Predicting the Spatial Distribution of Rain-Induced Shallow Landslides by applying GIS and Geocomputational Techniques: A Case Study from North East India

Parag Jyoti Dutta[1]*, Santanu Sarma[1], Jayanta Jivan Laskar[2]
[1]Department of Geology, Cotton University, Guwahati 781 001, India
[2]Department of Geological Sciences, Gauhati University, Guwahati 781 014, India
***Author for Correspondence** *Email: paragjdutta@gmail.com*

Abstract:This study presents a case of statistical modelling, by applying GIS and geocomputational techniques, to predict areas that are susceptible to future rain-induced shallow landslides. The statistical prediction model is based on the observed relationships between the spatial distribution of past landslideevents and environmental (causal) factors that are associated with such phenomena. The study also evaluates the predictive performance of a *nonlinear regression model*, namely the *Generalized Additive Model*(GAM),applied for the analysis. The study area comprises a residual hill of ≈ 6 Km^2 area situated in the heart of Guwahati (capital city of Assam in NE India). We exploited the geoprocessing functions of SAGA GIS to derive nine different terrain attributesfrom a *digital elevation model* (DEM) processed by *synthetic aperture radar interferometry* (InSAR). The terrain attributes along with land use classes, in raster grid format, constitute the predictor variables. An inventory of the locations of *eighty-two* past occurrences of shallow landslide events constitutes the response. We performed the modelling and statistical geocomputation entirely in the open-source *R* language and software environment. The procedure comprises the following three steps: (1) *Collinearityanalysis* to discard redundant predictors. (2) 100-fold *bootstrap resampling* to fit the GAM by a random selection of 2/3 of the landslide pixels ("training" subset) and validate the GAM by the remaining 1/3 ("test" subset). (3) Estimate model accuracy (*true error rates*) by a repeated 100-fold '*hold-out validation*' method and evaluate the predictive performance of the model by the *Area under the ROC curve* (AUROC) computed for 100 independently trained models. The *mean* and *standard deviation* of accuracy on training sets are 0.80 and 0.01, and that on test sets are 0.79 and 0.02 respectively. The AUROC corresponding to the *mean*of landslide probabilities is 0.87, and that of the 95% *Confidence Intervals* (CI) is between 0.86 and 0.88. Thevalues of these quality measures indicate that a data-driven model, such as the GAM, is efficient regarding its predictive performance, to highlight the unstable areas in the study area. We subsequently used the *mean* values of the landslide probability (*susceptibility*) estimates corresponding to each mapping unit (grid cell) to construct the landslide susceptibility map, which can be used for land use planning and hazard mitigation.

Keywords:*Generalized Additive Models*, *Landslide Susceptibility*, *True Error Rates*, *ROC curve*, *Guwahati*

*****

I. INTRODUCTION

Rain-induced shallow landslides pose a common natural hazard in many parts of the world. This kind of phenomena is also prevalent in Guwahati, the capital city of the state of Assam in India’s North East Region. The landscape features some residual hills (nineteen) scattered in and around the city. Every year during the monsoon season, the media is replete with tales of landslides, of huge chunks of earth hurtling down, of people getting buried and frantic rescue operations to save lives and property. However, illegal occupation of land in the hills continues still unabated. Shallow landslides are slope failures of soil cover, a few meters thick, called regolith which rests above bedrock. In Guwahati, these are, more often, triggered by intense than prolonged rainfall. Although they thrust relatively small volumes of soil, they produce a high impact resulting in significant damages to infrastructure and sometimes, fatalities.

Managing hazards associated with shallow landslides, at first, requires an understanding of where such landslides may occur.The spatial prediction of this occurrence event (i.e. “where” it is likely to happen) has been an interesting subject of research throughout the last three decades and is termed as *landslide susceptibility* (LS) assessment. However, the frequency or temporal probability of the future landslides (i.e. “when” it is likely to occur) is not assessed within the susceptibility component of landslide risk [1]. LS studies result in maps which classify the terrain into zones of relative probability estimates of failure. These maps are required in establishing standards for land-use planning and hazard mitigation. Previous studies on LS conducted in Guwahati [2] – [6] were able to produce maps only on a regional scale of 1:50,000. In practice, only limited detail can be shown on such small scale maps. Producing large-scale susceptibility maps requires several types of input data viz. topography, regolith thickness, mechanical and hydraulic parameters of soil, pore-water pressures and other time-variant environmental information of sufficient spatial resolution which are difficult to acquire on limited budgets. Hence, we considered that a landslide susceptibility map at a local scale of 1:10,000 would be an effective compromise between the cost involvement and requirement for land use planning and disaster preparedness in hillslopes within Guwahati. We should note that the production of most regulatory landslide hazard and risk maps in Europe is on a scale of 1:10,000 [7].

Currently, there is a huge array of quantitative LS modelling techniques – empirical (statistical), deterministic (process-based), and now machine learning, available for the spatial prediction of rain-induced shallow landslides [8 and references therein]. Statistical techniques are data-driven and are applied particularly to areas where sufficient geotechnical data are not available for deterministic assessment of LS. These are quantitative and objective methods based on a statistical estimation of the possible future locations of landslides according to the underlying principle that landslides are more likely to occur under similar terrain and environmental conditions of the past [9]. The spatial distribution is predicted by the identification of the relationships between independent predisposing factors ("predictors"), such as terrain attributes and hydrological indices, and a dependent variable ("response"), represented by an inventory of past landslide locations. The susceptibility model output is a prediction surface or map that spatially represents the distribution of predicted values, usually as probabilities distributed across grid cells.

More than a dozen classes of statistical models exist in the literature, but there is no best model or technique [8], [10], [11], [12]. Various criteria such as model accuracy, error rates estimated from cross-validation, robustness to sampling variation, and adequacy to describe landslide processes, exist for model selection in the context of landslide susceptibility [13], [14]. Recently, a research study was conducted in the province of Lower Austria to test a set of six statistical and machine learning modelling techniques [8]. Research outputs indicated that the most interpretable (i.e. interpretable results that can shed light on landslide conditioning factors) and visually appealing (i.e. had a smooth prediction surface) technique is the *Generalized Additive Model* (*GAM*). The primary objective of this study is to evaluate the performance of the GAM in predicting the locations of shallow landslides on residual hillslopes, in a tropical climate setting, using terrain attributes and land use as predictors. The implementation of a statistical model for susceptibility mapping requires a map of past landslide locations (inventory) as the necessary input. However, workforce constraints permitted us to create a landslide inventory map of only one of the hills, called the *Narakasur Hill*, which is considered as the study area for this work (Fig. 1). We assume that the results and procedure could be directly transferable to the other hills within the city.

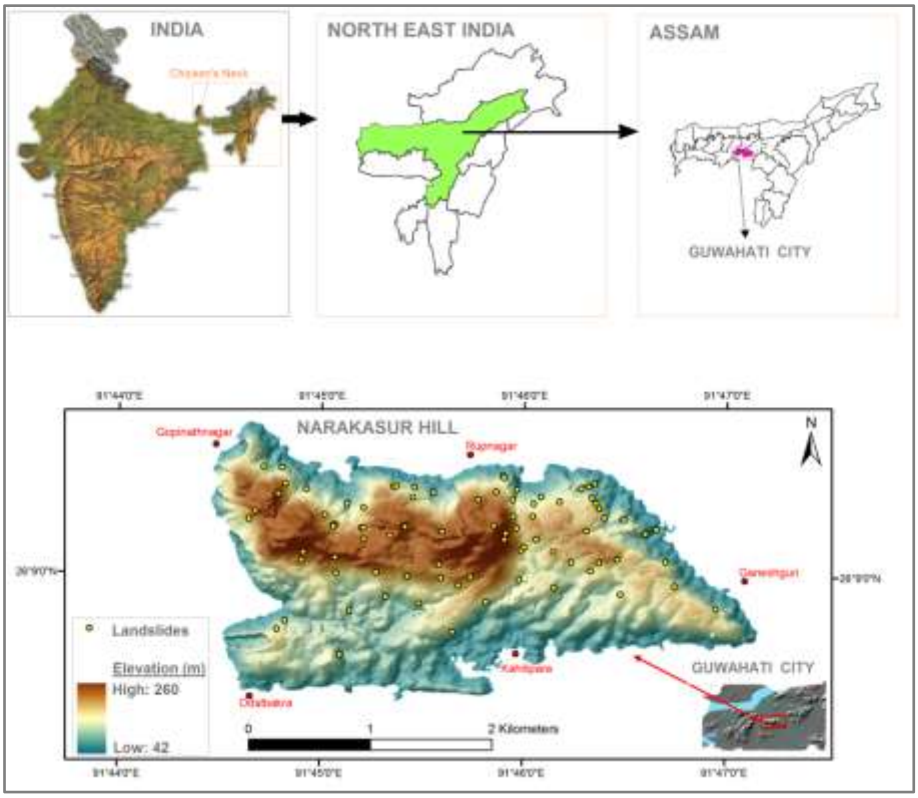


Figure 1: Location map of the Narakasur Hill with the spatial distribution of shallow landslides

II. STUDY AREA

A. Geology

Guwahati city is situated between 26º 05′ – 26º 10′ N and 91º 30′ – 91º 50′ E, with an aerial extent of ≈ 217 Km^2 under Guwahati Municipal Corporation (GMC), spread across both banks of the river, Brahmaputra. There are nineteen residual hills within the GMC limits, which are moderately high with a general E-W trend. These hills are the erosional remnants of the Precambrian *Basement Gneissic Complex* (BGC) of the Meghalaya Plateau, which is the NE extension of the Indian Peninsular Shield. The dominant bedrock within these hills is *Quartzo Feldspathic Gneiss* (QFG)[15]. Several Neoproterozoic granitoids, featuring porphyritic texture, have intruded the BGC at various locations. A few hills within the Guwahati landscape also bear patches of these granitoids exposed as tors, boulders and corestones on hilltops and slopes [16]. Among these, the *Narakasur Hill* (Fig. 1) is one which bears the intrusive signature of grey *porphyritic granite* (PG) within the QFG basement rocks. The areal extent of the hill is ≈ 6 Km^2 and is easily accessible as it lies in the heart of Guwahati city.

B. Climate

The Narakasur Hill is characterised by the highest altitude of 260 m above m.s.l. and slopes ranging 10° – 20°, although slopes with a steep gradient are widespread. Hillslopes are characterised equally by convex and concave curvatures. The climate is humid subtropical with 1,717.7 mm average yearly rainfall and 128 average rainy days per year (Source: *World Meteorological Organization*). The SW monsoons begin in Assam from the middle of June. Most of the monsoon rains fall between June and September with peaks in June and July. Thunderstorms known as *Bordoicila* are common in the season. During monsoon storms, cumulated daily rainfall can exceed 100 mm, with extreme events of +200 mm in 24 hours.

C. Land use

We mapped the distribution of the different land use classes in the Narakasur Hill by GPS survey aided by high resolution satellite images (*QuickBird*). These images have a spatial resolution of ≈ 2 m. The land use classes over the study area are: (1) *Barren Land* (2) *Forest Land* (3) *Forest Land with sparse settlement* (4) *Open Field* (5) *Shrubs* (6) *Urban Settlement* (7) *Vegetated Area with settlement* and (8) *Water Body (Pond)*. The land use distribution (Fig. 2) shows that urban and built-up areas (forest and vegetated areas with settlements) cover more than 74.1% of the study area. *Built-up areas* increase to the detriment of *forests*, *shrubs*, and *open fields*.

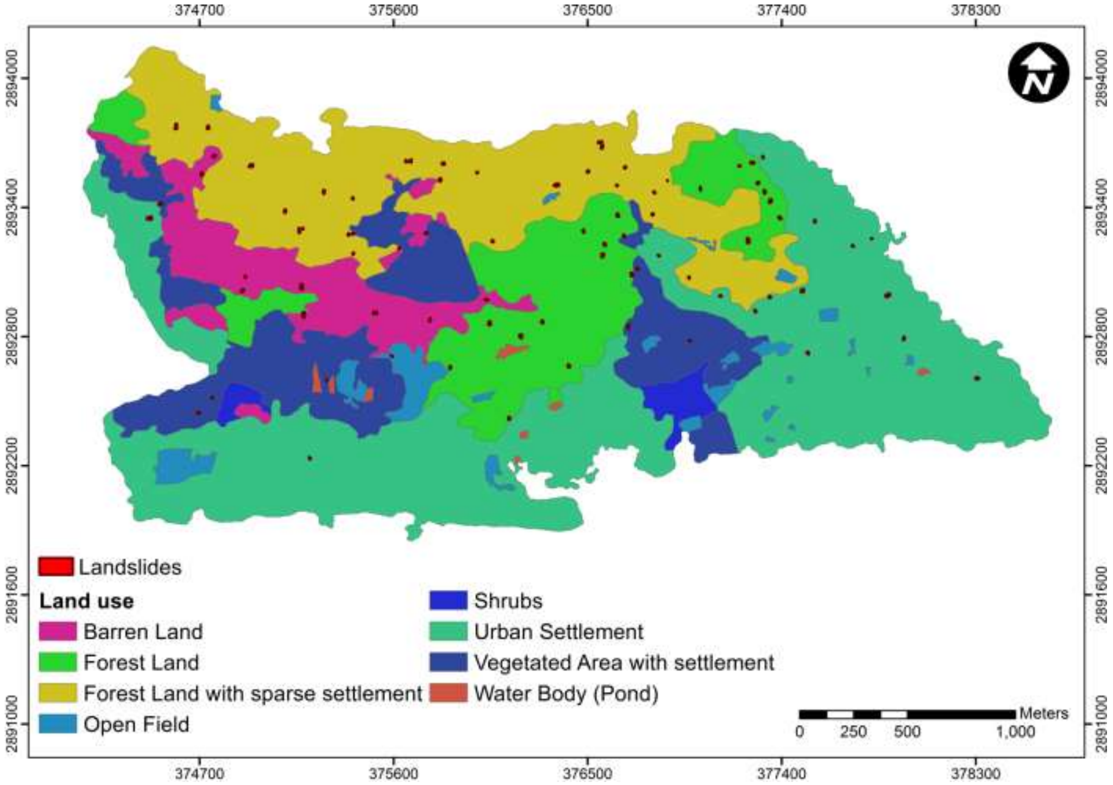


Figure 2: Distribution of Land use classes within the study area

D. Landslide Features

We were able to identify a total of 82 scars of past occurrences of shallow landslides in the Narakasur Hill during field surveys carried out in the years 2011-2014 (Fig. 1). As per the classification of Cruden and Varnes [17], we observed three types of failures: (1) *incipient translational slides*, where the displaced mass has limited movement (Fig.3a); (2) *translational soil slides*, where the mass has moved exposing the failure surface (Fig.3b & Fig.3c); and (3) *roto-translational slides*, adjacent to road-cuts (Fig. 3d). Among the three, *type (2)* failures are predominant. A shallow depth (1.0 – 1.5 m) below the ground surface of the sliding surface is a characteristic of these failures. Landslide bodies have lengths typically lower than 10 m, and volumes lower than 100 m^3. Most of these have occurred on hillslopes with a gradient of 30° – 45°.

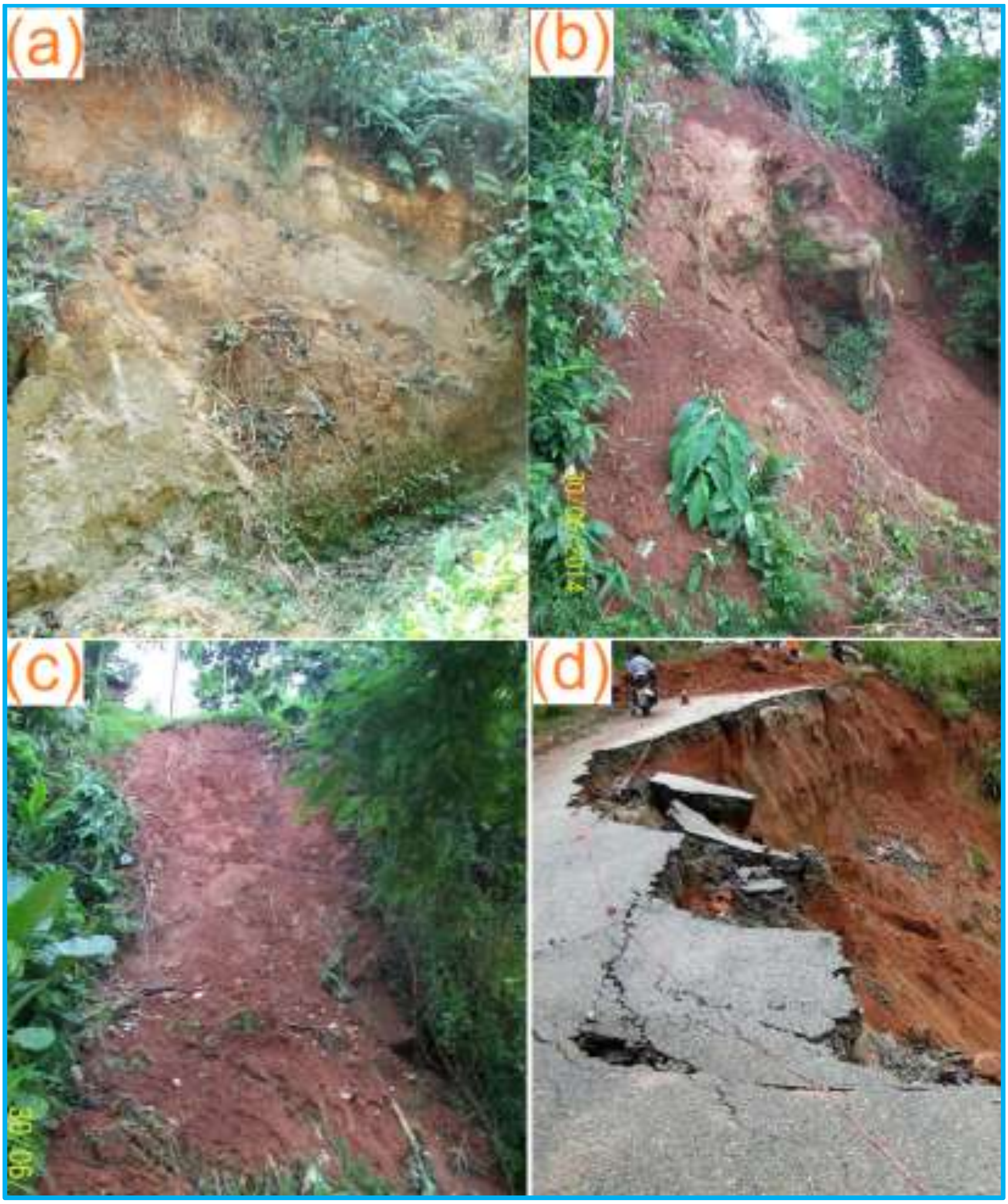


Figure 3: Photos of landslides within the study area: **(a)** Incipient Translational Slide; **(b)** Translational Soil Slide; **(c)** Translational Soil Slide; **(d)** Roto-translational Slide, which caused the destruction of a road

III. MATERIALS AND METHODS

A. Selecting the correct mapping unit

Evaluation of LS requires the preliminary selection of a suitable mapping unit. The term refers to a portion of the land surface containing a set of ground conditions which differ from the adjacent units across definable boundaries. At the scale of the analysis, a mapping unit represents a domain that maximizes internal homogeneity and heterogeneity between units. Various methods have been proposed to partition the landscape for distinct, clearly definable mapping units. The units fall into one of the following five types: (i) grid cells (ii) terrain units (iii) unique condition units (iv) slope units and (v) topographic units. Among the different types, the most frequently and widely used in statistical LS modelling are grid cells (pixels) and slope units. A grid cell mapping unit should correspond to a cell size that matches the contributing area of the most typical landslides in the study area. The final percentage of landslide cells in the study area should accurately reflect the actual portion of the study area covered by landslides. We have chosen the cell size from bothreference to the cartographic scale and the size of the smallest landslides in the study area. Hengl [18] suggests that the cell size should be the equivalent of 0.0005 × the scale number. Accordingly, mapping at 1:10,000 scale leads us to work with a 5 m × 5 m cell size.

B. Data Preparation

Susceptibility Assessment and mapping of rain-induced shallow landslides are commonly carried out in a GIS. In this type of analysis, the digital elevation model (DEM) representing the topography of the study area forms the key raster dataset as input. Terrain attributes, derived from a DEM, function as surrogates for surface processes and geophysical site conditions to simplify complex, meaningful geomorphological relationships [11], [19], [20]. We created a DEM (Fig. 1) with a cell size of 5 m × 5 m by *interferometric processing* of *Synthetic Aperture Radar* (SAR) data from the German *TanDEM-X*satellite mission. Thereafter, we derived nine different morphometric terrain attributes viz. *slope angle*, *slope aspect*, *plan curvature*, *profile curvature*, *catchment area*, *catchment slope*, *topographic wetness index*[21], *topographic position index*[22], [23], and *terrain ruggedness index*[24], from the DEM for using them as predictors in this study. We exploited the geoprocessing functions of SAGA GIS [25], for extracting the grids of these terrain attributes. These parameters, in the context of hillslope geomorphological processes, are defined [26], [20].

C. Predictors

Slope angle, *slope aspect*, *plan curvature*, and *profile curvature*are calculated based on local polynomial approximations, according to [27]. The *slope angle* is one of the most important factors, as it strongly controls the shear forces acting on hillslopes and the water distribution [28]. *Slope aspect* can play a fundamental role in landslide susceptibility, as it may have an influence on the response of the terrain to different amounts of rainfall and solar radiation, as expressed by local temperature and evaporation, and eventually through soil moisture content and vegetation growth [29]. The *slope aspect* was transformed into a categorical variable, to avoid the misclassification of flat areas as "No Data" areas. The *plan curvature* and *profile curvature* represent the topographic influence of local morphology on slope hydrology and soil erosion and deposition. In particular, *plan curvature* is the curvature of the surface perpendicular to the direction of the maximum slope which controls the convergence and divergence of topography and the sub-surface water flow. *Profile curvature* is the curvature of the surface parallel to the direction of the maximum slope, which characterizes the near-surface acceleration or deceleration of flow down a slope, influencing the potential erosion or deposition rate and, consequently, the soil depth [11]. The *catchment area* and *catchment slope* are derived using the *multiple-flow-directionalgorithm*[30]. We transformed the *catchment area* to its natural logarithm for reducing skewness [31] and used it as a proxy for soil moisture and soil depth. The *catchment slope* is another important factor that influences the intensity of the destabilizing forces upslope [31]. The *Topographic Wetness Index* (TWI) highlights the tendency of water to accumulate at points in the drainage basin and the trend of the water to move along a slope by the action of the gravitational forces. The index can be correlated to the soil moisture content and the groundwater conditions [32]. The *Topographic Position Index* (TPI) provides a simple proxy to study the effects of the location of objects, such as landslides, on a landscape. In the case of landslides, it relates slope elevation to their location. It is calculated by comparing the elevation of a cell to the mean elevation of the surrounding cells in a circular buffer of around 1000 m radii [22], [23]. The *Terrain Ruggedness Index* (TRI), calculated as the sum of the change in altitude between a cell of the grid and its eight neighbouring cells [24], is used to quantify the landscape heterogeneities, which could have effects on the locations of shallow landslide triggering areas. We inserted *land use* (Fig. 2) as a predictor variable in the GAM to consider the role of vegetation and anthropogenic activities on shallow landslides susceptibility [29]. We observed the occurrences of shallow landslides to be uniformly distributed in soils derived from both bedrock materials (PG and QFG). Hence, it was difficult to identify any significant relation between lithological features and the occurrence of shallow landslides in the study area. These reasons dictated our decision to exclude bedrock geology from the list of predictor variables.

D. Response

We conducted field surveys during the period 2011-2014 for mapping shallow landslide occurrences in the Narakasur Hill. We mapped a total of eighty-two landslide source areas, at a scale of 1:10,000, by walking a handheld GPS receiver along each landslide perimeter. For this operation, the GPS captured geographical coordinates every meter. No correction was applied to the GPS signal, and the expected planimetric error for the individual GPS measurements was ±5 m or less. The GPS data was later imported into a Geographical Information Software (GIS) and converted into shapefiles, as polygon features. The locations of the eighty-two landslides which serve as the response variable are shown in Fig.1. However, there are no records available on the triggering dates of these failures. We rasterized each landslide source area according to the size of a mapping unit (5m × 5 m). We followed this approach because of the limited coverage area of the landslides (generally less than 500 m^2) and due to the lack of field evidence for a distinct boundary between the landslide scarp and landslide accumulation zone. This is probably because of the superficial weathering of the landslide bodies.

E. Statistical Modelling

A statistical model for the spatial prediction of landslides is built on the assumption that the factors which caused slope-failure in a region are the same as those which will generate landslides in the future. The model is based on the concept of a "statistical sample", which is a subset of the population (which is usually unobservable), whose properties are "close" to those of the whole population. This means that the events observed during a fixed period do not represent the entire population of landslides but rather a sample of this population. Moreover, the explanatory variables related to geomorphometric and geo-environmental features influencing the spatial distribution of landslides (referred to as "*susceptibility*") are often *latent* (unobservable or not exactly measurable). Finally, the distribution of landslides, while related to the susceptibility, is partially dependent on random factors. Thus, the probability distribution of landslides, in terms of the general linear model, assumes the form[33]:

$$P[y=1] = \beta(z) + \varepsilon_k + \varepsilon$$
$$= \sum_{k=1}^{n} \lambda_k \ (x_k)$$
$$+ \varepsilon$$

where $P[y=1]$ is the probability of occurrence of an event (landslide); $\beta(z)$ is the parameter relating z and the observed probability;$z$is the latent variable (susceptibility); λ_k is the vector of weights; ε is the random error of the general linear model; and ε_k is the random error between the latent variable and the observable predictors. Neither z nor β are observable or exactly measurable, so we rely on k observable variables x_k (geology, topography, land use and others) to approximate z.

What is important in the above model is that we can have n possible different future outcomes of landslide distribution. This may seem an odd assumption, but we can consider that in a study area we may have some sectors in which the value of z is equivalent; thus, the fact that a landslide occurs in one equally susceptible zone and not in all of them is largely due to the stochastic term ε. Of course, ε may be physically determined, but we have to still consider it as stochastic because we cannot model its behaviour in a deterministic way. The above formulation introduces the concept of *uncertainty* in statistical LS models, which means that we can associate *prediction errors* and the *confidence interval* (CI) with our model outcomes.

Geomorphic processes involved in causing landslides can be considered to have nonlinearities. Phillips [34]opines that nonlinearity can occur in geomorphic systems that progress towards a critical point where, upon reaching, a system changes behaviour. In the case of hillslopes, this can be observed when it becomes unstable after a threshold point as a consequence of changing hydrological conditions. In such cases, application of a linear model for prediction would seem inappropriate and calls for a modification. The *Generalized Additive Model* (GAM) [35] is designed to replace the linear function of each covariate in the *Generalized Linear Model* (GLM) [36] with an empirically fitted smooth function to find the appropriate functional form for the data [35]. The GAM is a semi-parametric extension of the GLM or *Logistic Regression* (in case of a binary response). The GAM has been applied in the modelling of ecology [37], public health [38] and also in landslide susceptibility (LS) [11], [31], [39], [40], [41], [42], very recently.The GAM uses a link function to establish the relationship between the mean of the response variable and the sum of a set of smooth functions of predictor variables, as shown in the Eq. [2] below [41]:

$$g(\square) = \sum_{i=1}^{n} f_i(x_i) \qquad [2]$$

where μ = expected value (mean) of the response variable; *g*(μ) is the link function, and the $f_i\,(x_i)$ are smooth functions (typically, splines). Thus, the GAM allows a combination of linear and nonlinear smoothing functions in an additive manner to define the relationship between predictors and response. In the case of the logistic additive model for binary response variables, the response is modelled using the *logit* of the occurrence probability, *p*(*x*), conditional on predictor variables **x** = $(x_1,\ldots, x_m)^{T}$[35]:

$$\text{logit}(x) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 f_1(x_1) + .. + \beta_m f_m(x_m) \qquad [3]$$

where $\left(\frac{p(x)}{1-p(x)}\right)$ is referred to as the odds. *Odds* are defined as the ratio of the probability of success to the probability of failure and range between zero and infinity.

The functions, $f_i$in Eq. [2] are *non-parametric smoothers*. A smoother is a tool for summarising the *trend* of a response variable, *y*, as a function of a set of predictor variables, $x_{1,}x_2,\ x_2\ \ldots.x_p$. It gives an estimate of the trend that is more predictable than *y* itself. We call the estimate, a *smooth*. Smoother is very useful; first, it helps to pick out the trend from the plot and secondly, it estimates the dependence of the mean of *y* on the predictors. It is non-parametric in nature and is flexible in the case of the dependence of *y* on the predictors. It allows an '*approximation*' with a sum of functions.

The GAM is built using a *stepwise variable selection* using the *Akaike Information Criterion* (AIC) [43]. Each variable, starting from the null model, can be entered as *linear* (untransformed), *nonlinear* (transformed by smoothing splines of two equivalent degrees of freedom), or *not included*, using the AIC to choose the best model [40]. For instance, it may linearly integrate some predictors, while integrating the other ones by complex, smooth functions representing complex relations between the variables, and providing high flexibility [41]. The AIC is a measure of *goodness-of-fit* that penalises for model complexity to obtain a "*parsimonious model*" [44]. Smaller models help to keep the estimated coefficient standard errors low and prevent the model from *overfitting*, which occurs when the number of predictor variables in the model is larger than the number of samples of the response variable [45]. *Overfitting* refers to a model that performs well on the training dataset but poorly on the test dataset [45].

We applied the GAM through the open source ***R*** software (version 3.3.0), a free software environment for statistical computing[46], with the contributed packages '*gam*' (version 1.14)[47], '*boot*' (version 1.3-18)[48], [49], '*perturb*' (version 2.05) [50], and '*ROCR*' (version 1.0-7) [51]. The procedure comprises the following three steps: (i) *Collinearity analysis* to discard redundant predictors (ii) Selection of the most significant predictorsby a *100-fold bootstrap resampling procedure* to fit the GAM ("training") by a random selection of 2/3 of the landslide dataset and "testing" (*validation*) with the remaining 1/3 (iii) *Estimate model accuracy* (*true error rate*) by a '*repeated hold-out*' method and *evaluatepredictive performance* by ROC analysis of 100 independently trained models. The three-step procedure is followed by subsequentextraction of the shallow landslide susceptibility map. A flow chart of the incorporated methodology is presented in Fig. 4.

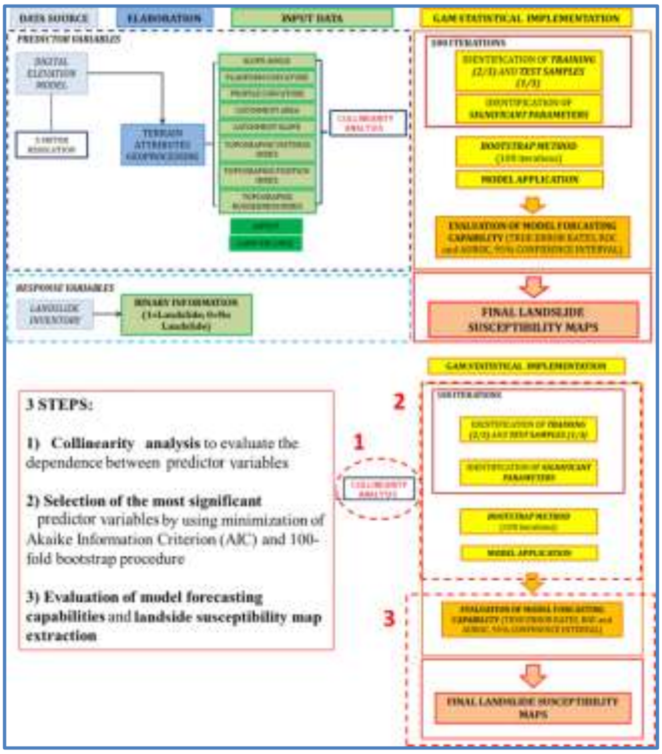

DATA SOURCE
ELABORATION
INPUT DATA
GAM STATISTICAL IMPLEMENTATION
PREDICTOR VARIABLES
DIGITAL ELEVATION MODEL
5 METER RESOLUTION
TERRAIN ATTRIBUTES GEOPROCESSING
SLOPE ANGLE
PLANFORM CURVATURE
PROFILE CURVATURE
CATCHMENT AREA
CATCHMENT SLOPE
TOPOGRAPHIC WETNESS INDEX
TOPOGRAPHIC POSITION INDEX
TOPOGRAPHIC RUGGEDNESS INDEX
ASPECT
LAND COVER
COLLINEARITY ANALYSIS
100 ITERATIONS
IDENTIFICATION OF TRAINING (2/3) AND TEST SAMPLES (1/3)
IDENTIFICATION OF SIGNIFICANT PARAMETERS
BOOTSTRAP METHOD (100 iterations)
MODEL APPLICATION
EVALUATION OF MODEL FORECASTING CAPABILITY (TRUE ERROR RATES, ROC and AUROC, 95% CONFIDENCE INTERVAL)
RESPONSE VARIABLES
LANDSLIDE INVENTORY
BINARY INFORMATION (1=Landslide; 0=No Landslide)
FINAL LANDSLIDE SUSCEPTIBILITY MAPS
GAM STATISTICAL IMPLEMENTATION
3 STEPS:
1) Collinearity analysis to evaluate the dependence between predictor variables
2) Selection of the most significant predictor variables by using minimization of Akaike Information Criterion (AIC) and 100-fold bootstrap procedure
3) Evaluation of model forecasting capabilities and landside susceptibility map extraction
1
2
3
COLLINEARITY ANALYSIS
100 ITERATIONS
IDENTIFICATION OF TRAINING (2/3) AND TEST SAMPLES (1/3)
IDENTIFICATION OF SIGNIFICANT PARAMETERS
BOOTSTRAP METHOD (100 iterations)
MODEL APPLICATION
EVALUATION OF MODEL FORECASTING CAPABILITY (TRUE ERROR RATES, ROC and AUROC, 95% CONFIDENCE INTERVAL)
FINAL LANDSLIDE SUSCEPTIBILITY MAPS

Figure 4: Flow chart representing the methodology and measures applied for quality assessment of the statistical model

IV. RESULTS AND DISCUSSION

A. Variable Importance and Nonlinearity

When predictor variables in a regression model are correlated with each other, this condition is referred to as *multicollinearity* (or *collinearity*). The presence of this problem hinders precise statistical explanation of the relationships between predictors and responses[52]. For each collinear relationship, there is one redundant variable. The first step in the GAM implementation is a *collinearity analysis*. The analysis is performed with the function *Colldiag*, available in the R package '*perturb*' [50]. It is an implementation of the regression collinearity diagnostic procedures[52]. These procedures compute the condition indexes of the matrix of independent variables. If the largest *condition index* (the condition number) is 30 or higher, then there may be collinearity problems. In the model application, preliminary collinearity analysis excluded the *catchment area* from the list of predictor variables.

In the second step, to avoid overestimation of non-landslide pixels, a 1:1 ratio is maintained between the landslide and non-landslide pixels. The examined data set is subdivided into two subsets: the *training* set and the *test* set. The training set, representing 2/3 of the dataset, is used to fit the model, while the test set, representing 1/3 of the dataset, is used to validate the fitted model. The random selection process for the training and test data sets is repeated in a 100-fold *bootstrap* procedure, aimed to identify the most frequent predictor variables. *Bootstrap* is a modern computationally intensive resampling-based technique for bias-reduced statistical error estimation [53]. The bootstrap draws independent samples (*with replacement*) from the available data to simulate the underlying data-generating distribution. The data themselves thus approximate this distribution without making any parametric distributional assumption. The most influential predictor variables, after the training phase, are *slope angle*, *slope aspect*, *profile curvature*, and *land use*, with variable-selection frequency ≥ 80 of the 100 bootstrap replications (Table 1a). According to the methodology, we used these variables as predictors in the GAM. A statistical justification for choosing the GAM is the *nonlinear relationship* between the *slope angle* and the occurrence probability of the shallow landslides (Table 1b).

TABLE 1(a)

Absolute frequencies of predictor variables (linear or nonlinear) selected by the 100-fold bootstrap replications. *NA*: "*not available*" indicates parameters discarded by the collinearity analysis. The predictors with bootstrap variable-selection frequency **≥ 80** are highlighted in **bold red**. SL: Slope Angle; TRI: Terrain Ruggedness Index; TPI: Topographic Position Index; TWI: Topographic Wetness Index; PLA: Plan Curvature; PRO: Profile Curvature; CA: Catchment Area; CS: Catchment Slope; ASP: Slope Aspect; LU: Land Use.

SL	TRI	TPI	TWI	PLA	PRO	CA	CS	ASP	LU
100	10	27	25	32	**82**	*NA*	18	**89**	**83**

TABLE 1(b)

Bootstrap replication frequencies of continuous predictors (as *linear* or *nonlinear*)

Continuous Predictor Variables	Linear	Nonlinear
SL	0	**100**
PRO	**58**	42

B. *Prediction Accuracy*

In the third step, the prediction accuracy is evaluated through a repeated 100-fold '*hold-out validation*' for regression with a binary response [54]. The *repeated hold-out* method is a *k*-fold repetition which consists of a random sampling of different *training* and *test* sets, in the proportion of 2/3 for *training* and 1/3 for *testing*. The method provides an estimate of the *true error rate* (accuracy) of each iteration. The *true error rates* of the 100 different iterations are calculated for all training and test sets (Table 2a). The results are averaged to yield an overall accuracy and compared.

TABLE 2 (a)

True error rates of prediction accuracy on *training* and *test sets* of landslide pixels

Mean Accuracy on training sets	Mean Accuracyon test sets	Standard deviation of accuracy on training sets	Standard deviation of accuracy ontest sets
0.8	0.79	0.01	0.02

C. *Model Uncertainty*

An important model uncertainty which can be visualized in a susceptibility map is the prediction uncertainty arising from using a statistical model. The concept of uncertainty in landslide susceptibility models means that we can associate prediction errors and the *confidence interval* (CI) with our model outcomes. The output of a statistical model for spatial modelling is usually comprised of a single probability value for each mapping unit (grid cells in our case) of the prediction surface. These individual probability values represent an estimated conditional mean value of the predicted probability [45]. Therefore, there is a prediction uncertainty as determined by the standard error of the predicted probability estimates for each unit of the susceptibility map [55]. The analysis of the standard error of the predicted probabilities and the prediction uncertainty analysis are independent of any class thresholds. The estimates for the model error in each mapping unit (grid cell) were obtained adopting a 100-fold “bootstrap” re-sampling procedure. The mean, standard deviation and other descriptive statistics of the probability (susceptibility) estimates were obtained for each grid cell from the ensembles of bootstrap model runs. To investigate the reliability of landslide probability associated with each grid cell, the 95% *Confidence Intervals* (CI) of each of the 100 landslide probability estimates are also computed. The amplitudes of 95% CI are shown in Fig. 5 signifying that the prediction uncertainty is low. The amplitude of a CI is defined as the difference between its upper and bottom limits. The low prediction uncertainty of landslide occurrence probability signifies that a susceptibility map built with the mean values of the probability estimates would be representative of the GAM results.

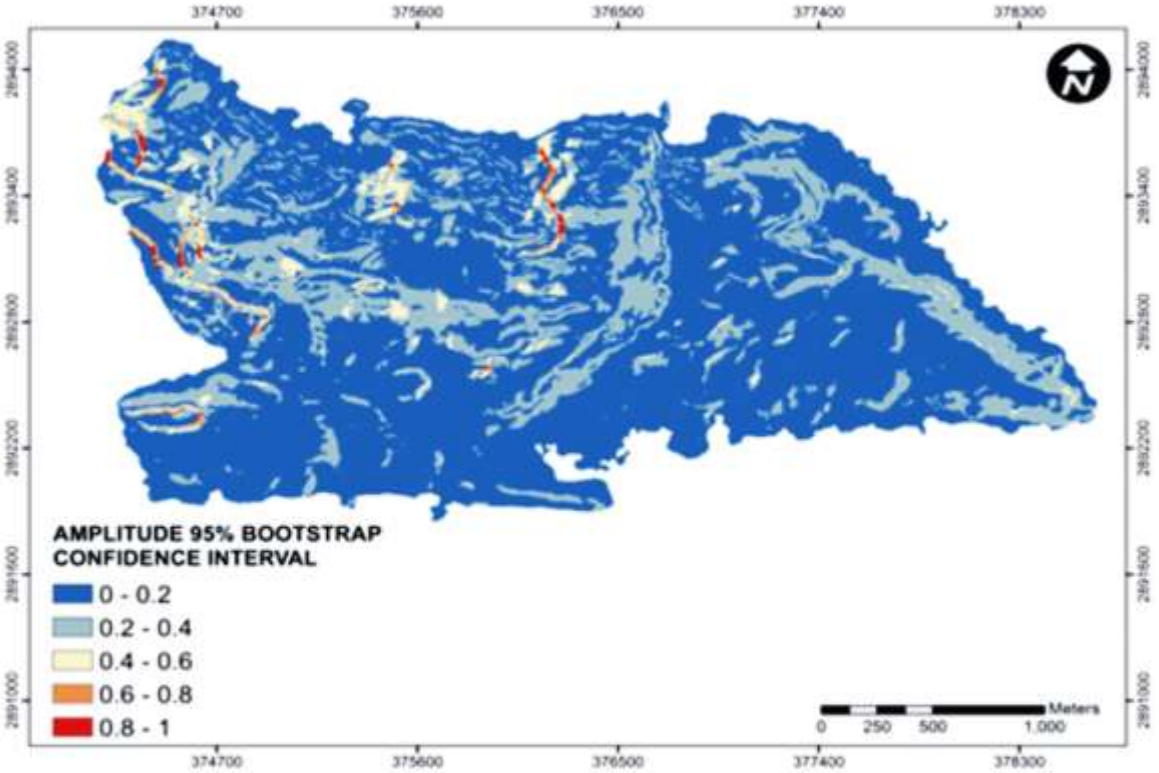


Figure 5: The amplitudes of 95% bootstrap confidence intervals (CI) of probability estimates assigned to each grid cell reflecting the reliability of the estimates

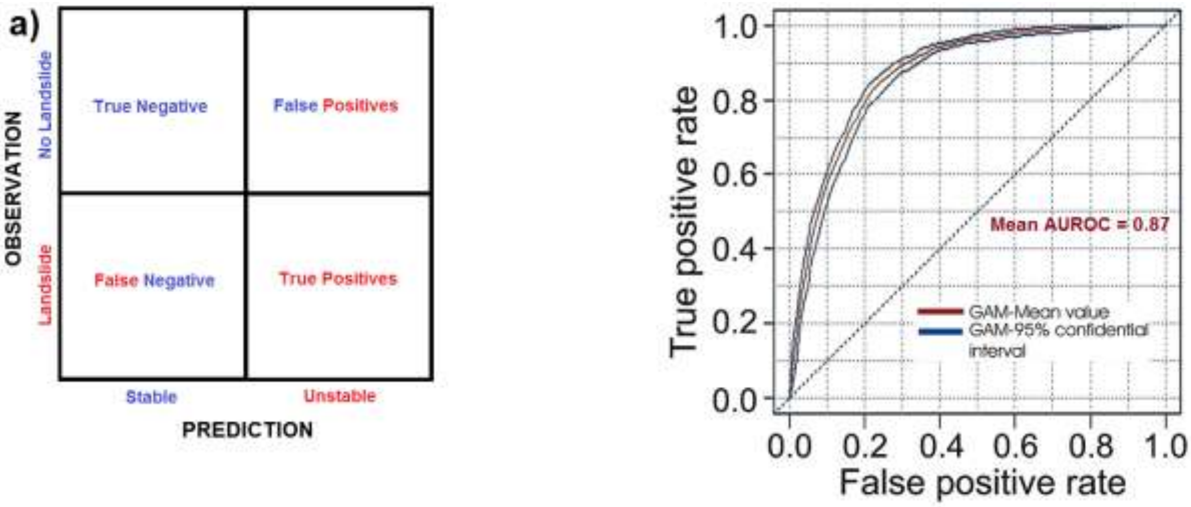
D. Prediction Performance

The most popular metric for evaluating the performance of landslide susceptibility models is to plot the receiver operator characteristic (ROC) curve [13]. The ROC chart illustrates the performance of classification models where the response variable is binary; in this case, the states are either "unstable" or "stable". There are four possible outcomes from a binary classifier (Fig. 6a) that can be formulated in a 2 × 2 contingency table (confusion matrix) [56]. If a grid cell is modelled as unstable and corresponds to a mapped landslide cell, it is considered a "True Positive" (TP); but if it corresponds to a non-landslide cell, it is regarded as a "False Positive" (FP). If the grid cell is modelled as stable and falls outside a mapped landslide cell, it is considered a "True Negative" (TN); but if it corresponds to a mapped landslide cell, it is regarded as a "False Negative" (FN). A ROC curve is constructed by plotting the "True Positive Rate" (TPR) on the Y-axis and the "False Positive Rate" (FPR) on the X-axis as derived from different contingency tables created by applying different cutoffs (thresholds). The TPR is the fraction of TP out of TP + FN that represents the total number of landslide cells. The FPR is the fraction of FP out of FP + TN that represents the total number of non-landslide cells. The Area under the ROC Curve (AUROC) can be used as a metric to assess the overall quality of a model [57]; the larger the area, the better the performance of the model over the whole range of possible cutoffs. AUC can take values from 0.5 (no discrimination) to 1.0 (perfect discrimination).

We evaluated the predictive performance of the GAM by the computation of the area under the ROC curve (AUROC). We used 100 independently drawn bootstrap replications of the landslide test set to estimate the AUROC for 100 independently trained models. We calculated the mean of

ISSN: 2321-8169
1306 – 1319

the 100 AUROC samples obtained from the 100-fold bootstrap procedure. The plots of TPR and FPR values of the various probability thresholds of the GAM predictions shaped the receiver operating characteristic (ROC) curve (Fig. 6b). We also calculated the bootstrap 95% confidence bands of ROC and its AUROC values. The ROC curves demonstrate a good prediction performance with the AUROC of Mean equal to 0.87 and AUROC of 95% Confidence Intervals range equal to 0.86 – 0.88 (Table 2b).



(a) Confusion Matrix

(b) ROC Curve

Figure 6:
(a) Confusion Matrix (*Adapted from* Fawcett[55])
(b) Receiver Operating Characteristic (ROC) curves of the *Mean* and 95% *Confidence Interval (CI)*of landslide probabilities.

Table 2 (b)
Area under the ROC curve (AUROC) of *Mean* and *95% Confidence Intervals(CI)*of landslide probabilities

AUROC of Mean	AUROC of 95% CI
0.87	0.86 – 0.88

We created the landslide susceptibility map (Fig. 7) with the mean values (μ) of the landslide probability estimates corresponding to each grid cell. The mean values of these probability estimates ranged from 0.0001 – 0.97. The susceptible classes in the map are categorized according to the *natural break*[58] classification of the probability range as *Low* (0 < p < 0.25); *Medium-Low* (0 .25 < p < 0.50); *Medium-High* (0.50 < p < 0.75); and *High* (0.75 < p < 1). Accordingly, 4% of the study area falls within the *High* susceptible class, 13% of the study area falls within the *Medium-High* susceptible class, 17% of the study area falls within the *Medium-Low* susceptible class, and 66% of the study area falls within the *Low* susceptible class.

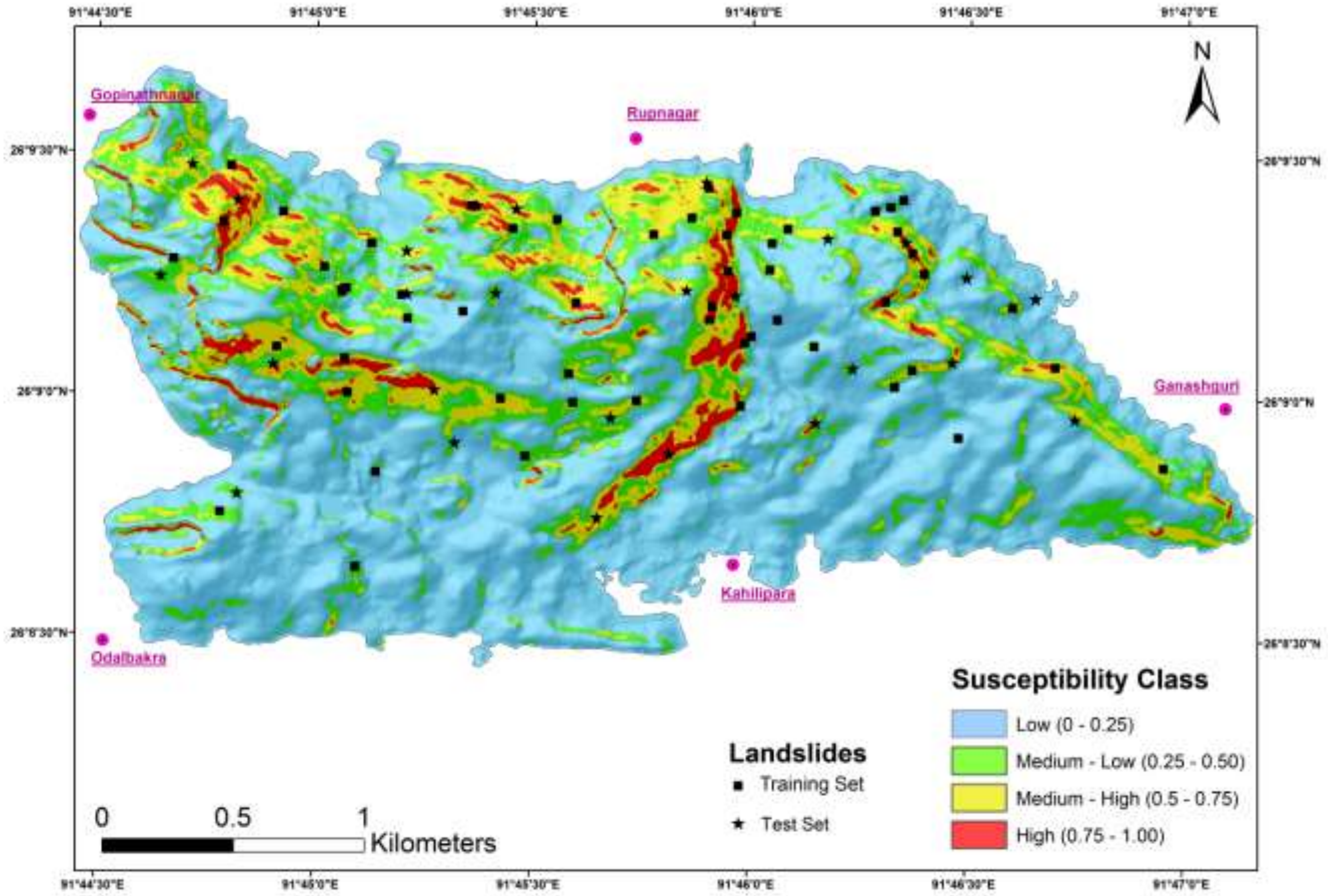


Figure 7: Landslide Susceptibility Map of the Narakasur Hill with the distribution training and test sets of landslides

The LS map classified 87% of the 25° – 40° and 75% of the > 40° slope angle categories as unstable, with probability values greater than 0.75. The 25° – 40° and > 40° slope angle categories constitute together 8% of the study area. Steep slope angles hold the control of the locations of shallow landslides in the study area as confirmed by the absence of a direct correlation between steepest slopes and specific classes of other predictors that seem to influence the distribution of failures (slope aspect, profile curvature, and land use). In fact, we mapped shallow landslides on hillslopes within most land use classes (Forest Land, Barren Lands, Forest Land with sparse settlements, Urban Settlements), profile curvature (> 0) categories and slope aspect (150°– 270°) classes.

Hillslopes with average gradient, generally between 15° and 25°, are also the sites of many failures. Urban Settlement and Forest Land with sparse settlements are the most general land use types in these slopes, while the other morphological and hydrological attributes do not have typical patterns. The percentage of correct prediction of the susceptibility of these hillslopes is 69%. This analysis implies that the GAM is efficient in identifying the relationships between morphological, hydrological, geological and land use features that cause shallow landslides on hillslopes characterised by moderate values of slope angles.

Spatial analysis of land use distribution and the LS map highlights that hillslopes with Urban Settlements and Forest Land with sparse settlements are significantly prone to shallow failures. Land-use planning initiatives in the study area have to account for this fundamental aspect: the ongoing rate of increase in urban settlements at the expense of natural forests is detrimental to the stability of slopes in the study area.

V. CONCLUSION

The present study, to the best of our knowledge, attempted for the first time to create an LS map in Guwahati on a local scale of 1:10,000, based on a data-driven statistical method. In the process, we mapped landslide occurrences to create an inventory for the first time in the Guwahati city area. However, workforce constraints for mapping at 1:10,000 scale permitted us to make a detailed field survey of only a part of the city, i.e. the Narakasur Hill, with an area of ≈ 6 Km^2.

The reasons for choosing the *Generalized Additive Model*(GAM), for the present study, is *three-fold*: (i) Unlike "black box" *machine learning algorithms*, the GAM is easy to interpret. (ii) It provides flexibility because the relationships between predictors and response are not assumed to be linear. (iii) As mentioned earlier, the GAM can control the smoothness of the predictor functions to prevent overfitting. We implemented the GAM in the R environment for statistical computing, using nine DEM-derived terrain attributes and land use as predictor variables. A multicollinearity analysis excluded the *catchment area*due to its correlation with other predictors. Bootstrap replications (100-fold) identified the *slope angle*, *slope aspect*, *profile curvature*, and *land use* as the predictors with maximum influence.

The prediction accuracy is evaluated through a repeated hold-out method with a random bootstrap sub-sampling of 100 different training and test sets in the proportion of 2:1 (55 landslides for training and 27 landslides for test). The true error rates (mean accuracy and standard deviation) on the training subsets are 0.80 and 0.01 respectively whereas those on the test subsets are 0.79 and 0.02 respectively. The mean, standard deviation and other descriptive statistics of the probability (susceptibility) estimates were obtained for each grid cell from the ensembles of 100 bootstrap model runs. The ROC curves plotted with the test subsets demonstrate a good prediction performance with the AUROC of Mean equal to 0.87 and AUROC of 95% CI ranging from 0.86 to 0.88. The four-fold plot of the LS scenario portray that the percentage of over-prediction is 16.3, and the percentage of under-prediction is 27.5. In contrast, the percentages of correct predictions are: 84.7 (stable) and 72.5 (unstable).

We created the final shallow LS map with the mean values (μ) of the landslide probability estimates corresponding to each grid cell. Accordingly, 4% of the study area falls within the High ($0.75 < p < 1$) susceptible class, 13% within the Medium-High (($0.50 < p < 0.75$) susceptible class, 17% within the Medium-Low ($0 .25 < p < 0.50$) susceptible class, and 66% within the Low ($0 < p < 0.25$) susceptible class.

The present study concludes that a data-driven model such as the GAM is more suitable due to its ability for determining most of the inter-relationships between morphological, hydrological, geological and land use features that cause shallow landslides occurrence. Moreover, the LS map produced by the GAM model completely reflects the negative role of new built-up areas on unstable slopes. Urbanisation without proper land use planning increases the probability of landslide occurrence even in stable sectors that were not affected by instability before anthropogenic impacts. The study, thus, demonstrates the need to hypothesise landslide susceptibility scenarios based on different rates of urbanisation as well as different return periods for rainfall events that act as triggering events for shallow landslide occurrences.

ACKNOWLEDGEMENT

This article is a part of the PhD research that is being carried out by the corresponding author at Gauhati University (Department of Geological Sciences). The authors are indebted to the University Grants Commission, New Delhi, India for financial support to carry out this research project. The authors acknowledge the provisions of the SAR (CoSSC) data of the TanDEM-X Mission by the German Aerospace Center (DLR) for the purpose of scientific research, without which this study would not have been possible.

REFERENCES

[1] Fell, R., Corominas, J., Bonnard, Ch., Cascini, L., Leroi, E. and Savage, W.Z. (2008) Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. Eng. Geol., 102: 85–98.

[2] Phukon, P., Chetia, D. and Das P. (2012) Landslide Susceptibility Assessment in the Guwahati City, Assam using Analytic Hierarchy Process (AHP) and Geographic Information System (GIS). International Journal of Computer Applications in Engineering Sciences, 2(1): 2231–4946.

[3] Dutta P.J. and Sarma S. (2013) Landslide susceptibility zoning of the Kalapahar hill, Guwahati, Assam state, (India), using a GIS-based heuristic technique. International Journal of Remote Sensing & Geoscience (IJRSG), 2 (2): 49-55.

[4] Bhusan K., Kundu S.S., Goswami K., Sudhakar S. (2014) - Susceptibility mapping and estimation of rainfall threshold using space based input for assessment of landslide hazard in Guwahati city in North East India. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol.XL-8, 15-19.

[5] Das S., Ray R.K. & Nain G. (2014) - GIS Based Landslide Hazard Zonation of Guwahati Region. International Journal of Engineering Development and Research, Vol. 2 (4), 4005-4014.

[6] Chiranjib Prasad Sarma, A. Murali Krishna, Arindam Dey (2015) Landslide hazard assessment of Guwahati region using physically based models. 6th Annual Conference of the International Society for Integrated Disaster Risk Management (IDRIM-TIFAC, January 2015), At NewDelhi, India

[7] Thiery, Y., Malet, J.P., Sterlacchini, S., Puissant, A., Maquaire, O., (2007) Landslide susceptibility assessment by bivariate methods at large scales: application to a complex mountainous environment. Geomorphology 92 (1–2): 38–59.

[8] Goetz, J.N., Brenning, A., Petschko, H., Leopold, P. (2015) Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Computers & Geosciences 81: 1–11.

[9] Varnes, D. J.: Landslilde hazard zonation: a review of principles and practice, United Nations Educational, Scientific and Cultural Organization, Paris, France, 1984.

[10] Brenning, A., Spatial prediction models for landslide hazards: review, comparison and evaluation (2005) Natural Hazards and Earth System Sciences, 5: 853–862.

[11] Goetz, J.N., Guthrie, R.H. and Brenning, A. (2011) Integrating physical and empirical landslide susceptibility models using generalized additive models. Geomorphology, 129 (3): 376–386.

[12] 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 & Geosciences, 51: 350–365.

[13] Frattini, P., Crosta, G., Carrara, A., (2010) Techniques for evaluating the performance of landslide susceptibility models. Eng. Geol. 111: 62–72.

[14] Brenning, A., (2012) Improved spatial analysis and prediction of landslide susceptibility: practical recommendations. In: Eberhardt E. Froese, C., Turner A.K., Leroueil S. (Eds.). Landslides and Engineered Slopes: Protecting Society through Improved Understanding. Proceedings ofthe 11th International and 2nd North American Symposium on Landslides and Engineered Slopes, vol. 1., Banff, Canada, 3–8 June 2012, CRC Press/Balkema Leiden, the Netherlands, pp. 789–794.

[15] Sarma, K.P. and Maswood, M.D. (1998) Structure controlled mode of emplacement of pegmatite around Guwahati, Kamrup district, Assam. Ind. J. Geochem., 13: 25-32.

[16] Migon, P. and Prokop, P. (2013) Landforms and landscape evolution in the Mylliem Granite Area, Meghalaya Plateau, Northeast India. Singapore J. Trop. Geo., 34: 206–228.

[17] Cruden, D.M., Varnes, D.J. (1996) Landslide types and processes. In Turner AK, Schuster RL (eds.) Landslides: investigation and mitigation, National Academy Press, Washington D.C., 36–75.

[18] Hengl, T. (2006) Finding the right pixel size. Computers & Geosciences, 32(9): 1283-1298.

[19] Moore, I.D., Grayson, R.B., Ladson, A.R., (1991) Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5 (1), 3–30.

[20] Wilson, J. P. (2012) Digital terrain modeling. Geomorphology, 137: 107–121.

[21] Beven, K.J., Kirkby, M.J., (1979) A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. Bull., 24: 43–69.

[22] Guisan A., Weiss S.B., Weiss A.D. (1999) GLM versus CCA spatial modeling of plant species distribution. Kluwer academic publishers. Plant Ecol., 143:107–122

[23] Weiss A (2001) Topographic position and landforms analysis. In: Poster presentation, ESRI user conference, San Diego, CA

[24] Riley, S.J., De Gloria, S.D., Elliot, R. (1999): A Terrain Ruggedness that Quantifies Topographic Heterogeneity. Intermountain Journal of Science, 5 (1-4): 23-27.

[25] Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J. (2015) System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev., 8: 1991–2007.

[26] Wilson, J.P., Gallant, J.C., (2000). Digital terrain analysis. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis: Principles and Applications. John Wiley and Sons, New York, pp. 1–27.

[27] Zevenbergen L.W., Thorne C.R. (1987) Quantitative analysis of land surface topography. Earth Surf. Proc. Landforms, 12: 47–56.

[28] Catani F, Lagomarsino D, Segoni S, Tofani V (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci 13: 2815–2831.

[29] Van Westen CJ, Castellanos E, Kuriakose SL (2008) Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. Eng. Geol 102: 112–131.

[30] Quinn P, Beven K, Chevallier P, Planchon O (1991) The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrol Proc 5:59–79

[31] Brenning A, Schwinn M, Ruiz-Páez AP, Muenchow J (2015) Landslide susceptibility near highways is increased by 1 order of magnitude in the Andes of southern Ecuador, Loja province. Nat. Hazards Earth Syst. Sci., 15: 45–57.

[32] Seibert J, Stendahl J, Sørensen R (2007) Topographical influences on soil properties in boreal forests. Geoderma 141: 139–148

[33] Sterlacchini, S., Ballabio, C., Blahut, J., Masetti, M. and Sorichetta, A. (2011) Spatial agreement of predicted patterns in landslide susceptibility maps. Geomorphology 125: 51–61

[34] Phillips, J.D., (2003) Sources of nonlinearity and complexity in geomorphic systems. Progress in Physical Geography 27: 1–23.

[35] Hastie, T.J., Tibshirani, R., 1990. Generalized Additive Models.. Chapman & Hall, London, p. 352.

[36] McCullagh, P. and Nelder, J. A. (1989) Generalized Linear Models, Second Edition, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, p. 532.

[37] Guisan, A., Edwards, T. C., and Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene, Ecol. Model., 157: 89–100.

[38] Min, J., Park, J.B., Lee, K., and Min, K. (2015): The impact of occupational experience on cognitive and physical functional status among older adults in a representative sample of Korean subjects. Annals of Occupational and Environmental Medicine 27:11

[39] Petschko, H., Bell, R., Brenning, A., and Glade, T. (2012) Landslide susceptibility modeling with generalized additive models – facing the heterogeneity of large regions, in: Landslides and Engineered Slopes, Protecting Society through Improved Understanding, Vol. 1, edited by: Eberhardt, E., Froese, C., Turner, A. K., and Leroueil, S., Taylor & Francis, Banff, Alberta, Canada, 769–777.

[40] Brenning A (2008) Statistical geocomputing combining R and SAGA: the example of landslide susceptibility analysis with generalized additive models. In: Böhner J, Blaschke T, Montanarella L (eds) SAGA — Seconds Out (= Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie, 19), Hamburg, pp 23–32.

[41] Jia, G., Tian, Y., Liu, Y., and Zhang, Y.: A static and dynamic factors-coupled forecasting model of regional rainfall-induced landslides: A case study of Shenzhen, Sci. China Ser. E, 51, 164–175, 2008.

[42] Petschko, H., Brenning, A., Bell, R., Goetz, J., Glade, T., 2014. Assessing the quality of landslide susceptibility maps – case study Lower Austria. Nat. Hazards Earth Syst. Sci. 14, 95–118.

[43] Akaike, H., (1974) A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723.

[44] Burnham, Kenneth P. and David R. Anderson (2002) Model Selection and Inference: A Practical Information-Theoretical Approach, 2nd Ed., New York: Springer-Verlag

[45] Hosmer, D.W., Lemeshow, S., (2000) Applied Logistic Regression, 2nd edition. John Wiley & Sons, New York, p. 373.

[46] R Core Team (2016) R: A language and environment for statisticalcomputing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[47] Hastie T. J. (2016) gam: Generalized Additive Models. R package version 1.14. http://CRAN.Rproject.og/package=gam.

[48] Angelo Canty & Brian Ripley (2016) boot: Bootstrap R (S-Plus) Functions. R package version 1.3 - 1.8.

[49] Davson, A. C. & Hinkley, D. V. (1997) Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge. ISBN: 0-521-57391-2

[50] Hendrickx J. (2015) perturb: Tools for evaluating collinearity. R package version 2.05. http://CRAN.Rproject.og/package=perturb

[51] Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T. (2015) ROCR: Visualizing the Performance of Scoring Classifiers. R package version 1.0-7. http://cran.r-project.org/package=ROCR.

[52] Belsley D.A., Kuh E., Welsch R.E. (1980) Regression diagnostics: identifying influential data and sources of collinearity. John Wiley & Sons, New York, pp. xv + 292. DOI: 10.1002/0471725153

[53] Efron, B. and Tibshirani, R.J. (1994): An Introduction to the Bootstrap. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, pp. 456.

[54] Maindonald J., Braun W.J. (2010) Data Analysis and Graphics Using R: An Example- Based Approach. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge, pp. 552.

[55] Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., and Galli, M. (2006) Estimating the quality of landslide susceptibility models, Geomorphology, 81: 166–184.

[56] Fawcett, T. (2006) An introduction to ROC analysis. Pattern Recogn. Lett., 27: 861–874.

[57] Hanley, J.A., McNeil, B.J., (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.

[58] Jenks G. (1967) The Data Model Concept in Statistical Mapping. International Yearbook of Cartography 7:186-90