

Flood susceptibility assessment using artificial neural networks in Indonesia

Stela Priscillia^{a,*}, Calogero Schillaci^b, Aldo Lipani^a

^a Department of Civil, Environmental and Geomatic Engineering, University College London (UCL), United Kingdom

^b European Commission, Joint Research Centre, Italy

ARTICLE INFO

Keywords:

Radar remote sensing
Topography
DEM
CNN
Segmentation
Flood
Water

ABSTRACT

Flood incidents can massively damage and disrupt a city economic or governing core. However, flood risk can be mitigated through event planning and city-wide preparation to reduce damage. For, governments, firms, and civilians to make such preparations, flood susceptibility predictions are required. To predict flood susceptibility nine environmental related factors have been identified. They are elevation, slope, curvature, topographical wetness index (TWI), Euclidean distance from a river, land-cover, stream power index (SPI), soil type and precipitation. This work will use these environmental related factors alongside Sentinel-1 satellite imagery in a model intercomparison study to back-predict flood susceptibility in Jakarta for the January 2020 historic flood event across 260 key locations. For each location, this study uses current environmental conditions to predict flood status in the following month. Considering the imbalance between instances of flooded and non-flooded conditions, the Synthetic Minority Oversampling Technique (SMOTE) has been implemented to balance both classes in the training set. This work compares predictions from artificial neural networks (ANN), k-Nearest Neighbors algorithms (k-NN) and Support Vector Machines (SVM) against a random baseline. The effects of the SMOTE are also assessed by training each model on balanced and imbalanced datasets. The ANN is found to be superior to the other machine learning models.

1. Introduction

When compared to other meteorologically driven natural disasters, flooding is one of the most disastrous phenomena as it can lead to massive damage to both properties and fatalities worldwide (Hinkel, 2020; Tehrany et al., 2015). Indonesia is a country that has significant flood risk in most of its provinces, with the islands of Java, Bali, and Sumatera as the largest at-risk areas (Bappenas, 2010). Jakarta, officially called Special Capital Region of Jakarta, is the capital city of Indonesia and is located on Java Island. Jakarta has a population of over 10 million inhabitants (as of 2019) within a 662.2 km² area, as a result Jakarta is the most densely populated province in Indonesia (BPS Jakarta, 2019).

Jakarta has been known to flood since the 5th century AD (Widyaningrum, 2019) and modern Jakarta is still at risk due to being located on a deltaic plain crisscrossed by 13 natural rivers and more than 1400 km of man-made waterways. About 40 percent of the city is below sea level, which contributes to flood risk (Baker, 2012). A range of factors including climate change accelerated sea-level rise, extreme meteorological events, along with land subsidence from multiple causes are worsening the flood risk in Jakarta (Abidin et al., 2011; World Bank, 2019).

On January 1, 2020, Jakarta was hit by the most intense rainfall event (peaking at 377 mm/day) since records began in 1866 (BMKG, 2020). This several-day rainfall event resulted in widespread flooding causing 16 deaths, and displacing more than 30,000 people (BNPB, 2020). The severity of the flood event caught city officials off-guard, with the huge number of affected persons paralyzing Indonesia's main economic and political hub. Therefore, to mitigate such incidents and allow a timely preparation, it is paramount to develop an ability to predict and evaluate the areas of Jakarta that are most susceptible to flooding.

In recent years, machine-learning models, particularly artificial neural networks (ANN), have been increasingly used for flood susceptibility modelling. Previous research has used precipitation and other environmental factors as input to ANNs and to output water levels and an inferred flood map (Kia et al., 2012). Although the previously mentioned flood events were mainly caused by the heavy rainfall (BMKG, 2020), a range of researchers have proven that several other factors have contributed to the disastrous flooding in the area. These include elevation, curvature, stream power index (SPI), topographic wetness index (TWI), soil type, distance from rivers, and others (Kia et al., 2012; Tehrany et al., 2015).

* Corresponding author.

E-mail addresses: stela.priscillia.20@ucl.ac.uk (S. Priscillia), calogero.schillaci@ec.europa.eu (C. Schillaci), aldo.lipani@ucl.ac.uk (A. Lipani).

<https://doi.org/10.1016/j.aiig.2022.03.002>

Received 6 January 2022; Received in revised form 19 March 2022; Accepted 19 March 2022

Available online 8 April 2022

2666-5441/© 2022 The Authors. Publishing Services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

There are several factors that could cause flood which have been used as both fixed and temporal attributes to predict flood susceptibility. However, it is still unknown if those factors combined and inputted into an ANN model will be able to create a reliable result that could predict flood susceptibility in Jakarta. Therefore, the authors of this work seek to develop a suite of machine learning models that uses as inputs these environmental factors to predict flood susceptibility in Jakarta but in a novel manner with a temporal aspect.

A novel experimental methodology is used to predict flood susceptibility. Commonly used environmental flood risk factors (Tehrany et al., 2015; Shahabi, 2020) have been used in classification tasks to predict flood susceptibility for multiple areas using all historical flooding. The novel addition is that unlike these flood susceptibility studies; a rainfall time series has been included allowing susceptibility predictions across each area for each month. This temporal element is similar to the study of Kia et al. (2012), which utilizes the same environmental flood risk factors and a rainfall time series in a regression task to predict a flow hydrograph at the bottom of the study area when using an ANN. These flow predictions can then be converted to a river stage (water height) reading and then used to infer flooded areas through ‘bathtub approaches’ in a Geographic Information System (GIS). This approach is not possible in the Jakarta study area due to the shortage of gauging data (Kure et al., 2013) and insufficient lead times for predictions based on upstream water levels (Miyamoto et al., 2012). Therefore, this study uses ANNs, and other MLs to skip the prediction of the hydrograph and go straight to predicting whether an area was flooded or not.

This work aims to create a tool which can be used to prioritize allocation of flood defense resources for city officials and inform civilians of flood risk when extreme rainfall is forecasted. This research will also aim to find an optimal model through comparing: ANNs, k-NN, and SVM models against a random baseline. Novel approaches such as those presented in this work are encouraged in natural hazard susceptibility modelling (Chen et al., 2019a) as there is no universal consensus about what the best modelling tool is (Chen et al., 2019b). An overview of models previously employed in flood susceptibility studies can be found in Islam et al. (2021). The work comprises three key topics; the

environmental flood risk factors in the study area, Sentinel-1 Synthetic Aperture Radar (SAR) imagery used to monitor floods, and the machine learning models used to predict flood susceptibility.

2. Material and methods

2.1. Environmental flood risk factors

The environmental flood risk factors can be divided according to their temporally static or varying natures. The eight static factors identified in similar studies (Kia et al., 2012; Tehrany et al., 2015; Vojtek, 2019) are (see Fig. 1 for the maps produced for each attribute): elevation, slope, curvature, stream power index (SPI), topographic wetness index (TWI), distance from river, land cover and soil type. The main temporally variable factor determining flood-susceptibility is precipitation recorded as daily time steps (Bappenas, 2010).

Elevation. It is a key factor in determining whether land is susceptible of flooding. Data source is the National DEM (see Table 1 for details). Regions at higher elevations are less likely flooded when compared to those at lower elevations (Botzen et al., 2012). Globally, regions with low-lying areas flood much more frequently, with average annual risk of large flood at around 4.9% for cities that are up to 10 m above sea level (Kocornik-Mina, 2016). This is a much higher average annual chance of flooding than cities located above 10 mamsl which only have a 1.3% chance of flooding each year (Kocornik-Mina, 2016).

Slope Degree. It is calculated as the first-order derivative of elevation and it is another key terrain-derived factor affecting flood risk (Tehrany et al., 2015). As ground slopes increase and runoff travels faster downhill this results in a decrease in the time available for surface infiltration (Schillaci et al., 2015). The decrease in surface infiltration allows more runoff to enter the rivers and drainage systems and can be the most critical factor in determining flood risk (Vojtek, 2019).

Slope Curvature. The curvature of the slope, the second order derivative of elevation, plays an important part in determining flood susceptibility. Curvature contains information on the erosive characteristics of water that include topography-controlled runoff conditions and it has

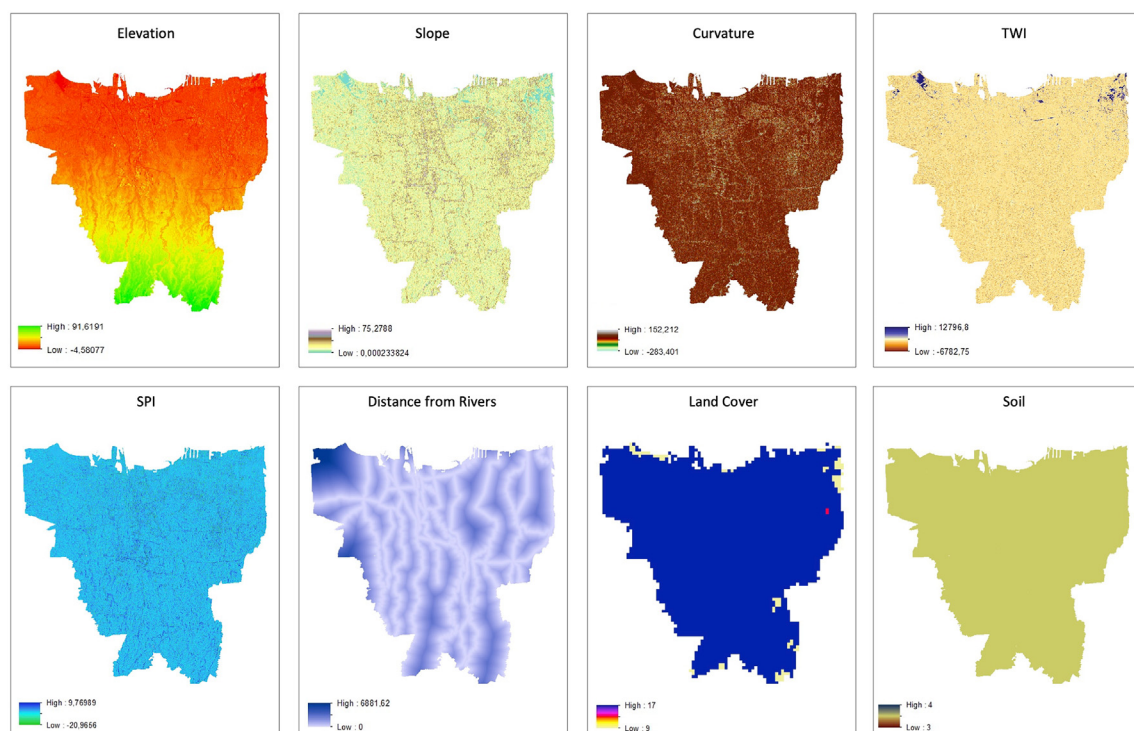


Fig. 1. Raster image for fix attributes.

Table 1
Summary of Environmental Factors.

Factor	Description	Source
Elevation	Terrain profile represented as height above reference datum.	2015 Jakarta DEM from Jakarta's Regional Disaster Relief Agency (BPBD) Geoportal.
Slope Degree	The first-order derivative of elevation.	Derived from Jakarta DEM.
Slope Curvature	The second order derivative of elevation.	Derived from Jakarta DEM.
TWI	A metric of moisture likelihood based on the principle of upstream contributing area. $TWI = \ln \frac{\alpha}{\tan \beta}$ α (m ² m ⁻¹) is the local upslope area draining through a certain point per unit contour length. $\tan \beta$ (radian) is the local slope gradients. (Kirkby and Beven, 1979; Moore et al., 1991).	Derived from Jakarta DEM.
SPI	A metric of water flow strength for erosive processes. $SPI = \alpha \times \tan \beta$ α and $\tan \beta$ same as above	Derived from Jakarta DEM.
Dist. from River	Euclidean planar distance.	Derived from Jakarta River Map from BPBD Geoportal.
LC	Categorizing the Global Land Cover Characterization (GLCC) types in the study area.	USGS, https://www.usgs.gov/special-topics/land-use-land-cover-modeling .
Soil Type	Soil Units in the Jakarta region, Inceptisol; Oxisol; Ultisol; Entisol; Histosol.	FAO Digital soil map report, polygons rasterized, scale 1:5000000.
Precipitation Time Series	Jakarta monthly average rainfall November 2015 until December 2020.	Indonesian Agency of Meteorology, Climatology and Geophysics (BMKG) database.

been found to be positively correlated with flood susceptibility in similar studies (e.g., Mind'je et al., 2019).

Stream Power Index. The stream power index (SPI) is a metric of water flow strength for erosive processes. SPI is determined by processes of catchment hydrology, hydraulic channel-geometry and water's conservation of mass, shear stress, flood interval, climate, and bedrock erodibility among others (Irawan et al., 2021). The SPI will be estimated based on the slope at each location and the contributing upstream area.

Topographical Wetness Index. The topographic wetness index (TWI) is a metric of moisture likelihood based on the principle of upstream contributing area. The TWI is calculated through spatially distributed terrain models or by using contours derived from these models. The TWI is also proven to be an early-stage indicator of flood-prone areas in residential development planning (Pourali et al., 2016). More advanced applications of the TWI include a calculation method using a maximum likelihood estimation procedure (Motevalli and Vafakhah, 2016). In this study, the TWI was shown to be a useful light weight alternative to hydraulic modelling as a high correlation between TWI with flood depth from a hydraulic model were observed (Motevalli and Vafakhah, 2016).

Distance from rivers. Across natural catchments, the distance from rivers or waterbodies has been proven to be the most dominating factor that influences flood-susceptibility (Giovannettone et al., 2018). In urbanized catchments with more complex drainage, the distance from waterbodies is still the second-most important factor behind elevation (Giovannettone et al., 2018). The flood susceptibility mechanism relates to the planar distance the waterbody must grow by to encroach on land, where the closer to an area is to a river or a body of water, the more likely that area is to flood (Glenn et al., 2012).

Land Use Land Cover. The Global Land Cover Characterization (GLCC) is another factor potentially linked to flood susceptibility. The data source is the United States Geological Survey (USGS) global land

cover product at 1 km spatial resolution, accessed from: <http://earthexplorer.usgs.gov/>. Flood likelihood and areas of high vegetation density have been noted to be negatively correlated (Tehrany et al., 2015). Vegetated areas have higher interception and infiltration rates of rain-water resulting in less surface runoff. Conversely non-vegetated areas tend to have higher surface runoff and therefore greater flood susceptibility (Lee et al., 2012). Cases of vegetation decrease, or urbanization, have been proven to be positively correlated to flood frequency and damage costs (Adnan et al., 2020).

Soil Type. Working on the same principle of controlling runoff rates, the soil type is another important factor affecting flood susceptibility. The data source is the FAO Soil Map of the World at 1:5,000,000 scale. The project was completed over a span of twenty years and uses the Soil taxonomy framework to classify soil units, accessed from: <https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/faounesco-soil-map-of-the-world/en/>. In the Jakarta study area however, the flood mechanism soil type exerts is through the differing subsidence rates of different soils (Abidin et al., 2011). Natural consolidation of alluvial soils coupled with heavy loads due to construction results in highly compressed soil during subsidence (Abidin et al., 2011). In addition to reducing elevation, the compression of sediments results in less infiltration (Herda Adeline et al., 2020) and therefore greater likelihood of flooding (van Leeuwen et al., 2019).

Precipitation Time Series. Differing from the aforementioned eight factors, the final factor of precipitation is a temporal factor. According to Indonesian National Atlas (ANI), Java has the most extreme precipitation rates of up to 400 mm/month (Bappenas, 2010). As mentioned earlier the January 2020 floods occurred during the most extreme rainfall in the city since 1866 (BMKG, 2020). Furthermore, monthly rainfall intensity in Jakarta has increased 100 mm from 1900 to 1970. These monthly increases have been consistent throughout the year, as well as during peak rainfall seasons (Bappenas, 2010). Finally in pluvial and fluvial flooding, the rate of water input to the system drives the response from the catchment. Therefore, rainfall intensity has been included as a time series rather than as a 'static' yearly/decadal average.

The temporally fixed environmental flood risk factors for the study area are displayed in Figs. 1 and 2.

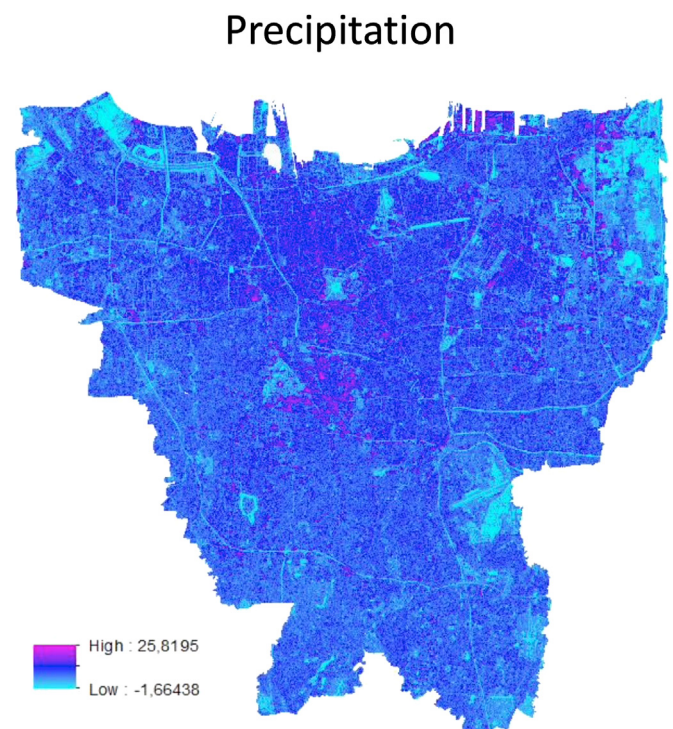


Fig. 2. Sentinel-1 (mean) Jan 2020.

2.2. Sentinel-1 imagery

Imagery from the Sentinel-1 C-band radar imaging satellite has been used to provide flood extent information during the January 2020 flood event. This active remote sensing system provides regular coverage through both day and night, and it is not affected by atmospheric processes such as cloud cover and so has been widely used to monitor flood events (Conde and Munoz, 2019; Tsyganskaya et al., 2019; Martinis et al., 2018; Qiu et al., 2021).

The Sentinel-1 imagery has been sourced through Google Earth Engine (GEE, Gorelick, 2017). Using the GEE cloud platform's data catalogue allows rapid and computational light access to the study area. These characteristics mean that GEE is suitable for emergency response applications, and it has been used in a range of formats in flood response systems (Cheng-Chien et al., 2018; De Vries, 2020; Pourghasemi et al., 2021).

Sentinel-1 imagery has been included as an input variable because could be used by the model as a proxy to sense the overall presence of water across the study area. This is through soil moisture of non-urbanized areas, vegetation intensity, and in terms of prevalence of open bodies of water. Antecedent soil moisture, or channel-capacity conditions are particularly important in determining the response of an area during successive rainfall events (Truckenbrodt et al., 2019). Sentinel-1 scenes of each month were averaged and as a result the value for each individual pixel is different.

2.3. Machine learning models for classification

Machine learning (ML) is a group of computational algorithms and statistical models which are designed to learn from their environment (Naqa and Murphy, 2015). Under a supervised learning paradigm, a model learns to improve its performance at a given task through experience. The following models are commonly used in predictive classification tasks and a brief overview of their learning mechanisms are provided. These models are Support Vector Machines (SVMs), k-Nearest Neighbors algorithms (k-NNs), and Artificial Neural Networks (ANNs).

2.3.1. Support Vector Machines

SVM is a model suitable for two-class classification problems and is based on the principle of the minimization of structural risk (Belousov et al., 2002). SVMs aim to create a hyper-plane to separate the two classes in a n-dimensional space, as determined by n-attributes (Marjanović et al., 2011). The SVM can be mathematically represented below as a primal optimization problem (Chang and Lin, 2011):

$$\min_{w,b,\epsilon} \frac{1}{2} w^T w + C \sum_{i=1}^l \epsilon_i$$

$$\text{subject to } y_i(w^T \Phi(x_i) + b) \geq 1 - \epsilon_i, \\ \epsilon_i \geq 0, i = 1, \dots, l,$$

Where the $y \in \mathbb{R}^l$ response vector can assume values -1 or 1 . $\Phi(x_i)$ maps each attribute, $x_i \in \mathbb{R}^n$, into a higher dimensional space. $C \geq 0$ is the penalty parameter on the error term.

In addition to finding a maximum margin of separation between classes (Pradhan, 2013), a unique advantage of the SVM is that a tolerance on misclassification can be specified for non-linearly separable problems. This width of the soft margin tolerance is a tunable cost (C) parameter where decision boundaries with narrower margins utilize fewer data points (support vectors) to separate the two classes.

Another way of solving non-linear problems is by employing a non-linear kernels in SVMs to plot non-linear decision boundaries (Cao et al., 2008). By replacing the $x_i^T x_j$ term with a kernel function, $k(x_i^T x_j)$,

the data can be mapped into a higher dimensional space. One such kernels which has been found to be successful in many problem domains is the Radial Basis Function (RBF) kernel (Yao et al., 2008), where $(x_i^T x_j) = \exp(-\gamma \|x_i - x_j\|^2)$. The RBF kernel has a tunable parameter (γ) controlling the kernel width and sphere of influence of individual data points upon the decision boundary (Chang and Lin, 2011).

2.3.2. k-Nearest Neighbors Algorithms

k-NN models simply classify new data points according to the most similar data point observed during training. Multiple neighboring points can be used to measure similarity through a distance weighted score or majority voting (Cunningham and Delany, 2007). The probability of a new instance belonging to a class corresponds with the proportion of k-nearest neighboring points with each class label, as shown below with Ω_0 being the set of k-nearest neighbors to x (Aristizabal et al., 2020).

$$P(Y=y|X=x) = \frac{1}{k} \sum_{j \in \Omega_0} I(y_j=y)$$

$$\text{where } I(y_j=y) = \begin{cases} 0 & \text{when } y_j \neq y \\ 1 & \text{when } y_j = y \end{cases}$$

The tunable hyperparameter k , the number of neighboring data points to consult, is commonly determined by using the square root of N , where N is the number of points in the dataset. The other hyperparameter that needs to be tuned is the distance metric used to find the nearest data points. Four distance metrics have been tested. They are the Euclidean, Manhattan, Cosine, and Minkowski distances. An overview of these metrics is provided in Alfeilat et al. (2019).

2.3.3. Artificial Neural Networks

ANNs are complex ML models which are comprised of networks of interconnected nodes (or neurons). Each neuron takes a vector of feature values as input. Then, it multiplies a column-vector of weights to this input vector and adds an additional weight named bias. Finally, the result of this multiplication is input to an activation function and its result returned as output. In a network of connected neurons information flows between each neuron from the model inputs to the model classification output. By incorporating multiple neurons to build networks of arbitrary width and depth, ANNs of arbitrary complexity can be built in response to the complexity of the problem. The network structure is only one of the tunable hyperparameters (Atkinson and Tatnall, 1997). Other hyperparameters include the activation function used within each neuron to combine the inputs into an output (Kia et al., 2012). The Sigmoid function is commonly used (Bishop, 1994). Other activation functions which were tested during hyperparameter-tuning include the; Identity, Logistic, ReLU, and Tanh functions (Agatonovic-Kustrin and Beresford, 2000).

The weights for each neurons are learned during training through backpropagation. Two more hyperparameters of an ANN are the learning rate and the α regularization term. The learning rate controls the size of the update to neuron weights during training (Gurney, 1997). The α regularization term constrains the size of the weights to prevent overfitting to the training data (Gurney, 1997).

ANNs are favored in a range of problem domains due to their robust capability in modelling nonlinear relationships between explanatory variables and target variables (Kia et al., 2012). The ANN differs from the other two ML models as an ANN, when having many layers, can learn new features automatically (deep learning). The SVM and k-NN models treat each data point's attributes as a vector with equal weight assigned to each. This means that these two learners can suffer from the curse of dimensionality (Bellman, 1958) but also cannot make context-aware predictions that ANNs can. This is where in different contexts and areas of the parameter space, different attributes can have differing levels

of importance on the final prediction (Anderson-Bell et al., 2021; James et al., 2021).

2.4. Methodology

2.4.1. Experiment design

The models take as input the environmental factors flood risk factors for each month (T) and will make predictions as to whether a location flooded or did not flood in the following month ($T + 1$). The input attributes span from November 2014–November 2020 and the label attributes span from December 2014–December 2020.

As computational resource is limited, rather than running the model for every grid cell in the study area for five years, the modelling has been undertaken at the polygonal administrative unit of the village. Jakarta is comprised of 260 administrative village units as displayed in Fig. 3.

To prove that the models are better than random, they will be compared against a random baseline, i.e., that randomly guesses whether a location was flooded or not. This model represents the theoretical worst performance a ML model can attain on a binary classification task.

2.4.2. Data preparation

Because only monthly precipitation data is available the modelling has been undertaken at the monthly timestep. Choosing this unit allowed the inclusion of composite Sentinel-1 images as acquisition dates varied for each month in each year. For each calendar month between November 2014 and December 2020, available Sentinel-1 scenes have been averaged to monthly composites. Because of the strong meteorological signal of the Monsoon season in Jakarta (BMKG, 2021), each month can be broadly typified as in the ‘dry season’ (May–September) or ‘rainy season’ (October–April).

The labels for the dataset were provided through private communication with the BMKG. Each administrative village area is classified as flooded or not flooded in each month. The centroids of the 260 Villages were matched up with the static environmental flood risk factors to assign attributes to each village. The temporal rainfall values and spatio-temporal composite Sentinel-1 values at each village centroid were added to create a monthly time series for each village. The 10-attribute time series for each village has been combined into a single dataset with an example shown below.

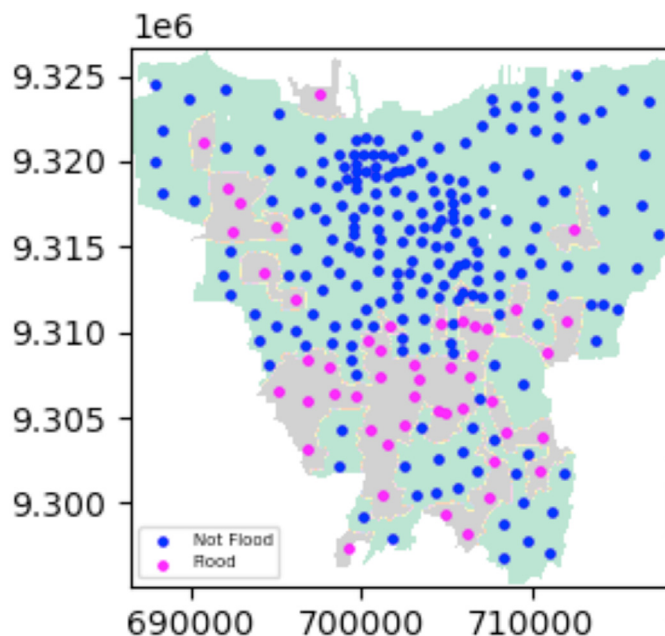


Fig. 3. Distribution of the 260 villages in the study area. Each dot is a polygon centroid.

As the models used cannot handle missing values, records which did not have full attribute values have been dropped. This has led to underrepresentation in 2014 where only two months (November and December) contained full data. This resulted in a slightly smaller dataset ($\sim 0.08\%$).

We examined variable collinearity and only a moderate correlation (0.63) was noted between variables. Temporal autocorrelation was not assessed as the models used both temporal and fixed variables to make their predictions.

As raw values range dramatically between attributes, min-max scaling has been used for all models to improve model performances, particularly for the ANN. This is needed as the gradient descent algorithm converges much faster when feature scaling or normalization is implemented (Ioffe, 2015).

2.4.3. Model training

The final dataset of 18,965 records (in Table 2 we show two samples) has been divided into three sets of: training, validation, and test sets. The training and validation sets are randomly sampled from records of the 73 months between 2014 and 2019 and comprise of 67.93% and 16.98% of the full dataset. The test set has been created from records in the 12 months of 2020, which represents 15.08% of the full dataset.

As floods do not happen every day, the combined training and validation datasets contains an imbalance with significantly more instances of the non-flood class (91.15%), than the flood class (8.85%). As highly imbalanced training datasets produced biased classifiers (Jiawei et al., 2012) the Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla et al., 2002) was implemented to oversample the minority class to the same prevalence as the majority class. SMOTE firstly identifies examples of the minority class that are nearby in the attribute feature space. Secondly SMOTE draws lines between them, and samples new synthetic points along these lines. New examples are generated as a combination of two chosen examples. A downside of this approach is the potential creation of ambiguous examples should the two classes overlap in the attribute feature space. For more information see Chawla et al. (2002).

All SVM and k-NN model hyperparameters have been tuned via a grid search using a 5-fold cross validation combining the training and validation set. As ANNs take longer to train, the tuning of their hyperparameters has been performed by using the validation set. See Table 3 and 4 for the details about the tuned hyperparameters.

2.4.4. Performance metrics

Given the sparsity of the natural phenomenon we want to predict – the dataset contains far fewer flood instances (positive class) than non-flood instances (negative class) – we evaluated our classifiers using Precision, Recall and the F1-Score.

The F1-Score is the harmonic mean of Precision and Recall with 0 being the worst value and 1 the best value. Precision is defined as the proportion of predicted positive samples that are true positive. Recall is proportion of true positive that have been correctly predicted as positive class.

Following we report the equations used to compute Precision, Recall and the F1-Score, where we consider as the positive class the flood instances:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = \frac{2TP}{2TP + FP + FN} = 2 \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP, FN and FP are the true positives, false negatives, and false positives.

Table 2

Two example records in the dataset for two different villages taken during November 2014.

Long	Lat	Year	Mon.	Alt.	Slope	Curv.	SPI	TWI	Soil	Riv. Dist.	LULC	Rain-fall	Sentinel-1	Flood
xx	xx	2014	11	12	2.5	0.8	0.2	89	3	913	13	0.6	13.45	1
xx	xx	2014	11	9	8.9	0.3	−1.1	−44	3	1319	13	0.6	16.83	0

Table 3

List of tuned hyperparameters.

Model	Tuned Hyperparameters
k-NN	<ul style="list-style-type: none"> Weight: uniform, distance Number of neighbors: range 1–11 Distance Metric: Euclidean, Manhattan, cosine, Minkowski
SVM	<ul style="list-style-type: none"> Kernel: linear, RBF C: range 1–11
ANN	<ul style="list-style-type: none"> Hidden layer sizes Maximum epochs Activation function Learning rate: Adaptive or Fixed

Table 4

The optimal hyperparameters used in each model.

Tuned Hyper parameters implements		
Dataset	Imbalanced Training data	Balance training data
k-NN	<ul style="list-style-type: none"> Weight: distance Number of neighbors: 1 Distance Metric: cosine 	<ul style="list-style-type: none"> Weight: distance Number of neighbors: 4 Distance Metric: Manhattan
SVM	<ul style="list-style-type: none"> Kernel: linear, RBF C: 9 	<ul style="list-style-type: none"> Kernel: linear, RBF C: 10
ANN	<ul style="list-style-type: none"> Hidden layer sizes: 1 hidden layer of 20 nodes Maximum epochs: 400 Activation function: logistic sigmoid Optimizer: Adam Alpha: 0.01 Learning rate: Adaptive 	

3. Results

The following results are reported on models trained on both the original imbalanced dataset and the balanced dataset using SMOTE. This has been done to assess the effect of SMOTE on the three models. Model results are presented on the training, validation, and test sets. The results for the random baselines have also been presented alongside.

In Table 5, we observe that the performance of all models using the imbalanced dataset is poor. In particular, both SVM and ANN are unable to generalize and learn to predict always that there is no flood. When trained on the balanced dataset, the ANN was found to performs the highest (0.45 F1) and the only model to perform better than the random classifier (0.24 F1).

4. Discussion

The effects of the SMOTE to rebalance the training datasets is important when predicting floods. The effect of this technique is clear when comparing the ANN trained with the imbalanced set vs. the balanced set. However, this contrasts with the SVM and k-NN models in which rebalancing the training data did not have any significant effect. The k-NN model performed minimally better in predicting flood when trained on the rebalanced data, while the k-NN model's score for the non-flood class remained the same. In the SVM model the scores for both classes remained unchanged when training on the balanced dataset.

The difference in performance we observe between the validation and test sets could be attributed to the fact that the validation set consists of datapoint between 2014 and 2019, while the test set consists only of datapoints in 2020. Another reason for this difference in performance could be attributed to the fact that the rainfall and flooding in the year

Table 5

Model performances. Note that the test set is never oversampled.

Model	Metric	Imbalanced Sets			Balanced Sets (with SMOTE)		
		Train	Valid.	Test	Train	Valid.	Test
Random	P	–	–	0.15	–	–	0.15
	R	–	–	0.52	–	–	0.52
	F1	–	–	0.24	–	–	0.24
k-NN	P	1.00	0.33	0.56	1.00	0.87	0.32
	R	1.00	0.31	0.01	1.00	0.96	0.02
	F1	1.00	0.32	0.02	1.00	0.92	0.03
SVM	P	0.73	0.92	0.00	0.77	0.76	0.00
	R	0.03	0.05	0.00	0.86	0.86	0.00
	F1	0.06	0.09	0.00	0.81	0.80	0.00
ANN	P	0.00	0.00	0.00	0.94	0.67	0.37
	R	0.00	0.00	0.00	0.66	0.75	0.59
	F1	0.00	0.00	0.00	0.75	0.71	0.45

2020 was higher than experienced previously. ML models struggle to extrapolate outside of their known range of their training data if the data distribution used for testing is different from the one used for training. Should ML models be used in a real scenario, it is important to recognize this limitation and retrain the model periodically.

Overall, only the ANN was able to perform better than random when predicting floods demonstrating that by using this model we could predict which villages are susceptible to flooding in the upcoming month. The novel nature of this work unfortunately limits its comparability to the wider literature. While the environmental input factors used are common throughout the ML flood susceptibility literature (e.g., [Kia, et al., 2012](#); [Shahabi, 2020](#); [Tehrany et al., 2015](#)) the setup of predicting flood susceptibility classification across time is novel. The higher performance of the ANN is suspected to be due to the complexity of the classification problem faced:

Firstly, the problem is non-linear and suited to a mixed-effects model where the relationships between input variables and output vary across space (each village) and across time (month during the monsoon cycle). Secondly, when predicting flood status in month $T + 1$ using environmental flood risk factors in month T , the strength of the relationship to the following month's flood status will vary depending on the month's position in the monsoon cycle. In months heading towards the monsoon peak the same volume of rainfall would be expected to be correlated with greater flooding than the same rainfall volume in a month heading towards the dry season peak.

The model might be improved in the future; this exercise offered an overview of the ANN application with selected attributes that are strictly related with floods in an area that suffered in the past for similar events. In Jakarta, the flood is predicted yearly with flooding risk rating all the villages together. The data shows that all villages have a medium risk for all month in year 2020. Using a model that can give predictions for each of the village on a monthly basis, government, firms, civilians can be on alert and prepare the disaster recovery plan.

5. Conclusions

A flood susceptibility study comparing k-NN, SVM, and ANN models was undertaken for the Jakarta study area. The model was trained on 2014–2019 monthly data and tested on 2020 data across Jakarta's 260 administrative villages. Two sets of models' groups were ran using the original dataset and the rebalanced dataset with SMOTE.

This analysis shows the importance of rebalancing the dataset. The best scores were obtained in the ANN model. The SVM and k-NN models performed similarly poorly across both imbalanced and balanced datasets. The novelty of the approaches used in this study limits its comparability to the literature but the context-specific spatial and temporal patterns within the problem are theorized to be the reason why the ANN model had the highest performance. The SVM and k-NN models are suspected to suffer from the curse of dimensionality.

Further work could examine a spatial breakdown of the models' performance to identify if some villages consistently perform worse than others. Similarly, a temporal breakdown could be undertaken to examine months or seasons of poorer performance.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful to the data providers, Indonesian Regional Disaster Relief Agency (BPBD) Geoportal, and the anonymous reviewers for their useful comments and suggestions.

References

- Abidin, H.Z., et al., 2011. Land subsidence of Jakarta (Indonesia) and its relation with urban development. *Nat. Hazards* 59, 1753.
- Adnan, M.S.G., Abdullah, A.Y.M., Dewan, A., Hall, J.W., 2020. The effects of changing land use and flood hazard on poverty in coastal Bangladesh. *Land Use Pol.* 99.
- Agatonovic-Kustrin, S., Beresford, R., 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharmaceut. Biomed. Anal.* 22 (5), 717–727.
- Alfeilat, Abu, et al., 2019. Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*. <https://doi.org/10.1089/big.2018.0175>.
- Anderson-Bell, J., Schillaci, C., Lipani, A., 2021. Predicting non-residential building fire risk using geospatial information and convolutional neural networks. *Remote Sens. Appl. Soc. Environ.* 21, 100470. <https://doi.org/10.1016/j.rsase.2021.100470>.
- Aristizabal, F., Judge, J., Monsivais-Huetero, A., 2020. High-resolution inundation mapping for heterogeneous land covers with synthetic aperture radar and terrain data. *Rem. Sens.* 12. <https://doi.org/10.3390/RS12060900>, 900, 2020.
- Atkinson, P.M., Tatnall, A.R.L., 1997. Neural network in remote sensing. *Int. J. Rem. Sens.* 18, 699–709.
- Baker, J., 2012. Climate Change, Disaster Risk, and the Urban Poor: Cities Building Resilience for a Changing World. World Bank, Washington, DC.
- Bappenas, 2010. Indonesia Climate Change Sectoral Roadmap. BAPPENAS, Jakarta.
- Bellman, Richard, 1958. Dynamic programming and stochastic control processes. *Inform. Contr.* [https://doi.org/10.1016/S0019-9958\(58\)80003-0](https://doi.org/10.1016/S0019-9958(58)80003-0).
- Belousov, A., Verzhakov, S., Von Frese, J., 2002. Application aspects of support vector machines. *J. Chemom.* 16 (8–10), 482–489.
- Bishop, C.M., 1994. Neural networks and their application. *Rev. Sci. Instrum.* 65 (6), 1803–1830.
- BMKG, 2020. Kompas.com [Online] Available at: <https://nasional.kompas.com/read/2020/01/03/10141971/bmkg-sebut-curah-hujandi-jakarta-awal-2020-paling-ekstrem-dalam-sejarah> [Accessed 8 August 2021].
- BMKG, 2021. Prakiraan Musim Hujan 2021/2022 Indonesia. BMKG, Jakarta.
- BNPB, 2020. tirto.id [Online] Available at: <https://tirto.id/penyebab-banjir-jakarta-2020-dan-daftar-jumlah-korban-meninggalp9p> [Accessed 8 August 2021].
- Botzen, W., Aerts, J., Bergh, V.D.J., 2012. Individual preferences for reducing flood risk to near zero through elevation. *Mitig. Adapt. Strategies Glob. Change* 18, 229224.
- BPS Jakarta, 2019. Penduduk, Laju Pertumbuhan Penduduk, Distribusi Persentase Penduduk Kepadatan Penduduk. Rasio Jenis Kelamin Penduduk Menurut Provinsi/Kabupaten/Kota, Jakarta: s.n.
- Cao, H., Naito, T., Ninomiya, T., 2008. The 1st International Workshop on Machine Learning for Vision-Based Motion Analysis - MLVMA'08. HAL, Marseille.
- Chang, C.C., Lin, C.J., 2011. LIBSVM. *ACM Trans. Intell. Syst. Technol.* 2, 27. <https://doi.org/10.1145/1961189.1961199>.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357.
- Chen, W., Panahi, M., Tsangaratos, P., Shahabi, H., Ili, I., Panahi, S., Li, S., Jaafari, A., Ahmad, B.B., 2019a. Applying population-based evolutionary algorithms and a neuro-fuzzy system for modeling landslide susceptibility. *Catena* 172, 212–231.
- Chen, W., Hong, H., Li, S., Shahabi, H., Wang, Y., Wang, X., Ahmad, B.B., 2019b. Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. *J. Hydrol.* 575, 864–873.
- Cheng-Chien, L., Ming-Chang, S., Ming-Syun, K., Wang, K., 2018. Flood prevention and emergency response system powered by Google Earth Engine. *Rem. Sens.* 10 (8), 1283.
- Conde, Francisco, Munoz, Maria, 2019. Flood monitoring based on the study of sentinel-1 SAR images: the Ebro river case study. *Water* 11 (12), 2454. <https://doi.org/10.3390/w11122454>.
- Cunningham, P., Delany, S., 2007. K-nearest neighbour classifiers. *Mult. Classif. Syst.* 54.
- De Vries, B., et al., 2020. Rapid and robust monitoring of flood events using Sentinel-1 and Landsat data on the Google Earth Engine. *Remote Sens. Environ.* 240, 111664.
- Giovannettone, J., Copenhaver, T., Burns, M., Choquette, S., 2018. A statistical approach to mapping flood susceptibility in the lower Connecticut river valley region. *Water Resour. Res.* 54 (10), 7603–7618.
- Glenn, E., et al., 2012. Roles of saltcedar (*Tamarix* spp.) and capillary rise in salinizing a nonflooded terrace on a flow-regulated desert river. *J. Arid Environ.* 79, 56–65.
- Gorelick, Noel, et al., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* <https://doi.org/10.1016/j.rse.2017.06.031>.
- Gurney, K., 1997. An Introduction to Neural Networks. UCL Press, London.
- Herda Adeline, E., Sari Hasibuan, H., Sarwanto Moersidik, S., 2020. Infiltration Capacity in Flood Mitigating Jakarta. <https://doi.org/10.4108/EAL22-10-2019.2292391>.
- Hinkel, Jochen, et al., 2020. Coastal flood damage and adaptation costs under 21st century sea-level rise. *PNAS*. <https://doi.org/10.1073/pnas.1222469111>.
- Ioffe, S.S.C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Cornell University.
- Irawan, L.Y., et al., 2021. Combination of machine learning model (LR-FR) for flash flood susceptibility assessment in Dawuan Sub watershed Mojokerto Regency, East Java. *IOP Conf. Ser. Earth Environ. Sci.* 739.
- Islam, A.R.M.T., Talukdar, S., Mahato, S., Kundu, S., Eibek, K.U., Pham, Q.B., Kuriqi, A., Linh, N.T.T., 2021. Flood susceptibility modelling using advanced ensemble machine learning models. *Geosci. Front.* 12 (3), 101075.
- James, T., Schillaci, C., Lipani, A., 2021. Convolutional neural networks for water segmentation using sentinel-2 red, green, blue (RGB) composites and derived spectral indices. *Int. J. Rem. Sens.* 42, 5342–5369. <https://doi.org/10.1080/01431161.2021.1913298>.
- Jiawei, Han, Pei, Jian, Kamber, Micheline, 2012. Data Mining Concepts and Techniques, 3 ed. Elsevier, Waltham.
- Kia, M.B., et al., 2012. An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia. *Environ. Earth Sci.* 67, 251–264.
- Kirkby, M., Beven, K., 1979. A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci. J.* 24 (1), 43–69.
- Kocornik-Mina, Adriana, et al., 2016. Flooded Cities. *Am. Econ. J.: Appl. Econ.* 12 (2). <https://doi.org/10.1257/app.20170066>, April 2020.
- Kure, S., Jang, S., Ohara, N., Kavvas, M.L., Chen, Z.Q., 2013. WEHY-HCM for modeling interactive atmospheric-hydrologic processes at watershed scale. II: model application to ungauged and sparsely gauged watersheds. *J. Hydrol. Eng.* 18 (10), 1272–1281.
- Lee, M., Kang, J., Jeon, S., 2012. Application of frequency ratio model and validation for predictive flooded area susceptibility mapping using GIS. *IEEE Int. Geosci. Remote Sens. Symp.* 895–898.
- Marjanović, M., Kovacević, M., Bajat, B., Voženilek, V., 2011. Landslide susceptibility assessment using SVM machine learning algorithm. *Eng. Geol.* 123 (3), 225–234.
- Martinis, Sandro, Plank, S., Čwik, K., 2018. The use of sentinel-1 time-series data to improve flood monitoring in arid areas. *Rem. Sens.* 10 (4), 583, 10.
- Mindje, R., et al., 2019. Flood susceptibility modeling and hazard perception in Rwanda. *Int. J. Disaster Risk Reduc.* 38.
- Miyamoto, M., Sugiura, A., Okazumi, T., Tanaka, S., Nabesaka, S., Fukami, K., 2012. July. Suggestion for an advanced early warning system based on flood forecasting in Bengawan Solo River basin, Indonesia. In: Proceedings of 10th International Conference on Hydroinformatics. IWA IAHR. No. 394).
- Moore, I., Grayson, R., Ladson, A., 1991. Digital terrain modeling: a review of hydrological, geomorphological, and biological applications. *Hydrol. Process.* 5, 3–30.
- Motevalli, A., Vafakhah, M., 2016. Flood hazard mapping using synthesis hydraulic and geomorphic properties at watershed scale. *Stoch. Environ. Res. Risk Assess.* 1889–1900.
- Naga, I.E., Murphy, M.J., 2015. *What is Machine Learning?* s.L. Springer, Cham.
- Pourali, S., Arrowsmith, C., Chrisman, N., 2016. Topography wetness index application in FloodRisk-based land use planning. *Appl. Spatial Anal.* 9, 39–54.
- Pourghasemi, H.R., et al., 2021. Assessment of urban infrastructures exposed to flood using susceptibility map and Google Earth Engine. *IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens.* 14, 1923–1937.
- 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. *Comput. Geosci.* 350 (365), 51.
- Qiu, J., et al., 2021. Flood monitoring in rural areas of the Pearl river basin (China) using Sentinel1 SAR. *Rem. Sens.* 13 (7), 1384.
- Schillaci, C., Braun, A., Kropáček, J., 2015. 2.4.2. Terrain Analysis and Landform Recognition. Available at: <https://www.tribunnews.com/nasional/2021/08/26/pr-ediksi-bmkg-puncak-musimhujan-akan-terjadi-pada-januari-februari-2022> [Accessed 5 September 2021].
- Shahabi, H., 2020. Flood detection and susceptibility mapping using sentinel-1 remote sensing data and a machine learning approach: hybrid intelligence of bagging ensemble based on K-nearest neighbor classifier. *Rem. Sens.* 12 (2), 266.
- Tehrany, M., Pradhan, B., Jebur, M., 2015. Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stoch. Environ. Res. Risk Assess.* 29 (4), 1149–1165.

- Truckenbrodt, J., Freemantle, T., Williams, C., Jones, T., Small, D., Dubois, C., Thiel, C., Rossi, C., Syriou, A., Giuliani, G., 2019. Towards sentinel-1 SAR analysis-ready data: a best practices assessment on preparing backscatter data for the cube. *Data* 4. <https://doi.org/10.3390/data4030093>.
- Tsyganskaya, V., Martinis, S., Marzahn, P., 2019. Flood monitoring in vegetated areas using multitemporal sentinel-1 data: impact of time series features. *Water* 11 (9), 1983.
- van Leeuwen, Kees, Hofman, Jan, Driessen, Peter, Frijns, Jos, 2019. The challenges of water management and governance in cities. *Water*. <https://doi.org/10.3390/w11061180>.
- Vojtek, M.V.J., 2019. Flood susceptibility mapping on a National scale in Slovakia using the analytical hierarchy process. *Water* 2 (364), 364.
- Widyaningrum, G., 2019. Sejarah Banjir di Jakarta, Sudah Terjadi Sejak Zaman Tarumanegara. National Geographic Indonesia, Jakarta.
- World Bank, 2019. IMPLEMENTATION COMPLETION AND RESULTS REPORT. Report No: ICR00004907 Available from. <https://documents1.worldbank.org/curated/en/153081567169469254/pdf/Indonesia-Jakarta-Urgent-Flood-Mitigation-Project.pdf> [Accessed 21 November 2021].
- Yao, X., Tham, L., Dai, F., 2008. Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. *Geomorphology* 101 (4), 572–582.