Analyzing the Impact of Forest Cover at River Bank on Flood Spread by using Predictive Analytics and Satellite Imagery

Muhammad Aneeq Yusuf[1], Dr. Muhammad Khalid Khan[2], Dr. Tariq Mahmood[3]
Muhammad Umer[4], Rafi Ullah Afridi[5]
College of Computing and Information Sciences
Karachi Institute of Economics and Technology, Karachi, Pakistan[1, 2, 4, 5]
Faculty of Computer Science, Institute of Business Administration, Karachi, Pakistan[3]

Abstract—**Floods have been a recurring problem for a number of countries around the world including Pakistan. It is believed that densely populated forests at river banks can prevent floods from spreading towards settlements and farmlands. The role of forest in flood spread has been an area of research for a while but the role of predictive modeling in this area is yet to be investigated in detail. In this study, we have used predictive analytics and satellite imagery to develop an environmental model that can predict the flood spread by considering forest cover at river bank and month of the year as parameters. We have used the satellite images of an area situated in the northern region of Pakistan i.e. Dera Ghazi Khan from the USGS's Land Sat program. These images comprised of a section of the Indus River and its adjoining areas. We want to analyze the forest bank at various section of the Indus River. We developed and trained our predictive model by using the satellite imagery data and tested it on a separate dataset to determine error percentage. The model showed significant promise and predicted the flood spread with an average accuracy of above 80%.**

Keywords—Floods; forests; predictive modeling; satellite imagery; environmental modeling

I. INTRODUCTION

Natural disasters are known to adversely affect population on a massive scale. It has been well documented that the impact of these disasters is more severe in developing countries, as compared to the better developed counterparts because of the disaster recovery services [1]. In such countries, large portion of the population lives below the poverty line and the impact of natural disasters is tremendous and everlasting [1]. Natural disasters can be of many types but in this study we are more focused on the impact of flood spread.

Floods have been a recurring problem for a number of countries around the world, particularly ones where farmers rely on irrigation for agriculture, such as Pakistan. Even though the country has a hot and dry climate, it receives ample rainfall during the months of July-August and December-January [2]. This often results in floods, as has been seen between the years 2003 and 2014. In a report by the Asian Development Bank, Pakistan is expected to experience increased variability in river flows. This is due to rise in

temperature and melting of glaciers. Moreover, the report also states that average annual rainfall is not expected to have any long-term trend [2]. This tends to create a high level of uncertainty for the country, with respect to floods. As such, Pakistan needs to take measures to prevent high-river flows from becoming disastrous.

The country has already experienced a devastating flood in 2010, which has been followed by a few smaller floods, in terms of impact [3]. The floods in 2010 alone were responsible for over 1,500 deaths and internal displacement of another 15 million people [3]. The main cause for this disaster was a few days of severe rainfall, at the end of July. It has been documented that this was mainly due to anomalous behavior, within the upper-atmospheric air. Moreover, it has been established that such atmospheric events, although rare, are a part of climatic events that occur over Pakistan [3]. However, this does not explain the floods that followed in subsequent years. As such, it would be correct to state that Pakistan is ill-equipped to deal with high-flows in its rivers, in times of heavy precipitation.

Furthermore, recently developed models of climatic conditions, particularly those of Pakistan modeled over a period of 30-years, have showed that the country is likely to experience wetter conditions. However, the coastal areas and the mountain ranges are expected to have a weak drying signal [3]. As the world moves towards a warmer climate, as a result of rising levels of greenhouse gases, the river-flows in the country are also expected to rise, as a result of faster glacial melting [4]. The importance of flood-risk assessment and flood prediction is of significant importance in such a hostile environment. The government has also been investing heavily in technology to help predict such disasters and provide better management skills, in case of such an event [4]. It is important to note here that other countries, with river systems, also face similar conditions [17].

It has been documented that the high-level of deforestation has been one of the major factors, contributing to the recurrence floods [1]. The country's population makes use of the wood for timber and fuel. However, the uncontrolled manner in which this is done, tends to greatly reduce the water retention capacity of the forests [1]. The forests which aren't dense enough fails to prevent the water from reaching the rural

populace. In addition to this, deforestation further leads to an increased surface water runoff and soil erosion, further clearing the way for the water to reach populated areas, in times of high-flows [1].

II.   APPROACH

By using various image processing and machine learning techniques, we can predict the flood spread. Here we introduce few techniques that are used in this work. Image masking technique is used to extract the area of interest from the satellite image (Fig. 1(a)).

This is done by identifying the start and end points of the area of interest, in the form of pixel-row and pixel-column number, while setting the values for all other pixels to 0 [5]. This tends to keep the pixel values of the area of interest intact, while neglecting the effect of everything else in the image. The technique is particularly useful for feature extraction and the processing of very large images. Next, the classification of image pixel is required and K-Nearest Neighbor Algorithm is used to achieve it.

K-Nearest Neighbor (Fig. 1(b)) classifies data sets based on their similarities with their neighbors and the k stands for number of data classes that are considered for classification [6]. This algorithm is based on supervised learning the basic idea of its working is:

- An object which is a new instance will be classified from majority votes given by its neighbors [7].

- It is measured by distant function so that the object is assigned to the most common class among its K nearest neighbors [7].
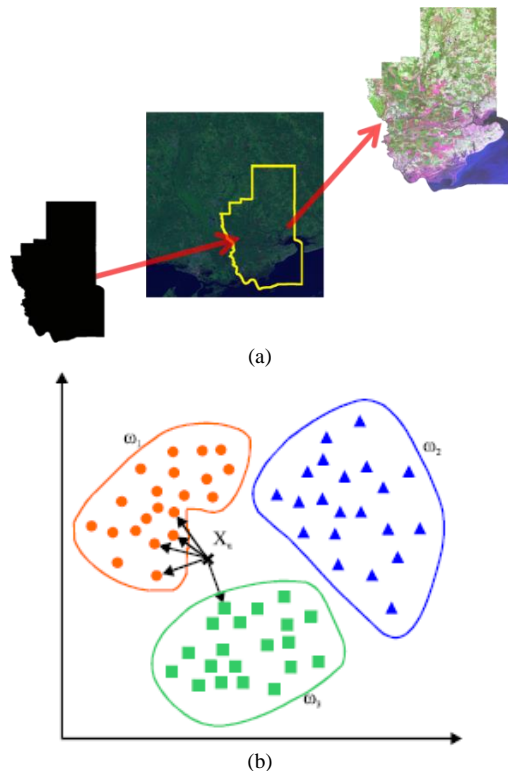

(a)

(b)

Fig. 1.   (a) Image Masking Visualization (b) KNN Visualization.

At the end, a model is developed by using predictive analytics techniques. Predictive Analytics can be defined as a set of processes that make use of data-mining and probability, to forecast outcomes [8]. This is done by identifying a set of predictors, also termed as known variables, and using them to develop a statistical model [8]. It should be noted here that once developed, the model can be extended and improved, once additional data, relating to the problem becomes available. Moreover, some models, excluding computational models, can be defined using a mathematical equation. This can then be used further to map out how exactly the predictors effect the outcome of a phenomenon, over time [8]. Computational models do not have a single defined equation, since they are developed using a number of different simulations, until a certain level of accuracy is reached.

It is clear that floods tend to be a devastating problem for Pakistan, recovery from which takes long-periods of time. It is also evident that various image processing and predictive analytics algorithms can help in analyzing the impact of forest cover at river banks. In this study, we have used predictive analytics and satellite imagery to develop a model that can predict the flood spread by considering forest cover at river bank and month of the year as parameters. We developed a model to predict how far the next flood would spread if given the area of the forest cover at river bank. In the subsequent section, we will present the related work and Section 4 explains the methodology in more detail. Sections 5 and 6, will discuss the results and conclusion respectively.

III.   RELATED WORK

Given Pakistan's history, with respect to floods, the natural disaster is well documented and extensively researched. However, the majority of these studies have been centered on determining the causes of the disaster, ways of better management, if and once the disaster strikes and determining the damage and patterns of floods in the past. Similarly, a number of researches have been conducted, focused on different approaches and machine learning methods for image processing and detecting and observing change in land-cover. A lot of this work is focused on detecting, quantifying and managing deforestation but currently, no known model exists, which makes use of forest cover around river bank to predict the risk of floods.

Author in [18], served as an example of how temporal changes, in land cover, are monitored using satellite imagery, from Landsat. The authors made use of imagery spread across a period of 40 years and performed a descriptive analysis of the changing land cover. However, in this case the researchers made use of cellular automata to develop a predictive model, for the Bakhtegan Lake. The authors clarified that the change in the ecology of the lake was mainly due to extreme climate change conditions. It should be noted here that, even though the imagery was taken between 1973 and 2013, only 12 images were used for the actual model development and analysis.

In [9], authors focused on assessing the usability and accuracy of developing linear and generalized linear mixed models (GLMM), using various machine learning techniques on freely available datasets. The study established that

although Artificial Neural Networks tend to outperform GLMM in certain areas, its accuracy suffers in other regions. On the other hand, Bayesian Networks were found to perform similarly across different test zones. Similarly in [10], assessment is made about the usability of clustering algorithms for the automatic detection of cultivable land, within a satellite image. The authors made use of a technique called 'Partitioning Around Medoids' for this purpose. The training and testing was done on single band images, with a particular focus on determining which bands are best suited to detect cultivable land, using unsupervised learning. It is important to point out here that the images used for our research were composed of images from the three visible bands (Red, Green and Blue), overlayed on each other, in a single image. However, the data was obtained from the Landsat 8 missions, similar to the ones used in this study. It is also important to note here that the authors identified the need for improved techniques for feature selection, in order to improve the accuracy of clustering algorithms.

When it comes to classification problems, machine learning is usually the de-facto approach used. However, accuracy of such classifications is crucial, in order to ensure that the results carry weight and are true in the specified situations. As a result, we made use of [19] to determine how accuracy of satellite imagery is performed. The authors made use of a data-assisted labeling approach, over unsupervised classified images. This was done by generated N-spectral class maps, using the ISODATA approach, followed by the development of maps with reference data and then assessment and selection of maps with the highest accuracy. It is important to mention here that the authors were able to report success with this approach and a similar approach was used in our own research, as well.

How ANNs are being successfully used in various image processing applications, many of which require a high-level of accuracy is discussed in [11]. The authors of this paper made use of the 'Sigmoidal function', for transfer and learning of the network. However, most of the application areas identified by the authors have access to a large amount of computational power. This was something that we lacked and therefore were not able to utilize ANNs, or any other form of Neural Networks, for our application. It should be noted here that the examples used in this paper do not identify the size of the images or number of pixels per image, which are processed in the mentioned applications. In our case, each of the Landsat images had approximately 56 million pixels, on average. As such, training a network using the computing power available was infeasible. Similarly, [21] was studied to determine the feasibility and benefits of using neural networks for satellite image classification. This paper was particularly selected due to the fact that the authors used Landsat imagery for testing purposes. The authors made use of a number of differently configured neural networks, in order to observe the changes of adding and removing nodes and layers of the network. The research resulted in the discovery of an optimal pairing, which was capable of delivering rapid convergence and better classification. The results were compared with a maximum likelihood classifier and the neural network was found to be equally good, when identifying urban areas. It is important to

note here that the authors made use of a section of the Landsat image. In our case, a machine learning technique was used, instead of neural networks, due to the processing requirements of classifying entire Landsat images, through a neural network.

A significant amount of work has also been done, during the last decade, to develop change detection models, for land cover mapping, using satellite imagery from MODIS and Landsat. It is important to note here that the spatial resolution of MODIS is $1km^2$, while that of Landsat is $30m^2$. In [12], the authors made use of the C5.0 algorithm (a variant of decision trees), to perform classification of regional land-covers across Central Asia, using time-series data from MODIS.

Furthermore, the authors analyzed data from the years 2001 and 2009, to observe seasonal features and how they change over-time. In addition to this, they also made use of high-resolution, remotely sensed imagery, to test their classifications. The results showed an accuracy of over 90%. In the following year, another study surfaced [13], where the authors used satellite imagery from Landsat 5 Thematic Mapper, to collect data from the Yellow River Delta, which is considered to be one of the fastest growing river deltas in the world. The research team made use of scenes from 1995, 2004 and 2010 to perform a comparative analysis of the area and detect any changes. Moreover, a per-pixel classification approach was employed using the C5.0 decision tree algorithm. By performing statistical analysis of the processed images, the research team was able to identify significant changes in the landscape of the delta. Moreover, the team was also able to determine intensified farming and urban sprawl as the main reasons for the said changes.

In [14], authors looked at the variations in stream flow and observed a number of records with anomalously high stream flow, which has resulted in flooding. It is important to note here that the hydrology of the region is dominated by snow accumulation and following melting processes, which result in times of high stream flow and low stream flow. The publication also pointed out that there was an imperfect understanding of the factors that contributed to the anomalies. The authors made use of the "Variable Infiltration Capacity" model to identify the key factors that contribute to this phenomenon, while also determining the influence of a set of predictors on these factors. The resulting model was then used to predict future stream flow in the basin. Similarly in [15], it is aimed to identify the synergies between the different types of changes in land-cover and how they affect the 'Catchment Scale'. Moreover, it was clarified that cross-disciplinary, long-term field studies are required to ensure major advances. It should be noted here that the research identifies forest management as a key-factor in drivers of flood-regime changes, which is also the main variable in our own prediction model. In addition to this, since we made use of a regression based approach in our research, [20] was reviewed to determine the outcome of using various types of regression techniques, to map landslide susceptibility. As is the case with most regression techniques, an inventory of the event occurrence and factors leading to that occurrence were used to develop the mapping models. The researchers identified the selection of events as an important factor in the accuracy of

the resultant models. A lot of emphasis, in this research, has been imposed on random sampling for model fitting. Following this, in our own research, we made use of sequential data, with both positive and negative points for flood events, to train our regression model.

In [16], authors focused on the use of large ensemble climate model simulations, in order to determine the causes of floods, in England. The authors were able to identify increased capacity for moisture in the atmosphere and anthropogenic warming as the main causes for increased westerly flow, in the Thames River. Combining these results with the catchment sensitivity results in a slight increase in properties, with respect to riverine flooding. However, this is also accompanied by a significant level of uncertainty. Furthermore, in [22] a group of researchers made an effort to improve the optimum-path forest clustering algorithm, by altering one of the characteristics of the main algorithm. Through this technique, the authors had aimed to make improvements to image segmentation techniques, for satellite imagery. It should be noted here that image segmentation is a key characteristic, when extracting features for target identification and object description.

It is evident that majority of the current work done is limited to observing the effects of floods and developing better mechanisms for flood management. However, significant progress is yet to be made, with respect to developing models for predicting future floods. Hence, predicting the flood spread by analyzing the forest cover at river bank is logical step in this direction.

IV. METHODOLOGY

The main objectives of this research is to convert information from satellite images into a quantifiable form and then use that data to develop and test a predictive model. We have used the satellite images of an area situated in the northern region of Pakistan i.e. Dera Ghazi Khan from the USGS's Land Sat program. These images comprised of a section of the Indus River and its adjoining areas as depicted in Fig. 2 but the problem is that these images covers lots of area which is of no interest to us and increase the processing burden. We used an image masking technique to extract our area of interest as depicted in Fig. 3. To get the pixel wise classification of the image, we used KNN algorithm and the output is depicted in Fig. 4. It should be noted here that in Fig. 4 the 'Water Class' is represented by blue, the 'Forest Class' has been represented by green and the 'Barren Land Class' has been represented by red.

The entire process was divided into a series of sequential steps, which are defined as follows; a graphical representation of the steps can be seen in Fig. 5:

- Feature Extraction: Once a satellite image is loaded into the system, the user identifies the X and Y coordinates of the training pixels, to be used for each of the three classes, namely, Water, Vegetation and Non-vegetation areas. This is done for each image.

- Image Slicing: Following the feature extraction step, the user would identify the limits of the X and Y

coordinates, for his area of interest (the river in question and its immediately surrounding areas). Using these values, only the area of interest is extracted from the overall image, significantly reducing the image size and number of pixels to classify.

- Pixel Classification: This step makes use of the values identified in the first step and trains the pixels extracted in the second step, color coding them into the three classes required. This is done using the K-Nearest Neighbors classifier and results in a uniformly colored image.

- Pixel Count: The output from the previous step is then passed to the Pixel Count module, which identifies the number of pixels in each class. The values are then multiplied by the spatial resolution of the satellite image (mentioned in section 1.4), to provide the area covered by each of the three classes, in meters. This results in the information from the satellite imagery to be available in a numerical form, which can now be modelled.

- Data Analysis and Predictive Modeling: Now that the data is available in a quantifiable form, it can be modeled using a regression model. The model is then tested for accuracy and used for future predictions.

It was discussed in the related work that neural networks are infeasible to be used with large sale images; we have decided to use machine learning technique to conduct pixel wise classification of the satellite images. Further, clustering also didn't work for large scale images on average computer systems; we decided to make use of supervised learning technique, rather than unsupervised learning to maintain a higher degree of accuracy and speed-up the learning process. The K-Nearest Neighbors algorithm was used, keeping in mind the sheer size of the satellite images and the computational power available. Each of the images was passed through the first four steps, mentioned above, with the numerical data being compiled in a tabular format; the resultant table can be seen below, Table I. It is important to note here that the feature extraction and classification was done on each of the images, so as to maintain a high-level of accuracy. As such, we were able to obtain over 85% classification accuracy, in almost every image.

Once the data was converted to a numerical form, it was then visualized graphically to determine the patterns and scatter in the data, the resultant plots can be seen in Fig. 6, 7, 8 and 9. The results of the visualizations are given in section 4. After careful analysis of the visualizations, it was decided to make use of the 'Ridge's Regression' model due to scatter of data.

It is important to state here that given the spread of the data points in the scatter plots, it can be inferred that a linear regression model would not be an ideal fit. This is mainly due to the fact that fitting a linear regression line to such a plot would mean greater variations in error bars. However, Ridge's Regression fits a curve that is closer to the data points, resulting in lower error bars and hence a more accurate predictive model.
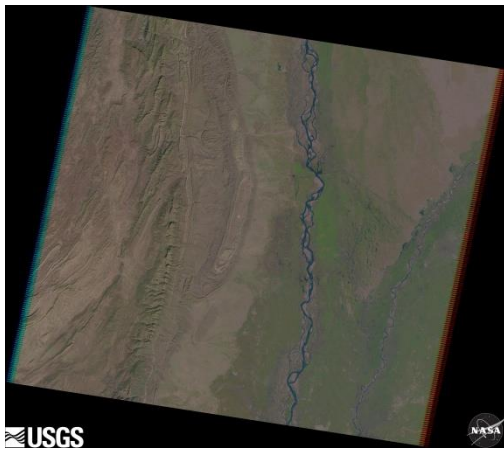
Fig. 2. Raw Satellite Image of North of Dera Ghazi Khan, Pakistan Showing River Indus and Associated Areas (Landsat).
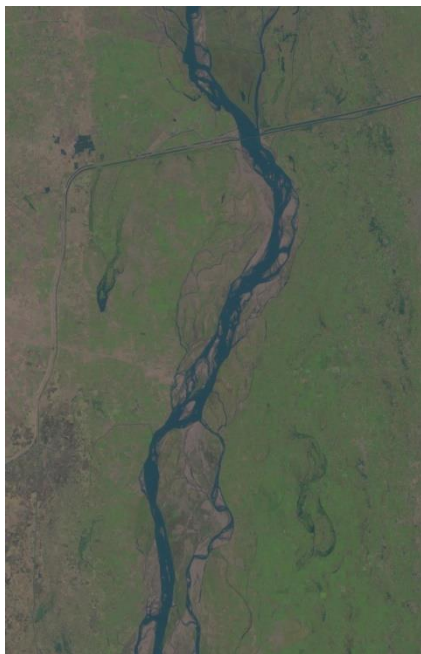

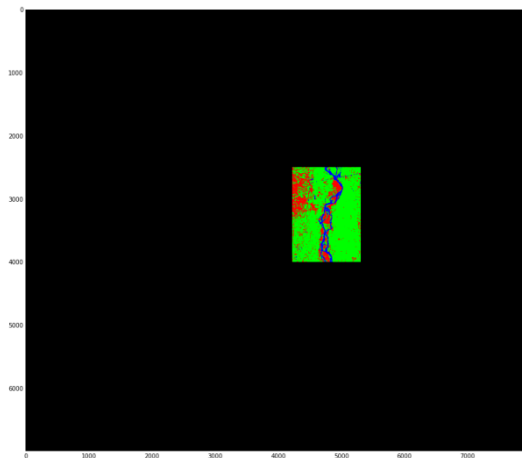Fig. 3. Extracted Area of Interest using Image Masking Technique.

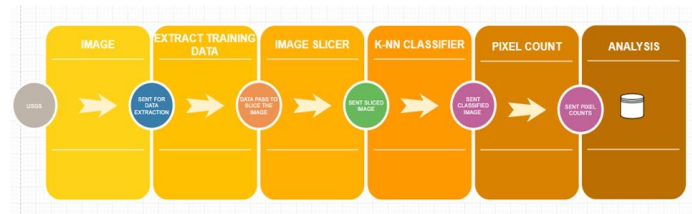Fig. 4. Image Classification using KNN Algorithm.

Fig. 5. Data Pipeline Diagram.

TABLE. I. DATA CREATED FROM SATELLITE IMAGES

	Date	Month	Soil Area (meters)	Vegetation Area (meters)	Water Area (meters)
1	2009-05-21	5	789138000	567756900	98784000
2	2009-06-06	6	787755600	538948800	128974500
3	2009-06-22	6	714393900	606996900	134288100
4	2009-07-08	7	414722700	887888700	153067500
5	2009-08-09	8	165952800	1142939700	146786400
6	2009-08-25	8	161583300	1096844400	197251200
7	2009-09-10	9	200644200	1138447800	116586900
8	2009-09-26	9	339753600	1019788200	96137100
9	2009-10-12	10	599943600	767756700	87978600
10	2009-10-28	10	739434600	643428000	72816300
11	2009-12-15	12	362314800	1043644500	49719600
12	2010-02-17	2	368703000	1025220600	61755300
13	2010-04-06	4	1193170500	205023600	57484800
14	2010-04-22	4	660550500	669238200	125890200
15	2010-05-24	5	268866000	1103776200	83036700
16	2010-06-09	6	1192885200	171452700	91341000
17	2010-07-11	7	625924800	557286300	272467800
18	2010-08-28	8	248742000	707883300	499053600
19	2010-11-16	11	267040800	1101443400	87194700
20	2010-12-02	12	244593000	1135185300	75900600
21	2011-01-19	1	345753900	1066625100	43299900
22	2011-02-20	2	591701400	753785100	110192400
23	2011-03-08	3	371706300	988533000	95439600
24	2011-03-24	3	738674100	637821900	79182900
25	2011-04-25	4	1260012600	110147400	85518900
26	2011-07-14	7	576993600	605738700	272946600
27	2011-10-02	10	354651300	979584300	121443300

In addition to this, a total of 27 data points are available to us for the training purpose as shown in Table I. During the analysis it was also revealed that forest cover around a river bank alone, would not be sufficient to accurately predict when and by how much a flood is expected to occur. This was mainly due to the fact that rivers do not flood their banks, unless there is heavy rainfall or extreme glacial melting at its source. As such a second independent variable was identified, in the form of "Estimated Rainfall". However, rainfall data for our specific area wasn't readily available. We made use of the "Month", when the image was taken, to serve as a supplemental variable for "Estimated Rainfall". This was mainly to accommodate seasonal variations, as closely as possible.

A total of three models were developed, two uni-variable regression models and one multivariate regression model. The aim here was to compare the three models and select the most optimal one for future use. Moreover, this allowed for determining the overall effect each variable has on the final model. Finally, another 6 data points were created, using the same techniques as identified earlier, for testing purposes. The accuracy of each of the three models was determined and averaged-out for better comparison. Table II shows the end results of this analysis and testing phase.

A. Training Data Blue (Water) and Green (Vegetation)

The histogram in Fig. 6 shows a graphical representation of the data extracted from the processing images, from the training set. There was an abundance of vegetation in the concerned area, as compared to water. However, during the exploratory data analysis, it was seen that readings no. 16 and 17 show an increase in water area, whilst there was a decline in vegetation around the river bank. However, this cannot be taken as a causation effect and further analysis and modeling is necessary.

B. Testing Data Blue (Water) and Green (Vegetation)

A similar histogram was developed for the test dataset and the results are shown in Fig. 7. As can be seen, despite the decline in vegetation, in readings no. 3 and 4, there was not a significant change in the area of water. This is why it was necessary to add another variable, in the form of months. It should be noted here that this was done to include the effect of seasonal change.
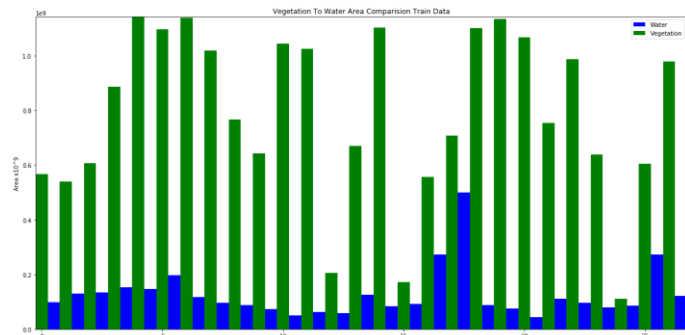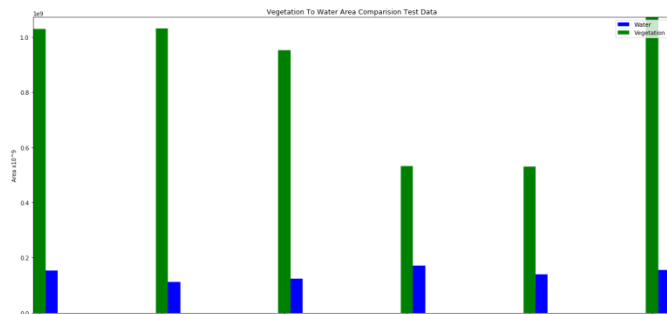

Fig. 7.   Test Data Visualization.

C. Scatter Plot (Water - Vegetation)

In order to further analyze the data, a scatter plot (Fig. 8) of the dependent and independent variables was also visualized. It was revealed that the data was skewed towards the y-axis, with just three outliers. This meant that there were just three instances in the data which indicated flood conditions in the region. However, there was significant variation in vegetation.

D. Scatter Plot (Water - Month)

Much like the plot above, the scatter plot of months against water area (Fig. 9) showed little variability. It is important to note here that for the purpose of analysis the water area in similar months was averaged out; hence there are just 12 data points, instead of 27, for the training data. Further analysis of the plot revealed that seasonal variations do seem to contribute to the phenomenon and are mostly active between June and September.
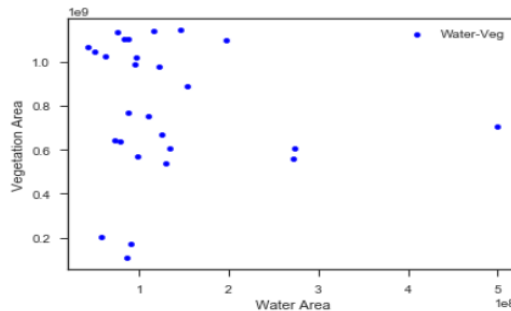

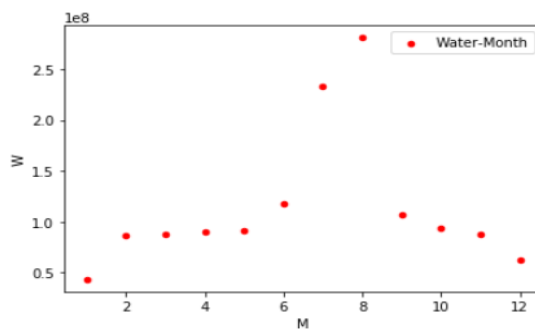Fig. 8.   Scatter Plot Vegetation Area against Water Area.

Fig. 9.   Scatter Plot Water Area against Month.

Fig. 6.   Train Data Visualization.

V. RESULTS

The following table (Table II) highlights the results of each of the six test cases. The percentage accuracy of each model, for every data point can be seen. As is visible, there was significant variation in the predicted and actual values of the test points, in all of the three models. The final row contains the percentage average error of each of the three models. When averaged out, the multiple-ridge-regression model was the most accurate one. However, to justify the model, an R-Squared test was also conducted on each of the three models. The 'Ridge Vegetation' model had an R-Squared score of 0.0082, as compared to 0.0180 for 'Ridge Month' and 0.0223 for 'Ridge Multivariable' models. Although the R-Squared values were quite low for each of the three models, it can largely be attributed to the small number of data points available.

Moreover, it was also the model with the least errors for individual tests, with the exception of reading no. 1. It can also be concluded from the given table, that month alone cannot be used as a predictor variable, since it has a high level of variation, in percentage error. Moreover, it can also be concluded that forest cover alone, can be used as a predictor variable, but it will contain a high degree of error. Therefore, a secondary variable needs to be added to make the model more accurate.

TABLE. II. PERCENTAGE ERROR IN PREDICTIVE MODELS; RV=RIDGE VEGETATION, RM=RIDGE MONTH, MVM=MULTIVARIATE MODEL

Date	Vegetation Area (1000's)	Water Area (1000's)	% Error (RV)	% Error (RM)	% Error (MVM)
23/08/14	10301163	153527	21.23%	13.35%	30.30%
24/09/14	1032350	111824	8.09%	22.53%	3.43%
10/10/14	952204	123246	0.14%	14.41%	6.40%
07/06/15	532970	171257	21.42%	26.97%	17.99%
26/08/15	530257	139664	3.60%	4.75%	0.56%
27/09/15	1072591	156467	23.46%	12.43%	10.24%
Mean Error:			12.99%	15.74%	11.49%

VI. CONCLUSIONS

From the previously stated results and careful analysis, it can be concluded that the "Multiple-Regression-Model" does show a lot of promise for accurately predicting how far a flood is expected to spread, given the amount of forest cover on the river bank and seasonal conditions. It is important to note here that there have not been any significant flood conditions in the region, following the floods in 2010. As such, it is yet to be tested if the model would hold true, in such a condition.

However, given that the number of data points used is quite small, further improvements can be made with the model, by fitting it over a larger dataset. It should be noted here that with the amount of computing power available to us, it would have been difficult to create more datapoints in a quick and efficient way. As such, datapoints from a 3-year period was used to perform model fitting, which could have resulted in over-fitting of the data. It is recommended that datapoints from a 10-year period be used for the creation of a more accurate model.

Moreover, given the budget constraints, it was not possible to make use of high-resolution satellite imagery, which led to the use of intuitions, rather than concrete data. This could have also resulted in over or even under-fitting of the data. Added to this, the constraints also made it difficult to obtain the complete rainfall data for the said 3-year period. It is recommended that satellite imagery with a spatial resolution of at least 10x10 meters be used for improving the model. In addition to this, the use of rainfall data, instead of months, would help improve the quality of the results.

REFERENCES

[1] Sayeeda Amber Sayed, Pedro Arcos González. (2014). Flood disaster profile of Pakistan: A review. Science Journal of Public Health. 2 (3), 144-149.

[2] Asian Development Bank (2017). Climate Change Profile of Pakistan. Chaudry, Qamar Uz Zaman.

[3] G. van der Schrier, L. M. Rasmijn, J. Barkmeijer, A. Sterl, W. Hazeleger. (2018). The 2010 Pakistan floods in a future climate. Climatic Change. 148, 205-218.

[4] Muhammad Iqbal Javed Akhter, Muhammad Irfan, Naeem Shahzad, Rehan Ullah. (2017). Community Based Flood Risk Reduction: A Study of 2010 Floods in Pakistan. American Journal of Social Science Research. 3 (6), 35-42.

[5] "Area/mask processing methods", Trucco, Chapter 3.

[6] Tavish Srivastava. (2014). Introduction To K-Nearest Neighbors. Available: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/. Last accessed 11th Nov 2017.

[7] David Cournapeau. (2007). Nearest Neighbors. Available: http://scikit-learn.org/stable/modules/neighbors.html. Last accessed 2nd Oct 2017.

[8] Math Works. Create models and forecast future outcomes. Available: https://www.mathworks.com/discovery/predictive-modeling.html. Last accessed August 15, 2018.

[9] Helen Mayfield, Carl Smith, Marcus Gallagher, Marc Hockings. (2017). Use of freely available datasets and machine learning methods in predicting deforestation. Environmental Modelling & Software. 87, 17-28.

[10] R.B. Arango, A.M. Campos, E.F. Combarro, E.R. Canas, I. Díaz. (2016). Mapping cultivable land from satellite imagery with clustering algorithms. International Journal of Applied Earth Observation and Geoinformation. 49, 99-106.

[11] Alexandrina-Elena Pandelea, Mihai Budescu, Gabriela Covatariu. (2015). Image processing using artificial neural networks. Universitatea Tehnică „Gheorghe Asachi" din Iaşi. 61 (65).

[12] Igor Klein, Ursula Gessner, Claudia Kuenzer. (2012). Regional land cover mapping and change detection in Central Asia using MODIS time-series. Applied Geography. 35 (-), 1-16.

[13] Marco Ottinger, Claudia Kuenzer, Gaohuan Liu, Shaoqiang Wang, Stefan Dech. (2013). Monitoring land cover dynamics in the Yellow River Delta from 1995 to 2010 based on Landsat 5 TM. Applied Geography. 44, 53-68.

[14] Charles L. Curry, Francis W. Zwiers. (2018). Examining controls on peak annual streamflow and floods in the Fraser River Basin of British Columbia. Hydrology and Earth System Sciences. 22, 2285-2309.

[15] M. Rogger, F. Gallart, M. Agnoletti, A. Alaoui, J.C. Bathurst, G. Bodner, M. Borga, V. Chaplot, G. Glatzel, J. Hall, J. Holden, L. Holko, R. Horn, A. Kiss. (2017). Land-use change impacts on floods at the catchment scale – Challenges and opportunities for future research. Water Resources Research. 53 (7), 5209-5219.

[16] Nathalie Schaller, Alison L. Kay, Rob Lamb, Neil R. Massey, Geert Jan van Oldenborgh, Friederike E. L. Otto, Sarah N. Sparrow, Robert Vautard, Pascal Yiou, Ian Ashpo. (2016). Human influence on climate in

the 2014 southern England winter floods and their impacts. Nature Climate Change. 6, 627-634.

[17] Top River Flood countries accessed at https://www.wri.org/blog/ 2015/03/world-s-15-countries-most-people-exposed-river-floods.

[18] Taghi Jokar Arsanjani, Reza Javidan, Mohamad Jafar Nazemosadat, Jamal Jokar Arsanjani, Eric Vaz. (2015). Spatiotemporal monitoring of Bakhtegan Lake's areal fluctuations and an exploration of its future status by applying a cellular automata model. Computers & Geosciences. 78 (-), 37-43.

[19] Ruili Lang, Guofan Shao, Bryan C.Pijanowski, Richard L.Farnsworth. (2008). Optimizing unsupervised classifications of remotely sensed imagery with a data-assisted labeling approach. Computers & Geosciences. 34 (12), 1877-1885.

[20] Arzu Erener, A. AbdullahSivas, A. Sevtap Selcuk-Kestel, H. Sebnem Düzgün. (2017). Analysis of training sample selection strategies for regression-based quantitative landslide susceptibility mapping methods. Computers & Geosciences. 104 (-), 62-74.

[21] C.H.Jarvis, N.Stuart. (1996). The sensitivity of a neural network for classifying remotely sensed imagery. Computers & Geosciences. 22 (9), 959-967.

[22] Siya Chen, Tieli Sun, Fengqin Yang, Hongguang Sun, Yu Guan. (2018). An improved optimum-path forest clustering algorithm for remote sensing image segmentation. Computers & Geosciences. 112 (-), 38-46.