Overcoming Data Scarcity Related Issues for Landslide Susceptibility Modeling with Machine Learning

Anika Braun, Katrin Dohmen, Hans-Balder Havenith, and Tomas Fernandez-Steeger

Abstract

Landslide susceptibility maps can be a useful tool to support holistic urban planning in mountainous environments. Data-driven methods for landslide susceptibility modeling work well even in data scarce areas, and there is an increasing relevance of machine learning methods that help analyze efficiently large and complex datasets. In this contribution we present some of our study examples to show how data quality, quantity, complexity, and preparation can have major effects on the outcomes of landslide susceptibility modeling. The aforementioned aspects are too often neglected in spite of their relevance, both in data scarce, but also data rich areas. We also use these examples to discuss the way we evaluate landslide susceptibility models, as the spatial performance of landslide susceptibility maps often differs from the mathematical performance. We finally discuss the necessity of standards for input data, modeling results and result communication to improve the usability of landslide susceptibility models in urban planning.

Keywords

Landslide susceptibility • Data quality • Machine learning

Introduction

With current trends of increasing intensity of climatic events and rapid urbanization in many areas of the world, it is becoming more important to systematically consider geological hazards in urban and land-use planning and build more resilient cities (UNISDR 2015). Landslide susceptibility mapping, which is defined as the analysis of the spatial probability of landslide occurrence, is a valuable tool for implementing slope stability in urban and land-use planning in inhabited mountainous areas (Fell et al. 2008). Data-driven, or statistically based methods work very well in data scarce areas, because although they require certain amounts of data, the required data can be relatively simple and is easily accessible as compared for instance to physically based methods. Usually, a landslide inventory, morphological and hydrological information derived from a digital elevation model (DEM), and some geological information are the minimum of data used in landslide susceptibility modeling (Reichenbach et al. 2018). Increasing availability of satellite data and methods for automatic landslide detection, as well as increasing availability of elevation data, make it possible to generate landslide susceptibility models even in data scarce areas.

The goal of data-driven or statistically based landslide susceptibility modeling is to quantify the spatial relationship between occurred landslides and related factors and to ultimately identify locations likely to be affected by future landslides by using only information about the factors. This can be achieved with simple bivariate statistical or multivariate methods, while in the last decade machine learning methods have increasingly proven to be useful for this task, as they can be employed to yield high accuracies while being able to handle large datasets. They are particularly relevant considering the increasing collection and accessibility of data. One main idea of data-driven methods is to overcome subjectivity based on decisions of the operator (van Westen et al. 2006). Landslide susceptibility can furthermore help to

A. Braun (✉) · K. Dohmen · T. Fernandez-Steeger
Engineering Geology Department, Institute of Applied Geosciences, Technische Universität Berlin, Ernst-Reuter-Platz 1, 10587 Berlin, Germany
e-mail: anika.braun@tu-berlin.de

H.-B. Havenith
Department of Geology, Georisks and Environment, University of Liège, Liège, Belgium

© Springer Nature Switzerland AG 2021

F. Guzzetti et al. (eds.), *Understanding and Reducing Landslide Disaster Risk*,
ICL Contribution to Landslide Disaster Risk Reduction,
https://doi.org/10.1007/978-3-030-60227-7_26

develop process understanding for landslides in a regional or even wider scale.

There is a huge amount of publications on the topic of statistically based landslide susceptibility modeling, as Reichenbach et al. (2018) demonstrated in their extensive review of 565 articles published between 1983 and 2016. As many of these authors showed, the methods for data-driven landslide susceptibility modeling work very well. They yield impressive performances measured by skill scores such as the area under the receiver operator characteristics (AUROC) curve. However, apart from tweaking models to reach better performances, two important topics are rarely ever discussed: data quality and usefulness of results.

As the name already suggests, in data-driven modeling we are replacing the expert reasoning with data, not algorithms. No matter how complex or advanced an algorithm is, it can only discover effects that are in the data. This point is often neglected in the literature, and until now, there exists no standard or unified approach for assessing the required quality, consistency, quantity, and preparation of input data for landslide susceptibility modeling.

The other point is the usefulness of resulting models. Evaluation criteria for landslide susceptibility models should depend on the goals of a study (Hearn and Hart 2019; Teimouri and Kornejady 2019). This point also implies the issue of the spatial plausibility of the resulting model and a reality check by local experts or through field verification, which could even help to develop new process understanding. To sum it up, we need better, standardized tools to define goals and assess the usefulness of landslide susceptibility models.

In this contribution we want to discuss the aspects data quality, consistency, quantity, and pre-processing of landslide inventories and input factor datasets, respectively, as well as result usefulness in landslide susceptibility modeling and their implications for data scarce areas using some examples from our research with machine learning tools on different scales.

Data and Methodology

Study Areas

In the following sections we want to use examples from three of our studies related to landslide susceptibility mapping with machine learning methods to discuss the above-mentioned aspects. The studies differ in the size of the study area and their environment, as well as in their specific goals and challenges. Two rather small study areas are the urban area of Tegucigalpa, capital city of Honduras (353 km^2, see Table 1), and the rural area of Ningnan County in southwestern China (726 km^2). The third study with a very large extent covers a good part of the Kyrgyz and Tajik Tien Shan Mountains (115,000 km^2).

The goal of the Tegucigalpa study was to generate a susceptibility map for urban planning with the challenge related to the availability of input data. The goal of the Ningnan study was to model landslide susceptibility based on parameters expressing rock quality and weathering, with a challenge regarding the landslide inventory quality. The goal of the Tien Shan study was to try to grasp the big picture, while the challenge was to handle a large dataset with uncertainties regarding the consistency of data for a large area covering two different countries.

Input Data

For all studies referred to in this contribution the freely available ALOS 30 m World DEM has been used (JAXA 2015–2019) as the basis to derive primary and secondary terrain and hydrological parameters with ArcGIS and SAGA GIS. In terms of data quality of the DEM all our studies are comparable.

The landslide inventory of the Ningnan study was mapped during field campaigns in 2015 and 2017. It will be discussed in more detail in the following section. The Tegucigalpa landslide inventory was generated in 2013 in collaboration with the Japan International Cooperation Agency (JICA) and the National Autonomous University of Honduras (UNAH), based on stereoscopic aerial image interpretation and field surveys (Braun et al. 2019). The Tien Shan landslide inventory has been created by Havenith et al. (2015) using Google Earth imagery.

For Ningnan a geological map was available that was reclassified regarding the geotechnical properties of the lithologies. For Tegucigalpa, different geological maps were available. For the Tien Shan only a classification into soft or hard rock was available ready-to-use.

All data was prepared on a pixel basis matching the ALOS 30 m DEM that was resampled to 30 m cells in UTM projection.

Modeling

For all three study areas landslide susceptibility was analyzed with the IBM SPSS Modeler. First, the data was transferred from a GIS into a table format and imported to the modeller. In a first step, the datasets were explored regarding their completeness and quality, distributions and inter-correlations of variables. In order to optimize the mathematical representation of the data, the variables were

Table 1 Key figures about the study areas presented in the examples in this contribution

Location	Area (km^2)	Landslides (%)	Variables (n)
Tegucigalpa	353	7	19
Ningnan	726	15	8
Tien Shan	115,000	0.8	25

transformed, scaled and recoded if necessary. More details about this workflow can be found in Braun et al. (2019).

In all studies, different classifiers, mainly Artificial Neural Networks (ANN) and Decision Trees (DT), were explored for their capabilities to model landslide occurrence in the different study areas using separate training, test, and sometimes validation subsets of the data using different compositions of the parameter sets. The models are described in Braun et al. (2019). The modeling results were evaluated with different skill scores, such as the total percentage of correct classifications, the percentage of correctly classified landslides (hit rate), false positives (false alarms), and false negatives (misses). Moreover, the results, namely the raw propensities (confidence that a cell is a landslide) and the binary classification result were plotted back in a GIS to evaluate the spatial quality, plausibility, and usefulness of the resulting models.

Landslide Inventories

Inventories of past landslides are the most crucial input for landslide susceptibility modeling, and also a great source of uncertainty and bias. Until recently, landslide inventories have mainly been created by experts, e.g. in field reconnaissance and through aerial or satellite image interpretation. Depending on experience, method, data quality, goal, scale, landslide type and landslide discretization (e.g. point or polygon) the results can vary greatly, e.g. in terms of accuracy and completeness, making it difficult to compare studies.

However, with the increasing availability of satellite data, techniques for automatic landslide detection, such as InSAR (Schlögel et al. 2015) or optical object recognition (Behling et al. 2014), are advancing. Another way to make landslide inventories more comparable is the use of slope units to discretize landslide occurrence (Alvioli et al. 2016; Schlögel et al. 2018).

In this section, we want to show how we used slope units to fix a landslide inventory with high uncertainties, and we want to discuss the problem of underrepresentation of landslides in datasets and how it can be solved using sampling or balancing techniques.

Improving a High Uncertainty Landslide Inventory with Slope Units

The landslide inventory available for the Ningnan study had spatial uncertainties. Landslides were mapped during an extensive field campaign, but mainly based on communications from local villagers. The inventory was updated in another field campaign and using Google Earth, but the spatial discretization of most landslides remained difficult. The reports of the villagers were however believed to be reliable, and thus a slope unit approach was implemented to spatially discretize slopes where failures have occurred in the past.

A slope unit is a region of space delimited between ridges and valleys under the constraint of homogenous slope aspect distribution (Carrara et al. 1991), which corresponds to either the left or right side of a sub-basin of any order into which a watershed is subdivided. Slope units were delineated in the study area by further subdividing watersheds computed in ArcGIS according to the main slope aspect and three geotechnical classes. Each pixel in a slope unit containing a landslide location was classified as a landslide pixel. The landslide susceptibility analysis was then carried out on a pixel basis. The advantage was that by assigning an event to an entire slope unit, a landslide to non-landslide pixel ratio of 15/85 could be achieved, which made it possible to model without any further balancing or sampling steps (see explanation below).

Underrepresentation of Landslide Cases

A very common problem in modeling with real-world datasets is the underrepresentation of the target class, making it hard for machine learning algorithms to capture it. In general, ratios of the target class of 20/80 to 30/70 are considered ideal (Pyle 1999).

There are different techniques to enhance the representation of the target class, such as balancing and under-sampling. In balancing techniques, duplicates of target datasets are generated for the training of the models. In under-sampling, a fraction of the non-target class is sampled to reach the desired distribution. Balancing is useful in small study areas to maintain a maximal sample size. For large

areas, under-sampling is ideal, as it helps reduce the computational effort by decreasing the total sample size.

When working with these techniques, care has to be taken to maintain the original distributions of the variables within the dataset. Also, with a higher representation of the target class in the training data, models might tend to overestimate its occurrence and produce increased numbers of false alarms.

We have worked with balancing in the Tegucigalpa study (Braun et al. 2019) and with under-sampling in the Tien Shan study (Dohmen 2019). Exploring different ratios of landslides to non-landslides revealed that a 30/70 ratio was ideal to increase the hit rates of models but still keep the false alarms reasonably low. Figure 1 shows for the Tien Shan and the Tegucigalpa studies how the hit rate (hr) of ANN increased significantly with an increasing ratio of landslides. At the same time, the number of false positives (fp) increased, which can be useful to a certain degree depending on the goal of the study (Teimouri and Kornejady 2019). For DT on the other hand we found they responded less to balancing, it even promoted over fitting (Braun et al. 2019).

Input Factor Sets—Sometimes More Is More

When it comes to the composition of the input dataset, we have made the experience that more is sometimes really more. In science we like to follow the principle of parsimony and strive for the simplest solution, which means in landslide susceptibility modeling to create a model with a minimum of influence factors. This makes a lot of sense for more basic bivariate statistical methods for landslide susceptibility mapping, as it also helps improve the interpretability by experts and end-users. In data mining however, the main idea is to use as much data as possible to maximize the chances that patterns can be discovered. Some algorithms, such as ANN, depending on their complexity actually require a certain number of input variables and samples, although there is no general rule as to how much data is enough.

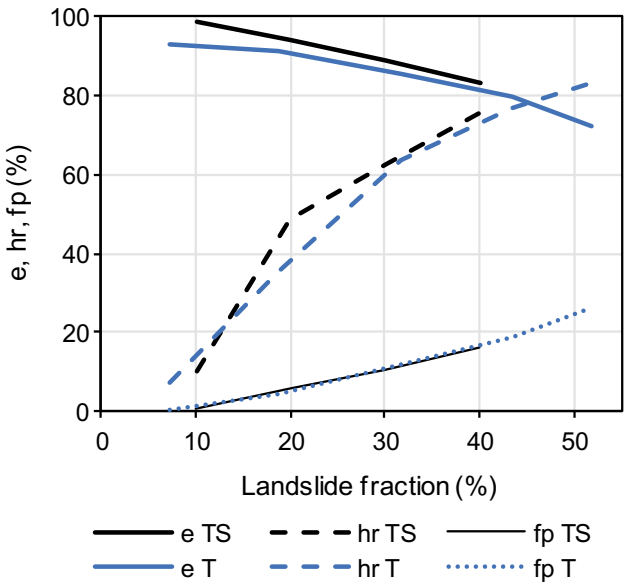

Fig. 1 Effect of balancing. Hit rates (hr), efficiency (e), and false positives (fp) for test run of ANN models trained with different landslide to non-landslide ratios for the Tien Shan (TS, black) and Tegucigalpa (T, blue)

Maximizing Information

In the Tien Shan study, we explored the response of ANN and DT models to the complexity of the set of input variables in a very large dataset. We used different combinations of input parameter sets, with a very basic set of parameters that is usually used in bivariate statistical modeling (elevation, slope, aspect, landforms, distance to rivers and faults, lithology), a complex set of 25 parameters (morphological, hydrological, climatic, geological, seismic, and anthropogenic factors), and a set of the 10 most important parameters as identified in the latter model. Then, to examine the effect of the continuous vs. nominal nature of the data, sets with only nominally coded data and with only continuous data were tested, respectively.

Interestingly, as it can be seen in Fig. 2, while the performance in terms of hit rate (hr) of the ANN increased with increasing complexity of the input parameter set, the DT already reached their optimal performance with 10 parameters. Two interesting aspects could be concluded from this. For the ANN more data really meant a better performance, even though most of the added data was derived from the same DEM. Moreover, it could be concluded that the DT model is an interesting choice when a simpler model is anticipated. Then again, the DT models produced significant artefacts wherever a variable was split into different branches of the DT. In the spatial validation, it also turned out that the ANN containing both, continuous and nominally discretized parameters, had a tendency to form spatial artefacts as well, while the artefacts did not occur when only continuous parameters were considered (Fig. 4). This underlines how skill scores alone do not enable a meaningful evaluation of landslide susceptibility models.

The Role of Lithological Information—Can Less Be Ok?

One parameter that is not always available in some areas, or not in a consistent form, but which is used in most

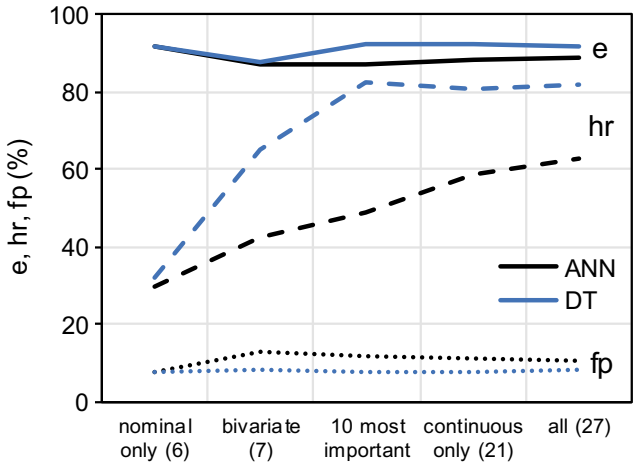

Fig. 2 Effect of dataset complexity. Hit rates (hr), efficiency (e), and false positives (fp) for validation run of ANN and DT models trained with different input parameter sets in the Tien Shan

publications (Reichenbach et al. 2018), is lithology. For Tegucigalpa, different geological maps were available, but they were not consistent. With some "expert interpolation" we tried to fix the geological information into a consistent map. In order to find out the effect of this highly uncertain information we generated ANN and DT with and without lithology as an input. We found that while the ANN performed a little poorer in terms of hit rate without the information about the lithology, the DT actually performed better without the lithology information (Fig. 3).

One explanation is that the nominal nature of the lithology variable is unfavourable for the DT because the dataset can only be split between classes, giving only few options

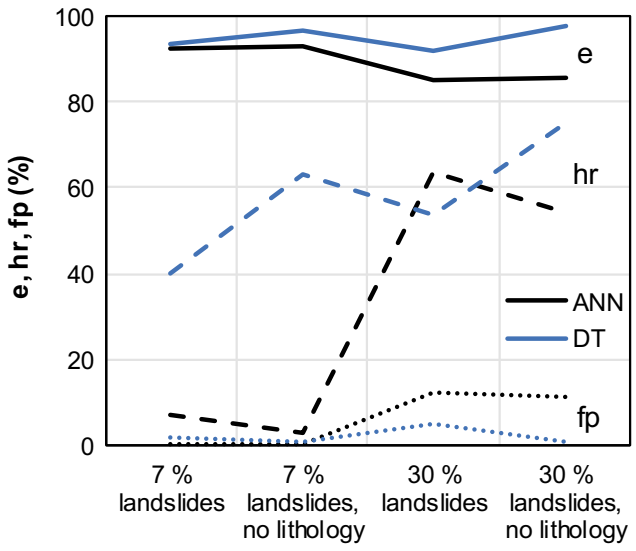

Fig. 3 Hit rates (hr), efficiency (e), and false positives (fp) for validation run of ANN and DT models trained with different input parameter sets in Tegucigalpa

for the tree to split with this variable. It should be noted that in the Tegucigalpa study ANN and DT developed quite different skills. While the DT developed very high accuracies with a tendency to over-fitting, the ANN produced more false positives, which formed however consistent patterns that seemed to be more useful for zoning. In the spatial context, the ANN trained with lithological information showed a more concise and coherent pattern than the one trained without it. Thus, in this case, in spite of being noisy, information on lithology could improve the model.

In the large study area of the Tien Shan on the contrary, as already explained above, the ANN had a better, artefact-free, spatial performance when no nominal parameters, such as the lithology class, were included (Fig. 4). However, the spatial evaluation of the entire, very large area showed that their accuracy differs strongly in different areas, which might be related to local effects. The analysis of this large area showed in the end that some landslides can be explained with such a general model, while others cannot. It will be interesting to explore this in future research to distinguish local from global effects.

Conclusions

In this contribution we presented three studies of landslide susceptibility mapping with machine learning methods to discuss the importance of different aspects regarding data quality, quantity, complexity, and preparation for data-driven landslide susceptibility modeling. We showed how we used slope units to fix a messy landslide inventory. We also showed how balancing, dataset complexity, and scale can have significant effects on both, the mathematical and spatial performance of different types of machine learning models of landslide susceptibility. The message of this is how important the data itself is, but also that there are techniques to overcome data quality related issues.

Perspectives

One point we want to discuss is the usefulness of landslide susceptibility models. There is a clear lack of studies regarding the implementation of landslide susceptibility maps in urban planning (Hearn and Heart 2019). In spite of the great potential that has been demonstrated in hundreds of publications, it remains difficult to implement machine learning methods for landslide susceptibility assessment into planning due to the lack of transparency for the end-user. Decision makers are most likely not impressed by high AUROC numbers that they are unable to interpret. In order to build confidence in landslide susceptibility maps as a tool for

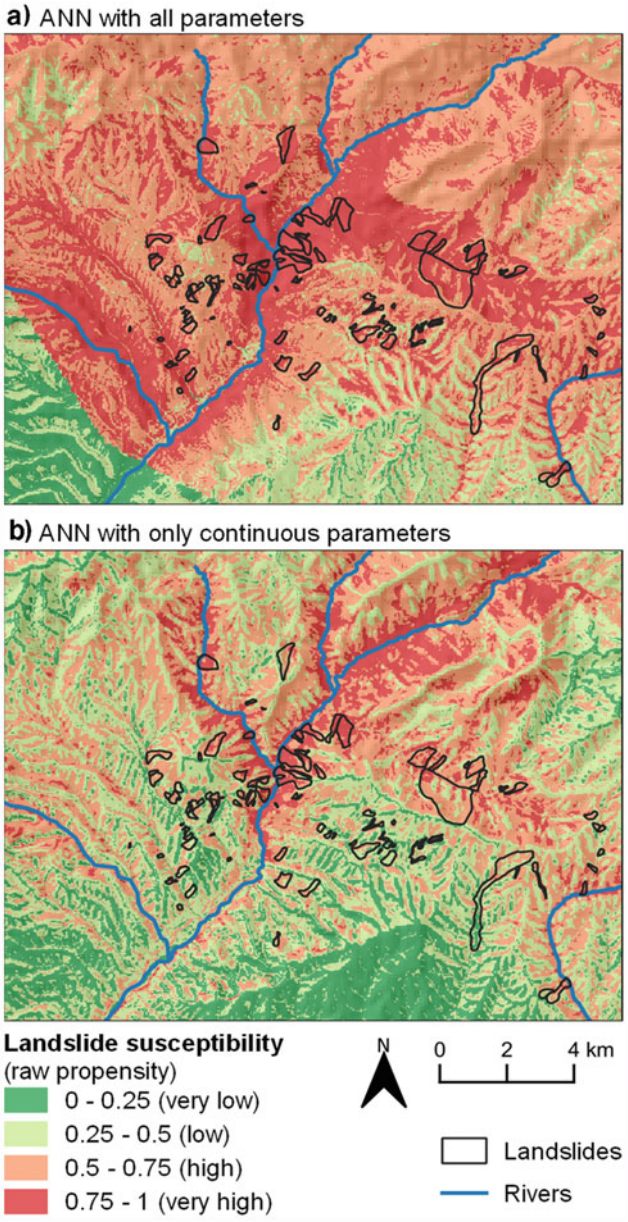

Fig. 4 Landslide susceptibility maps for the Maily Say area in the Tien Shan study

holistic urban planning, we need to define universal standards for input data, methodology, and evaluation criteria.

We also need to discuss more the way we communicate landslide susceptibility to make it more accessible for target audiences. The final medium of communication is the resulting map, which should look plausible and reliable to the end-user. Maps like the ones in Fig. 4 may be accurate but seem rather difficult to communicate. Slope units could be a tool that should be explored more for risk communication, as has been done in Italy for decades. It is easier to communicate that a whole slope is susceptible, rather than explain to the citizen why his house is on a red pixel, while his neighbours house is on a green pixel.

With the increasing challenges we are facing with the ongoing climate change and increasing urbanization we need to work on the way we do landslide susceptibility modeling to create a more useful tool for planning resilient cities. The data and methods are there.

Acknowledgements Many people contributed to the presented case studies, namely Luqing Zhang, Xueliang Wang, Zhenhua Han, and Jian Zhou, Elias Leonardo Garcia Urquia, Rigoberto Moncada Lopez, and Hiromitsu Yamagishi. Part of this work was funded by the Natural Science Foundation of China (Grant No. 41402285) and the Chinese Academy of Sciences President's International Fellowship Initiative (Grant No. 2016PZ032).

References

Alvioli M, Marchesini I, Reichenbach P, Rossi M, Ardizzone F, Fiorucci F, Guzzetti F (2016) Automatic delineation of geomorphological slope units with r.slopeunits v1.0 and their optimization for landslide susceptibility modeling. Geoscientific Model Dev 9 (11):3975–3991

Behling R, Roessner S, Kaufmann H, Kleinschmit B (2014) Automated spatiotemporal landslide mapping over large areas using rapideye time series data. Remote Sens 6(9):8026–8055

Braun A, Garcia Urquia EL, Lopez RM, Yamagishi H (2019) Landslide susceptibility mapping in Tegucigalpa, Honduras, using data mining methods. In IAEG/AEG Annual Meeting Proceedings, San Francisco, California, vol 1, pp 207–215 (2018)

Carrara A, Cardinali M, Detti R, Guzzetti F, Pasqui V, Reichenbach P (1991) GIS techniques and statistical models in evaluating landslide hazard. Earth Surf Proc Land 16(5):427–445

Dohmen K (2019) Landslide factors and susceptibility analysis using data mining methods for large study areas: a case study from the Tien Shan Mountains, Central Asia. Master thesis, Technische Universität Berlin, Berlin, Germany

Fell R, Corominas J, Bonnard C, Cascini L, Leroi E, Savage W (2008) Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. Eng Geol 102(3–4):99–111

Havenith HB, Torgoev A, Schlögel R, Braun A, Torgoev I, Ischuk A (2015) Tien Shan geohazards database: landslide susceptibility analysis. Geomorphology 249:32–43

Hearn GJ, Hart AB (2019) Landslide susceptibility mapping: a practitioner's view. Bull Eng Geol Env 78(8):5811–5826

JAXA (Japan Aerospace Exploration Agency) ALOS Global Digital Surface Model 'ALOS World 3D—30 m' (AW3D30 DSM Ver.1.0, 2.0 and 2.1), data available from the JAXA web interface (2016). http://www.eorc.jaxa.jp/ALOS/en/aw3d30/data/index.htm

Pyle D (1999) Data Preparation for Data Mining. Morgan Kaufmann, San Francisco. https://books.google.de/books?hl=de&lr=&id=hhdVr9F-JfAC&oi=fnd&pg=PA6&dq=Data+Preparation+for+Data+Mining&ots=6h9RbOGw5w&sig=elC6MNiVplWM14O5nx1ajw6drj0#v=onepage&q&f=false

Reichenbach P, Rossi M, Malamud B, Mihir M, Guzzetti F (2018) A review of statistically-based landslide susceptibility models. Earth Sci Rev 180:60–91

Schlögel R, Marchesini I, Alvioli M, Reichenbach P, Rossi M, Malet J (2018) Optimizing landslide susceptibility zonation: effects of DEM spatial resolution and slope unit delineation on logistic regression models. Geomorphology 301:10–20

Schlögel R, Doubre C, Malet J, Masson F (2015) Landslide deformation monitoring with ALOS/PALSAR imagery: a D-InSAR geomorphological interpretation method. Geomorphology 231: 314–330

Teimouri M, Kornejady A (2019) The dilemma of determining the superiority of data mining models: optimal sampling balance and end users' perspectives matter. Bull Eng Geol Environ

UNISDR (United Nations International Strategy for Disaster Reduction) (2015) Sendai framework for disaster risk reduction 2015–2030. http://www.wcdrr.org/uploads/Sendai_Framework_for_Disaster_Risk_Reduction_2015-2030.pdf. Accessed 28 Jan 2020

Van Westen CJ, Van Asch TW, Soeters R (2006) Landslide hazard and risk zonation—why is it still so difficult? Bull Eng Geol Env 65 (2):167–184