Ikbal_2021_Prediction of flood areas using the logistic regression method

by Suroso Suroso

Submission date: 28-Mar-2023 04:42PM (UTC+0700)
Submission ID: 2048922458
File name: dayat_Jati_2019_Prediction_of_flood_areas_using_the_logistic.pdf (559.48K)
Word count: 4247
Character count: 21993

Prediction of flood areas using the logistic regression method (case study of the provinces Banten, DKI Jakarta, and West Java)

Muhammad Ikbal Hidayat Jati[1], Suroso[2], Purwanto Bekti Santoso[3]

[1]Department of Civil Engineering, Faculty of Engineering, Jenderal Soedirman University, Purwokerto, Indonesia
[2]Department of Civil Engineering, Faculty of Engineering, Jenderal Soedirman University, Purwokerto, Indonesia
[3]Department of Civil Engineering, Faculty of Engineering, Jenderal Soedirman University, Purwokerto, Indonesia

E-mail: surosotsipil@unsoed.ac.id

Abstract. Disasters that occur in Indonesia continue to increase from year to year. The National Disaster Management Agency (BNPB) noted that throughout 2017 there had been 2,862 disasters. Of these, almost 99 percent are hydrometeorological disasters, namely disasters that are affected by weather and surface flow. The details of the disaster include floods (979), tornadoes (886), landslides (848), forest and land fires (96), drought (19), earthquakes (20), tidal waves and abrasions (11), dam volcanic eruptions (3). The western part of Java Island includes three provinces namely West Java Province, DKI Jakarta Province, and Banten Province. Of the three provinces, they are no stranger to hearing floods, especially in the capital city of Jakarta and West Java. Flood problems until now have not been resolved completely even the problem of floods is the tendency to increase both in terms of intensity, frequency, and distribution due to climate change. Based on the above conditions, it is necessary to do a study that can provide information about the main causal factors and predict areas that are likely to experience floods. In achieving these objectives, this study uses a mathematical model of logistic regression analysis and application of Geographic Information Systems (GIS) conducted based on the variables that cause flooding, namely rainfall, topography, slope, flow accumulation, land use, and distance to the nearest river. The modelling results obtained an accuracy rate for predicting flood disasters in the study area using logistic regression which was between 85.05% - 94.39%.
Keywords: Flood, Logistic Regression, Flood Parameters, GIS

1. Introduction

Indonesia is an archipelago that is geographically located in an area prone to natural disasters. Disasters that occur in Indonesia continue to increase from year to year [2]. The National Disaster Management Agency [9] noted that throughout 2017 there had been 2,862 disasters. Of these, almost 99 percent are hydrometeorological disasters, namely disasters that are affected by weather and surface flow. The details of the disaster include floods (979 events), tornadoes (886 events), landslides (848 events), forest and land fires (96 events), drought (19 events), earthquakes (20 events), tidal waves and abrasions (11 events), dam volcanic eruptions (3 events).

Flood is defined as an event of inundation in flat areas around the river as a result of overflowing river water that is not able to accommodate the river flow. In addition, flooding is the interaction between humans and nature and the natural system itself. Flood disaster is an aspect of the interaction between humans and nature that arises from a process where humans try to use beneficial nature and avoid nature that harms humans [4].

The most important factor in increasing the intensity of floods is global warming which results in extreme climate change [5]. This climate change causes a striking change, namely the longer duration of the dry season while the duration of the wet season is getting shorter. This results in higher rainfall intensity, the higher the intensity of the rain, the higher the possibility of a flood that will occur [6].

Flood problems until now have not been resolved completely even the problem of floods is the tendency to increase both in terms of intensity, frequency, and distribution due to climate change. One of the efforts to cope with the flood hazard is to develop a model that can predict flood-prone areas by exploiting the application of Remote Sensing and Geographic Information Systems (GIS). This can reduce flood risk due to flood disasters. This study aims to develop a model of flood prediction using the logistic regression method.

2. Data and study location

2.1. Study location

In this study, the specified location is in the western region of Java island which includes three provinces, namely Banten, the Special Capital Region of Jakarta, and West Java Province in the range of coordinates 5 ° 50 '- 7 ° 50' LS and 104 ° 48 '- 108 ° 48 'BT. Among the three provinces, they are no stranger to hearing floods, especially in the capital city of Jakarta. Flooding in Jakarta is a routine flood, which means that every rainy season flood occurs in the area. West Java Province is one of the regions with high potential for floods. The distribution of 2017 disaster data shows that West Java Province is a region that has quite a lot of flood disasters [9].

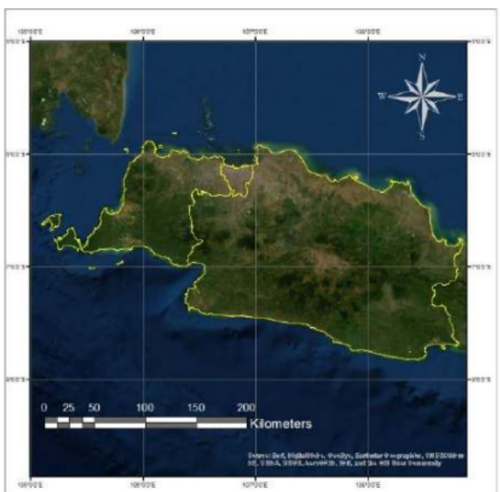

Figure 1. Map of the study area

2.2. Data collection
Data used in this study are collected from both official authorities and satellite based data including historical flooded area data, Digital Elevation Model (DEM), TMPA (TRMM-based Multi-satellite Precipitation Analysis), and MODIS (Moderate Resolution Imaging Spectrometer) as can bee seen in Table 1. The data is processed with the help of ArcGIS, Microsoft Excel applications, as well as R program.

Table 1. Data layer of Study Area.

Parameter	**Sub-Classification**	**Data Type**	**Scale**
Historical Flooded Area	Flood Extent	Point	-
DEM	Topography	Grid	30 x 30
	Slope	Grid	30 x 30
	Flow Accumulation	Grid	30 x 30
TMPA	Precipitation	Grid	0.25^{O} x 0.25^{O}
MODIS	Land Cover	Grid	463 x 463
Stream Map	*Distance from the Nearest Stream*	Grid	15 x 15

2.2.1. History of flood. The historical flood data is a basic reference to calibrate a model developed in this study. The data include specific time and locations where flood events have occurred and recorded by BNPB as can be shown in Figure 2 below. Based on the results of identification that has been carried out, it can be seen that there are 250 total points of historical flood events. Of these, there were 46 floods in Banten Province, 51 floods in DKI Jakarta Province, and 154 floods in West Java Province.

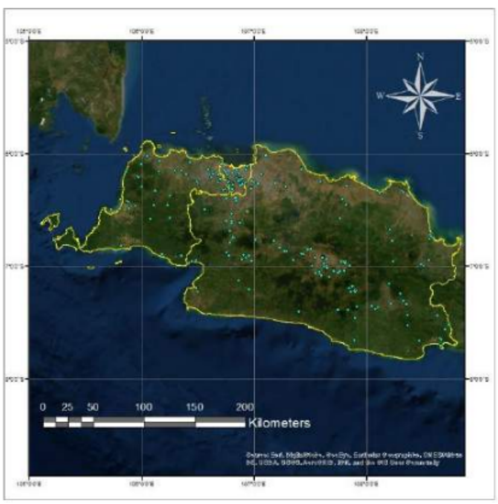

Figure 2. Distribution of Flood Points

Journal of Physics: Conference Series **1367** (2019) 012087 doi:10.1088/1742-6596/1367/1/012087

By analyzing this flood event data we can find out the intensity of flood events in each region. The results of these three provinces show that the areas recorded as having the highest flood events were the Bandung area (29 flood events), Bandung City (21 flood events), East Jakarta City (21 flood events), South Jakarta City (18 flood events), Garut (14 flood events), Lebak (13 flood events), and Sukabumi (12 flood events). For more details, the map of the intensity of flood events can be seen in Figure 3.

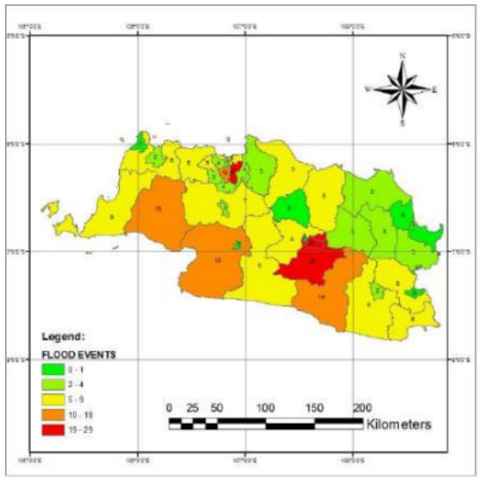

Figure 3. Historical Flood Map of the Study Area

In addition to flood event data, we also need event data not to flood so that in the modelling process we can evaluate whether our modelling is in accordance with the results in the field (in this case the occurrence of flooding and not flooding). This data is taken randomly in several places which are predicted to have a small possibility for the occurrence of floods, one of which is in the highlands, so that these points are taken randomly in areas that have a high elevation, amounting to 108 points so that there are 358 pieces coordinate points, for more details see in Figure 4.

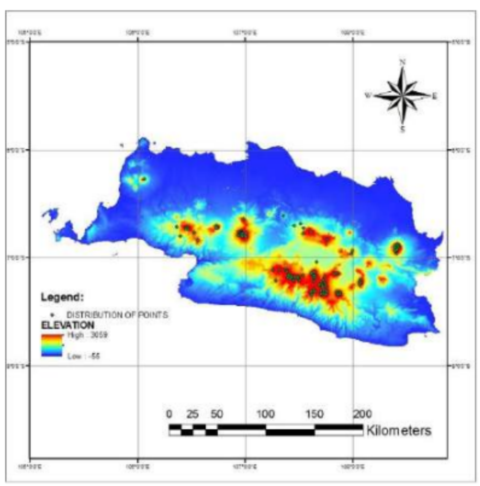

Figure 4 Distribution of Points are not flooded

Journal of Physics: Conference Series **1367** (2019) 012087 doi:10.1088/1742-6596/1367/1/012087

2.2.2. Digital elevation model (DEM). Digital elevation model (DEM) is one of the basic data in this study. The DEM used is derived from the Shuttle Radar Topography Mission (SRTM) data that belongs to the United States National Aeronautics and Space Administration (NASA). This DEM is used to find out some information, namely elevation (Figure 5), slope (Figure 7), and flow accumulation (Figure 9).

According to Figure 6 proves that flood disasters often occur in areas that have low topographic elevations or downstream areas. This is evidenced by 53.2% of the data on flood events recorded at low altitudes ranging from -1 m to 199 masl. From this data shows that areas that have low altitude have a greater likelihood of flooding, so there is a need for special handling of floods in these areas.

Based on Figure 8, it was found that 84.4% of the flood events occurred on low slopes or could be said to be flat areas. This proves that areas including flat or sloping slopes have a high probability of occurrence of floods.

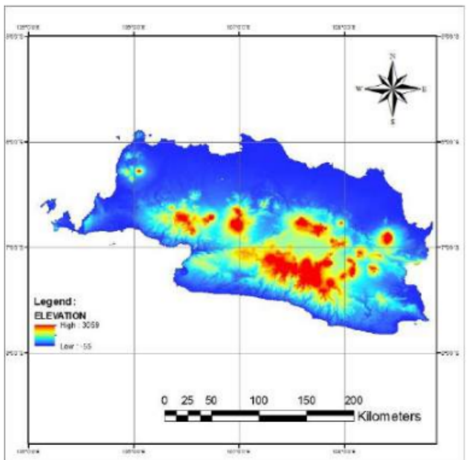

Figure 5. The relationship between elevations and flood events

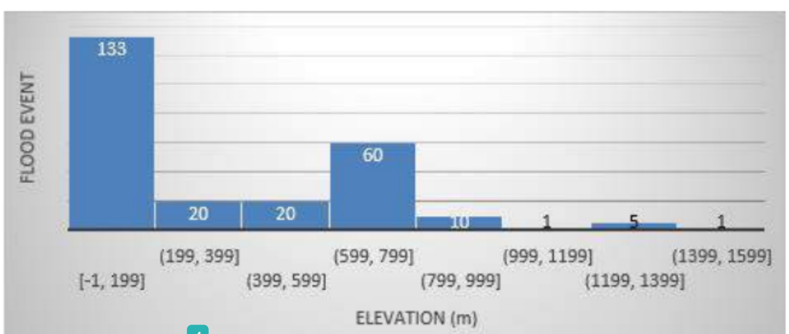

Figure 6. The relationship between elevations and flood events

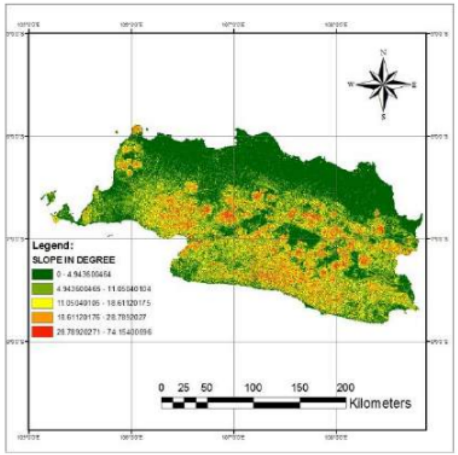

Figure 7. Slope Map

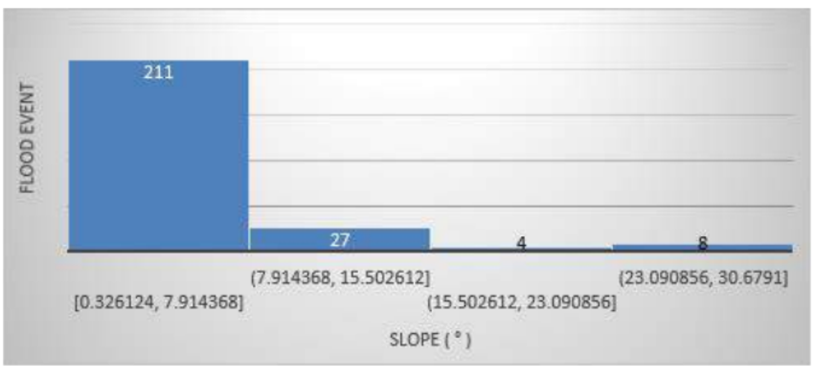

Figure 8. Relationship between Slope and Flood Event

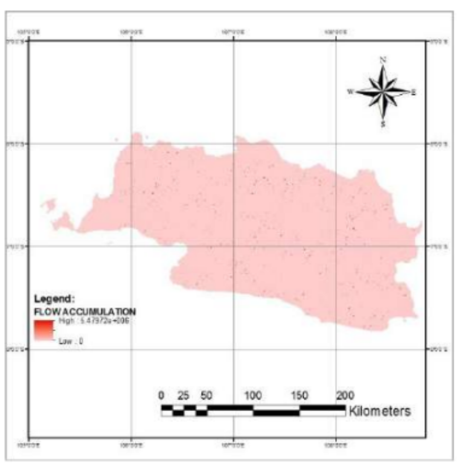

Figure 9. Flow Accumulation Map

2.2.3. Precipitation. Historical flood events are adjusted to the time where rainfall occurs. The rainfall data sets used in this study derived from TMPA (TRMM-based Multi-satellite Precipitation Analysis) which is a raw product from TRMM (Tropical Rainfall Measuring Mission) [8]. As an example, Figure 10 shows spatial precipitation values derived from TMPA as grids over the study area.

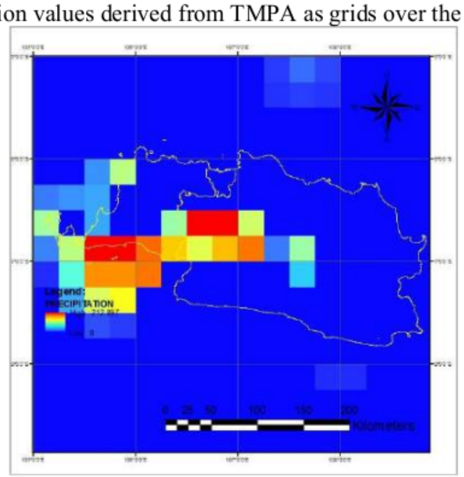

Figure 10. Precipitation amount on 23rd September 2018.

2.2.4. Land use land cover (LULC). MODIS (the Moderate Resolution Imaging Spectrometer) is a key instrument of the Terra and Aqua satellites [3]. Time-series MODIS-NDVI data have been proven to be useful for cover type characterization [7]. This land cover data from MODIS can be used to identify the type of vegetation and land cover on the surface of the earth, especially in the scope of the study, namely Banten, DKI Jakarta, and West Java. The MODIS data used in this study is the annual land use data in HDF-EOS format with ranges from 2001 to 2017. Figure 11 shows the MODIS land use in 2017.

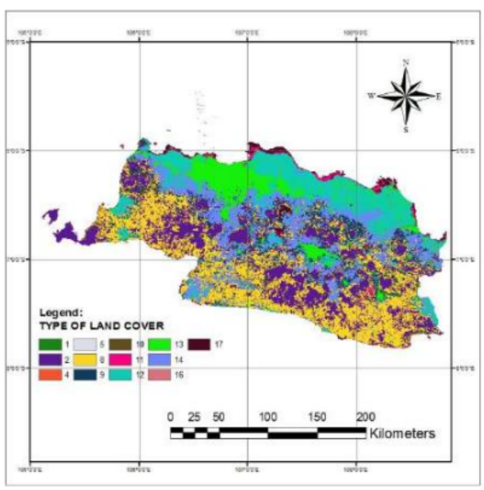

Figure 11. LULC in the study area in 2017

Land cover contributes significantly to the run-off. Thus, it is necessary to determine coefficient runoff based on the land cover. The Table 2 below is used to calculate coefficient runoff using the type of MODIS land cover data.

Table 2. Coefficient Runoff.

Class	Class Name	Description	Coefficient Runoff
1	Evergreen needleleaf forest	Lands dominated by needleleaf woody vegetation with a percent cover >60% and height exceeding 2 m. Almost all trees remain green all year. Canopy is never without green foliage.	0.1 – 0.6
2	Evergreen broadleaf forest	Lands dominated by broadleaf woody vegetation with a percent cover >60% and height exceeding 2 m. Almost all trees and shrubs remain green year round. Canopy is never without green foliage.	0.1 – 0.6
3	Deciduous needleleaf forest	Lands dominated by woody vegetation with a percent cover >60% and height exceeding 2 m. Consists of seasonal needleleaf tree communities with an annual cycle of leaf-on and leaf-off periods.	0.1 – 0.6
4	Deciduous broadleaf forest	Lands dominated by woody vegetation with a percent cover >60% and height exceeding 2 m. Consists of broadleaf tree communities with an annual cycle of leaf-on and leaf-off periods.	0.1 – 0.6
5	Mixed forests	Lands dominated by trees with a percent cover >60% and height exceeding 2 m. Consists of tree communities with interspersed mixtures or mosaics of the other four forest types. None of the forest types exceeds 60% of landscape.	0.1 – 0.6
6	Closed shrublands	Lands with woody vegetation less than 2 m tall and with shrub canopy cover >60%. The shrub foliage can be either evergreen or deciduous	0.1 – 0.4
7	Open shrublands	Lands with woody vegetation less than 2 m tall and with shrub canopy cover between 10% and 60%. The shrub foliage can be either evergreen or deciduous.	0.1 – 0.4
8	Woody savannas	Lands with herbaceous and other understory systems, and with forest canopy cover between 30% and 60%. The forest cover height exceeds 2 m.	0.1 – 0.6
9	Savannas	Lands with herbaceous and other understory systems, and with forest canopy cover between 10% and 30%. The forest cover height exceeds 2 m.	0.3 – 0.6

Class	Class Name	Description	Coefficient Runoff
10	Grasslands	Lands with herbaceous types of cover. Tree and shrub cover is less than 10%.	0.4 – 0.6
11	Permanent wetlands	Lands with a permanent mixture of water and herbaceous or woody vegetation. The vegetation can be present either in salt, brackish, or fresh water.	0.5 – 0.85
12	Croplands	Lands covered with temporary crops followed by harvest and a bare soil period (e.g., single and multiple cropping systems). Note that perennial woody crops will be classified as the appropriate forest or shrub land cover type.	0.1 – 0.25
13	Urban and built-up lands	Land covered by buildings and other man-made structures.	0.3 – 0.7
14	Cropland/natural vegetation mosaics	Lands with a mosaic of croplands, forests, shrubland, and grasslands in which no one component comprises more than 60% of the landscape.	0.1 – 0.25
15	Snow and ice	Lands under snow/ice cover throughout the year.	1
16	Barren	Lands with exposed soil, sand, rocks, or snow and never have more than 10% vegetated cover during any time of the year.	0.3 – 0.6
17	Water bodies	Oceans, seas, lakes, reservoirs, and rivers. Can be either fresh or salt- water bodies.	0

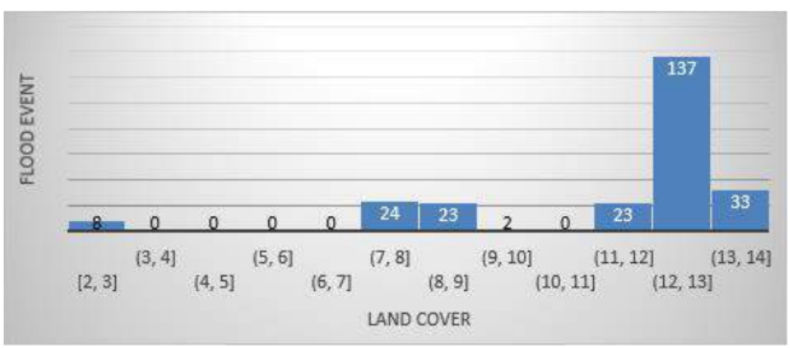

Figure 12. Graph of Relationship between LULC and Flood Event

According to the graph above, it shows that 54.8% of the total flood incidence data is found in the 13th type of land cover, where the type of land cover is a residential area and a built area. This proves that changes in land use affect the possibility of flooding. So, the increasing function conversion of land, especially from vegetated areas into built-up areas, will result in an increase in the probability or possibility of floods. This is because rainwater that falls to the surface will be difficult to absorb into the soil or called infiltration, so most of the volume of water going down will overflow the surface

2.2.5. River network data. The river network data used in this study is collected from The Geospatial Information Agency (Figure 13). This data is used to determine the closest distance the location of the flood events that have occurred from the nearest river. The distances are saved in the grid format as can be seen in Figure 14.

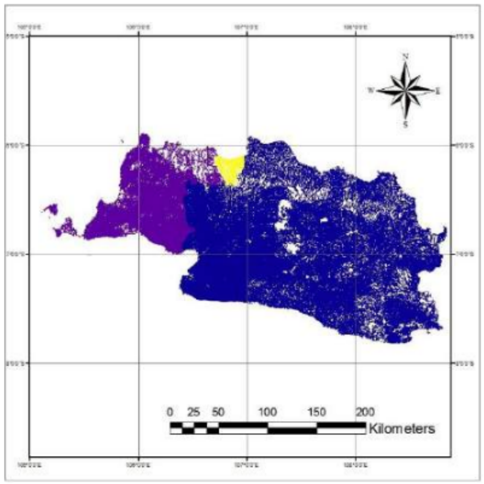

Figure 13. Stream Map of The Study Area

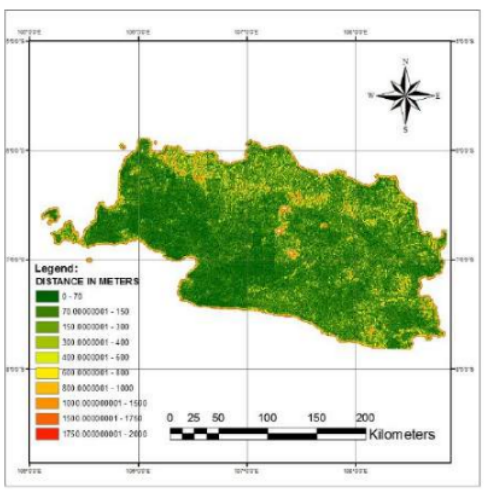

Figure 14. DNS Map

From the results of the analysis of the distance to the river, it was found to be 50.4% of the flood events that occurred at relatively close distances of 0 to 86 meters. This proves that the proximity of a region to a river can affect the possibility of a flood occurring. So, the closer a region is to a river, the greater the likelihood of a flood occurring. For this reason, areas that are close to the river or drainage need special handling to deal with flood management

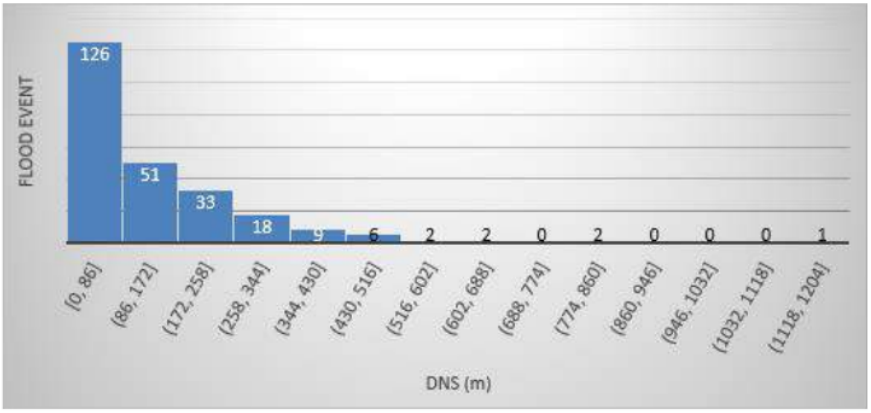

Figure 15. The relationship between DNS and flood event

3. Logistic regression model

3.1. Basic theory

Logistic regression analysis is a regression method used to find the relationship between categorical response variables with nominal, ordinal scale with one or more continuous and categorical explanatory variables. Logistic regression is a basic classification method that is used to find the relationship of the response variable (y) that is dichotomous and polycotomous with a predictor variable (x) that is polycotomous. The response variable (y) from binary logistic regression consists of two categories namely "success" and "failure", where the notation from y = 1 for the "success" category and y = 0 for the "fail" category. So that the response variable y follows the Bernoulli distribution for every single observation. The probability functions for each observation are as follows:

$$f(y) = \pi^y (1-\pi)^{1-y} \;; y = 0, 1 \tag{3.1}$$

Where if y = 0 then f (y) = 1 – π and y = 1 then f (y) = π, so the logistic regression function is obtained as follows:

$$f(z) = \frac{e^z}{1+e^z}; z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{3.2}$$

where n is many predictor variables. The value f (z) is between 0 and 1 for each given z value, because the value of z itself is between -∞ and ∞. The logistic regression model actually describes a probability of an object, in this case is a flood event. The logistic regression model is as follows:

$$\pi(x) = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}} \tag{3.3}$$

The function π (x) is a nonlinear function so it needs to be transformed by using logit transformations to obtain linear functions so that the relationship between the response variable y and the predictor variable can be seen.

$$\begin{aligned} g(x) &= logit\pi(x) \\ &= log\frac{\pi(x)}{1-\pi(x)} \\ &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \\ &= \alpha + \sum_{i=1}^{n} \beta_i x_i \\ &= X\beta \end{aligned} \tag{3.4}$$

In this study, flood disaster data was used as input of non-independent variables in logistic regression analysis. Thus the resulting flood hazard map will be representative for the region. This is because the determination of the weight value is based on the existing flood event data and does not use the conventional method where the weight has been previously determined.

The flood event data used for this modelling is 70% of the total number of random flood events. This aims to ensure that the coefficients of the equation can be continuous with the actual flood event data that has occurred.

This study uses computational assistance to model the variables. For this modelling using the Rstudio programming language. To run this program must already have data for each distribution of flood events that have been treated previously such as rainfall data, elevation, slope, runoff coefficient, and the closest distance to the river.

The data used in this modelling is the data distribution of flood coordinate points and data on coordinate points that have a small probability of flooding. The points are 250 flood points and 108 points that have a small chance of flooding.

Data that already contains flood parameters are stored in the Comma Separated Values (CSV) format. This data contains parameters that are used as modelling that has been adjusted to the coordinates and time of the flood event. This data is used as input data for modelling using logistic regression. The data used is 70% of the total data totaling 358, which is 251 data for this modelling. This data is randomly selected using the help of the Rstudio program, then the data is used for logistic regression modelling. The output of this modelling is the coefficient data from the parameters that we specified earlier. Table 3 below is the output of logistic regression programming.

3.2. Developing model

To find out the performance of this modelling, it is necessary to repeat the modelling. This modelling is carried out repeatedly 500 times. After completing the model repeatedly, we can find out the coefficients of the parameters of the flood event. The coefficients are sorted from the smallest to the largest. This aims to determine the uncertainty of modelling using logistic regression. Of the 500 models that have been sorted, they take 5% as the lower limit, 50% as the middle limit, and 95% as the upper limit.

Table 3. Coefficient of Logistic Regression to flooded areas

Parameter	**Coefficient of Logistic Regression**		
	5%	**50%**	**95%**
(Interception)	5.38328627721561	8.92077080302076	14.885488352785
Precipitation	-0.00858129471319676	0.0156654046354833	0.0450427170912787
Elevation	-0.00720598863261365	-0.00529532261788892	-0.00416033344926504
Slope	-0.302567439717362	-0.16851258016166	-0.0945001391429277
Coefficient Runoff	-9.09726512880789	-1.27434994746956	4.36197681012406
DNS	-0.00804182415433407	-0.00277991447925938	0.00263320067939322
Flow Accumulation	$3.15997403560425e^{-06}$	0.000130278891448994	0.00158313585774173

3.3. Validation Model

As a model validation, the data used for validation is 30% of the total data that has not been used for modelling, namely 107 pieces of data. This is done to prove whether the modelling that has been done is in accordance with what happened in the field, in this case is a flood.

To find out the performance or reliability in the calibration model, measurements were taken by looking at the correlation value. After completing the calculation, a comparison is made to see the success of the model that has been made so that the results of this modelling can be known to have good reliability or not. From the results of the comparison of the reliability level, the corresponding data results were 84.06542% for the lower limit (5%), 90.65421% for the middle limit (50%), and 94.39252% for the upper limit (95%). So, the level of accuracy for predicting floods in the study area using logistic regression is between 85.05% - 94.39%.

Table 4. Value of Validation.

	Value of Validation	
	False	True
5%	14.953271 %	85.04673 %
50%	9.345794 %	90.65421 %
95%	5.607477 %	94.39252 %

4. Conclusions

Based on the results of the research that has been done, it can be concluded several things as follows:

a. Logistic regression methods can predict flood events
b. The level of accuracy for predicting floods in the study area using logistic regression is between 85.05% - 94.39%
c. Low rainfall does not rule out the possibility to avoid floods. So, high rainfall is not the only benchmark for the occurrence of floods but from other aspects must also be taken into account
d. Areas that have low altitude have a greater likelihood of flooding, so there is a need for special handling of floods in these areas
e. Areas that include flat or sloping slope have a high probability of flooding
f. Increasing land use change, especially from vegetated areas to built areas, will result in the increased probability or the possibility of flooding
g. The closer a region is to a river the greater the likelihood of a flood occurring

References

[1] Arbain A A 2011 *Detection of landslide-prone areas using geospatial data and satellite-based geographic information system* (case study of Banten, DKI Jakarta, dan West Java provinces) Indonesia University Indonesia

[2] Buchori I, Sugiri A, Mussadun M, Wadley D, Liu Y, Pramitasari A, and Pamungkas I T 2018 *A predictive model to assess spatial planning in addressing hydro-meteorological hazards: a case study of Semarang City* Indonesia *International journal of disaster risk reduction* **27** 415-426

[3] Duveiller G, Hooker J, and Cescatti A 2018 *A dataset mapping the potential biophysical effects of vegetation cover change Scientific data* **5** 180014

[4] Haghizadeh A, Siahkamari S, Haghiabi A H, and Rahmati, O 2017 *Forecasting flood-prone areas using Shannon's entropy model Journal of Earth System Science* **126(3)** 39

[5] Li X, Wang X, and Babovic V 2018 *Analysis of variability and trends of precipitation extremes in Singapore during 1980–2013 International Journal of Climatology* **38(1)** 125-141

[6] Li X, Meshgi A, and Babovic V 2016 *Spatio- temporal variation of wet and dry spell characteristics of tropical precipitation in Singapore and its association with ENSO International Journal of Climatology* **36(15)** 4831-4846

[7] Shao Y, Lunetta R S, Wheeler B, Iiames J S, and Campbell J B 2016 *An evaluation of time-series smoothing algorithms for land-cover classifications using MODIS-NDVI multi-temporal data Remote Sensing of Environment* **174** 258-265

[8] Tan M L, Chua V P, Tan K C, and Brindha K 2018 *Evaluation of TMPA 3B43 and NCEP-CFSR precipitation products in drought monitoring over Singapore International journal of remote sensing* **39(8)** 2089-2104

[9] Rosyida A, Nurmalasari R, and Suprapto 2017 *Disaster Data in 2017* Book of The National Agency for Disaster Management (BNPB) **6**

Ikbal_2021_Prediction of flood areas using the logistic regression method

ORIGINALITY REPORT

9%	1%	8%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	J. Y. LIU. "Land-cover classification of China: integrated analysis of AVHRR imagery and geophysical data", International Journal of Remote Sensing, 1/20/2003 Publication	3%
2	Dodi Sudiana, Ardhi Adhary Arbain. "Geographic Information System (GIS)-based landslide susceptible area detection using geospatial and satellite data", TENCON 2011 - 2011 IEEE Region 10 Conference, 2011 Publication	2%
3	F. Hannerz. "Assessment of remotely sensed and statistical inventories of African agricultural fields", International Journal of Remote Sensing, 3/3/2008 Publication	1%
4	Anang Wahyu Sejati, Imam Buchori, Nattaya Mlatti Lakshita, I.G. Andika Wiratmaja, Desyta Ulfiana. "A 50-years flood frequency in Coastal area of Java, Indonesia: Integrating	1%

historical hotspot and Spatial Autocorrelation Analysis", Research Square Platform LLC, 2022
Publication

5	Bitu Babu Shreevastav, Krishna Raj Tiwari, Ram Asheshwar Mandal, Bikram Singh. ""“Flood risk modeling in southern Bagmati corridor, Nepal” (a study from Sarlahi and Rautahat, Nepal)", Progress in Disaster Science, 2022 Publication	<1 %
6	Yang Shao, Ross S. Lunetta, Brandon Wheeler, John S. Iiames, James B. Campbell. "An evaluation of time-series smoothing algorithms for land-cover classifications using MODIS-NDVI multi-temporal data", Remote Sensing of Environment, 2016 Publication	<1 %
7	www.scirp.org Internet Source	<1 %
8	Y. Pan. "An integrative classification of vegetation in China based on NOAA AVHRR and vegetation-climate indices of the Holdridge life zone", International Journal of Remote Sensing, 3/10/2003 Publication	<1 %
9	pingpdf.com Internet Source	<1 %

10	Anandharuban Panchanathan, Ali Torabi Haghighi, Mourad Oussalah. "A multi-criteria approach for improving streamflow prediction in a rapidly urbanizing data scarce catchment", Research Square Platform LLC, 2022 Publication	<1 %
11	Andrea Pelizzari, Ricardo Armas Goncalves, Mario Caetano. "Chapter 12 Information Extraction for Forest Fires Management", Springer Science and Business Media LLC, 2008 Publication	<1 %
12	Stefania Marzo, Silvia Natale, Stefano De Pascale. "Language attitudes among mobile speakers", John Benjamins Publishing Company, 2021 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		