LANDSLIDE SUSCEPTIBILITY ANALYSIS USING MACHINE LEARNING TECHNIQUES IN PENANG ISLAND, MALAYSIA

GAO HAN

UNIVERSITI SAINS MALAYSIA

2021

LANDSLIDE SUSCEPTIBILITY ANALYSIS USING MACHINE LEARNING TECHNIQUES IN PENANG ISLAND, MALAYSIA

by

GAO HAN

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

October 2021

ACKNOWLEDGEMENT

My work in research as a Ph.D. student at the School of Mathematical Sciences in USM was an amazing and beautiful experience. What I achieved here is, of course, *not only* the knowledge *but also* something intangible but valuable that words cannot describe fully.

I am very pleased that this thesis has become a reality with the dedicated support and help from my three supervisors. First and foremost, I would like to express my deepest gratitude to my main supervisor, **Dr. Fam Pei Shan**, for her patience, motivation, dedication and enthusiasm during my whole academic journey, and financial support during my first two academic years. Undoubtedly, she is a teacher as well as a friend to me. My sincerest thanks then go to my co-supervisor, **Professor Dr. Low Heng Chin**, who is a thoughtful and helpful lady. She always gives me confidence and courage to persevere in difficult times, no matter in life or in study. What I have learnt from her is research work is more than just pursuing the knowledge itself. I would also like to express my deepest thanks to my co-supervisor, **Dr. Tay Lea Tien**, who is a very kind and helpful lady as well. She provided me many technical support and valuable comments during my Ph.D. study, which improved my thesis a lot.

In addition, I would like to thank my parents, Mr. Gao and Mrs. Qiao, who always give me unconditional support and love during my life. Last but not least, I would like to express my special thanks to God for the spiritual support and love during my life in Malaysia.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

α	Learning rate
β	Coefficient parameter in SVM
C	Penalty parameter in SVM
d	Degree parameter in SVM
FR_i	Frequency ratio of the *i*th influencing factor
FR_{min}	Minimum value of the frequency ratio
FR_{max}	Maximum value of the frequency ratio
$\mathfrak{H}$	Feature space of SVM
i_k	Number of categories in logistic regression
P_{LO}	Percentage of landslide occurrence in each sub-category of a factor
P_{LF}	Percentage of each factor sub-category of a factor
$\sigma(\cdot)$	Logistic regression function
γ	Gamma parameter in SVM
X	A random variable
x_0	Original value of variable X
x_1	New value of variable X
$\mathbf{x}_{cat.}$	Vectors of categorical variables
$\mathbf{x}_{con.}$	Vectors of continuous variables
x_{min}	Minimum value of variable X

x_{max}	Maximum value of variable X
$\mathbf{x}_{old}$	Old variable of a landslide influencing factor
μ_{AND}	Fuzzy AND operator
μ_{GAMMA}	Fuzzy GAMMA operator
μ_i	Normalized frequency ratio of the *i*th influencing factor
μ_{OR}	Fuzzy OR operator
$\mu_{PRODUCT}$	Fuzzy PRODUCT operator
μ_{SUM}	Fuzzy SUM operator
ω_0	Coefficient of the constant in logistic regression
$\omega_{variable_{name}}$	Coefficients of the corresponding continuous variables in logistic regression

LIST OF ABBREVIATIONS

AHP	Analytic Hierarchy Process
AI	Artificial Intelligence
ANN	Artificial Neural Network
ADT	Alternating Decision Tree
AUC	Area Under Curve
CART	Classification and Regression Trees
CDA	Canonical Discriminant Analysis
CHAID	Chi-squared Automatic Interaction Detection
DA	Discriminant Analysis
DM	Data Mining
DS	Discriminant Score
DT	Decision Tree
EBF	Evidence Belief Function
EILs	Earthquakes Induced Landslides
ETC	Extra Tree Classifier
FA	Factor Analysis
FL	Fuzzy Logic
FR	Frequency Ratio
GIS	Geographic Information System

GDM	Gradient Descent with Momentum
JKR	Jabatan Kerja Raya
LightGBM	Light Gradient Boosting Model
LDA	Linear Discriminant Analysis
LIF	Landslide Influencing Factors
LN	Linear Function
LR	Logistic Regression
LSA	Landslide Susceptibility Analysis
LSI	Landslide Susceptibility Index
LSM	Landslide Susceptibility Map/Mapping
LSDM	Landslide Susceptibility Modelling
MLP	Multi-layer Perceptron
NASA	National Aeronautics and Space Administration
NBT	Naive-Bayes Tree
NML	Non-tree-based Machine Learning
OR	Odds Ratio
PCA	Principal Component Analysis
PL	Polynomial Function
RILs	Rainfall Induced Landslides
RBF	Radial Basis Function
RFC	Random Forest Classifier

ROTE	Random Oversampling Technique
RS	Remote Sensing
RUTE	Random Undersampling Technique
SCG	Scaled Conjugate Gradient
SCOTE	Self-Creating Oversampling Technique
SHALSTAB	Shallow Slope Stability
SIG	Sigmoid Function
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TGB	Tree-based Gradient Boosting
Tol	Tolerance value
VIF	Variance Inflation Factor
Wi	Statistical index
WoE	Weight of Evidence
WOM	Weighted Overlay Model
Wf	Weighting factor
XGBoost	eXtreme Gradient Boosting

LIST OF APPENDICES

ANALISIS KERENTANAN TANAH RUNTUH DENGAN MENGGUNAKAN TEKNIK-TEKNIK PEMBELAJARAN MESIN DI PULAU PINANG MALAYSIA

ABSTRAK

Tanah runtuh adalah bahaya semula jadi yang akan menyebabkan kehilangan nyawa dan harta benda yang besar. Analisis kerentanan tanah runtuh (LSA) sangat penting untuk pengurusan dan pengurangan tanah runtuh. Kajian ini bertujuan untuk meningkatkan prestasi ramalan reruang LSA dengan menggunakan teknik-teknik pembelajaran mesin. Oleh kerana sampel tanah runtuh menyumbang peratusan yang kecil dalam data mentah, pemilihan nisbah sampel yang optimum sebelum melatih model pembelajaran mesin dan peningkatkan sampel tanah runtuh dengan cara yang efisien adalah masalah-masalah utama penyelidikan ini. Di satu pihak, tiga jenis nisbah sampel dirancang untuk meningkatkan prestasi ramalan spasial melalui analisis perbandingan. Set data dengan nisbah sama didapati sebagai nisbah optimum dalam LSA. Selain itu, tiga teknik persampelan berlebihan, iaitu persampelan berlebihan rawak (ROTE), persampelan berlebihan minor sintetik (SMOTE) dan persampelan berlebihan pencipta diri (SCOTE), telah digunakan unutk menambah sampel tanah runtuh. Hasil yang setanding diperolehi dan ini menunjukkan kecekapan penambahan sampel tanah runtuh. Akhirnya, model peningkatan kecerunan gradien dibangunkan dengan pergabungan SMOTE dengan SCOTE di LSA. Kawasan di bawah lengkung (AUC) dianggap sebagai sukatan utama untuk menilai prestasi model-model. Hasilnya menunjukkan terdapat suatu peningkatan dalam prestasi dengan nilai AUC tertinggi 0.9525. Kesimpulannya, peta yang dihasilkan dalam kajian ini dapat memberikan maklumat berguna untuk pengurusan dan mitigasi tanah runtuh tempatan.

LANDSLIDE SUSCEPTIBILITY ANALYSIS USING MACHINE LEARNING TECHNIQUES IN PENANG ISLAND, MALAYSIA

ABSTRACT

Landslides are a natural hazard which cause great losses of lives and properties. Landslide susceptibility analysis (LSA) is of great importance for landslide management and mitigation. This study mainly aims to improve the spatial prediction performance of LSA using machine learning techniques. Since landslide samples account for a small percentage in the raw data, selecting an optimal sample ratio before training machine learning models and increasing the landslide samples in an efficient way are the main research problems. On the one hand, three types of sample ratios are designed to increase the spatial prediction performance through comparative analysis. The equal ratio for datasets is found as the optimal ratio in LSA. Additionally, three oversampling methods, random oversampling technique (ROTE), synthetic minority oversampling technique (SMOTE) and self-creating oversampling technique (SCOTE), are applied to augment the landslide samples. A comparable result is obtained which indicates the efficiency of the augmented landslide samples. Finally, gradient boosting models are developed to integrate with SMOTE and SCOTE in LSA. The area under the curve (AUC) values are considered as the key metric for evaluating the models' performance. The results show an enhancement in the performance with the highest AUC value of 0.9525. To summarise, the maps produced in this study can provide useful information for the local landslide management and mitigation.

CHAPTER 1

INTRODUCTION

1.1 General Introduction

The general idea of this thesis will be displayed in this chapter. It mainly focuses on landslide susceptibility analysis (LSA) in Penang Island, Malaysia. More emphasis will be put on the landslide susceptibility modelling (LSMD). Various models, including statistical models, machine learning models as well as gradient boosting models, are considered in the analysis. Studies based on different datasets, algorithms, and oversampling techniques are conducted in Penang Island, Malaysia. The research problems and research objectives are shown in Sections 1.4 and 1.5, respectively. The arrangement of the thesis is displayed in Section 1.6.

1.2 Research Background and Significance

Landslide is considered as the second natural hazard since it causes great losses of lives and properties, and damages to natural resources and environment around the globe every year (Hilker *et al.*, 2009; Kanungo *et al.*, 2009; Malamud and Turcotte, 2006; Gill and Malamud, 2016; Haque *et al.*, 2016; Ladds *et al.*, 2017; Scaringi *et al.*, 2018). Petley (2012) indicated that 2,620 fatal landslides in total were recorded worldwide between 2004 and 2010, causing a total of 32,322 recorded fatalities. Klose *et al.* (2015) estimated the transportation infrastructure losses in their research and found that the landslide loss for highways in the US totally amounted to USD23.5 million from 1980 to 2010.

Malaysia is a landslide-prone country. According to the data from the US National Aeronautics and Space Administration (NASA) (http://www.nasa.gov), Malaysia experienced 171 landslides from 2007 to 2016, which made the country rank

one of the top 10 countries in frequency of landslides. Moreover, the National Slope Master Plan 2009-2023 prepared by Jabatan Kerja Raya (JKR) Malaysia identifies Penang Island as a landslide prone area in 2009. The pictures of landslide occurrence in Penang Island are displayed in Plate 1.1.

As a landslide-prone area in Malaysia, Penang Island experiences various landslides every year, especially in the monsoon season from April to May and from October to November. For example, Penang Island has seen three deadly landslide incidents in less than two years - the Oct 21, 2017, Granito landslide; the Oct 19, 2018 Bukit Kukus landslide; and the June 28, 2019 Batu Ferringhi landslide. It is reported that a total of 24 lives have been lost as a result of the three landslide incidents (https://www.nst.com.my). Recently, a landslide was reported on Penang Hill on 7th October 2020 following continuous heavy rain by New Straits Times. Fortunately, no one was injured in the landslide incident as reported (https://www.nst.com.my).

Conducting landslide spatial research in Penang Island is of great importance and significance, which can help local authorities manage and mitigate the landslide risks to reduce or minimize the losses of lives and property. Furthermore, landslide is a global issue rather than a local problem. Therefore, the research outcomes can be generated to other landslide-prone area around the world to a high degree.

Figure 1.1 Landslide occurrences in Penang Island, Malaysia

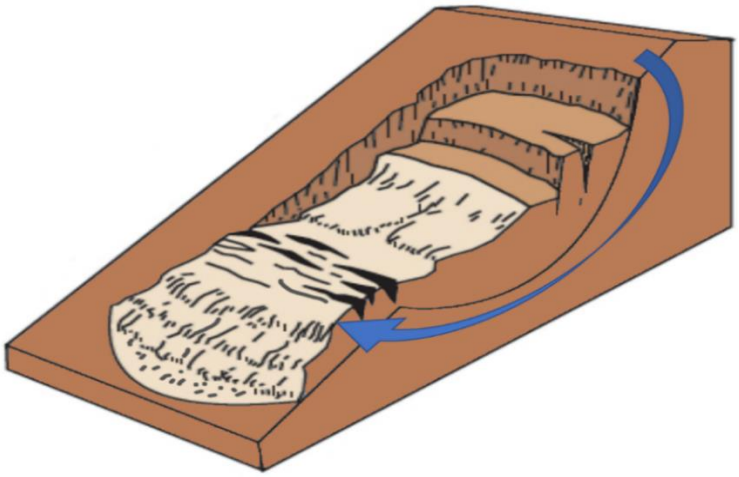
1.3 What is a Landslide?

A landslide is usually defined as the movement of a mass of rock, debris, or earth down a slope (Cruden, 1991). The term 'Landslide' usually encompasses five slope movements, namely, (1) falls, (2) topples, (3) slides, (4) spreads, and (5) flows (Varnes, 1978). Usually, slides can be classified into two categories: rotational and translational. Rotational slides commonly show slow movement along a curved rupture surface while translational slides usually denote rapid movements along a plane of distinct weakness between the overlying slide material and the more stable underlying material (Varnes, 1978). Figure 1.1 displayed the difference between rotational and translational landslides. The arrows denote the directions of the landslide motions.

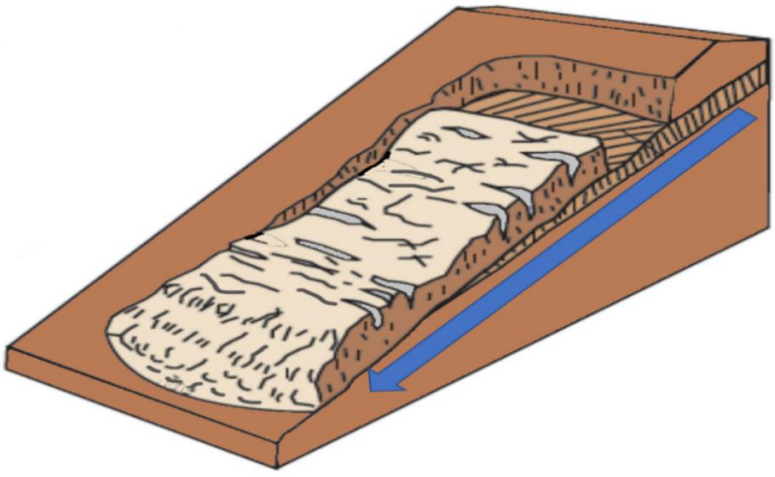
Moreover, a complex slope movement is defined as the combination of any two movements (Varnes, 1978). There are two main classes of landslide materials: rock and engineering soil which can be further divided into debris and earth (Varnes, 1978). According to the slope movement research by Varnes (1978), the combination of landslide types and materials is displayed in Table 1.1. Since the research resolves around the improvement of landslide spatial prediction, detailed information about the slope movement is not provided. Varnes (1978), Crozier (1986) and Wieczorek and Snyder (2009) described a comprehensive understanding in their studies.

Two types of factors, namely, triggering factors and influencing factors, are commonly considered to initiate landslides (Li *et al.*, 2012). According to the triggering factors, landslides can be broadly classified into rainfall induced landslides (RILs) and earthquakes induced landslides (EILs). The differences between RILs and EILs can be found in the studies by Chang *et al.* (2007) and Meunier *et al.* (2008).

Since the research area of this study, Penang Island, Malaysia, is located in a free-earthquake zone, more emphasis will be put on the RILs instead of EILs. Besides that, human activities, such as construction and extraction, are considered as a contributing factor to landslides as well (Haigh and Rawat, 2011; Skilodimou *et al.*, 2018). The introduction to landslide influencing factors (LIFs) will be discussed in Section 1.1.3 when dealing with LSA.



(a) Rotational landslide



(b) Translational landslide

Figure 1.2 Two types of landslides: (a) Rotational and (b) Translational

(**Source**: Arizona Geology e-Magazine, 2020)

Table 1.1 Classification of movements and material

Type of Movement		Type of Material		
		Bedrock	Engineering Soils	
			Predominantly coarse	Predominantly fine
Falls		Rock fall	Debris fall	Earth fall
Topples		Rock topples	Debris topple	Earth topples
Slides	Rotational	Rock slump	Debris slump	Earth slump
		Rock block slide	Debris block slide	Earth block slide
	Transitional	Rock slide	Debris slide	Earth slide
Lateral Spreads		Rock spread	Debris spread	Earth spread
Flows		Rock flow (deep creep)	Debris flow (slow creep)	Earth flow (slow creep)
Complex		Combination or two or more types of slope movement		

(**Source**: Varnes, 1978)

1.3.1 Landslide Susceptibility Analysis (LSA)

Landslide susceptibility refers to the likelihood of a landslide occurring in an area based on the local terrain and environmental conditions (Brabb, 1984; Reichenbach *et al.*, 2018). Susceptibility analysis aims to study where the landslides are likely to occur and what influencing factors are likely to cause them (Guzzetti *et al.*, 2005). In other words, it is a science of studying spatial landslide prediction rather than temporal prediction (Kanungo *et al.*, 2009).

There are four fundamental assumptions which are widely accepted in LSA when applying models for spatial prediction (Varnes, 1984; Hutchinson, 1995; Aleotti and Chowdhury, 1999; Reichenbach *et al.*, 2018):

1) The past and present landslides are considered keys to the future landslide occurrences which implies that the future landslide occurrences are more likely to occur under similar geological and geomorphological conditions.

2) Landslide occurrences are controlled by identifiable influencing factors which can be mapped from field surveys or remote sensing imagery.

3) Landslides would leave discernible signs which can be identified and classified through the analysis of remote sensing image interpretations.

4) The degree of landslide susceptibility can be measured, and different susceptibility classes can be zoned and mapped according to the degrees.

LSA is mainly based on the first assumption that landslides occur in the same or similar areas where they have taken place previously (Murillo and Alcántara, 2015). The whole research process in this thesis is developed based on this assumption.

Landslide influencing factors play an essential role in susceptibility analysis and mapping. Usually, a landslide is seldom attributed to a single influencing factor (Kanungo *et al.,* 2009). The most commonly considered influencing factors in LSA are included but not limited to geology, soil type, curvature, slope angle, precipitation, distance to road and distance to fault. The relationship between landslide occurrence and influencing factors in a specific area can be investigated using statistical and machine learning models, and the results may vary from model to model. However, there is no general agreement about which model possesses the best performance in LSA (Zhou *et al.,* 2018). That is why various studies are needed to be investigated in order to determine the optimal model for LSA in a specific area that is the main objective to be achieved in this research. More objectives of this research will be discussed in Section 1.3 in a comprehensive way after stating the research problems in Section 1.2.

1.3.2 Landslide Susceptibility Modelling (LSMD) and Mapping (LSM)

Landslide susceptibility modelling (LSMD) is usually regarded as the initial step to LSA (Fell *et al.,* 2008; Zhou *et al.,* 2018). The objective of LSMD is to identify the areas that are prone to landslide occurrences based on the knowledge of past landslide events as well as geological factors that are associated with landslide occurrence or non-occurrence (Brenning, 2005) using various kinds of models. The models used in LSA will be displayed in Chapter 2 in a thorough and detailed way.

Statistical models, including bivariate and multivariate models, and machine learning algorithms, including artificial neural network (ANN), support vector machine (SVM), and tree-based gradient boosting models, are widely used in LSA in recent decades (Reichenbach *et al.,* 2018). Machine learning is an evolving branch of

artificial intelligence (AI) and computer science that are designed to mimic human intelligence by learning from the surrounding environment (El Naqa & Murphy, 2015). Generally, the machine learning models are considered as more efficient with better prediction performance than other models (Goetz *et al.*, 2015; Pham *et al.*, 2016; Zhou *et al.*, 2018). The detailed literature review of the models used in LSA will be given in Chapter 2. The landslide susceptibility index (LSI) will be obtained during LSMD, which is further used to produce maps using ArcGIS.

Landslide susceptibility mapping (LSM) is an important tool for disaster management and planning development activities in landslide prone areas (Dahal *et al.*, 2008). It works on visualizing the LSA according to LSI obtained from LSMD by means of Geographic Information System (GIS) and remote-sensing data (Pradhan *et al.*, 2008). LSM is also called landslide susceptibility zonation or zonation mapping in previous literature (Clerici *et al.*, 2002; Chauhan *et al.*, 2010a; Shano *et al.*, 2020). Those terms are considered as identical in this research. More information about LSM will be discussed in Chapters 4-6.

1.3.3 Geographic Information System (GIS)

GIS stands for Geographic Information System. The definition of GIS varies from researcher to researcher since it is a very broad idea. Generally, a GIS can be defined as a special type of computer-based information system for capturing, storing, checking, and displaying data related to positions on Earth's surface (Worboys and Duckham, 2004).

GIS is everywhere when it comes to any information that includes location, of course, landslide spatial analysis included. Huabin *et al.* (2005) provided an overview to the landslide hazard assessment based on GIS techniques. During the process of

LSA in this research, various types of maps are needed, including landslide inventory map, landslide factor maps and landslide susceptibility maps as well. GIS technique involves all the map production using ArcGIS software. More details about the maps are discussed in Chapter 3.

1.4 Research Problem Statement

The fundamental research problem to be addressed in this research work is how to improve the landslide spatial prediction using machine learning models in Penang Island, Malaysia. In recent decades, machine learning models show powerful prediction abilities in various areas. For a specific prediction task, however, it is difficult to say one model has better performance than another without any verifications. Therefore, the central research problem is how to determine the optimal model for the research area, namely, Penang Island.

Moreover, there is a severe imbalanced sample ratio between landslide and non-landslide samples. In other words, the number of landslide samples is far less than the number of non-landslides, which is because Penang Island is a residential area, most of the area is free of landslides. Due to this reason, it may cause prediction inaccuracies when using the original dataset directly to train models. Thus, the second research question is how to determine an optimal sample ratio as well as sample size for landslide spatial prediction.

As mentioned before, the number of non-landslide samples is far larger than the number of landslide samples. However, the purpose of landslide spatial research is to predict the potential high-risk landslide area. In addition to determine the sample ratio in a proper way, how to increase the number of the landslide samples in an effective way is also considered as a research question to be addressed. Therefore, the

third research question is how to effectively augment landslide samples using oversampling techniques.

1.5 Research Objectives

According to the research problems discussed in Section 1.2, there are four objectives to be achieved in this research which are as follows:

1. To improve the landslide spatial prediction performance using machine learning models in Penang Island, Malaysia. Different types of models will be applied and discussed for this fundamental purpose and the optimal model will be developed finally.

2. To determine the optimal sample size and sample ratio for the imbalanced landslide dataset. For this purpose, experimental research in different dimensions is conducted to obtain the optimal model. Different data dimensions are considered in the research as well.

3. To develop a new oversampling method to augment landslide samples in an effective way. A comparative study will be conducted to compare the efficiency with existing oversampling methods.

4. To visualize the results in landslide susceptibility maps using ArcGIS software. The maps will be divided into different susceptibility zones according to the LSI obtained from LSMD, which can help local authorities manage and mitigate the landslide hazard.

The core of the four objectives is to improve the landslide spatial prediction and to provide guidance to local authorities to reduce the landslide risks. In Chapters 4-6, the

research storyline will unfold regarding to the four objectives through conducting various sets of studies. The results will be visualized in LSMs.

1.6 Scope of the Study

This research focuses on the landslide susceptibility analysis in Penang Island, Malaysia. The machine learning models are applied to improve the spatial prediction performance after using statistical models. Different sample ratios and sample sizes and oversampling techniques are combined with machine learning models to improve the spatial prediction performance. The time series analysis would not be considered in this research due to the real-time data unavailability.

1.7 The Arrangement of Thesis

There are a total of seven chapters in this thesis. Chapter 2 provides an overview of the literatures in landslide spatial research field, which covers from prediction models to evaluation methods. In Chapter 3, the research area, namely, Penang Island in Malaysia, and the data available for the research including landslide inventory and influencing factors are introduced to the readers in detail. Moreover, the preliminary analyses using traditional statistical models are considered in this chapter as well.

From Chapters 4 to 6, three studies are discussed to improve the prediction performance and produce LSMs, which is the main part of the thesis. The three studies share similarities in methodologies and writing structures but with different key points. Specifically, Chapter 4 discusses about the machine leaning models trained using the datasets with different sample sizes and sample ratios on LSA. Chapter 5 mainly focuses on the oversampling techniques. Chapter 6 provides the application of gradient

boosting models in Penang Island. Finally, Chapter 7 gives an overall conclusion as well as the limitations of the thesis.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The objective of this chapter is to gain an understanding of the existing research relevant to LSA and identify the research gap between previous research and what to be achieved in this thesis. Centering on this purpose, the literature with respect to different aspects of the landslide spatial research will be comprehensively reviewed. More emphasis will be put on the recent studies and significant studies as well.

In previous literature, confusion exists when it comes to 'landslide susceptibility' and 'landslide hazard'. As mentioned earlier, the definition of landslide susceptibility in this study is stated as the likelihood of a landslide occurring in an area based on the local terrain and environmental conditions (Brabb, 1984; Reichenbach *et al.*, 2018). It does not consider the size of the landslides (Carrara *et al.*, 1995; Reichenbach *et al.*, 2018).

Landslide hazard usually denotes the probability that a landslide of a given magnitude will take place in a given period and in a given area (Reichenbach *et al.*, 2018). Landslide hazard analysis has a strong demand to dynamic and real data, such as the dynamic rainfall data, to predict when a landslide may occur. In other words, landslide hazard is more difficult to ascertain than landslide susceptibility which can be considered as the spatial component of the hazard (Guzzetti, 2006a; Reichenbach *et al.*, 2018).

The term 'susceptibility' is used even when the original authors used 'hazard' but analyze the same thing with 'susceptibility' when discussing the previous literature in this chapter. It is consistent with the idea of the review paper by Reichenbach *et al.*

(2018) which was built on widely accepted works in landslide research area, such as the papers by Carrara *et al.* (1991, 1995), Guzzetti *et al.* (1999), Alvioli *et al.* (2016) and Rossi and Reichenbach (2016).

The methodologies used in LSA will be viewed from a broad view, such as qualitative and quantitative methods, to a narrow one that means a concrete model, such as ANN and SVM. The advantages and disadvantages of the models will be discussed.

2.2 Qualitative and Quantitative Methods

The methods applied in LSA can be broadly classified into two categories: qualitative and quantitative (Kanungo *et al.*, 2009). Qualitative approaches are considered subjective and describe the susceptibility levels in a qualitative way (Kanungo *et al.*, 2009). For example, assign ranks (low, moderate and high) to the locations with different landslide occurrence based on previous experiences. The results yielded by qualitative methods highly depends on the experts' knowledge and judgements (Dwikorita *et al.*, 2011; Saadatkhah *et al.*, 2014), such as heuristic analysis that is based on weights assigned by the expert's judgment (Abella and Van Westen, 2008). In order to remove the subjectivity in qualitative methods, various quantitative methods are applied in LSA in recent two decades (Saha *et al.*, 2005; Abella and Van Westen, 2008; Oh *et al.*, 2010; Fressard *et al.*, 2014; Guo *et al.*, 2015; Tsangaratos *et al.*, 2017; Hodasová and Bednarik, 2021).

Compared to qualitative approaches, quantitative approaches make use of a numerical assessment of the relationship between landslide occurrence and the influencing factors (Saadatkhah *et al.*, 2014) rather than experts' knowledge. Quantitative models are more important for scientists and engineers because it allows

the landslide susceptibility to be quantified in an objective and reproducible manner, and the results can be compared from one location to another (Corominas *et al.*, 2014).

Quantitative models, which is a broad term, can be further classified into two types of models: physical models and data-based models (Corominas *et al.*, 2014; Huang and Zhao, 2018). Physical models consider the failure mechanism of landslides when assessing the landslide susceptibility (Huang and Zhao, 2018). Physical approaches need sufficient geotechnical and soil depth data through a series of laboratory experiments that are suitable for small areas and hard to generate to large areas. Physical models belong to quantitative models but still contain subjectivity (Corominas *et al.*, 2014).

Along with the development of GIS and remote sensing techniques, more and more scientists focus on the data-based approaches, which are more suitable for LSA in large areas, and have obtained satisfactory outcomes (Ayalew and Yamagishi, 2005; Akgün and Bulut, 2007; Yalcin, 2008; Pourghasemi *et al.*, 2013a; Huang and Zhao, 2018; Bachri *et al.*, 2021; Chen and Chen, 2021). In this research, therefore, more emphasis will be put on data-based models. The available data for this research will be discussed in Chapter 3.

Data based models can be further classified into statistical, including bivariate and multivariate models, and machine learning models based on statistical theories (Reichenbach *et al.*, 2018). Before discussing the difference between statistical models and machine learning models, it is necessary to make clear that statistics differs from statistical models. Statistics can be regarded as the discipline that concerns the data collection, analysis, interpretation, and presentation (Romeijn, 2014), which is the mathematical study of data. However, a statistical model is a mathematical model that

is used either to infer something about the relationships between dependent and independent variables or to create a model that is able to predict future values and embodies a set of statistical assumptions concerning the generation of sample data (Cox, 2006).

Machine learning is as well a broad term that includes various types of models. In this research, the machine learning models are roughly classified into non-tree based and tree-based models. Compared to statistical models, machine learning models put more emphasis on how to obtain the optimal prediction rather than the relationship within the data and aim to equip computers with learning abilities without being explicitly programmed (Samuel, 2000). Table 2.1 provides a description on statistical and machine learning models. Figure 2.1 sketches the relationships between the models to get a clear picture of the models used in LSA. The models under the light gray box will be discussed in a detailed way in this research.

Table 2.1 Description on data-based models

Model type		**Description**
Statistical models	Bivariate models	Bivariate analysis explores the relationship between two variables, namely, landslide occurrence and each landslide influencing factor.
	Multivariate models	Multivariate analysis is simultaneous analysis of more than one influencing factors.
Machine learning models		Machine learning trains systems to have the ability to automatically learn from experience without being explicitly programmed.

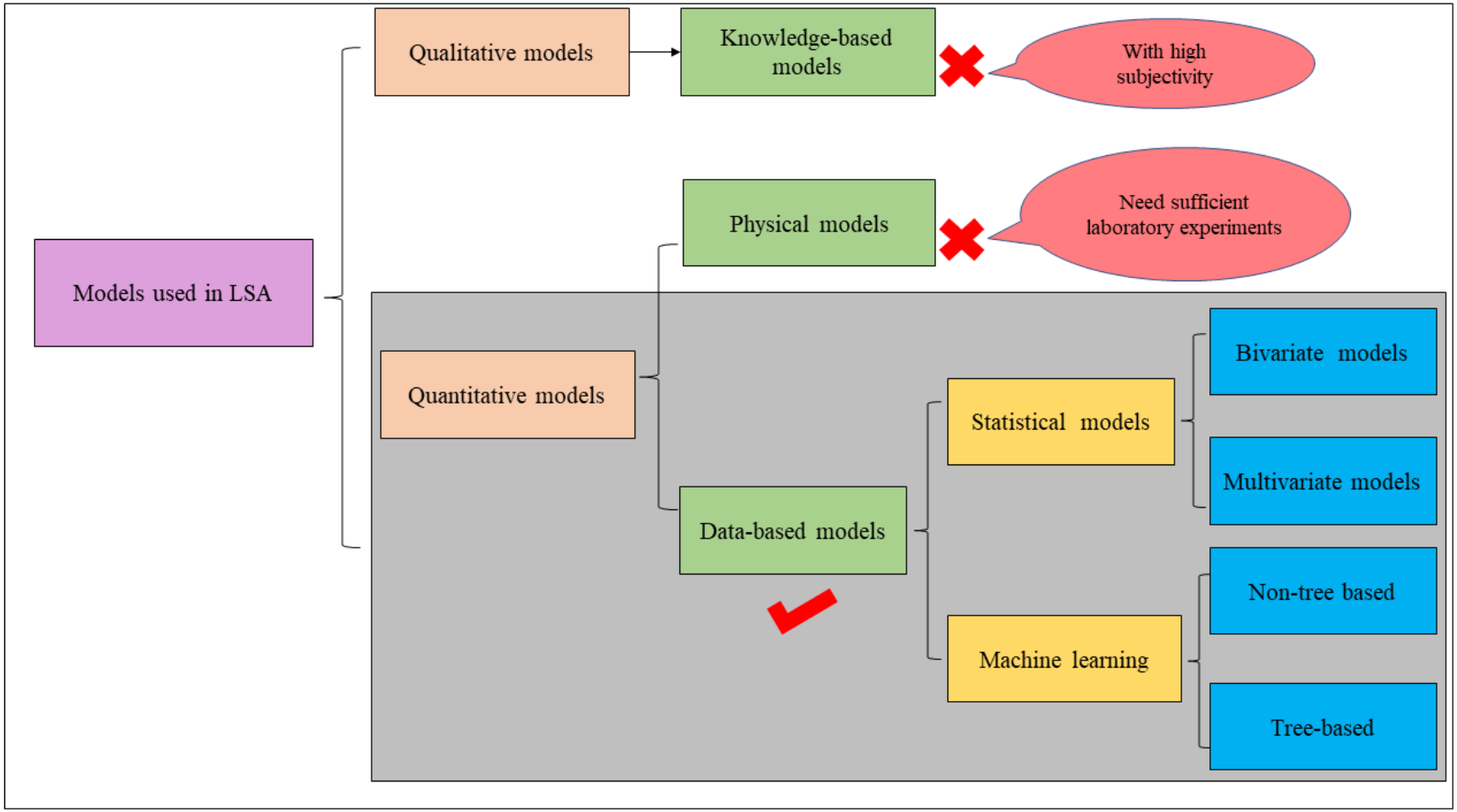


Figure 2.1 The models used in LSA

2.3 Bivariate Models used in LSA

Bivariate analysis is a simple statistical method which investigates the relationships between one dependent variable and one independent variable. Various bivariate models have been applied to LSA, such as frequency ratio (FR) (Lee and Sambath, 2006; Lee and Pradhan, 2007; Yilmaz, 2009; Mohammady *et al.*, 2012; Ozdemir and Altural, 2013; Shahabi *et al.*, 2014), fuzzy logic (FL) (Pradhan, 2010; Pourghasemi *et al.*, 2012; Zhu *et al.*, 2014; Shahabi *et al.*, 2015), weight of evidence (WoE) (Lee and Choi, 2004; Regmi *et al.*, 2010; Kayastha *et al.*, 2012), statistical index (Wi) (Bui *et al.*, 2011; Pourghasemi *et al.*, 2013a; Regmi *et al.*, 2014), weighted overlay model (WOM) (Basharat *et al.*, 2016; Shit *et al.*, 2016; Roslee *et al.*, 2017), and weighting factor (Wf) (Cevik and Topal, 2003; Bourenane *et al.*, 2016). Since FR and FL are much more popular and widely used among bivariate methods in LSA, more emphasis is placed on the application of the two methods in this overview.

2.3.1 Frequency Ratio (FR)

Frequency ratio is a simple and popular method used in LSA through dividing the occurrence ratio of landslide by the area ratio for each class of the influencing factor (Lee and Talib, 2005), which is easy to be applied in LSA (Ozdemir and Altural, 2013). It demonstrates the relationship between landslide distribution and each single conditioning factor (Balamurugan *et al.*, 2016). The detailed information about how to apply the method in LSA will be displayed in Section 3.5.

Shahabi *et al.* (2014) produced three landslide susceptibility maps using FR, logistic regression (LR), and analytic hierarchy process (AHP) models. The results in their study showed that the prediction power determined from area under curves (AUCs) are 0.8941, 0.8634, and 0.8115 obtained from LR, FR and AHP methods,

respectively. This indicates that FR has better predictive ability than AHP, but poorer than LR.

Huang *et al.* (2015) compared the LR and FR in producing landslide mapping and the validation results showed that LR (84.05%) had a better prediction performance than FR (76.64%) for the specific study area in Hong Kong. Meten *et al.* (2015b) also carried out research in central Ethiopia using LR and FR methods and obtained the same results with Huang *et al.* (2015), i.e., the map produced by FR has a lower success rate (FR: 74.8% & LR: 75.7%) and prediction rate (FR: 73.5% & LR: 74.5%) than the one by LR.

Landslide researchers usually favour in doing the comparison studies between FR and LR which is the most popular multivariate model used in LSM. The consistent comparison results, maps produced by LR performing better than those by FR no matter in success or accuracy rate or both, were obtained from the studies, such as Yilmaz (2009), Pradhan and Lee (2010a), Choi *et al.* (2012), Nourani *et al.* (2014), Sharma *et al.* (2014), Demir *et al.* (2015), Shahabi *et al.* (2015) and Wang *et al.* (2016a). Overall, most of the previous studies showed poor prediction performance for FR, except for the research by Solaimani *et al.* (2013), in which the FR showed a better performance than LR based on the accuracy values.

FR method has a relatively poorer prediction power in LSM than LR. It may be due to, on the one hand, the FR method has correlation only between landslide occurrence and conditioning factors while LR has both correlation and regression. Additionally, multivariate statistical methods, such as LR, are able to process more independent variables simultaneously than bivariate methods, like FR.

To sum up, FR analysis does not seem like a promising way in LSA. Many researchers, however, considered FR values as inputs when using other methods, such as FL (Pradhan and Lee, 2010b) and support vector machine (SVM) (Chen *et al.*, 2016b). In other words, FR model should be considered as a method to pre-process the raw data rather than a key method used to produce maps directly.

2.3.2 Fuzzy Logic (FL)

Fuzzy logic is widely applied in statistical models. It is based on the fuzzy set theory proposed by Zadeh (1965), which is relatively a young theory compared to other statistical theories, such as LR and discriminant analysis (DA) (Nedeljkovic, 2004). Although it shares some similarities to Boolean set theory, the key difference between them is that the membership of objects within a fuzzy set is defined (Dimri *et al.*, 2007).

FL method aims at modelling the imprecise modes of reasoning which are important for human beings to make rational decisions in a situation full of uncertainty and imprecision (Zadeh, 1988). There are five classical fuzzy operators, namely, fuzzy AND, fuzzy OR, fuzzy algebraic PRODUCT, fuzzy algebraic SUM, and fuzzy GAMMA operator. The mathematical expressions for the five operators will be displayed in Section 3.5.2.

Within the framework of fuzzy sets, overall, fuzzy GAMMA has the best prediction performance than fuzzy AND, fuzzy OR, fuzzy SUM and fuzzy PRODUCT (Lee, 2007; Pradhan *et al.*, 2009; Biswajeet and Saied, 2010; Ercanoglu and Temiz, 2011; Pradhan, 2010a; Kayastha, 2012; Pradhan, 2010b; Sharma *et al.*, 2013; Bortoloti *et al.*, 2015; Kumar and Anbalagan, 2015; Shahabi *et al.*, 2015; Rostami *et al.*, 2016). In some studies, the results show that the fuzzy PRODUCT and fuzzy GAMMA

possess comparable prediction power (Bui *et al.*, 2012a; Bui *et al.*, 2015; Bui *et al.*, 2017a).

Cervi *et al.* (2010) conducted a landslide spatial research in northern Apennines, Italy, using several models including the FL, WoE and shallow slope stability (SHALSTAB). The results showed that both the statistical models (0.77), namely FL and WoE, performed better than SHALSTAB (0.56) according to the global accuracy.

Bui *et al.* (2012b) evaluated and compared the landslide prediction performances of FL models and evidence belief function (EBF) in the Hoa Binh province of Vietnam. The results indicated that all models had good prediction capabilities. Furthermore, EBF model had the highest prediction ability while the model derived using fuzzy SUM had the lowest prediction capability. The fuzzy PRODUCT and fuzzy GAMMA models showed almost the same prediction capabilities.

Pourghasemi *et al.* (2012) compared the FL and AHP method and the verification results showed that the FL method (prediction power = 89.7%) had a better performance than AHP method (prediction power = 81.1%) in Haraz Watershed, Iran. There are other researchers who obtained better results for FL as well (Kanungo *et al.*, 2006; Tangestani, 2009; Ercanoglu and Temiz, 2011; Sahana and Sajjad, 2017).

The major advantage of FL, compared to general statistical methods, is that it allows the natural description of problems that should be solved in linguistic terms rather than in a precise way with numerical values (Nedeljkovic, 2004), which may be the main reason for those positive results produced using FL.

The negative results for FL, however, also exist in the literature (Biswajeet and Saied, 2010; Pradhan, 2010a; Bui *et al.*, 2012a; Meten *et al.*, 2015a; Vakhshoori and Zare, 2016), which might be due to the imprecision of FL theory. Therefore, numerous researchers tried to integrate FL with some other methods to produce LSMs in order to overcome the "weakness" of imprecision in FL.

Aghdam *et al.* (2016) proposed a hybrid model by integrating statistical index and adaptive neuro-fuzzy inference system. The validation results of the hybrid method showed that the AUC for success rate and prediction rate are 0.90 and 0.89, respectively, which can be used for land-use planning.

Bui *et al.* (2017a) proposed a novel fuzzy *k*-nearest neighbour inference model for LSM of rainfall-induced shallow landslides and obtained that the fuzzy *k*-NN model performed better with a high accuracy rate and prediction rate after comparing it to the SVM and decision tree (DT) models.

Feizizadeh *et al.* (2014) carried out a landslide prediction research in south-western Iran using a new method by integrating the fuzzy set theory with AHP method, which produced high accuracy and high level of reliability in the landslide susceptibility map. There are other landslide prediction studies integrating FL and other statistical methods, such as Gorsevski *et al.* (2006b), Vahidnia *et al.* (2010), Kanungo *et al.* (2011), Sdao *et al.* (2013), Anbalagan *et al.* (2015), Bui *et al.* (2015), and Lee *et al.* (2015). The results showed integrating FL with another method had a better prediction performance, which may be due to the relatively precise methods, such as ANN and DT, overcoming the imprecision "weakness" to some degree. The development of hybrid models can be considered as a promising way to solve natural hazard problems.

2.3.3 Summary of Bivariate Models

In Sections 2.3.1 and 2.3.2, two basic but significant bivariate methods used in LSA, namely, FR and FL, are discussed comprehensively. Beside the two mentioned methods, there are other bivariate methods which were less widely applied in LSA, such as WoE, a statistical approach based on Bayes' theorem (Vakhshoori and Zare, 2016; Lee and Choi, 2004), Wi (Bui *et al.*, 2011; Pourghasemi *et al.*, 2013a; Aghdam *et al.*, 2016), WOM (Gurugnanam *et al.*, 2012; Shit *et al.*, 2016), Wf (Yalcin, 2008; Yalcin *et al.*, 2011), and bivariate LR (Pradhan and Lee, 2010b).

In general, the bivariate models used in previous LSA studies showed good but not outstanding prediction performance. More different types of models deserve to be employed in LSA, such as multivariate and machine learning models.

2.4 Multivariate Models used in LSA

Multivariate analysis which involves more than one independent variable can provide more exact and reliable results than bivariate analysis which involves only one independent variable. This is because the interrelation among independent variables have non-negligible effect on the dependent variable. The most commonly used multivariate methods in LSA are logistic regression (LR) and discriminant analysis (DA). Minority of researchers implement other multivariate methods to produce LSMs, such as factor analysis (Komac, 2006). Therefore, the two models will be discussed in detail.

2.4.1 Logistic Regression (LR)

Logistic regression, developed by Cox (1958), is a widely used statistical model when the dependent variable is dichotomous, usually labelled '0' (non-