RGNZ/MGPB

The Performance of Bivariate Statistical Models in Landslide Susceptibility Mapping (Case Study: Cisangkuy Sub-Watershed, Bandung, Indonesia)

Rudarsko-geološko-naftni zbornik
(The Mining-Geology-Petroleum Engineering Bulletin)
UDC: 551.3:551.5
DOI: 10.17794/rgn.2024.5.2

Original scientific paper

Sukristiyanti[1]; Pamela[2]; Sitarani Safitri[3]; Ahmad Luthfi Hadiyanto[4]; Adrin Tohari[5]; Imam Achmad Sadisun[6]

[1] National Research and Innovation Agency (BRIN), Research Center for Geological Disaster, Sangkuriang, Dago, Bandung, Indonesia 40135. ORCID 0000-0002-1945-1878

[2] Geological Agency of Indonesia (BGL), Diponegoro No.57, Cihaur Geulis, Kec. Cibeunying Kaler, Bandung, Indonesia, 40122. ORCID 0000-0002-9598-2187

[3] National Research and Innovation Agency (BRIN), Research Center for Geoinformatics, Sangkuriang, Dago, Bandung, Indonesia 40135. ORCID 0000-0003-4821-6479

[4] National Research and Innovation Agency (BRIN), Research Center for Geoinformatics, Sangkuriang, Dago, Bandung, Indonesia 40135. ORCID 0000-0001-8943-470X

[5] National Research and Innovation Agency (BRIN), Research Center for Geological Disaster, Sangkuriang, Dago, Bandung, Indonesia 40135. ORCID 0000-0002-6662-0697

[6] Bandung Institute of Technology (ITB), Faculty of Earth Sciences and Technology, Ganesha 10 Bandung, Indonesia 40132. ORCID 0000-0001-7808-1427

Abstract

Landslide occurrences are common in hilly and mountainous areas, especially in tropical countries with high rainfall and intensive weathering. Landslide susceptibility mapping (LSM) is an initial effort to mitigate landslide hazards. This research conducted a comparative study of four LSM maps, namely frequency ratio (FR), information value model (IVM), weight of evidence (WoE), and Shannon entropy (SE), for the Cisangkuy Sub-watershed, West Java. Those models determine the relationship between the landslide density and the causative factors. The model utilized 76 landslide pixels and 15 causative factors. 70% of the landslides were used as training data, and the remaining was used for validation. The 15 factors were selected from 27 causative factors. The highly correlated causative factors were removed to address multicollinearity. In addition, only causal factors related to landslide data are involved in the modelling. The receiver operating characteristics (ROC) curve and the landslide density index (LDI) method were used for model validation. All models indicate appropriate prediction rates for FR, IVM, WoE, and SE, which are 0.770, 0.790, 0.793, and 0.788, respectively. Based on the LDI analysis, the LDI values did not increase gradually from very low to very high susceptibility classes for each LSM map. However, the maps are still favorable because the classes that are most susceptible in all models have the highest LDI. The performance of the models may be influenced by the number of classes and classification methods used to categorize each continuous parameter, as well as the small quantity of landslide inventory data.

Keywords:
landslide susceptibility modelling; bivariate statistical method; ROC curve; landslide density index

1. Introduction

Landslides are rock mass, debris, or earth material that move down a slope due to gravity. Landslides are not only limited to slide types but also falls and flows (**Sassa, 2007**). Landslide susceptibility mapping (LSM) can be a first step in disaster mitigation and supporting spatial planning. Expert-driven and data-driven models can be utilized to accomplish this objective. The expert-driven model depends on expert judgment, which each expert might use to form their own opinion, which is somewhat subjective (**Barredo et al., 2000**).

In contrast, the data-driven models are quantitative models based on the spatial relationship between landslide occurrences and landslide causative factors (**Pradhan, 2010**). A data-driven model is not subjective, but it is very responsive to the quality of the data (**Lee and Talib, 2005**), either landslide inventory or causative factors. **Nohani et al. (2019)** suggested that the identification of landslide locations should rely on high-resolution satellite imagery. However, the availability of such images is still very limited, especially in developing countries (**Batar and Watanabe, 2021**). Regarding a number of causative factors for LSM, there is currently no consensus about the quantity of the factors (**Juliev et al., 2019; Melati et al., 2024**). The primary variables contributing to landslides include geology, topography, hy-

Corresponding author: Sukristiyanti
e-mail address: sukris.tiyanti@gmail.com

drology, and anthropogenic factors. Consequently, all potential maps that contained these factors were included for modelling. Despite their typically lower scale, geological maps remained to be utilized as a data source. For instance, **Pourghasemi et al. (2014)** employed a geology map with a scale of 1:100,000, while other data sources had a resolution of 1:25,000. In addition, **Juliev et al., (2019)** used a geological map with a scale of 1:500,000, while the digital elevation model (DEM) they used had a spatial resolution of 30 m. Fortunately, the small scale of their geological maps did not affect the result.

Data-driven models consist of statistical and machine learning-based methods (**Lee and Talib, 2005; Reichenbach et al., 2018; Yao et al., 2008**). This study did not employ machine learning; bivariate statistical techniques were used instead. Bivariate statistical methods were developed earlier than machine learning. While machine learning, especially artificial neural networks (ANN), is viewed as a "black box" with a difficult-to-understand procedure (**Umar et al., 2014**), bivariate statistical methods can explicitly explain the relationship between landslides and their causative factors. Although machine learning models can handle a wide range of data sizes, they require two types of data responses: landslide and non-landslide. Landslide data shows the distribution of locations recorded as experiencing landslides, while landslide data requires a method to obtain it, and there is no standard procedure. The quality of non-landslide data itself is questionable (**Zhu et al., 2018**). Regardless of the approach used to acquire non-landslide data, it is essential to have access to information on landslide data. The reason is that non-landslide data is not located in the same area as landslide data. As the quality of landslide data worsens, the quality of non-landslide data also gets worse.

This research aims to apply four bivariate statistical methods, namely frequency ratio (FR), (2) information value model (IVM) or statistical index (SI), (3) weight of evidence (WoE), and (4) Shannon entropy (SE) or index of entropy (IoE) to study the relationship between landslide and causative factors and to produce the landslide susceptibility maps (LSM maps) in the Cisangkuy Subwatershed. The research area is predominantly characterized by hills and mountains composed of tuff, a type of rock that is susceptible to landslides. Besides, the study area has a critical infrastructure, i.e. a Geothermal Power Station and a number of tourist attractions. Therefore, it is crucial to conduct LSM in this area. As the initial phase of landslide hazard mitigation, LSM should provide crucial information to facilitate decision-making for spatial planning, thereby significantly minimizing the potential damage caused by landslides. LSM maps are created with the purpose of assisting individuals in identifying and implementing measures to mitigate the risks associated with landslides (**Rasyid et al., 2016**).

Comparative studies can be carried out using these four models. Some of those models were used and compared in the previous studies, but based on the literature review, no comparative study has been found between the four comparable models. The three models within the four models were employed by **Nohani et al. (2019)**, **Liu and Duan (2018)**, and **Razavizadeh et al. (2017)**. **Nohani et al. (2019)** examined the FR, SE, and WoE models, **Liu and Duan (2018)** conducted a comparison study between IV, SE, and WoE, while **Razavizadeh et al. (2017)** compared the FR, IV, and WoE. Meanwhile, other similar studies used one of those models (**Alfarabi et al., 2019; Arifianti et al., 2023; Sadisun et al., 2021**).

By conducting this study, it is possible to assess the accuracy of LSM maps produced from bivariate statistical models and ascertain their reliability. Furthermore, in order to improve maps that need to be produced, it is necessary to collect information on the specific locations that require concern for enhancement. It is necessary to investigate the effectiveness of bivariate statistical models for LSM due to their simplicity and ability to explain the relationship between causative factors and landslides explicitly. This article comprises several sections, including an introduction, material and methods, results, discussion, and conclusion.

2. Materials and methods

2.1. Study area

The Cisangkuy Sub-watershed is one of the sub-watersheds in the Bandung basin, located in Bandung Regency, West Java, Indonesia. The study area is 308.14 km², with an elevation range of 648 m to 2334 m and a slope that ranges from 0 to 184.6%. It is located between 6.976º – 7.232ºS and 107.482º – 107.656ºE. The study area is surrounded by volcanic mountains (**Abidin et al., 2013**) and experiences intensive landslides. **Figure 1** shows the research area.

Two large landslides occurred in this area. The first occurred in the complex of a vital object, viz. the Wayang-Windu Geothermal Power Station, Pangalengan, on May 5, 2015. The affected area was 700 m long and 200 m wide, hitting geothermal pipelines and buried settlements. This landslide caused seven deaths, two missing persons, and ten injuries. The landslide was triggered by a long rainfall period with a total of five months of rainfall from January 1 to May 5, 2015, of about 1,200 mm (**Rahardjo et al., 2017**). Another big landslide occurred on the bank of the Cibintinu River in Arjari on October 6, 2016. The landslide caused 400 people to be evacuated due to the landslide material blocking the Cibintinu River flow. The Cibintinu landslide also damaged some agricultural lands. In the last five years, there have also been frequent landslides; for example, the landslide that occurred in Sukaluyu village, Pangalengan subdistrict, in 2022 caused one house to be damaged and one person died, and most recently (January 8, 2024) there was a flash flood due to a landslide in Margamulya village,

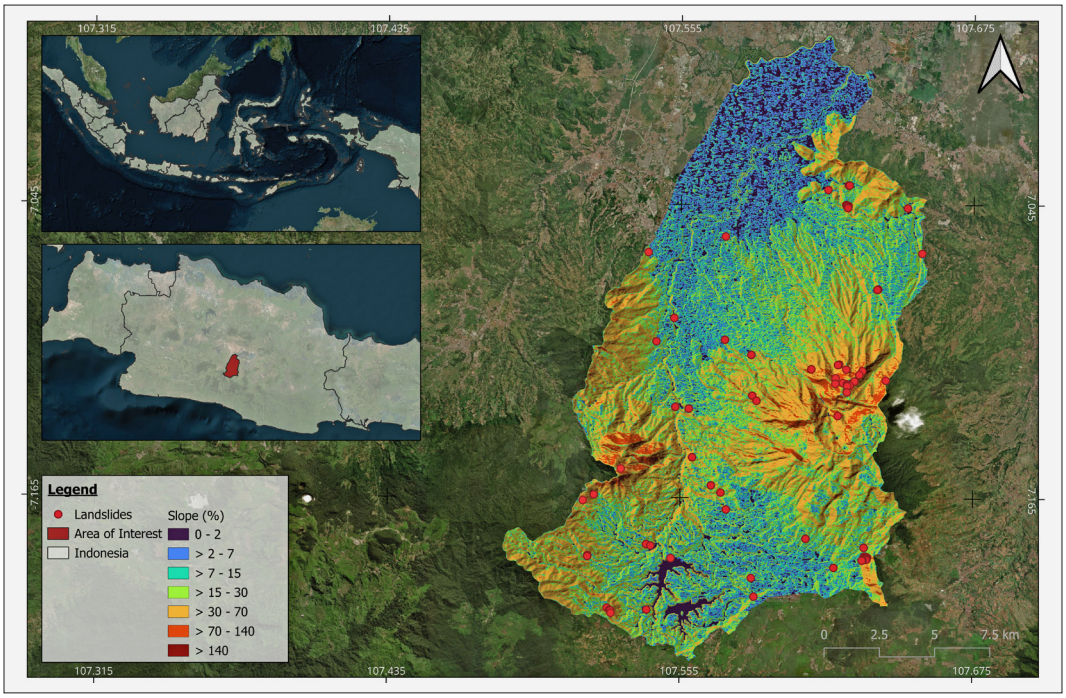

Figure 1: The Cisangkuy Sub-watershed with its landslide locations and inset maps

subdistrict Pangalengan which caused 19 residents' houses to be submerged and the 15-meter retaining wall around the village office to collapse.

The lithology of the research area is dominated by tuff, laharic breccia, and lava. Meanwhile, the geomorphology is dominated by young volcano cones. Most of the study areas were formed by volcanic activities except the plain area, which was formed by fluvial-lacustrine deposits. The precipitation ranges from 1868 to 2558 mm/year. The highest annual rainfall is in the southeastern part of the study area, namely the Bedil and Wayang Mountains. Generally, rainfall in the eastern part is higher than in the western part. Two large landslides occurred at the geothermal power plant on Mount Bedil and in Arjasari on Mount Pipisan in the east, which affected 10.8 ha and 4.53 ha, respectively.

2.2. *Data source*

2.2.1. Landslide inventory data

Landslide inventory data is mandatory data for LSM using bivariate statical methods, which are represented in a landslide distribution map (**Arifianti et al., 2023**). The landslide inventory data in this study were created using various data sources, including high-spatial-resolution Google Earth images obtained between 2010 and 2019, official landslide records from 2013 to 2020, and several field observation data from 2017 to 2020.

2.2.2. Landslide causative factors

Landslide causative factors are predisposing elements that result from direct and indirect environmental and human influences and ultimately define the trigger factor for landslides (**Arifianti et al., 2023**). This study created four groups (geological, topographical, hydrological, and anthropological consisting) consisting of 27 landslide causative factors. The main consideration used to determine the causative factors is the availability of a data source over the entire study area. There is a lack of consensus regarding the factors responsible for landslides or the precise number that should be selected to create LSM maps (**Melati et al., 2024**). However, not all of the provided variables will be utilized as input in the modelling process, but they do need to be selected. It is recommended to avoid using variables that have a strong correlation (**Evans, 1996**).

The 27 factors include (1) lithology, (2) distance to faults, (3) fault density, (4) distance to lineaments, (5) lineament density, (6) peak ground (PGA), (7) elevation, (8) slope, (9) topographic position index (TPI), (10) terrain ruggedness index (TRI), (11) hypsometric integral (HI), (12) relative relief (RR), (13) geomorphology, (14) aspect, (15) slope curvature, (16) plan curvature, (17) profile curvature, (18) flow direction, (19) stream power index (SPI), (20) topographic wetness index (TWI), (21) distance to rivers, (22) river density, (23) rainfall, (24) land cover, (25) normalized difference vegetation index (NDVI), (26) distance to roads, and (27) road density.

Some of those landslide causative factors are index maps, i.e. TPI, TRI, SPI, TWI, and NDVI. TPI has been identified as a contributing factor for landslides in earlier research conducted by **Othman et al. (2018)**, **Pourghasemi et al. (2014)**, and **Shahzad et al. (2022)**. In their LSM study, **Sadisun et al. (2021)** employed the usage of three indices: TRI, SPI, and TWI. On the other hand,

Table 1: Sources of data for preparing landslide causative factors

Sources		Landslide causative factors
Maps or imageries	**Scale or cell size**	
Topographic map from the Geospatial Information Agency (BIG)	1: 25,000	(1) distance to rivers, (2) river density, (3) distance to roads, (4) road density
Geological map from the Geological Agency	1: 50,000	(5) lithology, (6) distance to faults, (7) fault density
Geomorphological map from the Geological Agency	-	(8) geomorphology
DEMNAS (DEM from BIG)	8.34 m	(9) elevation, (10) slope, (11) TPI, (12) TRI, (13) HI, (14) RR, (15) aspect, (16) slope curvature, (17) plan curvature, (18) profile curvature, (19) flow direction, (20) SPI, (21) TWI
Pan-sharpened Landsat 8 imagery: the RGB 564 bands	15 m	(22) distance to lineaments, (23) lineament density
Sentinel 2A: bands 2, 3, 4, and 8	10 m	(24) land cover, (25) NDVI
Tabular rainfall data from Meteorological Agency (BMKG)	-	(26) rainfall
PGA map from the Geological Agency	-	(27) PGA

Melati et al. (2024) utilized TWI and NDVI in their research. In addition to utilizing TPI, **Shahzad et al. (2022)** utilized TWI and NDVI. NDVI was derived from Sentinel 2A satellite imagery, while the four index maps were produced using DEM. Each corresponds to topographical, hydrological, or anthropogenic aspects that could potentially impact the occurrence of landslides. All parameters were proceeded from the available topographical and geological maps, DEM, and remote sensing imageries, which have different formats and scales (see **Table 1**).

2.3. Methods

This study has four main steps for LSM, including (1) data preparation for obtaining landslide inventory data and causative factors, (2) parameter selection, (3) bivariate statistics analysis, and (4) model validation. All those steps can be seen in **Figure 2**.

2.3.1. Data preparation

Landslide objects are interpreted visually using Google Earth images (**Pham et al., 2018**; **Wati et al., 2010**). This study involved the source area of landslides since the source area is the part that experiences collapse (**Schlögel et al., 2018**). The use of the source area of landslides can increase the accuracy of landslide susceptibility modelling (**Schlögel et al., 2018**).

The polygons representing the source area of major landslides (>100 m^2) were created through visual interpretation, as described by **Sukristiyanti et al. (2021)**. However, small landslides (<100 m^2) could not be identified on Google Earth images. We collected small landslides by analyzing the official landslide report and conducting field surveys. Every landslide polygon was converted into raster data with an 8.34 m cell size, enabling one polygon to be represented by one or more landslide pixels. Meanwhile, each small landslide that was originally depicted as a point feature is then converted into a single landslide pixel. The cell size of 8.34 m refers to the spatial resolution of the DEM, and other data also resampled to the size. We selectively utilized representative landslide pixels from big landslide polygons rather than including all pixels from landslide polygons as our data input. If the polygon's area is smaller than 50 m x 50 m, it is denoted by a single pixel. If the polygon is more than that size, it is denoted by two or more pixels, depending on its area.

All landslide pixels were randomly partitioned into two groups: 70% of the pixels were assigned as training data, and the remaining pixels were allocated as test data (**Chen et al., 2017**) by utilizing a subset feature tool in ArcGIS. It was ensured that the training and test data were not sourced from the same landslide polygon to prevent biased outcomes and potentially inaccurate model validation. The training data is data used to build the model, and the test data is data that is not used to build a model but instead used to measure the predictive rate (**Umar et al., 2014**).

In addition to compiling landslide inventory data, the causative factors of landslides are also prepared. All causative factors were prepared in raster maps with a pixel size of 8.34 m × 8.34 m, referring to the spatial resolution of the DEM. The generated LSM maps are considered medium-scale due to the utilization of medium-scale maps, such as geological maps, in the modelling process. The geological map provided several landslide causative factors.

The study utilized 27 landslide causative factors, which are categorized into four groups: geological, topographical, hydrological, and anthropological. Five geological factors were obtained from a geological map,

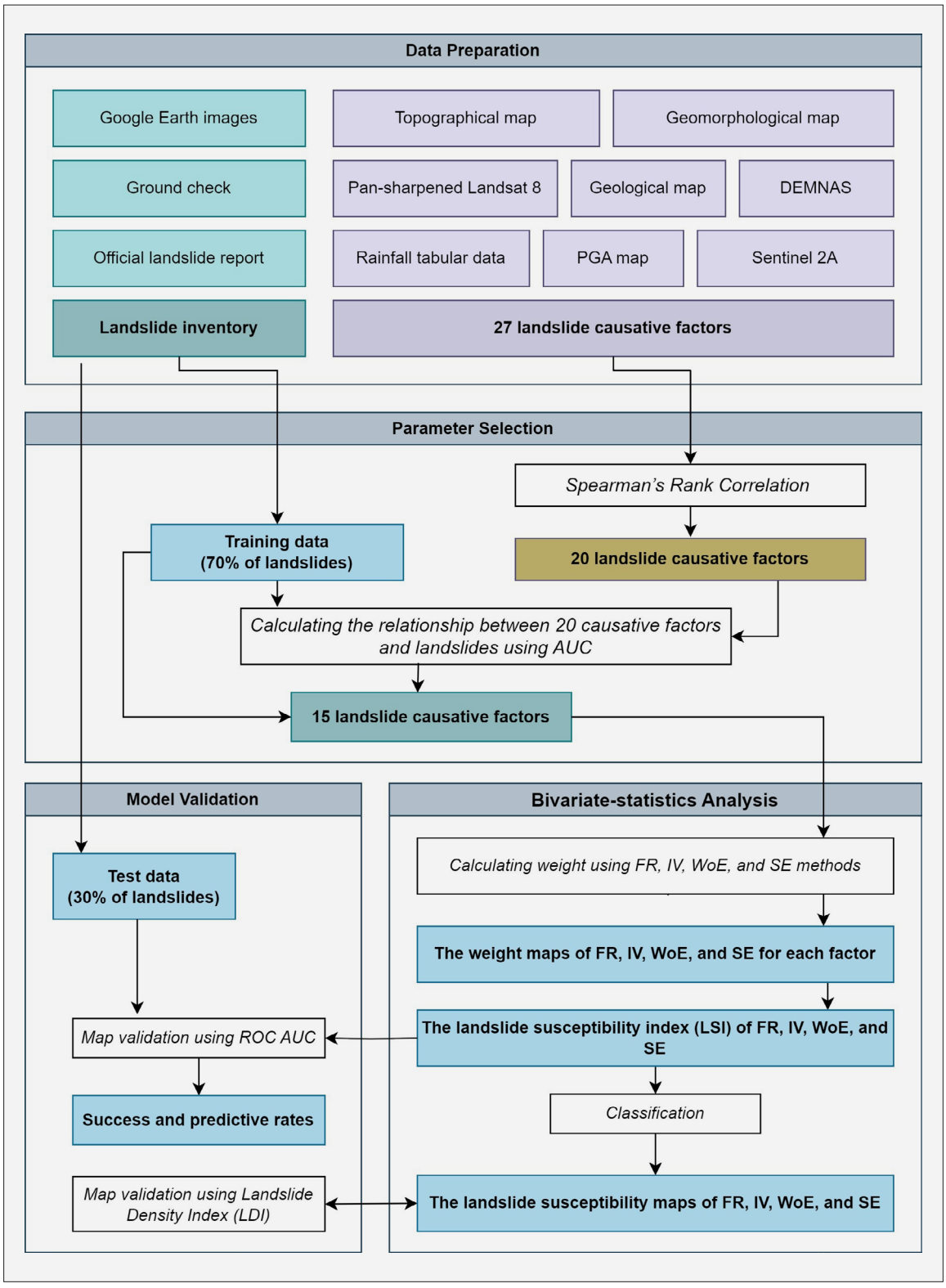

Figure 2: Flowchart to compare four different bivariate statistical-based models

Landsat images, and a PGA map. Those factors are lithology, distance to fault, fault density, distance to lineaments, lineament density, and PGA. PGA is associated with characteristics of earthquakes that are relevant to the circumstances where the earthquake takes place, like in this study area (**Arifianti et al., 2020**). Earthquakes are frequently a trigger factor for landslides. This study, involving a long-period assessment for landslide susceptibility, uses not only PGA but also rainfall as a trigger factor for landslides in this study area.

A different lithological unit can show a different mechanical and hydrological characteristic (**Juliev et al., 2019**) that can influence material strength and stability (**Othman et al., 2018**). The lithology of the research area is dominated by tuff, laharic breccia, and lava. The lithological factor is obtained from the lithological unit layer from the geological map by converting it from vector format to raster, and the PGA factor is obtained from the PGA map by converting it from vector format to raster.

Faults are correlated with the extent of fracture and weathering. Both faults and lineaments are related to tectonic activity and determine weak zones. Therefore, they significantly influence the likelihood of landslides occurring (**Melati et al., 2024**). Regarding faults, distance to fault and fault density factors generated from the fault polyline layer in the geological map were utilized. The "Euclidean distance" tool and the "line density" tool in ArcGIS generate the distance to faults and the fault density factors. A vector map of lineaments was generated using the automatic extraction method in the Catalyst-PCI Geomatics trial version on Landsat 8 images. The image was acquired on July 17, 2018, with a path/raw of 122/065 (**URL 1**). The images in composite 564 were created in a pan-sharped version with a spatial resolution of 15 m. Lineament vector maps are processed to produce distance to lineaments and lineament density using the "Euclidean distance" and "line density" tools, respectively.

The seven topographic factors are elevation, slope, TPI, TRI, HI, RR, and geomorphology. All those factors except geomorphology are derivatives from DEM, while the geomorphology factor was gained from the geomorphology map. Elevation is related to land cover or vegetation types, which often correlates with landslides. The slope is correlated to shear stress, which increases with an increasing slope angle. TPI is the difference in elevation between a specific location (cell) and its surroundings. TPI was counted in the present study using a moving window consisting of 100 cells, or roughly 834 m. In the meantime, TRI displays the terrain heterogeneity level distribution (**Riley, 1999**). TRI was calculated using neighborhood 3 x 3 data, which is the difference between the height at a specific site and the mean height of the eight surrounding cells (**Rózycka et al., 2017**). HI, a parameter related to morphometric features, is an index for identifying the evolutionary stage of landscape development. A high HI value indicates a terrain with great relief, whereas a low value indicates a flat landscape (**Othman et al., 2018**). In the current study, HI was created using a moving window of 100 cells, or about 834 m. The height difference between a unit area's highest and lowest points is known as RR (**Mandal and Mandal, 2018**). In this study, a unit area is 2000 m × 2000 m. Individual centroids were interpolated from each unit area to determine the RR.

Ten hydrological factors are aspect, slope curvature, plan curvature, profile curvature, flow direction, SPI, TWI, distance to rivers, river density, and rainfall. All those factors except rainfall are derivatives maps of DEM. Regarding river factors, the river layer was obtained from the topographical map. Subsequently, the data is transformed into the distance to the rivers and river density by utilizing the "Euclidean distance" and "line density" tools within ArcGIS Desktop.

Aspect affects soil moisture since it is related to solar radiation and evapotranspiration (**Samodra et al., 2017**). Slope curvature, profile curvature, and plane curvature are three forms of curvature that can be distinguished based on their curvature value. In the plan curvature and slope curvatures, a negative value denotes a concave surface, and a positive value represents a convex surface. Conversely, a negative value indicates a convex surface for profile curvature, and a positive value indicates a concave surface. The zero values indicate a flat surface in all three curvature types (**ESRI, 2014; Mandal and Mandal, 2018**). SPI estimates erosion capacity, while TWI indicates flow accumulation and water saturation (**Moore et al., 1991**).

Four anthropogenic factors used in this study are land cover, NDVI, distance to roads, and road density. They impact soil moisture, erosion and runoff, and slope stability. According to **Cao et al. (2019)**, vegetation can improve soil shear strength and slope stability. Road building, on the other hand, is closely linked to slope-cutting and land-filling operations. Sentinel-2A imagery bands 2, 3, 4, and 8, acquired on November 16, 2019, were used to obtain land cover and vegetation indices. The supervised classification technique was employed to generate land cover, as well as the NDVI formula, which calculates the ratio between the difference in reflectance values of the near infra-red (NIR) and red bands and the sum of reflectance values from the same band, to construct the vegetation indices. Road construction diminishes slope stability and shear strength. As well as the river layer, the road layer was obtained from the topographical map. The distance to roads and road density factors were processed using the "Euclidean distance" tool and the "line density" tool in ArcGIS Desktop.

In the bivariate statistical method, each causative factor must be categorical or nominal data. Lithology, geomorphology, aspect, flow direction, land cover, plan curvature, profile curvature, and slope curvature are all known to be nominal data types. The 19 remaining parameters are continuous data and thus need to be classified into several classes. They are often categorized based on the data histogram (**Samodra et al., 2017**). The present study employed various classification techniques, including manual, equal interval, natural break, quantile, and geometric interval. There is a lack of agreement over the classification method used to categorize each parameter or determine the number of classes. **Umar et al. (2014)** used the quantile classification method to divide all continuous data into different classes.

Samodra et al. (2017) divided the elevation map into seven classes based on a 100-m interval. **Melati et al. (2024)** reclassified the map into six classes based on a 300-m interval. **Arifianti et al. (2023)** classified the map into 12 classes based on a 200-m interval.

Distance to rivers, TWI, TRI, SPI, distance to lineaments, and lineament density were some parameters that applied the quantile classification method because the distribution of the data was highly skewed or not normally distributed. Slope, elevation, HI, distance to faults, distance to roads, and rainfall were classified using the manual method. According to Van Zuidam (1985) (**Noor, 2010**), the slope in the percentage was divided into seven classes. HI parameter was classified into three classes following **Othman et al. (2018)**: HI > 0.35 (a flat landscape), 0.35 < HI < 0.6 (a significantly eroded landscape), and > 0.6 (a landscape that has an entrenched drainage network). River density has a data distribution close to a normal curve, so it was classified using equal intervals. There are eight classes for river density. NDVI was classified into eight classes using the natural break method, while TPI was classified into eight classes using geometrical intervals based on its data distribution curve.

2.3.2. Parameter selection

During the initial phase of this study, 27 landslide causative factors were identified and prepared. Parameter selection was employed to address multicollinearity, reach the model's effectiveness, and increase the model's accuracy. Spearman's rank correlation was used to calculate the association between each parameter and landslide data to choose the parameters. Correlation calculations are carried out by creating a correlation matrix as a heatmap. A correlation matrix is a simple table that shows the correlation coefficient between two variables. A strong correlation indicates a significant relationship between two parameters, which can be either positive or negative. This study used threshold correlation values of -0.8 and +0.8 (**Evans, 1996**). Heatmap was created using Pandas and Seaborn packages in a Python environment.

Later, the selected parameters were filtered using the relationship value between each selected parameter and landslide data. The relationship value is an area under the curve (AUC) between the cumulative percentage of an area on the horizontal axis and the cumulative percentage of landslides on the vertical axis based on the bivariate statistical analysis result. The threshold value of AUC is 0.6 (**Sumaryono et al., 2015**). The causative factors included in the correlation calculation in the heatmap are a mixture of continuous data and nominal data. The nominal data are lithology, geomorphology, aspect, flow direction, slope curvature, plan curvature, profile curvature, and land cover. Meanwhile, in the AUC calculation, all causative factors are categorical data, and those continuous data are classified.

2.3.3. Bivariate statistical-based analysis

The bivariate statistical-based methods assess the influence of each class of landslide causative parameters on landslide occurrence (**Umar et al., 2014**). In the bivariate statistical method, each causative factor is categorical data. This study used four bivariate statistical methods: FR, IV, WoE, and SE. These four methods are the most widely used and produce balanced accuracy (**Hadmoko et al., 2017b; Razavizadeh et al., 2017; Liu and Duan, 2018; Othman et al., 2018; Riaz et al., 2018; Juliev et al., 2019; Nohani et al., 2019**).

The FR method is a ratio between a landslide density in a class and a landslide density in the whole study area. The formula is as follows:

$$FR = \frac{a/b}{c/d} \quad (1)$$

where *a* is the number of landslide pixels in a class, *b* is the total of landslide pixels in the whole study area, *c* is the number of pixels in a class, and *d* is the total pixels in the entire study area.

Another straightforward approach is IVM, which is simply the natural log product of FR (**Hadmoko et al., 2017**). On the other hand, the assessment of landslide susceptibility by using WoE requires specific values, namely W^+ (true-positive), W^- (true-negative), and W_c (weight contrast). The weight contrast, which is the difference in value between W^+ and W^-, shows the measure of the correlation strength between the analyzed variables and landslides (**Sumaryono et al., 2015**). The formulas to determine W^+ and W^- are given below:

$$W^+ = \ln\left(\frac{a/b}{e/f}\right) \quad (2)$$

where W^+ is the true-positive (a location is predicted as a prone area, and in reality, it is true), *a* is the number of landslide pixels in a class, *b* is the total of landslide pixels in the whole study area, *e* is the number of non-landslide pixels in a class, *f* is the total of non-landslide pixels in the entire study area.

$$W^- = \ln\left(\frac{g/b}{h/f}\right) \quad (3)$$

where W^- is the true-negative (a location is predicted as a non-landslide area, and in reality, it is true), *g* is the number of landslide pixels outside class, *b* is the total of landslide pixels in the whole study area, *h* is the number of non-landslide pixels outside class, and *f* is the total of non-landslide pixels in the entire study area.

The SE is the most complex of the four methods, with a coefficient for each parameter. The coefficient is represented by an entropy index (**Nohani et al., 2019**) and represents how much each parameter contributes to de-

termining the landslide occurrence. The total of twenty-three parameter coefficients is 1.0. The SE model was calculated using the following formulas:

$$E_{ij} = \frac{FR}{\sum_{j=1}^{M_j} FR} \quad (4)$$

where E_{ij} is the true-negative (a location is predicted as a non-landslide area, and in reality, it is true), FR is the frequency ratio, and j is the parameter 1, ...

$$Hj = -\sum_{i=1}^{M_j} E_{ij} \log_2 E_{ij}, j = 1,\ldots,n$$

$$H_{jmax} = \log_2 M_j, M_{j-\text{number of classes}}$$

$$I_j = \left(H_{jmax} - {Hj}/{Hjmax}\right), I = (0,1), j = 1,\ldots.$$

$$SE_{ij} = I_j FR_{ij} \quad (5)$$

where E_{ij} is the true-negative, H_j and H_{jmax} are the entropy values, Ij is the coefficient of each parameter, Mj is the number of classes in each parameter, and SE is the weight of a class (i) of a parameter (j). The weight of SE is the result of the parameter coefficient of SE multiplied by FR weight.

The landslide probability (the weight) of all factors was summed up to obtain the landslide susceptibility index (LSI) (**Mandal and Mandal, 2018; Pradhan, 2013**). Later, the LSI is classified into five classes to show the zonation of landslide susceptibility, namely very low, low, moderate, high, and very high.

2.3.4. Model validation

Model validation is essential for the appropriate interpretation of LSM maps. Consequently, the validation of the model is a crucial stage in the process of LSM (**Mersha and Meten, 2020**). Model validation is the final step in modelling. It assesses the performance of a model and the quality of the result. Model validation was performed using two methods, namely: (1) the ROC curve analysis to obtain the AUC values and (2) the landslide density index (LDI). ROC analysis validates the LSI, and the LDI validates LSM maps. LSM is a categorical map of LSI.

AUC values can be classified into five classes, i.e. 0.5 – 0.6 (bad), 0.6 – 0.7 (moderate), 0.7 – 0.8 (good), 0.8 – 0.9 (very good), and 0.9 – 1.0 (excellent) (**Pourghasemi et al., 2013**). Two AUC values were produced for model validation, namely success and predictive rates. A validation using the training data produces the success rate, and a validation using the test produces the predictive rates. If the training dataset was utilized to evaluate the model, it cannot accurately reflect the real power of the generated model. The predictive rate measures the accuracy of the model in predicting the occurrence of landslides in a given area (**Umar et al., 2014**).

The LDI is calculated as the proportion of landslide pixels to the total number of pixels in each class on a landslide susceptibility map. The landslide susceptibility map is divided into five classes in this study. The landslide susceptibility map was valid if the LDI increased from very low to very high susceptibility classes for each of the maps (**Mersha and Meten, 2020**).

3. Results

3.1. Landslide training data and test data

A total of 25 landslide polygons were obtained from the interpretation of Google Earth data. In contrast, 33 additional landslide points were identified by analyzing the official landslide records and carrying out field surveys. All landslide objects were converted into raster data with an 8.34 m cell size, enabling each polygon to be represented into one or more landslide pixels, and each landslide point was converted into a single landslide pixel. We selectively utilized representative landslide pixels from big landslide polygons rather than including all pixels from landslide polygons as our data input. If the polygon's area is smaller than 50 m x 50 m, it is denoted by a single pixel. If the polygon is more than that size, it is denoted by two or more pixels, depending on its area. As a result, 76 landslide pixels were utilized for modelling.

The research area is highly susceptible to sliding. However, the shortage of a landslide inventory can be attributed to the lack of landslide reports in this country. The majority of landslide incidents in the study area lack

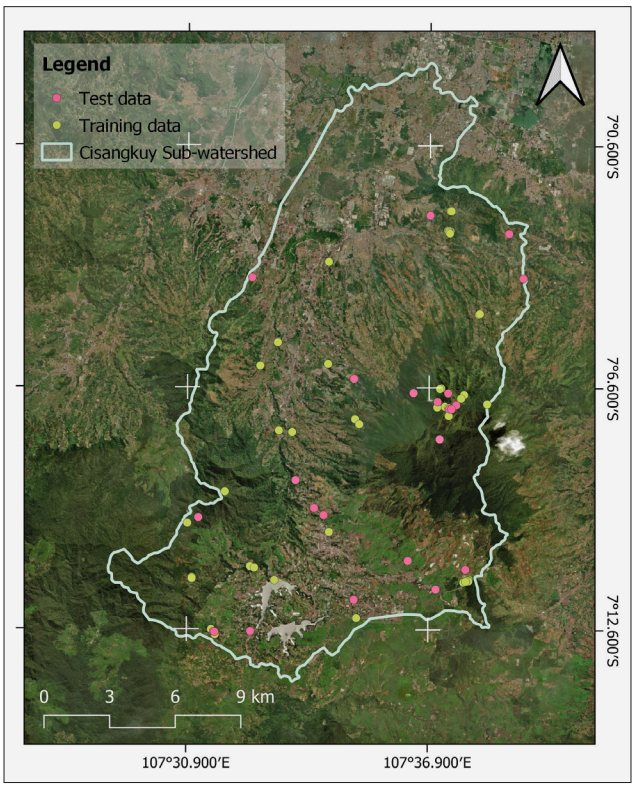

Figure 3: Training and test data of landslides for bivariate statistical-based modelling

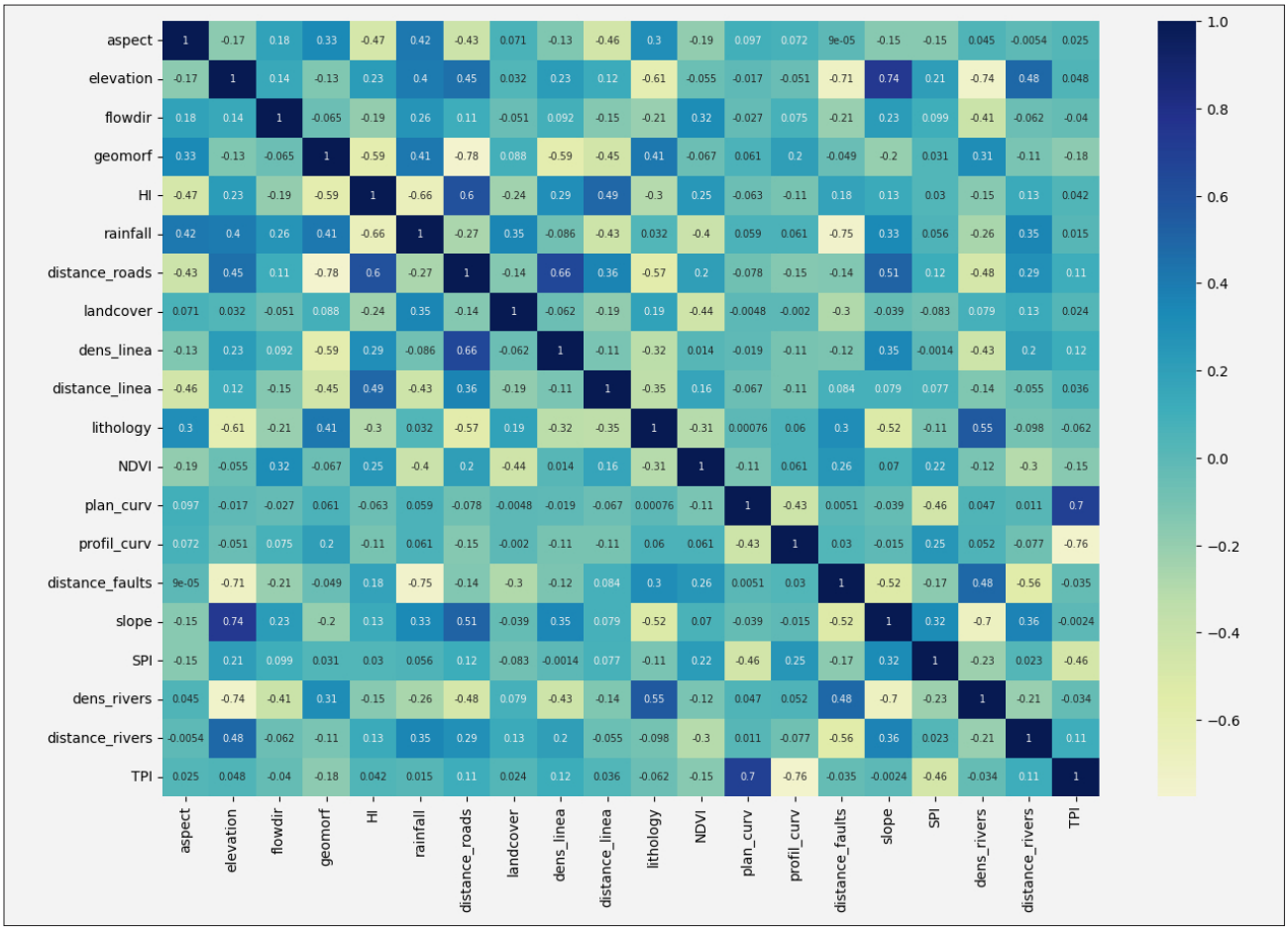

Figure 4: Correlation heatmap of the 20 selected landslide causative factors using Spearman's method

latitude and longitude coordinates, and the names of local areas are used instead (**Cepeda et al., 2010**). The utilization of all landslide reports does not yield representative data due to the fact that the mapping of these landslide locations relies only on administrative regions rather than precise geographical coordinates. Thus, only landslides with geographical coordinates from the official landslide report were employed. In addition, there is limited availability of high-resolution time-series imagery, and obtaining complete landslide inventory data is becoming more challenging in tropical regions like Indonesia, where rapid vegetation growth often obscures previous landslide features (**Samodra et al., 2018**).

Landslide types in this study were debris/earth slide, debris/earth fall, and creep. The landslide pixels were divided randomly into two groups: 70% as training data (53 pixels) and 30% as test data (23 pixels). As a note, the training data and test data do not come from the same landslide polygon to avoid biased results and potentially non-representative model validation. The distribution of landslides used for training and test data can be seen in **Figure 3.**

3.2. Selected parameters

There were 27 landslide causative factors, and parameter selection was performed on them by constructing a correlation heatmap and computing AUC for their remaining parameters. Based on the heatmap, seven parameters that have a strong correlation were eliminated, namely road density, TWI, fault density, RR, PGA, TRI, and slope curvature. They have correlation values of more than 0.8 or less than -0.8. After being eliminated, there is no strong correlation among the 20 factors in the heatmap (see **Figure 4**).

The remaining 20 parameters are evaluated to establish the correlation between variables (factors) and landslide training data using AUC values derived from bivariate analysis. Those factors are (1) lithology, (2) distance to faults, (3) distance to lineaments, (4) lineament density, (5) elevation, (6) slope, (7) TPI, (8) HI, (9) geomorphology, (10) aspect, (11) plan curvature, (12) profile curvature, (13) flow direction, (14) SPI, (15) distance to rivers, (16) river density, (17) rainfall, (18) land cover, (19) NDVI, and (20) distance to roads. For bivariate analysis, all continuous data were classified.

The result shows that five parameters have an AUC of less than 0.6, namely HI, land cover, plan curvature, profile curvature, and TPI, with AUC values of 0.568, 0.587, 0.509, 0.540, and 0.583, respectively. Therefore, a total of 15 factors were employed for LSM, namely, (1) lithology, (2) geomorphology, (3) aspect, (4) flow direc-

tion, (5) elevation, (6) slope, (7) distance to faults, (8) distance to lineaments, (9) lineament density, (10) distance to rivers, (11) river density, (12) SPI, (13) distance to roads, (14) NDVI, and (15) rainfall. All those selected factors can be observed in **Figure 5**.

Four of those 15 selected factors are nominal data, and the other 11 factors are categorical data derived from continuous data. The nominal data are lithology, geomorphology, aspect, and flow direction. The lithology of the Cisangkuy Sub-watershed consists of (1) tiff, laharic breccia, and lava, (2) crystal tuff, tuff breccia, and old laharic deposit, (3) intercalated lava, lahar, and breccia, (4) intercalated lava, breccia, and tuff, (5) clay and silt, as well as (5) tuffaceous breccia and lava. Based on the lithology map, there is no landslide in the clay and silt. The study area has six geomorphology units, i.e. a

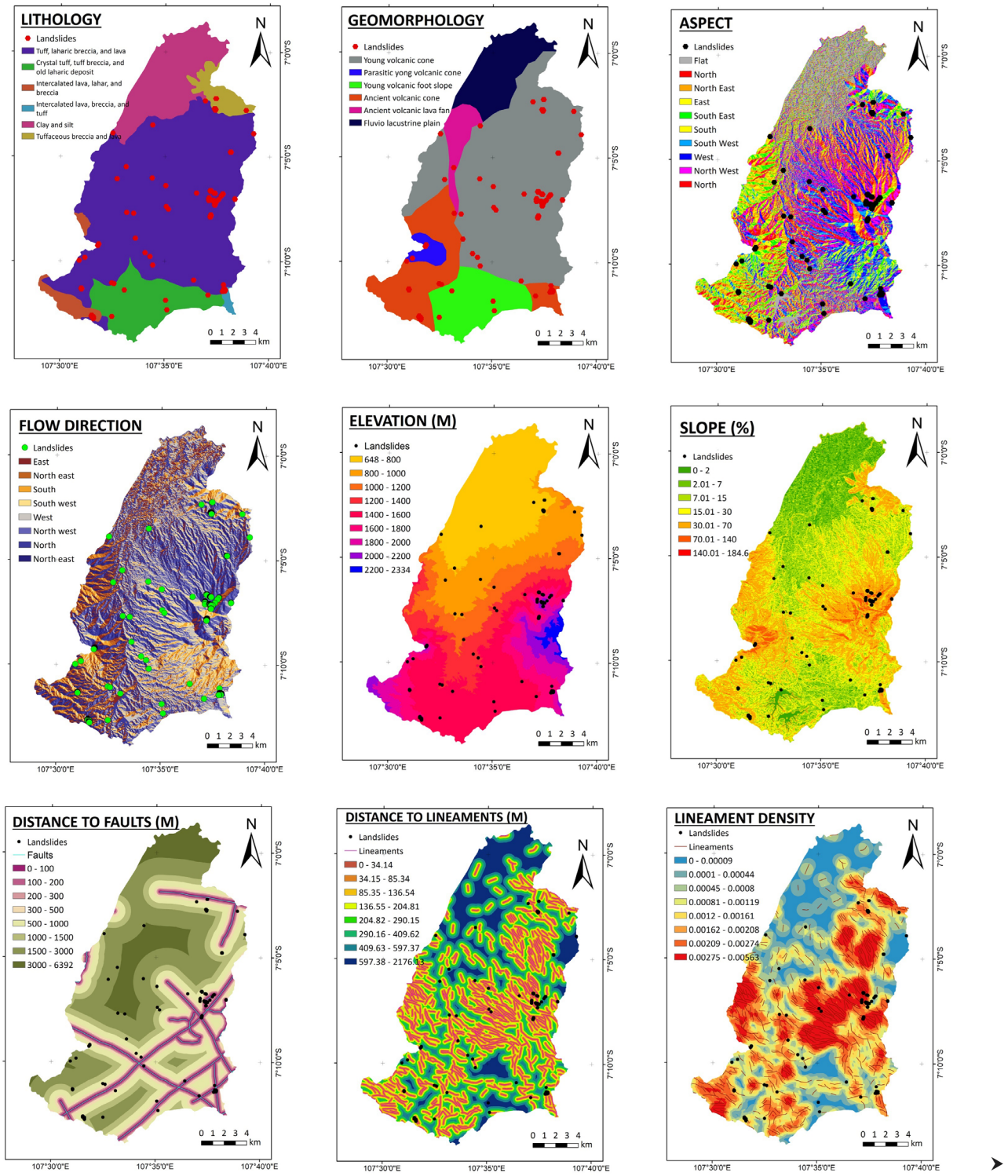

Figure 5: The 15 selected landslide causative factors superimposed with landslide distribution

young volcanic cone, parasitic young volcanic cone, young volcanic foot slope, ancient volcanic cone, ancient volcanic lava fan, and fluvio lacustrine plain. There is no landslide in the fluvio lacustrine plain. Aspect consists of nine directions of the slope, namely flat, north, northeast, east, southeast, east, southwest, west, and northwest. The land unit of the slope is similar to the land unit of the flow direction, but in the flow direction, there is no flat. Landslides were found in all land units with aspects and flow directions.

3.3. Landslide susceptibility analysis

This landslide susceptibility analysis was performed using four bivariate statistical methods: FR, IVM, WoE, and SE. Every class of landslide causative factor was linked to landslides. By knowing the number of landslides in every class and the area of each class, the probability of landslides can be obtained using each method, either FR, IVM, WoE, or SE. The results of the bivariate statistical modelling for each class from each parameter can be seen in **Table 2**. The probability of a landslide increases with increasing weight (**Lee and Talib, 2005**), either the weight of FR, IVM, WoE, or SE.

In the aspect factor, the highest landslide probability is in the West class for all models. Regarding the elevation factor, class 1600 – 1800 has the highest probability of landslides. The elevation class is not the highest in the study area; the highest class is 2200 – 2334. Based on the flow direction factor, the most significant possibility of landslides is in the Southwest class. An ancient volcanic cone is a geomorphology unit with the highest landslide probability. The rainfall class of 2400 – 2450 mm/year, which represents high rainfall in the study area, has the highest probability of landslide, with a weight that really stands out compared to other classes.

Regarding the distance to roads, the 1600 – 2400 class has the highest weight; this indicates that the construction of roads in the study area does not significantly impact the stability of the slope. The road construction shows some resemblance to the presence of lineaments, as well as the presence of rivers. The class with the highest lineament density has weights of 0.0012 – 0.00161,

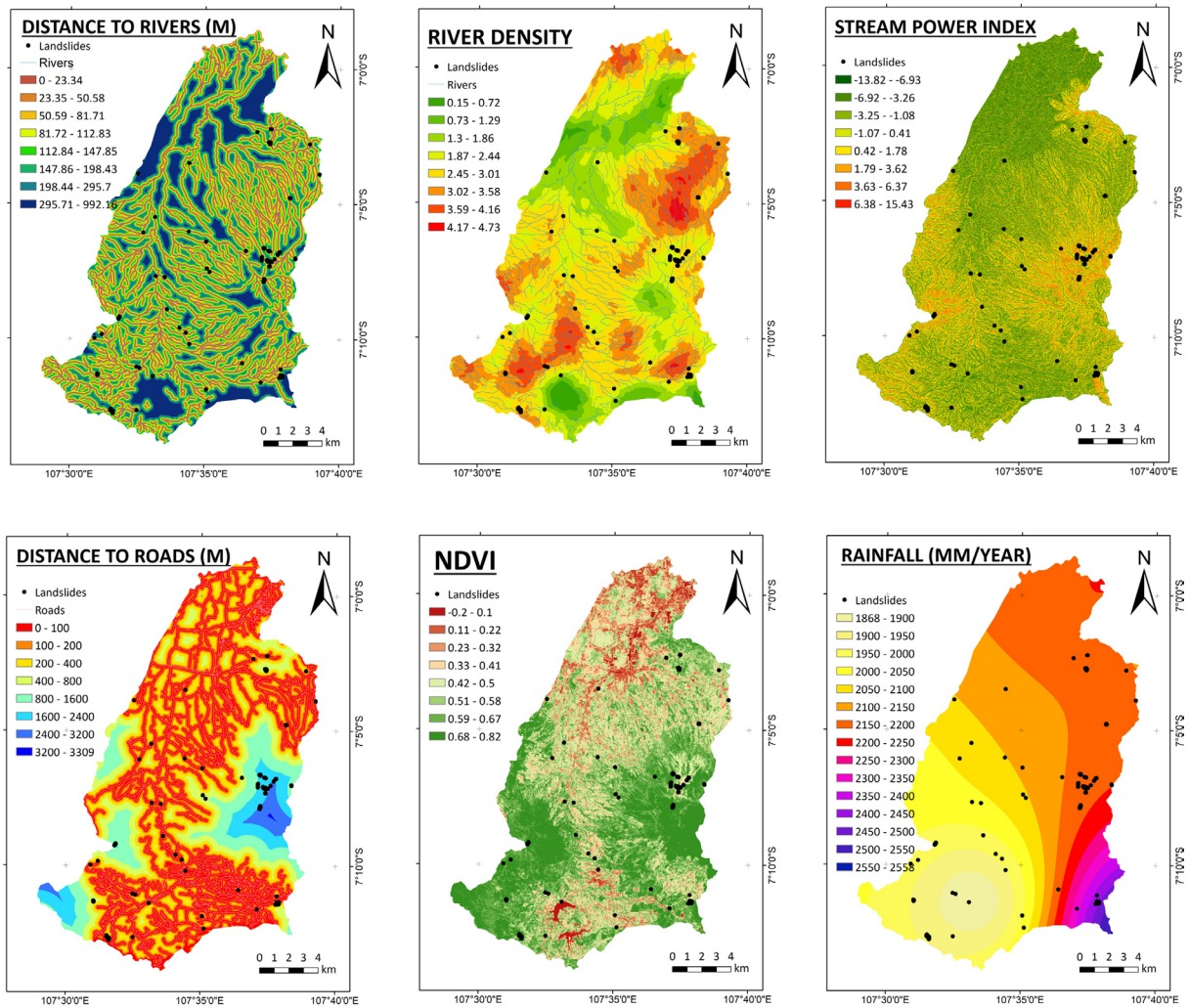

Figure 5: Continued

Table 2: The weights for classes of 15 landslides causative factors based on FR, IVM, WoE, and SE

Data layers	Number of pixels	Number of training landslide pixels	FR	IVM	WoE	SE
Aspect						
Flat (-1)	513492	2	0.338	-1.085	-1.195	0.006
North (0°to 22.5°)	277158	1	0.315	-1.157	-1.229	0.006
Northeast (22.5° to 67.5°)	573009	5	0.734	-0.309	-0.375	0.014
East (67.5° to 112.5°)	356541	1	0.249	-1.392	-1.483	0.005
Southeast (112.5° to 157.5°)	340361	3	0.742	-0.299	-0.348	0.014
South (157.5° to 202.5°)	366423	5	1.138	0.129	0.114	0.021
Southwest (202.5° to 247.5°)	501097	7	1.165	0.152	0.146	0.022
West (247.5° to 292.5°)	547356	12	1.817	0.597	0.692	0.034
Northwest (292.5° to 337.5°)	667948	12	1.492	0.400	0.464	0.028
North (337.5° to 360°)	286630	5	1.450	0.371	0.375	0.027
Elevation (m)						
648 – 800	1075656	9	0.705	-0.350	-0.469	0.040
800 – 1000	669502	4	0.509	-0.676	-0.760	0.029
1000 – 1200	478770	2	0.361	-1.018	-1.093	0.021
1200 – 1400	448213	3	0.568	-0.566	-0.614	0.033
1400 – 1600	968519	7	0.611	-0.492	-0.595	0.035
1600 – 1800	341338	19	4.585	1.523	1.878	0.263
1800 – 2000	263724	5	1.574	0.454	0.491	0.090
2000 – 2200	137478	4	2.405	0.878	0.924	0.138
2200 – 2334	46815	0	0.019	-3.989	-3.999	0.001
Flow direction						
East	459761	1	0.197	-1.625	-1.727	0.003
South East	290732	3	0.865	-0.145	-0.168	0.012
South	420543	7	1.384	0.325	0.353	0.020
South West	446817	8	1.487	0.397	0.440	0.021
West	787743	13	1.372	0.317	0.387	0.020
North West	595630	10	1.396	0.333	0.383	0.020
North	909058	6	0.560	-0.580	-0.701	0.008
North East	519731	5	0.808	-0.214	-0.252	0.012
Geomorphology						
Young volcano cone	2445504	26	0.891	-0.116	-0.279	0.070
Parasitic young volcanic cone	90534	1	0.925	-0.078	-0.117	0.073
Young volcano foot slope	506798	2	0.342	-1.072	-1.191	0.027
Ancient volcanic cone	703271	23	2.701	0.994	1.344	0.212
Ancient volcanic lava fan	186876	1	0.458	-0.782	-0.843	0.036
*Fluvi*o lacustrine plain	497032	0	0.019	-3.989	-4.143	0.001
Rainfall (mm/year)						
1868 – 1900	185976	3	1.342	0.294	0.537	0.256
1900 – 1950	393370	4	0.853	-0.159	0.054	0.163
1950 – 2000	415032	3	0.612	-0.492	-0.304	0.117
2000 – 2050	456846	3	0.557	-0.585	-0.407	0.106
2050 – 2100	475396	5	0.881	-0.126	0.087	0.168
2100 – 2150	799438	1	0.121	-2.111	-2.060	0.023
2150 – 2200	1343027	19	1.179	0.165	0.473	0.225
2200 – 2250	130374	0	0.019	-3.989	-3.791	0.004
2250 – 2300	68038	0	0.019	-3.989	-3.777	0.004

Table 2: Continued

Data layers	Number of pixels	Number of training landslide pixels	FR	IVM	WoE	SE
2300 – 2350	52518	0	0.019	-3.989	-3.773	0.004
2350 – 2400	45346	0	0.019	-3.989	-3.772	0.004
2400 – 2450	37697	15	32.662	3.486	4.031	6.237
2450 – 2500	17463	0	0.019	-3.989	-3.765	0.004
2500 – 2550	9126	0	0.019	-3.989	-3.764	0.004
2550 – 2558	368	0	0.019	-3.989	-3.762	0.004
Distance to roads (m)						
0 – 100	1628993	20	1.026	0.025	0.023	0.079
100 – 200	887181	12	1.128	0.121	0.136	0.087
200 – 400	748091	1	0.128	-2.054	-2.235	0.010
400 – 800	471194	9	1.585	0.461	0.516	0.122
800 – 1600	415539	2	0.413	-0.883	-0.960	0.032
1600 – 2400	198068	9	3.746	1.321	1.441	0.288
2400 – 3200	79627	0	0.019	-3.989	-4.024	0.001
3200 – 3309	1322	0	0.019	-3.989	-4.007	0.001
Lineaments density						
0 – 0.00009	527937	0	0.019	-3.989	-4.193	0.001
0.0001 – 0.00044	572826	3	0.448	-0.803	-0.961	0.023
0.00045 – 0.0008	579228	6	0.868	-0.141	-0.240	0.044
0.00081 – 0.00119	570701	8	1.169	0.156	0.102	0.059
0.0012 – 0.00161	545269	21	3.178	1.156	1.442	0.160
0.00162 – 0.00208	543395	9	1.377	0.320	0.295	0.069
0.00209 – 0.00274	544007	1	0.169	-1.776	-1.965	0.009
0.00275 – 0.00563	546652	5	0.769	-0.263	-0.374	0.039
Distance to lineaments (m)						
0 – 34.14	464488	9	1.608	0.475	0.507	0.051
34.15 – 85.34	686948	16	1.929	0.657	0.802	0.061
85.35 – 136.54	552478	10	1.503	0.408	0.440	0.048
136.55 – 204.81	582344	7	1.005	0.005	-0.037	0.032
204.82 – 290.15	557070	3	0.460	-0.776	-0.893	0.015
290.16 – 409.62	558238	6	0.900	-0.105	-0.162	0.029
409.63 – 597.37	515182	2	0.337	-1.088	-1.214	0.011
597.38 – 2176.13	513267	0	0.019	-3.989	-4.152	0.001
Lithology						
Tuff, laharic breccia, and lava	2921875	27	0.155	-1.864	-2.834	0.017
Intercalated lava, breccia, and tuff	23735	15	29.693	3.391	3.561	3.195
Clay and silt	662202	0	1.064	0.062	0.074	0.115
Tuffaceous breccia and lava	159641	7	5.151	1.639	1.808	0.554
Intercalated lava, lahar, and breccia	118765	2	7.206	1.975	2.163	0.775
Crystal tuff, tuff breccia, and old laharic deposit	543797	2	1.635	0.492	0.585	0.176
NDVI						
-0.2 – 0.1	63057	0	0.019	-3.989	-4.011	0.001
0.11 – 0.22	224763	0	0.019	-3.989	-4.048	0.001
0.23 – 0.32	292594	11	3.103	1.132	1.285	0.175
0.33 – 0.41	478656	9	1.561	0.445	0.507	0.088
0.42 – 0.5	656485	7	0.893	-0.113	-0.140	0.050
0.51 – 0.58	722085	9	1.041	0.040	0.040	0.059

Table 2: Continued

Data layers	Number of pixels	Number of training landslide pixels	FR	IVM	WoE	SE
0.59 – 0.67	933288	13	1.161	0.149	0.185	0.065
0.68 – 0.82	1059087	4	0.328	-1.114	-1.313	0.018
Distance to faults (m)						
0 – 100	263495	15	4.689	1.545	1.804	0.283
100 – 200	252447	9	2.943	1.080	1.198	0.178
200 – 300	232409	1	0.372	-0.990	-1.031	0.022
300 – 500	411262	2	0.417	-0.874	-0.938	0.025
500 – 1000	809152	9	0.931	-0.071	-0.094	0.056
1000 – 1500	652349	9	1.150	0.140	0.160	0.070
1500 – 3000	1288490	7	0.464	-0.767	-0.973	0.028
3000 – 6392	520411	1	0.176	-1.736	-1.847	0.011
Slope (%)						
0-2	359716	0	0.019	-3.989	-4.352	0.002
2-7	603748	1	0.154	-1.868	-2.273	0.015
7-15	885918	1	0.111	-2.197	-2.678	0.011
15-30	1255909	13	0.868	-0.142	-0.473	0.085
30-70	1202005	34	2.339	0.850	1.260	0.230
70-140	122618	4	2.695	0.991	0.761	0.265
140-184.6	101	0	0.019	-3.989	-4.269	0.002
SPI						
-13.82 - -6.93	519831	0	0.019	-3.989	-4.144	0.001
-6.92 - -3.26	586919	8	1.137	0.128	0.116	0.043
-3.25 - -1.08	529848	0	0.019	-3.989	-4.147	0.001
-1.07 – 0.41	580808	5	0.725	-0.322	-0.395	0.027
0.42 – 1.78	530627	9	1.410	0.344	0.368	0.053
1.79 – 3.62	547999	9	1.366	0.312	0.332	0.051
3.63 – 6.37	578082	9	1.296	0.259	0.272	0.048
6.38 – 15.43	555901	13	1.937	0.661	0.773	0.072
River density						
0.15 – 0.72	64288	0	0.019	-3.989	-4.315	0.002
0.73 – 1.29	245001	0	0.019	-3.989	-4.356	0.002
1.3 – 1.86	585880	2	0.299	-1.209	-1.622	0.031
1.87 – 2.44	963905	10	0.870	-0.140	-0.487	0.090
2.45 – 3.01	1373724	35	2.109	0.746	1.124	0.217
3.02 – 3.58	850676	6	0.597	-0.516	-0.918	0.061
3.59 – 4.16	318126	0	0.019	-3.989	-4.374	0.002
4.17 – 4.73	28415	0	0.019	-3.989	-4.307	0.002
Distance to rivers (m)						
0 – 23.34	513894	3	0.497	-0.698	-0.784	0.009
23.35 – 50.58	652854	9	1.149	0.139	0.143	0.021
50.59 – 81.71	593156	8	1.125	0.118	0.115	0.021
81.72 – 112.83	590553	12	1.686	0.522	0.611	0.031
112.84 – 147.85	518247	9	1.443	0.367	0.405	0.026
147.86 – 198.43	537009	8	1.241	0.216	0.227	0.023
198.44 – 295.7	523421	3	0.489	-0.716	-0.804	0.009
295.71 – 992.16	500881	1	0.182	-1.702	-1.823	0.003

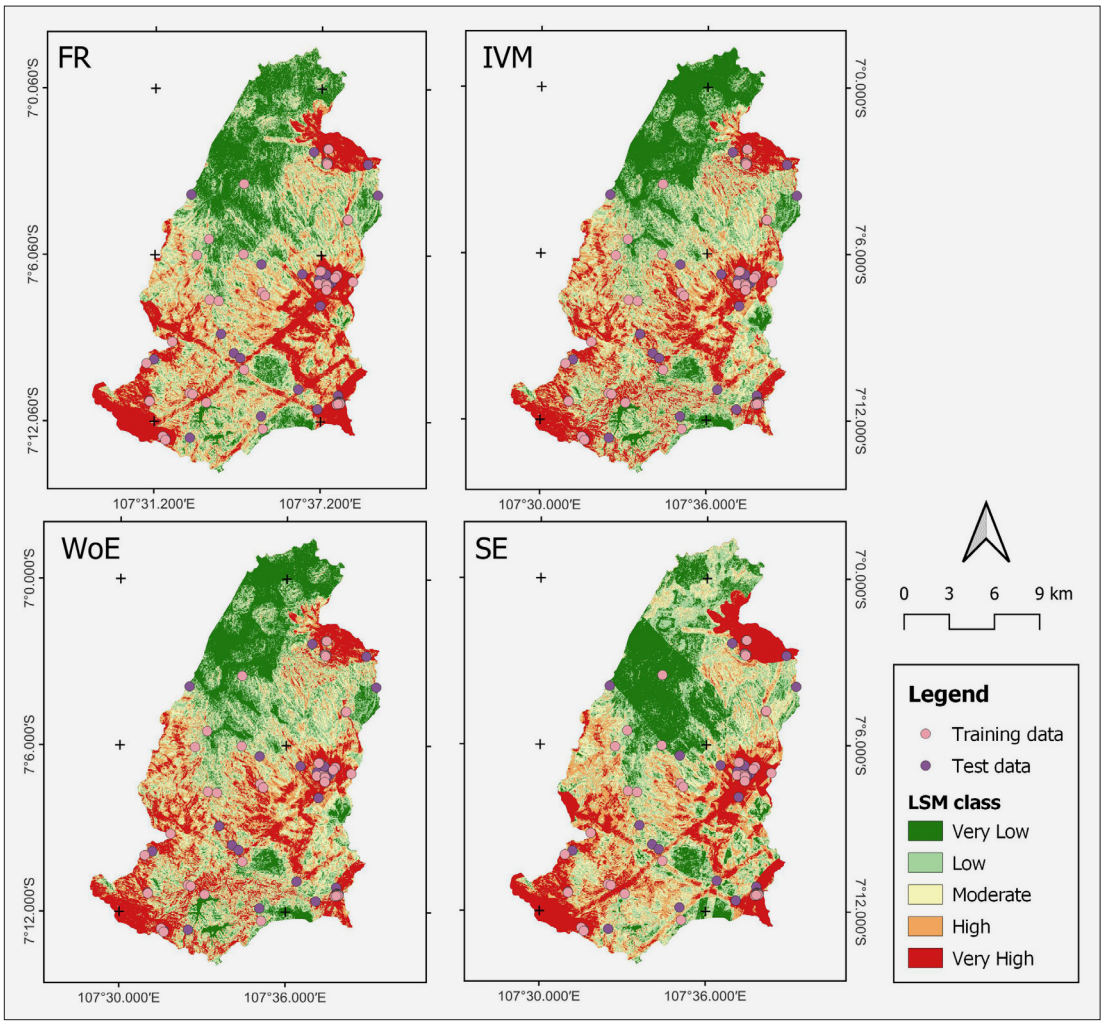

Figure 6: Maps of LSM for the study area using FR, IVM, WoE, and SE models

ranking fifth out of the eight classes. Similarly, the distance to rivers does not have the most significant weight in the class closest to the river, but rather in the fourth highest ranked class out of eight classes. The occurrence of road constructions, lineaments, and rivers differs from the occurrence of faults. The class closest to faults within a range of 0 to 100 meters has the highest likelihood of landslides, indicating that faults play a substantial role in causing landslides. Regarding SPI, SPI significantly affects landslides; the highest SPI has the highest weight of landslide susceptibility.

The intercalated lava, breccia, and tuff lithology classes have the highest probability of landslides. The vegetation index ranging from 0.23 to 0.32 has the most significant weight; this is reasonable, as lower vegetation index values may indicate the presence of water bodies or open areas with flat terrain.

3.4. LSM maps of FR, IVM, WoE, and SE

The LSI maps of the study area by four models were produced based on the weights from the 15 causative factors and the landslide training data (see **Table 2**). The LSI maps of the Cisangkuy Sub-watershed were obtained by summing the weight of all the 15 causative factors. The LSI for the FR model in the study area ranges from 3.179 to 94.599. The LSI for IVM, WoE, and SE are -36.637 to 17.458, -38.755 – 20.801, and 0.199 – 11.276, respectively. The data distribution type of each LSI is skewed. The LSI values of FR and SE have a data distribution type skewed right, while the LSI values of IVM and WoE have a data distribution left skewed. All those LSI values were categorized into five susceptibility classes using the quantile classification method (**Umar et al., 2014**). The LSI values for FR were categorized into five classes: very low (3.179 – 10.399), low (10.935 – 13.171), moderate (13.171 – 15.148), high (15.148 – 17.804), and very high (17.804 – 94.599). The LSI values of IVM are categorized into five groups based on their magnitude: very low (-36.637 – -2.942), low (-12.941 – -7.919), moderate (-7.919 – -4.890), high (-4.890 – -2.161), and very high (-2.161 – 17.458). The WoE values are classified into the following categories: very low (-38.755 – -14.643), low (-14.643 – -9.400), moderate (-9.400 - -6.014), high (-6.014 – -2.836), and very high (-2.837 – 20.801). The SE values are classified

into the following categories: very low (0.199 – 0.677), low (0.677 – -0.815), moderate (0.815 – 0.945), high (0.945 – 1.128), and very high (1.128 – 11.276). All resulting LSM maps can be seen in **Figure 6**.

In **Figure 6**, it can be seen that areas that have a high level of landslide susceptibility are in the western Pangalengan region, which consists of tea and coffee plantations, and the eastern Pangalengan region, which consists of mountains (Mount Wayang, Mount Malabar, and Mount Puntang), as well as in the Arjasari Hill region which is in the northeast of the Cisangkuy Sub-Watershed. In fact, landslides occur every year and can even occur several times a year. For example, on November 4, 2021, a landslide occurred in Cinangka, Wargaluyu village, Arjasari District, which caused the house to be damaged and seven residents to be buried. In 2022, on October 11, a landslide occurred in Pasirsari, Patrolsari village, Arjasari District, which caused a land retaining wall to collapse and block road access for residents. In the same year (2022), there was also a landslide that caused one person to die, namely on January 15 in Giri Awas, Sukaluyu village, Panglangengan District. Both the Arjasari and Pangalengan areas have clayey soil, which is identical to the mineral montmorillonite (**Wibawa et al., 2018**), which has the property of crumpling and unfolding causes potential landslides.

Areas with very low landslide susceptibility are mostly found in the northern part of the Cisangkuy Sub-watershed, namely in the Baleendah and Banjaran areas. Gentle slopes dominate both areas. Historically, in the last decade, there has only been one landslide, namely in Cipancur, Baleendah village, on March 18, 2023. However, in general, this area is prone to flooding caused by the overflowing of the Citarum River (**Nurwulandari and Rismana, 2021; Theophilus et al., 2022**).

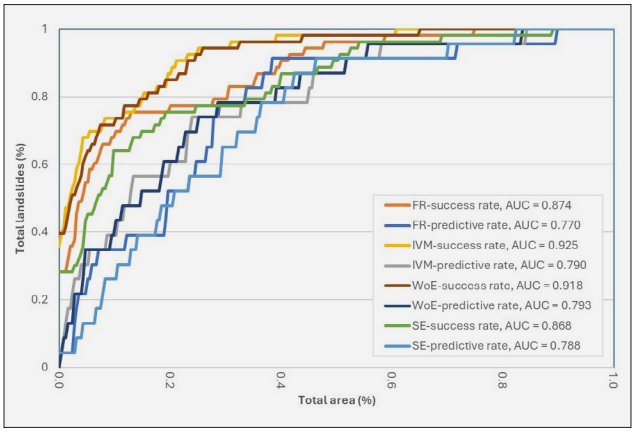

Figure 7: The AUC values of ROC identify model performances: FR, IVM, WoE, and SE

Table 3: Validation model using landslide density index (LDI)

Landslide susceptibility class	Number of landslide training pixels	Number of landslide test pixels	Number of pixels	LDI using training data	LDI using test data
FR					
Very low	0	1	886281	0.0000000	0.0000011
Low	1	1	886310	0.0000011	0.0000011
Moderate	4	0	886133	0.0000045	0.0000000
High	7	9	885698	0.0000079	0.0000102
Very high	41	12	885593	0.0000463	0.0000136
IVM					
Very low	0	1	886107	0.0000000	0.0000011
Low	1	0	886121	0.0000011	0.0000000
Moderate	0	4	886174	0.0000000	0.0000045
High	6	4	885864	0.0000068	0.0000045
Very high	46	14	885749	0.0000519	0.0000158
WoE					
Very low	0	1	886007	0.0000000	0.0000011
Low	1	0	886137	0.0000011	0.0000000
Moderate	1	3	886127	0.0000011	0.0000034
High	6	5	885908	0.0000068	0.0000056
Very high	45	14	885836	0.0000508	0.0000158
SE					
Very low	1	1	891294	0.0000011	0.0000011
Low	1	1	887333	0.0000011	0.0000011
Moderate	4	0	888531	0.0000045	0.0000000
High	6	8	882122	0.0000068	0.0000091
Very high	41	13	880735	0.0000466	0.0000148

3.5. *Model validation*

Model validation in this study used ROC analysis and the LDI method. AUC of ROC evaluated the performance of each model by comparing the LSI with test data of landslides to produce a predictive rate. However, the success rate was calculated by comparing the LSI with training data of landslides. The findings indicate that all success rates are above the corresponding prediction rates for each model by a small margin; it suggests that the models are working accurately. The success rates of FR, IVM, WoE, and SE are 0.874, 0.925, 0.918, and 0.868, respectively.

In comparison, the predictive rates are as follows: 0.770 for FR, 0.790 for IVM, 0.793 for WoE, and 0.788 for SE (see **Figure 7**). WoE has the best performance because it has the highest predictive rate compared to the others. According to the AUC values classified by **Pourghasemi et al. (2013)**, all four models have good performance, which is in the range of 0.7 – 0.8.

The LDI method is the ratio between landslide pixels to the total number of pixels in each class on a landslide susceptibility map; in this study, LSM maps have five classes, namely very low, low, moderate, high, and very high. LDI was performed either using training data or test data of landslides for all LSM maps (see **Table 3**). **Table 3** shows that the LDI values using test data for very high susceptibility classes are 0.0000136, 0.0000158, 0.0000158, and 0.0000148 for the FR, IVM, WoE, and SE models, respectively. These values are significantly higher than those of the other classes. LDI, which utilizes training data of landslides, experiences a similar phenomenon. Those values for very high susceptibility classes can indicate that the LSM maps are valid if there is a gradual decrease in LDI values, ranging from high to low susceptibility. The results of this study show there is no gradual decrement in landslide density values from very high to very low susceptibility (see **Figure 8**). Nevertheless, the outcomes remain favorable as the classes that are most susceptible in all models exhibit the highest LDI.

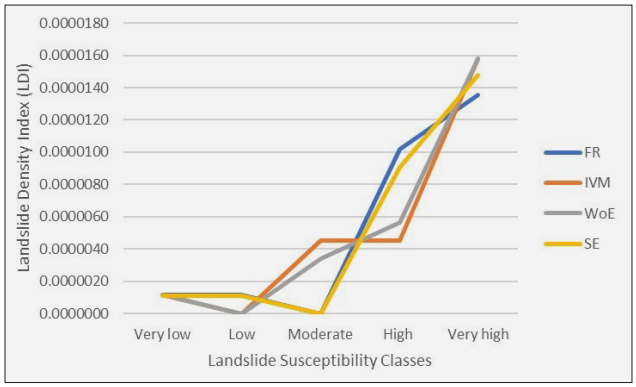

Figure 8: Landslide density index for FR, IVM, WoE, and SE using test data of landslides

4. Discussion

Most of the Cisangkuy Sub-watershed area is prone to landslides due to its rugged terrain. Hence, it is essential to have an understanding of particular areas prone to landslides. There is an urgent need for detailed information that includes exact spatial information. Bivariate statistical methods correlate the classes of landslide causative factors with landslides. This work utilized various bivariate statistical models (FR, IVM, WoE, and SE) to evaluate and generate LSM maps using 76 landslide pixels.

The scarcity of landslide data in this context is understandable due to the limited availability of high-resolution time-series images. The limited amount of landslide data is often encountered in data-scarce environments, like those used by some previous studies. **Sun et al. (2018)** utilized a dataset consisting of 40 landslide occurrences for their models. **Sun et al. (2018)** achieved a model accuracy of 84.9% using the logistic regression approach. **Mallick et al. (2018)** employed a dataset of 38 landslide occurrences. They produced a high-reliability susceptibility map (89.1%) using an integration method between fuzzy and AHP-MCDA (analytical hierarchy process – multi-criteria decision analysis).

Twenty-seven landslide causative factors were prepared in the study. The primary determinant for preparing those is the presence of data sources throughout the study area (**Melati et al., 2024**) related to geological, topographical, hydrological, and anthropological aspects. There is less agreement on the landslide causative factors or the specific number that should be chosen to generate LSM maps. Based on the analysis of those 27 factors, seven factors have a strong correlation with each other, and five factors have a weak relationship with landslides, leaving only 15 causative factors. Only the 15 chosen parameters were utilized for modelling.

Bivariate statistical methods require categorical data for causative factors. The continuous causal factors need to be classified as categorical data. Currently, there is no consensus on the specific number of classes and classification methods that should be used to categorize the continuous causative factors. It is typically a heuristic categorization (**Chen et al., 2017**). It can cause the LSM maps produced to be less accurate. Upon closer inspection of the LSM maps, it becomes noticeable that there is a remarkable pattern present in all of the maps (see **Figure 6**). This pattern is caused by the factor of distance to fault, as depicted in **Figure 5.** The map of distance to faults is divided into eight classes using manual classification. The first class, which represents distances from 0 to 100 m, has the highest number of landslides and carries the greatest weight in terms of landslide susceptibility across all models. Conversely, the fifth class, which ranges from 1500 to 3000 m, contains the largest portion of the entire area, accounting for 29.1% of the total area. The first class of fault distance carries an FR weight of

4.689, whereas the fifth category of fault distance carries an FR weight of 0.464. The FR's weight is in line with the other models, namely IVM, WoE, and SE. The condition results in the LSM map having a prominent look, with the distance to the fault factor having the largest weight. A noticeable appearance on the final map could indicate that the model is less satisfactory in producing the resulting map. The model's performance may be influenced by the number of classes and classification methods used for categorizing each continuous parameter. Similarly, likewise when it regards landslide inventory data, enhancing both the accuracy and amount of the data also has an impact on the effectiveness of the model.

5. Conclusions

Four bivariate statistical models (FR, IVM, WoE, and SE) were used to analyze the relationship between landslide occurrences and landslides causative factors in the Cisangkuy Sub-watershed. The model utilized 76 landslide pixels and 15 causative factors. The landslide inventory data was derived from 25 landslide polygons identified through visual interpretation of Google Earth imagery, as well as 33 landslide points obtained by investigating official landslide records and conducting field surveys. The 15 factors were selected from 27 causative factors. The highly correlated causative factors were removed to address multicollinearity. In addition, only causal factors related to landslide data are involved in the modelling. The aim was to produce maps that show the landslide susceptibility level. The success rates of the different models are as follows: 0.874 for FR, 0.925 for IVM, 0.918 for WoE, and 0.868 for SE. While the predictive rates are as follows: 0.770 for FR, 0.790 for IVM, 0.793 for WoE, and 0.788 for SE. All four models indicate good and balanced performance, with results ranging from 0.7 to 0.8. The WoE model shows a small superiority over the other three models. The results of this study show there is no gradual decrement in landslide density values from very high to very low susceptibility. LSM maps are considered valid when there is a gradual decrease in LDI values, ranging from high to low susceptibility. In this case, the resultant maps are still favorable because the classes that are most susceptible in all models have the highest LDI.

A noticeable appearance on the final map might indicate that the model is less successful in generating the resulting map. The performance of the model can be affected by the number of classes and classification methods employed to categorize each continuous parameter. Similarly, when it relates to landslide inventory data, improving both the precision and quantity of the data significantly affects the efficiency of the model. This study recommended future research regarding the classification method and the number of classes to categorize the continuous landslide causative factor to obtain the most optimal for a single causative factor. In addition, this study advises that there should be an increased effort to collect landslide data in order to get more representative landslide data. This could be an attempt to enhance the accuracy and reliability of the bivariate statistic models.

Acknowledgement

This study is supported by the Research Center for Geological Disaster, National Research and Innovation Agency (BRIN).

6. References

Abidin, H. Z., Gumilar, I., Andreas, H., Murdohardono, D., and Fukuda, Y. (2013): On causes and impacts of land subsidence in Bandung Basin, Indonesia, Environmental Earth Sciences, 68(6), 1545–1553. https://doi.org/10.1007/s12665-012-1848-z

Alfarabi, M. S., Supriatna, Manessa, M. D. M., Rustanto, A., and Ristya, Y. (2019): Geomorphology and Landslide-Prone Area in Cisolok District, Sukabumi Regency, E3S Web of Conferences, 125(201 9), 1–6. https://doi.org/10.1051/e3sconf/201912501005

Arifianti, Y., Pamela, Agustin, F., and Muslim, D. (2020): Comparative study among bivariate statistical models in landslide susceptibility map, Indonesian Journal on Geoscience, 7(1), 51–63. https://doi.org/10.17014/IJOG.7.1.51-63

Arifianti, Y., Pamela, Iqbal, P., Sumaryono, Omang, A., and Lestiana, H. (2023): Susceptibility Assessment of Earthquake-induced Landslides: the 2018 Palu, Sulawesi Mw 7.5 Earthquake, Indonesia, Rudarsko-Geolosko-Naftni Zbornik, 38(3), 43–54. https://doi.org/10.17794/rgn.2023.3.4

Barredo, J. I., Benavides, A., Hervás, J., and Van Westen, C. J. (2000): Comparing heuristic landslide hazard assessment techniques using GIS in the Tirajana basin, Gran Canaria Island, Spain, International Journal of Applied Earth Observation and Geoinformation, 2000(1), 9–23. https://doi.org/10.1016/s0303-2434(00)85022-9

Batar, A. K., and Watanabe, T. (2021): Landslide Susceptibility Mapping and Assessment Using Geospatial Platforms and Weights of Evidence (WoE) Method in the Indian Himalayan Region: Recent Developments, Gaps, and Future Directions, ISPRS International Journal of Geo-Information, 10(3), 114. https://doi.org/10.3390/ijgi10030114

Cao, J., Zhang, Z., Wang, C., Liu, J., and Zhang, L. (2019): Susceptibility assessment of landslides triggered by earthquakes in the Western Sichuan Plateau, Catena, 175(December 2018), 63–76. https://doi.org/10.1016/j.catena.2018.12.013

Cepeda, J., Smebye, H., Vangelsten, B., Nadim, F., and Muslim, D. (2010): Landslide risk in Indonesia, October, retrieved from internet: http://www.preventionweb.net/english/hyogo/gar/2011/en/bgdocs/Cepeda_et_al._2010.pdf, (October), 20.

Chen, W., Pourghasemi, H. R., and Naghibi, S. A. (2017): A comparative study of landslide susceptibility maps pro-

duced using support vector machine with different kernel functions and entropy data mining models in China, Bulletin of Engineering Geology and the Environment, 77(2), 647–664. https://doi.org/10.1007/s10064-017-1010-y

ESRI (2014): Curvature function, , retrieved January 20, 2021, from internet: https://desktop.arcgis.com/en/arcmap/10.3/manage-data/raster-and-images/curvature-function.htm.

Evans, J. D. (1996): Straightforward Statistics for the Behavioral Sciences, California, 600.

Hadmoko, D. S., Lavigne, F., and Samodra, G. (2017): Application of a semiquantitative and GIS-based statistical model to landslide susceptibility zonation in Kayangan Catchment, Java, Indonesia, Natural Hazards, 87(1), 437–468. https://doi.org/10.1007/s11069-017-2772-z

Juliev, M., Mergili, M., Mondal, I., Nurtaev, B., Pulatov, A., and Hübl, J. (2019): Comparative analysis of statistical methods for landslide susceptibility mapping in the Bostanlik District, Uzbekistan, Science of the Total Environment, 653, 801–814. https://doi.org/10.1016/j.scitotenv.2018.10.431

Lee, S., and Talib, J. A. (2005): Probabilistic landslide susceptibility and factor effect analysis, Environmental Geology, 47(7), 982–990. https://doi.org/10.1007/s00254-005-1228-z

Liu, J., and Duan, Z. (2018): Quantitative assessment of landslide susceptibility comparing statistical index, index of entropy, and weights of evidence in the Shangnan Area, China, Entropy, 20(11), 9–11. https://doi.org/10.3390/e20110868

Mallick, J., Singh, R. K., AlAwadh, M. A., Islam, S., Khan, R. A., and Qureshi, M. N. (2018): GIS-based landslide susceptibility evaluation using fuzzy-AHP multi-criteria decision-making techniques in the Abha Watershed, Saudi Arabia, Environmental Earth Sciences, 77(7), 1–25. https://doi.org/10.1007/s12665-018-7451-1

Mandal, S., and Mandal, K. (2018): Bivariate statistical index for landslide susceptibility mapping in the Rorachu river basin of eastern Sikkim Himalaya, India, Spatial Information Research, 26(1), 59–75. https://doi.org/10.1007/s41324-017-0156-9

Melati, D. N., Umbara, R. P., Astisiasari, A., Wisyanto, W., Trisnafiah, S., Trinugroho, T., Prawiradisastra, F., Arifianti, Y., Ramdhani, T. I., Arifin, S., and Anggreainy, M. S. (2024): A comparative evaluation of landslide susceptibility mapping using machine learning-based methods in Bogor area of Indonesia, Environmental Earth Sciences, 83(3). https://doi.org/10.1007/s12665-023-11402-3

Mersha, T., and Meten, M. (2020): GIS-based landslide susceptibility mapping and assessment using bivariate statistical methods in Simada area, northwestern Ethiopia, Geoenvironmental Disasters, 7(1). https://doi.org/10.1186/s40677-020-00155-x

Moore, I. D., Grayson, R. B., and Ladson, A. R. (1991): Digital terrain modelling: A review of hydrological, geomorphological, and biological applications, Hydrological Processes, 5(1), 3–30. https://doi.org/10.1002/hyp.3360050103

Nohani, E., Moharrami, M., and Sharafi, S. (2019): Landslide Susceptibility Mapping Using Different GIS-Based Bivariate Models, Water MDPI, 11(1402), 1–22. https://doi.org/10.3390/w11071402

Othman, A. A., Gloaguen, R., Andreani, L., and Rahnama, M. (2018): Improving landslide susceptibility mapping using morphometric features in the Mawat area, Kurdistan Region, NE Iraq: Comparison of different statistical models, Geomorphology, 319, 147–160. https://doi.org/10.1016/j.geomorph.2018.07.018

Pham, B. T., Jaafari, A., Prakash, I., and Bui, D. T. (2018): A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling, Bulletin of Engineering Geology and the Environment, 78(4), 2865–2886. https://doi.org/10.1007/s10064-018-1281-y

Pourghasemi, H. R., Moradi, H. R., and Fatemi Aghda, S. M. (2013): Landslide susceptibility mapping by binary logistic regression, analytical hierarchy process, and statistical index models and assessment of their performances, Natural Hazards, 69(1), 749–779. https://doi.org/10.1007/s11069-013-0728-5

Pourghasemi, H. R., Moradi, H. R., Fatemi Aghda, S. M., Gokceoglu, C., and Pradhan, B. (2014): GIS-based landslide susceptibility mapping with probabilistic likelihood ratio and spatial multi-criteria evaluation models (North of Tehran, Iran), Arabian Journal of Geosciences, 7(5), 1857–1878. https://doi.org/10.1007/s12517-012-0825-x

Pradhan, B. (2010): Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches, Journal of the Indian Society of Remote Sensing, 38(2), 301–320. https://doi.org/10.1007/s12524-010-0020-z

Pradhan, B. (2013): A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS, Computers and Geosciences, 51, 350–365. https://doi.org/10.1016/j.cageo.2012.08.023

Rahardjo, P. P., Hosoda, T., and Handoko, A. (2017): Investigation of Landslides and Monitoring of The Subsequent Ground Movement and Geothermal New Pipe Lines Foundation in West Java, f the 19th International Conference on Soil Mechanics and Geotechnical Engineering, retrieved from internet: https://www.issmge.org/uploads/publications/1/45/06-technical-committee-02-tc102-20.pdf, 635–638.

Rasyid, A. R., Bhandary, N. P., and Yatabe, R. (2016): Performance of frequency ratio and logistic regression model in creating GIS based landslides susceptibility map at Lompobattang Mountain, Indonesia, Geoenvironmental Disasters, 3(1). https://doi.org/10.1186/s40677-016-0053-x

Razavizadeh, S., Solaimani, K., Massironi, M., and Kavian, A. (2017): Mapping landslide susceptibility with frequency ratio, statistical index, and weights of evidence models: a case study in northern Iran, Environmental Earth Sciences, 76(14), 1–16. https://doi.org/10.1007/s12665-017-6839-7

Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., and Guzzetti, F. (2018): A review of statistically-based landslide susceptibility models, Earth-Science Reviews, 180 (March), 60–91. https://doi.org/10.1016/j.earscirev.2018.03.001

Riaz, M. T., Basharat, M., Hameed, N., Shafique, M., and Luo, J. (2018): A Data-Driven Approach to Landslide-Suscepti-

bility Mapping in Mountainous Terrain: Case Study from the Northwest Himalayas, Pakistan, Natural Hazards Review, 19(4), 05018007. https://doi.org/10.1061/(asce)nh.1527-6996.0000302

Riley, S. (1999): Index that quantifies topographic heterogeneity, Intermountain Journal of Sciences.

Rózycka, M., Migoń, P., and Michniewicz, A. (2017): Topographic Wetness Index and Terrain Ruggedness Index in geomorphic characterisation of landslide terrains, on examples from the Sudetes, SW Poland, Zeitschrift Fur Geomorphologie, 61, 61–80. https://doi.org/10.1127/zfg_suppl/2016/0328

Sadisun, I. A., Telaumbanua, J. A., Kartiko, R. D., Dinata, I. A., and Pamela (2021): Weight of Evidence Method for Landslide Susceptibility Mapping in Sigi Biromaru, Central Sulawesi, IOP Conference Series: Earth and Environmental Science, 830(1), 6–12. https://doi.org/10.1088/1755-1315/830/1/012029

Samodra, G., Chen, G., Sartohadi, J., and Kasama, K. (2018): Generating landslide inventory by participatory mapping: an example in Purwosari Area, Yogyakarta, Java, Geomorphology, 306, 306–313. https://doi.org/10.1016/j.geomorph.2015.07.035

Samodra, Guruh, Chen, G., Sartohadi, J., and Kasama, K. (2017): Comparing data-driven landslide susceptibility models based on participatory landslide inventory mapping in Purwosari area, Yogyakarta, Java, Environmental Earth Sciences, 76(4), 1–19. https://doi.org/10.1007/s12665-017-6475-2

Sassa, K. (2007): Landslide science as a new scientific discipline, Progress in Landslide Science, (1978), 3–11. https://doi.org/10.1007/978-3-540-70965-7_1

Schlögel, R., Marchesini, I., Alvioli, M., Reichenbach, P., Rossi, M., and Malet, J. P. (2018): Optimizing landslide susceptibility zonation: Effects of DEM spatial resolution and slope unit delineation on logistic regression models, Geomorphology, 301, 10–20. https://doi.org/10.1016/j.geomorph.2017.10.018

Shahzad, N., Ding, X., and Abbas, S. (2022): A Comparative Assessment of Machine Learning Models for Landslide Susceptibility Mapping in the Rugged Terrain of Northern Pakistan, Applied Sciences (Switzerland), 12(5). https://doi.org/10.3390/app12052280

Sukristiyanti, S., Wikantika, K., Sadisun, I. A., Yayusman, L. F., and Telaumbanua, J. A. (2021): Polygon-based Landslide Inventory for Bandung Basin Using Google Earth, Indonesian Journal of Geography, 53(2), 285–294. https://doi.org/http://dx.doi.org/10.22146/ijg.58014

Sumaryono, S., Muslim, D., Sulaksana, N., and Dasatriana, Y. (2015): Weights of Evidence Method for Landslide Susceptibility Mapping in Tandikek and Damar Bancah, West Sumatra, Indonesia, International Journal of Science and Research (IJSR), 4(10), 1283–1290.

Sun, X., Chen, J., Bao, Y., Han, X., Zhan, J., and Peng, W. (2018): Landslide susceptibility mapping using logistic regression analysis along the Jinsha river and its tributaries close to Derong and Deqin County, southwestern China, ISPRS International Journal of Geo-Information, 7(11), 1–29. https://doi.org/10.3390/ijgi7110438

Umar, Z., Pradhan, B., Ahmad, A., Jebur, M. N., and Tehrany, M. S. (2014): Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio and logistic regression models in West Sumatera Province, Indonesia, Catena, 118(September 2009), 124–135. https://doi.org/10.1016/j.catena.2014.02.005

URL 1: https://earthexplorer.usgs.gov/ (accessed 29th September 2020)

Wati, S. E., Hastuti, T., Widjojo, S., and Pinem, F. (2010): Landslide susceptibility mapping with heuristic approach in mountainous area a case study in Tawangmangu Sub District, Central Java, Indonesia, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science, XXXVIII, 248–253. https://doi.org/10.1016/j.geomorph.2006.10.032

Yao, X., Tham, L. G., and Dai, F. C. (2008): Landslide susceptibility mapping based on Support Vector Machine: A case study on natural slopes of Hong Kong, China, Geomorphology, 101(4), 572–582. https://doi.org/10.1016/j.geomorph.2008.02.011

Zhu, A. X., Miao, Y., Wang, R., Zhu, T., Deng, Y., Liu, J., Yang, L., Qin, C. Z., and Hong, H. (2018): A comparative study of an expert knowledge-based model and two data-driven models for landslide susceptibility mapping, Catena, 166(April), 317–327. https://doi.org/10.1016/j.catena.2018.04.003

SAŽETAK

Učinkovitost bivarijantnih statističkih modela u kartiranju podložnosti na klizanje (studija slučaja: podslijev Cisangkuy, Bandung, Indonezija)

Pojave klizišta česte su u brdsko-planinskim područjima, posebice u tropskim zemljama s velikom količinom oborina i intenzivnim vremenskim utjecajima. Kartiranje podložnosti na klizanje (LSM) prvi je korak u ublažavanju opasnosti od klizišta. Ovim istraživanjem provedena je usporedba četiriju LSM modela, odnosno metode omjera učestalosti (FR), metode informacijske vrijednosti (IVM), modela težinskih faktora (WoE) i Shannonove entropije (SE), za podslijev Cisangkuy, Zapadna Java. Navedene metode uzimaju u obzir odnos između gustoće klizišta i faktora klizanja. Za modeliranje je korišteno 76 piksela klizišta i 15 faktora klizanja. 70 % klizišta korišteno je kao podatak za treniranje modela, a ostatak je korišten za validaciju modela. Od 27 analiziranih preduvjeta klizanja odabrano je ukupno 15 faktora klizanja. Faktori koji pokazuju visoku kolinearnost uklonjeni su kako bi se izbjegla multikolinearnost podataka. Osim toga, u modeliranje su uključeni samo faktori klizanja koji se odnose na prikupljene podatke o klizištima. Za validaciju modela korištena je krivulja radnih karakteristika prijamnika (ROC) i metoda indeksa gustoće klizišta (LDI). Svi modeli pokazuju odgovarajuće stope predviđanja za FR, IVM, WoE i SE metode, koje iznose 0,770, 0,790, 0,793 odnosno 0,788. Na temelju LDI analize, LDI vrijednosti nisu postupno rasle od vrlo niske do vrlo visoke klase osjetljivosti za svaku LSM kartu. Međutim, karte su još uvijek povoljne jer klase koje su najpodložnije na klizanje u svim modelima imaju najviši LDI. Na učinkovitost modela može utjecati broj klasa i metoda klasifikacije koje se koriste za kategorizaciju kontinuiranih parametara, kao i mala količina podataka o klizištima u inventaru.

Ključne riječi:
modeliranje podložnosti na klizanje, bivarijantna statistička metoda, ROC krivulja, indeks gustoće klizišta

Author's contribution

Sukristiyanti (1) (Dr., junior researcher, GIS specialist for landslides) provided the four bivariate statistical analyses for landslide susceptibility mapping and wrote the paper. **Pamela (2)** (Master, first mapping surveyor, GIS specialist for landslides) was involved in the development of the research methodology. **Sitarani Safitri (3)** (Dr., senior researcher, GIS specialist) contributed in preparing the manuscript, presentation of the results, and proof-reading works. **Ahmad Luthfi Hadiyanto (4)** (Dr., junior researcher, remote sensing specialist) reviewed and gave feedback for the manuscript. **Adrin Tohari (5)** (Dr., principal researcher, geotechnical specialist) reviewed and gave feedback for the manuscript. **Imam Achmad Sadisun (6)** (Dr., associate professor, specialist on engineering geology, rock mechanics, and landslide geology) is a supervisor and reviewed the manuscript.