A Machine Learning-Based Prediction and Analysis of Flood Affected Households: A Case Study of Floods in Bangladesh

Kishan Kumar Ganguly*, Nadia Nahar, B M Mainul Hossain

Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh

Abstract

Floods are one of the most frequently occurring disasters in Bangladesh that cause small to large scale damage every year. Most of the studies in the literature provide a flood damage prediction or inference model at individual building level. Some of the works that adopt a higher spatial scale such as households conduct their analysis on a few specific regions. This paper presents a household-level flood damage analysis performed on 2004-2009 flood data from 64 districts of Bangladesh. The study focuses both on prediction and determination of influencing factors because both of these facilitate flood damage reduction programs. A machine learning driven approach has been taken for prediction where three learning algorithms namely linear regression, random forest and artificial neural network are fitted to the data and compared. In this work, linear regression performed better than the other two because its assumptions were considered. A regression analysis showed the significance of the relationship between predictors and damage.

*Corresponding author

Email addresses: kkganguly@iit.du.ac.bd (Kishan Kumar Ganguly), nadia@iit.du.ac.bd (Nadia Nahar), mainul@iit.du.ac.bd (B M Mainul Hossain)

Apart from the significant hydrologic predictors, literacy, flood awareness, house structure and disaster management knowledge were found to be influential. Preparedness was observed to be statistically insignificant unless it was combined with disaster management knowledge. A principal component analysis was further performed to cluster different variables into predictor groups and inspect their effect on flood damage. According to this analysis, hydrologic and environmental predictors, literacy, land ownership and house structure were found to be highly important where precaution and disaster related factors showed less significance.

Keywords: regression, flood damage, parametric method

---

1. Introduction

Being part of the Ganges-Brahmaputra Delta, 405 rivers flow through Bangladesh [1]. As it is an agricultural country, 20-25% monsoon flood inundation is considered useful [1]. However, inundation more than this causes damage to individuals and the society as a whole. Bangladesh faces floods every year ranging from flash floods to major ones causing damage to large areas throughout the country. According to the latest disaster report by Bangladesh Bureau of Statistics, 31 catastrophic floods occurred from 1787-2000 where 68% and 52% of the country went under water in 1998 and 1988 respectively [2]. Furthermore, 4361261 households were affected in floods that occurred in the 2009-2014 period [2]. Studies show that this large damage is significantly related to factors such as socio-economic structure, awareness, house structure etc. [3–5]. Appropriate prediction of the damage and identification of the damage factors can assist in taking precaution measures for

reducing any future risk.

Bangladesh consists of seven divisions which are further divided into 64 districts [2]. Divisions are larger units with an average area of 21081.42 square kilometers where districts have an average area of 2305.78 square kilometers [6]. Therefore, districts have more uniform geographical characteristics. Moreover, districts in Bangladesh have an active District Disaster Management Committee that governs Upazila and Union Disaster Management Committee [7]. Hence, flood damage prediction in district-level can provide administrative benefits to mitigate flood impact. Flood damage mitigation is related to the decision making regarding precaution measures, preparedness etc. In Bangladesh, households are the main source of decision-making. Therefore, a household-level flood damage analysis should provide more insight into the flood damage reduction rather than a building-level one. In addition, several studies have shown that it is more useful to analyze flood damage mitigation measures from household and more local perspectives [8–10].

Traditional flood damage models rely on stage-damage functions that express damage as a function of flood characteristics in a specific location. Multiparameter models are more useful than single parameter ones [3]. Some renowned flood damage models are multiparameter, for example, Flood Loss Estimation Model for Private Households (FLEMOps) used three individual building types, and five inundation depth, two building quality, three contamination and three private precaution classes for estimating damage [4]. Zhai et al. proposed a regression model with inundation depth, house ownership, house structure, length of residence and household income for

predicting damage for households in Japan [5]. Thieken et al. discussed the influence of flood warning, knowledge, preparedness, experience and socio-economic variables on flood damage from building-level [4]. They further performed a principal component analysis to measure the grouped behavior of the variables. In the analysis, hydrologic and socio-economic variables showed significant correlation with building damage where knowledge and precaution were less significant. Bubeck et al. explored the impact of risk perception, coping appraisals and several other factors on flood mitigation behavior that help to reduce damage [9]. Their findings include a positive correlation of risk perception and coping appraisal with mitigation behavior. They further found that knowledge about flood hazard shows a weak relationship with mitigation behavior. Merz et al. proposed a regression tree-based model considering multiple factors divided into hydrologic, emergency measure, experience, building characteristic and socio-economic status classes [3]. Poussin et al. studied the relationship between different factors and flood damage mitigation behavior [10]. They concluded that factors such as flood experience, ownership of home, education level and household size are positively related to flood mitigation measures. Wagenaar et al. compared several supervised learning algorithms for damage prediction where only residential damage was considered. Although some of the studies analyze flood damage from household-level, these have some drawbacks and opportunities for further enhancements. Firstly, these studies are conducted on data collected from either a specific flood or a few specific areas. None of these use local-level (districts) flood data of an extended time period. Secondly, most of the works directly use a specific technique such as linear regression with-

out checking its underlying assumptions. Thirdly, the validation method for checking the model performance has drawbacks of using the training data to construct the test set [3], using a single test set throughout the experiment [11] etc. These may wrongly conclude optimistic performance values from the model. Another issue is that most of these studies aim for either the detection of influencing factors or prediction. In order to reduce the flood damage, both predicting the damage and understanding the factors affecting the damage are required. Finally, in some of the studies such as [3, 5], a specific model is directly used without comparing it with the other existing machine learning techniques.

Some works on flood damage assessment of Bangladesh is present in the literature. Tingsanchali et al. used 1988 flood data to produce a hazard map that categorized land units into three hazard zones, namely high, medium and low [12]. A similar type of study on flood hazard and risk map derivation for Dhaka, Bangladesh was presented by Dewan et al. [13]. Yang et al. utilized water level data of Ganges, Brahmaputra, and Meghna basins to evaluate three forms of damage functions which are linear, logistic and exponential where the logistic functions outperformed the others [14]. Although studies related to flood damage map derivation and damage function estimations have been done for Bangladesh, these are mostly based on hydrologic predictors. As shown in the studies mentioned in the previous paragraph, other factors along with hydrologic ones can provide insight into the prediction and assessment of damage. K. M. Nabiul Islam presented a study to assess flood loss in urban areas considering five urban sectors namely residential, business, manufacturing, office and public buildings, and roads [15]. As

this study covers only urban areas, it excludes factors such as income from agriculture, water usage etc. Additionally, it does not perform district-level analysis to measure flood damage impact over whole Bangladesh. It further does not calculate loss from household-level which has been mentioned to be beneficial.

The objective of this study is to assist in flood damage reduction by providing a machine learning-based prediction and inference model derived from the district-level data of 2009-2014 floods [2]. In this study, the flood damage is represented by the ratio of flood affected households. Prediction is done by choosing the strongest machine learning algorithm according to some performance metrics which are calculated using k-fold cross-validation, a well-known technique for validation. During the comparison of performance, we prioritize parametric methods where all the methods perform equally. This is because predictor-response relationship can be more easily interpreted in parametric methods. Moreover, parametric methods perform well if assumptions of parametric methods hold. For this reason, the parametric method used in this study (linear regression) has been tuned so that its assumptions are satisfied. Linear regression has been observed to be the strongest algorithm in our study. Hence, the influence of the predictors on flood damage is determined using regression analysis. The predictors for the study are selected based on the literature. As poverty-stricken people tend to be more affected by floods, factors related to their lifestyle, education and economic condition are considered [16]. Moreover, environmental and hydrologic features such as rainfall, length of major rivers in each district, water level, discharge etc. are considered. Studies show that preparedness

and knowledge are closely related to the resultant damage [17]. This is why factors about precaution and awareness are included. During regression analysis, the findings are compared to the literature to determine whether the factors in this district-level analysis show any different influence on household damage. All these aspects of the study contribute to overcoming the aforementioned problems in the existing literature. In addition, a principal component analysis is performed to group the variables into specific classes called principal components. A regression analysis is conducted using these principal components as predictors. This helped to determine the influence of the predictor classes on flood damage.

2. Methodology

The data was collected from the latest disaster report by Bangladesh Bureau of Statistics, Statistical Year Book Bangladesh 2015, District Statistics and Bangladesh Water Development Board (BWDB) [2, 6, 18, 19]. For each district, predictor values were collected and integrated with the response (the ratio of total affected household) to construct the training data. In the data, the collected predictors have different units. For example, the river length is in miles and the rainfall is in millimeters. These predictors need to be re-scaled because a predictor may have larger impact on the response than another one due to its scale. In this work, re-scaling is done using Z-score standardization. Equation 1 shows the calculation of Z-scores.

$$Z_x = \frac{x_i - \bar{x}}{s_x} \tag{1}$$

The equation shows that standardization involves subtracting the mean and dividing by the standard deviation. Consequently, all the predictors are re-

scaled with a mean of zero and a standard deviation of one.

Feature selection was performed to select the best performing subset of features. This work uses the stepwise selection technique with Akaike's Information Criterion (AIC) [20]. In this technique, features are included and eliminated at each stage based on some criterion which is, in this case, AIC [21]. AIC is calculated using Equation 2.

$$AIC = -2 \log L(\hat{y}) + 2c \tag{2}$$

$L(\hat{y})$ is the maximized likelihood function and c is the number of predictors. Hence, AIC balances the number of predictors with the predictability of the model. As seen from this equation, lower AIC indicates better model. In the stepwise feature selection technique, each predictor that reduces AIC is added one by one. Backward elimination is then executed where AIC reducing predictors are removed at each step. The remaining predictors are considered for prediction. The data and the selected predictors are further discussed in Section 3 and 4.1.

Using the training data, three multivariate machine learning algorithms namely linear regression, random forest and artificial neural network have been applied to predict the ratio of total affected households. These algorithms are compared using a Paired T-Test where Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Correlation Coefficient are used as metrics. These metrics are calculated by comparing the performance of the algorithm on a test data set. The test data is constructed from the training data using the k-fold cross-validation technique [22]. In this technique, the training data is randomized and divided into k groups where one group is used for testing and the remaining k-1 groups are utilized for training. This

process is repeated k times. The metric values for each of the test data are then averaged. The benefit of this technique is that training and test data are different data sets drawn from a randomly sampled data. Otherwise, the metric value would always indicate high performance where the true performance may be lower. Furthermore, k-fold cross-validation has lower variance. This is because the metric values are averaged over multiple test data sets. Using a single holdout set for validation would yield different results for different test data set. Some of the previous works have limitations from this perspective. Merz et al. used a random sample directly from the training data that was used for learning [3]. As a result, the external validity of the study was hampered. Wagenaar et al. held back a part of the data as the test data set and it was consistently used for validation throughout their study [11]. Utilizing such single set of test data has the aforementioned problems. The k-fold cross-validation based technique followed in this paper overcomes these drawbacks.

The Paired T-Test shows whether the difference among these three algorithms regarding RMSE, MAE and Correlation Coefficient is statistically significant. As parametric models are easily interpretable, linear regression has been used as the base classifier for comparing its performance with the other two. Here, the null hypothesis is that the true mean performance difference between two observations (linear regression and another learning algorithm) is zero. The alternative hypothesis is the true mean performance of linear regression is higher than the other algorithm. Paired T-Test is conducted to observe whether the null hypothesis can be rejected. Linear regression is selected in case of rejection. Otherwise, if the null hypothesis

cannot be rejected, there is no significant difference. In this case, linear regression is selected when its performance metric value is higher or equal to that of the other algorithm. The following sections briefly describe linear regression, random forest and artificial neural network.

2.1. Linear Regression

Linear regression assumes a linear model of format $\hat{y} = \alpha x + c$ where $\hat{y}$, x, c and α are the predicted value of the response variable y, predictor, intercept and slope respectively. The $(y - \hat{y})$ is called the residual of the model. Using a training data set, linear regression attempts to minimize the sum of squared residuals (RSS) from Equation 3.

$$RSS = \sum_{i=1}^{k} (y_i - \hat{y}_i)^2 \tag{3}$$

In order to make prediction using multiple predictors, the following multivariate linear regression model needs to be considered.

$$\hat{y} = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_n x_n + c \tag{4}$$

Linear regression requires satisfaction of six assumptions which are observation of predictors and responses without measurement error, linearity, no autocorrelation, homoscedasticity, normality of residuals and no or little multicollinearity [23]. Linearity assumption states that the predictor-response relationship must be linear. Autocorrelation means that residual values are dependent on each other. In this case, residuals show patterns when these are plotted. In the absence of autocorrelation, residuals appear to be random. Homoscedasticity is the uniformity of error term for all the values of the predictor. This error is the stochastic noise that arises from the effect

of unspecified predictors or randomness involved in the experiment. Linear regression further requires residuals to be normally distributed. The sixth assumption, no or little multicollinearity, states that predictors must be linearly independent. To say practically, the goal is to minimize multicollinearity as much as possible [23].

2.2. Random Forest

Random forest is an improvement over bagged decision trees [24]. The decision tress that handles continuous variables are known as regression trees. Regression trees work similarly as decision trees by splitting each node based on a node impurity measure. Here, the node impurity is measured using errors sum of squares as shown in Equation 5.

$$ESS = \sum_{i=1}^{n(S_1)} y_i - \bar{y_1} + \sum_{i=1}^{n(S_2)} y_i - \bar{y_2} \tag{5}$$

S_1 and S_2 are the two children after splitting. The variables are exhaustively searched to find the split with minimum ESS.After regression tree has been constructed, the values of the given predictors help to conditionally reach a leaf node. The predicted response value is the average value of the leaf node.

Overfitting is a major problem of the regression tree. Bagging aims to avoid overfitting by reducing variance [25]. If the training data is used to fit only a single decision tree, overfitting may arise. To solve this, the data is bootstrapped, that is, a sample is repeatedly taken from the data and a decision tree is fitted for each of these. During prediction, results over all these decision trees are averaged, which causes lower variance of the prediction. This is because the variance of the mean of n independent observations $x_1, x_2, x_3, \cdots, x_n$ with σ^2 variance is $\frac{\sigma^2}{n}$. However, if there is a very strong

predictor in the data, the bagged trees will look similar because the strong predictor will be on the root of the trees. Therefore, variance will not be significantly reduced by averaging due to correlation among the trees. Random forest improves this by considering a subset of the predictors at each split. Generally, $k = \sqrt{p}$ is chosen at each split where p is the total number of predictors.

2.3. Artificial Neural Network (Multilayer Perceptron)

A multilayer perceptron is a feedforward artificial neural network which consists of an input layer, one or multiple hidden layer and an output layer [26]. Input layer receives the predictor values and output layer provides the prediction. Hidden layers combine input predictors to detect features. Learning can be done using backpropagation algorithm [27]. At every neuron, output can be computed using Equation 6.

$$y_k(i) = \theta \times \Big[\sum_{j=1}^{n} x_j(i) \times w_{jk}(i) - th_k \Big] \tag{6}$$

Here, $y_k(i)$ is the output of neuron k in the i^{th} iteration. $x_1, x_1, \cdots, x_n$ are the inputs from the previous layer. $w_{jk}(i)$ is the weight between input x_j and output neuron y_k, th_k is the threshold and θ is the activation function. For regression the threshold is omitted. This paper uses the sigmoid function as the activation function as it is vastly used in the literature. The sigmoid function is shown in Equation 7.

$$y_{sig} = \frac{1}{1 + e^{-x}} \tag{7}$$

In the backpropagation network, the error at the output layer is calculated and propagated backwards to update the network weights. The error is

calculated as difference between output from the output layer and the data, as in Equation 8.

$$e_o(i) = y_{od}(i) - y_o(i) \tag{8}$$

Here, $y_{od}(i)$ is the desired output and $y_o(i)$ the output produced by the network in the output layer at iteration i. The weight at the output layer is updated using this error with Equation 9, 10 and 11.

$$w_{jo}(i+1) = w_{jo}(i) + \delta w_{jo}(i) \tag{9}$$

$$\delta w_{jo}(i) = \alpha \times y_j(i) \times \delta_o(i) \tag{10}$$

$$\delta_o(i) = y_o(i) \times (1 - y_o(i)) \times e_o(i) \tag{11}$$

However, for the hidden layer, weights are updated as follows.

$$w_{sj}(i+1) = w_{sj}(i) + \delta w_{sj}(i) \tag{12}$$

$$\delta w_{sj}(i) = \alpha \times x_s(i) \times \delta_j(i) \tag{13}$$

$$\delta_j(i) = y_j(i) \times (1 - y_j(i)) \times \sum_{k=1}^{o} \delta_k(i) \times w_{jk}(i) \tag{14}$$

Here, $x_s(i)$ corresponds to the input of the s^{th} neuron in the i^{th} iteration and $y_j(i)$ is the output of the neuron. The summation indicates that errors from the output layer are backpropagated where o is the number of neurons in the output layer. After updating all the weights of the network, next iterations are consecutively performed until a specific error criterion is satisfied. Generally, sum of squares is used as the error criterion where it needs to be under a prespecified value.

Apart from prediction, as linear regression was the strongest predictor from the analysis (Section 4.1), regression analysis was performed to analyze

the impact of the predictors on the response. A one sample T-Test was performed on the regression model to determine which predictors significantly impacted the response. The null hypothesis is that the slope or the coefficient of a predictor is zero, that is, there is no relationship between the predictor and the response. If the P-value from the test is less than a prespecified significance level, the null hypothesis is rejected. Regression analysis captures the impact of an individual predictor on the response. If a group-wise impact needs to be analyzed, for example, the statistical significance of the relationship between hydrologic predictors and the response, this type of regression analysis does not suffice. In this paper, we use Principal Component Analysis (PCA) for this purpose. PCA expresses data in lower dimensions that explains the highest variations of the data. The first principal component explains the maximum variance and expressed as Equation 15.

$$PC_1 = l_{11}x_1 + l_{21}x_2 + l_{31}x_3 + \cdots + l_{n1}x_n \tag{15}$$

Here, l_{11}, l_{11} etc. are loadings of PCA. The loading of a predictor indicates how much variance is explained by it within that principal component. The second principal component is defined similarly and it explains the highest variance from the remaining variance. These principal components can be used as predictors in a regression analysis technique to analyze its relationship with the response.

3. Data

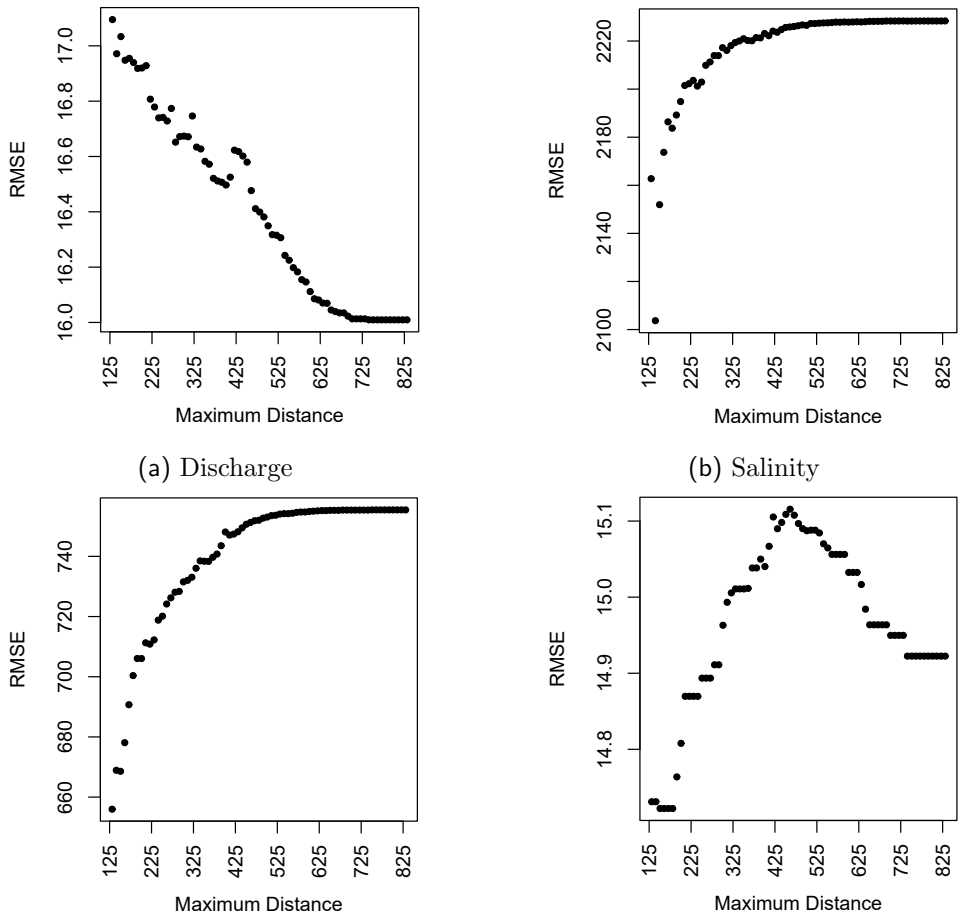
The name and description of the predictors and the response are given in Table 1. The table shows that 29 predictors were collected which were classified into five classes. The economic predictors are related to land ownership,

Table 1: Predictor and Response Classes, Names and Description (Acronyms are given for each of the predictors for representation purposes only)

Type	Predictor Class	Name	Acronym	Definition
Predictor	Economic	Having Own Land	ol	The ratio of households having own land to total households
		Having Pucca House	ph	The ratio of households with pucca house to total households
		House Ownership Ratio	or	The ratio of households having own house to households living in rented house
		5 Acres+ Operated land	oln	The ratio of households operating more than 5 acres land to total households
		Income Source Business	isb	The ratio of households with income source as business to total households
		Income Source Agriculture	isa	The ratio of households with agricultural income source to total households
		Income Per Household Member	ihm	The ratio of average total income of a household to average household size
	Lifestyle	Household Size	hs	The average size of a household
		Household Head Male	hhm	The ratio of households with male household head to households female head
		Households Using Tube Well and Supply Water	htw	The ratio of households using tube well and supply water to total households
		Households Using Natural Water Sources	hnw	The ratio of households using water from rivers, ponds and rain to total households
		Households Using Other Water Sources	hws	The ratio of households using other water sources except for supply, tube well and river to total households
	Educational	Male Literacy Rate	mlr	The ratio literate male to population
		Literacy Rate S.S.C/ H.S.C	lrs	The ratio of persons studied up to S.S.C/ H.S.C to total population
		Literacy Rate Grad and Above	lrg	The ratio of persons completed graduation or more to total population
		Literacy Rate	lr	The ratio of literate people to total households
	Environmental and Hydrologic	Low Land	ll	The ratio of the area covering low land to total area
		River Length (miles)	rl	Total length of major rivers through the district
		Rainfall (mm)	rf	Average yearly rainfall
		Humidity	hmd	Maximum yearly humidity
		Water Level	wl	The maximum average river water level of a district
		Discharge	dcr	The maximum average river discharge of a district
		Salinity	sal	The maximum average river water salinity of a district
		Sediment	sed	The maximum average river sediment of a district
	Precaution and Awareness	Households Having Climate Change Knowledge	cck	The ratio of households with climate change knowledge (long term, regional or sudden) to total households
		Households Having Disaster Management Knowledge	dmk	The ratio of households with disaster management knowledge (pre-, post- and during-disaster) to the households having knowledge about management only during disaster
		Households Having Sea Level Rise Awareness	sra	The ratio of households having knowledge and perception about the impact of sea level rise to total households
		Households Having Flood Awareness	fa	The ratio of households having knowledge and perception about impact of floods to total households
		Preparedness	prp	The ratio of households that took action (precaution) during flood period until normal situation to total households
Response		Ratio of Total Affected Household	affh	The ratio of flood affected households to total households

house structure and income. Masozera et al. showed that these predictors play a significant role to cope with natural disasters [28]. These predictors were collected from the latest disaster report of Bangladesh [2]. The literature further shows that vulnerability to flood damage is associated with the source of drinking water [29]. Moreover, gender and household size contribute to understanding flood damage and coping behavior [10, 28, 30]. All these constitute the lifestyle related predictor class. The data for this class was collected from the Bangladesh Disaster Related Statistics 2015 and Statistical Year Book Bangladesh 2015 [2, 6]. The educational predictors are correlated to flood damage as seen from previous studies [3, 4, 10]. These were collected from the Statistical Year Book Bangladesh 2015 [6]. From the environmental and hydrologic factors, low land, river length (miles), rainfall (mm) and humidity data were collected from Statistical Year Book Bangladesh 2015 and District Statistics [6, 18]. Bangladesh Water Development Board (BWDB) has been collecting Water level, discharge, salinity and sediment data from 366, 154, 138 and 21 stations throughout the country which were incorporated in our study [19]. Furthermore, high preparedness and precaution measures such as disaster related knowledge, flood awareness and perception, climate change knowledge etc. are associated with flood damage according to the literature [31] [3, 4, 9]. Hence, predictors corresponding to precaution and awareness are considered. The response is the ratio of total affected household to total households.

Among the predictors, water level, discharge, salinity and sediment had missing data. This is because stations for collecting these have not yet been constructed in every district. Out of 64 districts, stations measure discharge



(a) Discharge (b) Salinity

(c) Sediment (d) Water Level

Figure 1: The Alteration of RMSE Values with Increasing Maximum Distance for Discharge, Sediment, Salinity and Water Level

in 49, water level in 63, salinity in 29 and sediment in 15 districts. We estimate the missing data by following the Inverse Distance Weighting (IDW) interpolation technique. IDW has been used successfully in the literature to estimate missing data values of attributes such as rainfall [32]. IDW assumes that the nearby points contribute more than the distant ones. According to

Table 2: Descriptive Statistics (Minimum, Maximum, Mean and Standard Deviation) of the Data

Type	**Predictor Class**	**Name**	**Min**	**Max**	**Mean**	$\boldsymbol{\sigma^2}$
Predictor	Economic	Having Own Land	0.937	1	0.982	0.017
		Having Pucca House*	0.017	1.971	0.498	0.364
		House Ownership Ratio	6.765	770.32	225.22	236.15
		5 Acres+ Operated land	0.084	50.29	4.938	9.855
		Income Source Business	0.081	0.311	0.145	0.048
		Income Source Agriculture	0.182	0.616	0.384	0.1
		Income Per Household Member	0.023	0.837	0.048	0.101
	Lifestyle	Household Size*	3.91	5.86	4.655	0.412
		Household Head Male	5.71	36.74	19.34	7.885
		Households Using Tube Well and Supply Water	0.347	0.999	0.845	0.22
		Households Using Natural Water Sources	0	0.377	0.036	0.084
		Households Using Other Water Sources	0	0.068	0.008	0.015
	Educational	Male Literacy Rate*	0.464	0.765	0.624	0.069
		Literacy Rate S.S.C/ H.S.C*	0.043	0.169	0.091	0.027
		Literacy Rate Grad and Above	0.004	0.027	0.012	0.006
		Literacy Rate	34.98	70.54	50.17	7.651
	Environmental and Hydrologic	Low Land*	0	72	21.719	19.394
		River Length (miles)*	15	180	71.189	32.863
		Rainfall (mm)*	90	344	180.484	46.83
		Humidity*	62.458	80.331	71.812	4.504
		Water Level*	1.37	90.81	16.146	16.654
		Discharge	2.7	83.3	11.692	12.574
		Salinity	115.577	5287.659	1357.66	1829.731
		Sediment*	106.98	1539.12	736.2037	487.0911
	Precaution and Awareness	Households Having Climate Change Knowledge	0.444	0.988	0.821	0.112
		Households Having Disaster Management Knowledge*	0.099	6.816	1.767	1.112
		Households Having Sea Level Rise Awareness*	0.068	0.662	0.301	0.155
		Households Having Flood Awareness*	0.009	0.41	0.171	0.091
		Preparedness*	0	0.958	0.3	0.301
Response		Ratio of Total Affected Household	0	0.86	0.335	0.268

* Significant predictors

IDW, the missing values are estimated using Equation 16.

$$\hat{x} = \sum_{i=1}^{n} w_i x_i \tag{16}$$

Equation 17 shows how the the weight w_i is calculated.

$$w_i = \frac{d_i^{\gamma}}{\sum_{k=1}^{n} d_k^{\gamma}} \tag{17}$$

Here, x_i is the value of the predictor for nearby districts and w_i is the weight calculated from Equation 17 and d_i is the distance from each nearby district. Two parameters need to be specified for IDW which are the size of the neighborhood n and the value of γ. In our experiments, we empirically selected γ as 2. In case of neighborhood selection, some authors suggest to select a fixed number of nearby districts [33]. Another technique is to choose neighbors within a prespecified maximum distance. In our work, the maximum distance is selected as follows. We partition the data into two sets which are the data set with unknown values, S_u and known values, S_k. The distances inside the range (d_{min}, d_{max}) of S_k are selected one by one as the maximum distance for neighborhood construction. For each maximum distance, IDW is utilized to estimate the missing values in S_u. These estimated values are used to further estimate the values of S_k which are compared to its known values by RMSE. The maximum distance with the lowest RMSE is selected for defining the neighborhood. Figure 1 shows the RMSE values for different maximum distance selection. The pattern of the alteration of RMSE depends on the data and the missing values. Hence, it is different for each of the predictors. From the figure, RMSE is lowest for discharge at 741, for salinity at 131, for sediment at 131 and for water level at 151.

Table 2 shows the mean, standard deviation, minimum and maximum values of the data. House ownership ratio, 5 Acres+ operated land, household head male, households using natural water sources and other water sources, low Land, river length (miles), water level, discharge, salinity, sediment and preparedness have comparatively larger standard deviation regarding their range. This increases the risk of non-normality of these predictors. However,

In Section 4.1, it is showed that after transforming the predictors that exhibit non-linear relationship with the response, the normality assumption is satisfied.

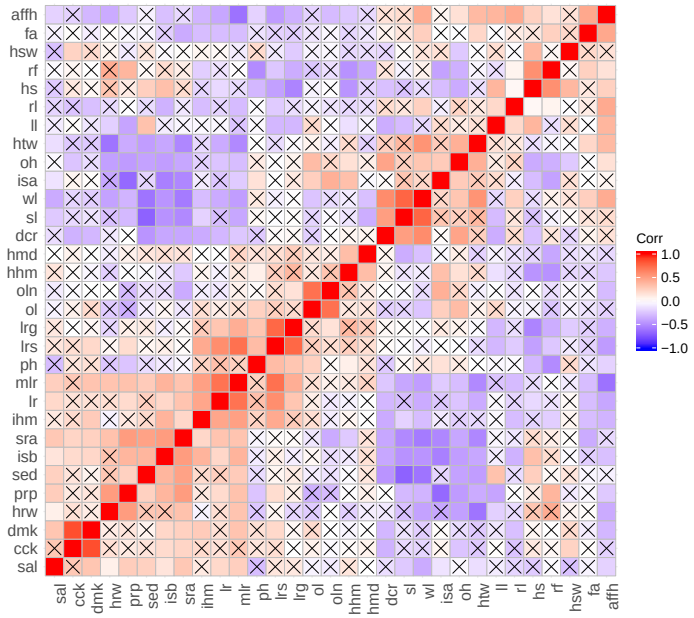


Figure 2: Spearman's Rank Correlation of the Features at 0.1 Significance Level (Non-Significant Features are Marked with a Cross (X))

Figure 2 depicts the Spearman's Rank Correlation values of the variables. Each cell depicts the correlation between the variables that form the cell. The non-significant predictors are crossed. For the ratio of total affected household, 14 features are statistically significant at 0.1 confidence level. Among these, male literacy (mlr) has the highest correlation value (-0.61) followed by households having flood awareness (fa) (0.45), river length (rl) (0.44), water level (wl) (0.43), literacy rate S.S.C/ H.S.C (-0.43), literacy rate (lr) (-0.38), low land (ll) (0.38), households using natural water sources (hnw) (-0.36), households using tube well and supply water (htw) (0.35), literacy rate

grad and above (lrg) (-0.34), income per household member (ihm) (-0.32), households having disaster management knowledge (dmk) (-0.31), 5 acres+ operated land (oln) (-0.29) and Income Source Business (isb) (-0.27). The correlation coefficients of some predictors are counterintuitive. For instance, salinity and sediment are negatively correlated with flood damage. Moreover, some important predictors do not show significant correlation. These issues occur due to two reasons. The first issue is data quality, for example, the hydrologic predictors had missing values and required to be imputed. Secondly, correlation does not consider the effect of other predictors opposed to multiple regression. A predictor may become significant when other predictors are held constant which is captured by multivariate regression. For example, Section 4.2 shows that these predictors are associated with the damage in an intuitive way. Along with the Spearman's Rank Correlation values, the Pearson correlation coefficient values were calculated and compared to the Spearman's correlation values. It was observed that both of these correlation values lie very close for most of the predictors. However, the Spearman's correlation values for 17 predictors are greater than Pearson correlation values. Out of these 17 predictors, 7 predictors are part of the selected features. Therefore, some of the selected features have a non-linear relationship with the response. As Random forest and multilayer perceptron models are non-parametric, this nonlinearity does not affect their performance. Nevertheless, as linearity is one of the assumptions of linear regression, data is transformed to satisfy this assumption (Section 4.1).

Table 3: MAE, RMSE and Correlation Coefficient values (* indicates statistically significant at 0.05 level compared to Linear Regression)

Metrics	**Linear Regression**	**Random Forest**	**Multilayer Perceptron**
MAE	0.13	0.16*	0.22*
RMSE	0.15	0.20*	0.28*
Correlation Coefficient	0.80	0.67*	0.61

4. Result Analysis and Discussion

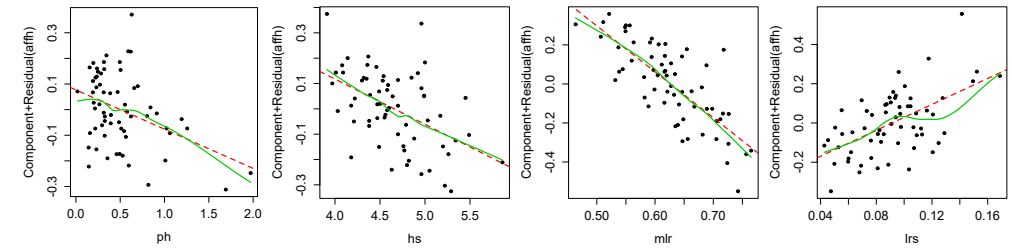
4.1. Paired T-Test

Prior to the Paired T-Test, feature selection was performed following stepwise selection technique with AIC (Section 2). 14 features are selected which are shown in Table 2. After feature selection, the data is checked for linear regression assumptions and appropriate measures are applied when an assumption is violated. For example, Tukey's transformation is applied to improve the linearity of the predictors. The data is further standardized as mentioned in Section 2. A paired T-Test with linear regression as the base predictor is executed using this data. Table 3 shows the Paired T-Test result of linear regression, multilayer perceptron and random forest at 0.05 confidence level. The RMSE values are 0.15, 0.20 and 0.28 for linear regression, random forest and multilayer perceptron respectively where 0.20 and 0.28 are significantly larger compared to linear regression. Similarly, for MAE, linear regression produces significantly better results than the other two. For correlation coefficient, linear regression performs significantly better than the other two. Therefore, the parametric model, linear regression is

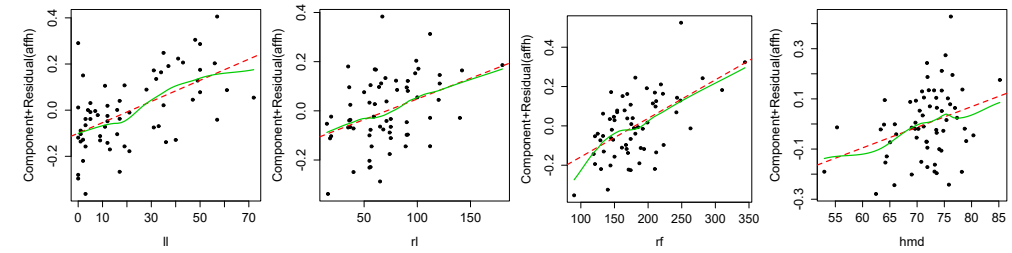
chosen for better explainability with acceptable prediction power.

In previous studies such as [10], [11] and [34], linear regression was used without considering the six assumptions mentioned previously. As a result, Wagenaar et al. and Dawson et al. found linear regression to be outperformed by regression trees. The performance of a machine learning algorithm is highly dependent on predictor selection and preprocessing applied on data. Nevertheless, in our study, we ensure that the six assumptions are met for linear regression. Hence, the performance indicator values are more valid ones than those of previous studies. The linear regression assumptions are inspected and met as follows.

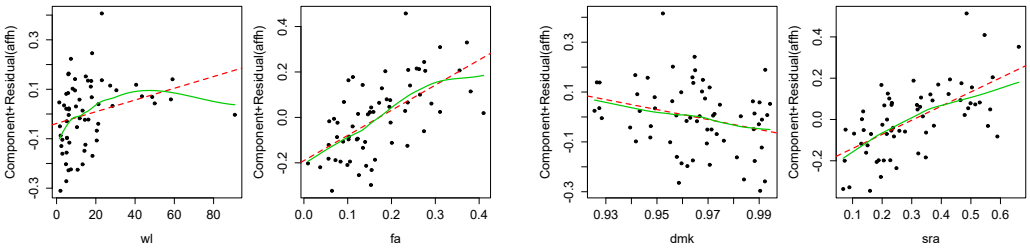
Although linearity assumption can be checked by plotting the predictors against the response, it does not consider the effect of a predictor with respect to the other predictors. In multivariate regression, the predictor-response relationship is assessed considering the presence of other predictors. Partial residual plots help to achieve this by plotting $residual+\alpha_i x_i$ versus x_i which is the green solid line in Figure 3. The red dotted line is the linear fit line which represents a least squares regression line. Linearity assumption is verified by comparing how closely the first line lies to the least squares regression line. To improve the conformance to linearity assumption, the predictors can be transformed. This work uses the Tukey's power transformation where a linear model $y = a + bx$ is re-expressed as $y = a + bx^{\lambda}$ [35]. The transformation is performed by continuously testing λ with different values and observing whether the predictor-response relationship becomes linear. A series of λ values can be tested using a table called Tukey's Ladder of Transformation [35]. In this case, $\lambda = -2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$ and 2 are used. Furthermore, $\lambda = 0$



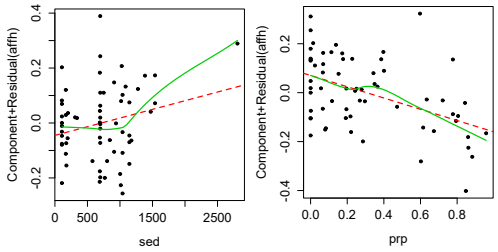
(a) Having Pucca House (b) Household Size (c) Male Literacy Rate (d) Literacy Rate S.S.C / H.S.C



(e) Low Land (f) River Length (g) Rainfall (h) Humidity



(i) Water Level (j) Households Having Flood Awareness (k) Households Having Disaster Management Knowledge (l) Households Having Sea Level Rise Awareness



(m) Sediment (n) Preparedness

Figure 3: Component+Residual Plots for the Predictors (The Green Solid Line is the Partial Residual Line and the Red Dotted Line is the Linear Fit Line)

is replaced by logarithmic transformation where $log(x)$ is used in place of x. We applied all the λ values from the Tukey's Ladder of Transformation to all the predictor and examined linearity both graphically and quantitatively. Improved linearity was observed after transforming wl, sed, hmd, dmk, lrs and prp. Logarithmic transformation was applied to wl, dmk and hmd, and $\lambda = 2$ was used for sed, lrs and prp. To say quantitatively, before applying the transformation, the values of R^2 and Adjusted R^2 were 0.75 and 0.68 respectively. After transformation, R^2 and Adjusted R^2 values increased to 0.8 and 0.74 respectively which indicates improved linearity.

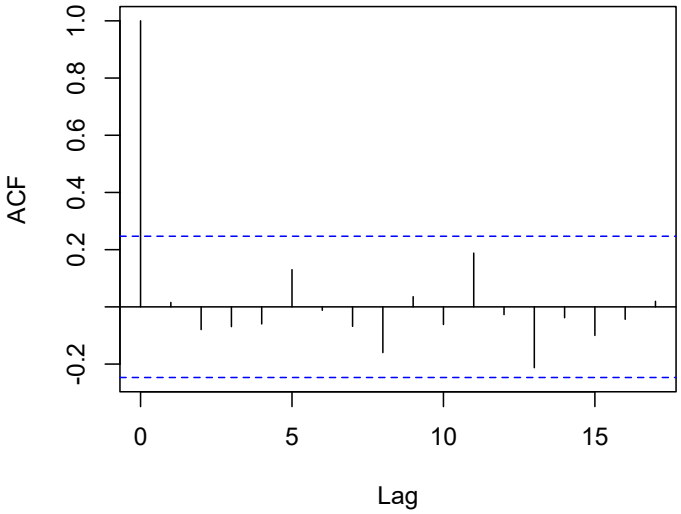


Figure 4: Autocovariance and Autocorrelation Functions Plot (The Blue Dotted Lines Indicate 95% Confidence Limits)

Figure 4 shows the autocorrelation function for residuals. The autocorrelation function measures the correlation among the residuals which can be expressed using Equation 18.

$$F_{auto} = F_{cor}(r_i, r_{i-n}) \tag{18}$$

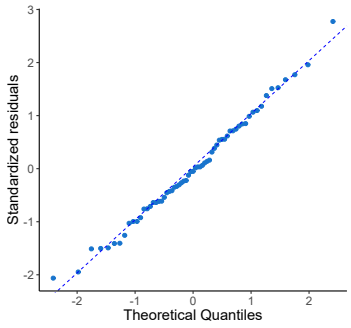


Figure 5: Quantile-Quantile Plot

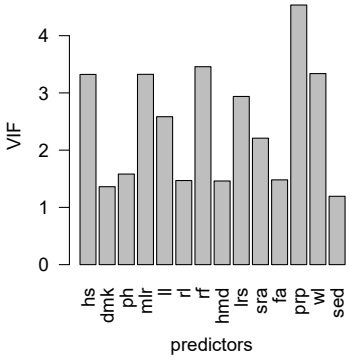


Figure 6: VIF Plot

F_{cor} measures the correlation between i^{th} and $(i-n)^{th}$ residuals. The n is the lag that represents the correlation between values in n time periods distance. In the figure, the first vertical line with acf value 1 indicates the correlation of the residual with itself. The blue dotted lines are the 95% confidence limits. If the values of the lags cross these lines, autocorrelation is present. From the figure, no autocorrelation is present as all the acf values of the lags (except lag 0) are within the confidence limits. This was further confirmed by Durbin-Watson test for autocorrelation [36]. The p-value (0.2187) indicates that we cannot reject the null hypothesis of independence among residuals, that is, the residuals are not autocorrelated.

To detect heteroscedasticity, the Breush-Pagan test was conducted [37]. Heteroscedasticity means that the variance of the residuals of a model is not equal across all the values of the response. The null hypothesis of the Breush-Pagan test is that the variances of the residuals are equal [37]. The p-value of 0.6163 indicates that the null hypothesis cannot be rejected. Hence, heteroscedasticity is not present. For testing the normality assumption, the normal Q-Q plot is used which is shown in Figure 5. The standardized resid-

uals are plotted against quantiles of the standard normal distribution. The linear trend of the points indicates that the distribution of the residuals is approximately normal. This is further supported by a Shapiro-Wilk normality test where the null hypothesis is that the residuals come from a normally distributed population. The p-value of the test is 0.91. Therefore, the null hypothesis cannot be rejected. To examine multicollinearity, Variance Inflation Factor (VIF) is considered [38]. Multicollinearity is the phenomenon when the predictor variables are intercorrelated. Multicollinearity is significant if VIF is greater than 10 [38]. As Figure 6 depicts, VIF is less than 10 for all the predictors. Hence, multicollinearity is not significantly high.

4.2. Regression Analysis

As discussed previously, the data was fit to a linear regression model. To evaluate the model, cross-validation was performed. The R^2 and Adjusted R^2 are 0.8 and 0.74 which indicates data fits well to the linear model. The F-Statistic is 13.77 on 14 and 48 degrees of freedom. The p-value of the F-Test is 2.589e$-$12 which is much less than the 0.05 significance level. This p-value signifies that the null hypothesis, the fit of the intercept-only model (the model without any predictors) and the fitted model is equal, can be rejected.

Table 4 shows the regression analysis results. The estimate column contains the coefficient values. These are standardized for comparability which constitute the beta column. For significance level 0.001, 0.01, 0.05 and 0.1, the probability of observing any value larger than T is calculated. At 0.001 significance level, male literacy rate (mlr), sea level rise awareness (sra), flood awareness (fa) and low land (ll) predictors are related to the response. The negative slope of mlr indicates that increasing male literacy rate contributes

Table 4: Coefficients of Linear Regression Analysis (****p<0.001, ***p<0.01, **p<0.05, *p<0.1)

Predictors	**Estimate**	**beta**	**Std. Error**	***T* value**	***Pr*(> |*T*|)**
Intercept	0.331****		0.017	19.4	<2E-16
hs	-0.080**	-0.3	0.031	-2.6	0.013
dmk	-0.037*	-0.14	0.02	-1.9	0.067
ph	-0.061***	-0.23	0.022	-2.8	0.007
mlr	-0.118****	-0.44	0.031	-3.8	4.00E-04
ll	0.099****	0.37	0.028	3.6	7.00E-04
rl	0.048**	0.18	0.021	2.3	0.026
rf	0.080**	0.3	0.032	2.5	0.016
hmd	0.056***	0.21	0.021	2.7	0.01
lrs	0.062**	0.23	0.029	2.1	0.041
sra	0.118****	0.44	0.025	4.6	3.00E-05
fa	0.102****	0.38	0.021	4.9	1.00E-05
prp	-0.053	-0.2	0.036	-1.4	0.157
wl	0.084***	0.32	0.031	2.7	0.01
sed	0.047**	0.18	0.019	2.5	0.016

to decreasing damage. To further analyze the effect of gender and literacy on damage, we replaced male literacy with female literacy rate and re-conducted the analysis. The female literacy rate was observed to be negatively related to damage at 0.001 significance level with a standardized coefficient of -0.478. We inspected whether there is a statistically significant difference between the standardized regression coefficient of male and female literacy. To do this, a boolean indicator called f (0 for male and 1 for female) and an interaction term f $\times$ lrmf were included as predictors, where lrmf indicates literacy rate of male or female. Equation 19 and 20 show that the coefficient of the

interaction term α_3 captures the difference between the coefficients of male and female literacy.

$$affh = \alpha_1 \times f + \alpha_2 \times lrmf + \alpha_3 \times f \times lrmf + \alpha_4 \times hs + \cdots + c \tag{19}$$

$$affh = \begin{cases} (\alpha_2 + \alpha_3) \times lrmf + \alpha_4 \times hs \\ + \cdots + c + \alpha_1, & f = 1 \\ \alpha_2 \times lrmf + \alpha_4 \times hs + \cdots + c, & f = 0 \end{cases} \tag{20}$$

A regression analysis was performed including these predictors where α_3 was observed to be statistically insignificant. Therefore, there is no statistically significant difference between the contribution of male and female literacy on damage risk reduction. The finding that literacy reduces damage is consistent with the findings by Thieken et al., poussin et al. and Merz et al. [3, 4, 10]. Thieken et al. and Merz et al. showed that higher socio-economic status by Plapp [39], decided by education, job position, income etc., is correlated with lower damage [3, 4]. Poussin et al. showed that education is positively associated with avoidance measures.

Districts with more low lands face more flood damage as indicated by the positive coefficient of ll. According to the coefficients of fa and sra, an increase of these variables results in increased damage. This counter-intuitive scenario happens because people in disaster-prone areas have higher flood and sea-level awareness through experiencing disasters over time. To examine this, we extracted the subset of highly affected households determined by affh values greater than the upper quartile. The regression analysis was rerun with this data. It was seen that both fa and sra were negatively related to

affh (not shown here). The predictor fa is associated with flood experience and flood knowledge. According to Thieken et al., Poussin et al. and Bubeck et al, knowledge on flood hazard and experience are significantly related to reduced flood damage, which is confirmed by the high significance of fa in our study.

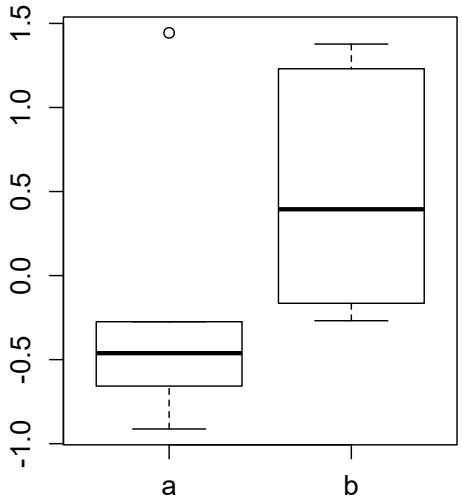


Figure 7: Distribution of the households having disaster management knowledge on two scenario- (a) high damage and preparedness (b) low damage and preparedness

At 0.01 and 0.05 confidence level, household size (hs), having pucca house (ph), river length (rl), rainfall (rf), humidity (hmd), water level (wl), sediment (sed) and literacy rate S.S.C/ H.S.C (lrs) are related to the damage. Larger households leads to lower flood damage. This is because household size positively influences precaution measures [17]. The ownership of a pucca house is positively related to damage similar to the literature [5, 40]. The environmental and hydrologic components namely rainfall, water level, sediment and humidity positively influence flood damage. Flood damage intensifies with increasing lrs. However, the literacy rate of grad and above (lrg)

is negatively correlated with flood affected households when it is fitted to a regression model with affh as response (not shown here). Hence, lrs can be regarded as lack of higher education. These results regarding lrs and lrg corresponds to the issue that higher education level is associated with lower flood damage risk which is supported by the literature [41]. Preparedness is considered to be an important factor in the literature [3, 4] which is, however, statistically insignificant in our analysis. This is because preparedness, which is the act of taking action regarding flood, is effective when it is supported by disaster management knowledge. This is explained using Figure 7 which shows the distribution of disaster management knowledge when (a) preparedness and damage is high and (b) preparedness and damage is low. Here, high and low correspond to the upper and lower quartiles respectively. In scenario a, when preparedness is high, damage can be high if disaster management knowledge is low. In scenario b, damage is low although preparedness is low because the number of households with disaster management knowledge is higher than the previous scenario (Figure 7). In our analysis, disaster management knowledge is negatively associated with the ratio of total affected household. However, it is less significant than the other predictors. This is similar to the findings by Thieken et al. and Bubeck et al. [4, 9].

4.3. Principal Component Analysis

To understand the grouped interaction of the variables on the flood damage, Principal Component Analysis (PCA) was performed. The number of significantprincipal components was chosen based on Kaiser criterion and the scree plot. According to Kaiser criterion, 8 principal components have been selected as seen from Figure 8. These explain 72% of the total variance.

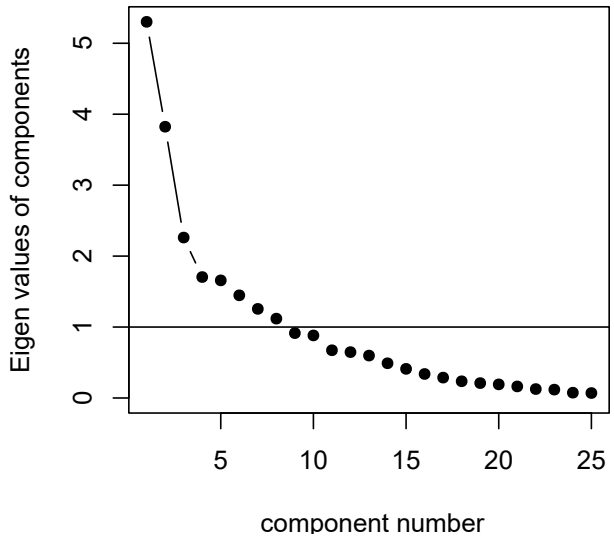


Figure 8: Scree Plot for Principal Component Analysis

Table 5 shows the varimax rotated loadings of these principal components. Using a cut-off factor of 0.5, only the significant variables are shown where their loadings are marked as bold. The first component represents literacy with high positive loadings of literacy rate (lr), Male Literacy Rate (mlr), Literacy Rate S.S.C/ H.S.C (lrs) and Literacy Rate Grad and Above (lrg). The second component represents income with income source business (isb), income source agriculture (isa) and income per household member (ihm), and preparedness (prp). The third principal component captures the water level (wl) and source of drinking water (households using tube well and supply water (htw), and households using natural source of water (hnw)). The fourth one is related to household structure consisting of household size (hs) and household head male (hhm), and rainfall (rf). The fifth one consists of households having climate change knowledge (cck) and disaster management knowledge (dmk). The sixth principal component represents land ownership

Table 5: Varimax Rotated Loading Vectors of First and Second Principal Components (Values with absolute loadings ≥ 0.5 are bolded)

Variables	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
ol	0.1	-0.1	0.03	-0.05	0.12	**0.84**	-0.1	-0.11
lr	**0.73**	0.19	-0.12	0.13	0.02	0	0.09	-0.36
isb	0.03	**0.76**	-0.03	0.09	0.15	-0.01	0.12	-0.09
hs	-0.39	0.27	0.04	**0.7**	0.11	0.15	0.23	0.1
ihm	0.44	**0.51**	0.13	0.03	0.22	-0.11	0.12	-0.21
hhm	-0.08	-0.04	0.07	**-0.86**	0.04	0.14	0.11	-0.04
cck	0.03	0.07	0.05	0.02	**0.91**	0.02	0.12	-0.04
dmk	0.11	0.04	-0.07	0.01	**0.91**	0.1	-0.05	0.05
ph	0.31	-0.29	0.24	-0.3	0.27	0.15	-0.29	**-0.5**
mlr	**0.73**	0.37	-0.25	-0.02	0.17	0.19	0.03	-0.13
ll	-0.2	-0.3	0.2	0.47	0.02	0.2	**0.57**	-0.31
rf	-0.21	0.41	-0.02	**0.55**	0.07	-0.02	-0.09	0.38
lrs	**0.88**	0.1	-0.01	-0.07	0.1	0.07	-0.05	0.04
lrg	**0.73**	-0.19	-0.11	-0.34	-0.04	0.04	0.08	0.12
oln	0.14	-0.09	0.06	0	0.04	**0.7**	0.14	0.32
isa	-0.08	**-0.8**	0.01	-0.26	0.05	0.24	0.1	0.03
fa	-0.1	-0.11	0.11	0.12	0.03	0.06	-0.04	**0.79**
htw	-0.14	-0.19	**0.89**	0.02	-0.03	-0.05	-0.05	0.14
hnw	0.04	0	**-0.91**	0.05	-0.02	-0.13	-0.08	0.09
prp	0.27	**0.77**	-0.43	-0.03	0.08	0.03	0.03	-0.01
wl	-0.04	-0.4	**0.5**	0.03	-0.06	-0.37	-0.34	0.31
sed	0.11	0.08	-0.04	-0.08	0.06	-0.06	**0.87**	0.03

with positive loadings of having own land (ol) and 5 acres+ operated land (oln). The seventh one is related to low land (ll) and sediment (sed), and the eighth one represents house structure with a negative loading of having

Table 6: Coefficients of Linear Regression Analysis with Principal Components (****p<0.001, ***p<0.01, **p<0.05, *p<0.1)

Predictors	**Estimate**	**Std. Error**	$\boldsymbol{T}$ **value**	$\boldsymbol{Pr(>\|T\|)}$
Intercept	0.3912	0.0281	13.9	0
PC1	-0.0786***	0.0238	-3.31	0.0017
PC2	-0.041*	0.0238	-1.72	0.0906
PC3	0.0899****	0.024	3.75	0.0004
PC4	0.0653***	0.0238	2.74	0.0083
PC5	-0.0507**	0.0237	-2.14	0.0373
PC6	-0.0616****	0.0157	-3.94	0.0002
PC7	0.0786***	0.0242	3.24	0.002
PC8	0.0753***	0.0254	2.97	0.004

pucca house (ph), and flood awareness (fa).

Table 6 shows regression analysis results using the principal components as predictors and the ratio of total affected households (affh) as response. Considering the coefficient estimates and the Pr values, the principal components can be ranked as PC3, PC6, PC1, PC7, PC8, PC4, PC5 and PC2. Flood damage increases with PC1 indicating that increased water level and the use of tube well and supply water is associated with higher damage where the use of natural water sources is negatively related to damage. This is because from Figure 2 we observe that people in areas with high water level tend to use tube well and supply water more than natural sources of water. The negative coefficient of PC6 shows that higher amount of land ownership is related to lower flood damage. Higher literacy rate contributes to lower affected households as seen from the coefficient of PC1. The positive

association between PC7 and affh is because ll and sed both are positively correlated with affh. For PC8, flood damage increases as ph and fa increase. The counter-intuitive case of positive association between affh and fa is due to the aforementioned reason (Section 4.2). According to the positive coefficient of PC4, damage increases with increasing household size and rainfall, and decreasing male to female household head ratio. PC5 and PC2 are negatively related to affh. For PC5, it means higher disaster related knowledge (dmk and cck) leads to lower damage. In case of PC2, isb leads to better income and literacy (Figure 2) resulting in decreased affh where isa indicates otherwise. Moreover, Higher preparedness is associated with lower damage.

Hydrologic and environmental components with wl, sed, ll and rf variables show highly significant correlation with damage according to PC3, PC7 and PC4. PC3 is significant at 0.001, and PC7 and PC4 significant at 0.01 significance level. Variables that influence ownership and socio-economic status such as literacy (PC1), land ownership (PC6) and house structure (PC8) are highly significant for predicting affh (Table 6). However, disaster knowledge (PC5) and preparedness (PC2) show comparatively less significance. These are consistent with the findings by Thieken et al. where loss ratio is significantly correlated with hydrologic and environment factors (flood impact items), ownership and socio-economic status, and less correlated with precaution and disaster knowledge related factors [4].

5. Conclusion

This paper has derived a machine learning based model for flood damage analysis from district-level data of affected households. Most of the studies

in the literature consider building-level damage and limited set of areas or floods. The paper differs as it adopts a higher spatial scale (households) and uses a 6-year flood data of 64 districts. Linear regression is the strongest prediction algorithm for this data with MAE 0.13, RMSE 0.15 and correlation coefficient 0.80. The findings from the regression analysis are mostly consistent with the literature. Inconsistencies occur in case of preparedness. According to our study, preparedness is effective in flood damage reduction when it is supported by disaster management knowledge. In addition, a few factors from PCA such as income per household member, household head male and source of drinking water show counter-intuitive behavior.

Some recommendations can be made from this study. Firstly, literacy is an important predictor of flood damage. The data shows that literacy rate degrades significantly from S.S.C/ H.S.C to higher education phase. This issue requires further attention. The importance of male and female literacy need to be considered equally. Secondly, preparedness is not enough for flood damage mitigation. Initiatives should be taken to both educate the people about disaster management before, during and after a disaster, and enable them to convert this knowledge into effective flood handling actions. As seen from the correlation coefficients, flood awareness is not significantly correlated with disaster management knowledge and preparedness. Therefore, the people who have experienced floods are required to be treated with equal attention regarding disaster management knowledge transfer and preparedness improvement.

References

[1] Flood Forecasting and Warning Centre, Bangladesh Water Development Board, Annual flood report 2014, Tech. rep. (2014).

[2] Bangladesh Bureau of Statistics, Bangladesh Disaster Related Statistics 2015. Climate Change and Natural Disaster Perspectives, http://www.indiaenvironmentportal.org.in/files/file/Disaster_Climate_Statistics.pdf (2015).

[3] B. Merz, H. Kreibich, U. Lall, Multi-variate flood damage assessment: a tree-based data-mining approach, Natural Hazards and Earth System Sciences 13 (1) (2013) 53–64.

[4] A. H. Thieken, M. Müller, H. Kreibich, B. Merz, Flood damage and influencing factors: New insights from the august 2002 flood in germany, Water resources research 41 (12).

[5] G. Zhai, T. Fukuzono, S. Ikeda, Modeling flood damage: case of tokai flood 2000, JAWRA Journal of the American Water Resources Association 41 (1) (2005) 77–92.

[6] Bangladesh Bureau of Statistics, Statistical Year Book Bangladesh 2015, http://203.112.218.65:8008/WebTestApplication/userfiles/Image/SubjectMatterDataIndex/YearBook15.pdf (2015).

[7] Center for Excellence in Disaster Management & Humanitarian Assistance, Bangladesh Disaster Management Reference Handbook 2017, https://www.cfe-dmha.org/LinkClick.aspx?fileticket=p1nOVyZSxVg%3d&portalid=0 (2017).

[8] W. Kellens, T. Terpstra, P. De Maeyer, Perception and communication of flood risks: a systematic review of empirical research, Risk Analysis: An International Journal 33 (1) (2013) 24–49.

[9] P. Bubeck, W. Botzen, H. Kreibich, J. Aerts, et al., Long-term development and effectiveness of private flood mitigation measures: an analysis for the german part of the river rhine, Natural Hazards and Earth System Sciences 12 (2012) 3507–3518.

[10] J. K. Poussin, W. W. Botzen, J. C. Aerts, Factors of influence on flood damage mitigation behaviour by households, Environmental Science & Policy 40 (2014) 69–77.

[11] D. Wagenaar, J. d. Jong, L. M. Bouwer, Multi-variable flood damage modelling with limited data using supervised learning approaches, Natural Hazards and Earth System Sciences 17 (9) (2017) 1683–1696.

[12] T. Tingsanchali, M. F. Karim, Flood hazard and risk analysis in the southwest region of bangladesh, Hydrological Processes: An International Journal 19 (10) (2005) 2055–2069.

[13] A. M. Dewan, M. M. Islam, T. Kumamoto, M. Nishigaki, Evaluating flood hazard for land-use planning in greater dhaka of bangladesh using remote sensing and gis techniques, Water resources management 21 (9) (2007) 1601.

[14] E. Yang, P. A. Ray, C. M. Brown, A. F. Khalil, H. Y. Winston, et al., Estimation of flood damage functions for river basin planning: a case study in bangladesh, Natural Hazards 75 (3) (2015) 2773.

673 [15] K. Islam, The impacts of flooding and methods of assessment in urban
674 areas of bangladesh, Ph.D. thesis, Middlesex University (1997).

675 [16] A. Dasgupta, Floods and poverty traps: Evidence from bangladesh,
676 Economic and Political Weekly (2007) 3166–3171.

677 [17] A. H. Thieken, H. Kreibich, M. Müller, B. Merz, Coping with floods:
678 preparedness, response and recovery of flood-affected residents in ger-
679 many in 2002, Hydrological Sciences Journal 52 (5) (2007) 1016–1037.

680 [18] Bangladesh Bureau of Statistics, District Statistics 2011,
681 http://www.bbs.gov.bd/site/page/2888a55d-d686-4736-bad0-
682 54b70462afda/District-Statistics (2015).

683 [19] Bangladesh Water Development Board (BWDB), Processing and Flood
684 Forecasting Circle, Bangladesh Water Development Board, http://
685 www.hydrology.bwdb.gov.bd (2015).

686 [20] H. Akaike, Information theory and an extension of the maximum like-
687 lihood principle, in: Selected papers of hirotugu akaike, Springer, 1998,
688 pp. 199–213.

689 [21] J. B. Kadane, N. A. Lazar, Methods and criteria for model selection,
690 Journal of the American statistical Association 99 (465) (2004) 279–290.

691 [22] F. Mosteller, J. W. Tukey, Data analysis, including statistics, Handbook
692 of social psychology 2 (1968) 80–203.

693 [23] M. A. Poole, P. N. O'Farrell, The assumptions of the linear regression

694 model, Transactions of the Institute of British Geographers (1971) 145–
695 158.

696 [24] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

697 [25] G. De'Ath, Boosted trees for ecological modeling and prediction, Ecol-
698 ogy 88 (1) (2007) 243–251.

699 [26] F. Rosenblatt, The perceptron: A probabilistic model for information
700 storage and organization in the brain., Psychological review 65 (6) (1958)
701 386.

702 [27] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal repre-
703 sentations by error propagation, Tech. rep., California Univ San Diego
704 La Jolla Inst for Cognitive Science (1985).

705 [28] M. Masozera, M. Bailey, C. Kerchner, Distribution of impacts of natural
706 disasters across income groups: A case study of new orleans, Ecological
707 Economics 63 (2) (2007) 299–306.

708 [29] M. Pelling, What determines vulnerability to floods; a case study in
709 georgetown, guyana, Environment and Urbanization 9 (1) (1997) 203–
710 226.

711 [30] F. Messner, V. Meyer, Flood damage, vulnerability and risk perception–
712 challenges for flood damage research, Flood risk management: hazards,
713 vulnerability and mitigation measures (2006) 149–167.

714 [31] H. Kreibich, I. Seifert, A. H. Thieken, E. Lindquist, K. Wagner, B. Merz,

715 Recent changes in flood preparedness of private households and busi-
716 nesses in germany, Regional environmental change 11 (1) (2011) 59–71.

717 [32] F.-W. Chen, C.-W. Liu, Estimation of the spatial rainfall distribution
718 using inverse distance weighting (idw) in the middle of taiwan, Paddy
719 and Water Environment 10 (3) (2012) 209–222.

720 [33] D. Zimmerman, C. Pavlik, A. Ruggles, M. P. Armstrong, An experimen-
721 tal comparison of ordinary and universal kriging and inverse distance
722 weighting, Mathematical Geology 31 (4) (1999) 375–390.

723 [34] R. Dawson, L. Speight, J. Hall, S. Djordjevic, D. Savic, J. Leandro,
724 Attribution of flood risk in urban areas, Journal of Hydroinformatics
725 10 (4) (2008) 275–288.

726 [35] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley, 1977.

727 [36] J. Durbin, G. S. Watson, Testing for serial correlation in least squares
728 regression. ii, Biometrika 38 (1/2) (1951) 159–177.

729 [37] T. S. Breusch, A. R. Pagan, A simple test for heteroscedasticity and
730 random coefficient variation, Econometrica: Journal of the Econometric
731 Society (1979) 1287–1294.

732 [38] S. Chatterjee, A. S. Hadi, Regression analysis by example, John Wiley
733 & Sons, 2015.

734 [39] T. Plapp, Wahrnehmung von Risiken aus Naturkatastrophen: eine em-
735 pirische Untersuchung in sechs gefährdeten Gebieten Süd-und West-
736 deutschlands, Vol. 2, VVW GmbH, 2004.

[40] H. Kreibich, A. H. Thieken, T. Petrow, M. Müller, B. Merz, Flood loss reduction of private households due to building precautionary measures–lessons learned from the elbe flood in august 2002, Natural Hazards and Earth System Science 5 (1) (2005) 117–126.

[41] M. Islam, T. Hasan, M. Chowdhury, M. Rahaman, T. Tusher, Coping techniques of local people to flood and river erosion in char areas of bangladesh, Journal of Environmental Science and Natural Resources 5 (2) (2012) 251–261.