Disaster Risk Reduction
Methods, Approaches and Practices

Raju Sarkar
Sunil Saha
Basanta Raj Adhikari
Rajib Shaw *Editors*

Geomorphic Risk Reduction Using Geospatial Methods and Tools

Springer

Disaster Risk Reduction

Methods, Approaches and Practices

Series Editor

Rajib Shaw, Keio University, Shonan Fujisawa Campus, Fujisawa, Japan

Disaster risk reduction is a process that leads to the safety of communities and nations. After the 2005 World Conference on Disaster Reduction, held in Kobe, Japan, the Hyogo Framework for Action (HFA) was adopted as a framework for risk reduction. The academic research and higher education in disaster risk reduction has made, and continues to make, a gradual shift from pure basic research to applied, implementation-oriented research. More emphasis is being given to multi-stakeholder collaboration and multi-disciplinary research. Emerging university networks in Asia, Europe, Africa, and the Americas have urged process-oriented research in the disaster risk reduction field. With this in mind, this new series will promote the output of action research on disaster risk reduction, which will be useful for a wide range of stakeholders including academicians, professionals, practitioners, and students and researchers in related fields. The series will focus on emerging needs in the risk reduction field, starting from climate change adaptation, urban ecosystem, coastal risk reduction, education for sustainable development, community-based practices, risk communication, and human security, among other areas. Through academic review, this series will encourage young researchers and practitioners to analyze field practices and link them to theory and policies with logic, data, and evidence. In this way, the series will emphasize evidence-based risk reduction methods, approaches, and practices.

Raju Sarkar · Sunil Saha · Basanta Raj Adhikari ·
Rajib Shaw
Editors

Geomorphic Risk Reduction Using Geospatial Methods and Tools

Springer

Editors
Raju Sarkar
Department of Civil Engineering
Delhi Technological University
New Delhi, India

Sunil Saha
Department of Geography
University of Gour Banga
Malda, West Bengal, India

Basanta Raj Adhikari
Department of Civil Engineering
Institute of Engineering
Tribhuvan University
Lalitpur, Nepal

Rajib Shaw
Graduate School of Media and Governance
Keio University
Fujisawa, Kanagawa, Japan

ISSN 2196-4106 ISSN 2196-4114 (electronic)
Disaster Risk Reduction
ISBN 978-981-99-7706-2 ISBN 978-981-99-7707-9 (eBook)
https://doi.org/10.1007/978-981-99-7707-9

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The increase of natural phenomena due to the advancement of human infrastructures is inevitable nowadays. The technology and methods to be used for geomorphic hazards risk reduction are absolutely dependent on the governance and cultural context. Mainly in the mountain, coastal, arid and semi-arid regions these natural hazards are the main barriers to the economic development. However, human pressure and specific human actions (such as deforestation, inappropriate land use, and farming techniques) have increased the danger of natural disasters and degraded the natural environment. This makes it more difficult for environmental planners and policymakers to develop appropriate long-term sustainability plans. Since previous several decades, geospatial technology has undergone dramatic advances, opening up new opportunities for handling environmental challenges in a more comprehensive manner.

The book is organized into three parts comprising a total of 16 chapters, each of which will emphasize the use of advanced geospatial techniques in geomorphic hazards modelling and risk reduction. This book also compares the accuracy of traditional statistical methods and advanced machine learning methods. This book also addresses the different ways of reducing the impact of geomorphic hazards.

Each chapter is written by scholars and/or practitioners with acknowledged expertise in the field and with adequate experience of working in the Asian region. The book is intended to cover all dimensions of geomorphic hazards and an interdisciplinary perspective is thus ensured.

The book is sub-divided into three parts and they are:

Part I: *Geomorphic Hazards and Machine Learning Techniques*: This part explores the application of different advanced machine learning techniques in modelling the geomorphic hazards like landslide, flood, soil erosion, river bank failure etc. This part also compares the result of advanced machine learning techniques with the traditional statistical methods. LULC changes have accelerated the intensity of different hazards. Future prediction of these geomorphic hazards also assessed in parity with LULC change in this part.

Part II: *Geomorphic Hazards and Multi-temporal Satellite Images*: Advancement in satellite systems has provided lots of advantages to planners and policymakers in detecting and modelling different hazards not only in accessible areas but also in inaccessible areas. This part assesses the application of multi-temporal high-resolution satellite images like Quickbird, Worldview 3, LiDAR, SPOT 5, Google Earth Engine etc. in the mapping of different geomorphic hazards.

Part III: *Geomorphic Hazards Risk Reduction and Management*: This part explores the techniques and methods that help in reducing the risk of geomorphic hazards. Resilience adopted by the inhabitants of the affected areas to different geomorphic hazards also discussed in this part.

New Delhi, India — Raju Sarkar
Malda, India — Sunil Saha
Lalitpur, Nepal — Basanta Raj Adhikari
Fujisawa, Japan — Rajib Shaw

About This Book

This book explores the use of advanced geospatial techniques in geomorphic hazard modelling and risk reduction. It also compares the accuracy of traditional statistical methods and advanced machine learning methods and addresses the different ways to reduce the impact of geomorphic hazards.

In recent years with the development of human infrastructures, geomorphic hazards are gradually increasing, which include landslides, flood and soil erosion, among others. They cause huge loss of human property and lives. Especially in mountainous, coastal, arid and semi-arid regions, these natural hazards are the main barriers to economic development. Furthermore, human pressure and specific human actions such as deforestation, inappropriate land use and farming have increased the danger of natural disasters and degraded the natural environment, making it more difficult for environmental planners and policymakers to develop appropriate long-term sustainability plans. The most challenging task is to develop a sophisticated approach for continuous inspection and resolution of environmental problems for researchers and scientists. However, in the past several decades, geospatial technology has undergone dramatic advances, opening up new opportunities for handling environmental challenges in a more comprehensive manner.

With the help of geographic information system (GIS) tools, high and moderate-resolution remote sensing information, such as visible imaging, synthetic aperture radar, global navigation satellite systems, light detection and ranging, Quickbird, Worldview 3, LiDAR, SPOT 5, Google Earth Engine and others deliver state-of-the-art investigations in the identification of multiple natural hazards. For a thorough examination, advanced computer approaches focusing on cutting-edge data processing, machine learning and deep learning may be employed. To detect and manage various geomorphic hazards and their impact, several models with a specific emphasis on natural resources and the environment may be created.

Contents

Editors and Contributors

About the Editors

Prof. Raju Sarkar is a professor in the Department of Civil Engineering and Coordinator of Centre of Excellence in Disaster Risk Reduction (CoEDRR), Delhi Technological University (DTU). Prior to returning to his parent organization, DTU, Dr. Sarkar was a professor in the Department of Civil Engineering and Architecture, College of Science and Technology, Royal University of Bhutan, Bhutan, under the Ministry of External Affairs, Government of India, on deputation to Bhutan. During his tenure in Bhutan, he established the Center for Disaster Risk Reduction and Community Development Studies and also worked as the team leader to start the new undergraduate programme of engineering geology at the Royal University of Bhutan. Currently, he is also the chair of the Commission on Education and Outreach and the co-chair of the Commission on Earthquake Hazard, Risk and Strong Ground Motion in the International Association of Seismology and Physics of the Earth's Interior (IASPEI). He has vast experience in working in the Hindu-Kush Himalayas region both at the government and community level. He has published many original research articles in peer-reviewed journals, books, book chapters and proceedings of international societies, and he serves as an editorial member of several journals. Professor Sarkar is collaborating in a number of research projects funded by the International Council for Science and the World Bank, among others. He has a keen interest in geotechnics for natural disaster mitigation, geohazards risk management, landslides, seismology, community resilience against cataclysmic events, vulnerability and risk assessment and disaster management education.

Dr. Sunil Saha is an assistant professor of geography at the University of Gour Banga, India, and is specialized in the field of environmental geography. His research interests include fluvial landforms, natural hazards (gully erosion, landslides and floods), surface and sub-surface hydrological conditions and their impacts on land use in regard to both time and space. He is also interested in the applications of remote sensing and GIS in geohazard modelling and management. He has eight years of teaching and research experience. His name has been listed in the top 2% of world scientists (2021, 2022 & 2023) published by Stanford University, USA. He has guided many Ph.D. Scholars in the field of geo-environmental hazard management. More than 36 students have completed their dissertations under his guidance at the post-graduate level. He has published 70 peer-reviewed research papers in international journals of Elsevier, Springer, Taylor and Francis and the Multidisciplinary Digital Publishing Institute (MDPI). He has also published five book chapters and has presented his research at some 12 national and international conferences. He is a reviewer of research articles assigned by leading international journals such as the *Journal of Cleaner Production*, *Science of the Total Environment*, *Environmental Development and Sustainability* and *Geocarto International*, among others.

Dr. Basanta Raj Adhikari is a Deputy Director at the Centre for Disaster Studies, Institute of Engineering, Tribhuvan University, Nepal, and is also an assistant professor of engineering geology at the same university. Moreover, he is also a high-level foreign associate professor at Sichuan University, China. Dr. Adhikari has a research interest in the tectonics of the Himalayas, climate change, hill-slope movement and human interaction, Himalayan sediment flux generation and disaster risk reduction. He is the author of more than 40 scientific articles and book chapters and has received various expressions of recognition for his work in the field of earth science including Young Scientist (Integrated Research on Disaster Risk), "young affiliates" (The World Academy of Sciences) and Sichuan 1000 Talents (Sichuan Province, China). Currently, he serves as an editor of the *Journal of Coastal and Riverine Flood Risk* and the thematic article collections of *Environmental Change Driven by Climate Change, Tectonism and Landslide.*

Rajib Shaw is a professor at the Graduate School of Media and Governance of Keio University, Japan. He is also a distinguished professor at the Institute of Disaster Management and Reconstruction (IDMR) of Sichuan University and Indian Institute of Technology (IIT) Guwahati, Indian Institute of Science (IISc), Indian Institute of Management Kozikode and Indian Institute of Technology Hyderabad. He did his studies at Yokohama National University, Osaka City University in Japan, University of Allahabad and Burdwan University of India. He is a co-founder of a Delhi-based social entrepreneur startup, the Resilience Innovation Knowledge Academy (RIKA), and the chair of the board of two Japanese non-government agencies: SEEDS Asia and CWS Japan. He is the co-chair of the Asia-Pacific Scientific and Technical Advisory Group (AP-STAG). He is also the coordinating lead author (CLA) for the Asia chapter's 6th Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). Professor Shaw has published 70 books and over 480 research papers in the field of environment, disaster management and climate change. Professor Shaw is the recipient of "Pravasi Bharatiya Samman Award (PBSA)" in 2021 for his contribution to the education sector. PBSA is the highest honour conferred on overseas Indian and person of Indian origin by the President of India. He also received the United Nations Sasakawa Award in 2022 for his lifetime contributions to the field of disaster risk reduction.

Contributors

Abhishek Agarwal Department of Geography, University of Gour Banga, Malda, West Bengal, India

Nasibul Alam Department of Environmental Science, The University of Burdwan, Burdwan, West Bengal, India

Dhruv Bhardwaj Department of Civil Engineering, Delhi Technological University, New Delhi, India

Nathaniel Chaney Department of Civil and Environmental Engineering, Duke University, Durham, USA

Ankit Chaudhary Department of Civil Engineering, Delhi Technological University, New Delhi, India

Baofeng Di Institute for Disaster Management and Reconstruction, Sichuan University-Hong Kong, Polytechnic University, Chengdu, Sichuan, China; Department of Environmental Science and Engineering, Sichuan University, Chengdu, Sichuan, China

Leki Dorji Civil Engineering Department, College of Science and Technology, Royal University of Bhutan, Thimphu, Bhutan

Amiya Gayen Department of Geography, University of Calcutta, Kolkata, India; Department of Geography, Midnapore College (Autonomous), Midnapore, India

Priyanka Gogoi Department of Geography, University of Gour Banga, Malda, West Bengal, India

Ashok Kumar Gupta Department of Civil Engineering, Delhi Technological University, New Delhi, India

Srimanta Gupta Department of Environmental Science, The University of Burdwan, Burdwan, West Bengal, India

Sk. Mafizul Haque Department of Geography, University of Calcutta, Kolkata, India

Wen He Institute for Disaster Management and Reconstruction, Sichuan University-Hong Kong, Polytechnic University, Chengdu, Sichuan, China

Harjeet Kaur National Institute of Disaster Management, Ministry of Home Affairs, Government of India, New Delhi, India

Suktara Khatun Department of Geography, University of Gour Banga, Malda, West Bengal, India

Ankit Kumar Department of Civil Engineering, Delhi Technological University, New Delhi, India

Rachael Lau Department of Civil and Environmental Engineering, Duke University, Durham, USA

Jierui Li Institute for Disaster Management and Reconstruction, Sichuan University-Hong Kong, Polytechnic University, Chengdu, Sichuan, China; Department of Land Surveying and Geo-Informatics and Research Institute for Land and Space, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

Kaustuv Mukherjee Department of Geography, Chandidas Mahavidyalaya, Birbhum, West Bengal, India

Sujata Pal Department of Geography, University of Gour Banga, Malda, West Bengal, India

Gopal Chandra Paul Department of Geography, University of Gour Banga, Malda, West Bengal, India

Aman Pawar Department of Civil Engineering, Delhi Technological University, New Delhi, India

Lingke Qiu Department of Environmental Science and Engineering, Sichuan University, Chengdu, Sichuan, China

Nerusupalli Dinesh Kumar Reddy Department of Civil Engineering, Delhi Technological University, New Delhi, India

Bishnu Roy Department of Geography, University of Gour Banga, Malda, West Bengal, India

Biswajit Roy Department of Geography, University of Gour Banga, Malda, West Bengal, India

Anik Saha Department of Geography, University of Gour Banga, Malda, West Bengal, India

Sunil Saha Department of Geography, University of Gour Banga, Malda, West Bengal, India

Anil Kumar Sahu Department of Civil Engineering, Delhi Technological University, New Delhi, India

Raju Sarkar Department of Civil Engineering, Delhi Technological University, New Delhi, India

Carolina Seguí Department of Civil and Environmental Engineering, Duke University, Durham, USA

Baboo Chooreshwarsingh Sujeewon Department of Civil Engineering, Delhi Technological University, New Delhi, India

Raju Thapa National Institute of Disaster Management, Ministry of Home Affairs, Government of India, New Delhi, India

Manolis Veveakis Department of Civil and Environmental Engineering, Duke University, Durham, USA

Tyler Waterman Department of Civil and Environmental Engineering, Duke University, Durham, USA

Wen Zeng Institute for Disaster Management and Reconstruction, Sichuan University-Hong Kong, Polytechnic University, Chengdu, Sichuan, China

Part I
Geomorphic Hazards and Machine Learning Techniques

Chapter 1
Landslide Susceptibility Assessment Based on Machine Learning Techniques

Jierui Li, Wen He, Lingke Qiu, Wen Zeng, and Baofeng Di

Abstract In this chapter, we will introduce the landslide susceptibility assessment (LSA) methods based on machine learning techniques. The economic loss or even casualties caused by landslides indicate the significance of LSA. LSA can be regarded as either regression or classification problems, which can be processed by machine learning techniques. LSA provides administrators or researchers with information on potential disaster areas, which can be an efficient way to relieve the pressure of disaster reduction and mitigation. Several landslide inventories and disaster-related geo-environmental variable datasets were recommended. A total of 9 machine learning methods applied in LSA were simply introduced. The advantages and future work of LSA based on machine learning techniques were summarized from the aspects of scale, performance, modeling, and interpretability.

Keywords Landslide Susceptibility model · Machine learning techniques · Landslide conditioning factors · Validation · ROC curve

---

J. Li (✉) · W. He · W. Zeng · B. Di (✉)
Institute for Disaster Management and Reconstruction, Sichuan University-Hong Kong, Polytechnic University, Chengdu 610200, Sichuan, China
e-mail: falsejieruili.cd@gmail.com

B. Di
e-mail: dibaofeng@scu.edu.cn

J. Li
Department of Land Surveying and Geo-Informatics and Research Institute for Land and Space, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China

L. Qiu · B. Di
Department of Environmental Science and Engineering, Sichuan University, Chengdu 610065, Sichuan, China

R. Sarkar et al. (eds.), *Geomorphic Risk Reduction Using Geospatial Methods and Tools*, Disaster Risk Reduction, https://doi.org/10.1007/978-981-99-7707-9_1

1.1 Introduction

1.1.1 Landslides

Landslides are worldwide geo-hazards distributed on almost all continents (Fig. 1.1), causing economic loss or even casualties. Based on the Emergency Events Database (EM-DAT) (Guha-Sapir 2019), there were 789 large-scale landslides from 1910 to March 2022 on a global scale. These fatal landslides have affected 14.7 million people, causing 4.3 million people homeless and 67.2 thousand deaths. Each landslide event leads to an average of 226.2 thousand U.S. dollars in economic loss.

Totally 17 Sustainable Development Goals (SDG) with 169 targets and 247 indicators related to peace and prosperity of people and the planet were proposed in 2015 by the United Nations General Assembly (Zeng et al. 2020). SDGs are supposed to be achieved by 2030, during which SDG 13 focuses on climate change and one of its targets is to strengthen the climate-related disaster early warning (Campbell et al. 2018). Since many studies have revealed that the occurrence of landslides is linked to climate change (Peres and Cancelliere 2018; Scheidl et al. 2020), the prevention and mitigation of landslides will be a topic for achieving SDGs.

Landslide susceptibility assessment (LSA) is a commonly applied tool for predicting the occurrence probability of landslides (Huang et al. 2018). The LSA zonation maps provide government officials, decision-makers, and the public with information on areas under different levels of landslide susceptibility, facilitating the establishment of a local landslide early warning system for landslide prevention and mitigation (Guzzetti et al. 2020). LSA is thus considered a crucial step for effectively reducing the casualties and economic loss due to landslides.

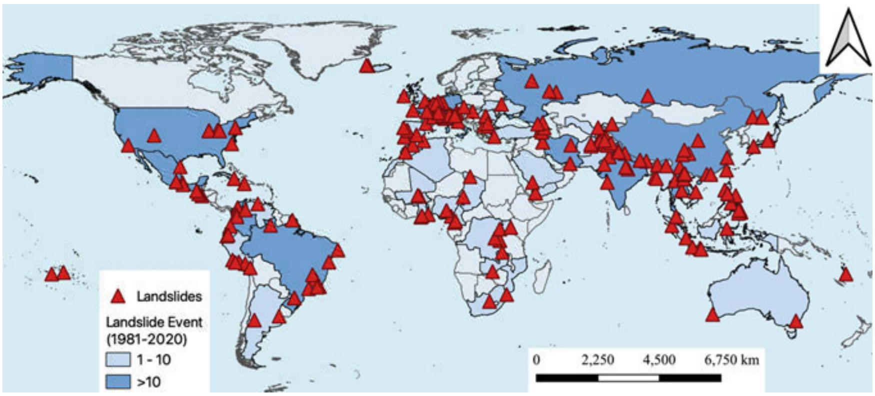

Fig. 1.1 Large-scale landslide spatio-temporal distribution around the world

1.1.2 Development of Landslide Susceptibility Assessment

Before the 1970s, scientists focused on the mechanism and process prediction of geological disasters. Since then, the loss caused by geological disasters continues to increase, which has seriously threatened the safety of human life and property. In this regard, disaster prevention and management began to be paid attention by scientists all over the world (Cui et al. 2011; Reichenbach et al. 2018). At the same time, a large amount of literature on landslide susceptibility began to be published. These studies used different mathematical methods to study the relationship between spatial distribution characteristics of landslides and assessment factors of landslide susceptibility in different assessment units (grid cells, slope units, or watersheds) (Guzzetti 2006). The first geological hazard assessment map in the world, namely landslide assessment map, was published in the mid-1970s (Neuland 1976). Neuland presented 250 stable and unstable slopes in the study area by means of cartographic synthesis through field investigation and experience, in which unstable regions refer to areas where landslides have occurred or are likely to occur. In addition, binary discriminant analysis was used to construct a specific stable/unstable landslide prediction model by selecting slope morphology, geomechanics, lithology, and structural characteristics (Neuland 1976).

The landslide assessment map should provide information on the probability and scale of landslide occurrence (Varnes 1984). However, early landslide assessment maps did not predict the probability of landslide occurrence. With the development of discriminant analysis (Carrara et al. 1991; Guzzetti 2006; Neely and Rice 1990), linear regression (Carrara 1983; Carrara and Guzzetti 1995), logistic regression (Ayalew and Yamagishi 2005; Lee 2005), and other conventional mathematical and statistical models, researchers began to attach importance to the application of mathematical models in disaster risk assessment. Meanwhile, Italian scholars developed a geographic information database to store landslide and geological environment information, and predicted landslide hazard of two watersheds in southern Italy by multiple regression analysis. The "landslide hazard" in this study is the early "landslide susceptibility," which promoted the prediction of regional landslides on a large spatial scale. Subsequently, according to the current situation of landslide distribution, the geological experts in the United States adopted topography, lithology, and other factors related to landslides, and divided the landslide risk into five grades (Wieczork 1984).

In the 1990s, with the development of geographic information technology, researchers analyzed various influencing factors of landslides through the Geographic Information System (GIS) platform, combined with statistical models to study landslide susceptibility (Ohlmacher and Davis 2003; Van Westen et al. 2008). Landslide susceptibility began to enter the quantitative assessment stage, which is an important milestone (Ghosh and Bhattacharya 2010; Gorsevski and Jankowski 2010). Subsequently, some scholars proposed the concept of landslide susceptibility in succession and considered the influence of driving factors on potential landslides in combination with historical landslide laws (Aleotti and Chowdhury 1999; Martin et al. 2002).

American scholars considered the bedrock and surface geology, tectonic geological conditions, climate, geomorphic units, land use, hydrological conditions, and other related factors to evaluate landslide susceptibility and made a zoning map for landslide risk assessment based on GIS (Mejia-Navarro and Wohl 1994). Korean scholars used GIS technology to extract 14 factors such as slope, aspect, slope angle, surface curvature, and lithology and analyzed the landslide susceptibility using the probability analysis and logistic regression algorithm (Lee and Min 2001).

In the past 20 years, with the development of computer science, machine learning technology has provided a new research method for landslide susceptibility. Its powerful nonlinear prediction ability improves the accuracy of regional LSA. The commonly used models include decision tree (Saito et al. 2009; Wu et al. 2014), Naïve Bayes (Tien Bui et al. 2012), random forest (Catani et al. 2013; Trigila et al. 2015), gradient boosting machine (Di et al. 2019), support vector machine (Peng et al. 2014; Yao et al. 2008), artificial neural network (Aditian et al. 2018; Ermini et al. 2005), convolutional neural network (Liu et al. 2022; Sameen et al. 2020), and so on. For example, the neural network model and logistic regression model were used to predict LSA, respectively, and the results showed that the evaluation results of the neural network model were more consistent with the actual situation (Yesilnacar and Topal 2005). Simultaneously, since there are no historical landslide samples in some areas, the transferability of LSA mapping is gradually developed (Sun et al. 2020; Zhou et al. 2021).

To sum up, there are many methods to evaluate landslide susceptibility. We need to know that there will be differences in LSA by selecting different models in the same area (Rossi et al. 2010). The uncertainty brought by these differences will affect the evaluation of assessment results. Therefore, appropriate susceptibility assessment methods should be carefully selected in the LSA (Huabin et al. 2005). At present, researchers prefer to use multiple assessment methods to calculate LSA and combine different assessment results into a comprehensive LSA map (Rossi et al. 2010; Tien Bui et al. 2016; Xiong et al. 2020). This will reduce the uncertainty brought by different methods and improve the reliability of assessment results. Of course, a disaster assessment system and method with universal principles are also worth developing in the future.

1.1.3 Machine Learning for Dealing with Regression and Classification Problems

A regression model or a classification for LSA can be trained and validated based on a large number of landslide records along with the geo-environmental factors. Landslide susceptibility can be described by either probability (value) or category (level), and thus both regression and classification methods are available for LSA (Sahin 2020; Wang et al. 2020).

Regression is a statistical method applied to simulate the relationships between a dependent variable and independent variables. For LSA, landslide susceptibility is considered the dependent variable, while the geo-environmental factors related to landslides are regarded as the independent variables (Mandal and Mandal 2018). The built regression model can be applied to predict the probability of landslide susceptibility in an area where the geo-environmental factors are known.

Classification is a method applied to identify the category of an item according to its properties (Shahabi et al. 2020). The landslide susceptibility level can be defined into different categories, and the geo-environmental factors related to landslides are the properties. The constructed classification model can determine the landslide susceptibility level based on the geo-environmental factors in the area. Furthermore, for landslide susceptibility zonation mapping, the regression results generally will be classified based on standards (natural breaks, quantiles, etc.) or classification algorithms (Guo et al. 2021).

1.1.4 Supervised and Unsupervised Learning

Machine learning can be generally divided into three types: supervised learning, unsupervised learning, and reinforcement learning, during which supervised learning and unsupervised learning are generally applied in LSA (Chang et al. 2020; Mandal and Mondal 2019).

Supervised learning requires the preset of labels for all items (Merghadi et al. 2020). In LSA, the dependent variable (landslide occurrence) should be labeled based on a certain assessment unit. All geo-environmental factors will then be extracted based on the assessment unit (Kalantar et al. 2020). The values of geo-environmental variables will be mapped into the labels through the training process. The output of a supervised learning model will contain the landslide susceptibility value or susceptibility level.

Unsupervised learning can be conducted without labeling the items with the prior knowledge (Pokharel et al. 2021). In other words, it will be more suitable for the situation when the landslide inventory is not sufficient. However, the good performance of an unsupervised learning LSA model is decided by the characterizations contained in the geo-environmental databases (Liang et al. 2021). Thus, the unsupervised learning for LSA requires more comprehensive variables based on the formation and initiation mechanisms of landslides. The output of an unsupervised learning model will divide the assessment region into different categories. The landslide susceptibility of each category can be further validated based on the existing landslide records.

However, since there is no prior information and training model, the prediction accuracy of unsupervised learning is generally not very high (Hakim et al. 2022). With the prior information on supervised learning, a more accurate LSA can be conducted (Tien Bui et al. 2016). A "semi-supervised" method was thus proposed for LSA (Huang et al. 2020). This method overcomes the shortcomings of the traditional supervised and unsupervised machine learning models in landslide susceptibility prediction.

1.2 Landslide Inventory and Disaster-Related Geo-Environmental Variable Dataset

This section provides shared online resources of landslide inventory and geo-environmental variable database, which can all be required in constructing the machine learning models for regional LSA. For geo-environmental datasets, referring to the most commonly applied variables based on the formation and initiation of landslides (Cui et al. 2014; Di et al. 2019; Gao and Sang 2017), we provided the online database for terrain, meteorological, soil, land use/cover, and geological data. We only displayed the global dataset, although there are still many comprehensive datasets in certain countries or regions. The resolution of the geo-environmental variable database directly determines the precision of LSA.

1.2.1 Landslide Inventory

1.2.1.1 Global Landslide Catalog

Global Landslide Catalog (GLC) was developed by the National Aeronautics and Space Administration (NASA) to identify landslides around the world (https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog/h9d8-neg4). GLC covers 6,788 landslide events from 2003 to 2015 all over the world, with detailed information including the triggers (precipitation, earthquake, or human activities) and subtype (debris flows, mudflow, creep, rockfall, snow avalanche, lahar, and so on) of the landslides. Fatalities and injuries are also contained in the dataset.

1.2.1.2 Emergency Events Database

Emergency Events Database (EM-DAT) was a comprehensive database including natural, technological, and complex disasters (www.emdat.be). The landslide was a natural disaster classified in the hydrological subtype. The EM-DAT covers the landslide records all over the world from 1900 up to now. Users can download the dataset from different continents or countries and during any period separately. The subtype of landslide records in the EM-DAT contains avalanche, mudslide, rockfall, and subsidence. Fatalities and injuries are all recorded as well as the total economic loss.

1.2.2 Geo-Environmental Database

We first revealed the contribution of the geo-environmental variables based on the formation and initiation mechanisms of landslides and then listed some free online resources for downloading the geo-environmental data including terrain, meteorological, soil, land use/land cover, and geological data.

1.2.2.1 Terrain Data

The terrain directly related to the energy of landslides is a crucial factor in disaster occurrences. Landslide is a kind of mass movement consisting of various sediments due to gravity. The area with a large elevation difference has high potential energy, while the area with a steep slope has the condition of energy conversion (Huang et al. 2012; Luo et al. 2019). Therefore, the topographic factors such as elevation difference and slope jointly determine whether landslides will occur in an area.

Space Shuttle Radar Topography Mission (SRTM) conducted by NASA in the year 2000 offers the digital terrain model (DTM) and digital elevation model (DEM) with 30-m resolution (https://earthexplorer.usgs.gov). The Global ALOS world 3D digital surface model (DSM) provided by the Japan Aerospace Exploration Agency (JAXA) (https://www.eorc.jaxa.jp/ALOS/en/index_e.htm) also has a resolution of 30 m. Compared to DTM or DEM, DSM includes the height information of buildings and trees on the surface.

1.2.2.2 Meteorological Data

Meteorological factors can form and initiate a landslide from different mechanisms. Water is considered the most common trigger of landslides. Precipitation, glacial meltwater, glacial lake outburst, and floods highly related to meteorological conditions can all initiate a landslide (Borga et al. 2014; Guo et al. 2020). In addition, a prolonged drought period results in the shrinkage of soil and the soil particles expand rapidly under heavy rainfall, increasing pore pressure and decreasing slope stability (Chen et al. 2014). Climatic variables related to dry–wet status thus contribute much to the landslide formation and initiation.

World Meteorological Organization (WMO) has developed a climate explorer tool for investigating and deriving meteorological data (http://climexp.knmi.nl/selectfield_obs2.cgi). The database covers the global scale from the earliest 1850 up to now, including temperature, precipitation, solar radiation, cloud cover, wind, drought index, potential evaporation, and so on. These datasets include both grid data with a minimum resolution of 0.25 degrees and station recorded data.

1.2.2.3 Soil Data

Soil provides the essential materials for landslides and soil properties are highly related to slope stability. Soil type, soil erodibility, soil liquidity index, and soil moisture are generally concerned factors reflecting the slope stability (Nandi and Shakoor 2010; Scheidl et al. 2020; Sun et al. 2020).

International Soil Reference and Information Centre (ISRIC) has been developing Soil and Terrain (SOTER) (https://www.isric.org/explore/soter) databases at a scale of 1:1 million with global coverage and the databases in most countries or regions have been developed. All properties in SOTER are compiled in GIS polygon format. Each vector layer includes items of soil types and physical/chemical properties of soil.

1.2.2.4 Land Use and Land Cover Data

Land use and land cover affect the initiation of slope failures (Roccati et al. 2021). Landslides are strongly influenced by land degradation along with land use and land cover change in mountainous regions (Galve et al. 2015). A higher vegetation coverage can promote soil strength (Li et al. 2021), and generally, land use and land cover datasets can also reflect the vegetation coverage.

European Space Agency (ESA) has provided a global land cover product for the year 2020 at a 10-m resolution (https://viewer.esa-worldcover.org/worldcover/). The land cover was divided into 10 categories including tree cover, shrubland, grassland, cropland, built-up, bare/sparse vegetation, snow/ice, permanent water bodies, herbaceous wetland, mangroves, and moss/lichen. Esri provides land use and land cover time series from 2017 to 2021 (https://www.arcgis.com/home/item.html). The product classifies the land use into water, trees, flooded vegetation, crops, built areas, bare ground, snow/ice, clouds, and rangeland.

1.2.2.5 Geological Data

Earthquake is another trigger of landslides. The size of landslides enlarges with the increase of ground motion (Valagussa et al. 2019). The distances to seismic sources, including epicenter and faults, are generally considered in LSA (Valagussa et al. 2019; Xi et al. 2019). It has been revealed that the tectonic and lithologic characteristics are also important in landslide occurrences (Bahrami et al. 2020).

United States Geological Survey (USGS) offered World Geologic Maps in 2000 (https://certmapper.cr.usgs.gov/data/apps/world-maps/). The geologic database has not yet included all the countries in the world. Faults and geologic-type maps are accessible only to some countries.

1.3 Machine Learning Methods for Landslide Susceptibility Assessment

1.3.1 Linear Regression

Linear regression estimates the equation that best describes the association between a continuous response variable and a single or multiple predictor variable. Simple Linear Regression (SLR) estimates the linear correlation between a single predictor (x) and a single criterion variable (y), while Multiple Linear Regression (MLR) predicts the linear correlation between two or more predictors ($x_1, x_2, \ldots, x_i$) and one criterion variable (y). MLR is applied more frequently in linear regression analysis, and each value of the independent variable is affiliated with the value of a dependent variable. The least squares approach is commonly applied to fit a linear regression model (Khademi et al. 2016).

MLR estimates the level of correlation between a single response variable (dependent variable) from multiple predictors (independent variable). The normal form of an MLR model is as shown in Eq. (1.1) (Chou and Tsai 2012):

$$\hat{y} = a_0 + \sum_{i=1}^{m} a_i x_i \tag{1.1}$$

where $\hat{y}$ is the output, x_i is the independent input variable, and a_i is the partial regression coefficient.

To analyze the relationship between the landslides and possible causative factors in the Kankai watershed, Nepal, the bivariate frequency ratio technique and the MLR method were applied, while the outcomes of the MLR method were less consistent and reliable than those of the frequency ratio technique (Kayastha et al. 2013). MLR and random forests were applied to predict two phenomenological models, the groundwater level and the landslide velocity models, for the predicting the movement of the Kostanjek landslide, the largest landslide in the Republic of Croatia, and the results obtained from random forests are just slightly better than those from MLR, in both models, proofing that multiple linear regression has a possibility for predicting landslide movement (Krkac et al. 2020). MLR analysis was used to determine the most influential factors among eight landslide factors toward landslide vulnerability levels through remote sensing data in Purworejo, and the results showed that the most significant factors were elevations, with regression values that were quite dominant among other variables (Sudaryatno et al. 2020).

1.3.2 Logistic Regression

Logistic Regression employs the use of independent variables to create a mathematical formula that predicts the probability of occurrence of an event. The dependent variable is dichotomous, while the independent variables could be interval, dichotomous, or categorical (Atkinson and Massari 1998). The relationship between the dependent variable and independent variables is nonlinear.

Assumptions for Logistic Regression are as follows: the dependent variable must be categorical; the independent variables should not have multi-collinearity.

The logistic model can be expressed as (Devkota et al. 2013):

$$p = \frac{\exp(z)}{(1 + exp(z))} \tag{1.2}$$

where p is defined as the probability of an event occurrence, such as landslide, which varies from 0 to 1 on a curve-shaped S; Z is the following equation (linear logistic model), whose value varies from $-\infty$ to $+\infty$:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + +\beta_n X_n \tag{1.3}$$

where 0 represents the intercept of the model, 1, 2, …, n, are the partial regression coefficients, and X_1, X_2, …, X_n represent the independent variables (Fig. 1.2).

The Logistic Regression technique, widely used in LSA, has been confirmed as reliable and high performing in terms of prediction (Trigila et al. 2015). Logistic Regression, Multi-criteria Decision Analysis, and a Likelihood Ratio Model were applied for landslide susceptibility mapping inside and outside the city of Izmir, Turkey, during which Logistic Regression performed the best based on Area Under Curve (AUC) (Akgun 2012). Logistic Regression is often regarded as a benchmark in LSA (Di et al. 2019; Xiong et al. 2020). LSA on the Mugling–Narayanghat road with peripheral zone was done using bivariate (Certainty Factor and Index of Entropy) and multivariate (Logistic Regression) models, as a result, all the models applied showed

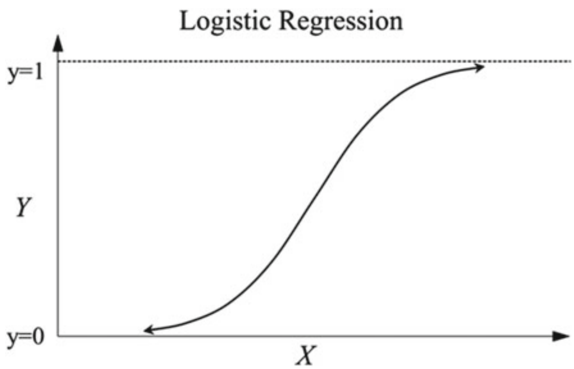

Fig. 1.2 Image of Logistic Regression. Logistic regression's is ranged between 0 and 1. Due to using a nonlinear log transformation to the odds ratio, logistic regression does not require a linear relationship between inputs and output variables

reasonably good accuracy (83.57%, 90.16%, and 86.29%, respectively) based on receiver operating characteristics (ROC) (Devkota et al. 2013).

1.3.3 Naïve Bayes

A Naïve Bayes classifier is built based on Bayes' theorem. The word "Naïve" indicates the conditional independence assumption between predictors (Soria et al. 2011). Naïve Bayes is easy to construct without complicated iterative parameter estimation schemes (Wu et al. 2008).

If we sort a test case *x*, the probability of each class given the vector of observed values for the predictive attributes may be obtained using Bayes' theorem:

$$p(C = c \mid X = x) = \frac{p(C = c)p(X = x \mid C = c)}{p(X = x)} \tag{1.4}$$

where $p(C = c \mid X = x)$ is known as posterior probability; $p(X = x \mid C = c)$ is known as a likelihood probability; $p(C = c)$ is termed as prior probability, and $p(X = x)$ is the probability of evidence.

Since the event integrates attribute value assignments, and due to the assumption of attribute conditional independence, the subsequent equation may be written:

$$p(X = x \mid C = c) = \prod_{i} p(X_i = x_i \mid C = c) \tag{1.5}$$

which is fairly simple to compute for training and testing data (John and Langley 1995).

In terms of LSA, given an occurrence that comprises k landslide-related variables, y_j is a Boolean output of the estimation of landslide or non-landslide areas. According to the subsequent equation, the prediction is made for the class with the biggest posterior probability (Tien Bui et al. 2012):

$$y_j = argmax\, P(y_j) \prod_{i=1}^{k} P(\frac{x_i}{y_j}) \tag{1.6}$$

where j indicates landslide or non-landslide.

It can be assumed that when models with low complexity are initialized and data sets with low size are given, the Naive Bayes classifier could be considered a reliable tool for LSA. Naïve Bayes outperformed Logistic Regression in LSA based on 116 sites located in the mountainous regions of Epirus, Greece (Tsangaratos and Ilia 2016). Using an ensemble classifier framework could improve the performance of Naïve Bayes classifier (Pham et al. 2017a). A novel ensemble classifier model combining Naïve Bayes classifier and Rotation Forest ensemble for LSA at the Luc

Yen district, Viet Nam, was generated, which outstripped other landslide models, e.g. AdaBoost, Bagging, MultiBoost, and Random Forest (Pham et al. 2017b).

1.3.4 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is one of the simplest machine learning algorithms based on the supervised learning technique (Al-Hadidi et al. 2016). KNN algorithm was developed for discriminant analysis for a situation where reliable parametric estimation of probability densities was unknown or difficult to determine (Kramer 2013). The dataset is stored in KNN models with categories for classifying new input data based on similarity (Taunk et al. 2019). KNN is applied more for classification problems (Zhang et al. 2017).

The value of K defines the area size (neighborhoods) when the KNN algorithm searches. (Bottou and Vapnik 1992). When K = 1, small neighborhoods appear throughout the large area, and the points from different categories are relatively scattered. When K = 20, the points with labels in the minority are ignored during generating neighborhoods, and only large categories are clustered together. Figure 1.3 is a schematic diagram of data classification when the KNN algorithm takes different k values (K = 1 or 20). When the KNN algorithm classifies two categories of points, the blue point category is represented by a blue background, and the red point category is represented by a white background. For example, in Fig. 1.3a, the blue domain category appears near some of the blue dots, but not in Fig. 1.3b.

With the development of the algorithm, the KNN model has overcome the limitations of time and memory, and has been practically applied in the field of big data (Maillo et al. 2017). KNN algorithm is often applied to LSA in combination with other machine learning algorithms, e.g., Decision Tree (DT), Random Forest (RF),

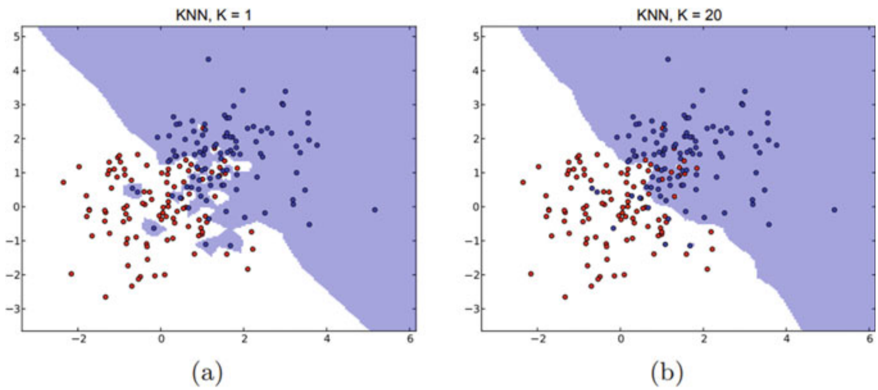

Fig. 1.3 Schematic diagram of algorithm classification of different neighborhoods (K value)

and Support Vector Machine (SVM) (Park and Lee 2021). KNN algorithm is robust to the noisy training data and can be more effective if the training data is large.

1.3.5 K-Means Clustering

K-means Clustering (KMC) is a common classical clustering analysis algorithm that solves specific problems through iteration. The main parameter of the algorithm is K, that is, there are some sample data, and we don't know what category each sample is, but we know how many categories all samples are or should be classified into. The calculation process is that, based on the parameter K, the inputted *n* data objects are divided into K categories (groups). Then clustering is performed centered on K points in the space, and the objects closest to these points are classified. In an iterative manner, the value of each cluster center is updated successively until the best clustering effect is obtained. In the iterative process, the similarity between objects in the same class is high, and the similarity between objects in different classes is low.

Figure 1.4 describes the basic flow of the KMC algorithm. The frame (a) represented the initial dataset. In frame (b), the centroids corresponding to the two values of k (suppose $k = 2$), the red centroid and the blue centroid, are randomly selected. Then calculate the distance from all other points to the two centroids, and select the closest centroid for classification. In frame (c), we classified the dataset after the first iteration of all sample points. New centroids for different class points are shown in frame (d), and the locations of new centroids have changed. The frames (e) and (f) repeated the process of the frames (c) and (d). The final two categories were shown in the frame (f).

Previous studies have shown that the landslide susceptibility mapping framework based on the KMC algorithm can give a more general characterization of landslide susceptibility and can provide effective solutions for landslide mitigation and management (Wang et al. 2017). Comparisons between the traditional equidistant classification method and the KMC algorithm in LSA showed that the overall performance of KMC algorithm was better (Tianlun et al. 2021).

1.3.6 Random Forest

Random Forest (RF) (Breiman 2001) is an ensemble learning method based on decision trees. Decision tree is a common machine learning algorithm (Quinlan 1986). The nodes of the decision tree class are divided into leaf nodes and branch nodes. The former represents a certain category after data classification, and the latter represents the test features of data division. The classification rules are the internal nodes, and the completed classification nodes are the leaf nodes. Node division and selection are based on criteria, such as information gain, gain rate, Gini coefficient,