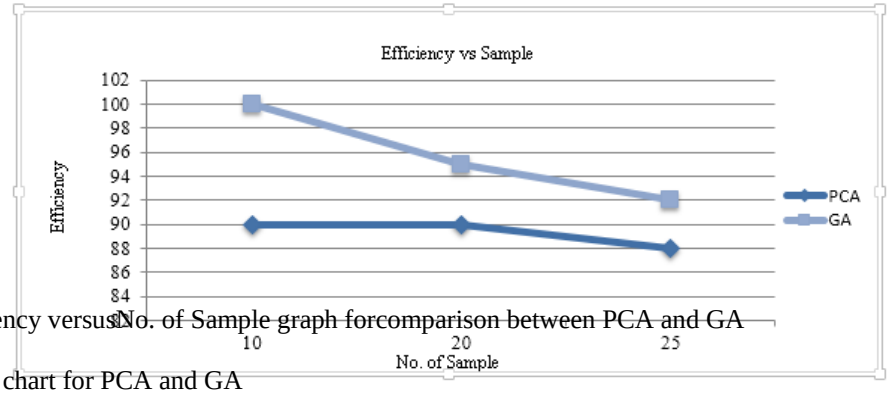
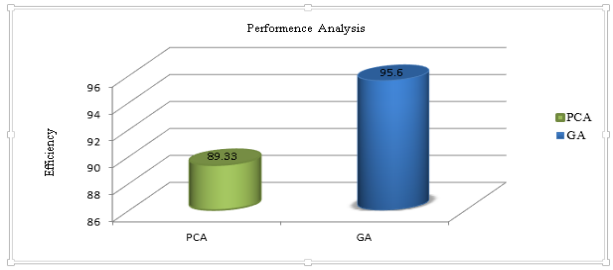


Fig.12: Efficiency versusNo. of Sample graph forcomparison between PCA and GA

Fig.13: Performance Analysis chart for PCA and GA

Draw the Performance analysis graph for Face Recognition System using PCA and GA. In above figure shown that GA is better than PCA. This Efficiency is so much higher than any other research [5] [2] [8].

5. Limitation



In **PCA,** the dimension of the resized image is critical: If it is too large, the system may diverge, whereas if it is too small, the images cannot be discriminated from each other. **GA** is a run time searching technique which takes a lot of time to search a image from train database.

6. Summary And Conclusions

Face Recognition System using the concept of Principle Component Analysis and Genetic Algorithm has been discussed. The maximum efficiency for Face Recognition System by using Genetic algorithm is 95.6% and the maximum efficiency for Face Recognition System by using PCA is 89.33%.The whole software is dependent on the database and the database is dependent on resolution of camera. So if goodresolution digital camera or good resolution analog camera is used, the results could be increased.

References

1. Manal Abdullah[1], Majda Wazzan[1] and Sahar Bo-saeed[2], "Optimizing Face Recognition using PCA",InternationalJournal of ArtificialIntelligence& Applications (IJAIA), March 2012.
2. MahimaAgrawal[1],Shubangi.D.Giripunje[2],P.R.Bajaj[3], "Recognizing Facial Expression using PCA andGeneticAlgorithm", International Journal of Computer & Communication Technology, April 2011.

3.Jonathon Shlens ,"A Tutorial on Principal Component Analysis", Systems Neurobiology Laboratory, Salk Insitutefor Biological Studies , April 22, 2009.

4. M. Thomas[1], S. Kumar[2], and C. Kambhamettu[3],"Face Recognition Using a Color PCA Framework", Universityof Delaware, Newark, Bell Laboratories, Lucent Technologies,November 05, 2007.

5. S.Venkatesan and Dr.S.Srinivasa Rao Madane,"Face Recognition System with Genetic Algorithm and ANT Colony Optimization",International Journal of Innovation, Management and Technology, December 2010.

6.Luigi RosaL.S. Ettore Majorana, "Face Recognition Based on Genetic Algorithms ForFeature Correlation", URL:http://www.advancedsourcecode.com

7. C.R Vimal Chand,"Face and gender Recognition Using Genetic Algorithm and Hopfield Neural Network", Global Journal of Computer Science and Technology, April 2010.

8. Sarawat Anam, Md. Shohidul Islam, M.A. Kashem, M.N. Islam, M.R. Islam, M.S. Islam, "Face Recognition Using Genetic Algorithm and Back Propagation Neural Network",Proceedings of the International MultiConference of Engineers and Computer Scientists (.Hong Kong),March 18 - 20, 2009

9. Ajoy Kumar Dey, Susmita Saha, Avijit Saha, Shibani Ghosh, "A Method of Genetic Algorithm (GA) for FIR Filter Construction: Design and Development with Newer Approaches in Neural Network Platform", (IJACSA)International Journal of Advanced Computer Science and Applications, May 11, 2010.

10. Rafael C. Gonzalez and Richard E Woods ,"Digital Image Processing"(2nd Edition).

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Robust and Diagnostic Statistics: A Few Basic Concepts in Mobile Mapping Point Cloud Data Analysis

Abdul Nurunnabi[†], David Belton[‡], Geoff West[‡]
Department of Spatial Sciences, Curtin University, Perth, Australia
Cooperative Research Centre for Spatial Information (CRCSI)
[†]abdul.nurunnabi@postgrad.curtin.edu.au, [‡]{d.belton, g.west}@curtin.edu.au

Abstract. It is impractical to imagine point cloud data obtained from laser scanner based mobile mapping systems without outliers. The presence of outliers affects the most often used classical statistical techniques used in laser scanning point cloud data analysis and hence the consequent results of point cloud processing are inaccurate and non-robust. Therefore, it is necessary to use robust and/or diagnostic statistical methods for reliable estimates, modelling, fitting and feature extraction. In spite of the limitations of classical statistical methods, an extensive literature search shows not much use of robust techniques in point cloud data analysis. This paper presents the basic ideas on mobile mapping technology and point cloud data, investigates outlier problems and presents some applicable robust and diagnostic statistical approaches. Importance and performance of robust and diagnostic techniques are shown for planar surface fitting and surface segmentation by using several mobile mapping real point cloud data examples.

Keywords: 3D modelling, covariance technique, feature extraction, laser scanning, M-estimator, mobile mapping technology, outlier detection, PCA, plane fitting, robust statistics, segmentation.

1. Introduction

The goal of Mobile Mapping (MM) is to acquire a 3D-survey of the environment and objects in the vicinity of the mapping vehicle accurately, quickly and safely. Applications of this emerging technology include 3D city modelling, corridor (e.g. road and rail) asset and inventory maintenance and management, abnormal load routes, environmental monitoring, accidental investigation, industrial control, construction management, digital terrain modelling, archaeological studies, marine and coastal surveying, telegeoinformatics, change detection for military and security forces, man-induced and natural disaster management, Geographical Information Systems (GIS) applications and simulation. The necessity and prospects of this technology related to the geospatial industry has been mentioned by Jack Dangermond (founder and president, ESRI) in his speech 'The geospatial industry empowers a billion+ people' [1].

Understanding, visualization, analysis and modelling from MM data rely on laser scanning point cloud data processing. Point cloud data acquired from MM Systems (MMS) are usually unorganized, multi-structured, incomplete and sparse. This type of data has no knowledge about statistical distribution and specific surface shape, has complex topology, geometrical discontinuities and inconsistent point density, may have sharp features and may even be missing pieces (such as holes). Besides, the physical limitations of the sensors, boundaries between 3D features, occlusions, multiple reflectance and noise can produce off-surface points that appear to be outliers [2]. Most of the classical statistical techniques work well only for high-quality data and fail to perform adequately in the presence of outliers. Robust and diagnostic statistics deal with the problem of outliers. The necessities of robust and diagnostic methods in computer vision, pattern recognition, photogrammetry, remote sensing and statistics have been well described in the literature (e.g. [3,4,5,6,7,8,9]). Stewart [9] states “It is important for the reader to note that robust estimators are not necessarily the only or even the best technique that can be used to solve the problems caused by outliers and multiple populations (structures) in all contexts”. In spite of the inevitable recognition and reality of the outlier (gross errors) problem in point cloud data processing, we see frequent use of non-robust statistical techniques for data analysis. The use of non-robust techniques causes undue influence on the estimators and hampers the accuracy.

The detection of outliers (or irregular data structures) and parameter estimation without the effects of outliers is one of the fundamental tasks in statistics and is as old as statistics. This paper does not cover robust statistics nor MM in detail. This paper has two main goals: one is to present the basic ideas of MM technology, laser scanning point cloud data and robust statistics, and the other is to point out the role and applications of some of the most applied robust and diagnostic statistical approaches in MM point cloud data analysis (e.g. surface fitting, segmentation and reconstruction).

2. Mobile Mapping Technology and Point Cloud Data

Mobile Mapping (MM) (i.e. mapping from moving vehicles (Fig. 1)) Systems (MMS) have been around for at least as long as photogrammetry has been practiced [10]. In the early 1990s, it was possible to think about MM in a different way when for the first time, GPS was able to provide reasonably accurate positioning of mobile sensor platforms [11]. The first research systems were simultaneously developed at the Center for Mapping, Ohio-State University, and the Department of Geomatics Engineering, University of Calgary [12]. MM is a non-invasive, state-of-the-art solution that incorporates various navigation and remote sensing technologies on a common moving platform. On board (Fig. 2), are advanced imaging and ranging devices (cameras, laser scanners or Light Detection and Ranging (LiDAR)) and navigation/positioning/geo-referencing devices (Global Navigation Satellite System (GNSS)) for the determination of the position of the moving platform, Inertial Measurement Unit (IMU) that contains sensors to detect rotation and acceleration and used for determining the local orientation of the platform, and odometer/Distance Measurement Instrument (DMI) typically connected to a wheel of the vehicle to provide linear distance in case of GNSS outage. The sensor arrangement works to maintain the alignment and accuracy between the sensors and the navigation equipment. The vehicle also includes a computer, storage and operational software to control the mapping mission. The mission is typically performed by a 2-person crew, one for driving the vehicle and the other for operating and managing the sensors. The two main components of MMS are geo-referencing based on navigation sensors (for details see e.g. [13]), and kinematic modelling of imaging sensors. Mid- and long-range laser scanners are usually based on time-of-flight technology, which measures time delays created by light waves (rather than radio waves) travelling in a medium from a source to a target surface and back to the source. Laser scanners mounted on the platform usually at a 45^o angle to the vehicle track swing the laser beam through 360^o. The unit can rotate at 80 to 200 revolutions per second. The laser is pulsed with the frequencies up to 200 kHz. Performance include spatial resolution up to 1 cm at 50 km/hour, range > 100 meters, measurement precision 7 mm (1 sigma), at operating temperatures -20^o C to + 40^o C (see Optech LYNX MMS). These configurations and advantages vary for different systems. High-speed counters measure the time-of-flight of the high energy and short length light waves. To get the desired results and to extract 3D coordinates of mapping objects from the geo-referenced images, modelling and data fusion required. Data fusion is necessary for merging data from various sources (sensors) of a different nature (see Fig. 3).

Figure 1.LYNX MMS vehicle [1]

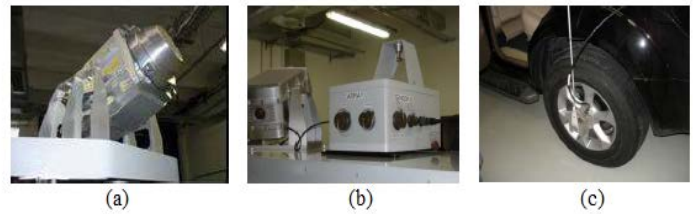
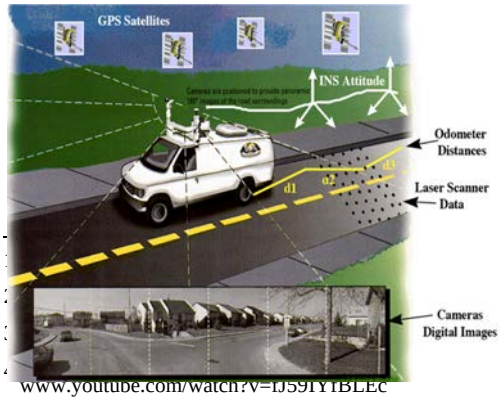


Figure 2. Configuration of the LYNX MMS [2] (a) LIDAR sensor (b) IMU (c) DMI

MMS provides efficient acquisition of multi sensor data of terrabytes of 3D points (geospatial data) defined by their *x*, *y*, and *z* coordinates (latitude, longitude and elevation) for each laser footprint called a point cloud (Fig. 3). Point cloud data may have colour (r, g, b) information with intensity (return energy) values. The output point cloud data is stored in a general industry standard format called 'LAS', which encodes the data into a point based binary file. The complexity of MMS has increased, reflecting the stronger dependency on technology and the more sophisticated design, data processing and information extraction process [10]. MMS are now able to collect more than 600,000 points per second. According to Graham [14], the achievable absolute accuracy of the 3D data can be as good as 2 cm (following adjustment to control). MMS significantly improves safety for data collectors (a major concern in highway work). No longer is 10 days of good weather needed to collect data for a 20-mile highway corridor. The same data can be acquired with a MMS in 30 minutes and all the data processing can be performed in the back office. Interested readers can acquire more information about MMS [12, 14, 15, 16]. There are also resources for understanding of data collection and visualization of point cloud data [3,4]. MMS data collection to output workflow is summarized in Figure 5.



www.youtube.com/watch?v=1J59IYIBLEc

Figure3. MMS data fusion [13]

pdf/Lynx

Figure 4. Point cloud data (with a single scanner, the main street of Wollongong, NSW, Australia) was scanned in under 10 minutes by StreetMapper. Features such as road markings, curbs, manholes and overhead wires are visible in the data [5]

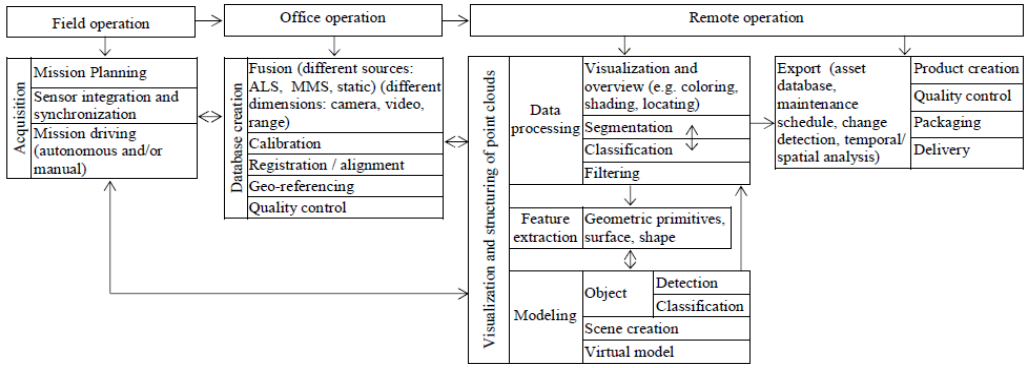


Figure 5. MMS workflow

3. Outlier, Robust and Diagnostic Statistics

People in different scientific disciplines (e.g. statistics, computer vision, pattern recognition, machine learning, data mining, photogrammetry and remote sensing) define the term 'outlier' in many ways (e.g. [2, 5, 6, 17]). A good answer of what are outliers and what are the problems of outliers is in statistics "In almost every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions, and the introduction of which into the investigations can only serve … to perplex and mislead the enquirer" [18]. Usually, when the sample size increases the sampling variability decreases. Large datasets such as mobile mapping point clouds may have very small variance and error may occur due to systematic bias of model misspecification. Agostinelli et al. [19] mention that two problems may arise to deal with large and complex data by classical statistical methods: (i) it may not be easy to fit simple and parsimonious models that reflect equally well all the data; and (ii) the sampling variability for such large datasets can be very small, to the extent that, the possible model misspecification bias dominates the statistical error, and may put into question the validity of the analysis. Robust statistical methods can deal with the above two challenges in the presence of outliers in a dataset. That means, when outliers and regular observation do not follow the same model.

There are two well known complementary approaches in statistics with the same objective for dealing with outliers: one is robust statistics and the other is diagnostics. Box [20] first introduces the technical terms 'robustness' and 'robust', and the subject matter was recognised as a legitimate topic in the mid-sixties, due to the pioneering work of Tukey [21], Huber [22] and Hampel [23]. The first monograph is 'Robust Statistics' [5]. The basic philosophy of robust statistics is to produce statistical procedures which are stable with respect to small changes in the data or model and even large changes should not cause a complete breakdown of the procedures [24]. Robustness of an estimator is a property, usually measured by the Breakdown Point (BP; the

[5] www.streetmapper.net/applications/citymodelling.htm

smallest fraction of outlier contamination that can cause an estimator to be arbitrarily far from the real estimate) and influence function, which measures the effect of an outlier. Diagnostics have taken a different view from robust statistics. Rather than modifying the fitting method, diagnostics condition on the fit using standard methods to attempt to diagnose incorrect assumptions, allowing the analyst to modify them and refit under the new set of assumptions [25]. According to Rousseeuw and Leroy [8], the purpose of robustness is to safeguard against deviation from the assumptions; the purpose of diagnostics is to find and identify deviation from the assumptions. It means that each views the same problem from opposite sides, and completing the metaphor, the more opaque the problem is, the more important it is to view the problem from the all sides. Fung [26] expresses the necessity of each other as: robust and diagnostic methods do not have to be competing but the complementary use of highly robust estimators and diagnostic measures provides a very good way to detect multiple outliers and leverage points. Hence, the goal of robust and diagnostic methods should be twofold: to identify outliers and to provide an analysis that has greatly reduced sensitivity to outliers.

3.1 Location and Scatter

The mean and standard deviation (SD) are the two most well known location and scale/scatter estimates respectively of the true values of an univariate random variable X. They are defined as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \mathrm{SD} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)} \,. \tag{1}$$

Suppose, we have five values of a sample: 5.56, 5.25, 5.38, 5.86 and 5.43. We want to estimate the true value, we usually get the mean = 5.496 from Eq. (1). If the 1st value is wrongly observed as 55.6 then we get mean = 15.504, which is far from the true value. We could consider another location measure e.g. median defined as the middle most observation of the sorted values. The median with an even number of sorted values is the average of the two elements in the middle. We calculate the median for both the cases (real set and after introducing the wrong value). To find the median, the arrangements are: for the real set $5.25 \le 5.38 \le 5.43 \le 5.56 \le 5.86$, and after contaminating by the wrong (outlying) value $5.25 \le 5.38 \le 5.43 \le 5.86 \le 55.6$. We get the median (middle; 3rd in sorted position) value unchanged, 5.43. We say median is a robust estimator whereas the mean is sensitive to outliers. We have to contaminate the dataset by at least 50% outlying observations to change the median value from real uncontaminated data. In the case of SD, we get the values 0.2318 and 22.4155 respectively for the real and contaminated datasets. That means SD is extremely non-robust. Therefore the BP of the mean and SD is merely 1/*n*, which tends to 0 if *n* tends to a large value. The robust alternative of SD is the Median Absolute Deviation (MAD) defined as:

$$\text{MAD} = a.\underset{i}{median}\ |x_i - \underset{j}{median}(x_j)|, \tag{2}$$

where*a* = 1.4826 is used to make the estimator consistent for the parameter of interest [27]. We calculate the MAD values and get 0.1927 and 0.2669 for the real and contaminated sample respectively. The result is quite reasonable in the presence of the outlier. Both the sample median and MAD have 50% BP. There are many scale estimators. A popular one is Inter Quartile Range (IQR) with 25% BP, which is the difference between the 3rd and 1st quartile of the data.

In the multivariate case, we can consider the data of *n* points of *m* dimensions that can be arranged in a matrix (*P*) of $n \times m$. The location and scatter (covariance matrix) for such multivariate data are defined as:

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i \quad \text{and} \quad \Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(p_i - \bar{p})^T(p_i - \bar{p});\ p_i \in R^m \tag{3}$$

They have 0% BP similar to the univariate location and scatter in Eq. (1). The estimation of multivariate location and scatter is one of the most important and challenging problems, especially in the presence of multiple/cluster outliers [3, 8, 28]. One of the main problems is most of the methods breakdown if the outlier contamination is larger than 1/(*m*+1). Many robust alternatives have been developed that can be categorized mainly into three approaches consisting of projection estimators (e.g. [29, 30]), combinatorial estimators such as Minimum Volume Ellipsoid (MVE), and Minimum Covariance Determinant (MCD) based estimators [28, 31] and iterative estimators such as Maximum Likelihood (ML) and M-estimators.

The idea of projection estimators is discussed in the next section. The MVE method finds the $h(h > n/2)-$ subset of data points for the ellipsoid with the smallest volume. Its BP is $(n-h)/n$. MCD finds observations in the dataset that have a covariance matrix with the minimum determinant. MCD has better statistical efficiency than MVE because it is asymptotically normal. The MCD estimators have a bounded influence function: $\text{BP} = (n-h+1)/n$, and attain maximum BP when $h = \lfloor (n+m+1)/2 \rfloor$. The MCD location and scatter are time intensive because ${}^{n}c_h$ subsets have to be investigated. However later the estimators have gained much popularity after the introduction of the fast-MCD [28]. Many other multivariate robust alternatives (e.g. M-estimators, S-estimators and MM-estimators) have been introduced in the literature [8, 32]. Some of them are based on the idea of a measure of outlyingness and use the re-weighted locations and scatters [33].

The M-estimator has been devised by Huber [22], and many others have enhanced it. M stands for Maximum Likelihood (ML) and the estimator robustifies the ML by down-weighting the extreme (outlying) values using a weight function. It is computationally more efficient and robust than the ML estimator. Assume x_i is the i^{th} observation of a random variable *X*, *T* is a location estimator, the residual r_i=x_i-*T* and a function defined as $\rho(r)$. Then the M-estimate of *T* can be obtained by minimizing

$$\sum_{i=1}^{n}\rho(r_i), \tag{4}$$

where $\rho(r)$ is continuous, positive definite $(\rho(r) \ge 0)$, symmetric $(\rho(r) = \rho(-r))$ and generally with a unique minimum at 0 (i.e. $\rho(r)$ increases as *r* increases from 0, but does not get too large). If $\rho = f$ (a probability density function) then the M-estimator becomes MLE. Differentiating w.r.t. the parameter *T* (location) yields:

$$\sum_{i=1}^{n}\psi(r_i)=0 \quad . \tag{5}$$

Different $\rho(r_i)$ and $\psi(r_i)$ yields different M-estimators. Huber [22] introduces the $\psi(r_i)$ as:

$$\psi(r)=\begin{cases}-k, & r<-k\\ r, & -k\le r\le k\\ k, & r>k,\end{cases} \tag{6}$$

where*k* is the tuning constant. The M-estimator has many applications in different multivariate techniques (e.g. robust regression). It has been used to develop the well known RANSAC algorithm [34] and its family of variants in computer vision (e.g. MSAC, MLESAC; [35]). M-estimator based segmentation has been also been used in data visualization.

3.2 Outlier Detection

The outlier detection approaches can be categorized into two: univariate and multivariate. A simple rule for outlier identification in *R* (set of real numbers) is to find observations that are three SD apart from the mean of the dataset. If all observations are more than *b.*SD (*b* is a positive real number) from the mean are classified as outliers then this rule has failed to identify a proportion of $b/(1+b^2)$ outliers with the same sign [24]. Univariate location and scatter are often used to get the so-called *z*-score for identifying outliers defined as:

$$z_i=\frac{|x_i-\bar{x}|}{\mathrm{SD}(x)} \quad . \tag{7}$$

The robust alternative of the *z*-score is the insertion of robust location and scale to form:

$$Rz_i=\frac{|x_i-median(x)|}{\mathrm{MAD}(x)} \quad . \tag{8}$$

The outlying observations greatly exceed 2.5 or 3 times the cut-off value given the normality assumptions of the *z*-scores.

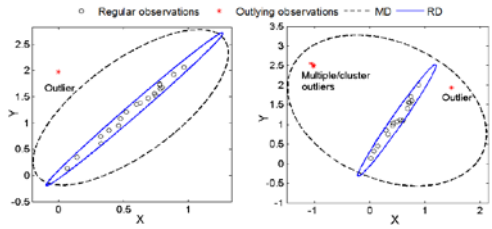
Outliers in multivariate (e.g. point cloud) data can be hard to detect because the dimensionality exceeds 2, and we can not rely always on human interpretation and analysis. These are inconsistent with the correlation structure in the data. The Mahalanobis Distance (MD; [36]) is probably the most popular multivariate outlier detection method defined as:

$$\mathrm{MD_i}=\sqrt{(p_i-\bar{p})^T\Sigma^{-1}(p_i-\bar{p})} \quad . \tag{9}$$

This method no longer suffices in the presence of multiple (cluster) outliers because of masking and swamping effects. Masking occurs when an outlying subset goes undetected because of the presence of another, usually adjacent, subset. Swamping occurs when good observations are incorrectly identified as outliers because of the presence of another, usually remote subset of observations [37]. The weakness of MD is the use of non-robust location and scatter. Robust Distance (RD) has been introduced by the alternative inclusion of robust (e.g. MVE and MCD based) locations and scatters. MCD based RD is more precise than MVE based RD and better suited to expose multivariate outliers [28]. The most popular one is the fast-MCD based RD, defined as:

$$\mathrm{RD_i}=\sqrt{(p_i-\bar{p}_{MCD})^T\Sigma_{MCD}^{-1}(p_i-\bar{p}_{MCD})} \quad , \tag{10}$$

where $\bar{p}_{MCD}$ and $\Sigma_{MCD}$are the MCD based location and scatter (covariance matrix) respectively. MD and RD both follow a Chi-square distribution [31]. Observations that have MD or RD values greater than $\sqrt{\chi^2_{m.0.975}}$ are treated as outliers. Fig. 6 illustrates outlier detection in two dimensions with MD and RD ellipses, the results based on 20 observations with a single outlier (red point; Fig. 6(a)) and 4 multiple outliers (Fig. 6(b)). MD (black-dotted ellipse) and RD (blue ellipse) are both successful in identifying one single outlier but MD totally failed in presence of multiple outliers and even its ellipse' direction is affected to the outliers. RD is successful at identifying all 4 outliers without its direction being affected (Fig.6(b)). The figures show that both MD and RD perform well when the regular observations follow an elliptical symmetric pattern but in reality this does not happen all the time. This has been addressed by Hubert and Veeken [38] who propose robust outlier detection methods for skewed data.



(a) (b)

Figure 6. Outlier (red) detection by MD and RD (a) one outlier (b) multiple/cluster outliers

Among many other multivariate outlier detection techniques, one well known method is the projection pursuit method, which tries to analyse the structure of a dataset by various projections into a lower dimensional space [8]. The idea is if a point is a multivariate outlier, then there will be a projection into one dimension in which the point will be a univariate outlier [33]. Stahel [29] and Donoho [30] independently developed this technique. The so-called Stahel-Donoho outlyingness measure looks for a univariate projection for multivariate data that makes an observation an outlier. It is defined as:

$$out(p_i) = \max_{\|v\|} \frac{|p_i v^T - median_j(p_i v^T)|}{\mathrm{MAD}_j(p_i v^T)}, \quad (11)$$

where v is a direction expression, $p_i v^T$ denotes a projection of the i^{th} point (p_i) onto the direction v and the max is over all possible projections v in R^m. One can use alternative robust locations and scatters to get the variants of the measure. This type of measure is also used for getting re-weighted robust location and scatter.

Apart from the above methods, there are many multivariate statistical techniques (e.g. regression analysis, support vector machines, classification and clustering) that are often used in point cloud data analysis. We briefly discuss here Principal Component Analysis (PCA) and its robust counterpart Robust PCA (RPCA) because PCA has many uses in point cloud processing for saliency feature (e.g. normal and curvature) estimation, surface reconstruction, geometric primitives fitting, feature extraction and visualization [39, 40].

3.3 Principle Component Analysis and Robust Principal Component Analysis

PCA is the most popular dimension reduction and visualization non-parametric multivariate statistical technique. It works to identify a smaller number of mutually orthogonal variables than the original set of variables. Principal Components (PCs) are the linear combinations of the original variables that rank the variability in a dataset through the variances, and produces corresponding directions using the eigenvectors of the covariance matrix. Every PC describes a part of the data variance not explained by the others, and those corresponding to the largest eigenvalues describes larger variances than the smaller ones. The well known mathematical technique: Singular Value Decomposition (SVD) is used to get the required number of PCs (eigenvectors) and corresponding eigenvalues explains how much variance is explained by the corresponding PC. Although PCA has many applications in point cloud processing, it is well known that it is sensitive to outliers.

Many robust PCA variants exist in the literature [41, 42, 43]. We briefly discuss the RPCA method [42] which is used later in this paper for planar surface fitting and segmentation in point cloud data. The authors combine the idea of Projection Pursuit (PP) [44] with the fast-MCD [28]. The PP is used to transform the data so that it lies in a subspace whose dimension is less than the total number of observations, and then the fast-MCD estimator is used to get the robust location and covariance matrix. Computationally, first the data is compressed to the PCs defining potential directions. Then, each i^{th} direction is scored by its corresponding value of outlyingness:

$$w_i = \arg\max_{\|v\|=1} \frac{\left|p_i v^T - \overline{p}_{MCD}(p_i v^T)\right|}{\Sigma_{MCD}(p_i v^T)}, \quad (12)$$

where the maximum is over all directions, v is a univariate direction and $p_i v^T$ denotes a projection of the i^{th} observation on the direction v. For every direction a robust location ($\overline{p}_{MCD}$) and scatter (Σ_{MCD}) of the projected data points ($p_i v^T$) is computed. Second, a fraction ($h > n/2$) of observations with the smallest values of w_i are used to construct a robust scatter Σ. Finally, robust PCA projects the data points onto the k-dimensional subspace spanned by the k largest eigenvectors (PCs) (depends on the objective, e.g. k=2 for plane fitting) of the Σ and computes their location and scatter by the re-weighted MCD.

While computing the robust PCA, we can detect two types of outliers in a diagnostic way (see Fig. 7). One outlier is an orthogonal outlier that lies away from the subspace spanned by the k (e.g. k=2 for plane) PCs and is identified

by a large orthogonal (residual) distance (OD) that is the distance between the observation p_i and its projection $\hat{p}_i$ in the k-dimensional PCA subspace:

$$\mathrm{OD}_i = \| p_i - \hat{p}_i \| . \tag{13}$$

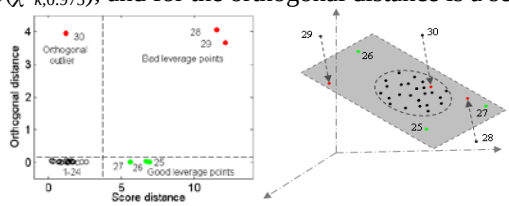
The other type of outlier is identified by the Score Distance (ScD) that is measured within the PCA subspace and defined as:

$$\mathrm{ScD}_i = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{l_j}} , \tag{14}$$

where t_{ij} is the ij^{th} element of the score matrix:

$$T_{n,k} = (P_{n,m} - 1_n \bar{p}_{MCD}) R_{m,k} , \tag{15}$$

where $P_{n,m}$ is the data matrix, 1_n is the column vector with all n components equal to 1, $\bar{p}_{MCD}$ is the robust location, $R_{m,k}$ is the matrix generated by the robust PCs, and l_j is the j^{th} eigenvalue of the robust scatter matrix Σ_{MCD}. The cut-off value for the score distance is $\sqrt{(\chi^2_{k,0.975})}$, and for the orthogonal distance is a scaled version of χ^2 [42].



(a) (b)

Figure 7. (a) Diagnostic plot; orthogonal distance versus score distance (b) fitted plane, green points are outliers in terms of score distance, and red points are orthogonal outliers

3.4 Covariance and PCA for Planar Surface Fitting in Point Cloud Data

It is highly probably that the plane is the most observed geometric shape among man-made objects. Here, we consider planar surface fitting as a part of feature extraction in point cloud data analysis. PCA has been used for planar surface fitting and local saliency features estimation in point cloud data [39, 40]. In the case of plane fitting, the first two PCs can be used to form a basis for the plane, and since the third PC is orthogonal to them, it defines the normal to the fitted plane. The first two PCs explain the larger variability as much as possible in two directions, hence the fitted plane is the best linear approximation to the data. The third PC explains the least amount of variation for 3D point cloud data and could be used to estimate the parameters for the fitted plane. Assume a sample of the data points $\{p_i(x_i, y_i, z_i); i = 1,2,\cdots,n;$ i.e. $p_i \in P \in R^3\}$ in a point cloud (P) is used to fit a plane. The plane equation is defined as:

$$ax + by + cz + d = 0 , \tag{16}$$

where a, b, c and d (distance from origin to plane) are the plane parameters. The Total Least Squares (TLS) method can estimate the parameters by minimizing the sum of squared orthogonal distances between the point and the plane, i.e.

$$\min_{c,\|n\|=1} \sum_{i=1}^{n} ((p_i - \bar{p})^T . \hat{n})^2 , \tag{17}$$

where $(p_i - \bar{p})^T . \hat{n}$ is the orthogonal distance between a point and the plane, $\bar{p}$ is the centre and $\hat{n}$ is the unit normal to the plane.

PCA is performed to get the parameters based on the covariance matrix in Eq. (3). The points in Σ may be considered as a local neighbourhood if the user's interest is local planar surface fitting or determining local saliency features (e.g. normal and curvature). For the local neighbourhood, Σ can define local geometric information of the underlying surface of the point cloud. By PCA, Σ is decomposed into PCs arranged in decreasing order of the eigenvalues: $\lambda_2 \geq \lambda_1 \geq \lambda_0 \geq 0$ with the corresponding eigenvectors v_2, v_1 and v_0. Thus v_0 approximates the surface normal $\hat{n}$ for the plane and could be used as the estimates of the plane parameters. Based on the λ values, Pauly et al. [40] define the curvature (rate of change of surface normal) as:

$$\sigma_p = \frac{\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2} . \tag{18}$$

3.5 Point Cloud Segmentation

Segmentation is the process of separating and labelling the most homogenous/similar and spatially close points into a number of separate surfaces. It plays a key role in surface reconstruction, modelling, feature extraction and object recognition, interrelated tasks in areas including computer vision, pattern recognition, reverse engineering, photogrammetry and remote sensing [45, 46, 47].

The region growing segmentation approach is one of the most popular segmentation methods in point cloud processing. Region based approaches use local surface neighbourhoods to combine nearby points that have similar properties (e.g. orientation) to obtain homogeneous regions and consequently find dissimilarity between the different regions [48, 49]. In this section, we briefly discuss a recently proposed region growing based segmentation algorithm [49]. The algorithm begins by searching a seed point which has the least curvature value among the point cloud (P) data of interest. Since the algorithm considers local surface point proximity and coherence criteria, it needs to find k neighbours (k neighbourhood NP_i) of the seed point. It calculates the Orthogonal Distance (OD) for the i^{th} seed point (p_i) to its best-fit-plane and the Euclidian Distance (ED) between the seed point (p_i) and one of its neighboursp_j. OD is defined as:

$$\mathrm{OD}(p_i) = (p_i - \overline{p})^T . \hat{n} \quad , \tag{19}$$

where $\overline{p}$ and $\hat{n}$ are the centroid and the unit normal of the best-fit-plane. ED is defined as:

$$\mathrm{ED}_{ij} = \left\| p_i - p_j \right\| \quad , \tag{20}$$

wherep_i and p_j are the seed point and its j^{th} neighbour respectively. Besides local surface point proximity criteria, the algorithm considers the angle between two neighbours to determine points on the same surface.
The angle (θ) between two points ($p_i$and p_j) is defined as:

$$\theta_{ij} = \arccos \left| \hat{n}_i^T . \hat{n}_j \right| \quad , \tag{21}$$

where $\hat{n}_i$ and $\hat{n}_j$ are the unit normals for the i^{th} seed point and one of its neighbours p_j. Two spatially close points are considered as co-surface points if θ_{ij} is less than a predefined angle threshold θ_{th}. Finally, the algorithm grows region by adding more neighbour points to the current region R_c and to the current seed point list S_c using the following three conditions:

$$\text{(i) } \mathrm{OD}(p_i) < \mathrm{OD}_{\mathrm{th}} \quad \text{(ii) } \mathrm{ED}(p_i) < \mathrm{ED}_{\mathrm{th}} \quad \text{and (iii) } \theta_{ij} < \theta_{th} \quad , \tag{22}$$

where threshold $\mathrm{OD}_{\mathrm{th}} = median\{\mathrm{OD}(NP_i)\} + 2 \times \mathrm{MAD}\{\mathrm{OD}(NP_i)\}$ and $\{\mathrm{OD}(NP_i)\}$ is the set of all ODs for all the points in the neighbourhood, and MAD is defined in Eq. (2), threshold $\mathrm{ED}_{\mathrm{th}} = median\{\mathrm{ED}(p_{ij})\}$; $\{\mathrm{ED}(p_{ij})\}$ is the set of all EDs between the seed point and its neighbours.

Now the point p_j is removed from P and considered as the next seed point for the region R_c, R_c then grows until no new point is available in S_c. After completing region R_c, the next seed point is selected for a new region from the remaining points in P with the least curvature σ_p value. If a region size is less than R_{min} (minimum number of points) then the region is considered as an insignificant segment.This process of region growing continues until the P is empty.

4. Experiments

The following experiments explore the limitations of classical PCA and the advantages of robust PCA for local planar surface fitting and point cloud segmentation.

4.1 Planar Surface Fitting

Artificial data

To illustrate outlier effects on PCA, we simulate a 3D dataset of 20 points shown in Fig. 8 (a), 19 of which (black points) follow a multivariate Gaussian distribution with means (x=2, y=8, z=6) and variances (x=15, y=15, z=0.01). We include just 1 (5%) artificial outlier (red point) that is generated from a different multivariate Gaussian distribution with means (x=10, y=15 z=10) and variances (x=10, y=2, z=1). Fig. 8 (a) clearly shows the outlier is far form the bulk of the data. The fitted planes show that PCA is highly influenced by the outlier. The outlier effect is so severe for PCA that the robustly fitted plane (with or without outliers) is almost perpendicular to the PCA fitted plane. It shows that only one outlier may have sufficient influence to significantly affect the PCA estimation. To compare the accuracy between robust (RD and RPCA) and non-robust methods, we simulate 1000 datasets of 50 points with different percentage (1 to 20) of outliers from 3D Gaussian distributions as in earlier fashion. Fig. 8 (c) shows RD and RPCA based bias angles are significantly less than for PCA based results. Box plots of average bias

angles from 1000 samples for the dataset of 50 points with 15% outliers, Fig. 8 (b), show the robustness of the methods and explore the advantage of robust methods.

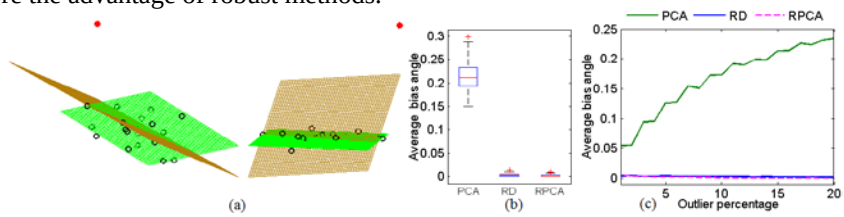


Figure 8. (a) Outlier (red point)'s influence on PCA for plane fitting at two different views (b) box plots of average θ (c) line plot of average θ versus outlier percentages

Real Point Cloud Data

We used laser scanner 3D point cloud data (Fig. 9(a)) with 2,747 points. This dataset, acquired using a MMS, represents a road side wall that we consider a planar surface. Fig. 9(b) shows the orientations of the fitted planes. The orientation of the plane determined by PCA (green) is significantly different because of its sensitivity to outliers. The plane fitted by RPCA is close to the same orientation as the real plane. In Fig. 9(d) the points fitted by the plane from PCA show that all outliers are taken as regular points in the plane. The RPCA based diagnostics finds 236 outliers (red points) by score distance and 713 outliers (blue points) by orthogonal distance (Fig. 9(c)). With the remaining inlier (regular) points we get the RPCA plane (Fig. 9(e)), which is free from outliers.

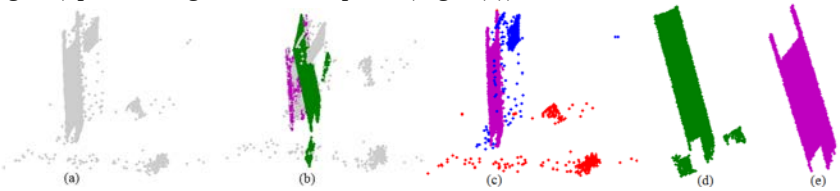


Figure 9. (a) Real point cloud data (b) plane orientation by PCA (green) and RPCA (magenta) (c) Red and blue points (outliers) detected by SD and OD respectively (d) PCA plane (e) RPCA plane

4.2 Point Cloud Segmentation

This dataset consists of 18,191 points for one road side building (a shop), again captured from a MMS. We add 2% outliers (Gaussian noise; red points) to the real data (Fig. 10(a)). We use the segmentation algorithm [49] described in Section 3.5. We set k=30, θ_{th}=10°, and R_{min}=10. Fig. 10(b) shows the presence of both over and under segmentation for PCA based results. We employed the RD based outlier detection method to find outliers in a local neighbourhood NP_i. After the deletion of outliers in a local neighbourhood we calculate necessary normals and curvatures based on the regular points. Using these local saliency features we grow the regions. Results in Fig. 10(c) show better segmentation for RD. Fig. 10(d) shows RPCA gives consistent segmentation and significantly better results than PCA and even RD. The different features around the windows as well as the vertical window bars have been separated from the main building and the umbrella poles have been segmented as complete features. Segmentation performance in Table 1 show RPCA performs significantly better than RD and PCA, and robust statistical approach based segmentation outperforms classical PCA. Robust approaches reduce over and under segmentation with high accuracy.

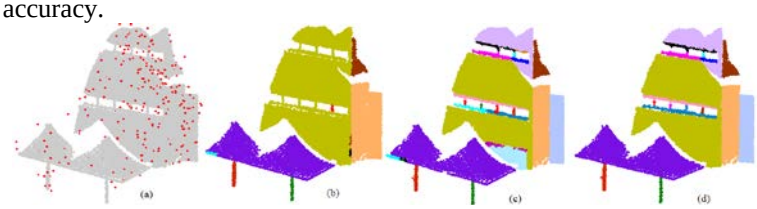


Figure 10. (a) Real point cloud data with 2% outliers (noise) (b) PCA segmentation (c) RD segmentation (d) RPCA segmentation

Table 1. Segmentation performance

Total segments	Segmentations	Methods	Proper segments	Over segments	Under segments
20	Figure 10. (b)	PCA	4 (20%)	3	15
	Figure 10. (c)	RD	14 (70%)	7	0
	Figure 10. (d)	RPCA	16(80%)	0	2

5. Conclusions

This paper introduces mobile mapping technology as a source for spatial information and some basic concepts about robust and diagnostic statistical methods that are commonly used in point cloud data analysis. Such methods are needed because of noisy data and the large volumes of data that render manual techniques for feature extraction impractical. Results show that robust PCA based local (planar) surface fitting and segmentation outperforms classical PCA based methods. Classical methods are affected by outliers and give unreliable and non-robust results. Results demonstrate that noisy point cloud data is infeasible to produce useful results. Hence using robust statistical methods is recommended for reliable analysis and more accurate feature extraction and point cloud processing. Further work is being carried out on extending the approach to deal with non-planar and free-form surfaces for more comprehensive and useful results.

Acknowledgements

This study has been carried out as a PhD research supported by a Curtin University International Postgraduate Research Scholarship (IPRS). The work has been supported by the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme. We are also thankful to McMullen Nolan and Partners Surveyors for the real point cloud datasets.

References

[1] 'Spatial Sustain', http://www. sensysmag.com; spatialsustain/the-geospatial-industry-empowers-a- billion-people.html, 2012, accessed: 27-05-2012

[2] Sotoodeh, S.: 'Outlier detection in laser scanner point clouds', In IAPRS, Dresden, XXXIV, 297–301, 2006

[3] Hampel, F., Ronchetti, E., Rousseeuw, P. J., and Stahel, W.: *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley, 1986

[4] Hampel, F. R.: 'Robust statistics: a brief introduction and overview', Invited talk in the Symposium "Robust Statistics and Fuzzy Techniques in Geodesy and GIS" held in ETH Zurich, March 12–16, 2001

[5] Huber, P. J.: *Robust Statistics*, New York: John Wiley, 1981

[6] Meer, P.: 'Robust techniques for computer vision', In *Emerging Topics in Computer Vision*, (Eds). Medioni, G. and Kang, S. B., Prentice Hall, 2004

[7] Nurunnabi, A., Belton, D., and West, G.: 'Diagnostic-robust statistical analysis for local surface fitting in 3d point cloud data', *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 1-3, 2012, 269–274, and presented in the XXII Congress of the ISPRS, 25 August –02 September, Melbourne, Australia, 2012

[8] Rousseeuw, P. J., and Leroy, A.: *Robust Regression and Outlier Detection*, New York: John Wiley, 1987

[9] Stewart, C. V.: 'Robust parameter estimation in computer vision', *SIAM Review*, 41(3), 513–537, 1999

[10] Schwarz, K. P. and EI-Sheimy, N.: 'Digital Mobile Mapping Systems - state of the art and future trends', In *Advances in Mobile Mapping Technology*, Eds., Tao, C. V. and Li, J. , London: Taylor and Francis Group, 3–18, 2007

[11] Novak, K.: 'Data collection for multi-media GIS using mobile mapping systems', GIM, 7(3), 30–32, 1993. Optech Inc., http://www.optech.ca/prodaltm.html

[12] Toth, C. K.: 'R & D of mobile LiDAR mapping and future trends', In ASPRS annual conference, Baltimore, Maryland, 9–13 March, 2009.

[13] El-Sheimy, N.: 'An overview of mobile mapping system', From Pharaohs to Geoinformatics, FIG working week 2005 and GSDI-8, Cairo, Egypt, April 16–21, 2005

[14] Graham, L.: 'Mobile mapping system overview', Photogrammetric Engineering and Remote Sensing, March 2010, 222–228, 2010

[15] Petrie, G.: 'An Introduction to the Technology Mobile Mapping Systems', *Geoinformatics*, January /February 2010, www.geoinformatics.com.

[16] Tao, C. V., and Li, J.: *Advances in Mobile Mapping Technology*, London: Taylor and Francis Group, 2007

[17] Breuning, M. Kriegel, H. P., Ng, R., and Sander, J. L. O. F.: 'Identifying density-based local outliers', In Proceeding of the 2000 ACM SIGMOD International Conference on Management of Data, New York: ACM Press, 93–104, 2000

[18] Barnett, V., and Lewis, T. B.: *Outliers in Statistical Data*, New York: John Wiley, 1995

[19] Agostinelli, C., Filzmoser, P., and Salibian-Barrera, M.: 'Final report', International Workshop on Robust Statistics and R, October 28-November 2, 2007, www.birs.ca/workshops/2007/07w5064/report07w5064.pdf

[20] Box, G. E. P.: 'Non-normality and tests on variance', *Biometrika*, 40, 318–335, 1953

[21] Tukey, J. W.: *A survey of sampling from contaminated distributions: Contribution to Probability and Statistics*, Eds., I. Olkin et al., California, Stanford: Stanford University press, 1960

[22] Huber, P. J.: 'Robust estimation of location parameter', *Annals of Mathematical Statistics*, 35, 73–101, 1964

[23] Hampel, F. R.: *Contributions to the Theory of Robust Estimation*, Ph D Thesis, University of California, Barkeley, 1968

[24] Davies, P. L., and Gather, U.: 'Robust Statistics', Papers/Humboldt-Universitat Berlin, Center for Applied Statistics and Economics (CASE), No. 2004, 20, 2004, http://hdl.handle.net/10419/22194

[25] Stahel, W., and Weisberg, S.: *Direction in Robust Statistics and Diagnostics*, Preface, New York:Springer-Verlag, 1991

[26] Fung, W-K.: 'Unmasking outliers and leverage points: a confirmation', *Journal of the American Statistical Association*, 88 (422), 515–519, 1993

[27] Rousseeuw, P. J., and Croux, C.: 'Alternative to the median absolute deviation', *Journal of the American Statistical Association*, 88 (424), 1273–1283, 1993

[28] Rousseeuw, P. J., and Driessen, K. V.: 'A fast algorithm for the minimum covariance determinant estimator', *Technometrics*, 41(3), 212–223, 1999

[29] Stahel, W. A.: 'Breakdown of Covariance Estimators', Research Report 31, Fachgruppe fur Statistik, E.T.H. Zurich, 1981

[30] Donoho, D. L.: *Breakdown Properties of Multivariate Location Estimators*, Ph.D. qualifying paper, Harvard University, 1982

[31] Rousseeuw, P. J., and van Zomeren, B. C.: 'Unmasking multivariate outliers and leverage points', *Journal of the American Statistical Association*, 85(411), 633–639, 1990

[32] Maronna, R. A., and Yohai, V. J.: 'Robust estimation of multivariate location and scatter', *Encyclopedia of Statistics*, 2, New York: John Wiley, 1998
[33] Maronna, R., A., and Yohai, V. J.: 'The behavior of the Stahel-Donoho robust multivariate estimator', *Journal of the American Statistical Association*, 90 (429), 330–341, 1995
[34] Fischler, M. A., and Bolles, R. C.: 'Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography', *Communications of the ACM*, 24, 381–395, 1981
[35] Torr, P. H. S., and Zisserman, A.: 'MLESAC: A new robust estimator with application to estimating image geometry', *Journal of Computer Vision and Image Understanding*, 78(1), 138–156, 2000
[36] Mahalanobis, P. C.: 'On the generalized distance in statistics', In Proceedings of The National Institute of Science In India, 12, 49–55, 1936
[37] Hadi, A. S., and Simonoff, J. S.: 'Procedures for the identification of outliers', *Journal of the American Statistical Association*, 88, 1264–1272, 1993
[38] Hubert, M., and Veeken, S. V.: 'Outlier detection for skewed data', *Journal of Chemometrics*, 22, 235–246, 2008
[39] Hoppe, H., De Rose, T., and Duchamp, T.: 'Surface reconstruction from unorganized points', InProceedings of ACM SIGGRAPA, 26(2), 71–78, 1992
[40] Pauly, M., Gross M., and Kobbelt, L. P.: 'Efficient simplification of point sample surface',InProceeding of the Conference on Visualization,Washington, D.C., 163–170, 2002
[41] Maronna, R. A.: 'Principal components and orthogonal regression based on robust scales', *Technometrics*, 47(3), 264–273, 2005
[42] Hubert, M., and Rousseeuw, P. J.: 'ROBPCA: A new approach to robust principal component analysis', *Technometrics*, 47(1), 64–79, 2005
[43] Hubert, M., Rousseeuw, P. J., and Verdonck, T.: 'Robust PCA for skewed data and its outlier map', *Computational Statistics and Data Analysis*, 53, 2264–2274, 2009
[44] Friedman, J., and Tukey, J.: 'A projection-pursuit algorithm for exploratory data analysis', *IEEE Transaction on Computers*, 23, 881–889, 1974
[45] Besl, P. J., and Jain, R. C.: 'Segmentation through variable- order surface fitting', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10 (2), 167–192, 1988
[46] Fan, T-J., Medioni, G. and Nevatia, R.: 'Segmented descriptions of 3-D surfaces', *IEEE Journal of Robotics and Automation*, RA-3 (6), 527–538, 1987
[47] Liu, Y., and Xiong, Y.: 'Automatic segmentation of unorganized noisy point clouds based on the Gaussian map', *Computer-Aided Design*, 40, 576–594, 2008
[48] Nurunnabi, A., Belton, D., and West, G.: 'Robust segmentation for multiple planar surface extraction in laser scanning 3D point cloud data', In the Proceedings of 21st International Conference on Patten Recognition (ICPR'12), 11–15 November, 2012, Tsukuba, Japan, 1367–1370, 2012
[49] Nurunnabi, A., Belton, D., and West, G.: 'Robust segmentation in laser scanning 3D point cloud data', International Conference on Digital Image Computing: Techniques and Application (DICTA'12), 3–5 December, 2012, Fremantle, Australia, In Press, 2012

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Performance of Wavelet transformation, Fourier transformation, and Singular Value Decomposition (Classical and Robust) in Image Compression and Denoising

Md. Tofazzal Hossain[1], Nishith Kumar[2] and Mohammed Nasser[3]

[2]DepartmentofStatistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University

[3]Department of Statistics, University of Rajshahi

Abstract. Now-a -days for several purposes we have to store, transmit and analyze huge images that necessitate image compression and denoising. There are many techniques for image compression and denoising such as Wavelet transformation, Fourier transformation, Classical and Robust Singular value decomposition, Neural networks, Principal Component Analysis etc. Since in literature no comparison among Wavelet transformation, Fourier transformation, Singular value decomposition have been found, in this paper, an attempt has been made to compare Fourier transformation, Wavelet transformation and singular value decomposition (SVD), the three most popular as well as influential techniques in image compression. Again in image denoising, the performance of Fourier transformation, Wavelet transformation and Classical and Robust singular value decomposition has been compared. As there are many Wavelet functions namely Haar wavelet, Daubechies wavelets, Symlets, Coiflets, Discrete approximation of Meyer wavelet etc., to choose the best wavelet function, the performance of these wavelet functions has been compared. And it is found that Daubechies of order 2 and Symlets of order 2 are jointly better than the other wavelet functions. It is observed that the performance of Wavelet transformation based on Daubechies of order 2 is the best for image compression and Fourier transformation is the second. In image denoising, Wavelet transformation based on Daubechies of order 2 is also the best performer for random denoising but for text denoising robust singular value decomposition (RSVD) is the best. Therefore, among these techniques, we recommend to use Wavelet transformation for image compression and random denoising and robust singular value decomposition (RSVD) for text denoising.

Keywords: Wavelet Transformation, Fourier Transformation, Classical and Robust Singular Value Decomposition, Image Compression and Denoising.

1 Introduction

Image compression is very important for storage of images in computer and transmission of images through Internet. Image compression plays a vital role in several important and diverse applications, including televideoconferencing, remote sensing, medical imaging and magnetic resonance imaging and many more [1]. With the use of digital cameras, requirements for storage, manipulation, and transfer of digital images, has grown explosively [2]. These images contain large amounts of information that requires much storage space, large transmission bandwidths and long transmission times. Therefore it is advantageous to compress the image by storing only the essential information needed to reconstruct the image.

For image and video compression, different types of preprocessing are very importent. When we introduce the inputs from the analog sources, different noises can appear in the raw images. To enhance the image quality, denoising is very important. Many denoising methods like [3-12], sparse representation [13] and K-SVD [14] methods, curvelet [15] and ridgelet [16] based methods, shape-adaptive transform [17], bilateral filtering [18,19], non-local mean based methods [20,21], non-local collaborative filtering [22] and Two stage principal component analysis[23] have been proposed. The performance of these methods should be compared. Though Zhang (2009) has proved Wavelet [24] Transformation is more effective than [1-10] but the comparison of four most popular techniques like Fourier transformation, Wavelet transformation, SVD and RSVD has not yet been found in the literature. So in this paper we have introduced wavelet transformation, Fourier transformation SVD and RSVD [25] for performance analysis of image denoising.

There are many techniques for image compression such as Wavelet transformation [26, 27, 28], Fourier transformation [29, 30, 31], Singular Value Decomposition [32, 33, 34], Neural networks [35] etc. Wavelet transformation is a modern technique used for signal processing, denoising, data compression, image analysis etc. A wavelet transformation is a lossless linear transformation of a signal into coefficients on a basis of wavelet functions. Wavelet transformation represents data in terms of wavelet basis which is orhtonormal and both spatially and frequency localized. A vector z is localized in space near n_0 if most of the components z(n) of z are 0 or at least relatively small, except for a few values of n near n_0. .And a vector is frequency localized if its discrete Fourier

transformation is localized in space. Wavelet basis, being both spatially and frequency localized, wavelet transformation represents local as well as global information very well. Wavelet transformation is used in many fields such as signal and image processing, time series analysis, microarray data analysis [36, 37] etc.

Fourier transformation is also a technique used for signal processing, denoising, data compression, image analysis etc. Fourier transformation is also a linear transformation of a signal into coefficients on a basis of Fourier functions. Fourier basis is orthogonal and frequency localized but not spatially localized. It transforms a signal from time domain to frequency domain. It gives frequency analysis of a signal. Fourier transformation is an old technique and used in many fields which is known to all.

Singular Value Decomposition was originally developed for solving system of linear equations. But it is used for many purposes such as data reduction both variables and observations, solving linear least square problems, image processing and compression, K-selection for K-means clustering, multivariate outlier detection, microarray data analysis etc. It decomposes a rectangular matrix X into three matrices: a matrix U whose columns are orthonormal, a diagonal matrix Λ and a matrix V whose rows are orthonormal such that $X=U\Lambda V^T$.

The performance of the old image compression technique Fourier transformation is not quite satisfactory due to large MSE between original and reconstructed image. But the performance of modern image compression technique Wavelet transformation is good enough and this technique is being widely used by Mathematicians, Physicists and Computer scientists in recent years. Again Singular value Decomposition is also a modern technique for image compression and is being used in recent years. This article attempts to compare the performance of wavelet transformation, Fourier transformation and classical and robust singular value decomposition in image compression as well as denoising.

2 Image Compression

To store a digital image in computer needs some memory space. Image compression reduces the file size of image without degrading the quality of the image to an unacceptable level. Image compression allows to store more images in a given memory space. It also reduces the amount of time required for sending images through internet or for downloading images from web pages. There are two types of image compression: 1. Lossless compression and 2. Lossy compression. If the original image can be reconstructed completely without losing any information, then it is called lossless compression. Again if the image can be reconstructed by permanently eliminating certain information, especially redundant information, then it is called lossy compression. In this paper lossy compression is used.

3 Methodology

In this section we describe how wavelet transformation, Fourier transformation and classical and robust singular value decomposition works in image compression and denoising.

3.1 Wavelet Transformation

First we convert the study image in data matrix say ***A***. Then we represent this matrix with respect to wavelet basis. That is we get some wavelet coefficients and among these coefficients we choose the largest k coefficients and applying inverse wavelet transformation we approximate the image matrix. Let $B= V_1, V_2, \dots, V_{mn}$ be the wavelet basis. Thus $A=\sum_{i=0}^{mn-1} a_i V_i$ for some scalars $a_0, a_1, \dots, a_{mn-1}$. Let ***S*** be the set of K largest scalars. Then Approximate ***A*** by $\tilde{A}=\sum_{i\in S} a_i V_i$.

3.2 Fourier Transformation

For the Fourier basis ***F*** the procedure is same as the wavelet transformation.

3.3 Singular Value Decomposition

First we decompose the image matrix ***A*** into three matrices: a column orthonormal matrix, a diagonal matrix and a row orthonormal matrix. The diagonal elements of the diagonal matrix are called singular values. Then we choose the largest ***l*** diagonal elements and select *l* columns of the row and column orthonormal matrix and multiplying these three matrices we approximate the image matrix. Decompose matrix ***A*** as $A=U_{m\times k}\Lambda_{k\times k}V_{k\times n}^{T}$ where *k*=rank of ***A***. Select the largest *l* diagonal elements of Λ. Redefine *U* and *V* by taking only *l* columns of them. Then approximate the matrix ***A*** by $\tilde{A}=U_{m\times l}\Lambda_{l\times l}V_{l\times n}^{T}$.

3.4 Robust Singular Value Decomposition

For image compression and denoising, in this article we used M-estimation based RSVD proposed by Fernando De la Torre and Michael J. Black [25]. It can do both compression and denoising by taking smaller number of singular values that accounts a large amount of variation of data. It can also handle missing values. The procedure of robust singular value decomposition is given in detail. Let $\boldsymbol{X}$ be a data matrix $\boldsymbol{X}=(x_1, x_2, \ldots, x_{n)=} (x^1, x^2, \ldots x^m)^T$ and $\boldsymbol{X} \in R^{m \times n}$. We assume that the data is zero mean, if not then we can convert the data matrix with zero mean by subtracting the mean. Robust mean must be explicitly estimated along with the bases for robust cases.

In case of SVD the data matrix X can be decomposed as $X=U\Lambda V^T$ where $U^TU=I$, and the columns of U that are associated with nonzero λ_i's can make a basis of X. If we approximate the column space of X with $k<<m$ rank, then the data x_i can be approximated by $x_i^{rec}=UU^Tx_i$. Let $c_i=U^Tx_i$are the linear coefficients obtained by projecting the column data on to the principal subspace that is $C=(c_1, c_2, \ldots, c_n)=U^TX$. We can approximate the least square estimate of SVD by minimizing

$$\min \sum_{i=1}^{n} \left\| x_i - UU^T x_i \right\|_2^2$$ subject to the constraint $U^TU=I$.

For noisy image it is observe that the noise of each pixel is assumed to be Gaussian ($e_{pi} \sim N(0; \sigma^2)$). But noise can vary for every row of data ($e_{pi} \sim N(0; \sigma_p^2)$). So Torre and Black [25] minimized the following function

$$\min \sum_{i=1}^{n} \sum_{p=1}^{m} \rho(x_{pi} - \mu_p - \sum_{j=1}^{k} u_{pj} c_{ji} \sigma_p)$$

for a particular class of robust ρ function (Geman-McClure error function), using M-estimation process.

4 Analysis

This section consists of four subsections: i) Comparison of different wavelet functions ii) Image compression using wavelet transformation, Fourier transformation and singular value decomposition iii) Image denoising using wavelet transformation, Fourier transformation and classical and robust singular value decomposition : Random Denoising. iv) Image denoising using wavelet transformation, Fourier transformation and classical and robust singular value decomposition: Text Denoising. We measure the performance of the techniques by relative error defined as

$$\text{Relative error} = \frac{\left\| A - \tilde{A} \right\|}{\left\| A \right\|} \quad (1)$$

where $\tilde{A}$ is the approximation of the original image matrix $\boldsymbol{A}$ and $\|\cdot\|$ stands for norm. Here Frobenius-norm is used.

4.1 Comparison of Different Wavelet Functions

We use four images for comparison of nine wavelet functions namely Haar wavelet, Daubechies of order 2, 4 and 6, Symlet of order 2 and 4, Coiflet of order 2 and 4 and discrete approximation of Meyer wavelet.The comparison is based on the relative error defined in equation (1). The images used for different wavelets comparison is given in Figure 1.

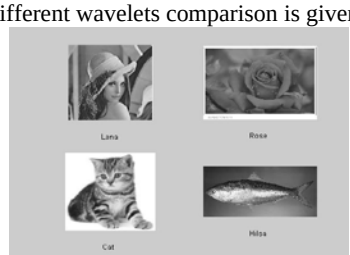


Fig. 1.Images for comparison of different wavelet functions.

Figure 2, Figure. 3, Figure 4 and Figure 5 shows the relative error of different wavelet functions for Lena, Rose, Cat and Hilsa images respectively

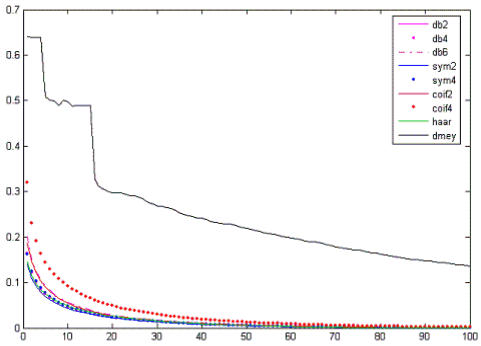


Fig. 2. Comparison of different wavelet functions for Lena image

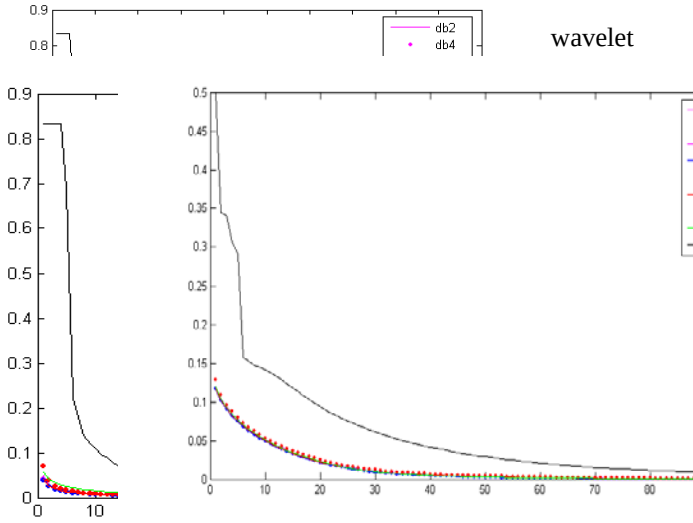


Fig. 4.Comparison of different wavelet functions for Cat image

Fig. 5.Comparison of different wavelet functions for Hilsa image

From figure 2 it is observed that Daubechies 2 and Symlets 2 are jointly better than the other wavelets, and then Haar wavelet is better. Discrete approximation of Meyer wavelet is the worst performer. From figure 3 it is observed that again Daubechies 2 and Symlets 2 are jointly better than the other wavelets and discrete approximation of Meyer wavelet is the worst performer. From figure 4 it is observed that Daubechies 4 and Symlets 4 are jointly better than the other wavelets. From figure 5 it is observed that except discrete approximation of Meyer wavelet, performance of all wavelets are approximately equal.

Among four, Daubechies 2 and Symlets 2 are jointly better than the other wavelets for three images. Moreover, for image compression and denoising we use Lena image and for this image Daubechies 2 and Symlets 2 are jointly better than the other wavelets. Therefore in this paper we use Daubechies 2 wavelets.

4.2 Image compression using Wavelet Transformation, Fourier Transformation and Singular Value Decomposition

We use Lena image for our analysis that is very renowned image for image processing. The dimension of the image matrix is 256× 256. We want to compress the image by wavelet transform, Fourier transform and SVD and compare their performance. Figure 6 shows the original image and the compressed image by wavelet transformation, Fourier transformation and SVD with 5% observation.

Fig. 6. Compression performance among wavelet, Fourier and SVD.

We know that for grey image 1 pixel= 1 byte. The following table shows the memory of original image and compressed image with 5% observation.

Table 1. Memory of Original and Compressed image

Image Status	Memory
Original Image	65536 bytes
Compressed Image	3277 bytes
Memory reduction	**62259 bytes**

From Figure 6 it is observed that wavelet is better than Fourier and SVD. Between Fourier and SVD, Fourier is better.

Now we increase the percentage of observation and compare the performances of the three techniques. Figure 7 shows the original image and the compressed image by using wavelet transformation, Fourier transformation and SVD with 15% observation.

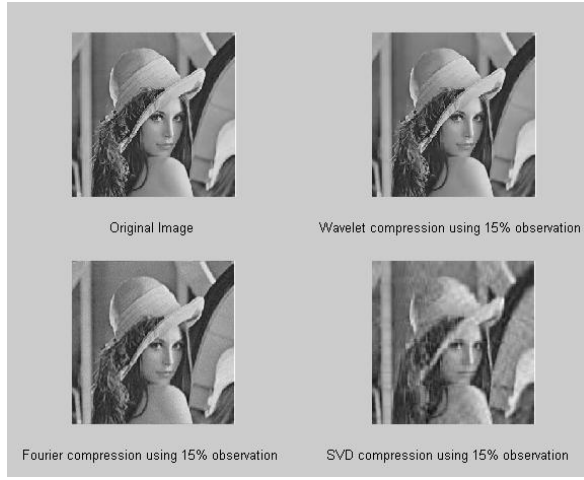


Fig. 7. Compression performance among wavelet, Fourier and SVD.

The following table shows the memory of original image and compressed image with 15% observation.

Table 2. Memory of Original and Compressed image

Image Status	Memory
Original Image	65536 bytes
Compressed Image	9830 bytes
Memory reduction	**55706 bytes**

From figure 7 it is observed that although only 15% observation is used, there is no significant difference between the original image and the image of wavelet compression with 15% observation. But the image of Fourier compression and SVD compression with 15% observation differ from the original image. Thus we can say that wavelet is better than Fourier and SVD.

Now we compare the performance of the three techniques by using relative error. Figure 8 shows the plot of relative error against the percentage of observation.

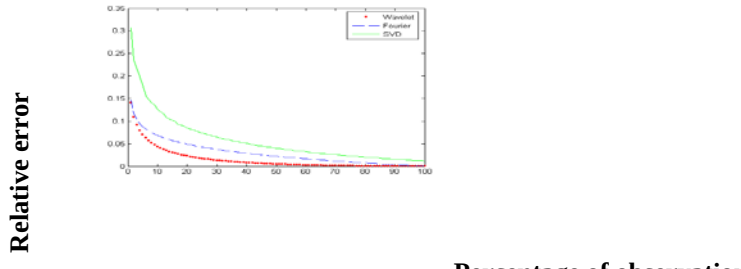


Fig. 8. Relative error of wavelet, Fourier and SVD compression.

From figure 8 it is observed that wavelet is better than Fourier and SVD and between Fourier and SVD, Fourier is better.

4.3 Image Denoising using Wavelet Transformation, Fourier Transformation and Classical and Robust Singular Value Decomposition: Random Denoising

In this section we use wavelet transformation, Fourier transformation, SVD and RSVD to denoise image and compare their performance by visualization. We create a random matrix of the same order as the original matrix by normal random number and add to the original image matrix. In this way we create the noisy image. Figure 9 shows the original image and noisy image of Lena.

Fig. 9. Original and Noisy image of Lena.

Now we denoise the noisy image using wavelet, Fourier, classical and robust SVD and compare their performance. Figure 10 shows the denoising performance among Wavelet, Fourier, SVD and RSVD.

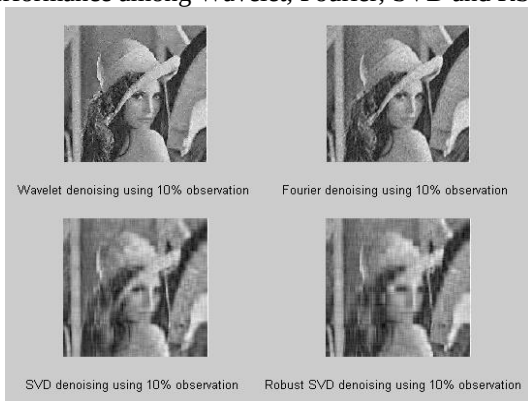


Fig. 10. Denoising performance among wavelet, Fourier, SVD and RSVD.

The following table shows the relative error wavelet, Fourier, SVD and RSVD.

Table 3.Relative error of wavelet, Fourier, SVD and RSVD.

Techniques	Relative error
Wavelet	0.0153
Fourier	0.0185
SVD	0.0372
RSVD	0.0423

From figure 10 it is observed that the denoising performance of wavelet is better than Fourier, SVD and RSVD. And Fourier is better than SVD and RSVD. Again from the relative error the same scenario is observed. Here only 10% observation is used for denoising and wavelet has deleted almost all the noises and the image quality is also satisfactory.

4.4 Image Denoising using Wavelet Transformation, Fourier Transformation and Classical and Robust Singular Value Decomposition: Text Denoising

In this section we use wavelet transformation, Fourier transformation, SVD and RSVD to denoise text from the image and compare their performance by visualization. Suppose there is some text inside in an image and we have to denoise it. Figure 11 shows the original image and noisy image of Lena.

Fig. 11. Original and Noisy image of Lena.

Figure 12 shows the denoised image of the text noisy image using different methods.

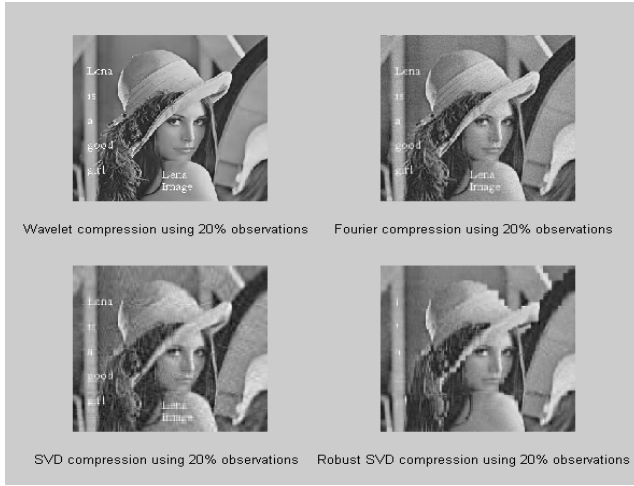


Fig. 12. Denoising performance among wavelet, Fourier, SVD and RSVD.

The following table shows the relative error wavelet, Fourier, SVD and RSVD.

Table 4.Relative error of wavelet, Fourier, SVD and RSVD.

Techniques	Relative error
Wavelet	0.0429
Fourier	0.0402
SVD	0.0421
RSVD	0.0264

The relative error of RSVD is the least and from figure 12 it is observed that RSVD deleted almost all the text noise. But Wavelet, Fourier and classical SVD could not delete the text noise from the image. Thus RSVD is better than wavelet, Fourier and SVD for text denoising.

5 Conclusion

Finally we can conclude that the performance of wavelet is better than Fourier, classical SVD and robust SVD in image compression as well as random denoising. But for text denoising RSVD is better than Wavelet, Fourier and classical SVD. Wavelet, Fourier and SVD takes only several seconds whereas RSVD takes several minutes for computation purposes. That is, the computation time of RSVD is much more than wavelet, Fourier and SVD but it is more powerful than wavelet transformation, Fourier transformation and classical SVD for text denoising. Therefore among these techniques, we recommend to use wavelet transformation for image compression and random denoising and to use robust SVD for text denoising.

References

1. Anuj Bhardwaj & Rashid Ali: Image Compression Using Modified Fast Haar Wavelet Transform, *World Applied Sciences Journal* 7 (5): 647-653, 2009

2. Bhawna Gautam: Image Compression Using Discrete Cosine Transform & Discrete Wavelet Transform, B.Sc. thesis, 2010

3. Donoho, D. L.: De-noising by soft thresholding, *IEEE Transactions on Information Theory* 41, 613-627, 1995

4. Coifman, R. R. & Donoho, D. L.: Translation-invariant denoising, in: Antoniadis, A.G., Oppenheim (Eds.), Wavelet and Statistics, Springer, Berlin, Germany, 1995

5. Mihcak, M. K., Kozintsev, I., Ramchandran, K. & Moulin, P.: Low-complexity image denoising based on statistical modeling of wavelet coefficients, IEEE Signal Processing Letters 6 (12) 300-303, 1999

6. Chang, S. G., Yu, B. & Vetterli, M.: Spatially adaptive wavelet thresholding with context modeling for image denoising, *IEEE Transaction on Image Processing* 9 (9) 1522-1531, 2000

7. Pizurica, A., Philips, W., Lamachieu, I. & Acheroy, M.: A joint inter and intra scale statistical model for Bayesian wavelet based image denoising, *IEEE Transaction on Image Processing* 11 (5) 545-557, 2002

8. Zhang, L., Paul, B., & Wu, X.: Hybrid inter and intra wavelet scale image restoration, *Pattern Recognition* 36 (8) 1737-1746, 2003

9. Hou, Z.: Adaptive singular value decomposition in wavelet domain for image denoising, Pattern Recognition 36 (8) 1747-1763, 2003

10. Portilla, J., Strela, V., Wainwright, M. J. & Simoncelli, E. P.: Image denoising using scale mixtures of Gaussians in the wavelet domain, *IEEE Transaction on Image Processing* 12 (11) 1338-1351, 2003

11. Zhang, L., Bao, P. & Wu, X.: Multiscale LMMSE-based image denoising with optimal wavelet selection, *IEEE Transaction on Circuits and Systems for Video Technology* 15 (4) 469-481, 2005

12. Pizurica, A. & Philips, W.: Estimating the probability of the presence of a signal of interest in multiresolution single and multiband image denoising, *IEEE Transaction on Image Processing* 15 (3) 654-665, 2006

13. Elad, M. & Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries, *IEEE Transaction on Image Processing* 15 (12) 3736-3745, 2006

14. Aharon, M., Elad, M. & Bruckstein, A. M.: The K-SVD: an algorithm for designing of over complete dictionaries for sparse representation, *IEEE Transaction on Signal Processing* 54 (11) 4311-4322, 2006

15. Starck, J. L., Candes, E. J., Donoho, D. L.: The curvelet transform for image denoising, *IEEE Transaction on Image Processing* 11(6) 670-684, 2002

16. Chen, G. Y., Ke gl, B., Image denoising with complex ridgelets, *Pattern Recognition* 40 (2) 578-585, 2007

17. Foi, A., Katkovnik, V. & Egiazarian, K.: Point wise shape adaptive DCT for high quality denoising and deblocking of gray

scale and color images, *IEEE Transaction on Image Processing* 16 (5), 2007

18. Tomasi, C. & Manduchi, R.: Bilateral filtering for gray and colour images, in: Proceedings of the 1998 IEEE International Conference on Computer Vision, Bombay, India, pp.839-846, 1998.

19. Barash, D.: A fundamental relationship between bilateral filtering, Adaptive smoothing, and the nonlinear diffusion equation, *IEEE Transaction on Pattern Analysis and Machine Intelligence* 24 (6) 844-847, 2002

20. Buades, A., Coll, B., Morel, J. M.: A review of image denoising algorithms, with a new one, Multiscale Modeling Simulation 4 (2) 490-530, 2005

21. Kervrann, C., Boulanger, J.: Optimal spatial adaptation for patch based image denoising, *IEEE Transaction on Image Processing* 15(10) 2866-2878, 2006

22. Dabov, K., Foi, A., Katkovnik, V. & Egiazarian, K.: Image denoising by sparse 3D transform-domain collaborative filtering, *IEEE Transaction on Image Processing* 16 (8) 2080-2095, 2007

23. Zhang, L., Dong, W., Zhang, D. & Shi, G.: Two stage image denoising by principal component analysis with local pixel grouping, *Pattern Recognition* 43 1531-1549, 2010

24. Mallat, S.: A Wavelet Tour of Signal Processing, *Academic Press, NewYork*, 1998

25. De la Torre, F. & Black, M. J.: Robust principal component analysis for computer vision. Int. Conf. on Computer Vision (ICCV'2001), Vancouver, Canada, July 2001

26. Singh, T., Chopra, S., Kaur, H. & Kaur, A.: Image Compression Using Wavelet and Wavelet Packet Transformation, International Journal of Computer Science and technology, Vol.1, Issue 1, 2010

27. Raviraj, P. & Sanavullah, M. Y.: The Modified 2D-Haar Wavelet Transformation in Image Compression, *Middle-East Journal of Scientific Research* 2 (2):73-78, 2007

28. Lawson, S. & Zhu, J.: Image compression using wavelets and JPEG2000: a tutorial, *Electronics Communication Engineering Journal*, 2002

29. Kaur, P.: Image Compression Using Fractional Fourier Transform, an M.Sc. thesis.

31. Almeida, L. B.: The fractional Fourier transform and time frequency representations, *IEEE Trans. Signal*

Process., vol. 42, pp. 3084-3091, 1994

32. Arnold, B.: An Investigation into using Singular Value Decomposition as a method of Image Compression, a thesis, 2000

33. Abrahamsen, A. & Richards, D.: Image Compression using Singular Value Decomposition, 2001

34. Kalman & Dan: A Singularly Valuable Decomposition, The College Mahtematics Journal, Vol.27 No.1, 2-23, 1998

35. Jiang, J.: Image compression with neural networks, *Signal Processing: Image Communication* 14, 737-760, 1999

36. Prabakaran, S., Sahu, R. & Verma, S.: An analysis of Microarray data using Wavelet Power Spectrum, *Engineering Letters*, 13:3, EL_13_3_17 (Advance online publication), 2006

37. Prabakaran, S., Sahu, R. & Verma, S.: A Clustering and Selection Method using Wavelet Power Spectrum, IAENG International Journal of Computer Science, 32:4, IJCS_32_4_6, 2006

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

A DWT-Based Statistical Image Fusion Technique for Noisy Source Images

Fatema Tuz Jhohura[1], Tamanna Howlader[21,2]Institute of Statistical Research and Training, University of Dhaka, Dhaka-1000, Bangladesh

Abstract. Image fusion deals with the integration of images from various sensors to obtain an image that has higher information content than the individual source images. Traditional image fusion algorithms assume that the input images are noise-free. However, in practice, this is rarely the case. The presence of noise in the source images introduces unwanted artefacts and distortion in the fused image. Thus fusion algorithms need be designed that operate robustly under noisy input conditions. The discrete wavelet transform is highly successful in the development of efficient image fusion and denoising algorithms. The purpose of this research work is to design a new DWT-based image fusion technique for noisy source images that takes into account the statistical dependency between the DWT coefficients of noisy source images as well as the noise-free fused image using a locally adaptive joint PDF. This is used as a prior function in the Bayesian MAP estimation technique to derive an estimator for the noise-free DWT coeffcient of the fused image. Experiments are carried out on a large number of test images to evaluate the performance of the proposed method as compared to commonly used fusion methods. Results show that the proposed method outperforms other methods in terms of standard performance metrics such as the structural similarity, peak signal-to-noise ratio, and cross-entropy.

Keywords: Discrete wavelet transforms, Image fusion, Maximum a posteriori estimation, Statistical model, Denoising, Multivariate Gaussian probability density function

1 Introduction

Image fusion can be defined as the process by which several images, or some of their features, are combined together to form a single image. It is often desirable to fuse images from different sources, acquired at different times, or otherwise having different characteristics because fused image from different sensors create new images that are more suitable for the purposes of human/machine perception, and for further image-processing tasks such as segmentation, object detection or target recognition [1]. The successful fusion of images acquired from different modalities or instruments is of great importance in many applications such as medical imaging, microscopic imaging, remote sensing, computer vision, concealed weapon detection, battle field monitoring and robotics [1]. Most of the existing methods for fusion assume that the source images are noise free. But images are often corrupted by noise. In image processing, noise is considered as undesirable information that contaminates the image data. It produces undesirable random variation in image brightness and may cause artefacts that distort the information contained in the image. Hence, development of an efficient fusion algorithm for noisy images is essential [2],[3].

Image fusion can be performed both in pixel domain and as well as in the transform domain. The pixel based fusion methods often produce poor results in comparison to the transform based methods because the salient features of an image are more clearly depicted in transformed domain and due to the nature of the transform, image processing tasks can be performed more efficiently. Transforms that have been used in image fusion include the pyramid transform [4], discrete wavelet transform (DWT) [5-8, 10 11], complex wavelet transform (CWT) [12-13], curvelet transform [14], morphological wavelet [15], maximal gradient wavelet [16] etc. In general, the DWT-based fusion methods perform better than any of the pyramid transform-based methods [17]. The success of the DWT based fusion method can be attributed to the efficient representation of significant features of images such as edge and texture, and the ease with which detailed information can be extracted from the source image to produce the fused image. Although the CWT, contourlet, and curvelet transforms possess shift-invariance property and improved directional selectivity as compared to the DWT and, therefore, are useful for the development of an efficient image fusion algorithm, the increased computational complexity of these transforms cannot be ignored. Hence, the DWT-based fusion techniques are still preferable when massive volumes of image data need to be merged quickly. Apart from its success in developing image fusion algorithms, the DWT has been widely used in performing other key image processing tasks such as image denoising and compression [18]. Thus, the routines for DWT-based image fusion can be seamlessly embedded into the routines of other image processing operations resulting in a fast image processing algorithm. Furthermore, any fusion rule designed in the DWT domain can be easily extended for application in other wavelet-like transform domains.

Fusion of noisy source images produces a noisy fused image from which accurate information cannot be obtained. Traditional methods for noise-free images could be used by first denoising the source images. However

[1]Corresponding author: Fatema Tuz Jhohura, E-mail: fjhohura@isrt.ac.bd.

er this does not ensure the best fusion results even with sophisticated denoising algorithms and requires more computer time. Moreover an increase in the number of sensors used in a particular application leads to the proportional increase in the amount of image data [19] for which a fast image fusion and denoising algorithm is required for real time applications. In this paper, we propose a new DWT-based fusion rule for two source images that requires a very simple or crude denoising algorithm to obtain a fused image of very good quality. The method can be extended for K source images. It involves less computational complexity and saves the time required for implementation. The method is based on a locally adaptive joint statistical model for the DWT coefficients of the noisy source images and the noise-free fused image. This PDF is then used to derive a Bayesian Maximum a posteriori (MAP) estimator for the noise-free DWT coefficient of the fused image. Extensive simulation experiments are performed to compute the performance of the proposed algorithm with commonly used methods for image fusion with respect to three standard performance matrices, namely, cross entropy (CEN) [21,22], structural similarity (MSSIM) [20] and peak signal to noise ratio (PSNR) considering various noise levels. It is expected that the method will be useful in real time applications. This paper is organized as follows. Section II includes a brief review of the 2D-DWT. In Section III, the proposed DWT-based fusion method for noisy images is given. Simulation results are presented in Section IV. Finally, Section V provides the conclusion.

2 The 2D-DWT: a brief review

The DWT is a particular form of multiresolution analysis that is used extensively in image processing. Let $g(i,j), i = 1,2,\cdots,N_1, j = 1,2,\cdots,N_2$represent a pixel of an image of size $N_1 \times N_2$, where *(i, j)*is the two-dimensional index. The DWT of the image is given by [23]

$$
\begin{aligned}
g(i,j) = \frac{1}{\sqrt{N_1 N_2}} \sum_{K_1=1}^{N_1} \sum_{K_2=1}^{N_2} x_J^A(k_1;\, k_2) \left\{ 2^{J/2} \phi(2^J i - k_1) \phi(2^J j - k_2) \right\} \\
+ \frac{1}{\sqrt{N_1 N_2}} \sum_{l=1}^{J} \Bigg[\sum_{K_1=1}^{N_1} \sum_{K_2=1}^{N_2} x_l^H(k_1;\, k_2) \{ 2^l \psi(2^l i - k_1) \phi(2^l j - k_2) \} \\
+ \sum_{K_1=1}^{N_1} \sum_{K_2=1}^{N_2} x_l^V(k_1;\, k_2) \{ 2^l \phi(2^l i - k_1) \psi(2^l j - k_2) \} \\
+ \sum_{K_1=1}^{N_1} \sum_{K_2=1}^{N_2} x_l^D(k_1;\, k_2) \{ 2^l \psi(2^l i - k_1) \psi(2^l j - k_2) \} \Bigg]
\end{aligned} \tag{1}
$$

wherex_J^Adenotes the approximate coefficients in the largestlevel J, $x_l^O (O \in H,V,D)$denote the detail coefficients inthe level $l(l \in 1,2,\ldots,J)$ of orientation O, and ϕ and ψ,respectively, are the scaling and wavelet functions. The coefficientsare arranged into groups or subbands of different levelsand orientations. The subbandsHL_l, LH_l, and $HH_l(l \in 1,2,\ldots,J)$, contain the detail coefficients of the horizontal, vertical, and diagonal orientations, viz.,x_l^H, x_l^V and x_l^D respectively. Here, H represents the high pass filter and L the low pass filter for realizing ϕ and ψ[23]. The subband$LL_J$is the lowest resolution residual that containsx_J^A. The functions ϕ and $\psi$are chosen in such a way that these subbands can be reassembled to reconstruct the original image without error.

3 Fusion of Noisy Images in DWT domain

Let $f_1(i,j)$and $f_2(i,j)$ be pixels of two noise free source images where, $i = 1,2,\cdots,N_1$and $j = 1,2,\cdot,N_2$. Then the noisy pixels may be represented as

$g_1(i,j) = f_1(i,j) + \epsilon(i,j)$

$g_2(i,j) = f_2(i,j) + \epsilon(i,j)$

where$\epsilon(i,j)$ are noise-samples at the reference location. We assumed that the source images were corrupted with white gaussian noise with zero mean and varianceσ_ϵ^2. The standard deviation $\sigma_\epsilon$indicates the strength of noise. Let $x_1(i,j)$ and $x_2(i,j)$denote the DWT coefficients of the noise-free source images respectively, at spatial location $(k_1,k_2)$of a given subband. Since the DWT is a linear transform, the noisy coefficients of the images at that spatial location can be written as

$$
\begin{aligned}
y_1(k_1,k_2) &= x_1(k_1,k_2) + n(k_1,k_2) \\
y_2(k_1,k_2) &= x_2(k_1,k_2) + n(k_1,k_2)
\end{aligned}
$$

where$n(k_1,k_2)$ represents the noise coefficients of the source images with variance σ_n^2. If σ_n is unknown, it may be estimated by applying the median absolute-deviation method [24] in the highest frequency subband of the noisy DWT coefficients. Since the DWT coefficients of images in a subband are spatially non-stationary [25], the random variables of the coefficients are index-dependent. Let $x_1(k_1,k_2)$and $x_2(k_1,k_2)$be the samples of the random variables $X_1(k_1,k_2)$and $X_2(k_1,k_2)$, respectively. Similarly, we define the random variables $Y_1(k_1,k_2)$and

$Y_2(k_1,k_2)$. On the other hand, the wavelet coefficients of noise are spatially stationary and, therefore, the corresponding random variable$N$ are index independent having i.i.d. N(0, σ_n^2) distribution. Let $X_f(k_1,k_2)$denote the random variable for the noise-free DWT coefficient of the fused image at the same spatial location. We would like to find an estimate of the detailed DWT coefficient of fused image $\hat{x}_f(k_1,k_2)$by utilizing a priori information regarding the random variables$Y_1(k_1,k_2)$, $Y_2(k_1,k_2)$, and $X_f(k_1,k_2)$via a Bayesian statistical estimation approach. Specifically, the maximum a posteriori (MAP) estimator is derived for $X_f(k_1,k_2)$, which is the mode of the posterior density function$p_{X_f|Y_1,Y_2}(x_f|y_1,y_2)$. For notational convenience, the indices $(k_1,k_2)$are suppressed in the remainder of the paper, unless stated otherwise. Thus, the fused wavelet coefficients of the detailed subbands of a given decomposition level can be obtained using the MAP estimate as

$$\hat{x}_f = \arg\max_{x_f} p_{X_f|Y_1,Y_2}(x_f|y_1,y_2)$$

$$= \arg\max_{x_f} \frac{p_{Y_1,Y_2,X_f}(y_1,y_2,x_f)}{p_{Y_1,Y_2,}(y_1,y_2)}$$

$$= \arg\max_{x_f} p_{Y_1,Y_2,X_f}(y_1,y_2,x_f) \quad (2)$$

wherey_1, y_2 and $x_f$are the observed values of the index dependent random variables Y_1, Y_2 and X_f respectively, and $p_{Y_1,Y_2,X_f}(y_1,y_2,x_f)$ is their joint PDF.

There are two very important issues that have to be considered in choosing an appropriate joint probabilistic model for $Y_1, Y_2$and X_f. First, the model should provide an adequate fit to the data to be fused. Secondly, it should be mathematically tractable so that a fast fusion algorithm can be obtained for merging several source images in a real-time application. Since source images to be fused are captured from the same scene using different sensors or modalities of an imaging system, intuitively, these images should be correlated with each other in the pixel domain. In such a case, it is expected that the local neighboring DWT coefficients of a given subband of the source images will be correlated due to the fact that the DWT is a linear transform. In order to verify the existence of such a linear dependency, the test of significance for the correlation between the local neighboring DWT coefficients of two noisy source images is performed using the standard test based on the *t* distribution [26]. Table 1 shows the average percentage of Y_1 and Y_2 in each subband that are significantly correlated for three commonly-used test images, namely, *MRI*, *SAR*, and *Clock* [27]. The test is performed for the correlation between Y_1 and Y_2at a given spatial location by using the local neighboring DWT coefficients within a 3 × 3 window centered at that location. The level of significance used for this test is 5%. It is seen from Table 1 that at least 50% of the coefficients within the subbands of source images are significantly correlated with each other at a decomposition level l = 4 for the images *Clock*, *MRI* and *SAR* at noise level σ_n=10. In addition, the percentages of coefficients having significant correlations in the subbands of other decomposition levels are also non-negligible. The results obtained using other noise levels (i.e.σ_n=20, 30,…), other window sizes and for other test images are similar to those given in Table 1. Thus, the correlation that exists between the local neighboring DWT coefficients of a given subband of the noisy source images cannot be neglected.

Table 1.Average percentage of Y_1 and Y_2 that are significantly correlated in various subbands using the test based on the *t* distribution [26] at noise level σ_n=10

Image		*Clock*				*MRI*				*SAR*			
Level		l=1	l=2	l=3	l=4	l=1	l=2	l=3	l=4	l=1	l=2	l=3	l=4
Subband	HH_l	6	8	17	53	6	30	36	50	6	10	28	50
	HL_l	5	9	23	56	7	29	31	55	6	16	41	67
	LH_l	6	10	21	54	6	29	30	52	6	12	31	61

Since the image is a 2D random signal, the DWT coefficients of images are also random variables. The PDF of the DWT coefficients can be used to describe the stochastic nature of images and to extract important features for constructing more efficient denoising and fusion algorithm. The PDF for local neighboring DWT coefficients of natural images is very often chosen as zero-mean Normal [28]. Roy et al. [9] showed that the joint PDF of X_1, X_2 and X_f follows trivariate Gaussian. Then the joint prior function of Y_1, Y_2 and X_f can be derived from $p_{X_1,X_2,X_f}(x_1,x_2,x_f)$ using transformation of variables. Since the noise is independent of the signal, we can write $p(x_1,x_2,n,x_f) = p(x_1,x_2,x_f) \times p_N(n)$where N is distributed as N$(0,\sigma_n^2)$. Hence by using transformation technique we obtain the joint density of $p(y_1,y_2,n,x_f)$ and then integrating out noise term the joint density of $p(y_1,y_2,x_f)$ becomes

$$p(y_1, y_2, x_f) = \frac{1}{(2\pi)^{3/2}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\sqrt{\sigma_{ff}}\sqrt{\vartheta}} \exp\left[-\frac{1}{2\vartheta}\left\{\frac{y_1^2}{\sigma_{11}}(1-\rho_{2f}^2) + \frac{y_2^2}{\sigma_{22}}(1-\rho_{1f}^2) + \frac{x_f^2}{\sigma_{ff}}(1-\rho_{12}^2)\right.\right.$$
$$\left.\left. - 2(\rho_{12}-\rho_{1f}\rho_{2f})\frac{y_1}{\sqrt{\sigma_{11}}}\frac{y_2}{\sqrt{\sigma_{22}}} - 2(\rho_{1f}-\rho_{12}\rho_{2f})\frac{y_1}{\sqrt{\sigma_{11}}}\frac{x_f}{\sqrt{\sigma_{ff}}} - 2(\rho_{2f}-\rho_{12}\rho_{1f})\frac{y_2}{\sqrt{\sigma_{22}}}\frac{x_f}{\sqrt{\sigma_{ff}}}\right\}\right]$$
$$\times m_1 \exp\left[\frac{m_2^2}{2m_3}\right] \qquad (3)$$

where $m_1 = \sqrt{\frac{\vartheta}{\left\{\frac{(1-\rho_{2f}^2)}{\sigma_{11}} + \frac{(1-\rho_{1f}^2)}{\sigma_{22}} - \frac{2(\rho_{12}-\rho_{1f}\rho_{2f})}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}\right\}\sigma_n^2 + \vartheta}}$,

$m_2 = \frac{y_1\left(\frac{1-\rho_{2f}^2}{\sigma_{11}} - \frac{\rho_{12}-\rho_{1f}\rho_{2f}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}\right) + y_2\left(\frac{1-\rho_{1f}^2}{\sigma_{22}} - \frac{\rho_{12}-\rho_{1f}\rho_{2f}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}\right) - x_f\left(\frac{\rho_{1f}-\rho_{12}\rho_{2f}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{ff}}} - \frac{\rho_{2f}-\rho_{12}\rho_{1f}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{ff}}}\right)}{\vartheta}$,

$m_3 = \frac{\frac{(1-\rho_{2f}^2)}{\sigma_{11}} + \frac{(1-\rho_{1f}^2)}{\sigma_{22}} - \frac{2(\rho_{12}-\rho_{1f}\rho_{2f})}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}}{\vartheta} + \frac{1}{\sigma_n^2}$, $\vartheta = 1 + 2(\rho_{12}\rho_{1f}\rho_{2f} - \rho_{12}^2 - \rho_{1f}^2 - \rho_{2f}^2)$, and

$\{\sigma_{11}, \sigma_{22}, \sigma_{ff}\}and\{\rho_{12}, \rho_{1f}, \rho_{2f}\}$ are the variance and correlation parameters, respectively, that are estimated using the local neighboring DWT coefficients of the noise-free images. The parameter ρ_{uv}in (3) measures the strength of linear dependency betweenthe DWT coefficients of the image $u$and v,where $(u, v) \in (1,2,f)$. We may now use the PDF in (3) to obtain the MAP estimator given in (2). The estimator is obtained by solving the likelihood equation, $\frac{\partial}{\partial x_f}\left[\ln p_{Y_1,Y_2,X_f}(y_1, y_2, x_f)\right] = 0$, which takes the form

$$-\frac{1}{\vartheta}\left[\frac{(1-\rho_{12}^2)}{\sigma_{ff}}x_f - \frac{(\rho_{1f}-\rho_{12}\rho_{2f})}{\sqrt{\sigma_{11}}\sqrt{\sigma_{ff}}}y_1 - \frac{(\rho_{2f}-\rho_{12}\rho_{1f})}{\sqrt{\sigma_{22}}\sqrt{\sigma_{ff}}}y_2 + \frac{m_2}{m_3}\left(\frac{\rho_{1f}-\rho_{12}\rho_{2f}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{ff}}} + \frac{\rho_{2f}-\rho_{12}\rho_{1f}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{ff}}}\right)\right] = 0 \quad (4)$$

Solving for $x_f$in the above equation gives

$$\hat{x}_f = \frac{1}{\frac{(1-\rho_{12}^2)}{\sigma_{ff}} - ab^2}\left[\frac{(\rho_{1f}-\rho_{12}\rho_{2f})}{\sqrt{\sigma_{11}}\sqrt{\sigma_{ff}}}y_1 + \frac{(\rho_{2f}-\rho_{12}\rho_{1f})}{\sqrt{\sigma_{22}}\sqrt{\sigma_{ff}}}y_2\right.$$
$$\left. - ab\left\{y_1\left(\frac{1-\rho_{2f}^2}{\sigma_{11}} - \frac{\rho_{12}-\rho_{1f}\rho_{2f}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}\right) + y_2\left(\frac{1-\rho_{1f}^2}{\sigma_{22}} - \frac{\rho_{12}-\rho_{1f}\rho_{2f}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}\right)\right\}\right] \qquad (5)$$

where $a = \frac{\sigma_n^2}{\left\{\frac{(1-\rho_{2f}^2)}{\sigma_{11}} + \frac{(1-\rho_{1f}^2)}{\sigma_{22}} - \frac{2(\rho_{12}-\rho_{1f}\rho_{2f})}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}\right\}\sigma_n^2 + \vartheta}$ and $b = \frac{(\rho_{1f}-\rho_{12}\rho_{2f})}{\sqrt{\sigma_{11}}} + \frac{(\rho_{2f}-\rho_{12}\rho_{1f})}{\sqrt{\sigma_{22}}}$ are constants.

The fused wavelet coefficients of the approximate subband in the J^{th}level of decomposition are calculated as

$$\hat{x}_f^A = \frac{1}{2}[y_1^A + y_2^A] \qquad (6)$$

wherey_1^A and y_2^A are the approximate coefficients of two noisy source images.

3.1 Estimation of parameters

In order to obtain detailed coefficients of the fused image using (5), the parameters of the joint PDF in (3) are required to be estimated. These parameters depend on the spatial index $(k_1, k_2)$and are estimated from the coefficients of a local neighborhood $\mathcal{S}(k_1, k_2)$ which is chosen as a square-shaped window centered at $x_f(k_1, k_2)$ or $y_k(k_1, k_2)$ $(k = 1, 2)$. For the purpose of estimation, it is assumed that the coefficients within $\mathcal{S}(k_1, k_2)$ are i.i.d. in nature. In other words, $\sigma_{rr}(w_1, w_2)$=$\sigma_{rr}(k_1, k_2)$ $(r \in (1, 2, f))$and $\rho_{rs}(w_1, w_2) = \rho_{rs}(k_1, k_2)(s \in$

$(1,2,f), r \neq s)$for all indices $(w_1, w_2) \in \mathcal{S}(k_1,\ k_2)$.In such a case, the parameters may be estimated using the method of maximum likelihood as [29]

$$\hat{\sigma}_{rr}(k_1,k_2) = \max\left(\frac{1}{M}\sum_{\mathcal{S}(k_1,k_2)} y_r^2(w_1,w_2) - \sigma_n^2, 0\right) \quad (7)$$

$$\hat{\rho}_{rs}(k_1,k_2) = \max\left(\min\left(\frac{1}{\hat{\sigma}_{rr}(k_1,k_2)\hat{\sigma}_{ss}(k_1,k_2)}\frac{1}{M}\times\sum_{\mathcal{S}(k_1,k_2)} y_r\,(w_1,w_2)y_s(w_1,w_2) - \sigma_n^2, 1\right), -1\right) \quad (8)$$

where$M$is the total number of coefficients in $\mathcal{S}(k_1,\ k_2)$. However, in order to estimate $\hat{\sigma}_{ff}^2(k_1,\ k_2)$ and $\hat{\rho}_{rf}(k_1,\ k_2)$ using (7) and (8), respectively, an initial estimate of the detailed coefficients of the noise-free fused image denoted as $\hat{x}_f^0$isrequired. In such a case, one may use any simple DWT-based fusion algorithm to obtain the initial estimate of fused image. In the proposed algorithm,we have obtained the initial estimate of fused coefficients according to the formula

$$\hat{x}_f^0 = \frac{1}{2}\sum_{k=1}^{2}\hat{x}_k(k_1;\ k_2) \quad (9)$$

where$\hat{x}_k(k_1;\ k_2)\ (k = 1,2)$are the estimated noise-free DWT coefficients of the source images which can be obtained by using any simple DWT- based denoising algorithm. A block diagram of the proposed method for obtaining the fused image $\hat{I}_f$from the twosource images $g_k(k = 1,2)$ is given in Fig. 1.

4 Experimental results

Extensive experimentations are carried out on a variety of commonly-used test images in order to compare the performance of the proposed fusion method with that of the recent methods in the literature. In this section, representative results are given for three sets of two source images, viz., *Clock*, and *SAR* and *MRI*. All the test sets comprise grayscale images. The source images can be obtained from the website: http://www.imagefusion.org. The test set *Clock* consists of multifocus images of size 256 × 256. The second set *SAR* represents a set of multisensored and multispectral images of size 512 × 512. The third set, *MRI* consists of two images of size 256×256 captured using the same imaging system, but at two modes.

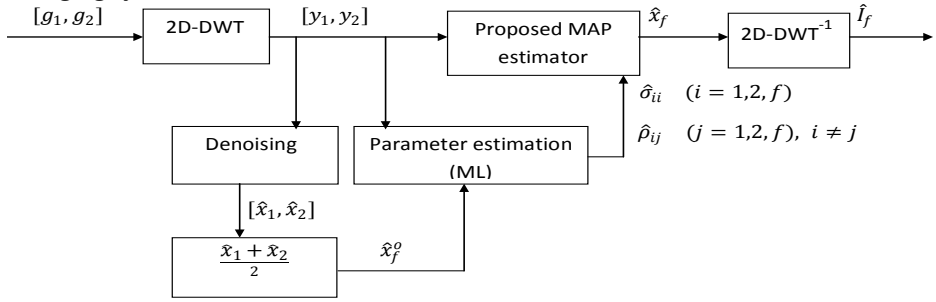


Fig. 1. Block diagram of the proposed fusion algorithm for the noisy detailed DWT coefficients

Here noisy images are generated synthetically by adding Gaussian noise to the noise free images considering three values of $\sigma_n$viz. 10, 20,and 30. The proposed fusion method is compared with four popular wavelet-based image fusion methods, namely, the wavelet maxima [33], variance feature [34], contrast measure [35], and adjustable parameter method [36]. The parameters of the fourth method are adjusted in such a way that the highest possible fusion performance of the method can be obtained. The wavelet coefficients in all the experiments are obtained by employing a 4-level DWT, wherein the orthogonal Symlet filter of length 8 is used. Results of the proposed method are given for a 3×3 window size for estimating the local variance and correlation parameters, since an increase of the window size does not provide any significant change in fusion performance except an increasing computational burden. For a fair comparison, the same window size is used for the variance feature [34] and contrast measure [35] methods to calculate the local variance. The performance of the fusion methods are compared with respect to three commonly used metrics, viz., SSIM [20], PSNR, and CEN [21,22]. For a given image set, these metrics are calculated from each pair of the fused and source image, and then the final value of the metric is obtained by averaging over the number of source images. A better fusion method should provide a higher value of SSIM and PSNR, and a lower value of CEN [37].

Table 2. Comparison of various DWT-based fusion methods for noisy source images with respect to standard metrics and implementation time at different noise levels[1]

Noise level	Images	Methods	Metrics			Time (sec.)
			MSSIM	CEN	PSNR	
σ_n=10	*Clock*	Wavelet maxima [33]	0.8791	8.0679	69.6124	1.4530
		Variance feature [34]	0.8783	8.0482	69.5426	1.4210
		Contrast measure[35]	0.8751	8.0069	70.9398	1.1930
		Adjustable parameter[36]	0.8668	7.9685	72.9184	1.1720
		Proposed method	**0.9241**	**7.7778**	**73.4304**	1.5000
	SAR	Wavelet maxima [33]	0.7854	7.0796	66.3452	1.5460
		Variance feature [34]	0.7824	7.0541	66.3949	1.4840
		Contrast measure[35]	0.7999	6.9836	67.4442	1.0780
		Adjustable parameter[36]	0.7718	6.9770	67.0277	1.2030
		Proposed method	**0.8337**	**6.6983**	**68.6716**	1.5470
	MRI	Wavelet maxima [33]	0.3879	6.9823	47.2304	0.3280
		Variance feature [34]	0.3711	7.0167	47.2211	0.2960
		Contrast measure[35]	0.4161	6.7765	48.4693	0.2650
		Adjustable parameter[36]	0.3374	6.7839	48.2418	0.3120
		Proposed method	**0.4405**	**6.1512**	**49.5394**	0.3130
σ_n=20	*Clock*	Wavelet maxima [33]	0.8310	8.2010	68.4236	1.4530
		Variance feature [34]	0.8352	8.2182	68.5149	1.4220
		Contrast measure[35]	0.7982	8.1892	69.3586	1.0160
		Adjustable parameter[36]	0.7907	8.1621	69.9170	1.1090
		Proposed method	**0.8977**	**7.9074**	**71.9082**	1.5000
	SAR	Wavelet maxima [33]	0.7357	7.1604	65.2886	1.4690
		Variance feature [34]	0.7348	7.1389	65.3448	1.4210
		Contrast measure[35]	0.7092	7.0997	65.9100	1.0160
		Adjustable parameter[36]	0.6875	7.0913	65.5114	1.1250
		Proposed method	**0.7850**	**6.8663**	**66.9599**	1.5150
	MRI	Wavelet maxima [33]	0.3669	7.0401	47.2719	0.3120
		Variance feature [34]	0.3542	7.0390	47.2565	0.2970
		Contrast measure[35]	0.3799	6.8767	48.3263	0.2350
		Adjustable parameter[36]	0.3144	6.8360	48.1361	0.2660
		Proposed method	**0.4234**	**6.3288**	**49.3878**	0.3120

Table 2 shows the values of these metrics that are obtained from the four DWT-based fusion algorithms considered in the experiments as well as the proposed method and the time required for implementation. It can be seen from the table that the proposed method consistently provides a higher value for SSIM and PSNR and lower value for CEN for all the test images. Hence the proposed method is superior to the other methods because it yields fused images with higher information content and less distortion by noise. Moreover, the time required for implementation is not significantly higher than the other methods. In contrast, the other methods require more sophisticated denoising methods to generate noise-free source images and to achieve the same initial estimates for noisy source images were obtained using SureShrink method level of fusion performance. This in turn would result in increased computational complexity and time for implementation. Thus, the method proposed in this paper holds great promise for realtime applications and for combining images of large size and poor quality. It may be mentioned that although the results in Table 2 were obtained by using the SureShrink method to estimate noise-free source images, other denoising methods, such as VisuShrink and NeighCoeff yielded similar results.

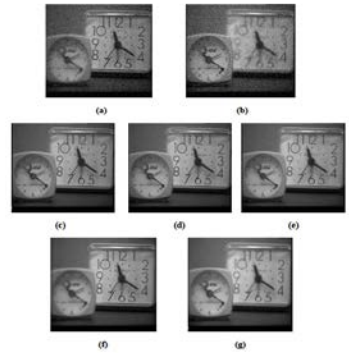


Fig. 2. Results of multifocus image fusion for noisy Clock images (σ_n=10) using various DWT-based fusion methods. The images are (a) Noisy source image (focused on right), (b) Noisy source image (focussed on right).

Fused images are obtained using (c) Wavelet maxima, (d) Variance features method, (e) Contrast measure, (f) Adjustable parameter method (g) proposed method.

Fusion performance may also be assessed through visual comparison of the fused image of the proposed method with that of the methods. Figure 2 shows the two source images of the test set *Clock* and their fused images obtained using the wavelet maxima [33] Variance features [34], Contrast measure [35], Adjustable parameter method [36], and proposed methods. It is apparent from this figure that the fused image obtained using the proposed method has all objects in focus and significantly less noise contamination compared to fused images

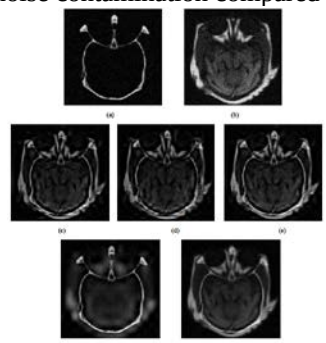


Fig. 3. Results of DWT-based multimodal image fusion for noisy *MRI* images (σ_n=20). (a) Noisy source image 1, (b) Noisy source image 2, (c) Noisy source image 3. Fused image obtained using (d) Variance features method, (e) Contrast measure, (f) proposed method

obtained using the other methods. On the other hand, Figure 3 shows the noisy source images of the test set *MRI* and the fused images obtained using the wavelet maxima [33] Variance featuresmethod [34], Contrast measure [35], Adjustable parameter method [36], and proposed methods. It can be seen from this figure that the proposed method combines the structural details of the source images and suppresses noise and artifacts more efficiently than the competing methods.

5 Conclusions

Image fusion techniques are widely used for combining several source images, captured using different sensors or at different modes of the imaging system, so that the resulting image contains more detail than any of the individual source images. But in practice, the images to be fused are noisy due to the non-ideal characteristics of imaging systems. Most of the image fusion methods in the literature assume that the source images are noise-free and produce poor quality images when used with noisy source images. The wavelet transform has shown significant success in the development of both the fusion and denoising algorithms for images. The motivation for the decimated DWT is that, it is non-redundant and therefore fast to implement. Most of the existing methods do not consider a probabilistic approach to developing a fusion algorithm. In addition, they do not give proper treatment to the correlation that exists between the noisy source images and in between the noisy source images and the fused image. In this paper we considered these aspects in designing an efficient fusion algorithm for two noisy source images in the DWT domain. More specifically, a joint PDF is derived for the DWT coefficients of the noisy source images as well as DWT coefficients of the noise-free fused image. This is used as a prior function in the Bayesian MAP estimation technique to derive an estimator for the DWT coefficient of the noise-free fused image. The unknown parameters in the formula for the estimator are determined locally using ML estimation. An important feature of the proposed estimator is that it has the effect of suppressing noise in addition to producing good fusion results and can easily be extended for K source images. Moreover existing methods require the noisy source images to be pre-processed using a sophisticated denoising algorithm while the proposed method can produce superior fusion results with a simple or crude denoising rule. Extensive experiments were conducted to compare the performance of the proposed method with commonly used image fusion techniques with respect to the image fusion performance metrics MSSIM, CEN and PSNR. The experiments which were performed on several types of source images at different noise levels showed that, the proposed technique provides better fusion results than the competing methods. As a final comment, it can be said that the proposed method is likely to be very useful for real-time implementation.

References

1. Amolins, K., Zhang, Y., and Dare, P.: 'Wavelet-based image fusion techniques- an introduction, review and comparison'. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(4): pages 249-263, 2007
2. V. S. Petrovi´c and C. S. Xydeas.: 'Sensor noise effects on signal-level image fusion performance' *Information Fusion*, vol. 4, no. 3, pages 167– 183, 2003

3. Z. Chen, Y. Zheng, B. R. Abidi, D. L. Page, and M. A. Abidi.: 'A combinational approach to the fusion, de-noising and enhancement of dual-energy X-ray luggage images' in *Proc. of IEEE Conference onComputer Vision and Pattern Recognition - Workshops*, vol. 3, San Diego, CA, pages 2–2, 2005
4. Toet, A.: 'Hierarchical image fusion'. *Machine Vision and Applications*, 3(1): pages 1-11, 1990
5. Koren, I., Laine, A., and Taylor, F.: 'Image fusion using steerable dyadic wavelet transform'. *In Image Processing.Proceedings., International Conference*, volume 3, pages 232-235. *IEEE*, 1995
6. Chipman, L., Orr, T., and Lewis, L.: 'Wavelets and image fusion'. *In Proceedings IEEE. International Conference on Image Processing.Washington D.C.*, volume 3, pages 248-251, 1995
7. Li, H., Manjunath, B., and Mitra, S.: 'Multisensor image fusion using the wavelet Transform'. *Graphical models and image processing*, 57(3): pages 235-245, 1995
8. Rockinger, O.: 'Pixel-level fusion of image sequences using wavelet frames'. *In Proc. 16th Leeds Annual Statistical Research Workshop*, pages 149-154. Citeseer, 1996
9. Roy, S., Howlader, T., and Rahman, S.M.M.: 'Image fusion technique using multivariate statistical model for wavelet coefficients'. *Signal, Image and Video Processing*, pages doi:10.1007/S11760-011-024-9, 2011
10. Zheng, H., Zheng, D., Hu, Y., and Li, S.: 'Study on the optimal parameters of image fusion based on wavelet transform'. *Journal of Computational Information systems*, volume 1: pages 131-137, 2010
11. Le Moigne, J. and Smith, J.: 'Image registration and fusion for nasa remotely sensed imagery'. In Information Fusion, 2000.*Proceedings of the Third International Conference on IEEE*, volume 1, pages TUB3-24, 2000
12. Qing, X., Shuai, X., Bing, T., Jiansheng, L., and Zexun, G.: 'Complex wavelets and its application to image fusion', 2004
13. Wan, T., Canagarajah, N., and Achim, A.: 'Segmentation-driven image fusion based on alpha-stable modeling of wavelet coefficients'. *Multimedia, IEEE Transactions*, 11(4): pages 624-633, 2009
14. Choi, M., Kim, R., Nam, M., and Kim, H.: 'Fusion of multispectral and panchromatic satellite images using the curvelet transform'. *Geoscience and remote sensing letters, IEEE*, 2(2): pages 136-140, 2005
15. Lin, P. and Huang, P.: 'Fusion methods based on dynamic-segmented morphological wavelet or cut and paste for multifocus images'. *Signal Processing*, 88(6): pages 1511-1527, 2008
16. Scheunders, P.: 'An orthogonal wavelet representation of multivalued images. Image Processing', *IEEE Transactions*, 12(6): pages 718-725, 2003
17. Zhang, Q.,Guo, B.: 'Multifocus image fusion using the nonsubsampled contourlet transform'. *Signal Process*. 89, pages 1334–1346, 2009
18. Rahman, S.M.M., Ahmad, M.O., Swamy, M.N.S.: 'Contrast-based fusion of noisy images using discrete wavelet transform'. *IET Image Process.*4(5), pages 374–384, 2010
19. Petrovic´, V.: 'Multisensor pixel-level image fusion'. doktorska teza, Manchester University, UK, 2001
20. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: 'Image quality assessment: from error visibility to structural similarity'. *IEEE Trans. Image Process*.3(4), pages 600–612, 2004
21. Zheng,Y.,Qin, Z., Shao, L., Hou, X.: 'A novel objective image quality metric for image fusion based on Renyi entropy'. *Inf. Technol. J.*7(6), pages 930–935, 2008
22. Zhang,Y.: 'Methods for image fusion quality assessment—a review, comparison and analysis'. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*. XXXVII(B7), pages 1101–1109, 2008
23. Mallat, S.: 'A Wavelet Tour of Signal Processing'. 2nd edn. Academic Press, San Diego, 1999
24. Arivazhagan, S., Ganesan, L., Kumar, T.G.S.: 'A modified statistical approach for image fusion using wavelet transform'. *Signal Image Video Process*. 3, pages 137–144, 2009
25. Cai, T. and Silverman, B.: 'Incorporating information on neighbouring coefficients into wavelet estimation' Sankhya: *The Indian Journal of Statistics*, Series B, pages 127-148, 2001
26.Urdan, T.: 'Statistics in plain English'. Lawrence Erlbaum, Mahwah, 2000
27. Image fusion web-site. [Online]. Available: http://www.imagefusion.org/
28. Mihcak, M., Kozintsev, I., Ramchandran, K., and Moulin, P.: 'Low-complexity image denoising based on statistical modeling of wavelet coefficients'. *Signal Processing Letters*, IEEE, 6(12): pages 300-303, 1999
29. Giri, N.C.: 'Introduction to Probability and Statistics', 2nd ed. M. Dekker, New York, 1993
30. Donoho, D. and Johnstone, J.: 'Ideal spatial adaptation by wavelet shrinkage'. *Biometrika*, 81(3): pages 425-455, 1994
31. MRI Database: National Center for Image Guided Therapy. [Online]. Available: http://www.ncigt.org/
32. De, I., Chanda, B.: 'A simple and efficient algorithm for multifocus image fusion using morphological wavelets'. *Signal Process*. 86, pages 924–936, 2006
33. Li, H., Manjunath, B.S., Mitra, S.K.: 'Multisensor image fusion using the wavelet transform'. *Graph.Models Image Process*.57(3), pages 235–245, 1995
34. Wu, J., Liu, J., Tian, J., Huang, H.: 'Multi-scale image data fusion based on local deviation of wavelet transform'. *In: Proceedings of the IEEE International Conference on Intelligent Mechatronics and Automation*, Chengdu, China, pages 677–680, 2004
35. Pu, T., Ni, G.Q.: 'Contrast-based image fusion using the discrete wavelet transform'. *Opt. Eng.*39(8), pages 2075–2082, 2000
36. Yunhao, C., Lei, D., Jing, L., Xiaobing, L., Peijun, S.: 'A new wavelet-based image fusion method for remotely sensed data'. *Int. J. Remote Sens*. 27(7), pages 1465–1476, 2006

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

A New Outlier Rejection Rule for Robust ICA and Its Application to Image Processing

Md. Shahjaman[1] and Md. Nurul Haque Mollah[2]

[1] Assistant Professor, Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh

Email<shahjaman_brur@yahoo.com>

[2] Professor, Department of Statistics, Rajshahi University, Bangladesh

Email<mnhmollah@yahoo.co.in>

Abstract. Independent Component Analysis (ICA) is a powerful statistical method for blind source separation (BSS) from the mixture data. It is widely used in signal processing like audio signal processing, image processing, biomedical signal processing as well as processing any time series data. However, most of the ICA algorithms are not robust against outliers. In this paper we propose a new outlier rejection rule for robustification of ICA algorithms using β-weight function. The values of the tuning parameter β play the key role in the performance of the proposed method. A cross validation technique is used as an adaptive selection procedure for the tuning parameter β. The performance of the proposed method is investigated in a comparison of the popular robust FastICA algorithms using natural image signals. Simulation and experimental results show that the proposed method improves the performance over the existing robust FastICA algorithms.

Keywords: Independent component analysis (ICA), Minimum β-Divergence Estimator, β-selection, β-Weight function, Outliers and Image signals.

1 Introduction

ICA aims to maximize the non-goussianity or minimize the dependency among the variables as it seeks to recover the sources that are as independent of each other as possible [5]. The independence is a much stronger property than uncorrelatedness [5]. Thus ICA becomes more superior to the principle component analysis (PCA). There are two popular model based robust ICA algorithms (a) FastICA algorithms [4] and (b) minimum β-divergence method [7]. In the case of minimum β-divergence method for ICA, Gaussianity of source signals should be known in advance, otherwise it may produce misleading results. This method suggests two contrast function, where one works to recover sub-Gaussian signals and the other one works to recover super-Gaussian signals. For example, image signals are sub-Gaussian signals and audio signals are super-Gaussian signals. Therefore, in the case of image processing or audio signal processing, minimum β-divergence method may works well. In other cases where source signals are unknown as sub-Gaussian or super-Gaussian, minimum β-divergence method may gives misleading results. On the other hand, robust FastICA algorithm suffers from the non-robust prewhitening procedure and some outliers those make perpendicular direction with recovering vectors, where non-robust prewhitening can be overcome by β-prewhitening [8], but the later one problem exist yet in the robust FastICAalgorithms. Therefore, a user or researcher may feel inconvenience to select an appropriate ICA algorithm in some situations. So, in this paper, our proposal is to use outlier rejection rule[10, 11] for robustification of ICA algorithms based on the β-weight function [9] instead of existing robust ICA algorithms.

2 Independent Component Analysis (ICA)

Independent component analysis (ICA) is a statistical tool for revealing hidden factors that underlie sets of random variables or signals. For given set of observations of random variables, $\boldsymbol{x}(t) = \left(x_1(t), x_2(t), \ldots\ldots, x_m(t)\right)^T$, where t is the time or sample index, assume that they are generated as a linear mixture of independent components then ICA model is given by:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ . \\ . \\ . \\ x_m(t) \end{pmatrix} = A \begin{pmatrix} s_1(t) \\ s_2(t) \\ . \\ . \\ . \\ s_q(t) \end{pmatrix}$$

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t), \quad t = 1,2, \ldots\ldots, n \qquad (2.1)$$

where, $A$is some unknown matrix. The aim of ICA is to estimate both the matrix $A$and the $\boldsymbol{s}(t)$, when we only observe the $\boldsymbol{x}(t)$. The ICA of a random vector $\boldsymbol{x}(t)$ consists of finding a linear transform

$$\boldsymbol{y}(t) = W\boldsymbol{x}(t), \quad t = 1,2, \ldots\ldots, n \qquad (2.2)$$

so that components of $\boldsymbol{y}(t)$ are as mutually independent as possible, where W transformation matrix obtained by ICA algorithm [1-6]. It is also known as recovering matrix or unmixing matrix or pseudo inverse of A or generalized inverse of A. The components of a random vector $\boldsymbol{y}(t)$are said to be independent of each other if and only if the density function of $y$is factorized as

$$p(\boldsymbol{y}) = \prod_{i=1}^{m} p_i(y_i)$$

where, $p_i(y_i) = \int p(\boldsymbol{y}),\ldots dy_1,\ldots dy_{i-1}\, dy_{i+1},\ldots\, dy_m$ is the marginal density of y_i, (i = 1, 2, ...,m). If components of y are independent of each other, then most important property of their independence is

$$E\left\{\prod_{i=1}^{m} h_i(y_i)\right\} = \prod_{i=1}^{m} E\{h_i(y_i)\}$$

where,$h_i(y_i)$ is any measurable function of y_i.

3 Outlier Rejection Rule for ICA Based on β-weight Function (New Proposal)

Mahalanobis distance (D^2) is a popular measure for detection of multivariate outliers. It works well in presence of few outliers. However, in presence of large number of outliers, it produces misleading results. To overcome this problem, in this paper we propose β-weight function as a new alternative measure for outlier detection.

This β-weight function is originated from the minimum β-divergence estimators for the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{V}$

$$\boldsymbol{\mu}_{t+1} = \frac{\sum_{i=1}^{n} \varphi_\beta(\boldsymbol{x}_i|\boldsymbol{\mu}_t, \boldsymbol{V}_t)\boldsymbol{x}_i}{\sum_{i=1}^{n} \varphi_\beta(\boldsymbol{x}_i|\boldsymbol{\mu}_t, \boldsymbol{V}_t)} \text{and}$$

$$\boldsymbol{V}_{t+1} = \frac{\sum_{i=1}^{n} \varphi_\beta(\boldsymbol{x}_i|\boldsymbol{\mu}_t, \boldsymbol{V}_t)(\boldsymbol{x}_i - \boldsymbol{\mu}_t)(\boldsymbol{x}_i - \boldsymbol{\mu}_t)^T}{(1+\beta)^{-1}\sum_{i=1}^{n} \varphi_\beta(\boldsymbol{x}_i|\boldsymbol{\mu}_t, \boldsymbol{V}_t)}$$

$$\text{where, } \varphi_\beta(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{v}) = \exp\left\{-\frac{\beta}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}$$

which is known a β-weight function [9]. It produces smaller weight for each contaminated data vector and larger weight for each uncontaminated data vector. Our intention is to use this weight function to separate the data into two parts bad (outliers/unusual) data points and good (usual) data points . To detect outliers, we compute the β-weight as follows:

$$\varphi_\beta(\boldsymbol{x}|\hat{\mu}_\beta, \hat{V}_\beta) = exp\left\{-\frac{\beta}{2}(\boldsymbol{x} - \hat{\mu}_\beta)^T \hat{V}_\beta{}^{-1}(\boldsymbol{x} - \hat{\mu}_\beta)\right\}$$

and then we construct a criteria to test the contaminacy of a data vector as follows:

$$\varphi_\beta(\boldsymbol{x}|\hat{\mu}_\beta, \hat{V}_\beta) = \begin{cases} larger\ (but \leq 1), & if\ \boldsymbol{x}\ is\ not\ contaminated \\ smaller\ (but \geq 0), & if\ \boldsymbol{x}\ is\ contaminated \end{cases} \qquad (2.4)$$

Using the proposed test criteria, we take the decision that a data vector $\boldsymbol{x}$ is said to be contaminated by outliers if

$$\varphi_\beta(\boldsymbol{x}|\hat{\mu}_\beta, \hat{V}_\beta) \leq \delta,$$

where we choose the threshold value of $\delta$by

$$\delta = (1-\eta)\ \min_{\boldsymbol{x} \in \boldsymbol{DS}} \varphi_\beta(\boldsymbol{x}|\hat{\mu}_\beta, \hat{V}_\beta) + \eta\ \max_{\boldsymbol{x} \in \boldsymbol{DS}} \varphi_\beta(\boldsymbol{x}|\hat{\mu}_\beta, \hat{V}_\beta)$$

with heuristically η = 0.10, where $\boldsymbol{DS}$ is the dataset. It was also used in [7,9] for choosing the threshold value. Then we reject or remove the contaminated data points from the dataset and we can apply any ICA algorithms to the clean dataset to recover source signals from the robustness points of view. In this paper we consider FastICA algorithm to demonstrate the performance of the proposed rejection rule in a comparison of the classical Mahalanobis distance approach.

4 Simulation Study

To demonstrate the performance of the proposed method in a comparison of some existing methods, we consider the following synthetic data sets.

1. Two-dimensional 1000 random samples were drawn from uniform distribution with mean zero and variance one. Figure 1a represents the scatter plot of this source data points. Then we mixed these data points by a random mixing matrix

$$A = \begin{bmatrix} 1.01 & 0.8 \\ 0.4 & 0.9 \end{bmatrix};$$

Figure 1b represents the scatter plot of this mixed data points.

2. 700 outliers (*) are added to data set 1to make 1700 samples in total. For convenience of presentation, we took last 700 observations as outliers out of 1700. Figure 2a represents the scatter plot of this data set.

Let us consider data set 1 described above. To remove outliers, first we select tuning parameter β using cross validation [8]. Figure 1c shows the plots of $\hat{D}_{\beta_0}(\beta)$. In this plot asterisks (*) are $\hat{D}_{\beta_0}(\beta)$ and a circle outside the asterisk indicates the smallest value. Dotted lines are $\hat{D}_{\beta_0}(\beta) \pm SD_{\beta_0}$. Plot of $\hat{D}_{\beta_0}(\beta)$ shown in the Figure 1c suggest β = 0 for β_0 = 0.5 by "One Standard Error" rule. Thus adaptive selection procedure suggests there is no outliers in the data set 1. Figure 1d shows the scatter plot of recovered sources by FastICA. Comparing Figures 1a and 1d, we see that recovered sources are independent with each other with non-Gaussian Structure.

Simulation with Uniformly Distributed Datasetin Absence of Outliers

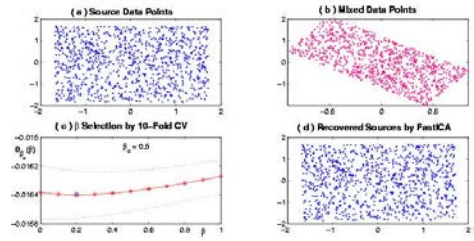


Figure 1. (a) Scatter plot of source signals generated from uniform distribution, (b) Scatter plot of mixed signals, (c) Plots of $\hat{D}_{\beta_0}(\beta)$ for selection of β, (d) Recovered Sources by FastICA.

To investigate the performance of the proposed method in presence of outliers (*), we consider data set 2 shown in figure 2a. To remove outliers by the proposed method, we select the values of the tuning parameter β by *K*-fold CV (k=10) as before. We computed $\hat{D}_{\beta_0}(\beta)$ for β varying from 0 to 1 by 0.5 with β_0 = 0.5. Figure 2b show the plots of $\hat{D}_{\beta_0}(\beta)$. In this plot the asterisk (*) are $\hat{D}_{\beta_0}(\beta)$ and the smallest value is indicated by a circle outside the asterisk. Dotted lines are $\hat{D}_{\beta_0}(\beta) \pm SD_{\beta_0}$. Plots of $\hat{D}_{\beta_0}(\beta)$ shown in the Figure 2b have elbow shape and suggested β = 0.8 for β_0 = 0.5 by "One Standard Error" rule. Thus adaptive selection procedure suggests that outliers corrupt the data set 2, which is true as shown in figure 2a. Figure 2c shows the recovered sources by FastICA using data set 2 before removing the outliers. Clearly we see that recovered sources are not similar to the original sources. Then we remove outliers from data set 2 using Mahalanobis distance. Figure 2d shows Mahalanobis D^2 for each data points. Figure 2e shows the mixed data after removing the outliers by Mahalanobis D^2 with $\chi^2_{1,0.95} = 3.84$. Clearly we see that all outliers are not removed based on Mahalanobis distance. Figure 2f shows the recovered sources by FastICA after removing outliers by Mahalanobis D^2. Clearly we see that recovered sources are not similar to the original sources as shown in 1a.

Simulation with Uniformly Distributed Datasetin Presence of Outliers

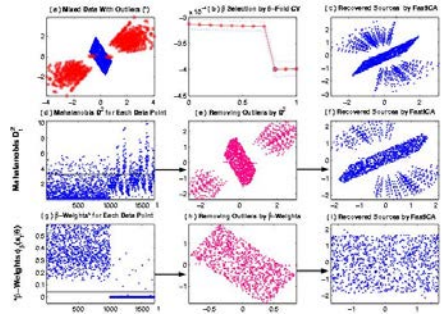


Figure 2. (a) Scatter plot of mixed signals with outliers (*), (b) Plots of $\hat{D}_{\beta_0}(\beta)$ for selection of β , (c) Recovered Sources by FastICA before removing outliers, (d) Mahalanobis distance for each data points with cut-off point $\chi^2_{1,0.95} = 3.84$, (e) Mixed data after removing outliers using Mahalanobis distance, (f) Recovered sources by FastICA after removing outliers by Mahalanobis distance, (g) β -weight for each data point, (h) Mixed data after removing outliers using β-weight function , (i) Recovered sources by FastICA after removing outliers by β-weight function.

Then we remove outliers from data set 2 using our proposed method. The Figure 2g shows the β-weight for each data points. Figure 2h shows the mixed data points after removing outliers by proposed method. Clearly we see that almost all outliers are removed by our proposed method. Figure 2i shows the recovered sources by FastICA after removing outliers by our proposed method. Clearly we see that recovered sources are similar to the originals sources as shown in 1a.

4.1 Images Processing

To demonstrate the performance of the proposed method for image processing, we considered two 256×256 pixels original images of flower and Gaussian noise as shown in figure (3a-3b), respectively. Figure 3c represents the scatter plot of these two original images. Then we mixed these two original images using the linear ICA model. Figures (3d-3e) represent the mixture of flower image and Gaussian noise image respectively. Figure 3f represents the scatter plot of these two mixed images. To recover original images from the mixture, we first apply well known FastICA algorithm. Figure (3g-3h) shows the recovered images of flower and Gaussian noise, respectively by FastICA. Figure 3i represents the scatter plot of these two separated images. Then we apply FastICA algorithm to recover the original images using our proposed method again from the same mixture. Figure (3j-3k) shows the recovered images of flower and Gaussian noise, respectively by our proposed method. In absence of outliers, clearly we see that performance of robust FastICA and our proposed method based FastICA is good and recovers almost similar images.

Image Source Separation in Absence of Outliers

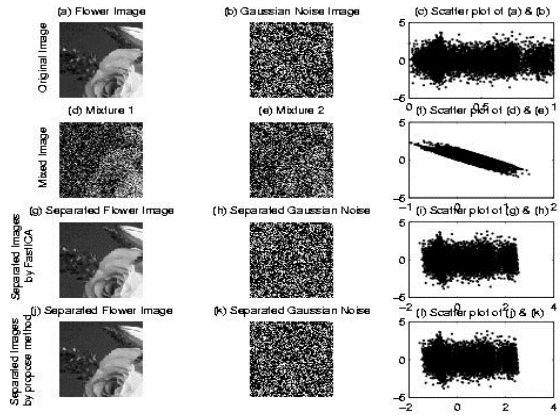


Figure 3. (a) Original flower image, (b) Original Gaussian noise image, (c) Scatter plot of original images, (d) Mixture of flower image, (e) Mixture of Gaussian noise image, (f) Scatter plot of two mixed images, (g) Flower image recovered from the mixed images by FastICA, (h) Gaussian noisy image recovered from the mixed images by

FastICA, (i) Scatter plot of two recovered images by FastICA, (j) Flower image recovered from the mixed images by proposed method, (k) Gaussian noisy image recovered from the mixed images by proposed method, (l) Scatter plot of two recovered images by proposed method.

To investigate the performance of FastICA in presence of outliers in a comparison of Mahalanobis distance, we added 256×6 pixel values with each mixed images from Gaussian distribution as outliers. Figures (4a-4b) represent the mixture of flower image and Gaussian noise image in presence of outlying pixel values (+), respectively. Figure 4c shows scatter plot of two mixed images in presence of outliers. To recover original images from the mixture in presence of outliers, we first apply FastICA algorithm after removing outliers by Mahalanobis distance. Figure (4d-4e) shows the recovered images by Mahalanobis D^2. Figure 4f shows scatter plot of two recovered images by Mahalanobis distance. Clearly we see that performance of FastICA using Mahalanobis distance is not good in presence of outliers and recovered images cannot recognize the original images of flower and Gaussian noise. Then we apply the FastICA algorithm after removing outliers by our proposed method with β= 0.1 to recover the original images again from the same mixture. Figure (4g-4h) shows the recovered images of flower and Gaussian noise, respectively. Obviously, it is seen that the performance of robust FastICA using our proposed method is good and recovered images can easily recognize the original image of flower and Gaussian noise.

Image Source Separation in Presence of Outliers

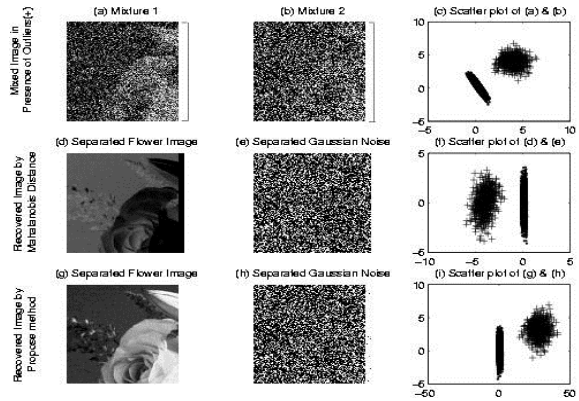


Figure 4. (a) Mixture 1 with 256×6 pixels outliers, (b) Mixture 2 with 256×6 pixels outliers, (c) Scatter plot of two mixed images in presence of outliers (+), (d) Flower image recovered by FastICA using Mahalanobis D^2, (e) Gaussian noise image recovered by FastICA using Mahalanobis D^2, (f) Scatter plot of two recovered images by FastICA using Mahalanobis D^2, (g) Flower image recovered from the mixed images by FastICA using proposed method, (h) Gaussian noisy image recovered from the mixed images by FastICA using proposed method, (i) Scatter plot of two recovered images by FastICA using proposed method.

5 Conclusion

The two popular robust ICA algorithms are minimum βdivergence method [7] and FastICA algorithm [5]. But when the data contains outlier they do not work well as usual. In this paper we discuss the robustification of ICA using an outlier rejection rule [10]-[11]based on β-weight function [9]. The values of the tuning parameter β play the key role in the performance of the proposed method. A cross validation technique is discussed as an adaptive selection procedure for the tuning parameter β [9]. If adaptive selection procedures produce $\beta > 0$, then data set is corrupted by outliers. If adaptive selection procedure produce $\beta = 0$, then data set is not corrupted by outliers. Both simulation and real image data results show that FastICA algorithm is able to recover all hidden signals properly after removing outliers by our proposed method. But FastICA algorithm is not able to recover all hidden signals properly after removing outliers by Mahalanobis distance.

References

[1] Cichoki, A. and Amari, S. (2002). *Adaptive Blind Signal and Image Processing.* Wiley,New York.

[2] Comon, P. (1994). Independent component analysis - a ew concept? *Signal Processing,*Vol.36, pp. 287-314.

[3] Hyv̈arinen , A. (1999): Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Network,* 10(3), 626-34.

[4] Hyv̈arinen , A. Karunen, J. and Oja, E. (2001): *Independent Component Analysis,* New York: Wiley.[5] Lee, T.-W. (2001): *Independent Component Analysis: Theory and applications,* Kluwer Academic Publishers.

[6] Minami, M. and Eguchi, S. (2002): Robust Blind Source eparation by beta-Divergence.*Neural Computation* 14, 1859-1886.

[7] Mollah, M.N.H., Minami, M. and Eguchi, S. (2006): Exploring latent structure of mixture ICA models by

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Topic Spotting on Bangla Document Corpus by Support Vector Machines and Compared to other Methods of Text Mining

Ashis Kumar Mandal[1,*], Md. Delowar Hossain[2], Rikta Sen[3] and Tangina Sultana[3]

[1] Department of CEN, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh.
[2]Department of CIT, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh.
[3] Department of ETE, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh.
[4] Department of TEE, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh.

Abstract. This paper explores the use of Support Vector Machines (SVM) to topic spotting. This is a kind of text categorization, or more precisely, the task of automatically sorting a set of documents into categories from a predefined set. Although recently SVM based topic spotting is used in English language for text categorization with good performance, relatively few studies have been done on Bangla language. Hence we analyze the efficiency of SVM for Bangla based topic spotting. In order to validate, Bangla articles corpus from various sites are used as examples for the experiment of this paper. For Bangla, empirical results support that SVM attain good performance over the currently other performing methods including decision tree, Naive Bayes (NB), K-Nearest Neighbor (KNN) and behave robustly in terms of high dimensional and relatively noisy document feature vectors.

Keywords: SVM, Text categorization, Bangla aricles corpus, topic spotting

1 Introduction

Topic spotting is an active research area of text mining where the documents are categorized with supervised knowledge. For this purpose a number of statistical and machine learning techniques has been developed including regression model, k-nearest neighbor , decision tree, Naïve Bayes, Support Vector Machines (Cortes and Vapnik, 1995), using n-grams and so on (Berry and Castellanos, 2007). Despite of the fact that these techniques are being used in English text analysis, many scopes have been revealed for Bangla language as Bangla based electronic documents both in online and offline have been being emerged very quickly.

In this paper, we study how the information from bangle online text documents can be categorized into five different classes and then classify using Support Vector Machine (SVM). In our experiments we use LIBSVM (Chang and Lin, 2001) package for the sigmoid kernels. For better topic spotting, we firstly preprocess training documents by some sequential steps: tokenization, digit removal, punctuation removal, and stop words removal. After that, each token is transformed by character n-gram technique. Secondly, feature extraction is performed by statistical approach. Finally, we construct models from the entire training data .For all data sets, 10-fold cross-validation is used for testing and training set. In our experiment we compare SVM with other three prevalent methods KNN, Decision Tree and NB in terms of categorization of these documents. Result shows that SVM outperform those two methods, achieving 91.18% accuracy.

2 Related Works

Traditionally, many text categorization tasks have been solved manually, but such manual classification is expensive to scale and also labor intensive. Therefore, most popular approach is to categorization of text using machine learning automatically (Sebastiani, 2002) .Considerable work has been done in text categorization of the English documents (Jin-Shu et al., 2006). In addition to English language there are many studies in European languages such as French, German, Spanish (Ciravegna et al., 2000) and in Asian languages such as Chinese and Japanese (Peng et al. 2003).For some Southern Indian Languages Naïve Bayes , Neural networks have been applied to news articles to automatically categorized predefine classes(Rajan et al. 2009). Nevertheless little works has been done on under resourced language Bangla due to lack of resources, annotated corpora, name dictionaries, morphological analyzer. One of the works is applying n-gram technique to categorize bangle news paper corpus (Monsur et al., 2006).

3 Proposed Dataset

A corpus we used for topic spotting contains 150 Bangla text documents- all in Unicode encoding. These documents are taken from the Bangla news web sources such as prothom-alo.com, online_dhaka.com, bdnews24.com, dailykalerkantha.com, bbc.co.uk/Bengali, ittefaque.com etc, but all the documents are specific subject related. As topic spotting is a supervised learning, meaning we have predefined classes, so we have 5 classes for this corpus, these are: বাণিজ্য (business), খেলা (sports), স্বাস্থ্য (health), প্রযুক্তি (technology), শিক্ষা (education) . Each class contains approx. 20-25 documents.

4 Methodology

4.1 Pre-processing

Choosing an appropriate representation of words in text documents is crucial to obtaining good classification performance. Therefore, feature extraction or transforming the input data into the set of features in indispensable. It is expected that proper feature extraction will extract the relevant information from the input data, and due to the high dimensionality of feature sets, feature extraction can be performed to reduce the dimensionality of the feature space and improve the efficiency (Forman and Kirshenbaum, 2008).

But before extracting features from the documents, we have to do some preprocessing. In preprocessing phase we represent each original text document as "Bag of words". Then following operations are done on each document:

- Tokenization:

  Tokenization is the process of breaking the sentences as well as the text file into word delimited by white space or tab or new line etc. Outcome of this tokenization phase is a set of word delimited by white space.

- Digit Removal:

  A general Bengali text file may contain Bengali as well as English digits. But as meaningful Bengali words do not contain digits, we remove these digits by using their Unicode representation.

- Punctuation Removal:

  Remove punctuation marks, special symbols (<,>, :,{,},[,],^,&,*,(,) , | etc.) from the Bengali text documents. Also, if a document contains excess use of spaces, tabs, shift, remove them.

- Stop words Removal:

  Stop words are the frequently occurring set of words which do not aggregate relevant information to the text classification task. Therefore, we have to remove these words from the text documents. We have made a list of Bengali language stop words from the dataset. This corpus contains around 364 words. Besides, there exist a lot of words having a single letter. Most of these Single-Letter-Words have little value. As a step of our processing the stop-words need to be removed before further processing. So the Single-Letter-Words are removed in this phase.

- n-Gram:

  Now these refined words in documents are transformed by n-Gram approach. Character n-Gram creates all possible n-Grams of each token in a document. HereCharacter sequence of length n is fixed for our corpus of documents and this is 3.

4.2 Feature Extraction

After pre-processing phase we have documents with less numbers of words, and extracting features from these words now become easier. The collection of words that are left in the document after all those steps are considered as a formal representation of the document, and we call the words in this collection terms. This is our text corpus. Various types of statistical approaches can be used to extract features from this text corpus. We use normalized (term frequency–inverse document frequency) TFIDF weighting with length normalization to extract the features from the document (Joho and Sanderson, 2007) as this method performs better than may other methods.

A combination of term frequency and inverse document frequency called TFIDF is commonly used to represent term weight numerically. The weight for a term *i* in terms of TF-IDF is given by

$$w_i = \frac{\left(TF_i \times \log\left(\frac{N}{n_i} \right) \right)}{\sqrt{\sum_{i=1}^{n} \left(TF_i \times \log\left(\frac{N}{n_i} \right) \right)^2}} \quad (1)$$

Where N is total number of documents and $n_i$is document frequency of term *i.*

4.3 Support Vector Machine Implementation

SVM has been successfully used on text classification. SVM was introduced by Cortes and Vapnik (1995) as a class of supervised machine learning techniques which is actually a binary classifier. It is based on the principle of structural risk minimization. In linear classification, SVM creates a hyper plane that separates the data into two sets with the maximum-margin. A hyper plane with the maximum-margin has the distances from the hyper plane to points when the two sides are equal. Mathematically, SVMs learn the sign function $f(x) = sign(wx + b)$, where

w is a *n* weighted vector in R^n . SVMs find the hyper plane $y = wx + b$ by separating the space R^n into two half-spaces with the maximum-margin. This Linear SVMs can be generalized for non-linear and multi class problem. The former is done by mapping data into another space H and performing the linear SVM algorithm over this new space, while the latter is done by decomposing the multi class problem into k binary problems.

Once the features are extracted and normalized, using SVM we have to train a data set to obtain a model and secondly, using the model to predict information of a testing data set. For implementing SVM, a Software called LIBSVM by Chung Chang and Chih-Jen Lin was used. LIBSVM is integrated software for support vector classification, regression and distribution estimation. This LIBSVM is also an implementation of the SVM classifier algorithm that supports multiclass classification. According to our preliminary tests, the best results were achieved by the C-SVC method with the sigmoid kernel and we use this configuration. The values from the testing file are fed into the LIBSVM tool for training and predicting the data set and analysis is done.We keep the parameters that obtain the best accuracy using a10-fold Cross Validation on the training set.

Like SVM we use popular text classification algorithms including probabilistic algorithms NB, instance based algorithm KNN, decision tree classifier based on C4.5 on same bangla text corpus so as to assess performances in comparison with SVM.

5 Experiments and Results

In this section some preliminary results will be illustrated and discussed. Tests were performed on four classifiers: NB, k-NN Classifier, Decision tree(C4.5) and SVM. For each experiment we measured the classification accuracy (ratio between correctly and incorrectly classified articles) and the evaluation was performed as "X-validation" (x=10) with stratified sampling. Among four classifiers applied on five corpora, SVMs achieved the highest average accuracy (91.18%), then NB with average accuracy of 87.22% and DT with average accuracy of 80.65%. KNN was the worst with average accuracy of 58.24%. Figure 1 shows the classifiers average performance.

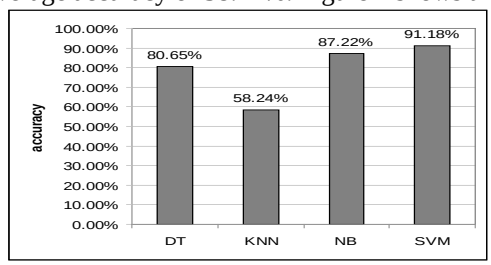


Fig 1 Average classification accuracy for four classifiers

Table 1 shows the percent accuracy obtained from 10-fold cross-validation by each method on the five corpora. These results demonstrate that topic spotting with our SVM improves text classification performance. This improvement is seen across different classification methods and different corpora sets.

Table 1 Text classification performance in percent accuracy

Classifiers	বাণিজ্য(business)	খেলা(sports)	স্বাস্থ্য(health)	প্রযুক্তি (technology)	শিক্ষা(education)
DT	82.35	86.49	80.56	63.29	90.57
KNN	54.23	59.33	51.69	51.59	74.24
NB	95.25	84.21	91.79	73.11	91.75
SVM	93.11	95	99.56	74.5	93.75

Training time is also an important factor for building any classification system. Due to nature of high dimensionality of text dataset, training takes time. Figure 2 shows training time in seconds for the four text classifiers. *SVMs* and *NB* variants classifiers take shortest time for training, while *DT* required the longest training time. DT is not scalable in high dimensional dataset, and it requires very long training time.

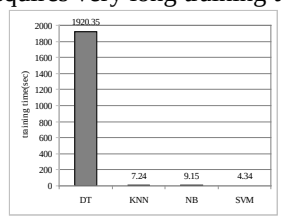


Fig 2 Average classifiers training time

6 Conclusions and Future Direction

This paper presented the results of topic spotting of Bangla text documents on five different Bangla corpora by using recognized machine learning techniques. A tool was implemented for feature extraction and selection. Besides, the performance of four popular classification algorithms C4.5, NB, KNN, and SVM has been evaluated on categorizing Bangla corpora; SVM classifier, in general, gives better performance. In our future work, we plan to introduce other classification algorithms in addition to those used here. Additionally, we plan to utilize other feature selection and weighting methods and compare them with the methods already used. Finally, we will continue to investigate the effect of each factor on the accuracy of the classification of Bangla text. Due to very high dimensionality of text data, popular dimensionality reduction techniques include term stemming and pruning might be added extra benefits of Bangla text mining.

References

Berry M. W. and Castellanos M. (2007) 'Survey of Text Mining II: Clustering, Classification and Retrieval' Springer

Ciravegna F., Gilardoni L., Lavelli A., Ferraro M., Mana N., Mazza, S., Matiasek J., Black W. and Rinaldi F. (2000) 'Flexible Text Classification for Financial Applications: the FACILE System*' In Proceedings of PAIS-2000, Prestigious Applications of Intelligent Systems sub conference of ECAI2000.*

Chang Chih-Chung, and Chih-Jen Lin (2001) 'LIBSVM: a library for support vector machines' Software available at http://www.csie.ntu.edu.tw/!cjlin/libsvm

Cortes, C. & Vapnik, V. (1995).Support-vector network.*Machine Learning*, 20, 1–25.

Forman G. and Kirshenbaum E. (2008) 'Extremely Fast Text Feature Extraction for Classification and Indexing', *In Proceeding CIKM '08 Proceedings of the 17th ACM conference on Information and knowledge management* ACM New York, NY, USA

Jin-Shu S., Bo-Feng Z., Xin X. (2006) 'Advances in Machine Learning Based Text Categorization' *Journal of Software*, Vol. 17, No.9, pp.1848-1859

Joho H. and Sanderson M.(2007) 'Document frequency and term specificity' i*n Large Scale Semantic Access to Content (Text, Image, Video, and Sound),*Paris, France pp. 350–359.

Mansur M., UzZaman N., and Khan M. (2006) 'Analysis of n-gram based text categorization for Bangla in a newspaper corpus' *In Proceedings of the 9th International Conference on Computer and Information Technology (ICCIT 2006).*

Peng F., Huang X., Schuurmans D. , and Wang S. (2003) 'Text Classification in Asian Languages without Word Segmentation*' In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL 2003), Association* for Computational Linguistics, Sapporo, Japan.

Rajan K., Ramalingam V., Ganesan M., Palanivel S., Palaniappan B.(2009) 'Automatic Classification of Tamil documents using Vector Space Model and Artificial Neural Network'*Journal Expert Systems with Applications: An International Journal,* Vol. 36 Issue 8.

Sebastiani F. (2002) 'Machine learning in automated text categorization' *ACM Computing Surveys,* Vol. 34 number 1. pp.1-47.

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Intrusion Detection System (IDS) Using Support Vector Machine with Different Kernels

Md. Al Mehedi Hasan,[1,*] Mohammed Nasser[2]

[1]Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi

[*]Ph.D. Student, Department of Computer Science & Engineering, Rajshahi University, Rajshahi

[2]Department of Statistic, Rajshahi University, Rajshahi

e-mail: mehedi_ru@yahoo.com, mnasser.ru@gmail.com

Abstract. The success of any Intrusion Detection System (IDS) is a complicated problem due to its nonlinearity and the quantitative or qualitative network traffic data stream with many features. To get rid of this problem, several types of intrusion detection methods have been proposed and shown different levels of accuracy. This is why, the choice of the effective and robust method for IDS is very important topic in information security. Support vector machine (SVM) has been employed to provide potential solutions for the IDS problem. However, the practicability of SVM is affected due to the difficulty of selecting appropriate kernel and its parameters. Thus, this paper is aimed to use different kernel on the NSL-KDD Dataset and find out which is best for SVM based intrusion detection system. NSL-KDD dataset is new version of KDD'99 dataset and has some advantages over KDD'99. The experimental results indicate that polynomial kernel can achieve higher detection rate and lower false negative rate than others kernel like Linear and RBF kernels in the same time.

Keywords: Intrusion detection, support vector machine, Kernel, KDD'99, NSL-KDD

1 Introduction

Along with the benefits, the Internet also created numerous ways to compromise the stability and security of the systems connected to it. Although static defense mechanisms such as firewalls and software updates can provide a reasonable level of security, more dynamic mechanisms such as intrusion detection systems (IDSs) should also be utilized [1]. Intrusion detection is the process of monitoring events occurring in a computer system or network and analyzing them for signs of intrusions. IDSs are simply classified as host-based or network-based. The former operates on information collected from within an individual computer system and the latter collect raw network packets and analyze for signs of intrusions. There are two different detection techniques employed in IDS to search for attack patterns: Misuse and Anomaly. Misuse detection systems find known attack signatures in the monitored resources. Anomaly detection systems find attacks by detecting changes in the pattern of utilization or behavior of the system [2].

As network attacks have increased in number and severity over the past few years, Intrusion Detection Systems (IDSs) have become a necessary addition to the security infrastructure of most organizations [3]. Deploying highly effective IDS systems is extremely challenging and has emerged as a significant field of research, because it is not theoretically possible to set up a system with no vulnerabilities [4]. Several machine learning (ML) algorithms, for instance Neural Network [5], Genetic Algorithm [6], Fuzzy Logic [4, 7, 8, 9], clustering algorithm [10] and more have been extensively employed to detect intrusion activities from large quantity of complex and dynamic datasets.

In recent times, Support Vector Machine (SVM) has been extensively applied to provide potential solutions for the IDS problem. But, the selection of an appropriate kernel and its parameters for a certain classification problem influence the performance of the SVM because different kernel function constructs different SVMs and affects the generalization ability and learning ability of SVM. There is no theoretical method for selecting kernel function and its parameters. Literature survey showed that, for all practical purposes, most of the researchers applied Gaussian kernel to build SVM based intrusion detection system [11, 12, 13, 14] and find its parameter value using different technique which is not unique and some research paper did not mention its value [13]and some uses the default value of the software package [15]. But there are many other kernel functions which are not yet applied in intrusion detection. Other kernels should also be used in comparison to find optimal results for applying SVM based approach depending upon the nature of classification problem [13]. This motivated us to apply different kernel functions of SVM apart from RBF for IDS classification purpose which may provide better accuracy and detection rate depending on different nonlinear separations. We have also tried to find out parameter value to the corresponding kernel. In this paper, we provide a review of the SVM and its kernel approaches in IDS for future research and implementation towards the development of optimal approach in intrusion detection system with maximum detection rate and minimized false alarms.

The remainder of the paper is organized as follows. Section 2 provides the description of the KDD'99 and NSL-KDD dataset. We outline mathematical overview of SVM in Section 3. Experimental setup is presented in Section 4 and Preprocessing, Evaluation Metrics and SVM model selection are drawn in Section 5, 6 and 7 respectively. Finally, Section 8 reports the experimental result followed by conclusion in Section 9.

2 KDDCUP'99 vs NSL-KDD Dataset:

Under the sponsorship of Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL), MIT Lincoln Laboratory has collected and distributed the datasets for the evaluation of researches in computer network intrusion detection systems [16]. The KDD'99 dataset is a subset of the DARPA benchmark dataset prepared by Sal Stofo and Wenke Lee [17]. The KDD data set was acquired from raw tcpdump data for a length of nine weeks. It is made up of a large number of network traffic activities that include both normal and malicious connections. A connection in the KDD-99 dataset is represented by 41 features, each of which is in one of the continuous, discrete and symbolic form, with significantly varying ranges. The KDD99 data set includes three independent sets; ''whole KDD'', ''10% KDD'', and ''corrected KDD''. Most of researchers have used the ''10% KDD'' and the ''corrected KDD'' as training and testing set, respectively [18]. The training set contains a total of 22 training attack types. Additionally the ''corrected KDD'' testing set includes an additional 15 attack types and therefore there are 37 attack types that are included in the testing set, as shown in Table I and Table II. The simulated attacks fall in one of the four categories [1, 18]: (a) Denial of Service Attack (DoS), (b) User to Root Attack (U2R), (c) Remote to Local Attack (R2L), (d) Probing Attack.

Table I. Attacks in KDD'99 Training dataset

Classification of Attacks	**Attack Name**
Probing	Port-sweep, IP-sweep, Nmap, Satan
DoS	Neptune, Smurf, Pod, Teardrop, Land, Back
U2R	Buffer-overflow, Load-module, Perl, Rootkit
R2L	Guess-password, Ftp-write, Imap, Phf, Multihop, spy, warezclient, Warezmaster,

Table II. Attacks in KDD'99 Testing dataset

Classification of Attacks	**Attack Name**
Probing	Port-Sweep, Ip-Sweep, Nmap, Satan, **Saint, Mscan**
DoS	Neptune, Smurf, Pod, Teardrop, Land, Back, **Apache2,Udpstorm, Processtable,Mail-Bomb**
U2R	Buffer-Overflow, Load-Module, Perl, Rootkit, **Xterm, Ps, Sqlattack.**
R2L	Guess-Password, Ftp-Write, Imap, Phf, Multihop, Spy, Warezclient, Warezmaster, **Snmpgetattack, Named, Xlock, Xsnoop, Send-Mail, Http-Tunnel, Worm, Snmp-Guess.**

Statistical analysis on KDD'99 dataset found important issues which highly affects the performance of evaluated systems and results in a very poor evaluation of anomaly detection approaches [Principle Component Analysis, 13]. To solve these issues, researchers have proposed a new data set, NSL-KDD, which consists of selected records of the complete KDD data set [15]. The NSL-KDD dataset does not include redundant records in the train set, so the classifiers will not be biased towards more frequent records. The numbers of records in the train and test sets are reasonable, which makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research works will be consistent and comparable.

3 SVM classification

The theory of Support Vector Machine (SVM) is from statistics and the basic principle of SVM is finding the optimal linear hyperplane in the feature space that maximally separates the two target classes [19, 20]. There are two types of data namely linearly separable and non-separable data. To handle these data, two types of classifier, linear and non-linear, are used in pattern recognition field.

3.1 Linear Classifier

Consider the problem of separating the set of training vectors belong to two linear separate classes, $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$where $x_i \in R^n, y_i \in \{-1, +1\}$ with a hyperplane $w^T x + b = 0$. Finding a separating hyperplane can be posed as a constraint satisfaction problem. For this problem, the constraint problem can be defined as follows find w and b such that

$$w^T x_i + b \geq 1 \; if \; y_i = +1$$
$$w^T x_i + b \leq -1 \; if \; y_i = -1$$
$$where \; i = 1,2,3, \ldots \ldots, n$$

Considering the maximum margin classifier, there is hard margin SVM, applicable to a linearly separable dataset, and then modifies it to handle non-separable data. This leads to the following constrained optimization problem:

$$minimize_{w,b} \frac{1}{2}\|w\|^2$$
$$\text{Subject to: } y_i(w^T x_i + b) \geq 1, i = 1,2,3, \ldots \ldots, n \qquad (1)$$

The constraints in this formulation ensure that the maximum margin classifier classifies each example correctly, which is possible since we assumed that the data is linearly separable. In practice, data is often not linearly separable and in that case, a greater margin can be achieved by allowing the classifier to misclassify some points. To allow errors, the optimization problem now becomes:

$$min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{Subject to: } \quad y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1,2,3, \ldots \ldots, n \qquad (2)$$
$$\xi_i \geq 0 \;\; , i = 1,2,3, \ldots \ldots, n$$

The constant C > 0 sets the relative importance of maximizing the margin and minimizing the amount of slack. This formulation is called the soft-margin SVM [19, 20]. Using the method of Lagrange multipliers, we can obtain the dual formulation which is expressed in terms of variables α_i [19, 20]:

$$maximize_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i \, y_j x_i^T x_j$$
$$\text{Subject to: } \quad \sum_{i=1}^{n} y_i \alpha_i = 0, 0 < \alpha_i < C \text{ for all } i = 1,2,3, \ldots \ldots, n \qquad (3)$$

The dual formulation leads to an expansion of the weight vector in terms of the input examples:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Finally, the linear classifier based on a linear discriminant function takes the following form

$$f(x) = \sum_i^n \alpha_i x_i^T x + b \qquad (4)$$

3.2 Non-linear Classifier

In many applications a non-linear classifier provides better accuracy. The naive way of making a non-linear classifier out of a linear classifier is to map our data from the input space X to a feature space F using a non-linear function $\emptyset: X \rightarrow F$. In the space F, the discriminant function is:

$$f(x) = w^T \emptyset(x) + b.$$

Now, examine what happens when the nonlinear mapping is introduced into equation (3). We have to optimize

$$maximize_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i \, y_j \emptyset(x_i)^T \emptyset(x_j)$$
$$\text{Subject to:} \sum_{i=1}^{n} y_i \alpha_i = 0, 0 < \alpha_i < C \quad \text{for all } i = 1,2,3, \ldots \ldots, n \qquad (5)$$

Notice that the mapped data only occurs as an inner product in the objectives. Now, we can apply a little mathematically rigorous magic known as kernels. By Mercer's theorem, we know that for certain mapping $\emptyset(x)$ and any two points x_i and x_j, the inner product of the mapped points can be evaluated using the kernel function without ever explicitly knowing the mapping [21] . The kernel function can be defined as

$$k(x_i, x_j) = \emptyset(x_i)^T \emptyset(x_j)$$

Substituting the kernel in the equation 5, the optimization takes the following form:

$$maximize_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i \, y_j k(x_i, x_j)$$
$$\text{Subject to: } \sum_{i=1}^{n} y_i \alpha_i = 0 \;, 0 < \alpha_i < C \text{ for all } i = 1,2,3, \ldots \ldots, n \qquad (6)$$

Finally, in terms of the kernel function the discriminant function takes the following form:

$$f(x) = \sum_{i}^{n} \alpha_i k(x, x_i) + b$$

3.3 Kernel and its parameters selection:

A kernel function and its parameter have to be chosen to build a SVM classifier [14]. In this work, three main kernels have been used to build SVM classifier. They are

1. Linear kernel: $K(x_i, x_j) =< x_i, x_j >$
2. Polynomial kernel: $K(x_i, x_j) = (< x_i, x_j > +1)^d$, d is the degree of polynomial.
3. Gaussian kernel: $K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma})$, σ is the width of the function.

Training an SVM finds the large margin hyperplane, i.e. sets the parameters α_i (c.f. Equation 6). The SVM has another set of parameters called hyperparameters: The soft margin constant, C, and any parameters the kernel function may depend on (width of a Gaussian kernel or degree of a polynomial kernel)[22]. The soft margin constant C adds penalty term to the optimization problem. For a large value of C, a large penalty is assigned to errors/margin errors and creates force to consider points close to the boundary and decreases the margin. A smaller value of C (right) allows to ignore points close to the boundary, and increases the margin.

Kernel parameters also have a significant effect on the decision boundary [22]. The degree of the polynomial kernel and the width parameter σ of the Gaussian kernel control the flexibility of the resulting classifier. The lowest degree polynomial is the linear kernel, which is not sufficient when a non-linear relationship between features exists. Higher degree polynomial kernels are flexible enough to discriminate between the two classes with a sizable margin and greater curvature for a fixed value of the soft-margin constant. On the other hand in Gaussian Kernel, for a fixed value of the soft-margin constant, small values of σ the decision boundary is nearly linear. As σ increases the flexibility of the decision boundary increases and large values of σ lead to over fitting [22].

A question frequently posed by practitioners is "which kernel should I use for my data?". There are several answers to this question. The first is that it is, like most practical questions in machine learning, data-dependent, so several kernels should be tried. That being said, we typically follow the following procedure: Try a linear kernel first, and then see if we can improve on its performance using a non-linear kernel [22].

3.4 Multiclass support vector machine

Support vector machines are formulated for two class problems. But because support vector machines employ direct decision functions, an extension to multiclass problems is not straightforward [12]. There are several types of support vector machines that handle multiclass problems. We used here only one-vs-all multiclass support vector machines for our research work. The One-Vs-All techniqueis extended from the binary two-class problem to perform classification tasks with k > 2 classes. In this approach, the base classifier (in our case - SVM) is trained on K copies of the K-class original training set, with each copy having the K-th label as the positive label, and all other labels as the negative label (combined class). We denote the optimal separating hyperplane discriminating the class j and the combined class as

$$g^j = \mathrm{x}^{\mathrm{T}}\hat{\mathrm{w}}^{\mathrm{j}} + \hat{\mathrm{b}}^{\mathrm{j}}, \qquad \mathrm{j} = 1,,2,3,\ldots,\mathrm{k}$$

where the superscript in $\hat{w}^j$ stands for the class which should be separated from the other observations. After finding the all k optimal separating hyperplanes, the final classifier has been defined by

$$f_k(x) = argmax_j(g^j(x))$$

In this approach the index of the largest component of the discriminant vector $(g^1(x), g^2(x), \ldots\ldots, g^k(x))$ is assigned to the vector x. In other words, each input is classified by all K models, and the output is chosen by the model with the highest degree of confidence.

4 Dataset and Experimental setup

Investigating the existing papers on the anomaly detection which have used the KDD data set, we found that a subset of KDD'99 dataset has been used for training and testing instead of using the whole KDD'99 dataset [13, 15, 23, 24]. Existing papers on the anomaly detection mainly used two common approaches to apply KDD [15]. In the first, KDD'99 training portion is employed for sampling both the train and test sets. However, in the second approach, the training samples are randomly collected from the KDD train set, while the samples for testing are arbitrarily selected from the KDD test set. The basic characteristics of the NSL-KDD and the original KDD 99 intrusion detection datasets in terms of number of samples is given in Table III. Although the distribution of the number of samples of attack is different on different research papers, we have used the Table I and II to find out the distribution of attack [1, 3, 18]. In our experiment, whole NSL-KDD train set (KDDTrain+) has been used to train our classifier and the NSL-KDD (KDDTest+) test set has been used to test the classifier. All experiments were performed using Intel core i5 2.27 GHz processor with 4GB RAM, running Windows 7.

To select the best model in model selection phase, we have drawn 10% samples from the training set (KDDTrain+) to tune the parameters of all kernel and another 10% samples from the training set (KDDTrain+) to validate those parameters, as shown in Table III. In our experiment, three different types of kernel have been used.

Table III. Number of Samples of Each Attack in Dataset

Dataset	Normal	DoS	Probing	R2L	U2R	Total
Whole KDD (Original KDD)	972780	3883370	41102	1126	52	4898430
10% KDD (Original KDD)	97278	391458	4107	1126	52	494021
KDD corr (Original KDD)	60593	229853	4166	16347	70	311029
KDDTrain+ (NSL-KDD)	67343	45927	11656	995	52	125973
KDDTest+ (NSL-KDD)	9711	7458	2421	2887	67	22544
Train Set (For Model Selection)	6735	4593	1166	100	6	12600
Validation Set (For Model Selection)	6735	4593	1166	100	6	12600

5 Pre-processing

SVM classification system is not able to process NSLKDD dataset in its current format. Hence preprocessing was required before SVM classification system could be built. Preprocessing contains the following processes: SVM requires that each data instance is represented as a vector of real numbers. The features in columns 2, 3, and 4 in the NSLKDD or KDD'99 dataset are the protocol type, the service type, and the flag, respectively. The value of the protocol type may be tcp, udp, or icmp; the service type could be one of the 70 different network services such as http and smtp; and the flag has 11 possible values such as SF or S2. Hence, the categorical features in the KDD dataset must be converted into a numeric representation. This is done by the usual binary encoding – each categorical variable having possible m values is replaced with m-1 dummy variables. Here a dummy variable have value one for a specific category and having zero for all category. After converting category to numeric, we got 119 variables for each samples of the dataset. Some researchers used only integer code to convert category features to numeric representation instead of using dummy variables which is not statistically meaningful way for this type of conversion [13, 18]. The final step of pre-processing is scaling the training data, i.e. normalizing all features so that they have zero mean and a standard deviation of 1. This avoids numerical instabilities during the SVM calculation. We then used the same scaling of the training data on the test set. Attack names were mapped to one of the five classes namely Normal, DoS (Denial of Service), U2R (user-to-root: unauthorized access to root privileges), R2L (remote-to-local: unauthorized access to local from a remote machine), and Probe (probing: information gathering attacks).

6 Evaluation Metrics

Apart from accuracy, developer of classification algorithms will also be concerned with the performance of their system as evaluated by False Negative Rate, False Positive Rate, Precision, Recall, etc. In our system, we have considered both the precision and false negative rate. To consider both the precision and false negative rate is very important in IDS as the normal data usually significantly outnumbers the intrusion data in practice. To only measure the precision of a system is misleading in such a situation [25]. The classifier should produce lower false negative rate because an intrusion action has occurred but the system considers it as a non-intrusive behavior is very cost effective.

7 SVM Model Selection

In order to generate highly performing SVM classifiers capable of dealing with real data an efficient model selection is required. In our experiment, Grid-search technique has been used to find the best model for SVM with different kernel. This method selects the best solution by evaluating several combinations of possible values. In our experiment, Sequential Minimization Optimization with the following options in Matlab, shown in Table IV, has been used. We have considered the range of the parameter in the grid search which converged within the maximum iteration using the train set (For Model Selection) and validation set (For Model selection) shown in Table III.

Table IV. Sequential Minimization Optimization Options

Option	Value
MaxIter	1000000
KernelCacheLimit	10000

For linear kernel, to find out the parameter value C, we have considered the value from 2^{-8} to 2^{8} as our searching space. The resulting search space for linear kernel is shown in Fig. I. We took parameter value C=16 for giving us 99.56% accuracy in the validation set to train the whole NSL-Kdd train data (KDDTrain+) and test the NSL-Kdd test data (KDDTest+).

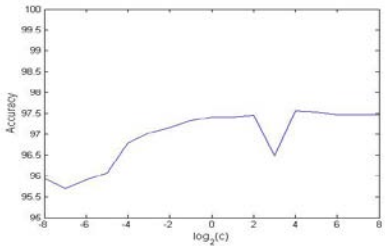


Fig. I: Tuning Linear Kernel

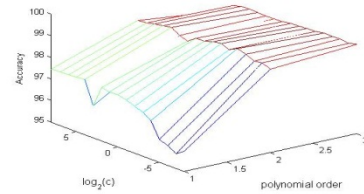


Fig. II: Tuning Polynomial Kernel

For polynomial kernel, to find the parameter value C (penalty term for soft margin) and d (poly order), we have considered the value from 2^{-8} to 2^{8} for C and from 1 to 3 for d as our searching space. The resulting search space for polynomial kernel is shown in Fig. II. We took parameter value d=2 and C=0.0625 for giving us 99.24% accuracy in the validation set to train the whole NSL-Kdd train data (KDDTrain+) and test the NSL-Kdd test data (KDDTest+).

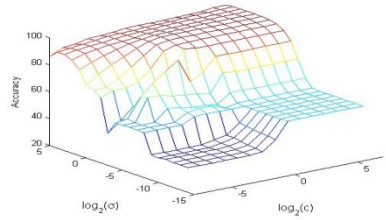


Fig. III: Tuning Radial Basis Kernel

For radial basis kernel, to find the parameter value C (penalty term for soft margin) and sigma, we have considered the value from 2^{-8} to 2^{6} for C and from 2^{-15} to 2^{5} for sigma as our searching space. The resulting search space for radial basis kernel is shown in Fig. III. We took parameter value C=32 and sigma=2 for giving us 99.01% accuracy in the validation set to train the whole NSL-Kdd train data (KDDTrain+) and test the NSL-Kdd test data (KDDTest+).

8 Obtained Result

The final training/test phase is concerned with the production and evaluation on a test set of the final SVM model created based on the optimal hyper-parameters set found so far in the model selection phase [19]. After finding the parameter, we built the model using NSL-Kdd train set for each of the kernel tricks and finally we have tested the model using NSL-Kdd test set. The training and testing results are given in Table V according to the classification accuracy. From the results it is observed that the test accuracy for polynomial kernel is better than linear and radial basis kernel.

Table V: Training and Testing Accuracy

Kernel	Training Accuracy	Testing Accuracy
Linear	86.56	63.60
Polynomial	99.31	73. 54
Radial Basis	99.80	56.88

For the test case**, t**he confusion matrix for each of the kernel is given in Table VI, VII and VIII respectively. Going into more detail of the confusion matrix, it can be seen that Linear kernel performs better on probing attack detection and polynomial kernel performs well on Dos, R2L, and U2R detection.

Table VI: Confusion matrix for Linear Kernel

		Actual					
		Dos	Normal	Probing	R2L	U2R	%
Prediction	Dos	3095	107	180	22	23	90.31
	Normal	1382	9324	331	2784	31	67.31
	Probing	2977	262	1910	51	5	36.70
	R2L	4	13	0	2	2	9.52
	U2R	0	5	0	28	6	15.38
	%	41.50	96.01	78.89	0.07	8.96	

Table VII: Confusion matrix for Polynomial Kernel

Prediction	Actual					
	Dos	Normal	Probing	R2L	U2R	%
Dos	5410	66	538	15	3	89.69
Normal	1222	9192	364	2304	34	70.08
Probing	823	431	1515	113	7	52.44
R2L	0	16	1	447	9	94.50
U2R	3	6	3	8	14	41.18
%	72.54	94.66	62.58	15.48	20.90	

Table VIII: Confusion matrix for Radial Basis Kernel

Prediction	Actual					
	Dos	Normal	Probing	R2L	U2R	%
Dos	2198	10	0	0	0	99.55
Normal	3903	9395	1250	2815	65	53.91
Probing	1357	304	1171	7	0	41.25
R2L	0	2	0	58	2	93.55
U2R	0	0	0	7	0	0
%	29.47	96.75	48.36	2.01	0	

We also considered the false negative rate and precision for each of kernel and as shown in Table IX and X respectively. The polynomial kernel gives lower average false negative rate and high precision than other kernels.

Table IX: False Negative Rate of Different Kernels for each of the attack types.

Kernel	Dos	Probing	R2L	U2R	Average False Negative Rate
Linear	18.53	13.67	96.43	46.27	43.73
Polynomial	16.36	15.04	79.81	50.75	40.49
Radial Basis	52.33	51.63	97.50	97.01	74.61

Table IX: Precision of Different Kernels for each of the attack types.

Kernel	Dos	Probing	R2L	U2R	Average Precision
Linear	0.90	0.37	0.10	0.15	0.38
Polynomial	0.90	0.52	0.96	0.41	0.70
Radial Basis	0.99	0.41	0.94	0	0.59

9 Conclusion

In this research work, we developed an intrusion detection system using support vector machines as classifier. The performances of the different kernel based approaches have been observed on the basis of their accuracy, false negative rate and precision. The results indicate that the ability of the SVM classification depends mainly on the kernel type and the setting of the parameters. Research in intrusion detection using SVM approach is still an ongoing area due to good performance. The findings of this paper will be very useful for future research and to use SVM more meaningful way in order to maximize the performance rate and minimize the false negative rate.

References

1. Kayacik H. G., Zincir-Heywood A. N., Heywood M. I., "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Benchmark", Proceedings of the PST 2005 – International Conference on Privacy, Security, and Trust, pp. 85-89, 2005.
2. Adetunmbi A.Olusola., Adeola S.Oladele. and Daramola O.Abosede. "Analysis of KDD ’99 Intrusion Detection Dataset for Selection of Relevance Features", Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I WCECS 2010, October 20-22, 2010.
3. Hesham Altwaijry ,Saeed Algarny, "Bayesian based intrusion detection system", Journal of King Saud University – Computer and Information Sciences, pp.1–6, 2012.
4. O. Adetunmbi Adebayo, Zhiwei Shi, Zhongzhi Shi, Olumide S. Adewale, "Network Anomalous Intrusion Detection using Fuzzy-Bayes", IFIP International Federation for Information Processing, Vol: 228, pp: 525-530, 2007.
5. Cannady J, “Artificial Neural Networks for Misuse Detection”, in Proceedings of the ’98 National Information System Security Conference (NISSC’98), pp. 443-456, 1998.

6. Susan M. Bridges and Rayford B.Vaughn, “Fuzzy Data Mining And Genetic Algorithms Applied To Intrusion Detection”, In Proceedings of the National Information Systems Security Conference (NISSC), Baltimore, MD, pp.16-19, October 2000.
7. Abadeh, M.S., Habibi, J., "Computer Intrusion Detection Using an Iterative Fuzzy Rule Learning Approach", in Proceedings of the IEEE International Conference on Fuzzy Systems, pp: 1-6, London, 2007.
8. Bharanidharan Shanmugam, Norbik Bashah Idris, "Improved Intrusion Detection System Using Fuzzy Logic for Detecting Anamoly and Misuse Type of Attacks", in Proceedings of the International Conference of Soft Computing and Pattern Recognition, pp: 212-217, 2009.
9. Yao, J.T., Zhao, S.L., Saxton, L.V., "A study on Fuzzy Intrusion Detection", Proc. of Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, pp. 23-30, 2005.
10. Qiang Wang and Vasileios Megalooikonomou, "A clustering algorithm for intrusion detection", in Proceedings of the conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, vol. 5812, pp. 31-38, March 2005.
11. Vipin Das, Vijaya Pathak, Sattvik Sharma, Sreevathsan, MVVNS.Srikanth, Gireesh Kumar T, "Network Intrusion Detection System Based On Machine Learning Algorithms", International Journal of Computer Science & Information Technology (IJCSIT), Vol 2, No 6, December 2010.
12. Arvind Mewada, Prafful Gedam, Shamaila Khan, M. Udayapal Reddy,"Network Intrusion Detection Using Multiclass Support Vector Machine", Special Issue of IJCCT Vol.1 Issue 2, 3, 4; 2010 for International Conference [ACCTA-2010], August 2010.
13. Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien, and Ajith Abraham, "Principle Components Analysis and Support Vector Machinebased Intrusion Detection System", *10th International Conference on Intelligent Systems Design and Applications, 2010*.
14. V.Jaiganesh, Dr. P. Sumathi, "Intrusion Detection Using Kernelized Support Vector Machine With Levenbergmarquardt Learning", International Journal of Engineering Science and Technology (IJEST), Vol. 4 No.03 March 2012.
15. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", in Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, pp. 53-58, Ottawa, Ontario, Canada, 2009.
16. MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation, http://www.ll.mit.edu/CST.html,MA, USA. July, 2010.
17. KDD’99 dataset, http://kdd.ics.uci.edu/databases, Irvine, CA, USA, July, 2010.
18. M. Bahrololum, E. Salahi and M. Khaleghi, "Anomaly Intrusion Detection Design Using Hybrid Of Unsupervised And Supervised Neural Network", International Journal of Computer Networks & Communications (IJCNC), Vol.1, No.2, July 2009.
19. Vladimir N. Vapnik, The Nature of Statistical Learning Theory, Second Edition, Springer, New York, ISBN 0-387-98780-0, 1999.
20. Bernhard Scholkopf, Alexander J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond", The MIT Press Cambridge, Massachusetts London, England, 2001.
21. Kristin P. Bennett, Colin Cambell, "Support Vector Machines: Hype or Hallelujah? ", SIGKDDExplorations, Volume 2, Issue 2 ,pp.1-13, 2000
22. A. Ben-Hur and J. Weston. "A User’s guide to Support Vector Machines", In Biological Data Mining. Oliviero Carugo and Frank Eisenhaber (eds.) Springer Protocols, 2009.
23. Fangjun KUANG, Weihong XU, Siyang ZHANG, Yanhua WANG, Ke LIU , "A Novel Approach of KPCA and SVM for Intrusion Detection", Journal of Computational Information Systems, pp. 3237-3244, 2012.
24. Shilpa lakhina, Sini Joseph and Bhupendraverma, "Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD", International Journal of Engineering Science and Technology Vol. 2(6), pp.1790-1799, 2010.
25. JingTao Yao, Songlun Zhao, and Lisa Fan, "An enhanced support vector machine model for intrusion detection", RSKT'06 Proceedings of the First international conference on Rough Sets and Knowledge Technology, Pages 538-543, 2006.

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment,*** 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Seasonal Cycle Extraction from Climate Signals Using Discrete Wavelet Transform

Md. Khademul Islam Molla[1, 2, *], A. T. M. Jahangir Alam[3], Munmun Akter[2], A. R. Shoyeb Ahmed Siddique[2] and M. Sayedur Rahman[4]

[1]Department of Information and Communication Engineering, The University of Tokyo, Tokyo, Japan
[2]Department of Computer Science and Engineering, The University of Rajshahi, Rajshahi, Bangladesh
[3]Department of Environmental Sciences, Jahangirnagar University, Savar, Dhaka, Bangladesh
[4]Department of Statistics, The University of Rajshahi, Rajshahi, Bangladesh
[*]Email: molla@gavo.t.u-tokyo.ac.jp

Abstract. This paper presents a data adaptive filtering technique to extract seasonal cycles of different climate signals using discrete wavelet transform (DWT). The fractional Gaussian noise (fGn) is used here to determine adaptive threshold without any prior training. The climate signal and fGn are decomposed into a finite number of subband signals using DWT. The subband energy of fGn and its confidence intervals are computed and the upper bound of the confidence interval is used as the threshold. The lowest order (higher frequency) subband of the climate signal exceeding the threshold is determined as the starting subband to represent signal trend i.e. seasonal cycle. All the higher order subbands are summed up to extract the seasonal cycle. The experimental results illustrate the efficiency the proposed data adaptive approach to separate the seasonal cycle of the climate signals.

Keywords: Climate signal, discrete wavelet transform, subband representation, time domain filtering.

1 Introduction

The term ‘climate variability’ denotes the inherent characteristic of climate which manifests itself in changes of climate with time. The degree of climate variability can be described by the differences between long-term statistics of meteorological elements calculated for different periods. Climate variability is often used to denote deviations of climate statistics over a given period of time such as a specific month, season or year from the long-term climate statistics relating to the corresponding calendar period. To mitigate the effects of climate change risk assessmentsare being required more frequently by policy makers [1]. Climate is currently changing inways that mirror the effects of global warming. There is also increasing demand forclimate change information, particularly from policy makers for impact assessment studies [2]. Several linear statistical models have been applied to climate records, but the answersare not conclusive due to the high sensitivity of model results to model parameters [3-5],especially when stochastic processes are taken into account. Various approaches havebeen employed to develop climate change scenarios at different scales [6].

The seasonal cycle of any climate signals represents are good measure of climate variability. The variation of such cycle from the mean stream can be used to illustrate the variation of climate. If we consider the seasonal cycle for a year it can be termed as annual cycle. It is commonly estimated from observational data or model output by taking the average of all Januaries, all Februaries, and so forth. If the observational record is long enough and conditions are stationary (i.e. there is no significant long-term trend), a meaningful seasonal cycle will result that can be used to calculate an anomaly time series [7]. In the analysis of climate variability, the annual cycle refers to every twelve month (consecutive) length of data.

In this paper, the seasonal (annual) cycle of several climate signals are extracted using data adaptive filtering approach. The annual cycle is treated as the trend in the climate signal. It is required to develop a method to extract the trend representing the cycle using a data adaptive technique without affecting the other parts of the signal [7]. It needs a mathematical function used to divide a given function or time series into different scale components i.e. sub-bands [8]. The traditional approach – Fourier transform can be used to decompose the climate time series into several sub-bands to extract the trend of the signal [9]. Fourier transforms for representing functions that have discontinuities and sharp peaks and for accurately deconstructing and reconstructing finite, non-periodic and non-stationary signals. The discrete wavelet transform (DWT) is the representation of a function by wavelet which is very efficient to decompose the signal in a data adaptive nature [10]. Wavelet transform have advantages over traditional Fourier transform. In this study we attempt to obtain a better understanding of the variability through the analysis of seasonal climate cycles of different climate signals say rainfall, humidity with wavelet transform. The climate signal is decomposed into several subbands using wavelet. The subbands representing the trend are determined by comparing their energies with that of the reference signal. The fractional Gaussian noise is used here as the reference signal to separate the seasonal cycle in a data adaptive way.

2 Wavelet Based Subband Decomposition

The climate signal is considered as fully non-linear and non-stationary and hence wavelet transform is more suitable technique to be analyzed. The wavelet transform is computed by changing the scale of the analysis window, shifting the window in time,multiplying by the signal, and integrating over all times. The filters of different cutofffrequencies are used to analyze the signal at different scales [11]. The coefficients are usually sampled on adyadic grid, i.e., $s_0 = 2$ and $t_0 = 1$, yielding $s=2^j$ and $t=k*2^j$.

The DWT analyzes the signal atdifferent frequency bands with different resolutions by decomposing the signal into a coarse approximationand detail information. DWT employs two sets of functions, called scaling functions and wavelet functions,which are associated with low pass and highpass filters, respectively. The decomposition of the signal intodifferent frequency bands is simply obtained by successive highpass and lowpass filtering of the timedomain signal. The original signal $x(n)$ is first passed through a halfband highpass filter $g(n)$ and a lowpassfilter $h(n)$. After the filtering, half of the samples can be eliminated according to the Nyquist's rule, since thesignal now has a highest frequency of $\pi/2$ radians instead of π. The signal can therefore be subsampled by2, simply by discarding every other sample. This constitutes one level of decomposition and canmathematically be expressed as follows:

$$y_{high}(k) = \sum_n x(n) \cdot g(2k - n)$$
$$y_{low}(k) = \sum_n x(n) \cdot h((2k - n) \quad (1)$$

where$y_{high}(k)$ and $y_{low}(k)$ are the outputs of the highpass and lowpass filters, respectively, aftersubsampling by 2.This decomposition halves the time resolution since only half the number of samples now characterizes theentire signal. However, this operation doubles the frequency resolution, since the frequency band of the signal now spans only half the previous frequency band, effectively reducing the uncertainty in the frequency by half.

The above procedure, which is also known as the subband decomposition, can be repeated for furtherdecomposition. At every level, the filtering and subsampling will result in half the number of samples (andhence half the time resolution) and half the frequency band spanned (and hence double the frequency resolution) [11]. The number of detail (d) subbands is equal to the decomposition levels and one approximation (a), hence total of (m+1) subbands for m decomposition levels. The frequencies that are most prominent in the original signal will appear as high amplitudes in that region of the DWT signal that includes those particular frequencies. The difference of this transform from the Fourier transform is that the time localization of these frequencies will not be lost. However, the time localization will have a resolution that depends on which level they appear. A toy signal and its different subbands obtained by applying DWT are shown in Fig. 1.

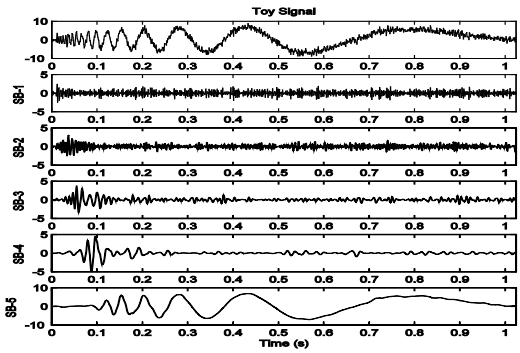


Fig. 1.A toy signal and different subbands obtained by DWT

The reconstructed signal by wavelet synthesis is illustrated in Fig. 2. The energy of the error signal (sample-wise difference) is negligible (in the order of 10^{-11}) and hence it is stated that the perfect reconstruction is possible from DWT based subband decomposition. From the above mentioned explanation, it is clear that the wavelet based subband decomposition is to be used as data adaptive filter bank technique with perfect reconstruction. The climate signal is a non-harmonic time series and hence this type of data adaptive method is suitable for decomposition.

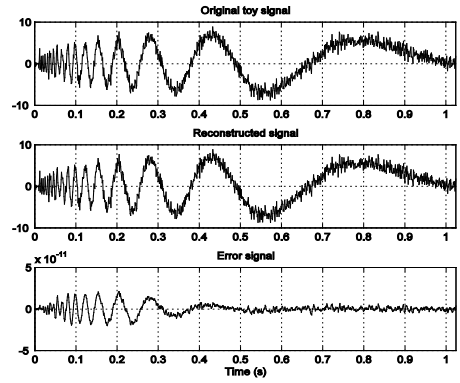


Fig. 2.The original toy signal (top), the reconstructed signal using wavelet synthesis (middle) and the error between those two (bottom)

3Seasonal Cycle Extraction Method

The seasonal (annual) cycle is considered as the low frequency and relatively higher energy trend of climate signals. The trends of the recorded climate signals are detected using the energy distribution of the signalover the individual subbands. The cycle is retrieved here by partial reconstruction ofthe subbands in the wavelet domain. Usually the modes contain a mixture of frequencies and thesemixed modes are much more difficult to interpret. Since we are not interested in intra-seasonalvariations, a minimal amount of pre-smoothing is performed so that single or multiple modescollectively contain the annual cycle. The subbands will generally be ordered from high to lowfrequency. It is important to determine the significanceof the subbands. We should not expect all bands to be significant to separate the seasonal cycle.

The aim of the first step of the analysis is to find low frequency signal from the original climate signal. We decomposed the original signal into subband signals using discrete wavelet transform. The analyzingclimate signal $s(n)$ consists a slowly varying trend superimposed to a fluctuating process $y(n)$, the trend is expected to be captured by subband signals of large indices. A process of de-trending $s(n)$, which corresponds to estimating $\zeta(n)$, may therefore relate to compute the partial, fine-to-coarse, reconstruction

$$\zeta(n) = \sum_{k=1}^{K} d_k(n) \tag{2}$$

where$K$ is the largestsubband index prior the remaining subbands representing signal trend contamination. For the subbands $d_k(n)$; k=1, 2, …,K, a rule of thumb, so the choice of K is based on observation of the evolution of the $\zeta(n)$ energy as a function of a test order k. The optimized k=K is chosen when the energy index departs significantly from the energy of the reference signals [12]. The starting index of the subband to separate the trend i.e. the low frequency components of the climate signal is determined by comparing the subband energy with that of the reference signal. The fractional Gaussian noise (fGn) is used here as the reference signal. There is a subband of climate signal exceeding the upper limit of 95% confidence interval of the corresponding subbands' energies of fGn. That subband is selected as the lower bound of the low frequency component of the climate signal. All the lower order subbands starting from the bound are summed to construct the low frequency trend $y(n)$=$s(n)$- $\zeta(n)$ representing the seasonal cycle.

The fractional Gaussian noise (fGn) is a generalization of ordinary white noise. It is a versatile model of homogeneously spreading broadband noise without any dominant frequency band, is an intrinsically discrete-time process, and may be described as the increment process of fractional Brownian motion (fBm) since fBm is the only self-similar Gaussian process with stationary increments [12]. Consequently the statistical properties of fGn are entirely determined by its second-order structure, which depends solely upon one single scalar parameter, Hurst exponent. The energies of the subbands of fGn are decreased linearly with increasing its order. Such property of fGn is very much applicable to determine the trend of the analyzing signal by comparing the energies of its subbans. The normalized fGn is decomposed only once and then each of the climate signals using DWT. The scaling factor of each signal is reused to obtain the original scale in the wavelet domain. The steps of the proposed algorithm are as follows:

1. Apply DWT based subband decomposition on the fGn. Compute Log energy of individual subband and its upper and lower bound with 95% confidence interval.
2. Apply the same decomposition on any climate signal. Compute the Log energy of its subbands.

Find the subband with energy exceeding the upper limit of 95% confidence interval derived in step 1 say it n^{th}subband. The selected n^{th}subband is the starting index of constructing trend i.e. seasonal cycle. The seasonal cycle is separated by summing up the higher order (lower frequency)subbands starting from n^{th}one of the climate signals. Thus obtained low frequency trend with higher energy represents the seasonal cycle of the respective climate signal. It illustrates the non-linear nature of the climate variable.

4Experimental Results and Discussion

The performance of the proposed seasonal cycle extraction method is evaluated by real climate data collected at different regions in Bangladesh. Only the data collected in Dhaka region from Jan, 1981 to Dec, 2011 are used in this analysis. Two climate variables – daily rainfall and humidity are considered here. All the weather parameters are measured with sufficiently efficient instruments. Both of the climate signals (daily rainfall and humidity) and the fGn are decomposed into 11 subband (10 levels decomposition) signals using DWT with db4 wavelet as illustrated in Fig. 3, 4 and 5 respectively.

According to the proposed algorithm, the energy of each subband of the fGn is computed then the 95% confidence interval (CI) of the energy curve (of fGn) is determined. The energy of each subband of the climate signal (daily rainfall) is compared with the upper bound of the CI which is used here as data adaptive threshold. Find the subband with energy exceeding the upper limit of CI; say it n^{th} band which is selected as the turning index of detecting the trend i.e. the seasonal cycle. The energy based index thresholding is shown in Fig. 6 for the daily rainfall. The 8^{th} subband of the daily rainfall signal is selected as the turning index of trend detection. Similarly, the turning index determination of daily humidity signal is illustrated in Fig. 7 where the 8^{th} one is selected as the starting subband of seasonal cycle.

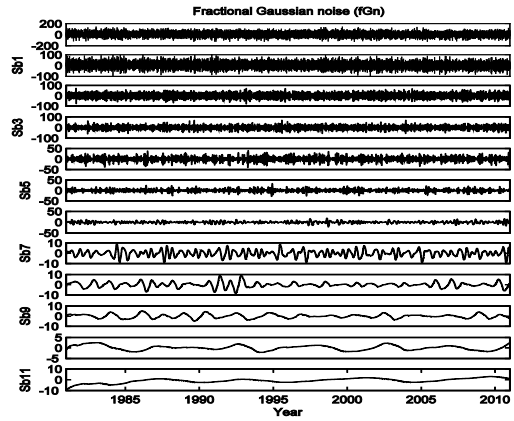


Fig. 3. The fGn and its different subbands obtained by applying DWT

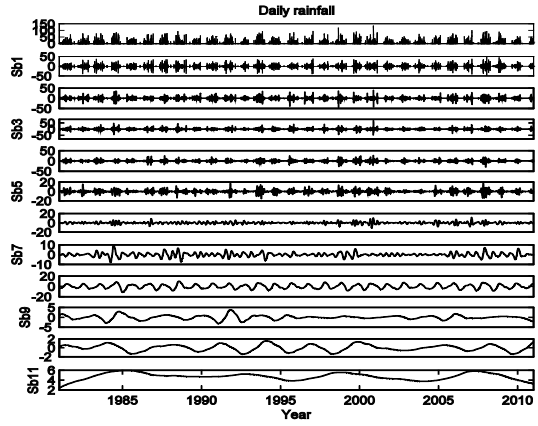


Fig. 4. The daily rainfall data and its different subbands obtained by applying DWT

The seasonal cycle is separated by summing up the higher order subband signals starting from n^{th} one up to the last subband of climate signals. The daily rainfall, the separated seasonal cycle and the residue signals (suppressing the seasonal cycle from the original signal) are illustrated in Fig. 8. The separation of seasonal cycle for daily humidity is shown in Fig. 9.

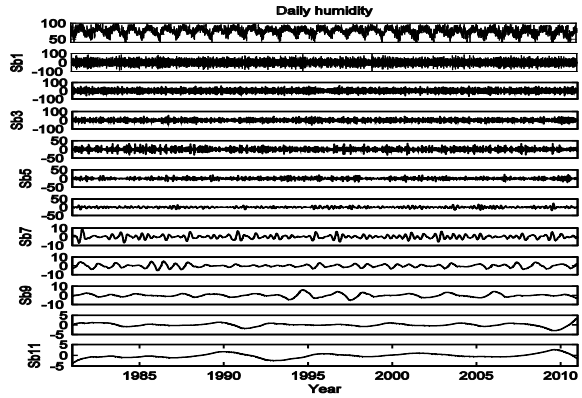


Fig. 5. The daily humidity data and its different subbands obtained by applying DWT

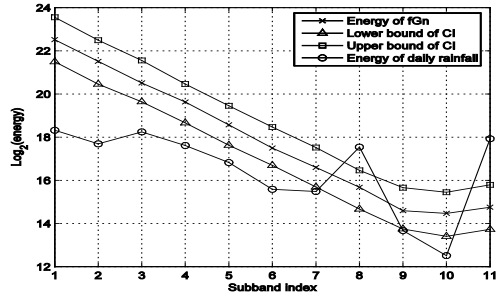


Fig. 6.Selection of starting subband to extract seasonal cycle of daily rainfall. The 8th subband is selected

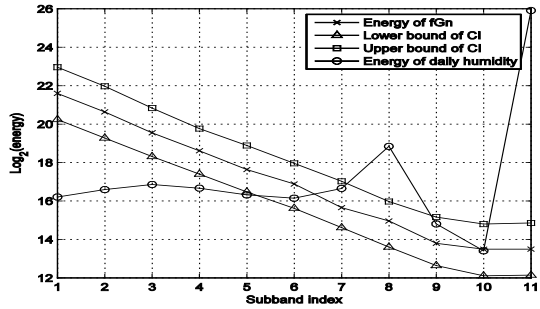


Fig. 7.Selection of starting subband to extract seasonal cycle of daily humidity. The 8th subband is selected

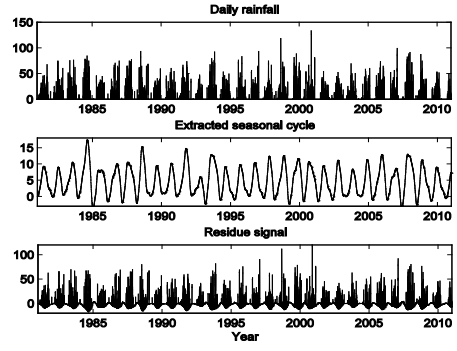


Fig. 8. Seasonal cycle separation results; daily rainfall data (top), extracted seasonal cycle (middle) and residue after suppressing seasonal cycle (bottom)

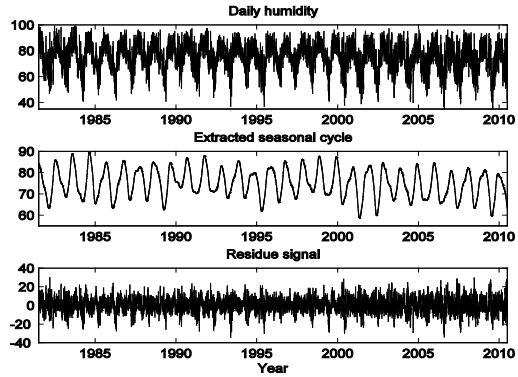


Fig. 9. Seasonal cycle separation results; daily humidity data (top), extracted seasonal cycle (middle) and residue after suppressing seasonal cycle (bottom)

The proposed seasonal cycle extraction method is fully data adaptive. No training is required to determine the threshold. Thus obtained seasonal cycle is carries necessary properties for further processing of the climate variables. Climate signals always represent non-stationary data. It is not possible to assume such data the sum of harmonics and hence Fourier based transformation is not suitable for analysis of climate signals. The DWT is highly data adaptive and efficient for nonlinear and non-stationary time series.

5Conclusions

In this paper wavelet based data adaptive time domain filtering technique is implemented to extract the seasonal cycle of climate signals. The main superiority of this method are to apply the DWT method yielding subband signals based on local properties of the signal, which eliminate the need for spurious harmonics to represent non-linear and non-stationary signals. The DWT is a well known approach to many researchers in climate research. This study plays a vital role for analysis the properties of non-linear and non-stationary daily rainfall and humiditytime series data. The analysis of the correlation among different subbanbs of the climate signals is the future extension of this study.

References

1.Dairaku, K., Emori, S., Nozawa, T., Yamazaki, N., Hara, M.and Kawase, H., "Hydrological change under the global warming in Asia with a regional climate model nested in a general circulation model," in *Proceedings of the 3rd International Workshop on Monsoons (IWM '04)*, vol. 56, 2004.

2. Bates, B. C., Charles, S. P.and Hughes, J. P., "Stochastic downscaling of numerical climate modelsimulations," *Environmental Modelling and Software*, vol. 13, no. 3-4, pp. 325–331, 1998.

3.Rajagopalan, B., Lall, U., and Cane, M. A., "Anomalous ENSO occurrences: an alternate view," *Journalof Climate*, vol. 10, no. 9, pp. 2351–2357, 1997.

4.Rajagopalan, B., Lall, U., and Cane, M. A., "Comment on reply to the comments of Trenberth and Hurrell," *Bulletin of American Meteorological Society*, vol. 80, pp. 2724–2726, 1999.

5.Harrison, D. E.and Larkin, N. K., "Darwin sea level pressure, 1876–1996: evidence for climate change?" *Geophysical Research Letters*, vol. 24, no. 14, pp. 1779–1782, 1997.

6. Mpelasoka, F. S., Mullan, A. B.and Heerdegen, R. G., "New Zealand climate change informationderived by multivariate statistical and artificial neural networks approaches," *International Journalof Climatology*, vol. 21, no. 11, pp. 1415–1433, 2001.

7. Mak, M., "Orthogonal wavelet analysis: Inter-annual variability in the sea surface temperature", Bull. Amer. Meteor. Soc., 76, 2179-2186, 1995.

8. Oh, H. S., Ammann, C. M., P. Naveau, Nychka, D.and Otto-Bliesner, B. L., "Multi-resolution time series analysis applied to solar irradiance and climate reconstructions", *Journal of Atmos. Solar-Terr. Physics*. 65, 191-201, 2003.

9. Broughton S. A.and Bryan, K. M., "Discrete Fourier Analysis and Wavelets: Applications to Signal and Image Processing", Wiley, 2008.

10. Mallat, S., "A Wavelet Tour of Signal Processing: Third Edition", 2008.

11. Strang G.and Nquyen, T., "Wavelets and Filter Banks", Willesley, Cambridge Press, Oct. 1996.

12. Flandrin, P., Rilling G.and Goncalves, P., "Empirical mode decomposition as a filter bank", *IEEE signal processing letters*, Vol. 11, No. 2, pp: 112-114, 2004.

Abstract for Oral Presentation

Entropy-Based Portfolio Models: Practical Issues

Yasaman Izadparast Shirazi[1], Md. Sabiruzzaman[1] and Nor Aishah Hamzah[1]
[1]Institute of Mathematical Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia
Email: yasi13132000@yahoo.com

Abstract: Entropy is a nonparametric alternative of variance and has been used as a measure of risk in portfolio analysis. Although attractive features of entropy are pronounced in a number of papers, it is still not popular among practitioners. In this paper, we compare entropy-based portfolio models and benchmark against Markowitz mean-variance portfolio. The computation of entropy from a given set of data is discussed through with illustration. We discuss density estimation and entropy calculation from real data in R environment. We find that entropy and variance may give different interpretation of risk and hence, the portfolio based on entropy may differ from that based on variance. We propose a natural extension of the mean entropy portfolio to make it more general and diversified. This new model performs equivalently to the mean-entropy portfolio when applied to real and simulated data, and offers higher return if no constraint is set for the desired return; also it is found as the most diversified portfolio model.

Key Words: Entropy, Kernel density estimation, Portfolio optimization, Diversification.

Selection of Optimal Follow-up Interval in Multi-state Model

Kishor Kumar Das[1], Md. Asaduzzaman[2] and Mahbub A.H.M. Latif[2]
[1]Centre for Communicable Diseases, icddr,b, e-mail: kdas@icddrb.org
[2]Institute of Statistical Research and Training, University of Dhaka, e-mail: asad@isrt.ac.bd

Abstract. In multi-state model interval censored data are widely available due to discrete follow-up of observation scheme. Interval of follow-up is arbitrarily chosen. In this paper an algorithm is presented to choose optimal interval using simulation based on magnitude of bias.

Some key words: Discrete follow-up; Continuous follow-up; Multi-state model.

Statistical learning algorithm for automatic Bat identification from motion sensor camera images

Jaynal Abedin*[1], M. A. Yushuf Sharker[1], Kishor Kumar Das[1], Shawkat Ali[2], Asfaqur Rahman[3], SalahUddin Khan[1], Emily Gurley[1], Stephen P. Luby[1]
[1]icddr,b, Bangladesh
[2]Central Queensland Universities, Australia
[3]Intelligent Sensing and Systems Laboratory, CSIRO, Hobart, Australia

Abstract

Bat, the natural reservoir of Nipah Virus (NiV), contaminates date palm sap while harvesting by licking the shaved part of tree. Raw sap consumption is the potential risk factor of NiV infection to human. To research onbat contaminating date palm sap at night, researchers from icddr,b used motion censored cameras which can capture 60-14000 photos per night per tree. They extract data on time and frequency

of Bat visit manually by observing each of the photos which is pain staking, time consuming and costly. Our objective was to develop a computer algorithm that will extract the data of interest from those images.

We developed an image processing rule and extracted features from the processed image. In pilot phase, we selected a random sample of images of size 500 from a single tree and labeled those photos as presence/absence and we extracted 2 features of each images. After that we developed a statistical learning model using double bagged logistic regression. We test our algorithm on a set of test images from that tree. In second phase, we repeat the procedure based on images on multiple trees of 42 different camera-nights. After that we tested our algorithm on full data set. .

In pilot phase our proposed algorithm identified the presence or absence of bats with 90% accuracy. In the testing phase for all data, our algorithm provided 78.2% accurate prediction of presence and absence of bat.

In pilot phase, the result seems promising regarding the reduction of time and cost of data extraction from images but, we need to explore further, the reasons of reduction of efficiency in generalizing the algorithms efficiency in broader perspective.

The θ-Centralizers of Semi-prime Gamma Rings

M. F. Hoque[1] and A. C. Paul[2]

[1]Department of Mathematics, Pabna Science and Technology University, Pabna-6600, Bangladesh

E-mail: fazlul math@yahoo.co.in

[2]Department of Mathematics, Rajshahi University, Rajshahi-6205, Bangladesh

Abstract

The main Results: Let M be a 2-torsion free semi prime Γ-ring satisfying a certain assumption and θ be an endomorphism of M. Let $T : M \to M$ be an additive mapping such that $T(x\alpha y\beta x) = \theta(x)\alpha T(y)\beta\theta(x)$ holds for all $x, y \in M$ and $\alpha, \beta \in \Gamma$. Then we prove that T is a θ-centralizer. We also show that T is a θ-centralizer if M contains a multiplicative identity 1.

Keywords: semi-prime Γ-ring, left centralizer, centralizer, Jordan centralizer, left θ-centralizer, θ-centralizer, Jordan θ-centralizer.

Abstract for Poster Display

Mother's Health Seeking Behavior: A Qualitative Study on after Childbirth

M. Monaemul Islam Sizear[1]

[1] District Coordinator of Research, NGO Forum for Public Health, and have completed MSS at 2011 from Anthropology Department, University of Rajshahi.

ABSTRACT

Women's health after childbirth is an area that is under-investigated. Although many studies have been carried out about health seeking behavior of women during pregnancy, none have focused on after childbirth, particularly regarding middle class women of urban areas. In this study an attempt has been made to explore the health seeking behavior of middle class women after childbirth in Rajshahi City Corporation. Qualitative method and techniques were used to explore health seeking behavior. All the respondents of the study were selected through the purposive sampling from the study area and semi structured interview, in-depth interview were conducted with one FGD. It is found that, the perception of health care after childbirth is not satisfactory and given little attention although they acknowledged the importance. In case of after childbirth, some physical health problems are common of women after childbirth and it may continue one-two months. Again, few physical symptoms have been suffered for long time in some cases. The health seeking behavior of women is influenced by mainly senior family members and it is seen that, in post-partum complications, most of the cases women don't go to doctor for treatment rather they taken advice from experienced women. In this decision making process and health seeking behavior of women after childbirth, some influential socio-cultural factors played the vital role which is upheld in this study.

KEYWORDS: health seeking behavior, childbirth, maternal health, postpartum health.

Surviving With Seasonal Extremeness: A Case Study of Slum Dwellers of Rajshahi City

Papia Sultana[1], Mahafuzr Rahman[2] and A. S. M. Enamul Kabir[2]

[1]Associate Professor, Department of Statistics, University of Rajshahi, Bangladesh
[2]MSc Thesis Student, Department of Statistics, University of Rajshahi, Bangladesh

ABSTRACT

Introduction: All major urban centers in Bangladesh have slums and squatter settlements, the largest concentrations being in Dhaka, followed by Chittagong, Khulna and Rajshahi. In Winter and in Rainy season, the weather becomes extreme at Rajshahi making people more difficult, even more for slum dwellers. We aimed to evaluate the extremeness of the seasonal variation to various seasonal diseases.

Key Words: Descriptive statistics, Chi-square test, Logistic Regression

Methods: Our data was from a pilot study of 250 respondents of a survey on "Health Status of Slum Dwellers in Rajshahi City". Original sampling method is two-stage PPS. First stage units are 35 wards of Rajshahi city and second units are the slum dwellers (per household basis) of the wards. Study area for the pilot survey was ward number 28 and 30. Study population was adult slum dwellers. Data collection period was Nov 2010 to Jan 2011.Descriptive statistics (mean with standard deviation, median with IQR or percentage where appropriate) were primarily observed. Using cross table technique, chi-square test (Pearson or LR whichever applicable) was used next to find association of various covariates to diseases. Logistic model was fitted to main outcome variable (diseases) for advanced analysis.

Results: We found that mean age of the respondent was 30.88 yrs with standard deviation10.68. Of the respondent female was 56%; maximum of the respondents were illiterate (56.08%) following with primary level 34.46% and with secondary level 9.46%; maximum of them are Muslim (97.30%). Some of their citable professions are daily

labor (21.62%), pulling rickshaw (10.81%) small business (7.43%) etc. Mean family income was tk 90493 per year with standard deviation (sd) 67068; on an average family savings per year was tk 1182 with sd 1705. Among the respondents, 93% manage their worm cloths by themselves; 95% store their worm cloths after winter; among the common communicable diseases, 88% suffer from Cold, 56% suffer from Flue, 14% suffer from Asthma in Winter, whereas 38% suffer from cold, 37% suffer flue, 12% suffer asthma in Rainy season. Results suggested that none of the covariates had significant effect on Cold in Winter, but in Rainy season, Profession had significant association with it (p-value=0.019). It was also found that respondent of profession daily labor (OR=0.25, 95%CI=0.06, 1.03) and pulling rickshaw (OR=0.57, 95%CI=0.13, 2.58) were less likely to suffer from it and of profession small business (OR=3.70, 95% CI=0.72, 18.94) was more likely than other professions. For the slum dwellers, profession and family status had significant effect on suffering from Flue in Winter season which also revealed that daily laborer (OR=17.41, 95%CI=3.04, 99.71), small businessmen (OR=2.63, 95%CI=0.36,19.17) and rickshaw puller (OR=6.55, 95%CI=1.05, 41.02) are more likely than other professions and respondents with unique family (OR=0.17, 95%CI=0.05,0.56) are less likely to suffer from it than combined family. But in Rainy season, it was found that only family status had significant effect on it and respondents with unique family (OR=0.25, 95%CI=0.08,0.76) are less likely to suffer from it than combined family. However, none of the covariates were found to have significant effect to Asthma in both Winter and Rainy season.

Conclusions: By means of feasible action plans focused on a broad urban development vision that will lead to lasting and meaningful improvements in the health and lives of slum dwellers.

Factors Associated with Contraceptive Use among Adolescent of Bangladesh

Md. Tahidur Rahman[1], Md. Nurul Islam[2], Md. Golam Hossain[2], Md. Rezaul Karim[3] and Md. Mizanur Rahman[3]

[1]Statistical Officer, Shafi Consultancy Bangladesh, E-mail: tohid_stat@yahoo.com

[2]Professor, Department of Statistics, University of Rajshahi, Bangladesh.

[3]Research Fellow, Department of Statistics, University of Rajshahi.

Abstract. Adolescence is a transitional period from childhood to adulthood characterized by significant physiological, psychological and social changes. In this paper we especially considered for the married female adolescents, as they have to enter early into marital life that pushes them to bear the consequence of childbearing. Although adolescents have knowledge on contraceptives, but their use of contraceptives is low. Quantitative data on 1,348 married female adolescents revealed that almost cent per cent of the married adolescent are aware of at least one contraceptive method. But the current use rate is low (27.0%). There are many socio-economic and demographic factors those are significantly associated with contraceptive use. In view of multiple logistic regression analysis, it is evident that the explanatory variables such as respondent's education, husbands lives in house, type of place of residence, region, husband's occupation, husband's education, age of respondents and access to mass media are important on the ever use of contraception. On the other hand, age of respondents and marital duration are important on the current use of contraception. Multiple logistic regression analysis estimates that the current age of respondent and marital duration have direct influences on contraceptive use.

Key words: Contraception, Adolescent, Knowledge and Bangladesh**.**

Female sex workers HIV/AIDS harm reduction for awareness: A review in Rajshahi City Corporation, Bangladesh

Sheikh Moin Uddin[1], Md. Ashraful Islam[2], Md. Nurul Islam[2] and Md. Golam Hossain[2]

[1]Population Services and Training Center (PSTC), Bangladesh

[2]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh

Abstract. Unsafe female sex trade and their clients who dominate unsafe sex are the major causes to spread out HIV/AIDS in Bangladesh. In under developing country like Bangladesh may be some significant factors are engaged to increase the number of HIV/AIDS infected people day by day. The purpose of the present study was to

know unsafe female sex trades HIV/AIDS harm reduction for awareness in Rajshahi City Corporation, Bangladesh. Primary data were collected from 200 female sex workers in Rajshahi City Corporation from February to September 2012. Samples were selected by simple random sampling. The age range of the female sex workers was 16 to 41 years with average age 24.52 ± 6.26 years, and 84.4% were married. More than 88% sex workers agreed that client dominate sex without condom, this factor was significantly associated with female marital status ($p<0.01$), and 62% females sexual contacted with client without condom for money and forced by client. More than 32% sex workers did not test HIV/VCT and this factor was significantly associated with education level ($p<0.01$), age ($p<0.01$) and economical condition ($p<0.05$). 89.5% respondents did not interest to ask their new client about HIV/VCT test and this variable was associated with sex workers residence ($p<0.05$), age ($p<0.05$) and economical condition ($p<0.01$). Education, age and economical condition are most important factors for making awareness about HIV/AIDS for female sex workers.

Keywords: Female sex worker (FSW), STI/STD, HIV/AIDS, Awareness, Female sex trade.

Projected population with non-communicable diseases in the perspective of Bangladesh

Aziza Sultana Rosy Sarkar, Md. Aminul Hoque and Md. Nurul Islam

Department of Statistics, Faculty of Science, University of Rajshahi, Rajshahi-6205, Bangladesh.

Abstract

Objectives: The projection of population affected with non-communicable diseases in Bangladesh and setting the targets to control non-communicable diseases as a future plan in health sectors.

Methods: Exponential model and polynomial regression model are used to project populations.

Results: Among the age groups, circulatory system related diseases and neoplasm greatly affect the age group 45-59 and above for male. More specifically, circulatory system related diseases have the highest percentage (34.01%) for the age group 70-79 and neoplasm contributes the highest percentage (34.38%) for the age group 60-69 for male.

Among the age groups like the male person circulatory system related diseases, respiratory related diseases and neoplasm greatly affect the age group 45-59 and above for female. More specifically, circulatory system related diseases have the highest percentage (38.65%) for the age group 70-79, neoplasm contributes the highest percentage (29.41%) for the age group 45-59 and the highest percentage is 32.69% for respiratory related diseases in the age group 60-69.

Conclusions: It is recognized that a huge number of population will die because of non-communicable diseases. The number of person rapidly increases year by year at a large scale. Among non-communicable disease; mortality population due to circulatory system (stroke, ischemic heart disease and hypertensive disease) is most prominent in Bangladesh. The second major cause of death is neoplasm for the national population.

Non-communicable diseases say, circulatory related disease was significantly higher as compared to other non-communicable diseases.

In case of national projected diseased population, the transition period is 2016 year. The diseased population shows decreasing trend before 2016 and after that year the diseased population gives increasing trend. Because the diseased population is a balance trend between national projected non-communicable male and national others population except non-communicable population.

Key Words: Non-Communicable diseases (NCDs), exponential model, mortality rates, diseased population and population projection.

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment***, 21-24 December 2012.

ISBN: **978-984-33-5876-9**

Multiple logistic regression analysis of socio-economic and demographic factors influencing early childbearing mother's health in Bangladesh

Md. Ashraful Islam[1], Md. Nurul Islam[2] and Md. Golam Hossain[2]

[1]Ph.D. Fellow, Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh
E-mail: sasraf29@yahoo.com
[2]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh

Abstract. Early childbearing is an important indicator for women's health. Body mass index (BMI) is a good indicator of nutritional status in a population. In underdeveloped countries like Bangladesh, this indicator provides a method that can assist intervention to help eradicate many preventable diseases. The aim of the present study was to find the effect of socio-demographic factors on early childbearing mother's health. Data was extracted from Bangladesh Demographic and Health Survey (BDHS)-2007. The age range of the sample was 15 to 24 years, with an average age of 20.29 ± 2.54 years. More than 33% early childbearing mothers have been suffering from chronic energy deficiency (CED), among them 35.4% from rural and 26.9% from urban. The coefficients and odds ratio of logistic regression analysis demonstrated that early childbearing mothers who were from rural areas, illiterate, employed, unemployed partner, poorest, non-caesarian, delivered at home, earlier age at first marriage, having two or more children were at higher risk for getting chronic energy deficiency.

Keywords: BMI, CED, Logistic Regression, Undernutrition, Early childbearing mother

Introduction

Early childbearing mothers are the young adolescent (teenage) women who get baby before she is physically and/ or mentally not ready to give birth. Usefully woman becomes mother before 20 years old is called early child bearing mother. Early childbearing is an important indicator for women health[1]. Early childbearing mother is at higher risk for poor prenatal outcomes such as gestational diabetes, gestational hypertension and preterm deliveries than the general population[2]. Many researches [2-5] have captured teenage pregnancy and early childbearing both developed and developing countries in recent decades due to the socioeconomic and health consequences for both mother and their child. In the recent past decades, the patterns of early marriage and early childbearing are still persistent in Bangladesh despite substantial development in Human Development Indicators (HDI). Early childbearing is often associated with higher than average mortality rates. In Bangladesh mortality rate for children born to adolescent mothers is 10.4%[5]. The World Health Organization estimated that the risk of death following pregnancy is twice as great for women between 15 and 19 years than for those between the ages of 20 and 24[6]. The maternal mortality rate can be up to five times higher for girls aged between 10 and 14 than for women of about twenty years of age[6]. Adolescent (teenage) childbearing is widely prevalent in Bangladesh: 76% of the first-born children were born to women before their 20th birthday[5].

Body mass index (BMI) is the best instrument for the assessment of health status of a given population[7]. A BMI value of over 30 kg/m^2 has been shown to be a risk factor for hypertension, heart disease, diabetes mellitus, cardiovascular disease, gall bladder disease and various types of cancer. On the other hand, a low BMI (underweight BMI $\leq$18.5 kg/m^2) has been associated with a higher risk of hip fracture in women[8]. Low birth weight and higher mortality rate has also been associated with a low BMI in pregnant mothers[9].

Many researchers have investigated the relationship between health status/nutritional status and mortality[9], socioeconomic and demographic factors[10-12] of reproductive women in Bangladesh. Among the reproductive women, special attention is needed to be paid to early childbearing mothers because they face a lot of health and social problems compare to mothers who give child birth at a later stage of life. A healthy mother puts a potential influence on the family and their contribution to the nation's workforce and productivity. It is important to detect the factors which are related to undernutritional/health problems (underweight) of early childbearing mothers in Bangladesh. To the best of our knowledge the study on health status of early childbearing mothers are poorly documented.

The purpose of the present study was to find the effect of socio-economic and demographic factors on the health status of early childbearing mothers in Bangladesh.

Materials and methods

Materials: The total simple used in the present study consisted of 1908 currently non-pregnant young mother (age 15-24 years). This cross sectional data extracted from a sample of 10,996 ever married Bangladeshi women (age 15-49 years) were collected by the 2007 BDHS. The data set was checked for outliers/abnormal by the present authors using statistical techniques[13], because these abnormal points can affect the interpretation of results. Some missing values were also detected, and these cases were excluded. After removing outliers, cases with incomplete data, and excluding women who were above 24 years old, currently pregnant, the data set was reduced to 1908 for the analysis in the present study. The sampling technique, survey design, survey instruments, measuring system and quality control have been described elsewhere[14]. Body mass index was defined and calculated as the ratio of weight in kilograms to height in meters squared.

Methods: The BMI was classified according to most widely used categories of BMI for adults. These were: underweight/undernutrition (BMI≤18.5 kg/m^2), normal weight (18.5 < BMI <25 kg/m^2), overweight 25≤BMI < 30 kg/m^2) and obese (BMI≥30 kg/m^2)[15]. The subjects were also classified on the basis of chronic energy deficiency (CED) grades as follows: (i) CED grade III (sever thinness) (BMI < 16.0 kg/m^2), (ii) CED grade II (moderate thinness) (16.0 ≤ BMI <17.0 kg/m^2), (iii) CED grade I (mild thinness) (17.0 ≤ BMI < 18.5 kg/m^2)[4].

Multiple logistic regression analysis: Multiple logistic regression was used to examine the relative importance of socio-demographic factors on early childbearing mothers' health. In this model, body size (BMI) was considered as a dependent variable coded as 0= Normal and overweight (BMI≥18.51) and 1= Underweight (BMI≤18.50/malnutional). The underlying multiple logistic regression models corresponding to each variable are:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} \quad \ldots [1]$$

where, p = the probability of underweight (BMI≤18.50) (coded 1)

1-p = the probability of normal and overweight (BMI≥18.51) (coded 0)
X_1= type of place of residence (coded; urban=0, rural=1)
X_2 = respondent's (woman) educational level (coded; no education=0, school education=1, higher education=2)
X_3= partner's (husband) educational level (coded; no education=0, school education=1, higher education=2)
X_4= respondent's occupation (coded; housewife=0, hard labor=1)
X_5 = partner's occupation (coded; employed=0, farmer/worker=1)
X_6 = wealth index (coded; poorest=1, poorer=2, middle=3, richer=4, richest=5)
X_7= delivery system (coded; non-caesarian=0, caesarian=1)

X_8 = place of delivery (coded; hospital/clinic=0, home=1)
X_9= age at first marriage
X_{10}= respondent age at first birth
X_{11}= total number of children ever born

β_0= intercept term, and β_1, β_2,…,β_{11} are unknown coefficients. Multicollinearity problem among the predictor variables were checked by standard error (SE). If the magnitude of the SE lies between .001 and .05, suggested that there is no evidence of Multicollinearity problem[16].

All the statistical analysis should carried out using Statistical Package for Social Scientists (SPSS version 17.0) software.

Results

A total of 1908 early childbearing currently non-pregnant young mothers were analyzed in the current study. The age of subjects varied from 15 to 24 years, with a mean age of 20.29 ± 2.54 years (95% CI: 20.17–20.40). The range of age of first birth was 12 to 19 years with mean age 16.46 ± 1.67 years (95% CI: 16.38-16.53). The average weight was 44.96± 7.09 kg (95% CI: 44.64-45.27) with range 27.50 kg to 88.00 kg. The mean height was150.32 ± 5.42 cm (95% CI: 150.08-150.56) with range 129.60 to 168.20 cm. BMI varied from 11.95 kg/m^2 to 37.79 kg/m^2, with mean of 19.86 ± 2.70 kg/m^2 (95% CI: 19.74-19.98) (Table1).

Table 1. Descriptive statistics for age of first birth, height, weight and BMI of early childbearing young mothers in Bangladesh

Variable	Mean	SD	SE	95% CI for mean		Minimum	Maximum
				Lower	Upper		
Age	20.29	2.54	0.06	20.17	20.40	15	24
Age of first birth	16.46	1.67	0.04	16.38	16.53	12	19
Weight(kg)	44.96	7.09	0.16	44.64	45.27	27.50	88.00
Height(cm)	150.32	5.42	0.12	150.08	150.56	129.60	168.20
BMI(kg/m^2)	19.86	2.70	0.06	19.74	19.98	11.95	37.79

Table 2 shows the frequency distribution of body size (BMI category). More than 60% women in the current study had normal weight (61.6%), while 33.3% were undernourished (underweight). Some women were overweight (4.3%) and a few (0.8%) participants were obese.

Table 2. Frequency distribution of body size (BMI category) of early childbearing mothers

BMI category	N (%)
Underweight (BMI ≤ 18.5 km/m^2)	636 (33.3%)
Normal weight (18.5 < BMI < 25 km/m^2)	1175 (61.6%)
Overweight (25≤BMI<30 km/m^2)	82 (4.3%)
Obese (BMI ≥30 km/m^2)	15 (0.8%)

The underweight and obese of the study population are depicted graphically by residence. The numbers (percentages) of underweight rural mothers were more than that of urban mothers, while the number of obese were higher in urban than in rural environment (Fig.1).

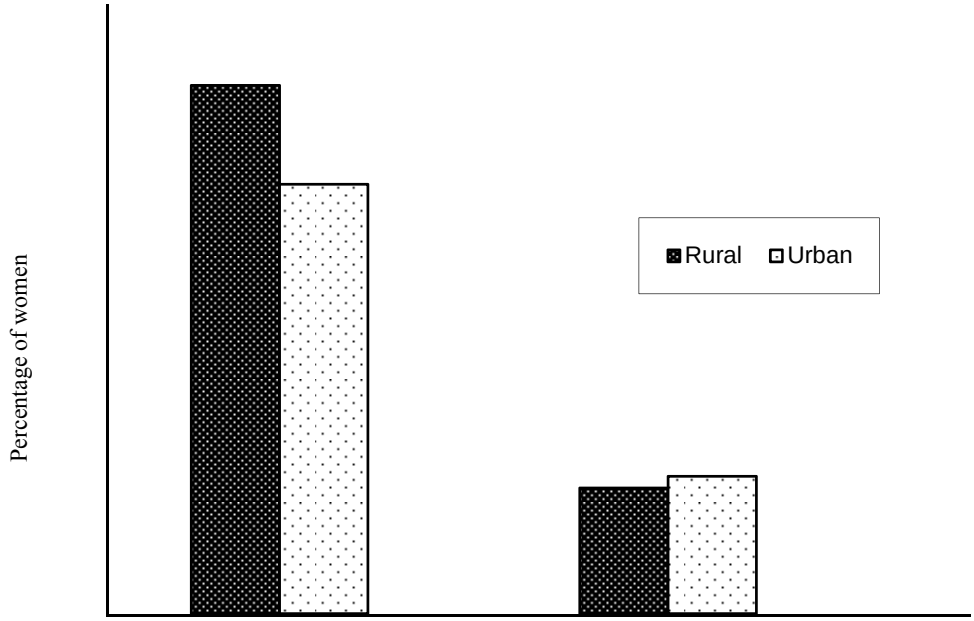


Fig1: Difference between urban and rural in the percentage of underweight and obese mothers

Frequency distribution of chronic energy deficiency (CED) showed that about 33.3% early childbearing mothers were suffering from chronic energy deficiency (BMI ≤ 18.5) (Table2). Among them 10.69% mothers had sever thinness (CED grade III), while 22.01% had moderate thinness (CED grade II) and 67.30 % had mild thinness (CED grade I) (Table 3).

Table 3. Frequency distribution of chronic energy deficiency (CED) mothers among CED grades

BMI category	N (%)
CED grade III (sever thinness) (BMI< 16.0 kg/m^2)	68 (10.69%)
CED grade II (moderate thinness) (16.0 kg/m^2≤BMI<17.0 kg/m^2)	140 (22.01%)
CED grade I (mild thinness) (17.0 kg/m^2≤BMI≤18.5kg/m^2)	428 (67.30%)

The logistic regression coefficients and odds ratios demonstrated that the participants who came from rural environment, had a chance to get undernutrition 41.6% [(1.416-1)*100] ($p<0.001$) higher than urban mothers. Illiterate mothers had a chance to get undernutrition 0.661 and 0.328 times higher than secondary educated ($p<0.01$) and higher educated ($p<0.01$) mothers, respectively. Similar result was found for partner's/husband's educational level (Table 4). These results suggest that both women's and her husbands' educations are important factors for women health status. Mothers who were involved with hard labor had [(1.298-1.00) x 100] = 29.8% higher risk of having undernourished than mothers who were housewife ($p<0.05$). On the other hand, the women whose partners were unemployed had [(1.466-1.00) x 100] = 46.6% high risk of getting undernutrition than mothers whose partners were employed ($p<0.01$). Regarding economic condition, poorest mothers had a chance to get undernutrition 0.703 ($p<0.05$), 0.663 ($p<0.05$), 0.615 ($p<0.05$), and 0.363 ($p<0.001$) times higher than poor, middle, rich and richest mothers, respectively. These results suggest that wealth index can play an important role among young mothers towards improvement of health status. The risk of becomes undernourished of non-caesarian mothers

were 0.335 times higher than caesarian mothers. Also, mother who delivered at home had a chance to get undernutrition [(2.219-2.00) x 100] = 21.9% higher than mothers who delivered at hospital/clinic. Age at first marriage and age at first birth was negatively related to underweight and both were statistically significant. Young mothers had risk for getting undernutrition [(1.057-1.00) x100] = 5.7% higher at every single birth compared to the previous birth.

Table 4: Estimates of the odds ratios for selected socio-economic and demographic characteristics through Logistic regression analysis

Variable	Coefficient	SE	Wald	p-value	Odds Ratio	95% CI for odds ratio	
						Lower	Upper
Place of Residence	0.346	0.107	10.401	0.001	1.414	1.146	1.745
Respondent education level Secondary Vs	---	---	23.495	0.000	---	---	---
No educated	-0.414	0.129	10.342	0.001	0.661	0.514	0.851
Higher Vs No educated	-1.114	0.371	9.040	0.003	0.328	0.159	0.678
Partner education level	---	---	30.446	0.000	---	---	---
Secondary Vs No educated	-0.359	0.107	11.155	0.001	0.699	0.566	0.862
Higher Vs No educated	-0.948	0.220	18.554	0.000	0.387	0.252	0.596
Respondent's Occupation	0.261	0.119	4.787	0.029	1.298	1.028	1.639
Partner's Occupation	0.382	0.098	15.306	0.000	1.466	1.210	1.775
Wealth Index	---	---	35.633	0.000	---	---	---
Poor Vs Poorest	-0.352	0.150	5.515	0.019	0.703	0.524	0.943
Middle Vs Poorest	-0.270	0.151	3.212	0.043	0.663	0.568	0.826
Rich Vs Poorest	-0.485	0.156	9.727	0.002	0.615	0.454	0.835
Richest Vs Poorest	-1.013	0.175	33.364	0.000	0.363	0.257	0.512
Delivery System	-1.093	0.248	19.424	0.000	0.335	0.206	0.545
Place of Delivery	0.756	0.152	24.708	0.000	2.129	1.580	2.868
Age at First Marriage	-0.071	0.022	10.522	0.001	0.932	0.893	0.972
Age at First Birth	-0.044	0.022	4.002	0.045	0.957	0.917	0.999
Children Ever Born	0.055	0.020	7.955	0.005	1.057	1.017	1.098

Discussion and Conclusions

The data used in this study, gathered by the 2007 BDHS, are nationally representative, covering both urban and rural areas. The present study demonstrated that early childbearing are still common and deeply entrenched among Bangladeshi women. Among Asian countries the incidence of early childbearing is highest in Bangladesh[3]. Bangladeshi women become mother at their very early ages with the large majority of women started bearing children before they reach at the age of twenty[17]. In the present study we looked the health status of early childbearing mother, age ranged of the subject 15 to 24 years. Previous studies in Bangladesh examined the factors affecting adolescent motherhood in Bangladesh using the 2007 Bangladesh Demographic and Health Survey data [3]reported that women's education, husband's education, place of residence, ever use of contraceptive method, religion, wealth and region were important determinants of adolescent motherhood in Bangladesh. Other two previous studies [11-12] examined the association of BMI with age, mortality, level of education, wealth index and other social variables. More recently[10] examined the association of BMI with socio-demographic factors and also

showed the trend in BMI over last three decades. The above studies, authors examined women health impact using the indicator BMI but their sample aged 15-49 years, not especially early childbearing young mothers.

The present study demonstrated that the mean BMI of Bangladeshi early childbearing young mother was 19.86 kg/m^2. More than half of the mothers (61.60%) were of normal weight. Undernourished mothers constituted 33.3%, while overweight mother constituted 4.30%. Only 0.8% was considered obese. This information is consistent with other studies on Bangladeshi women. A study on Bangladeshi women living in an urban area reported that 15.7% were overweight and 3.9% were obese[11], while another study on women living in the slum area of Dhaka reported that 54% of women were underweight[18]. More recently[10] investigated on BMI of Bangladeshi reproductive women (age 15-49) and found that 57.73% women were of normal weight, while 28.66% underweight, 11.45% overweight and 2.16% obese of their population. A relatively similar pattern was also observed in a large population study in neighboring India, where 56.9% of married women were reported to be of normal weight, 31.2% were underweight, 9.4% were overweight and 2.6% were obese[4].

The findings in the current study suggest that adolescent marriage is a common phenomenon in Bangladesh. Early childbearing is directly associated with early marriage. Poor economic conditions, illiteracy, early age at marriage, early age at first delivery, insufficient medical facilities lack of suitable work and tendency to getting more children in rural areas are the main causes of being underweight of early childbearing young mothers in Bangladesh.

Government and non-government origination should take care of women about their health especially in rural area and they may take policies;

- To make aware the general population for following the ordinance of legal age at marriage of Bangladesh Government should be properly implemented.
- Adolescents and their guardians should be made more aware of the adverse health outcome, social and economic consequences of early marriage and early childbearing.
- Develop a policy that will be helped to reduce the literacy rate.
- Increase medical facilities in rural area.
- To make aware the married people about family planning.
- Take policy to remove unemployment from Bangladesh.

Other possible influences on the young mothers health includes smoking habits, weight goals, weight-loss methods, body-shape perceptions, eating attitudes and behaviors, self-concept and physical activity and age at menarche. Clearly, more research is required.

Acknowledgments: We would like to thank the Bangladesh Demographic and Health Survey (BDHS) for providing nationally representative based data collected 2007.

Conflict of interest: All authors declared that there were no conflicts of interests in relation to this study.

Reference

1. Luker K.: *Dubious Conceptions: The Politics of Teenage Pregnancy*, Cambridge, Mass: Harvard University Press, 1996
2. Palacios J. and Kennedy H.P.: 'Reflections of Native American teen mothers',*JOGNN*, 39, 425–434, 2011
3. Kamal S.M. M.: 'Adolescent motherhood in Bangladesh: Evidence from 2007 BDHS data', *Can. Stu.Pop*, 39, 63–82, 2012
4. Bharati S., Pal M., Bhattacharya B.N. and Bharati P.: 'Prevalence and causes of chronic energy deficiency and obesity in Indian women', *Hum Biol.*, 79, 395-412, 2007
5. Maitra P. and Pal S.: 'Early Childbirth, Health Inputs and Child Mortality: Recent Evidence from Bangladesh', *IZA Discussion Paper Series*, No. 2841, 2007
6. Locoh T.: 'Early Marriage and Motherhood in Sub-Saharan Africa', *AJOL*, 3-4, 31-42, 2006
7. FAO.: 'The State of the World Population', *Rome, FAO*, 1993
8. Gnudi S., Sitta E. and Lisi L.: 'Relationship of body mass index with main limb fragility fractures in postmenopausal women', *J Bone Miner Metab.*, 27, 479-484, 2009
9. Hosegood V. and Campbell O.M.: 'Body mass index, height, weight, arm circumference, and mortality in rural Bangladeshi women: a 19-y longitudinal study', *Am J Clin Nutr.*, 77, 341-347, 2003
10. Hossain M.G., Bharati P., Aik S., Lestrel P.E., Abeer A. and Kamarul T.: 'Body mass index of married Bangladeshi women: trends and association with socio-demographic factors', *J Biosoc Sci.*, 44, 385-99, 2012
11. Khan M.M. and Kraemer A.: 'Factors associated with being underweight, overweight and obese among ever-married non-pregnant urban women in Bangladesh', *Singapore Med J.*, 50, 804-813, 2009
12. Shafique S., Akhter N., Stallkamp G., de Pee S., Panagides D. and Bloem M.W.: 'Trends of urder-and overweight among rural and urban poor women indicate the double burden of malnutrition in Bangladesh', *Int J Epidemiol.*, 36, 449-457, 2007
13. Dunn O.J. and Clark V.A.: *Applied Statistics: Analysis of Variance and Regression*, Toronto: John Wiley & Sons, 1974
14. NIPORT: 'National Institute of Population Research and Training, Bangladesh Demographic and Health Survey Dhaka, Mitra & Associates and ORC Macro, Bangladesh and Calverton, MD, USA., 2007
15. WHO: 'Diet, Nutrition and the prevention of Chronic diseases', Report of a Joint WHO/FAO Expert Consultation Technical Report, Series No. 916, Geneva, 2003
16. Chan Y.H.: 'Biostatistics 202: Logistic regression analysis', *Singapore Med J.*, 45, 149-153, 2004
17. Singh S.: 'Adolescent childbearing in developing countries: A global review', *Stud Fam Plann.*, 29, 117-136, 1998
18. Pryer J.A., Rogers S. and Rahman A.: 'Factors affecting nutritional status in female adults in Dhaka slums, Bangladesh', *Soc Biol.*, 50, 259-269, 2003