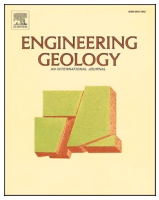
Engineering Geology 311 (2022) 106902

Contents lists available at ScienceDirect

Engineering Geology

journal homepage: www.elsevier.com/locate/enggeo



Predicting landslide runout paths using terrain matching-targeted machine learning

Lu-Yu Ju [a], Te Xiao [a,*], Jian He [a], Hao-Jie Wang [a], Li-Min Zhang [a,b]

[a] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[b] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China

ARTICLE INFO

Keywords:
Landslide
Landslide risk
Runout path
Terrain matching
Machine learning

ABSTRACT

Landslide debris will travel certain distances and threaten people and properties along its runout path, highlighting the importance of runout path prediction in landslide risk management. Conventional landslide runout models, either statistical or machine learning-based, only consider the geographic characteristics at the source without fitting the terrain features along the path. A novel terrain matching-targeted machine learning model is proposed for predicting landslide runout paths, in which a consistent terrain matching strategy is introduced for both training and prediction. The model first forward predicts multiple travel distances based on geographic characteristics of all cells along a possible runout path, and then determines the termination cell whose predicted travel distance fits the terrain features best to backward estimate model parameters. Such a terrain matching process not only accounts for geographic characteristics along the paths but also enables the incorporation of three-dimensional terrain reality into model training. A case study of natural terrain landslides in Hong Kong is conducted to validate the proposed machine learning model. Results indicate that the terrain matching-targeted machine learning models significantly outperform conventional statistical models in terms of prediction accuracy. The fall height and landslide scale are the most critical physical factors affecting travel distances of channelized landslides and open hillslope landslides, respectively. The landslide runout model is applied to the Mid-Levels at the foot of Victoria Peak to identify high-risk urban areas vulnerable to landslides, which provides guidelines for designing landslide prevention and mitigation measures.

1. Introduction

Landslides are one of the most severe geological disasters in mountainous areas (Crosta and Frattini, 2003; Huang and Fan, 2013; Tang et al., 2019; Tonini et al., 2022). As a vast number of landslides could be triggered by a heavy rainstorm or a strong earthquake, predicting landslide occurrence in such events is a hot topic in engineering geology (Dai and Lee, 2001; Ko and Lo, 2016, 2018; Allstadt et al., 2022; Xiao et al., 2022). These landslides often travel certain distances on hilly terrains or along stream channels, greatly threatening the people and properties not only within the landslide source areas but also along the larger landslide runout areas. It is, therefore, necessary to rapidly predict the post-landslide runout paths at a regional scale to better assess and manage landslide risks.

Numerical runout simulations considering landslide mobility mechanism and terrain reality are usually limited to an individual landslide (Hungr, 1995; Soga et al., 2016), while statistical methods are widely applied to regional landslide runout analysis due to the high efficiency. The basic idea of the latter is to establish a statistical relationship between the influential factors, such as fall height H (i.e., the elevation difference between landslide source and termination location), and the dependent variables related to runout distance, such as travel distance L (i.e., the horizontal projection of movement path along the longitudinal profile) or angle of reach $\alpha = \tan^{-1}(H/L)$, as shown in Fig. 1. From the perspective of energy conservation, it is understandable that a high correlation exists between the travel distance, denoting energy loss, and the fall height, indicating potential energy, as adopted in many statistical analyses (Nicoletti and Sorriso-Valvo, 1991; Chen et al., 2015; Mitchell et al., 2020; Gao et al., 2021; Zhao et al., 2022). Considering that both variables are indeed unknown before the stop of landslide mass movement, many studies also prefer to use the angle of reach to combine them into a single dependent variable for simplicity (Hsu,

* Corresponding author.
E-mail addresses: lju@connect.ust.hk (L.-Y. Ju), xiaote@ust.hk (T. Xiao), jhebl@connect.ust.hk (J. He), h.wang@connect.ust.hk (H.-J. Wang), cezhangl@ust.hk (L.-M. Zhang).

https://doi.org/10.1016/j.enggeo.2022.106902
Received 1 October 2021; Received in revised form 26 September 2022; Accepted 26 October 2022
Available online 30 October 2022

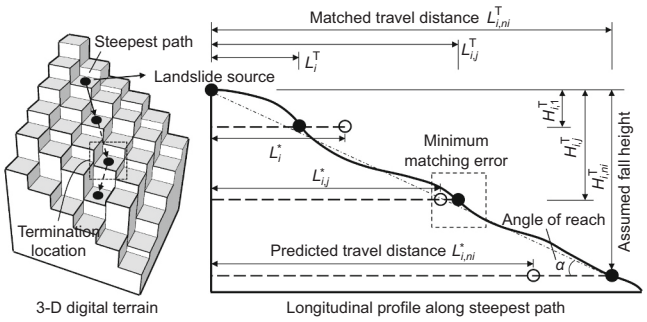


Fig. 1. Trial-and-error terrain matching for landslide runout path prediction.

1975; Corominas, 1996; Lau and Woods, 1997; Finlay et al., 1999; Zhang et al., 2012; Zhan et al., 2017).

On the one hand, the fall height, travel distance, and many other variables used in statistical models are mainly point variables that cannot preciously characterize the entire runout trace in a three-dimensional (3-D) space. For example, the geographic characteristics along the path (e.g., spatial changes in slope gradient and geomaterials), which determine internal energy loss and how far the landslide can move, are rarely considered in statistical models. On the other hand, the fall height and travel distance are coupled like a chicken-and-egg problem. With the unknown fall height, a trial-and-error terrain matching process is required to construct the 3-D runout path. As shown in Fig. 1, the longitudinal profile of a possible runout path starting from the landslide source is first extracted from a 3-D digital terrain model (DTM), and all cells along the path are visited one by one to determine where the landslide should terminate. For statistical models, such a terrain matching process can be adopted for path prediction after the model training but is hardly involved during the model training. The inconsistency in terrain matching may significantly reduce the accuracy of statistical models in landslide runout path prediction, despite a strong correlation between the travel distance and fall height.

Machine learning techniques have emerged as a powerful tool in many landslide problems (Goetz et al., 2015; Bui et al., 2016; Choubin et al., 2019; Yang et al., 2019; Merghadi et al., 2020; Amatya et al., 2021; Su et al., 2021, 2022; Wang et al., 2021a, 2021b; Kainthura and Sharma, 2022; Wu et al., 2022), particularly landslide identification and susceptibility mapping related to landslide occurrence. Their applications to post-landslide runout prediction are rare, maybe because most studies only focus on augmenting the nonlinear prediction capacity but handle the geographic characteristics in an identical way to statistical models, i.e., only considering the point characteristics at landslide sources. Such a machine learning manner cannot improve the runout prediction accuracy fundamentally. Hence, a new strategy is required to introduce geographic characteristics along unknown traces and enable consistent terrain matching in both model training and prediction, eventually toward more reliable landslide runout prediction.

This study aims to develop a runout path prediction model for regional landslides with a novel terrain matching-targeted machine learning technique. The extension of the conventional statistical landslide runout model from distance prediction to path prediction is introduced first, followed by the development of the terrain matching-targeted machine learning model, in which the geographic characteristics along the paths can be properly considered. A case study of natural terrain landslides in Hong Kong is conducted to validate the proposed machine learning model. The model is finally applied to the Mid-Levels at the foot of Victoria Peak to identify high-risk urban areas vulnerable to landslides.

2. Landslide runout path prediction

2.1. Statistical model: extending from distance to path

Traditionally, a statistical model only predicts the landslide travel distance; hence, a trial-and-error terrain matching process (Fig. 1) is required to convert the distance into a path. This study establishes a statistical model by taking the travel distance L as a dependent variable. As shown in Fig. 2(a), in addition to the critical fall height H, many other explanatory variables are also considered to capture the landslide mobility mechanism indirectly, including the geometry of the landslide source (i.e., length L_s and width W_s), and the geographic characteristics (i.e., slope gradient β_s, curvature γ_s, landcover C_s, and geomaterials G_s) and rainfall (i.e., maximum rolling 4-h and 24-h rainfall amounts, $R_{4,s}$ and $R_{24,s}$) at the landslide source. Applying the multivariate linear regression, the landslide travel distance can be predicted as:

$$log(L^*) = \boldsymbol{w}\boldsymbol{x}_s + b \tag{1}$$

where $\boldsymbol{x}_s = [\log(H), L_s, W_s, \beta_s, \gamma_s, C_s, G_s, R_{4,s}, R_{24,s}]$ are the source-dominated explanatory variables used in the statistical model; $\boldsymbol{\theta} = \{\boldsymbol{w}, b\}$ are model parameters; and L^* is the predicted travel distance. Both fall height H and travel distance L are taken logarithm to maintain similar magnitudes with other factors. The least-squares method can be adopted to obtain the optimal model parameters by minimizing a cost function defined as:

$$J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left[log\left(L_i^*\right) - log(L_i)\right]^2 \tag{2}$$

where N is the number of training samples; and L_i and L^*_i are the measured and predicted travel distances, respectively, of the i-th sample in the dataset.

Since fall height is unknown prior to landslide occurrence, a terrain matching process is needed to determine the termination location of landslide mass movement, no matter which prediction model is applied (e.g., statistical or machine learning model). As shown in Fig. 1, it requires a 3-D DTM and a landslide runout analysis model that predicts a travel distance (L) from a given fall height (H) and other factors. Take the i-th landslide sample as an example. A possible runout path starting from the source is identified first on the DTM by assuming that the landslide moves along the steepest path to an adjacent position. Such a steepest path assumption works for soil or soil-like landslides, but may be less reasonable for rock landslides. If the landslide stops at the j-th cell along the path ($j = 1, 2, \ldots, n_i$; and n_i is the total number of cells on the path), the actual fall height and travel distance on the DTM from the landslide source to the termination location are $H^T_{i,j}$ and $L^T_{i,j}$, respectively, while the runout analysis model predicts another travel distance $L^*_{i,j}$ from the given $H^T_{i,j}$. The two distances (i.e., $L^T_{i,j}$ and $L^*_{i,j}$) may not be identical. After going through all cells on the path, the cell with a minimum matching error between $L^T_{i,j}$ and $L^*_{i,j}$ can be taken as the termination location, and the runout path is accordingly determined.

Note that, when fitting a statistical model, the landslide trace information closely related to landslide mobility is out of consideration and the targets of model training and path prediction are inconsistent. For model training, the target is to fit the predicted travel distance (L^*) to the one recorded in the landslide database (L), while it is to match the predicted travel distance to the one derived from the DTM (L^T) in the path prediction. Ignoring trace characteristics and the target inconsistency between model training and prediction may reduce the accuracy of statistical models in landslide runout path prediction. Both deficiencies can be addressed by a terrain matching-targeted machine learning model in the following section.

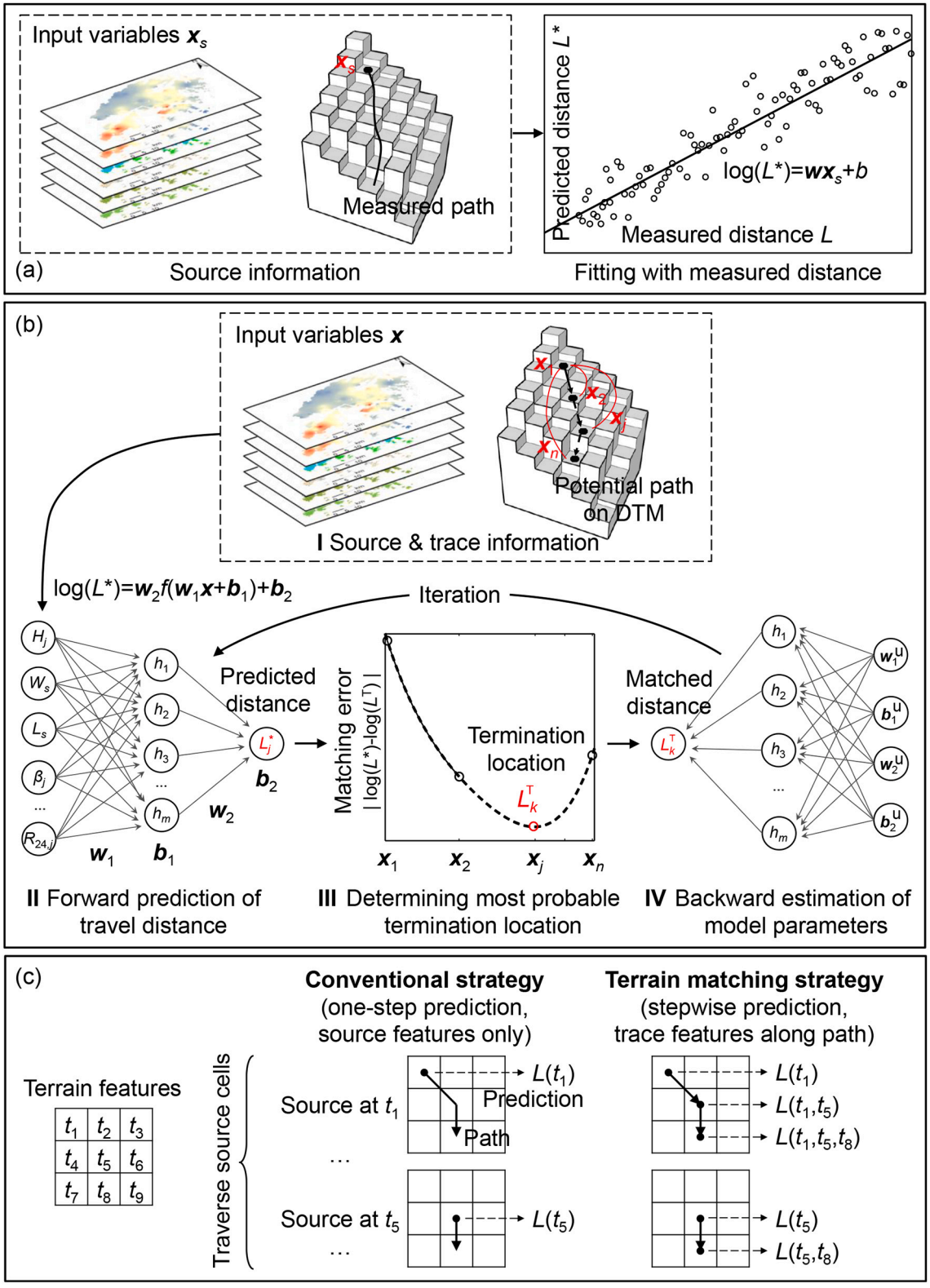


Fig. 2. Landslide runout path prediction: (a) statistical model; (b) terrain matching-targeted machine learning model; (c) stepwise prediction using trace features.

2.2. Terrain matching-targeted machine learning model

Simply substituting a machine learning model for a statistical model but with the same training strategy and explanatory variables is not helpful. Instead, a consistent terrain matching strategy in both model training and prediction is proposed for machine learning-based landslide runout path prediction, as shown in Fig. 2(b). Unlike conventional machine learning models, an additional step for determining the most probable termination location is inserted between the forward runout prediction and backward parameter estimation. The explanatory variables of each cell along the possible path are fed separately into the machine learning model to predict multiple travel distances $L^*_{i,j}$ (i.e.,

Table 1
Sources of databases.

Database	Source (accessed 05/01/2022)
ENTLI	Geotechnical Engineering Office (https://www.geomap.cedd.gov.hk/GEOOpenData/eng/ENTLI.aspx)
Terrain	Lands Department (https://www.landsd.gov.hk/en/spatial-data/opendata.html)
Landcover	Planning Department (https://www.pland.gov.hk/pland_en/info_serv/open_data/landu/index.html) or Gong et al. (2019a) (http://data.ess.tsinghua.edu.cn)
Geology	Geotechnical Engineering Office (https://www.geomap.cedd.gov.hk/GEOOpenData/eng/GeologicalMap.aspx)
Rainfall	Hong Kong Observatory (https://www.hko.gov.hk/en/cis/climat.htm) and Geotechnical Engineering Office (https://www.geomap.cedd.gov.hk/GEOOpenData/eng/Raingauge.aspx)

Step II). Afterward, the cell k with a minimum matching error between $L^*_{i,j}$ and the actual distance $L^{\mathrm{T}}_{i,j}$ on the DTM is regarded as a temporary termination cell (i.e., Step III), and the actual distance $L^{\mathrm{T}}_{i,k}$ is then fed back into the machine learning model to estimate model parameters (i.e., Step IV). These procedures are iteratively repeated until the optimal model parameters are obtained.

Take a widely-used machine learning method, i.e., neural network, as an example. If only one hidden layer is adopted, the network can be written as:

$$log(L^*) = \boldsymbol{w}_2 f(\boldsymbol{w}_1 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{3}$$

where $\boldsymbol{x}$ are the trace-dominated explanatory variables as described later; $\boldsymbol{\theta} = \{\boldsymbol{w}_1, \boldsymbol{b}_1, \boldsymbol{w}_2, \boldsymbol{b}_2\}$ are model parameters; and $f(\cdot)$ is the activation function (e.g., sigmoid function). One hidden layer is found satisfactory in this study because it can approximate arbitrarily with any function containing a continuous mapping from one finite space to another (Heaton, 2008). For the preceding terrain matching process, the cost function to be minimized can be written as:

$$J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left[log\left(L^{\mathrm{T}}_{i,k}\right) - log(L_i)\right]^2 \tag{4}$$

where $L^{\mathrm{T}}_{i,k}$ is the most probable travel distance at location k on the DTM for the i-th landslide and the termination location k can be determined according to the fitness of terrain as:

$$k = \underset{j=1,2,\ldots,n_i}{argmin}\left|log\left(L^*_{i,j}\right) - log\left(L^{\mathrm{T}}_{i,j}\right)\right| \tag{5}$$

where $L^*_{i,j}$ is predicted by Eq. (3) with $\boldsymbol{x} = \boldsymbol{x}_{i,j}$, the explanatory variables at the j-th cell on the i-th landslide path. Combining Eqs. (4) and (5) makes the predicted travel distance (L^*) first fit the one on the DTM (L^{T}) and then the one recorded in the landslide database (L). Through such a pseudo bi-objective optimization, the predicted landslide path can be more comparable with the terrain reality. As this terrain matching strategy only needs the updating of cost function, it can be readily applied to other advanced machine learning algorithms if necessary.

The geographic characteristics along the trace affect the landslide travel distance as they govern the landslide mobility mechanisms (e.g.,

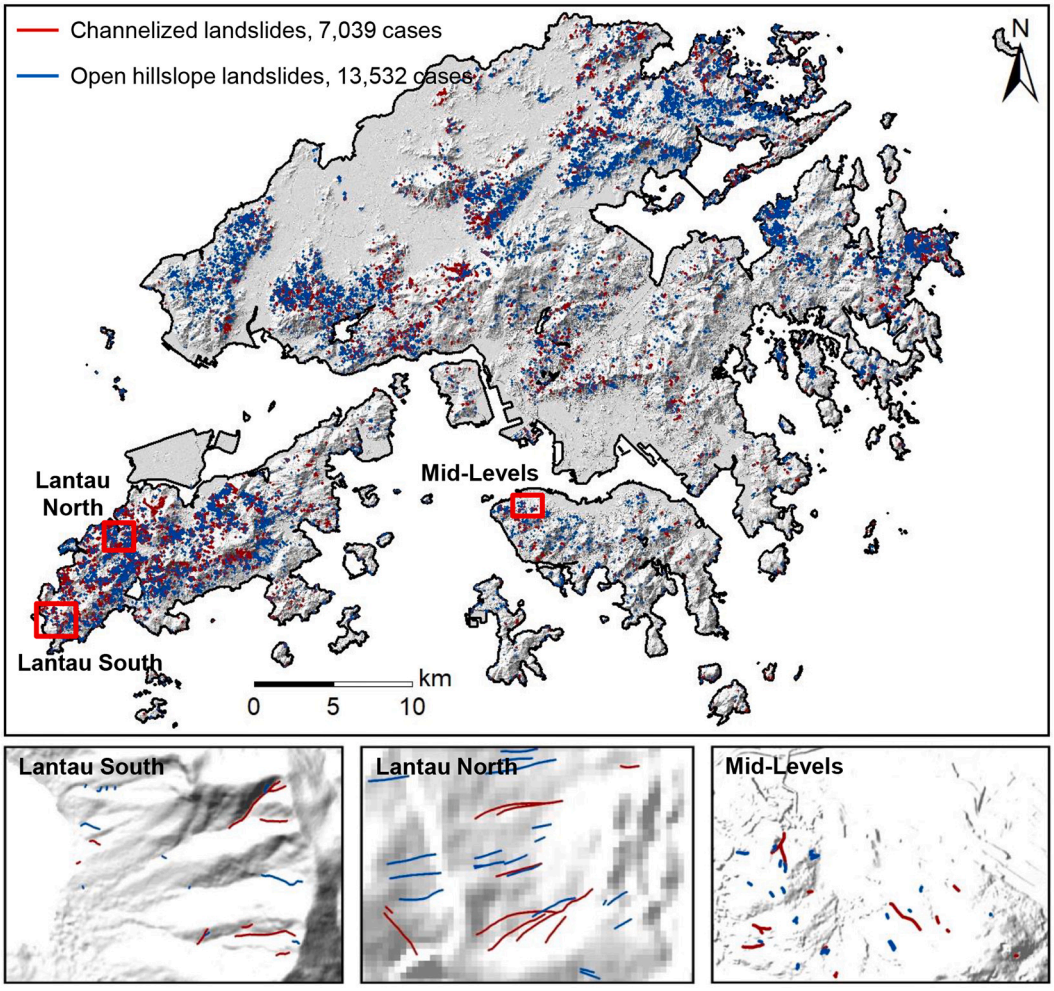


Fig. 3. Traces of channelized landslides and open hillslope landslides in ENTLI (1924–2016).

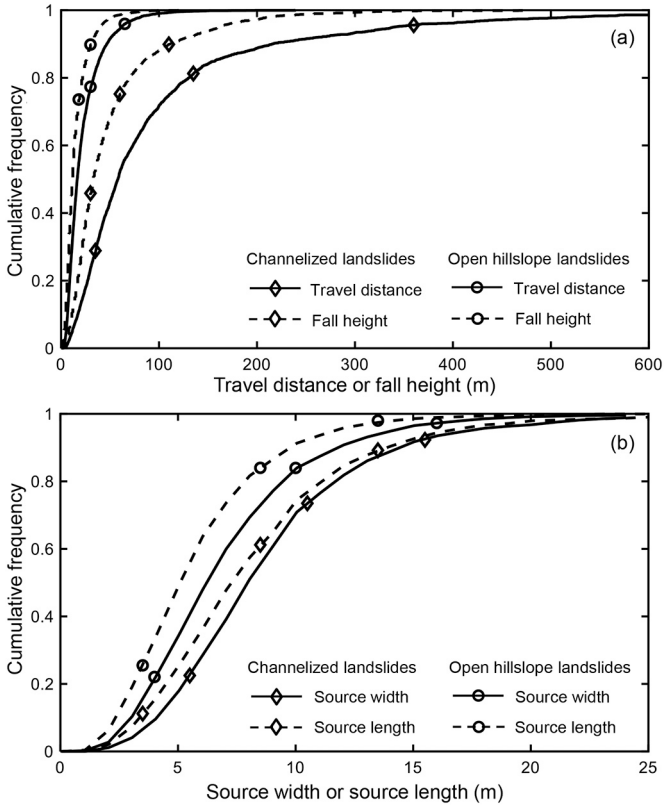


Fig. 4. Statistical distributions for landslides in ENTLI (2000–2016): (a) travel distance and fall height; (b) width and length of landslide source.

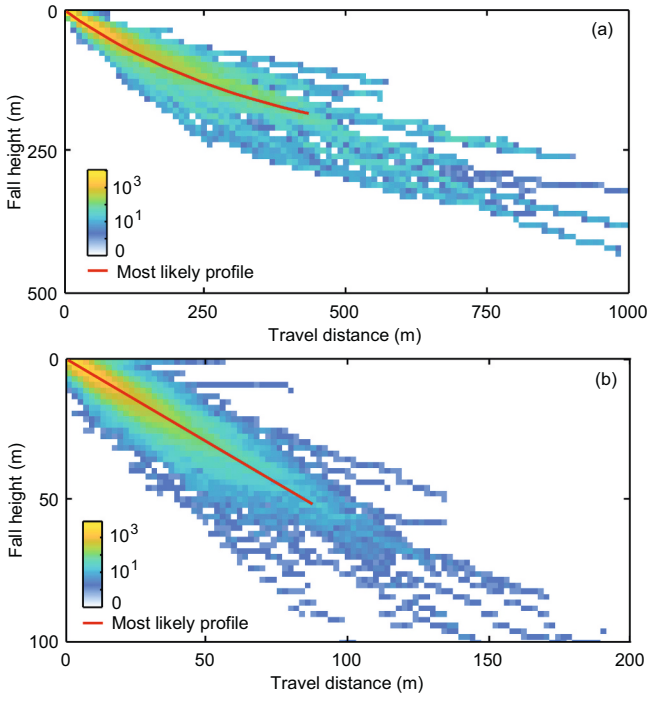


Fig. 5. Longitudinal profiles for landslides in ENTLI (2000–2016): (a) channelized landslides; (b) open hillslope landslides.

velocity and direction). The proposed terrain matching-targeted machine learning model can easily account for these trace characteristics when visiting all possible termination locations. Specifically, the explanatory variables $\boldsymbol{x}_{i,j}$ can include the trace information from the landslide source to the j-th cell, namely $\boldsymbol{x}_{i,j} = [\log(H_j), L_s, W_s, \beta_j, \gamma_j, C_j, G_j, R_{4,j}, R_{24,j}]$. Compared with those in a statistical model, the length and width of the landslide source (i.e., L_s and W_s) are consistent, and H_j is the fall height between the landslide source and the j-th cell. Regarding the trace information, averaged geographic characteristics along the path (from the landslide source to the j-th cell) can be taken into account: numerical variables like slope gradient and curvature take arithmetic averages along the trace, defined as β_j and γ_j, while for categorical variables like landcover and geomaterials, the proportions of the trace covered by each category are utilized instead, defined as C_j and G_j. With respect to the rainfall information, $R_{4,j}$ and $R_{24,j}$ are the maximum rolling 4-h and 24-h rainfall amounts, respectively, summed up from the landslide source to the j-th cell on the path. Fig. 2(c) further compares feature utilization manners in the conventional strategy, adopted in statistical and conventional machine learning models, and the terrain matching strategy. They are fundamentally different despite of the same databases utilized. The conventional strategy uses each feature once when that cell is selected as a landslide source, since the path is not concerned. In contrast, the terrain matching strategy enables stepwise prediction by adding more trace features along the path sequentially as the landslide moves from one cell to another. One feature can be repeatedly used for different paths passing through.

To sum up, the proposed terrain matching-targeted machine learning model has three major contributions. First, a screening process is developed between the forward runout prediction and backward model estimation in model training to capture the site-specific terrain effects on landslides at different locations. Second, the geographic characteristics along the trace can be properly quantified from a deterministic value of the source point to a varying representative value (i.e., average or sum) of the trace line to better capture the physical mechanisms of landslide mobility. Third, the distance on the DTM is adopted to represent the prediction rather than the direct output in standard machine learning algorithms to guarantee that the predicted path is more comparable with the DTM. By this means, the proposed machine learning model can not only provide rapid regional runout path predictions like a statistical model does, but also reasonably incorporate complex geographic characteristics along landslide traces and 3-D terrain reality like in a 3-D numerical simulation.

3. Case study in Hong Kong

This section applies both statistical and machine learning models to predict runout paths of natural terrain landslides in Hong Kong. Hong Kong is a typical mountainous city with more than half of the land area covered by steep natural hillsides. Intense rainfall is frequent from April to October and triggers around 380 natural terrain landslides annually. These landslides are classified as open hillslope landslides remaining wholly on a planar hillslope and channelized landslides moving along stream channels. Most of them are soil landslides.

3.1. Landslide runout path databases

Table 1 presents the sources of databases used in this study. The historical landslide runout paths are extracted from the Enhanced Natural Terrain Landslide Inventory (ENTLI) complied by the Geotechnical Engineering Office of the Hong Kong SAR Government, as shown in Fig. 3. The ENTLI contains >20,000 recent landslides from 1924 to 2016, including 7039 channelized landslides and 13,532 open hillslope landslides, in the form of polylines. This case study focuses on the landslides between 2000 and 2016 (2398 channelized landslides and 4100 open hillslope landslides) because the occurrence time of these landslides is close to the production time of the DTM (i.e., 2011). The ENTLI records four key features of landslides, namely travel distance, fall height, and length and width of the landslide source, as shown in Fig. 4. Statistically, channelized landslides have longer travel distances,

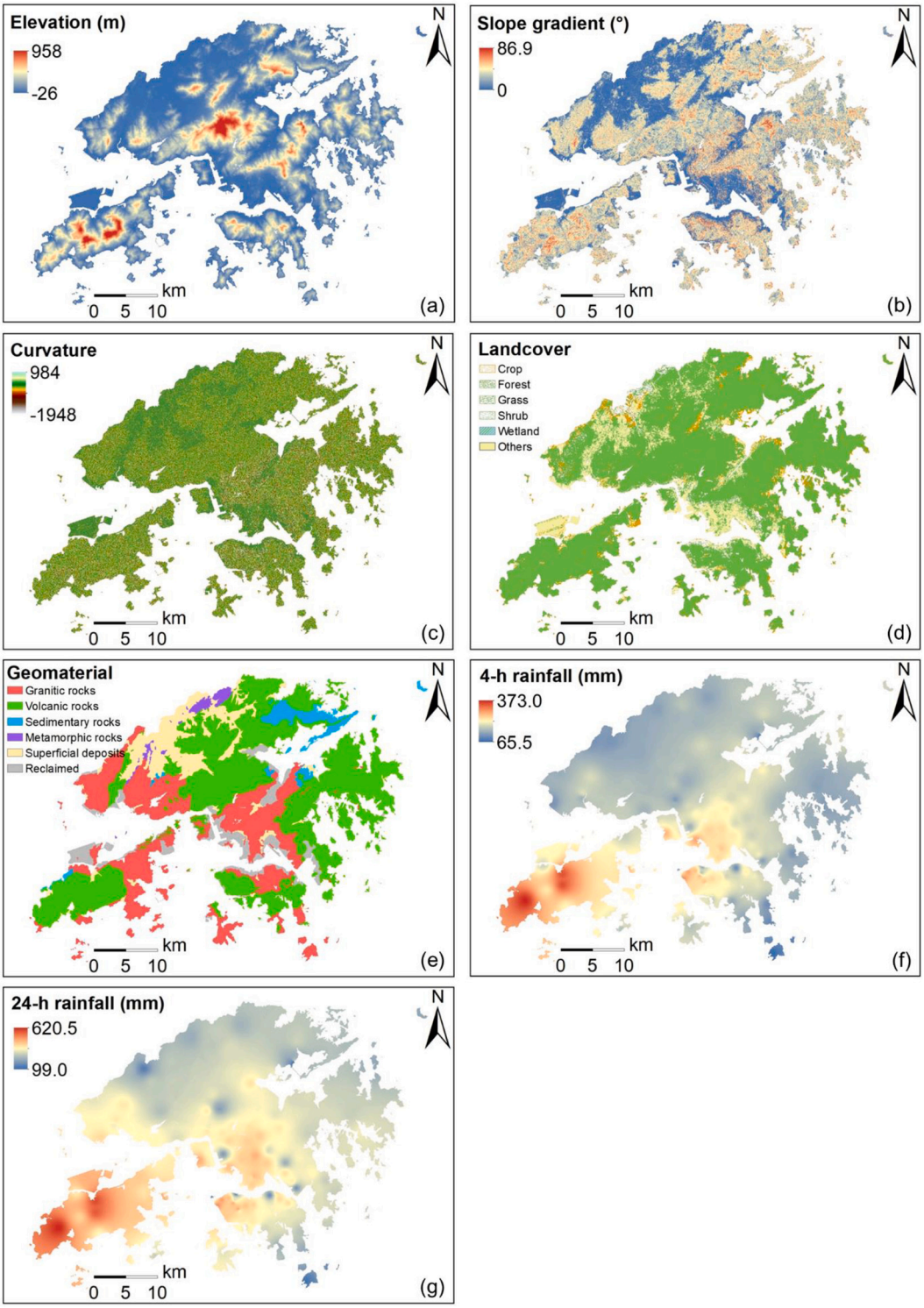


Fig. 6. Topographic, geological and rainfall factors: (a) elevation; (b) slope gradient; (c) curvature; (d) landcover; (e) geomaterials; and maximum rolling (f) 4-h and (g) 24-h rainfall in 2008.

higher fall heights, and larger source areas than open hillslope landslides. Fig. 5 plots longitudinal profiles of all landslides centered at landslide sources, in which the color denotes the number of landslides passing through and the solid lines are the most likely landslide profiles. Channelized landslides gradually move to flat areas after long-distance traveling, while open hillslope landslides are more likely to stay at hillsides (i.e., remaining the same sloping). As a result, two runout path prediction models will be developed separately to reflect the different initiation and mobility mechanisms of channelized landslides and open hillslope landslides.

In addition to the four features from ENTLI, other topographic, geological and rainfall factors influencing landslide travel distances are also considered, as shown in Fig. 6. The slope gradient and curvature are derived from a 5-m resolution DTM. A 10-m resolution landcover map (Gong et al., 2019a) is applied to obtain six major landcover categories: cropland, forest, grassland, shrubland, wetland, and others. Six primary

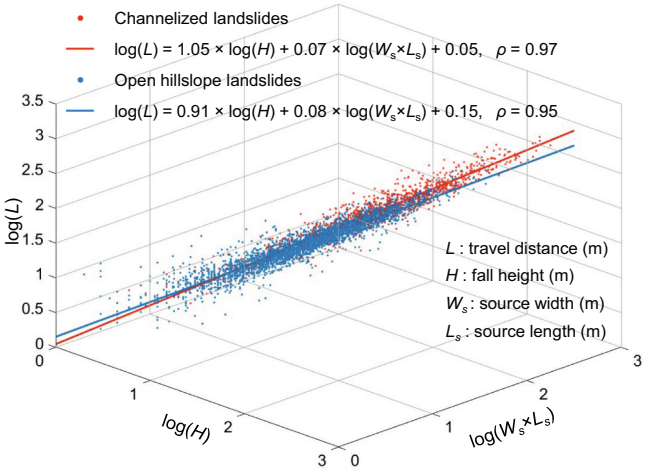


Fig. 7. Simplified statistical landslide runout models.

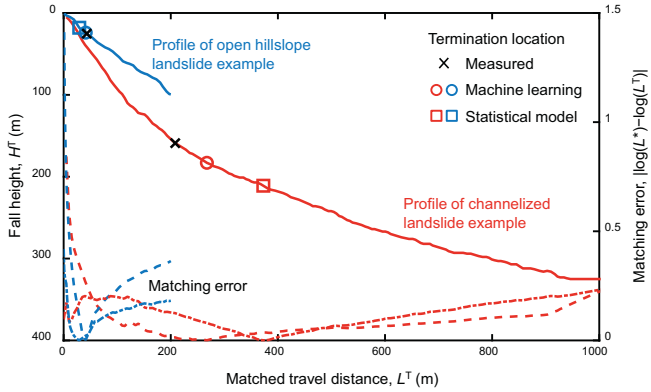


Fig. 8. Terrain matching for two landslide examples.

Table 2
Performance of statistical and machine learning models.

Type	Correlation coefficient			
	Training set		Test set	
	Statistical	Machine learning	Statistical	Machine learning
Channelized landslide	0.39	0.70	0.42	0.70
Open hillslope landslide	0.39	0.74	0.37	0.74

geomaterials are identified from the solid and superficial geological maps: granitic rocks, volcanic rocks, sedimentary rocks, metamorphic rocks, superficial deposits, and reclaimed land. The one-hot encoding is adopted to convert the categorical landcover and geomaterials into two sets of six binary variables. Hourly rainfall data from 141 rain gauges covering the entire Hong Kong are used to produce the maximum rolling 4-h and 24-h rainfall amounts in a specific year, since ENTLI only records the occurrence year of each landslide without knowing the specific failure date. They are interpolated from rain gauge locations to landslide locations using the inverse distance weighting approach. Although only two rainfall variables are needed for a specific year, such as Figs. 6(f) and 6(g) for landslides in 2008, a total of $17 \times 2 = 34$ rainfall variables is generated to cover the entire studied period (i.e., 2000–2016).

In total, nineteen explanatory variables are used for runout path prediction, including fall height, source length, source width, slope gradient, curvature, six landcover variables, six geomaterial variables, and maximum rolling 4-h and 24-h rainfall amounts. Those geographic characteristics are derived from landslide sources for statistical models and from landslide traces for machine learning models. All variables are resampled to a uniform spatial resolution, i.e., 5 m and 1 m for channelized and open hillslope landslides, respectively. This is because the travel distances of open hillslope landslides are much shorter than channelized landslides (Fig. 4(a)), and a higher spatial resolution is necessary.

3.2. Model training and comparison

The databases of channelized landslides and open hillslope landslides are randomly split by a ratio of 70/30 into training sets and test sets to train statistical and machine learning models and test their performances. For statistical models, the length L_s and width W_s of landslide source are integrated as one variable (i.e., $W_s \times L_s$) representing the scale of landslide source. Two simplified statistical models, namely $\log(L) = w_1 \times \log(H) + w_2 \times \log(W_s \times L_s) + b$, with less important explanatory variables fixed at mean values, are shown in Fig. 7. As expected, the correlation coefficients are higher than 0.95, indicating very high fitness. For machine learning models, the structure of neural networks is fixed at one hidden layer with 30 neurons, and the ridge regression technique is applied to avoid overfitting. Fig. 8 gives two examples of the terrain matching process in runout path prediction, in which the travel distance on the DTM (L^T) with minimum matching error is taken

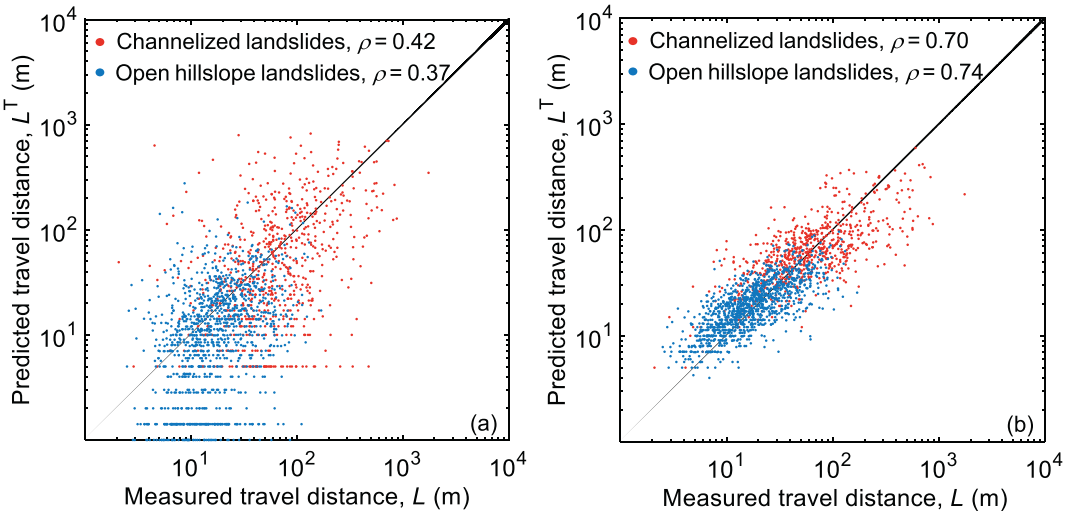


Fig. 9. Measured and predicted landslide travel distances on test sets: (a) statistical model; (b) machine learning model.

as the final predicted travel distance for both statistical and machine learning models.

The performance metrics of statistical and machine learning models are summarized in Table 2, in terms of the correlation coefficient (ρ) between the measured and predicted landslide travel distances. Both models have similar performances for training and test sets without overfitting. The correlation coefficients of machine learning models (i.e., $\rho > 0.7$) are significantly higher than those of statistical models (i.e., $\rho < 0.4$). Fig. 9 further compares the scatters of measured and predicted landslide travel distances on test sets. The results from machine learning models concentrate more along the 1:1 line (i.e., perfect prediction), regardless of the landslide type. Note that, due to the inconsistency of terrain matching in training and prediction, the prediction performances of statistical models (i.e., $\rho < 0.4$) are much worse than their fitness (i.e., $\rho > 0.95$). Therefore, applying statistical models to landslide path prediction should be cautious, even with very high fitness. Conventional machine learning models without the internal terrain matching step are also developed for reference. Their performances are almost consistent with the statistical models (i.e., $\rho = 0.42$ and 0.36 for channelized and open hillslope landslides, respectively). These results indicate that the proposed terrain matching-targeted machine learning model can properly incorporate the geographic characteristics along the paths and the 3-D terrain reality into model training and provide much more accurate predictions on landslide runout paths.

To be more visualized, several typical channelized landslides in Lantau South and open hillslope landslides in Lantau North (Fig. 3) are selected to showcase the predicted 3-D runout paths on satellite images. As shown in Fig. 10, machine learning models provide more consistent paths with the ground truth, while statistical models significantly overestimate the paths of channelized landslides (Fig. 10(b)) and underestimate the paths of open hillslope landslides (Fig. 10(d)). Some predicted landslide paths slightly deviate from the measured paths due to changes in DTM (i.e., these landslides occurred in 2008 but the used DTM is the post-landslide terrain in 2011). Open hillslope landslides are more sensitive to such changes than channelized landslides. Numerical landslide runout simulations are also conducted for comparison using a distributed model EDDA (Shen et al., 2018), in which model parameters have been calibrated for landslides in Hong Kong (Gao et al., 2016; Zhou et al., 2019), as shown in Figs. 10(c) and 10(f). Numerical simulations provide more runout details beyond the path, such as the flow width and depth, but undoubtedly take more computational time, which will grow exponentially when covering a large region. Based on the path prediction, further extending the machine learning model to cover the prediction of path width or impact area will be investigated in the future, with the help of satellite images and numerical simulations.

3.3. Most critical physical parameters

The impact of different explanatory variables on the performance of machine learning models is explored to identify the most critical physical parameters in landslide runout analysis. By extending the physical classification of Bathurst et al. (1997), the explanatory variables can be categorized into four classes with different physical mechanisms: (1) fall height, representing the conversion from potential energy to kinetic energy; (2) source information (i.e., width and length of landslide source), an indicator of momentum; (3) trace information (i.e., slope gradient, curvature, landcover, and geomaterials), reflecting the friction-induced internal energy loss; and (4) rainfall information (i.e., maximum rolling 4-h and 24-h rainfall), an external factor reducing the shear strength and generating surface runoff. Four machine learning models will be re-trained based on four groups of explanatory variables, each of which contains all variables from the previous group and one additional variable class, as shown in Fig. 11. For instance, Group 3 is comprised of Group 2 (i.e., fall height and source information) and trace information.

The correlation coefficient between measured and predicted travel distances is again used to quantify the model performance. Fig. 11 presents the variation of correlation coefficients for different variable groups. For channelized landslides, fall height plays a vital role in the landslide runout prediction. Using this variable only already reaches a relatively satisfactory correlation coefficient of 0.42. Source information helps to improve the prediction moderately, raising the correlation coefficient to 0.61, and slight improvements can be achieved by further adding trace and rainfall information. For open hillslope landslides, source information is more crucial than fall height. The involvement of source information significantly increases the correlation coefficient from 0.25 to 0.69 and a validation model using source information only also gives a high correlation coefficient of 0.64. Therefore, the fall height and landslide scale (i.e., source information) are the most critical physical factors affecting travel distances of channelized landslides and open hillslope landslides, respectively. This highlights the need to separately develop two machine learning models.

4. Identification of potential landslide-affected urban areas

With the developed prediction models of landslide runout paths, it is possible to identify potential landslide-affected urban areas and high-risk landslide-bearing elements. Consider the Mid-Levels at the foot of

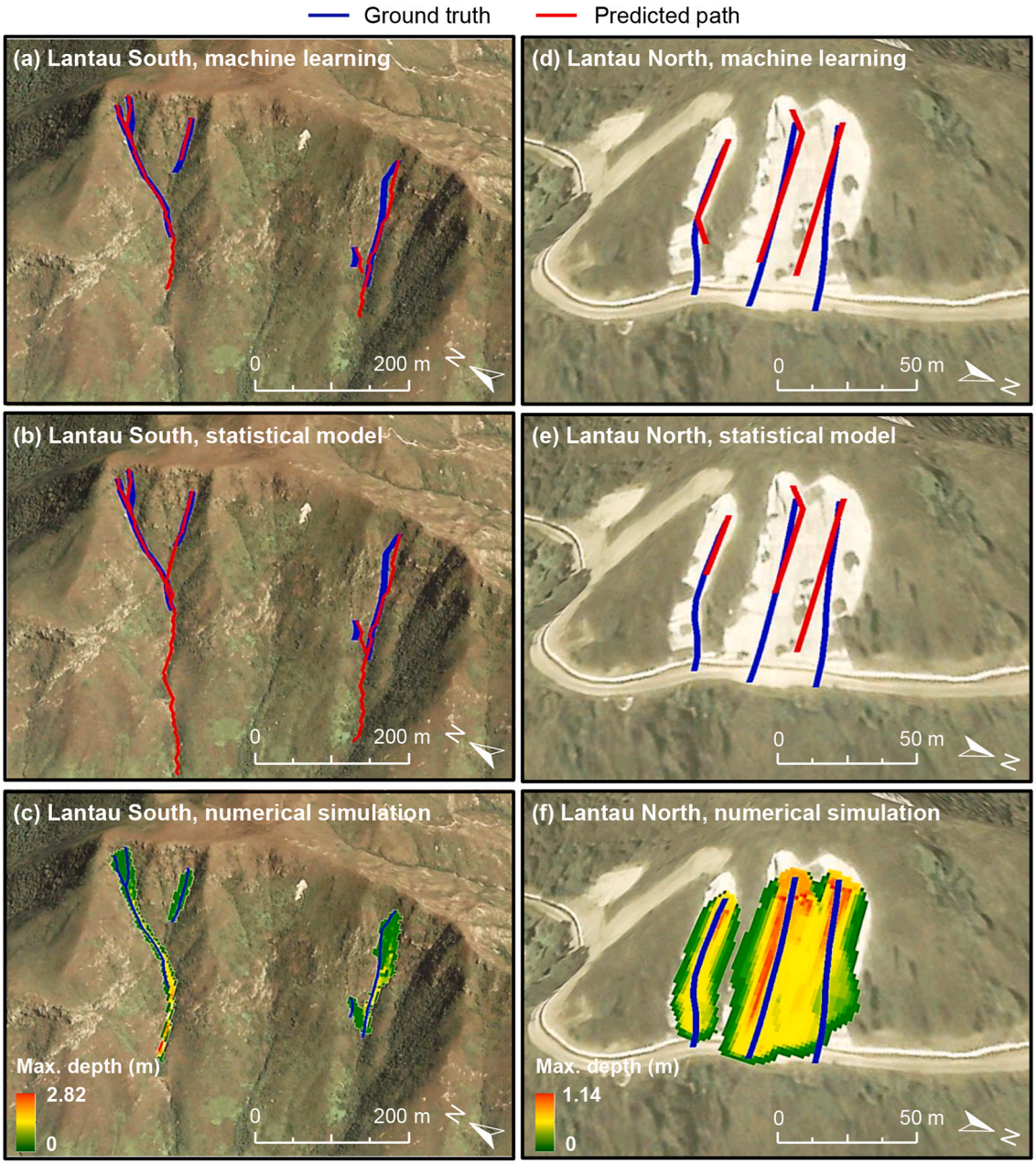


Fig. 10. Examples of landslide runout path prediction (image source: Google Earth): (a) machine learning, (b) statistical, and (c) numerical predictions for channelized landslides; (d) machine learning, (e) statistical, and (f) numerical predictions for open hillslope landslides.

Victoria Peak on Hong Kong Island as an example (Figs. 3 and 12). The Mid-Levels is a landslide-prone area with various buildings densely packed near the steep natural hillsides. One of the most severe rainstorms hitting the Mid-Levels in history is the August 16–21, 2005, rainstorm, with the maximum rolling 4-h and 24-h rainfall amounts being 171 mm and 567.5 mm, respectively.

The probabilities of landslide occurrence at all possible locations can be estimated from either landslide susceptibility analysis (Wang et al., 2019a; Marin and Mattos, 2020; Luo et al., 2021) or slope reliability analysis (Xiao et al., 2016; Gong et al., 2019b; Wang et al., 2019b; Liu et al., 2020). This study chooses the landslide susceptibility map produced by Wang et al. (2021b) as an indicator of landslide occurrence. The natural terrain area in the Mid-Levels is discretized into 30,933 cells (5 m × 5 m in size), and there are 2238 unstable cells with landslide susceptibility >0.5 (i.e., high and very high). Conservatively, only major channelized landslides are considered with landslide sources being 15 m in width and 14 m in length (i.e., 90th percentiles of those in ENTLI), and the rainfall information of the August 16–21, 2005, rainstorm is adopted. By applying the proposed machine learning model, 2238 landslide travel paths can be predicted, among which 1268 landslides rush into the urban area, as shown in Fig. 12(a). About 72% of these landslides stop within 30 m to the mountain foot, and only a few travel a long distance (e.g., 0.5% over 90 m to the mountain foot). The building at the lower right corner of Fig. 12(b) is at the highest risk of being attacked by several potential landslides and demands necessary slope stabilization measures. In fact, an 8 m high concrete retaining wall (11SW-A/R 162) has been constructed between the building and the mountain for safety reasons. Such a landslide runout model is beneficial to designing landslide prevention and mitigation measures for risk management.

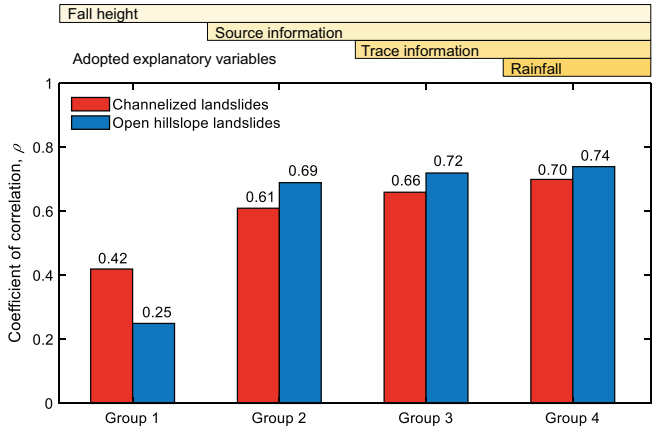


Fig. 11. Impact of explanatory variables on model performance.

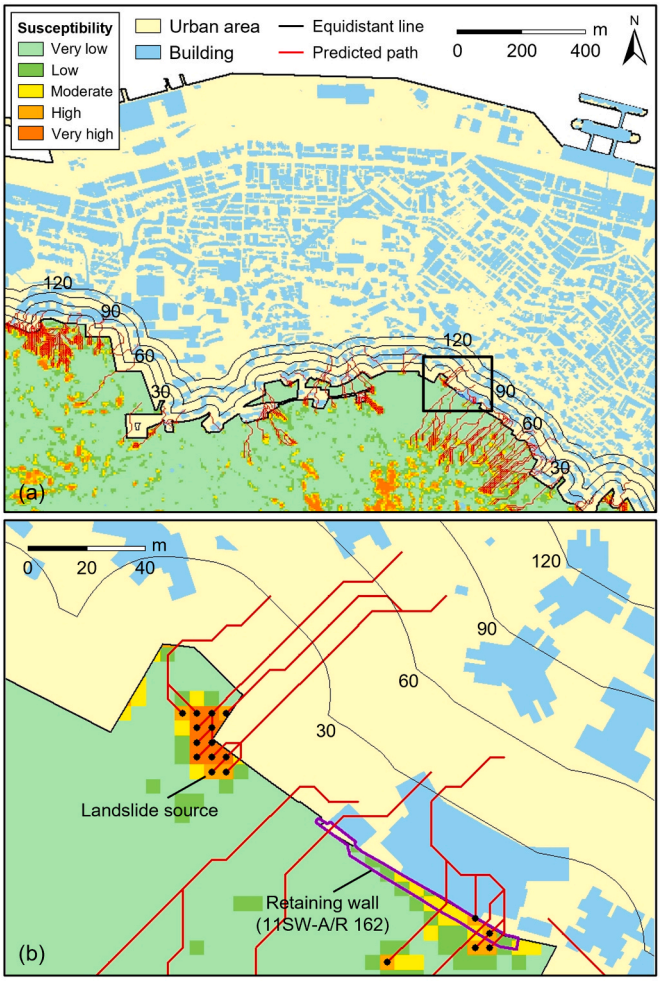


Fig. 12. Predicted landslide-affected areas in the Mid-Levels: (a) overview; (b) zoom-in view.

5. Conclusions

This study proposes a novel terrain matching-targeted machine learning model for landslide runout path prediction and applies the model to a case study of natural terrain landslides in Hong Kong. Major conclusions can be drawn:

(1) Through a terrain matching process inserted between forward runout prediction and backward parameter estimation, the proposed machine learning model can not only provide rapid regional runout path predictions like a statistical model does, but also reasonably incorporate complex geographic characteristics along landslide traces and 3-D terrain reality like in a 3-D numerical simulation.

(2) The machine learning models significantly outperform conventional statistical models in terms of prediction accuracy. The correlation coefficients between the measured and predicted landslide travel distances increase from 0.42 to 0.70 for channelized landslides and from 0.37 to 0.74 for open hillslope landslides.

(3) Two runout path prediction models are developed separately to reflect the different mechanisms of channelized landslides and open hillslope landslides. The fall height and landslide scale are found to be the most critical physical factors affecting travel distances of channelized landslides and open hillslope landslides, respectively.

(4) The landslide runout prediction model is applied to identify urban areas vulnerable to landslides in the Mid-Levels at the foot of Victoria Peak. About 72% of landslides stop within 30 m to the mountain foot, and only 0.5% travel over 90 m. The ability of the proposed models to identify high-risk landslide-bearing buildings helps to design effective landslide prevention and mitigation measures for risk management.

CRediT authorship contribution statement

Lu-Yu Ju: Methodology, Investigation, Software, Validation, Writing – original draft. **Te Xiao:** Methodology, Investigation, Writing – original draft, Writing – review & editing. **Jian He:** Validation, Investigation. **Hao-Jie Wang:** Validation, Investigation. **Li-Min Zhang:** Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

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.

Data availability

Data links have been provided in Table 1.

Acknowledgments

This work was supported by the Research Grants Council of the Hong Kong SAR Government (Project Nos. 16205719, 16203720, and AoE/E-603/18) and the Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (Project No. HZQB-KCZYB-2020083).

References

Allstadt, K.E., Thompson, E.M., Jibson, R.W., Wald, D.J., Hearne, M., Hunter, E.J., Fee, J., Schovanec, H., Slosky, D., Haynie, K.L., 2022. The US Geological Survey ground failure product: near-real-time estimates of earthquake-triggered landslides and liquefaction. Earthq. Spectra 38 (1), 5–36. https://doi.org/10.1177/87552930211032685.

Amatya, P., Kirschbaum, D., Stanley, T., Tanyas, H., 2021. Landslide mapping using object-based image analysis and open source tools. Eng. Geol. 282, 106000 https://doi.org/10.1016/j.enggeo.2021.106000.

Bathurst, J.C., Burton, A., Ward, T.J., 1997. Debris flow run-out and landslide sediment delivery model tests. J. Hydraul. Eng. 123 (5), 410–419. https://doi.org/10.1061/(ASCE)0733-9429(1997)123:5(410).

Bui, D.T., Tuan, T.A., Klempe, H., Pradhan, B., Revhaug, I., 2016. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and

logistic model tree. Landslides 13 (2), 361–378. https://doi.org/10.1007/s10346-015-0557-6.
Chen, H.X., Zhang, L.M., Gao, L., Zhu, H., Zhang, S., 2015. Presenting regional shallow landslide movement on three-dimensional digital terrain. Eng. Geol. 195, 122–134. https://doi.org/10.1016/j.enggeo.2015.05.027.
Choubin, B., Borji, M., Mosavi, A., Sajedi-Hosseini, F., Singh, V.P., Shamshirband, S., 2019. Snow avalanche hazard prediction using machine learning methods. J. Hydrol. 577, 123929 https://doi.org/10.1016/j.jhydrol.2019.123929.
Corominas, J., 1996. The angle of reach as a mobility index for small and large landslides. Can. Geotech. J. 33 (2), 260–271. https://doi.org/10.1139/t96-005.
Crosta, G.B., Frattini, P., 2003. Distributed modelling of shallow landslides triggered by intense rainfall. Nat. Hazards Earth Syst. Sci. 3 (1/2), 81–93. https://doi.org/10.5194/nhess-3-81-2003.
Dai, F.C., Lee, C.F., 2001. Frequency–volume relation and prediction of rainfall-induced landslides. Eng. Geol. 59 (3–4), 253–266. https://doi.org/10.1016/S0013-7952(00)00077-6.
Finlay, P.J., Mostyn, G.R., Fell, R., 1999. Landslide risk assessment: prediction of travel distance. Can. Geotech. J. 36 (3), 556–562. https://doi.org/10.1139/t99-012.
Gao, L., Zhang, L.M., Chen, H.X., Shen, P., 2016. Simulating debris flow mobility in urban settings. Eng. Geol. 214, 67–78. https://doi.org/10.1016/j.enggeo.2016.10.001.
Gao, L., Zhang, L.M., Chen, H.X., Fei, K., Hong, Y., 2021. Topography and geology effects on travel distances of natural terrain landslides: evidence from a large multi-temporal landslide inventory in Hong Kong. Eng. Geol. 292, 106266 https://doi.org/10.1016/j.enggeo.2021.106266.
Goetz, J.N., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput. Geosci. 81, 1–11. https://doi.org/10.1016/j.cageo.2015.04.007.
Gong, P., Liu, H., Zhang, M., et al., 2019a. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. Sci. Bull. 64, 370–373. https://doi.org/10.1016/j.scib.2019.03.002.
Gong, W.P., Tang, H.M., Wang, H., Wang, X.R., Juang, C.H., 2019b. Probabilistic analysis and design of stabilizing piles in slope considering stratigraphic uncertainty. Eng. Geol. 259, 105162 https://doi.org/10.1016/j.enggeo.2019.105162.
Heaton, J., 2008. Introduction to Neural Networks with Java, 2nd ed. Heaton Research, Inc, St. Louis.
Hsu, K.J., 1975. Catastrophic debris streams (sturzstroms) generated by rockfalls. Geol. Soc. Am. Bull. 86 (1), 129–140. https://doi.org/10.1130/0016-7606(1975)86%3C129:CDSSGB%3E2.0.CO;2.
Huang, R., Fan, X., 2013. The landslide story. Nat. Geosci. 6 (5), 325–326. https://doi.org/10.1038/ngeo1806.
Hungr, O., 1995. A model for the runout analysis of rapid flow slides, debris flows, and avalanches. Can. Geotech. J. 32 (4), 610–623. https://doi.org/10.1139/t95-063.
Kainthura, P., Sharma, N., 2022. Machine learning driven landslide susceptibility prediction for the Uttarkashi region of Uttarakhand in India. Georisk 16 (3), 570–583. https://doi.org/10.1080/17499518.2021.1957484.
Ko, F.W.Y., Lo, F.L.C., 2016. Rainfall-based landslide susceptibility analysis for natural terrain in Hong Kong-a direct stock-taking approach. Eng. Geol. 215, 95–107. https://doi.org/10.1016/j.enggeo.2016.11.001.
Ko, F.W.Y., Lo, F.L.C., 2018. From landslide susceptibility to landslide frequency: a territory-wide study in Hong Kong. Eng. Geol. 242, 12–22. https://doi.org/10.1016/j.enggeo.2018.05.001.
Lau, K.C., Woods, N.W., 1997. Review of Methods for Predicting the Travel Distance of Debris from Landslides on Natural Terrain. Technical Note No. 797. Geotechnical Engineering Office, the Hong Kong SAR Government.
Liu, X., Li, D.Q., Cao, Z.J., Wang, Y., 2020. Adaptive Monte Carlo simulation method for system reliability analysis of slope stability based on limit equilibrium methods. Eng. Geol. 264, 105384 https://doi.org/10.1016/j.enggeo.2019.105384.
Luo, J., Zhang, L., Yang, H., Wei, X., Liu, D., Xu, J., 2021. Probabilistic model calibration of spatial variability for a physically-based landslide susceptibility model. Georisk 1-18. https://doi.org/10.1080/17499518.2021.1988986.
Marin, R.J., Mattos, Á.J., 2020. Physically-based landslide susceptibility analysis using Monte Carlo simulation in a tropical mountain basin. Georisk 14 (3), 192–205. https://doi.org/10.1080/17499518.2019.1633582.
Merghadi, A., Yunus, A.P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D.T., Avtar, R., Abderrahmane, B., 2020. Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance. Earth-Sci. Rev. 207, 103225 https://doi.org/10.1016/j.earscirev.2020.103225.
Mitchell, A., McDougall, S., Nolde, N., Brideau, M.A., Whittall, J., Aaron, J.B., 2020. Rock avalanche runout prediction using stochastic analysis of a regional dataset. Landslides 17, 777–792. https://doi.org/10.1007/s10346-019-01331-3.
Nicoletti, P.G., Sorriso-Valvo, M., 1991. Geomorphic controls of the shape and mobility of rock avalanches. Geol. Soc. Am. Bull. 103 (10), 1365–1373. https://doi.org/10.1130/0016-7606(1991)103<1365:GCOTSA>2.3.CO;2.
Shen, P., Zhang, L., Chen, H., Fan, R., 2018. EDDA 2.0: integrated simulation of debris flow initiation and dynamics considering two initiation mechanisms. Geosci. Model Dev. 11 (7), 2841–2856. https://doi.org/10.5194/gmd-11-2841-2018.
Soga, K., Alonso, E., Yerro, A., Kumar, K., Bandara, S., 2016. Trends in large-deformation analysis of landslide mass movements with particular emphasis on the material point method. Géotechnique 66 (3), 248–273. https://doi.org/10.1680/jgeot.15.LM.005.
Su, Z., Chow, J.K., Tan, P.S., Wu, J., Ho, Y.K., Wang, Y.H., 2021. Deep convolutional neural network-based pixel-wise landslide inventory mapping. Landslides 18 (4), 1421–1443. https://doi.org/10.1007/s10346-020-01557-6.
Su, C., Wang, B., Lv, Y., Zhang, M., Peng, D., Bate, B., Zhang, S., 2022. Improved landslide susceptibility mapping using unsupervised and supervised collaborative machine learning models. Georisk 1-19. https://doi.org/10.1080/17499518.2022.2088802.
Tang, H.M., Wasowski, J., Juang, C.H., 2019. Geohazards in the three Gorges Reservoir Area, China–Lessons learned from decades of research. Eng. Geol. 261, 105267 https://doi.org/10.1016/j.enggeo.2019.105267.
Tonini, M., Pecoraro, G., Romailler, K., Calvello, M., 2022. Spatio-temporal cluster analysis of recent Italian landslides. Georisk 16 (3), 536–554. https://doi.org/10.1080/17499518.2020.1861634.
Wang, H.J., Xiao, T., Li, X.Y., Zhang, L.L., Zhang, L.M., 2019a. A novel physically-based model for updating landslide susceptibility. Eng. Geol. 251, 71–80. https://doi.org/10.1016/j.enggeo.2019.02.004.
Wang, Y., Qin, Z., Liu, X., Li, L., 2019b. Probabilistic analysis of post-failure behavior of soil slopes using random smoothed particle hydrodynamics. Eng. Geol. 261, 105266 https://doi.org/10.1016/j.enggeo.2019.105266.
Wang, H.J., Zhang, L.M., Yin, K.S., Luo, H.Y., Li, J.H., 2021a. Landslide identification using machine learning. Geosci. Front. 12 (1), 351–364. https://doi.org/10.1016/j.gsf.2020.02.012.
Wang, H.J., Zhang, L.M., Luo, H.Y., He, J., Cheung, R.W.M., 2021b. AI-powered landslide susceptibility assessment in Hong Kong. Eng. Geol. 288, 106103 https://doi.org/10.1016/j.enggeo.2021.106103.
Wu, J., Li, Y., Zhang, S., Oualembo Mountou, J.C.J., 2022. Early identification of potential loess landslide using convolutional neural networks with skip connection: a case study in northwest Lvliang City, Shanxi Province, China. Georisk 1-13. https://doi.org/10.1080/17499518.2022.2088803.
Xiao, T., Li, D.Q., Cao, Z.J., Au, S.K., Phoon, K.K., 2016. Three-dimensional slope reliability and risk assessment using auxiliary random finite element method. Comput. Geotech. 79, 146–158. https://doi.org/10.1016/j.compgeo.2016.05.024.
Xiao, T., Zhang, L.M., Cheung, R.W.M., Lacasse, S., 2022. Predicting spatio-temporal man-made slope failures induced by rainfall in Hong Kong using machine learning techniques. Géotechnique 1-17. https://doi.org/10.1680/jgeot.21.00160.
Yang, B., Yin, K., Lacasse, S., Liu, Z., 2019. Time series analysis and long short-term memory neural network to predict landslide displacement. Landslides 16 (4), 677–694. https://doi.org/10.1007/s10346-018-01127-x.
Zhan, W., Fan, X., Huang, R., Pei, X., Xu, Q., Li, W., 2017. Empirical prediction for travel distance of channelized rock avalanches in the Wenchuan earthquake area. Nat. Hazards Earth Syst. Sci. 17 (6), 833–844. https://doi.org/10.5194/nhess-17-833-2017.
Zhang, S., Zhang, L.M., Peng, M., Zhang, L.L., Zhao, H.F., Chen, H.X., 2012. Assessment of risks of loose landslide deposits formed by the 2008 Wenchuan earthquake. Nat. Hazards Earth Syst. Sci. 12 (5), 1381–1392. https://doi.org/10.5194/nhess-12-1381-2012.
Zhao, T., Lei, J., Xu, L., 2022. An efficient Bayesian method for estimating runout distance of region-specific landslides using sparse data. Georisk 16 (1), 140–153. https://doi.org/10.1080/17499518.2021.1952613.
Zhou, S.Y., Gao, L., Zhang, L.M., 2019. Predicting debris-flow clusters under extreme rainstorms: a case study on Hong Kong Island. Bull. Eng. Geol. Environ. 78 (8), 5775–5794. https://doi.org/10.1007/s10064-019-01504-3.