

THE USE OF QUANTILE AND LOGISTIC REGRESSION MODELS FOR  
MINIMUM OF VS30 CONDITIONED ON GEOLOGY AND TOPOGRAPHY IN  
TURKIYE

A THESIS SUBMITTED TO  
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES  
OF  
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AMIR JALEHFOROUZAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR  
THE DEGREE OF DOCTOR OF PHILOSOPHY  
IN  
ENGINEERING SCIENCES

JULY 2024



Approval of the thesis:

**THE USE OF QUANTILE AND LOGISTIC REGRESSION MODELS FOR  
MINIMUM OF VS30 CONDITIONED ON GEOLOGY AND TOPOGRAPHY  
IN TURKIYE**

submitted by **AMIR JALEHFOROUZAN** in partial fulfillment of the requirements  
for the degree of **Doctor of Philosophy in Engineering Sciences, Middle East  
Technical University** by,

Prof. Dr. Naci Emre Altun  
Dean, Graduate School of **Natural and Applied Sciences** \_\_\_\_\_

Prof. Dr. Murat Dicleli  
Head of the Department, **Engineering Sciences** \_\_\_\_\_

Assoc. Prof. Dr. Mustafa Tolga Yılmaz  
Supervisor, **Engineering Sciences, METU** \_\_\_\_\_

**Examining Committee Members:**

Prof. Dr. Ayşegül Askan Gündoğan  
Civil Engineering, METU \_\_\_\_\_

Assoc. Prof. Dr. Mustafa Tolga Yılmaz  
Engineering Sciences, METU \_\_\_\_\_

Assoc. Prof. Dr. Mustafa Kerem Koçkar  
Civil Engineering, Hacettepe University \_\_\_\_\_

Assoc. Prof. Dr. Mustafa Abdullah Sandikkaya  
Civil Engineering, Hacettepe University \_\_\_\_\_

Assoc. Prof. Dr. Zehra Çağnan Ertuğrul  
Engineering Sciences, METU \_\_\_\_\_

Date: 25.07.2024

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name Last name : Amir Jalehforouzan

Signature :

## **ABSTRACT**

### **THE USE OF QUANTILE AND LOGISTIC REGRESSION MODELS FOR MINIMUM OF VS30 CONDITIONED ON GEOLOGY AND TOPOGRAPHY IN TURKIYE**

Jalehforouzan, Amir  
Doctor of Philosophy, Engineering Sciences  
Supervisor: Assoc. Prof. Dr. Mustafa Tolga Yılmaz

July 2024, 137 pages

The main goal of this research is to develop regression models which can be used in the development of proxy-based Vs30 maps for Türkiye, so that the effect of site conditions can be considered in rapid earthquake damage and loss estimations. Using the dataset provided by AFAD, and the geological and topographical classifications developed within the project UDAP-Ç-20-01, this study explores quantile regression and logistic regression models to obtain probabilistic conditional estimations of Vs30 depending on geological and topographical properties. The geological class is found to be the prominent estimator for Vs30, and the topographical gradient contributes to the predictions secondarily.

Keywords: Proxy-Based  $V_{s30}$  Prediction, Quantile Regression Model, Logistic Regression Model, Geological and Topographical Properties

## ÖZ

# TÜRKİYE İÇİN MİNİMUM VS30 TAHMİNİNDE JEOLOJİ VE TOPOGRAFYA KOŞULLU KANTİL VE LOJİSTİK REGRESYON MODELLERİNİN KULLANIMI

Jalehforouzan, Amir  
Doktora, Mühendislik Bilimleri  
Tez Yöneticisi: Assoc. Prof. Dr. Mustafa Tolga Yılmaz

Temmuz 2024, 137 sayfa

Bu araştırmanın temel amacı, Türkiye için proxy tabanlı Vs30 haritalarının geliştirilmesinde kullanılacak regresyon modelleri geliştirmektir, böylece hızlı deprem hasarı ve kaybı tahminlerinde saha koşullarının etkisi dikkate alınabilir. AFAD tarafından sağlanan veri setini ve UDAP-Ç-20-01 projesi kapsamında geliştirilen jeolojik ve topografik sınıflandırmaları kullanarak, bu çalışma jeolojik ve topoğrafik özelliklere bağlı olarak Vs30 değerlerinin olasılıksal koşullu tahminlerini elde etmek için kantil regresyon ve lojistik regresyon modellerini araştırmaktadır. Jeolojik sınıfların Vs30 için asıl tahmin ediciler olduğu, topografik eğimin ise ikinci seviyede sonuçlara tesir ettiği görülmüştür.

Anahtar Kelimeler: Proxy -Tabanlı  $V_{s30}$  Tahmini, Kuantil Regresyon Modeli, Lojistik Regresyon Modeli, Jeolojik ve Topografik Özellikler

To My Beloved Family

## ACKNOWLEDGMENTS

The author wishes to express his deepest gratitude to his supervisor Assoc. Prof. Dr. Mustafa Tolga Yılmaz for his guidance, advice, criticism, encouragements and insight throughout the research.

I express my sincere thanks to my father Rahim Jaleh Forouzan, to my mother Hamideh Parnia and my brother Saeid Jaleh Forouzan for their boundless inspiration, encouragement, sacrifice and blessings. Without their support, love, patience and belief in me I would never have accomplished this. This thesis is dedicated to them.

In particular, I would like to express my sincere thanks to Assoc. Prof. Dr. Mustafa Kerem Koçkar, Gökhan Şahin, and Dr. Kıvanç Okalp for their invaluable assistance in preparing the geological and topographical dataset during my research.

This work is partially funded by Disaster and Emergency Management Presidency of Türkiye, AFAD, under grant number UDAP-Ç-20-01.

## TABLE OF CONTENTS

ABSTRACT.....	v
ÖZ.....	vi
ACKNOWLEDGMENTS .....	viii
TABLE OF CONTENTS.....	ix
LIST OF TABLES .....	xii
LIST OF FIGURES .....	xiv
LIST OF ABBREVIATIONS.....	xvii
LIST OF SYMBOLS .....	xviii
1 INTRODUCTION .....	1
1.1 General.....	1
1.2 Research Hypotheses .....	2
1.3 Research Scope .....	4
1.4 Outline of Thesis.....	5
2 LITERATURE REVIEW AND DATASET .....	7
2.1 Background.....	7
2.2 Dataset.....	12
2.2.1 Geological Restrictions .....	14
2.2.2 Topographic Restrictions .....	16
3 REGRESSION ANALYSIS .....	21
3.1 Introduction.....	21

3.2	Regression .....	23
3.3	Regression models classifications .....	24
3.3.1	Regression Models Classification Criteria .....	24
3.3.2	Assumptions and Constraints .....	26
3.3.3	Purpose of Regression Analysis .....	30
3.3.4	Types of Regression Models .....	31
3.4	Nonlinear Regression Models to Estimate $V_{s30}$ Values in Türkiye .....	36
4	QUANTILE REGRESSION .....	37
4.1	Introduction .....	37
4.2	Quantile Regression .....	41
4.2.1	Quantile Regression Model and Estimation .....	42
4.2.2	Goodness of Fit .....	45
4.3	Quantile Regression Model to Estimate $V_{s30}$ Values in Türkiye .....	48
4.3.1	Nonlinear Categorical Quantile Regression Model .....	48
4.3.2	Nonlinear Categorical Quantile Regression Models with Interaction Terms	52
4.3.3	Discussion of Results .....	56
5	LOGISTIC REGRESSION .....	69
5.1	Introduction .....	69
5.2	Logistic Regression .....	75
5.2.1	Different Types of Logistic Regression .....	78
5.2.2	Goodness of Fit .....	80
5.3	Application of Logistic Regression Model in Site Classification .....	81

5.3.1	Application of Ordinal Logistic Regression Model for Site Classification.....	81
6	CONCLUSION.....	103
6.1	General.....	103
6.2	Conclusions.....	104
6.3	Recommendations for Future Study .....	106
	REFERENCES .....	107
	APPENDICES .....	125
A.	Data transformation .....	125
B.	Tables.....	126
B. 1	Statistics of Geological Classes.....	126
B. 2	Quantile Regression Model Coefficient.....	127
C.	Figures.....	128
C.1	Scatter Plots of Geological Classes vs Topographic Attributes .....	128
C.1.1	Scatter Plots of Geological Class $C_1$ vs Topographic Attributes .....	128
C.1.2	Scatter Plots of Geological Class $C_2$ vs Topographic Attributes .....	130
C.1.3	Scatter Plots of Geological Class $C_3$ vs Topographic Attributes.....	132
C.1.4	Scatter Plots of Geological Class $C_4$ vs Topographic Attributes .....	134
	CURRICULUM VITAE.....	137

## LIST OF TABLES

### TABLES

Table 3.1 General types of variables .....	26
Table 4.1 Models with different numbers of terms vs goodness of fit criteria .....	50
Table 4.2 Model coefficients value .....	53
Table 4.3 Model Coefficient vs Estimation .....	55
Table 4.4 90% The proportion of $V_{s30}$ observations that are lower than (a) 16% quantile, and (b) $\mu$ - $\sigma$ estimation.....	60
Table 4.5 Summary of slope ranges for subdivided NEHRP $V_{s30}$ categories. ....	61
Table 4.6 Statistical analysis of NCQR at 50%, Iwahasi et al. (2010) and Nonlinear regression models (n=434).....	62
Table 4.7 Performance evaluation of method suggested by Allen and Wald (2007). 64	
Table 4.8 Statistical analysis of method suggested by Allen and Wald (2007).....	65
Table 4.9 Statistical analysis of method suggested by Allen and Wald (2007).....	65
Table 4.10 Statistical analysis of method suggested by Allen and Wald (2007).....	65
Table 4.11 Statistical analysis of method suggested by Allen and Wald (2007).....	66
Table 4.12 Statistical analysis of method suggested by Allen and Wald (2007).....	66
Table 4.13 Statistical analysis of method suggested by Allen and Wald (2007).....	66
Table 4.14 Statistical analysis of method suggested by Allen and Wald (2007).....	67
Table 4.15 Statistical analysis of method suggested by Allen and Wald (2007).....	67
Table 5.1 Türkiye's Soil Type Catalog .....	81
Table 5.2 Models with different numbers of terms vs goodness of fit criteria .....	83
Table 5.3 Coefficient's value of Equation 5.10 .....	83
Table 5.4 The sample for sites with $V_{s30} > 760$ m/s in Figure 5.11 .....	91
Table 5.5 The sample for sites with $V_{s30} > 760$ m/s in Figure 5.20 .....	92
Table 5.6 The sample for sites with $V_{s30} > 760$ m/s in Figure 5.13 .....	93
Table 5.7 The sample for sites with $V_{s30} > 360$ m/s in Figure 5.14 .....	94
Table 5.8 The sample for sites with $V_{s30} > 360$ m/s in Figure 5.15 .....	95

Table 5.9 The sample for sites with $V_{s30} > 360$ m/s in Figure 5.16.....	96
Table 5.10 The sample for sites with $V_{s30} > 360$ m/s in Figure 5.17.....	97

## LIST OF FIGURES

### FIGURES

Figure 2.1 AFAD’s strong ground motion sites in Türkiye (modified from Google Earth, 2023).....	13
Figure 2.2 $V_{s30}$ data descriptives for four geological period classes (According to studies conducted by members of the project UDAP-Ç-20-01) .....	16
Figure 2.3 Square cells used to calculate slope .....	17
Figure 4.1 $V_{s30}$ (m/s) vs Gradient (%) for all geological classes according to Equation 4.16.....	51
Figure 4.2 $V_{s30}$ (m/s) vs Gradient (%) for all geological classes according to Equation 4.17 .....	54
Figure 4.3 Histogram of residuals according to Equation 3.6.....	57
Figure 4.4 Histogram of residuals according to Equation 3.7.....	58
Figure 4.5 Histogram of residuals according to Equation 3.8.....	58
Figure 4.6 Histogram of residuals according to Equation 3.9.....	58
Figure 4.7 Estimated $V_{s30}$ (m/s) vs site’s $V_{s30}$ values for NCQR at 50%.....	62
Figure 4.8 Estimated $V_{s30}$ (m/s) vs site’s $V_{s30}$ values for model developed by Iwahashi et al. (2010).....	63
Figure 4.9 Estimated $V_{s30}$ (m/s) vs. site’s $V_{s30}$ values for nonlinear regression models developed by Gokhan et al. (2024) .....	63
Figure 5.1 The probability of $V_{s30} > 760$ m/s due to ordinal logistic regression model. ....	84
Figure 5.2 The gradient vs probability of $V_{s30} > 360$ m/s due to ordinal logistic regression model. ....	84
Figure 5.3 Gradient vs cumulative probability of ordinal logistic regression model to predict site with $V_{s30} > 760$ m/s for geological class $C_1$ .....	87
Figure 5.4 Gradient vs cumulative probability of ordinal logistic regression model to predict site with $V_{s30} > 760$ m/s for geological class $C_2$ .....	87

Figure 5.5 Gradient vs cumulative probability of ordinal logistic regression model to predict site with $V_{s30} > 760$ m/s for geological class $C_3$ .....	88
Figure 5.6 Gradient vs cumulative probability of ordinal logistic regression model to predict site with $V_{s30} > 760$ m/s for geological class $C_4$ .....	88
Figure 5.7 Gradient vs cumulative probability of ordinal logistic regression model to predict site with $V_{s30} > 360$ m/s for geological class $C_1$ .....	89
Figure 5.8 Gradient vs cumulative probability of ordinal logistic regression model to predict site with $V_{s30} > 360$ m/s for geological class $C_2$ .....	89
Figure 5.9 Gradient vs cumulative probability of ordinal logistic regression model to predict site with $V_{s30} > 360$ m/s for geological class $C_3$ .....	90
Figure 5.10 Gradient vs cumulative probability of ordinal logistic regression model to predict site with $V_{s30} > 360$ m/s for geological class $C_4$ .....	90
Figure 5.11 The change on proportion of sites with $V_{s30} > 760$ m/s according to the gradient intervals for geological class $C_2$ .....	91
Figure 5.12 The change on proportion of sites with $V_{s30} > 760$ m/s according to the gradient intervals for geological class $C_3$ .....	92
Figure 5.13 The change on proportion of sites with $V_{s30} > 760$ m/s according to the gradient intervals for geological class $C_4$ .....	93
Figure 5.14 The change on proportion of sites with $V_{s30} > 360$ m/s according to the gradient intervals for geological class $C_1$ .....	94
Figure 5.15 The change on proportion of sites with $V_{s30} > 360$ m/s according to the gradient intervals for geological class $C_2$ .....	95
Figure 5.16 The change on proportion of sites with $V_{s30} > 360$ m/s according to the gradient intervals for geological class $C_3$ .....	96
Figure 5.17 The change on proportion of sites with $V_{s30} > 360$ m/s according to the gradient intervals for geological class $C_4$ .....	97
Figure 5.18 NCQR vs ordinal regression model for geological class $C_1$ with $V_{s30} > 360$ m/s .....	100
Figure 5.19 NCQR vs ordinal regression model for geological class $C_1$ with $V_{s30} > 760$ m/s .....	100

Figure 5.20 NCQR vs ordinal regression model for geological class $C_2$ with $V_{s30} > 360$ m/s .....	101
Figure 5.21 NCQR vs ordinal regression model for geological class $C_2$ with $V_{s30} > 760$ m/s .....	101
Figure 5.22 NCQR vs ordinal regression model for geological class $C_3$ with $V_{s30} > 760$ m/s .....	102
Figure 5.23 NCQR vs ordinal regression model for geological class $C_4$ with $V_{s30} > 760$ m/s .....	102

## **LIST OF ABBREVIATIONS**

### **ABBREVIATIONS**

AFAD: Afet ve Acil Durum Yönetimi Başkanlığı

DEM: Digital Elevation Models

GIS: Geographic Information System GIS

MAE: Mean Absolute Error

MASW: Multichannel Analysis of Surface Waves

NCQRM: Nonlinear Categorical Quantile Regression Model

NEHRP: National Earthquake Hazard Reduction Program

OLS: Ordinary Least Squares

OLS: Ordinary Least Squares Regression

PGA: Peak Ground Acceleration

PGV: Peak Ground Velocity

QRM: Quantile Regression Model

RMSE: Root Mean Squared Error

SEM: Structural Equation Models

USGS: U.S. Geology Survey

## LIST OF SYMBOLS

### SYMBOLS

$\hat{Y}_i$ : Estimated Dependent Value

$\frac{\partial^2(e)}{\partial x^2}$ : The Second Partial Derivative of Elevation with Respect to the  $x$ -Coordinate

$\frac{\partial^2(e)}{\partial y^2}$ : The Second Partial Derivative of Elevation with Respect to the  $y$ -Coordinate

$\Delta^2(e)$ : The Laplacian Operator

$z_{\alpha/2}$ :  $Z$ -Value for the Corresponding Confidence Level

$H_5$ : Elevation of Central Point

$H_i$ : Elevation of Each Neighbor Point

$LL_{Model}$ : Natural Logarithm of the Likelihood of the Model with Predictors

$LL_{Null}$ : Natural Logarithm of the Likelihood of the Model Without Predictors

$Q^{(p)}(Y_i|X_i)$ : Quantile Regression For  $p$ th Conditional Quantile

$Q^{(p)}(\epsilon_i)$ : Error Term for the  $p$ th Quantile

$R^2$ : R-Squared

$R_{MF}^2$ : McFadden's  $R^2$

$R_{MFA}^2$ : McFadden's Adjusted  $R^2$

$S_i$ : Slope

$\bar{Y}$ : Sample Mean of Real Dependent Variables

$Z(x)_l$ : The Equation to Be Used in the Ordinal Logistic Regression Model for Each Class

$d_p$ : Weighted Distance

$\beta_0$ : Coefficient of Linear Regression Model

$\beta_0^p$ : Quantile Specific Coefficient of Quantile Regression Model

$\beta_1$ : Coefficient Of Linear Regression Model

$\beta_1^p$ : Quantile Specific Coefficient of Quantile Regression Model

$\epsilon_i^p$ : Quantile Specific Random Error Term of Quantile Regression Model

$0 < p < 1$ : Proportion of the Population with Dependent Values Below the Quantile  
At p

$D_k$ : The Cut-Point Value to Be Used for Each Class

$D_l$ : The Cut-Point Value to Be Used for Each Class (For Ordinal Logistic  
Regression)

$E(\epsilon_i)$ : Expected Value Of Random Error Term

$f(\cdot)$ : A Nonlinear Function

$f(X_i, Y_i)$ :  $Y_i$  is a Function of  $X_i$

$g(x)$ : The Model to Be Used to Predict the Success Probability in Multinomial Logistic  
Regression Model

$h(x)$ : The Model to Be Used to Predict the Cumulative Success Probability

$k$ : Number of Each Dependent Variable Class (For Multinomial Logistic Regression)

$K$ : Number Of the Dependent Variable Class Which Is Selected as Pivot Class (For  
Multinomial Logistic Regression)

$L$ : Number Of All Dependent Variable Classes That Have Defined Order (For  
Ordinal Logistic Regression)

$l$ : Number of Each Dependent Variable Class (For Ordinal Logistic Regression)

$n$ : Number of Observations

$N$ : Number of Parameter, Coefficients, In the Logistic Regression Model

$P(Y = k)$ : The Probability by Which the Dependent Variable Belongs to Pivot Class

$P(Y = k)$ : The Probability by Which the Dependent Variable Belongs to Other Classes

$P(Y=1)$ : Success Probability

$X_i$ : Independent Variables

$Y_i$ : Dependent Variables

$\beta_i$ : Coefficients

$\pi(X_i)$ : Expected Value of  $Y$  Given  $X$  For Logistic Distribution

$D$ : Size Of the Cell

$E(Y_i|X_i)$ : Expected Value of Dependent Variables According to Independent Variables

$E(\epsilon|X_i)$ : Expected Value of Random Error Term Given the Independent Variable

$N$ : Total Sample Size

$v$  : Number Of Independent Variables

$\epsilon_i$ : Random Error Term

# CHAPTER 1

## INTRODUCTION

### 1.1 General

Earthquakes stand as among the most perilous natural calamities, wielding profound impact on local communities and the constructed environment. The intensity of ground shaking is contingent on a confluence of factors, earthquake source mechanism, size, fault rupture distance, and subsurface geology. Significantly, the augmentation of ground motions by near-surface materials can crucially escalate the aftermath, often demarcating the line between minor and severe destruction. The response of a site to seismic activity, termed site response, is typically characterized by a classification scheme. Alas, comprehensive information on seismic site conditions remains sparse. Acknowledging this, the role of site topography and surficial geology, particularly the influence of soft sediments, assumes prominence in amplifying ground shaking, a pivotal element in forecasting ground motion levels at any site.

To predict seismic shaking or loss, it becomes imperative to estimate potential amplification stemming from site conditions across affected or potentially affected zones. This framework delineates how near-surface geology magnifies ground shaking relative to the bedrock's shaking level. Amidst this context  $V_{s30}$  emerges as a paramount parameter embodying site conditions.

Originally introduced by Borchardt (1994) as a fundamental measure for site-amplification factor estimation, this parameter has assumed a crucial role in seismic site classification as mandated by seismic design codes for structures. Prominent examples include the International Building Code, IBC, published in 2021, Eurocode 8, EC-8, issued by the European Committee for Standardization, CEN, in 2004, and

notably, the Turkish Seismic Code, AFAD, in 2018. Also, it finds application in ground-motion prediction equations (Abrahamson and Silva, 2008; Boore et al., 1993, 1994) and hazard maps. Given the arduous and costly nature of determining  $V_{s30}$ , the necessity to develop proxy-based  $V_{s30}$  maps becomes evident, serving as a practical means to bridge the gap in comprehensive  $V_{s30}$  data availability.

## 1.2 Research Hypotheses

Türkiye experiences a significant earthquake of magnitude 7 or higher approximately every decade, if not more frequently. This occurrence excludes the numerous impactful events of lower magnitudes. Notable instances of such impactful seismic events include the 1992 Otlukbeli (Erzincan), 1999 Golcuk (Kocaeli) and Duzce, 2011 Ercis (Van), and more recently, the 2023 Pazarcik (Kahramanmaras) and Elbistan (Kahramanmaras) earthquakes. These occurrences serve as illustrative examples of the destructiveness associated with such seismic events within the region. The AFAD-RED program, supported by the Disaster and Emergency Management Presidency of Türkiye, AFAD, serves as a pivotal tool for generating ShakeMap. These ShakeMaps are instrumental in assessing the spatial extent of damage caused by specific seismic events. To achieve this, the program necessitates a seismic site condition map to gauge the impact of site attributes on ground motion. While microzonation studies offer detailed seismic site condition maps for certain major cities within the country, the scenario is different for regions that encompass rural areas, towns, or smaller cities. In these locations, the availability of a comprehensive seismic microzonation is sporadic, and its potential completion remains uncertain. Consequently, a substantial number of damaged structures are situated in these areas, which underscores the importance of alternative methodologies, like the AFAD-RED program, to estimate seismic effects and guide disaster response efforts effectively. The necessity for a nationwide site-

conditions map became notably pronounced in the wake of the 2023  $M_w7.6$  and  $M_w7.7$  earthquakes that struck Kahramanmaras. These seismic events underscored the imperative of having a comprehensive understanding of site attributes to enhance disaster response and management strategies.

The main goal of the present study is to develop models that can be used through  $V_{s30}$  maps preparation for Türkiye, to be used in the AFAD Rapid Earthquake Damage and Loss Estimation Software (AFAD-RED, ShakeMap) in the aftermath of earthquakes occurring in Türkiye, so that the site effects can be considered in estimations. Spectral analysis of surface waves, SASW, seismic cone penetration test, SCPT, seismic refraction/reflection test, suspension logging test, downhole test, cross-hole test, and standard penetration test-N value, SPT-N, are commonly employed methodologies for local  $V_{s30}$  determination. In light of the fact that these surveys are not available everywhere, a country-wide simple and reliable model for mapping  $V_{s30}$  will be critically important. To achieve this goal, a detailed site-conditions map will be created using the nation's digital geological map and global elevation data. Through the application of empirical relationships, a digital  $V_{s30}$  map can be generated, facilitating further integration into Geographic Information Systems (GIS). As part of ongoing efforts, AFAD is progressively installing fresh strong-motion sites across the country to amass accelerometric data post-earthquakes.  $V_{s30}$  is systematically surveyed at these sites employing the MASW technique as explained by Xia et al. (1999). However, our preliminary analysis of the data highlights the inadequacy of existing literature-based relationships to provide accurate  $V_{s30}$  estimations for Türkiye. Notably, singular reliance on topographical slope yields statistically unsatisfactory results in estimating  $V_{s30}$ . To address this limitation, research was performed to forge novel correlations between the geological formations meticulously mapped across Türkiye, pertinent topographical parameters, and  $V_{s30}$ . The emphasis is on developing statistical models that improve the predication, thereby enabling the conditional estimation of  $V_{s30}$ . Subsequently, lower bounds for conditional  $V_{s30}$  values are proposed, bolstered by an alternative statistical model to substantiate the derived estimations.

The resultant estimation functions derived from this rigorous approach can be utilized to prepare the GIS-driven generation of  $V_{s30}$  maps designed to Türkiye's distinct seismic context.

### **1.3 Research Scope**

As previously mentioned, the development of a  $V_{s30}$  prediction model, that can be used AFAD, is the main goal of this research. Given the resource-intensive nature of measuring  $V_{s30}$  values in terms of both cost and time, a paramount approach involves constructing robust models useful at estimating  $V_{s30}$  values based on the inherent geological and topographic attributes at specific sites.

Steps to attain this objective include:

- 1) To explore the geological and topographic characteristics of locations where AFAD provides  $V_{s30}$  values.
- 2) To investigate nonlinear categorical quantile regression following geological and topographic constraints.
- 3) To investigate logistic regression models based on Türkiye's  $V_{s30}$  catalog, utilized for the purpose of site classification.
- 4) To conduct a comparative analysis between the model developed in this study and previously established models, aiming to assess their robustness and precision.

## 1.4 Outline of Thesis

This thesis is prepared in 6 chapters: Introduction; Literature Review; Regression Analysis; Quantile Regression; Logistic Regression; Analysis of Results and Discussion.

Chapter 1: This chapter introduces the fundamental necessity for the current research, elucidates the main objectives, outlines the study's scope, and provides a structural overview of the thesis.

Chapter 2: Chapter 2 presents an in-depth examination of existing literature pertaining to proxy-based  $V_{s30}$  maps. This literature review serves as a basic underpinning for the concepts and contributions presented within the thesis. It critically evaluates prior research relevant to the subject matter, culminating in a discussion that underscores the distinctiveness of the current study. This distinction is articulated to substantiate the research's significance and to underscore its novel contribution to the field. Additionally, the geological and topographic attributes of sites provided by AFAD, where  $V_{s30}$  values have been ascertained, are subject to comprehensive examination and discussion.

Chapter 3: This chapter provides an extensive explanation of the classifications of regression models. It comprehensively addresses the criteria and constraints that inform the categorization of regression models. The discussion encompasses various types of statistical data structures that play a crucial role in classification. Furthermore, the chapter introduces nonlinear regression models specifically designed for predicting  $V_{s30}$  values in Türkiye.

Chapter 4: This section investigates an in-depth exploration of quantile regression models, shedding light on their intricacies. It examines various dimensions of quantile regression models, encompassing discussions on the assessment of goodness of fit criteria specific to this regression approach. Additionally, the development of

nonlinear categorical quantile regression models to predict  $V_{s30}$  values according to geological and topographic properties takes center stage, with comprehensive details on their construction. An integral aspect of this section involves a thorough comparison between these nonlinear categorical quantile regression models and nonlinear categorical quantile regression models featuring interaction terms.

Chapter 5: This chapter delves into a comprehensive discussion of logistic regression models employed as classifiers. The exploration encompasses a detailed assessment of various types of this classification method, coupled with an elucidation of the goodness-of-fit measures tailored for this model. Moreover, the chapter proceeds to detail the development of logistic regression models under the constraints of topographic and geologic considerations. These models are used for categorizing sites into distinct classes based on Türkiye's seismic catalog.

Chapter 6: This section of the dissertation submits the summary, conclusions, and contribution of the study. The novality of the contributions provided by this thesis are presented in this chapter. Also, this chapter suggests some recommendations for future study.

## CHAPTER 2

### LITERATURE REVIEW AND DATASET

#### 2.1 Background

ShakeMaps are geographic products, which provide a spatial representation of the extent and distribution of ground shaking caused through an earthquake (Wald et al., 2008). These maps consider the intensity of earthquake following a ground shaking rather than parameters explaining the source of an earthquake. Thus, for an earthquake with a specific epicenter and magnitude different ranges of ground shaking levels are expected for a region based on the distance from the shaking source, site geological condition, the topographical condition of the site, and changes in the propagation of seismic waves from the ground shaking due to complexities in the structure of the earth's crust. In general, ShakeMaps are referred to determine instrumentally derived intensities and peak ground motion, PGM, parameters containing peak ground acceleration (PGA), peak ground velocity (PGV), and spectral acceleration response (Michellini et al., 2008). These maps are utilized by governmental organizations for post-earthquake response, public and scientific knowledge, as well as for preparation exercises and disaster planning.

In addition to source and path effects on earthquakes the local site conditions, such as sediment thickness and impedance contrast, have a key role in determining ground motion intensity during ground shaking (Boore et al., 1997; Field, 2000). Site topography, subsurface material geometry and properties, and input motions are the basic parameters influencing ground motion (Kramer, 1996). Commonly site classification schemes are used to describe how near-surface soil deposits influence the earthquake level, relative to the level of shaking of the underlying bedrock (Kelly, 2006). Recently, National Earthquake Hazard Reduction Program, NEHRP, developed a classification (adapted from D. J. Wald and Allen, 2007) that is used by the U.S.

Geology Survey, USGS, to generate ShakeMaps. Another common site classification scheme is the National Building Code of Canada, NBCC, which applies these site conditions within strict guidelines for buildings prone to earthquake, which must be designed to withstand (Hunter et al., 2010). International Building Code 2007, IBC, Uniform Building Code 1997 UBC, and Eurocode-8 are other examples through which site classification is used for building seismic design. Since soil rigidity correlates with shear-wave velocity,  $V_s$ , in geological materials, generally  $V_s$  has been accepted as a measure to define soil classification (Boore, 2006; Fumal, 1978). Although the variation of material properties at a depth of tens to hundreds of meters beneath the surface of earth, even deeper, is found to have a role in site amplification (Boore, 2004; Frankel et al., 2002; Holzer et al., 2005); several studies demonstrated an acceptable relation between site amplification and  $V_{s30}$  (Boore, 1997; Borchardt, 1994a). In addition, most of the earthquake acceleration (Holzer, Bennett, et al., 2005) prediction equations are developed according to seismic site-conditions, described with  $V_{s30}$  values (Abrahamson and Silva, 2008; Boore, 1997; Boore et al., 1993).

Since determination of  $V_{s30}$  is prohibitive and sparse, it is common to use geological and topographic properties as proxy to predict  $V_{s30}$ . Therefore, the aforementioned properties are used separately or in combination with geostatistical methods to estimate  $V_{s30}$  where there are no observations to develop proxy-based  $V_{s30}$  maps.

A number of proxy-based geologic maps have been developed based on the correlations between  $V_{s30}$  and geological structure. In the Los Angeles basin, by relating surface geology to the variation of  $V_s$  with grain size, age, and depth Tinsley et al. (1985) characterized the site classification for specific geologic units (Tinsley et al., 1985). Based on averaged  $V_{s30}$  seven broad geological units were proposed within California (Wills and Silva, 1998). Also, further refinements were performed with the California geological map according to  $V_{s30}$  and more detailed geological data (Wills et al., 2000; Wills and Clahan, 2006). For the central and eastern region of U.S., a geology map was prepared based on soil thickness and description (Fullerton et al., 2004). The proposed geologic units were used to develop a  $V_{s30}$  map (Withers, 2007). A 3-Dimensional model was derived from  $V_s$  and geological properties including the

extent, thickness, and velocity of each geologic units (Holzer et al., 2005). According to available  $V_{s30}$  data prepared by AFAD at strong ground motion sites, Yilmaz et al. (2014) investigated mean  $V_{s30}$  and standard deviation for five geological classes in Türkiye.

By using two geographic rules in the Geographic Information System (GIS), the distance-based rule and the slope-based rule, young alluvium regions were divided into smaller units (Wills et al., 2015) in order to evaluate the correlation developed by Wills and Clahan (2006). The other example of using GIS in  $V_{s30}$  map development was the one prepared by Matsuoka (1995) in Japan by using terrain-based classification and correspondent peak ground velocity. The other GIS-Class-based study in Japan considered site amplification estimation by attending to the geological and geomorphological data (F. Yamazaki et al., 2000). Recently site classifications have been constructed using available remote sensing data (Romero, 2001; Yong et al., 2008, 2012). For Portugal Vilanova et al. (2018) developed a  $V_{s30}$  site-condition map, a map showing the site effect on the ground motion which is based on  $V_{s30}$ .

In many seismically active zones, surficial geology information neither exists nor is easily accessible. Conversely, topographic information is existing for the globe. Actually, topographic variations demonstrate the geomorphology and lithology of near-surface layers (Wald and Allen, 2007). Indeed, the topographic slope derived from the Shuttle Radar Topography Mission, SRT, (Farr and Kobrick, 2000) was used to estimate  $V_{s30}$  in California and Taiwan (Allen and Wald, 2009; Wald and Allen, 2007). Wald and Allen (2007) studied the correlation between  $V_{s30}$  and slope for active and stable tectonic regions. Also,  $V_{s30}$  is divided into classified ranges according to Borchardt's (Borchardt, 1994) classification scheme. Intuitively, the greater the topographic slope, higher the  $V_{s30}$ , therefore, the stiffer bedrock associates with large topographic slopes. Based on empirical observations there is a nonlinear relationship between topographic slope and  $V_{s30}$  values. Magistrale (2012) proposed separate correlations between topographic slope and  $V_{s30}$  for tectonically stable central and eastern U.S., tectonically active western U.S. and for areas of the western U.S that hosted Pleistocene and younger lakes. For Europe and the Middle East Lemoine et al.

(2012) evaluated the potential applicability of topographic slope-based estimation of  $V_{s30}$  for active and stable sites. According to results for active sites, this method is better than blind chance, but for stable regions data is limited. Because of restrictions with resolution of the topographic data we face constrained  $V_{s30}$  estimation. Research on the effect of data resolution on  $V_{s30}$  estimation illustrates that using higher resolution data submits better results, in terms of  $V_{s30}$  estimation, in comparison with lower resolution data (Allen and Wald, 2009). The assumptions through which topographic slope is referred to as a proxy to develop the  $V_{s30}$  map are valid in many cases since steeper slopes tend to be solid rock while shallower slopes tend to be unconsolidated soils. However, under certain circumstances, such as South Table Mountain in Golden where there is a top flat with solid rock, the previous assumptions are not valid. The lack of high  $V_{s30}$  sample values, typically associated with hard rock sites, is the other weakness of slope-based  $V_{s30}$  maps. A regional slope-based  $V_{s30}$  predictive map was added by Heath et al. (2020) to the global  $V_{s30}$  map to develop a more accurate map. In addition to an assessment of the effect of digital elevation models, Dem's, resolution on DEMs-based  $V_{s30}$  maps, and topographic slope were used for  $V_{s30}$  mapping in Iran (Karimzadeh et al., 2019). Lin et al. (2019) performed a linear and nonlinear regression to investigate the slope-based  $V_{s30}$  map for California.

By investigating the hierarchical approach, and decomposition of complex problems by reducing them to a smaller set of interrelated problems, the  $V_{s30}$  map was derived from topographic slopes and geological observations (Wald et al., 2011). In addition, a hybrid model was developed by regressing  $V_{s30}$  to the geologic properties, topographic-slope, and cross-term coefficients. Because of the strong spatial correlation, demonstrated by residuals, the kriging-with-a-trend method, the trend is the hybrid model, was used for more refinement of the  $V_{s30}$  prediction map. Thompson et al. (2010, 2011, 2014) presented a  $V_{s30}$  map according to geological and topographical properties. In California young alluvium were divided into three classes with distinct  $V_{s30}$  range based on surface slope (Wills and Gutierrez, 2008). Applying the same system and utilizing more detailed geologic maps, from a 1:250,000 scale to 1:24,000 for much of California region, resulted in more acceptable results (Wills et

al., 2015). For eastern North and central part of America Parker et al. (2017) investigated a hybrid geology-slope approach for  $V_{s30}$  estimation using largescale geologic maps. For all of Japan Matsuoka (2005) performed a mapping of  $V_{s30}$  according to geomorphologic data, elevation and slope gradient, and geographic information, distance from hills, by using a multiple regression method. By utilizing global geomorphological categorization based on normalized topographical slope, local convexity, and surface texture, after Iwahashi et al. (2018), a large-scale mapping of  $V_{s30}$  was performed for Italy (Mori et al., 2020). The results obtained from studies demonstrate that hybrid models based on topography and geology result in more accurate estimations.

In keeping with the previous studies, to develop AFAD-RED, ShakeMap, after an earthquake in Türkiye, it is required to evaluate the site effect on ground motion, which should be performed by estimated  $V_{s30}$ . Mapping of  $V_{s30}$ , to develop a reliable model, must be performed based on geologic and topographic data. Since several weaknesses were observed in the topographic slope-proxy-based model it is needed to look for novel and more practical topographic functions for the estimation of  $V_{s30}$ .

## 2.2 Dataset

The current research incorporated three primary datasets sourced from AFAD: digital elevation information, a digital geological map, and records of  $V_{s30}$  values. The dataset, including topographic properties and geological classifications, was collaboratively prepared by members of the group involved in the project UDAP-Ç-20-01. Notably, the elevation data prominently featured the Multi-Error-Removed Improved-Terrain Digital Elevation Model, MERIT-DEM, which is openly accessible at 90-meter resolution (Yamazaki et al., 2017). Despite of potential inaccuracies, the MERIT-DEM presents a high-precision model that effectively mitigates substantial errors, by enhancing vertical accuracy, specifically within flat regions (Yamazaki et al., 2017). This aspect holds notable significance for the study, given that, a majority of the examined  $V_{s30}$  data originates from areas characterized by low slopes. Iwahashi et al. (2018) formulated a terrain classification technique reliant on the MERIT-DEM, subsequently facilitating the indirect prediction of  $V_{s30}$  values. Employing a comparable methodology, the current investigation studies the interrelation between slope, texture, and convexity parameters derived from the MERIT-DEM's 90-meter resolution digital elevation dataset, alongside the corresponding  $V_{s30}$  records obtained from the sites with strong ground motion.

In pursuit of this objective, several studies utilized the 2-D trend surface analysis methodologies by Davis (1986), Riley et al. (1999), Iwahashi and Pike (2007), Devore (2012) and Conrad et al. (2015) to investigate and enhance the relationship between topographic information and  $V_{s30}$  values through the utilization of second-order polynomials.

The second dataset encompasses a digital geological map of Türkiye, compiled by the General Directorate of Mineral Research and Exploration of Türkiye, MTA, at a scale of 1:500,000. This map plays a key role in characterizing the geological units and their respective periods, wherein  $V_{s30}$  measurement surveys were conducted. Through the precise assessment of the map, several geological units and their corresponding period

classes have come to light. Notably, there are 38 units categorized by code, 79 units indicated by symbol, 8 units classified by period, era, and class, and 36 units specified by period (Okalp, 2013). However, upon a comprehensive evaluation, it becomes evident that a substantial portion of these geological units lacks a sufficient sample size to draw dependable statistical inferences. As a result, the process necessitated reclassifications and more assessments to be undertaken.

The third dataset encompasses  $V_{s30}$  profiles gathered from a total of 511 ground motion sites by AFAD, through a range of studies (Akkar et al., 2010; Kurtuluş et al., 2020). These  $V_{s30}$  profiles are procured by surface seismic testing methodologies such as multichannel analysis of surface waves, known as MASW. This collection of samples possesses a fairly uniform dispersion, effectively encompassing Türkiye's diverse geological and lithological attributes (Figure 2.1). It is worth noting that a clustering of data emerges within regions characterized by low slopes when the focus solely centers on the topographic slope parameter. This clustering phenomenon can be attributed to the  $V_{s30}$  values gathered by AFAD in close proximity to ground motion sites, which are typically situated in flat zones.

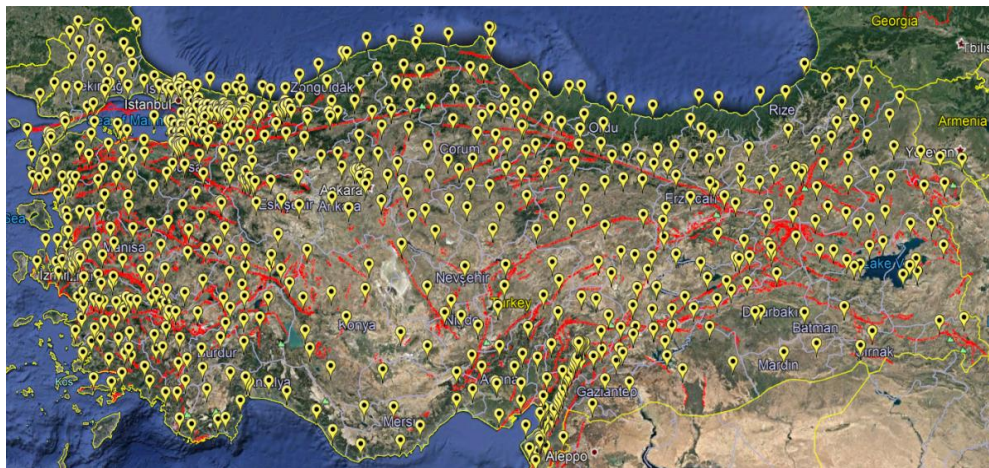


Figure 2.1 AFAD's strong ground motion sites in Türkiye (modified from Google Earth, 2023)

After evaluations, a total of 511 site data points were subject to assessment, encompassing both findings from scholarly investigations and compiled data prepared by AFAD. During this evaluation process, careful consideration was given to the geological periods and units associated with each site. To conduct a comprehensive analysis, the sites were systematically classified according to these geological factors.

### **2.2.1 Geological Restrictions**

During the studies conducted by members of the project UDAP-Ç-20-01, the  $V_{s30}$  dataset, comprising 511 samples along with site coordinates and supplementary attributes, underwent an initial digitization process within a GIS, Geographic Information System, environment. By superimposing the digitized site's locations onto the geological map, it became possible to integrate the surface geology information with the corresponding  $V_{s30}$  values of each site. Subsequent analysis revealed that among the total of 84 geological units, 511 sites with  $V_{s30}$  measurements were distributed across 38 distinct geological units (Okalp, 2013). Remarkably, more than two-thirds of these sites were positioned on Quaternary sediments. This inclination can be attributed to the strategic placement of sites in close proximity to densely inhabited areas situated on gently sloping terrain.

Acknowledging the extensive array of mapped geological units, totaling 84, whose presence consequently curtails the statistical robustness of the analyses due to the constrained sample size, a strategic reorganization was executed, reorganizing these units into 19 distinct classes. A careful evaluation of the regression model, founded upon the reorganized dataset, revealed a relatively modest goodness of fit as indicated by an  $R^2$  value of 0.243. This outcome prompted the recognition that enhancing the established correlation could be attainable through the exploration of alternative geological classifications. Thus, in pursuit of this objective, the geological period is

posited as a suitable alternative to the existing geological unit classification to formulate an improved regression model.

Subsequent to the survey of data from 511 sites, a stratification process ensued, segregating them into 5 distinct geological period classes: "Paleozoic and Paleozoic-Mesozoic (P/PM)", "Mesozoic and Mesozoic-Tertiary (M/MT)", "Tertiary and Tertiary-Quaternary (T/TQ)", and two classes categorized as "Quaternary (Qa)" and "Quaternary (Q1)". It's noteworthy that the designations "Q1" (encompassing basalts, andesites, pyroclastics, terrestrial clastic, travertines) and "Qa" (encompassing young sediments) serve as subclasses under the overarching class of "Quaternary (Q)". Considering both the dataset and geological classification, an in-depth comparison yielded no significant distinctions between the "Q1" and "T/TQ" classes. Subsequent to the development of multiple linear regression models, the  $R^2$  values falling below 0.299 indicated an insufficient association between these geological classes and the corresponding  $V_{s30}$  values. Upon closer examination of the distinct characteristics exhibited between the geological classes "Q1" and "T/TQ" as well as "Qa", it became evident that a finer distinction within the "Q1" class was necessary. This led to the subdivision of the "Q1" class into two subcategories: "Q1c" representing terrestrial clastic and travertines, and "Q1vm" encompassing basalt, andesite, and pyroclastic rocks. Subsequently, it was decided to streamline the original four classes into two more comprehensive categories: the amalgamation of "Qa" and "Q1c" gave rise to the category "Quaternary (Qa/Q1c)", while the consolidation of "Q1vm" and "T/T-Q" resulted in the category "Tertiary and Tertiary-Quaternary (T/TQ/Q1vm)". Subsequently, all the sites were categorized into four distinct geological classes (Figure 2.2): "P/PM (C<sub>4</sub>)", "M/MT (C<sub>3</sub>)", "T/TQ/Q1vm (C<sub>2</sub>)", and "Qa/Q1c (C<sub>1</sub>)".

An analysis of the dataset unveiled the presence of sites situated in close proximity to geological period boundaries on the map. In light of this, a decision was reached to exclude sites that lacked clear certainty regarding their placement within specific geological units from the ensuing analysis. Consequently, after the evaluation, the classification approach was deemed suitable for the remaining 434 data points.

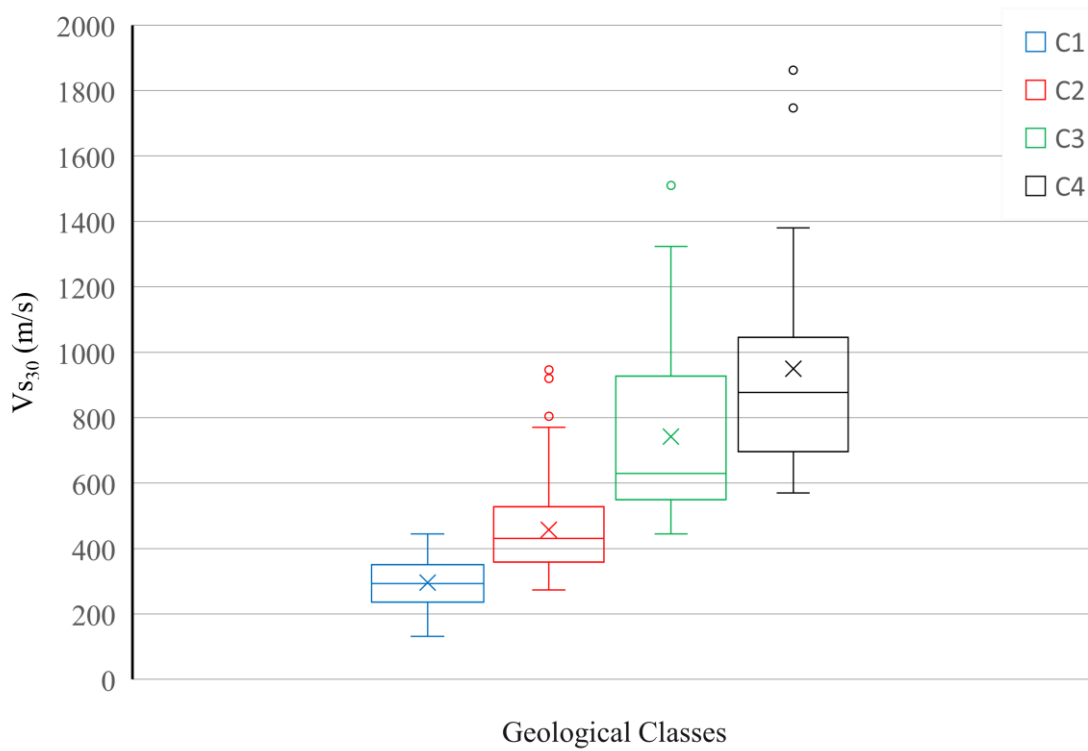


Figure 2.2  $V_{s30}$  data descriptives for four geological period classes (According to studies conducted by members of the project UDAP-Ç-20-01)

## 2.2.2 Topographic Restrictions

Topography is defined as the shapes and attributes characterizing land surfaces. When referencing the topography of a region, it could denote either the actual physical land formations and distinctive features or a portrayal and description captured within maps. In the scope of the current investigation, three specific topographic attributes, slope, surface terrain texture, and curvature, are taken into account as significant topographical parameters. These attributes, in conjunction with geological properties, are utilized to formulate proxy-based regression models for predicting  $V_{s30}$  values.

Within the context of the present research, various mathematical and computational algorithms are available for calculating slopes according to digital elevation model,

DEM. For this study, a rasterized DEM is employed, utilizing square cells (Figure 2.3). Employing the methodology outlined by Travs et al.(1975), the slope is established for the central point, denoted as  $Z_5$ . This is achieved by determining the slope from this central point to each of its eight neighboring points. This calculation involves computing the absolute difference in elevation between  $Z_5$  and each of its neighboring points and subsequently dividing this difference by the size of the cell. The maximum slope among the eight calculated values is then attributed to the cell marked as  $Z_5$  (Equation 2.1). Given that the ultimate slope derived from the method suggested by Travis et al. is the highest value among all the computed slopes, it is appropriately termed as the gradient. Computations for determining the slope at each site were carried out utilizing the MERIT-DEM digitized map. This was accomplished by employing the tools within the "SAGA-GIS" software.

$$S_i = \frac{|H_5 - H_i|}{D} \quad \text{Equation 2.1}$$

where:

$S_i$ : Slope

$H_5$ : Elevation of central point

$H_i$ : Elevation of each neighbor point

$D$ : Size of the cell

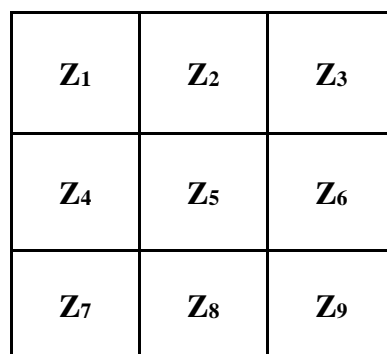


Figure 2.3 Square cells used to calculate slope

Landforms are referred to as intricate amalgamations of diverse surface configurations, constituting a fundamental cornerstone in the realm of geomorphology (Pain, 1985). The classification principle of landforms ought to encompass more than just the categorization of landform types; it should encompass the entirety of spatial arrangements, embodying their diversity. Additionally, it should intricately expound upon the mechanisms driving the formation of these landforms, which undergo transformation due to external forces in a sequence that unfolds over both time and space (Zhang et al., 2020). Geo-informatics graphical methodologies have emerged as prominent analytical techniques, encompassing pivotal methods such as slope spectrum analysis, analysis of profile spectrum and terrain texture. Terrain texture, as an integral facet of landform surfaces, has engendered distinct characteristics in landforms (Shang et al., 2020), evolving in accordance with discernible patterns through the influence of external forces (Li et al., 2015). Terrain texture, delineated with strict precision by both relief ( $Z$ ) and the spacing ( $X, Y$ ) between land features, can be effectively encapsulated through metrics that capture spatial intricacies. These metrics encompass factors such as drainage density in addition to alterations in the direction of slope aspect or curvature within a given area, normalized by unit area (Iwahashi and Pike, 2007). In essence, terrain texture provides insights into the intricacy and complexity of the topography, revealing how the land surface varies in terms of its features and attributes. Considering the research conducted by Iwashashi et al. (2007), the calculation of terrain texture involves the identification of grid cells, informally referred to as "pits" and "peaks." These grid cells delineate the distribution of existing valleys and ridges present within the Digital Elevation Model, DEM.

Slope gradient, gradient, and surface texture, when combined, play a foundational role in the automated classification of steep topographical features. However, their efficacy falls short when it comes to distinguishing between landforms characterized by low relief. To enhance the recognition of such landforms, an additional variable was introduced, positive surface curvature, also known as local convexity (Iwahashi and Pike, 2007). Surface curvature refers to the degree to which a surface deviates from being flat or planar. It quantifies how much a surface is curved or bent at a specific

point. In the context of terrain analysis, surface curvature provides information about the local shape of the land surface. It can be used to identify areas that are convex (curving outward) or concave (curving inward) and to characterize the overall shape of the landscape. Surface curvature is commonly computed by analyzing how the elevation values change around a particular point on a surface.

Surface curvature is gauged through the employment of the Laplacian filter, which is an image-processing technique commonly utilized for edge enhancement. In the context of elevation analysis, this filter approximates the second derivative of elevation, generating positive values in areas exhibiting convex upward curvature, negative values in concave areas, and yielding a value of zero on planar slopes (Equation 2.2). By analyzing the output of the Laplacian filter on elevation data, different landforms based on their curvature characteristics can be identified. This information is particularly useful in terrain analysis and landform classification.

$$\Delta^2(e) = \frac{\partial^2(e)}{\partial x^2} + \frac{\partial^2(e)}{\partial y^2} \quad \text{Equation 2.2}$$

where;

$\Delta^2(e)$ : The Laplacian operator

$\frac{\partial^2(e)}{\partial x^2}$  : The second partial derivative of elevation with respect to the  $x$ -coordinate

$\frac{\partial^2(e)}{\partial y^2}$  : The second partial derivative of elevation with respect to the  $y$ -coordinate

Much like the computation of slope, and gradient, in this case, texture, and local convexity were carried out on the digitized MERIT-DEM map. These calculations were performed using the tools available in the "SAGA-GIS" software to extract the pertinent information regarding texture and the presence of local convexity (Conrad et al., 2015). To optimize the alignment between the topographic and  $V_{s30}$  datasets, maps representing topographic slope, local convexity, and surface texture were systematically generated with varying cell sizes. The most favorable correlation

outcomes with  $V_{s30}$  were achieved from maps created at a cell size of 540 meters (yielding  $R^2$  values of 0.240 for texture and 0.241 for convexity). This determination was reached through a comprehensive linear regression analysis, evaluating the interconnection between  $V_{s30}$  measurements at site locations and the computed topographic attributes. Subsequent analyses were conducted, focusing on convexity and texture parameters, utilizing the specific cell size that demonstrated the highest correlation performance.

## CHAPTER 3

### REGRESSION ANALYSIS

#### 3.1 Introduction

Since determination of  $V_{s30}$  values is an expensive and time-consuming process, several studies were done to develop regression models to predict  $V_{s30}$ 's values. Multiple linear regression analysis was conducted by Matsuoka et al. (2005) to determine the correlation between  $V_{s30}$  values and topographical slope, surficial geology, and combinations of geomorphic parameters. Utilizing global 30 arc sec topographic information, Wald and Allen (2007) developed an alternative approach to correlate  $V_{s30}$  values against topographic slope by developing two sets of parameters, one for active tectonic regions and one for stable shields. Multiple linear regression analyses were evaluated in Japan to study the relationship between the logarithm of observed  $V_{s30}$  and three topographic attributes (Iwahashi et al., 2010). Several empirical regression models were studied using multivariable analysis for the Ilan area and Taipei basin to develop the best-fitting regression model between  $V_{s30}$  and soil index (Kuo et al., 2011). Using a hybrid model, Wald et al. (2011) regressed  $V_{s30}$  for the geologic  $V_{s30}$  medians, slope and cross-term coefficients. Combining both slope and geology. Applying the geostatistical approach of regression kriging, Thompson et al. (2014) constructed a model according to geological and topographic constraints to estimate the  $V_{s30}$  values for California. A power law relationship was fitted to data gathered through Greece by Stewart et al. (2014) to estimate  $V_{s30}$  values for different geomorphic categories based on gradient. After evaluation of previously constructed models, Mc Gann et al. (2017) developed CPT- $V_s$  correlation specific for Christchurch, which was based on multiple linear regression in natural log space. A hybrid geology-slope approach was developed for central and eastern North America using semi log, and log-log regressions (Parker et al., 2017).

By considering only slope values, as an independent variable, a nonlinear regression model was developed according to the dataset gathered through all seismic regions of Iran (Karimzadeh et al., 2019). Considering geomorphological classes, lognormal linear regression models were expressed by Mori et al. (2020) according to normalized slope, convexity, and surface texture for Italy. For the mountainous region of the Iberian Peninsula, a model was constructed as a function of three proxies including slope, geological age, and lithology (Crespo et al., 2022).

## 3.2 Regression

In the preceding section, 3.1, the significance of regression and model development in estimating  $V_{s30}$  values was highlighted through the presentation of previous studies conducted in this domain. Following that, the overall concept of regression and its fundamental components will be elucidated.

Regression analysis primarily aims to discover the functional connection between independent variables,  $X_i$ , and dependent variables,  $Y_i$ , which are related in a nondeterministic fashion to enable predictions or estimations according to detected data. In pursuit of this objective, regression analysis employs statistical techniques to find out the parameters of the developed regression model,  $\beta_i$  and  $\epsilon_i$ . This involves fitting a linear or nonlinear regression model to the available data, with the intention of achieving the closest possible alignment with statistical methodologies (Equation 3.1). Additionally, it helps in perception of how changes of the independent variables affect the dependent variable and provides insights into the strength and direction of the developed relationship.

$$Y_i = f_i(X_i, \beta_i) + \epsilon_i \quad \text{Equation 3.1}$$

where;

$Y_i$ : Dependent variables

$X_i$ : Independent variables

$\beta_i$ : Coefficients

$\epsilon_i$ : Random error term

### **3.3 Regression models classifications**

Various types of regression models are employed in order to elucidate the mathematical relationship between independent and dependent variables. The selection of an appropriate regression model necessitates careful consideration of multiple criteria to effectively capture the possible relationship within the dependent, and independent variables. In this chapter, several common regression classification criteria and different types of regression analysis techniques will be examined, to be considered when determining the type of regression analysis to employ. When evaluating these criteria, it is crucial to prioritize the main properties of the data and the study goal due to their elevated significance.

#### **3.3.1 Regression Models Classification Criteria**

Regression models can be categorized based on different criteria, which offer a framework for classifying these models from various perspectives (Rawlings et al., 1998). This section presents some of the most significant criteria that should be taken into consideration when determining the appropriate type of regression model to uncover the existing relationship between independent and corresponding dependent variables. By introducing these key criteria, we aim to provide a comprehensive understanding of the factors involved in selecting a regression model that best suits the research objectives.

Linearity is a significant criterion in regression classification, differentiating between linear regression, and nonlinear regression models, which allow for more complicated patterns between dependent and independent variables. Considering the linearity assumption is essential to select the appropriate regression model and ensure accurate predictions and reliable inferences (Schroeder., 2016). It is crucial to assess the linearity assumption by examining scatter plots, residual plots, and other diagnostic tools. If a linear relationship does not exist, nonlinear regression models or other techniques may be more appropriate to demonstrate the relationship between datasets.

The other criterion of classification is the count of independent variables, used in the regression construction. By considering the number of independent variables, regression models can be classified into simple regression or multiple regression models. The choice between these models depends on the research goal, the available data, and the complexity of the relationship being studied. Simple regression models are suitable when investigating the effect of a single independent variable, while multiple regression models are employed to detect how the combined effects of multiple independent variables influence the dependent variable. Also, they allow for the analysis of the joint influence of multiple independent variables on the target dependent variable.

The type of the dependent variable is the other criterion used to classify regression models. It refers to the nature or measurement scale of the variable that we are trying to predict or explain using the independent variables (Rawlings et al., 1998). Type of the dependent variable determines the appropriate regression model to be used.

In statistics, generally measured variables can be divided into two main classes including qualitative and quantitative. Quantitative variables, also known as numeric variables, can be measured and described using numbers. On the other hand, qualitative variables provide information about the qualities or characteristics of variables and are often represented using labels or categories (Peck et al., 2008). This broad classification stems from the fact that variables can be either measured or observed as a feature of interest. Within quantitative variables, there are further subdivisions known as discrete and continuous. Discrete variables refer to values that are distinct and separate, typically representing whole numbers or counts. On the other hand, continuous variables represent measurements that can take any value within a range. Qualitative variables, on the other hand, can be classified into two main types: nominal and ordinal. Nominal variables consist of labels or categories without any inherent order. Ordinal variables, instead, represent categories that can be ordered or ranked according to certain criteria (Table 3.1)

Table 3.1 General types of variables

Type of Data	Quantitative	Continuous
		Discrete
	Qualitative	Ordinal
		Nominal

### 3.3.2 Assumptions and Constraints

In the context of regression analysis, assumptions are underlying set of conditions or requirements that must be satisfied through the regression model construction. These assumptions submit basic foundation for the regression model development and guide our interpretation of the results. Constraints, on the other hand, are known limitations that we acknowledge as true and must be taken into account while developing regression models. These constraints act as boundaries or limitations that we need to work around to ensure the validity and reliability of the regression analysis. They can arise from various sources and have an impact on the modeling process or interpretation of the regression results (Rawlings et al., 1998).

Finally, risks in regression analysis represent factors that we are aware of but their occurrence and impact are uncertain. These uncertain factors introduce potential challenges or uncertainties that may influence the accuracy or generalizability of the results presented by developed regression. As part of the analysis, it is important to identify and mitigate these risks to minimize their potential impact on the validity and reliability of the regression analysis. In the development of each regression model, there are specific assumptions and constraints to consider correspondingly. Subsequently, this section introduces several common assumptions that underlie regression analysis.

Model assumptions and constraints must be considered while regression development. Complex real-world challenges require complex models to be built to give out

predictions with utmost accuracy. When developing a regression model, one important issue to consider is the model's form and complexity. Model complexity is a key consideration in regression model development. As models become more complex, they are more likely to overfit the dataset, which means, they may show performance well on the existing dataset but fail to generalize to new data. Commonly, in various model selection criteria, the degree of freedom, number of variables in the model, is often used to determine a model complexity measure (Ye., 1998). More generally, Ye (1998) introduced the generalized degrees of freedom, GDF, as complexity criteria, to decide about the overall sensitivity of the model fit to perturbation of corresponding observed value.

Regression model's complexity, including the number of independent variables and interaction terms, can be a constraint due to limited data, data dispersion, and the model's overfitting or underfitting concerns. Balancing model complexity with the available data is essential to avoid overfitting or underfitting problem. Also, excessive complexity can lead to unreliable results. In linear regression models, which is the simplest regression model, the assumed relationship between the dependent variable and independent variables is linear. This means changes in the dependent variable are directly proportional to corresponding changes in the independent variables. In contrast, nonlinear regression models capture relationships that do not follow a linear pattern (Rawlings et al., 1998.). These models allow for more complex and nonlinear relationships to be constructed between the independent and dependent variables. The relationship between the variables may exhibit curves, bends, or other nonlinear patterns that cannot be adequately represented by a straight line. Through the regression modeling process, it is crucial to assume linearity between the independent and dependent variables when selecting a linear regression model. However, if the developed linear regression model exhibits high residuals, it is advisable to use nonlinear regression models, which is more flexible.

The selection of an appropriate regression model, considering different complexities and adhering to the existing assumptions and constraints, is crucial for achieving an accurate representation of the relationship between independent and dependent

variables. This process is instrumental in making valid inferences based on the available data set. By understanding and selecting the correct model, we can effectively capture the underlying patterns and dynamics of the variables, leading to more reliable and insightful conclusions.

Additionally, there are data assumptions and restrictions that require to be considered through the regression construction. The observations in the data set are assumed to be independent of each other. This assumption, independent variables independency, ensures that the observations are not influenced by each other and that there is no autocorrelation.

Moreover, homoscedasticity assumes that the variability of the dependent variable is constant across all levels of the independent variables (Hoffmann., 2021) . In other words, the spread of the residuals, and errors, is same throughout the range of the predictors. Violations of this assumption, such as heteroscedasticity, can impact the reliability of regression estimates and may require appropriate adjustments or alternative modeling approaches.

In some cases, there are correlations between an independent variable and one or more other independent variables. This phenomenon, multicollinearity, can cause problems while making statistical inferences about the individual effects of independent variables on estimations (Schroeder., 2016). To solve this restriction, the variance inflation factor, VIF, can provide information about which variable or variables are redundant, and thus the variables that have a high VIF can be removed.

Another important assumption to consider in regression analysis is normality. Normality assumes that the observations of the dependent variable and independent variables, in real-world scenarios, tend to follow a normal distribution. It is commonly observed that a significant portion of datasets conform to this distribution. Specifically, in regression analysis, the assumption of normality is applied to the residuals, the differences between the predicted and observed values. It is assumed that these residuals follow a normal distribution. This assumption is fundamental as it allows for the utilization of inferential statistics and hypothesis testing based on the

properties of the normal distribution. However, it is worth noting that in certain regression models, quantile regression, the normal distribution of the dataset is not considered as an assumption during the development of the regression model (Davino et al., 2013.). In such cases, the focus is placed on estimating the conditional quantiles of the dependent variable, rather than assuming normality. Overall, while normality is a commonly made assumption in regression analysis, its applicability may vary depending on the specific regression model being used and the goals of the analysis.

Outliers are defined as abnormal values in a dataset that do not follow the regular distribution of data set (Devore., 2012) and have the potential to significantly distort any regression model. These observations are problematic for many statistical analyses. The presence of influential outliers exerts influence on the regression model and can impact the estimates and assumptions of regression analysis. To ensure accurate predictions, it is important to assume that these extreme observations do not have a strong influence on the ultimate analysis (Hao., 2007). If necessary, certain studies suggest that these extreme observations can be removed from the dataset. However, it is essential to apply caution when removing outlier observations and carefully consider the potential impact on the overall analysis.

Sample size serves as an important constraint in regression analysis (Schroeder., 2016). The size of the sample, the number of observations available for analysis, not only impact the reliability and precision of regression estimates but also affects the inferential statistic. For example, for a stated probability, the t-statistic depends on the degree of freedom, number of sample size minus the number of coefficients estimated. A smaller sample size may lead to increased uncertainty and wider confidence intervals, making it more challenging to draw meaningful conclusions from the analysis. With a limited sample size, there is a higher likelihood of sampling variability, where the observed data may not fully represent the population from which it was sampled. This can affect the generalizability of the results and the ability to make accurate predictions. Additionally, sample size affects the statistical power of the analysis. In regression analysis, larger sample sizes that represent the target population properly provide more statistical power, allowing for the detection of

smaller effects and more precise estimation of regression coefficients. Insufficient sample sizes may result in biased or unreliable estimates, while larger sample sizes can enhance the robustness and validity of the findings.

The other constraint which can be faced through regressing is corresponding to data quality. Data quality is generally recognized as a multidimensional concept. While no single definition of data quality has been accepted by researchers working in this area, there is agreement that data accuracy, currency, completeness, and consistency are important areas of concern. The reliability and accuracy of the data used for regression modeling are crucial for obtaining meaningful and valid results. Data quality constraints can arise from various factors, including errors, missing values, outliers, or inconsistencies within the dataset.

### **3.3.3 Purpose of Regression Analysis**

The specific purpose or objective of the regression analysis helps determine the appropriate regression technique to consider all criteria (Devore., 2012). In general, regression models are often used for predictive purposes, aiming to estimate or forecast the value of the dependent variable based on the independent variables. The primary focus is on developing a model that can accurately predict the outcome variable for new or unseen data. Techniques such as linear regression, polynomial regression, or machine learning algorithms like random forest regression or support vector regression can be employed for predictive modeling.

Although regression analysis is primarily used for predicting continuous numerical values, there are some instances where regression techniques can be adapted for classification problems (Hosmer et al., 2000). These adaptations are known as regression for classification or regression-based classification methods. Logistic regression and support vector regression, SVR, methods are two common regression-based classification approaches utilized to categorize the dataset into two or more

classes. Numerous goodness-of-fit methods guide researchers in selecting the appropriate regression model based on regressing purposes, continuous dependent value prediction, and discrete dependent classification.

By carefully considering the purpose or objective of the analysis, researchers can choose the most suitable regression technique or approach that aligns precisely with their specific goals. In the process of selecting regression models, a complete understanding of the dataset's characteristics and adherence to relevant assumptions are essential for obtaining reliable and interpretable results.

### **3.3.4 Types of Regression Models**

Within the traditional regression framework, standard regression models are accompanied by certain assumptions that must be fulfilled before conducting the regression analysis. These models exhibit distinctive characteristics. Firstly, attention must be given to the statistical properties of the data and the relationship between variables, encompassing linearity, independence, homoscedasticity, and normality. Secondly, the formulation or methodology of the model, including the functional form, parameter constraints, and distributional assumptions, should be taken into account. Lastly, the primary objective of the study holds the utmost significance in determining the appropriate regression model type. The remainder of this section will be dedicated to exploring distinct types of regression models. We will examine each model briefly, considering how they align with the aforementioned fundamental criteria. By delving into these various regression models, we aim to gain a comprehensive understanding of their characteristics and suitability for different research scenarios.

The linear regression model is the simplest form of regression model through which continuous dependent variables are related to one or more independent variables using a linear model. The goal of linear regression is to estimate the values of the coefficients that best fit the data, which minimizes the sum of squared differences between the

actual observed values and the predicted values based on the linear equation (Equation 3.2). In a simple linear regression model, we examine the relationship between one independent variable and one dependent variable. On the other hand, when the relationship involves multiple independent variables and one dependent variable, it is common to use multiple linear regression. The most common method for estimating the coefficients in linear regression is the ordinary least squares method, OLS. It finds the values of  $\beta_0, \beta_1, \dots, \beta_n$  that minimize the sum of squared residuals, and differences between observed and predicted values (Equation 3.3).

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n + \epsilon_i \tag{Equation 3.2}$$

$$\min_{\beta_0, \beta_1, \dots, \beta_n} \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n)]^2 \tag{Equation 3.3}$$

where:

$Y_i$ : Dependent variable

$X_1, X_2, X_3, \dots, X_n$ : Independent variables

$\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_n$ : Coefficients

$\epsilon_n$ : Random error term

$n$ : Number of observations

In linear regression models, the expectation of the random error, also referred to as the mean value of the random error, is assumed to be zero (Equation 3.4). No or little multicollinearity, homoscedasticity, and observation independence are the main assumptions that must be considered while developing simple linear or multiple linear regression models. Violations of any of these assumptions produce undesirable properties in the results obtained when regression coefficients are estimated without attention for these assumptions.

$$E(\epsilon_i) \equiv E(\epsilon_i|X_i)=0$$

Equation 3.4

where:

$E(\epsilon_i)$ : Expected value of random error term

$E(\epsilon_i|X_i)$ : Expected value of random error term given the independent variable

Similar to different types of linear regression models, which have been briefly explained, there are other models based on the linear form of regression that can be developed based on specific restrictions, assumptions, and the purpose of the regression model. However, since these models are not within the scope of this research, they will not be discussed here. It is essential for researchers to explore and select the appropriate linear or regression models that align with their specific objectives and data characteristics to achieve accurate and meaningful results.

Nonlinear regression is a type of regression analysis used when the relationship between the dependent variable and the independent variables cannot be adequately described by a linear model. Unlike linear regression, which assumes a linear relationship between dependent and independent variables, nonlinear regression allows for more complex, nonlinear relationships (Equation 3.5).

$$Y_i = f(\beta_0, \beta_1, \dots, \beta_n, X_1, X_2, \dots, X_n) + \epsilon_n$$

Equation 3.5

where:

$Y_i$ : dependent variable

$X_1, X_2, X_3, \dots, X_n$ : independent variables

$\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_n$ : Coefficients

$\epsilon_n$ : Random error term

$n$ : Number of observations

$f(\cdot)$ : A nonlinear function that defines the relationship between the variables.

Nonlinear regression models can take various forms, depending on the specific problem and the underlying relationship between the variables. Some common examples of nonlinear functions include polynomial functions, exponential functions, logarithmic functions, trigonometric functions, and power functions. The process of estimating the coefficients in nonlinear regression involves minimizing the sum of squared residuals. Unlike linear regression, there is no closed-form solution for the coefficients in most nonlinear regression models, so numerical optimization techniques such as gradient descent, the Levenberg-Marquardt algorithm, and genetic algorithms are used to find the best-fitting parameters (Gavin, 2022).

As it was stated before regression models are classified according to various criteria. Linear and nonlinear regression models are two main categories of regression models that are classified based on the assumption of the functional form of the relationship between the dependent variable and the independent variables. This classification falls under the criterion of linearity.

Also, as mentioned previously, one of the criteria for classifying regression models is their purpose through the analysis. Quantile regression is a notable example of this classification based on the intended goals of the regression. In the forthcoming chapters, this model will be elaborated in detail, exploring its unique characteristics and applications in  $V_{s30}$  values prediction.

Regression models can also be used as classifiers in statistical processes. Instead of predicting continuous numeric values as in traditional regression, expressed in the previous section, regression classifiers predict discrete categorical labels, making them suitable for classification tasks. The predicted label can represent different classes or categories, such as binary or multi-class. Some regression techniques, such as logistic and support vector regression can be adapted to perform classification by using a threshold or decision boundary to assign class labels. In the future chapter, logistic regression and its application in determining categorical  $V_{s30}$  values will be extensively explored. Logistic regression is a widely used statistical technique for modeling the

relationship between a binary or categorical dependent variable and one or more independent variables. Its relevance and utility in predicting categorical  $V_{s30}$  values will be thoroughly discussed, shedding light on its significance in geotechnical and seismic studies.

Additionally, orthogonal regression is a method used when the main purpose is to fit a line to data points, but unlike regular regression, it takes into account errors in both the x and y directions. This makes it useful when both variables are prone to measurement errors. This method was applied by Castellaro et al. (2008) to re-analyze the empirical bases of the correlation between  $V_{s30}$  and seismic amplification.

Non-parametric regression is the other type of regressing model through which no specific form is defined for the relationship between the dependent and independent variables. Multivariate Adaptive Regression Splines, MARS, is one of the flexible non-parametric regression development methods. Recently, this regression method was used to assess and predict the trend in mean kappa values for the Aegean Sea region in Türkiye (Kurtulmus et al., 2023). Conic Multivariate Adaptive Regression Splines, CMARS, is an extended version of MARS, which leads to smoother and more accurate fits. Using dataset from Türkiye and applying CMARS, research was conducted to develop GMPEs, Ground Motion Prediction Equations (Yerlikaya-Ozkurt et al., 2014). Unsupervised learning algorithms learn patterns from unlabeled data, when the dependent variables are not available. Gaussian Mixture Model, GMM, is one of the common unsupervised methods. It is a probabilistic model that assumes that the data is generated from a mixture of several Gaussian distributions with different means and variances. By using GMM sites can be clustered according to topographical and geological properties. Through the post-clustering analysis  $V_{s30}$  values can be estimated using empirical correlations based on the properties of each cluster.

### 3.4 Nonlinear Regression Models to Estimate $V_{s30}$ Values in Türkiye

Concurrently with the ongoing study, another research was undertaken to construct nonlinear regression models aimed at predicting  $V_{s30}$  values using the same dataset. In the course of this concurrent study, four distinct regression models were developed individually for each geological class,  $C_1$ ,  $C_2$ ,  $C_3$ , and  $C_4$ . According to these nonlinear regression models, for geological classes  $C_1$  and  $C_2$ , there is a good correlation between  $V_{s30}$  and topographic slope, gradient, and values (Equation 3.6 and Equation 3.7). On the other hand, for geological class  $C_3$  using slope gradient, convexity, and terrain surface texture submits better result. For geological class  $C_4$  slope gradient and curvature demonstrate better results (Equation 3.8 and Equation 3.9) (Sahin et al., 2024). The developed nonlinear regression models resulted in an  $R^2$  of 0.601.

$$Y = 350 \times 180 / ((350 - 180) \times \exp(-1.1943 \times g) + 180) \quad \text{Equation 3.6}$$

$$Y = 572 \times 273 / ((526 - 273) \times \exp(-0.225 \times g) + 273) \quad \text{Equation 3.7}$$

$$Y = 445 + (3.041 \times g) + 0.21 \times t \times c \quad \text{Equation 3.8}$$

$$Y = 570 + (3.73 \times c) + (11.58 \times g) \quad \text{Equation 3.9}$$

where:

$Y$ : Dependent variables,  $V_{s30}$

$g$ : Slope, gradient

$c$ : Convexity

$t$ : Terrain surface texture

The findings of that study are also considered for model development.

## CHAPTER 4

### QUANTILE REGRESSION

#### 4.1 Introduction

The quantile regression model, QRM, is a statistical tool which can be used to predict conditional quantiles of a dependent variable given independent variables. This method was introduced by Koenker and Bassett (1978), as an extension of the linear regression model, LRM, to be used when the conditions of linear regression are not met. Unlike regular linear regression method, which use the least squares technique to estimate the conditional mean of the dependent variable across different independent variables, quantile regression presents a more comprehensive picture of the correlation between the independent and different conditional quantiles of the dependent variables. This type of regression model finds applications in many fields including economics, health sciences, education, and engineering science due to its ability to provide valuable insights into different parts of the conditional distribution of the response variable.

In a significant advancement of the quantile regression model, Powell (1984, 1986) offered the censored quantile regression model. This type of quantile regression model provides a robust estimation of conditional quantiles in situations where observations on the dependent variable are censored. A recent collection of studies provides a comprehensive and highly informative examination of the quantile regression model in econometrics. Moreover, Fitzenberger et al. (2022) demonstrated the extensive applicability of quantile regression in economics through the illustration of its application in previous studies. Their work demonstrated how this method can be effectively employed to analyze diverse economic problems, making it a valuable tool for researchers and practitioners seeking to explore different aspects of economic relationships and outcomes.

Buhai (2005) demonstrated the successful implementation of the quantile regression technique in the selected applications of survival analysis and recursive structural models, models that study the causal relationship between variables. Somers (Somers and Whittaker, 2007) utilized quantile regression in retail credit risk evaluation to demonstrate the power of this model in the diverse distributions that arise in the financial service industry.

In addressing real-world health challenges, the associations between exposure and outcome variables can often be intricate and multifaceted. outcome variables can often be intricate and multifaceted. To navigate these complexities effectively, quantile regression models offer an ideal approach. In their work, Wei et al. (2019) introduced a diverse set of quantile regressing methods that prove invaluable for epidemiological studies, including the classical linear quantile regression, nonparametric quantile regression for growth trajectories, and quantile regression models for case-control design. Furthermore, quantile regression models are used in the field of genetics studies, large-scale analyses of various biological molecules or components within a cell or organism. Briollais and Durrieu (Briollais and Durrieu, 2014) conducted a comprehensive review of the application of quantile regression models in the field of -omics (a discipline in biology). In their study, Staffa et al. (Staffa et al., 2019) employed quantile regression models to investigate the relationship between ventilator dependence through 48 hours of surgery and total hospital length of stay. Additionally, they explored the relationship between physical status and total blood loss per kilogram using quantile regression.

In Japan, the occurrence of shallow landslides due to heavy rainfall presents a significant geotechnical challenge. To address this issue, empirical rainfall intensity and duration, I-D, threshold models have been constructed to identify the triggering of shallow landslides during rainfall events. Utilizing the quantile regression method and analyzing data from 1174 rainfall-induced shallow landslides, Saito et al. (2010) successfully developed new threshold models tailored specifically for Japan. In the context of coal resource estimation, the machine learning based quantile regression forest algorithm was utilized. To gauge the performance of the quantile regression

forest algorithm, the results were compared against those obtained from the inverse distance weighting and regression kriging methods. The study revealed that the model combining quantile regression and random forest methods outperformed the other approaches in terms of accuracy, bias reduction, and precision (Maxwell et al., 2021). To determine uncertainty related to the high variability of Mohr–Coulomb’s shear strength parameters, statistical analysis conducted based on quantile regression to investigate rainfall-induced shallow landslides in ash-fall pyroclastic soils resulting from the explosive activity of the Somma-Vesuvius volcano in southern Italy (Tufano et al., 2021). A hybrid computational intelligence approach, kernel-based support vector machine quantile regression, was utilized by Ma et al. (2020) to predict reservoir landslides in China. Considering different percentiles as warning levels and applying quantile regression methods, Kang and Kim proposed warning criteria for landslides resulting from rainfalls (Kang & Kim, 2016). Considering 92 landslides caused by rainfalls in the southern part of Thailand and using quantile regression Salee et al. (2022) proposed a landslide-triggering rainfall threshold. In South Korea, the effect of different geological conditions on shallow landslides triggered by rainfalls was discussed through the quantile regression analysis (Lee et al., 2022). To overcome the missing data at some depth intervals in well logs, which are used by geoscientists to infer and extrapolate the physical properties of subsurface rocks, machine learning algorithms were used by Feng et al. (Feng et al., 2021) also to quantify the uncertainty in the prediction’s quantile regression tree, combination of quantile regression and decision tree was used to determine prediction intervals. Bozorgzadeh and Harrison (2015) employed the quantile regression method to investigate and establish a criterion that represents the characteristic triaxial strength of intact rock. Moreover, a study focused on examining the scour around river-crossing bridges in relation to foundation width and flood conditions and estimations were made employing the quantile regression method. (Wang, 2021). Chen et al. (2022) considered long-term and complex deformation in reservoir bank slope using quantile regression and developed a fundamental tool for prediction and early caution of slope deformation. By utilizing quantile regression, in South Korea, the effect of antecedent rainfall conditions on

shallow landslides triggered by short-term high-intensity rainfall and long-term low-intensity rainfall was studied to determine the I-D threshold (Kim et al., 2021).

Applying quantile regression, Zijl et al. (2014) studied the combined effect of soil properties that govern gully erosion in a catchment located in Lesotho, Africa. In order to predict the subsurface properties around the well logs besides to machine learning model, random forests, RF, Sankaranarayanan, et al. (2021) utilized a quantile regression model to evaluate the uncertainty in the predicted properties in the form of prediction intervals. In the petroleum industry in order to optimize the drilling operations and for penetration rate improvement, an efficient predictive model of the ROP, rate of penetration, as a function of key drilling parameters is necessary. Ambrus et al. (2022) applied quantile regression deep neural networks to the penetration rate prediction problem, through which quantile regression models perform probabilistic forecast of ROP for a given range of drilling parameters.

## 4.2 Quantile Regression

In the real application of regression analysis, the dependent variable cannot be predicted exactly from independent variables, and for fixed values of independent variables the regression models submit random dependent variables (Devore., 2012). To this end, measures of central tendency such as mean, median, or mode are used to summarize the behavior of dependent variables according to the fixed value of independent variables. The idea of developing a regression model that describes the mean of the dependent variable using fixed independent variable values, conditional-mean, is the core of a broad traditional family of regression-modeling approaches. While conditional-mean models possess certain appealing characteristics, they also come with inherent limitations (Hao and Naiman, 2007). Firstly, these models cannot be extended to noncentral locations, and for research regarding noncentral locations on the dependent variables distribution they demonstrate inefficiency. The second constraint arises from the model assumptions, which often do not align with real-world scenarios. Assumptions such as homoscedasticity may not hold, and instances of heavy-tailed distributions and the presence of outliers within the dataset can lead to misleading outcomes. Thirdly, conditional-mean models inherently lack the capacity to effectively capture the intricate relationship between variations in the independent variables and the shape of the dependent variable's distribution. Conditional-median regression, or simply median regression, is an alternative to conditional-mean regression. In this method least-squares estimation is replaced by least-absolute distance estimation, which demands greater computing power (Koenker and Bassett 1978). Given that interpreting the mean can be challenging in the case of highly skewed distribution and the median remains highly informative, therefore, conditional-median regression models have the potential to be more useful.

Conditional-median regression is a special case of the quantile regression model, the conditional 0.5th quantile by relating it to covariates through a functional relationship. More generally, other quantiles can be used to describe the noncentral position of a distribution.

### 4.2.1 Quantile Regression Model and Estimation

The median regression model estimates the effect of a covariate on the conditional median, so it represents the central location even when the distribution is skewed, where the mean might be affected by extreme values.

Through the linear regression model development ( Equation 4.4) it is assumed that the expected value of the error term is zero ( Equation 4.2) and conditional mean, expected value, of dependent variables,  $Y_i$ , according to independent variables,  $X_i$ , are expressed as Equation 4.3.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{Equation 4.1}$$

$$E (\epsilon_i) = 0 \quad \text{Equation 4.2}$$

$$E (Y_i|X_i) = \beta_0 + \beta_1 X_i \quad \text{Equation 4.3}$$

where:

$Y_i$ : Dependent variables

$X_i$ : Independent variables

$\beta_0$  and  $\beta_1$ : Coefficients of linear regression model

$\epsilon_i$ : Random error term

$E (\epsilon_i)$ : The expected value of the error term

$E (Y_i|X_i)$ : Expected value of dependent variables according to independent variables

The quantile regression model, which is an extension of the linear regression model, was introduced by Koenker and Bassett (1978). In fact, this model is the general form of the median-regression model that estimates the potential differential effect of a covariate on various quantiles in the conditional distribution. The  $p$ th quantile signifies the value of the dependent variable below which a proportion of  $p$  within the population lies. Consequently, quantiles have the ability to specify various positions within a distribution. The quantile regression model corresponding to the linear regression model with one independent variable can be expressed by Equation 4.4. For the corresponding quantile regression model the  $p$ th conditional quantile, given  $X_i$ , is specified by Equation 4.5. Also, the  $p$ th quantile of the error term is zero (Equation 4.6).

$$Y_i = \beta_0^p + \beta_1^p X_i + \epsilon_i^p \quad \text{Equation 4.4}$$

$$Q^{(p)}(Y_i|X_i) = \beta_0^{(p)} + \beta_1^{(p)} X_i \quad \text{Equation 4.5}$$

$$Q^{(p)}(\epsilon_i) = 0 \quad \text{Equation 4.6}$$

Where:

$0 < p < 1$ : proportion of the population with dependent values below the quantile at  $p$

$Y_i$ : Dependent variables

$X_i$ : Independent variables

$\beta_0^p$  and  $\beta_1^p$ : Quantile specific coefficients of quantile regression model

$\epsilon_i^p$ : Quantile specific random error term of quantile regression model

$Q^{(p)}(Y_i|X_i)$ : Quantile regression for  $p$ th conditional quantile

$Q^{(p)}(\epsilon_i)$ : Error term for the  $p$ th quantile

After discussing the quantile regression model, it's important to delve into the estimation of model coefficients. When estimating the coefficients of a linear regression model, the common approach involves minimizing the sum of squared vertical distances between data points  $(X_i, Y_i)$  and the fitted line. This method is known as ordinary least squares, OLS, and it aims to find the best-fitting line that minimizes the squared differences between the observed values,  $Y_i$ , and the predicted values on the line (Equation 4.7). Through the linear regression model development, if the assumptions are correct, the fitted linear model approaches the population conditional mean as the sample size goes to infinity.

$$\min \sum_i (Y_i - (\beta_0 + \beta_1 X_i))^2 \quad \text{Equation 4.7}$$

A closed-form solution to the minimization problem is obtained by taking partial derivatives of Equation 4.7 with respect to the coefficients and setting each partial derivative equal to zero. Subsequently, we solve the resulting system of two equations with two unknowns. This process leads us to the derivation of the two estimators.

A notable distinction between the quantile regression model estimator and the linear regression model estimator lies in how distances of data points from a line are evaluated. In quantile regression models, the point-to-line distances are gauged using a weighted sum of vertical distances, without applying the squaring operation. The weight assigned to each distance is  $(1-p)$  for points situated below the fitted line and  $p$  for points located above the line. Each choice for this proportion  $p$  gives rise to a different fitted conditional-quantile function. The task is to find coefficients, and estimators, with the desired property for each possible  $p$ . To this end, quantile regression model coefficients are defined as the values that minimize the weighted sum of distances between fitted and actual dependent values (Equation 4.8). In linear regression models, the mean of a distribution can be considered as the point that minimizes the average squared distance over the population, whereas in quantile regression models a quantile  $q$  can be viewed as the point that minimizes an average

weighted distance with weights depending on whether the point is above or below the value  $q$ .

$$\min \sum_{i=1}^n d_p(Y_i, \hat{Y}_i) \quad \text{Equation 4.8}$$

$$= \min \left[ p \sum_{Y_i \geq \beta_0^{(p)} + \beta_1^{(p)} X_i} |Y_i - \beta_0^{(p)} - \beta_1^{(p)} X_i| \right. \\ \left. + (1 - p) \sum_{Y_i < \beta_0^{(p)} + \beta_1^{(p)} X_i} |Y_i - \beta_0^{(p)} - \beta_1^{(p)} X_i| \right]$$

where:

$\hat{Y}_i$ : Estimated dependent value

$d_p$ : Weighted distance

#### 4.2.2 Goodness of Fit

In regression analysis, the concept of goodness of fit serves as a crucial measure to assess how well a chosen regression model aligns with the observed data. It reflects the degree to which the model's predicted values closely match the actual observed values. A high goodness of fit indicates that the model captures the underlying relationships between the variables effectively, leading to accurate predictions. Various statistical metrics, such as the coefficient of determination, R-squared, root mean squared error, RMSE, and mean absolute error, MAE, are employed to quantify the goodness of fit. Commonly, in linear regression models, the goodness of fit is measured by the R-squared method (Equation 4.9). This quantity ranges from 0 to 1, with a larger value of  $R^2$  indicating a better model fit.

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} \quad \text{Equation 4.9}$$

where:

$R^2$ : R-squared

$\hat{Y}_i$ : Estimated dependent value

$\bar{Y}$ : Sample mean of real dependent variables

Analogous goodness of fit criterion to the  $R^2$  can be developed for quantile regression models, which is known as pseudo  $R^2$ . While the linear regression model fits according to the least-squares criterion, quantile regression models are founded on the principle of minimizing a sum of weighted distances (Hao and Naiman, 2007). To this end, by comparing the sum of weighted distances for the model of interest with the sum in which only the intercept parameter appears, goodness of fit can be measured for quantile regression (Koenker and Machado, 1999) (Equation 4.10).

$$R(p) = 1 - \frac{V^1(p)}{V^0(p)} \quad \text{Equation 4.10}$$

where:

$$\begin{aligned} V^1(p) &= \sum_{i=1}^n d_p(Y_i, \hat{Y}_i) && \text{Equation 4.11} \\ &= \sum_{Y_i \geq \beta_0^{(p)} + \beta_1^{(p)} X_i} p |Y_i - \beta_0^{(p)} - \beta_1^{(p)} X_i| \\ &+ \sum_{Y_i < \beta_0^{(p)} + \beta_1^{(p)} X_i} (1-p) |Y_i - \beta_0^{(p)} - \beta_1^{(p)} X_i| \end{aligned}$$

$$\begin{aligned}
V^0(p) &= \sum_{i=1}^n d_p(Y_i, \hat{Q}^{(p)}) && \text{Equation 4.12} \\
&= \sum_{Y_i \geq \beta_0^{(p)} + \beta_1^{(p)} X_i} p |Y_i - \hat{Q}^{(p)}| \\
&+ \sum_{Y_i < \beta_0^{(p)} + \beta_1^{(p)} X_i} (1-p) |Y_i - \hat{Q}^{(p)}|
\end{aligned}$$

$V^1(p)$ : Sum of weighted distances for the full  $p$ th quantile-regression model

$V^0(p)$ : Sum of weighted distance for the model that includes only a constant term

$\hat{Q}^{(p)}$ : Fitted constant which is the sample  $p$ th quantile  $\hat{Q}^{(p)}$  for the sample dependent variable,  $Y_1, Y_2, Y_3, \dots, Y_n$

It should be considered that, since the  $V^0(p)$  and  $V^1(p)$  are nonnegative,  $R(p)$  is at most 1. Also, because the sum of weighted distances is minimized for the full-fitted model,  $V^1(p)$  is never greater than  $V^0(p)$ , so  $R(p)$  is greater than or equal to zero.

Also, the adjusted  $R^2$ , an extended form of  $R^2$  or pseudo  $R^2$ , serves as a refined measure of goodness of fit in regression analysis. This goodness of fit criteria considers the complexity of the model concerning the number of predictor variables. While  $R^2$  indicates the proportion of variance in the dependent variable explained by the model, adjusted  $R^2$  takes into account the number of predictors used in the model, penalizing the addition of unnecessary variables that might overfit the model.

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - v - 1} \quad \text{Equation 4.13}$$

where:

$R^2$ :  $R$  squared or pseudo-  $R$  squared

$N$  : Total sample size

$v$  : number of independent variables

### **4.3 Quantile Regression Model to Estimate $V_{s30}$ Values in Türkiye**

In a previous chapter, the discussion emphasized the significance of developing distinct regression models tailored to specific purposes based on statistical data structure, along with data and model constraints and underlying assumptions. This section of the ongoing study extends these principles by introducing a novel approach, Nonlinear Categorical Quantile Regression Model, NCQRM. This new model is designed to accommodate a dataset encompassing a blend of quantitative continuous variables, as well as qualitative categorical variables. In this context, the independent variables include four geological classes, as geological properties of sites, in addition to terrain surface texture and gradient values, representing topographical aspects of sites, at which  $V_{s30}$  values are submitted by AFAD.

#### **4.3.1 Nonlinear Categorical Quantile Regression Model**

The inclination of topography, often referred to as the slope or gradient, holds the potential to provide valuable insights into the nature of  $V_{s30}$  values. This is attributed to the fact that materials possessing higher competence and velocity tend to exhibit steeper slopes. Conversely, environments characterized by exceptionally low gradients typically facilitate the deposition of deep basin sediments (Wald and Allen, 2007). This connection underscores the diagnostic capacity of slope, or gradient, in assessing  $V_{s30}$  variations. By comprehending the intricate interplay between gradient and  $V_{s30}$  values, it becomes evident that the selection of a nonlinear functional form (Equation 4.14) for the quantile regression model is of paramount importance. This choice is driven by the aim to establish a model that aligns with theoretical expectations, in regions characterized by elevated gradient values the model should predict correspondingly higher  $V_{s30}$  values. This thoughtful selection of a nonlinear functional form allows for capturing the intricate relationships between these variables, gradient, and  $V_{s30}$  values, and leveraging them to yield more accurate and realistic estimations, reflecting the anticipated theoretical associations between gradient and  $V_{s30}$ .

$$Y_i = a_0 \cdot a_1^{X_i} \quad \text{Equation 4.14}$$

where:

$Y_i$ : Dependent variable

$X_i$ : Independent variable

$a_0, a_1$ : Coefficients of nonlinear increasing function

Employing the STATA software, a linear categorical quantile regression model was constructed, focusing on the 16th percentile. However, considering the ultimate objective of crafting an ascending nonlinear categorical quantile regression model, certain preparatory steps are indispensable. This entails initially subjecting the dataset to transformations to render it compatible with the STATA software's requirements for formulating a linear categorical quantile regression model. Upon deriving the pertinent coefficients for the linear categorical quantile regression model, the subsequent step necessitates an additional transformation. These coefficients must be suitably converted to align with the structure of the upcoming nonlinear quantile regression model. This transformation is essential to ensure that the insights gleaned from the linear model can be seamlessly integrated into the nonlinear model's framework, ultimately facilitating the development of a robust, ascending nonlinear categorical quantile regression model.

Through this ongoing research endeavor, a comprehensive approach was taken to develop nonlinear categorical quantile regression models utilizing STATA software. These models were meticulously constructed by incorporating four distinct geological classes as fundamental geological properties, coupled with terrain surface texture and gradient, as topographical attributes. The formulation of these models also integrated interaction terms to encapsulate complex interactions within the dataset (Equation 4.15). In the pursuit of the robust model with fewer terms assessment, the research employed both pseudo  $R^2$  and adjusted  $R^2$  as a criterion for evaluating goodness of fit (Table 4.1). Notably, to decrease model terms, terms exhibiting inconsistency within

the regression model were discerningly disregarded. These excluded terms, while not profoundly influencing the pseudo  $R^2$  and adjusted  $R^2$ , were deemed unnecessary.

$$V_{s30} = a_0 \cdot a_1^g \cdot a_2^t \cdot a_3^{g.t} \cdot a_4^{C_2} \cdot a_5^{C_3} \cdot a_6^{C_4} \cdot a_7^{C_2.g} \cdot a_8^{C_3.g} \cdot a_9^{C_4.g} \cdot a_{10}^{C_2.t} \cdot a_{11}^{C_3.t} \cdot a_{12}^{C_4.t} \cdot a_{13}^{C_2.g.t} \cdot a_{14}^{C_3.g.t} \cdot a_{15}^{C_4.g.t} \quad \text{Equation 4.15}$$

where:

$V_{s30}$ :  $V_{s30}$  values at sites

$g$ : Gradient at sites

$t$ : Terrain surface texture

$C_2, C_3, C_4$ : Geological classes

$a_0, a_1, a_2, \dots, a_{15}$ : Nonlinear categorical quantile regression model coefficients

Table 4.1 Models with different numbers of terms vs goodness of fit criteria

NO. of terms in the model	Pseudo $R^2$	Adj. Pseudo $R^2$
16	0.386	0.364
15	0.379	0.358
14	0.378	0.358
...	...	...
7	0.367	0.358
6	0.361	0.353
5	0.350	0.343
4	0.272	0.267

Taking into account the recorded  $R^2$  and adjusted  $R^2$  values presented in Table 4.1, along with the observed behavior of the models (Figure 4.1), as a result, the nonlinear categorical quantile regression model characterized by five distinct terms (Equation 4.16) emerges as the most fitting and suitable option for estimating  $V_{s30}$  values. This selection is underscored by its remarkable adherence to theoretical expectations, sites with higher gradients demonstrating higher  $V_{s30}$  values, and its consistent ability to encapsulate the intricate interactions between gradient and  $V_{s30}$  values.

$$V_{s30} = a_0 \cdot a_1^g \cdot a_2^{C_2} \cdot a_3^{C_3} \cdot a_4^{C_4} \quad \text{Equation 4.16}$$

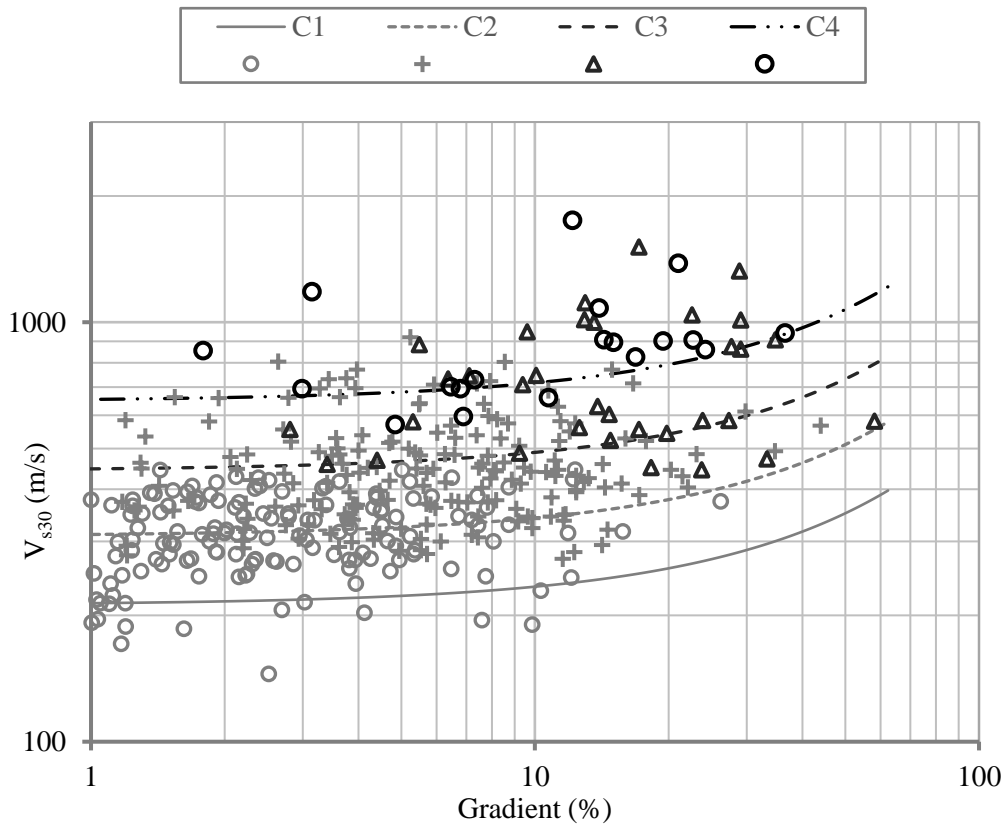


Figure 4.1  $V_{s30}$  (m/s) vs Gradient (%) for all geological classes according to Equation 4.16

### 4.3.2 Nonlinear Categorical Quantile Regression Models with Interaction Terms

When interaction terms are included in a regression model, they allow us to account for the fact that the impact or influence of one independent variable on the dependent variable changes based on the values of another independent variable. In essence, interaction terms in regression recognize that the relationship between two independent variables is not fixed or uniform across all scenarios, and they enable the model to incorporate these nuanced and context-dependent effects. In order to evaluate the combination effect of gradient or texture with each geological class on  $V_{s30}$  values, corresponding interaction terms are used while developing nonlinear categorical quantile regression models (Equation 4.17 and Equation 4.18)

$$V_{s30} = a_0 \cdot a_1^g \cdot a_2^{C_2} \cdot a_3^{C_3} \cdot a_4^{C_4} \cdot a_5^{C_2 \cdot g} \cdot a_6^{C_3 \cdot g} \cdot a_7^{C_4 \cdot g} \quad \text{Equation 4.17}$$

$$V_{s30} = a_0 \cdot a_1^g \cdot a_2^{C_2} \cdot a_3^{C_3} \cdot a_4^{C_4} \cdot a_5^{C_2 \cdot t} \cdot a_6^{C_3 \cdot t} \cdot a_7^{C_4 \cdot t} \quad \text{Equation 4.18}$$

Considering Equation 4.17 and the corresponding goodness of fit criteria,  $R^2 = 0.36$  and adjusted  $R^2 = 0.35$ , it becomes evident that the incorporation of interaction terms, gradient with geological class, in a complex regression model does not yield a significant enhancement in model performance. This notion is further substantiated when examining the coefficient values of the interaction terms (Table 4.2) which exhibit proximal proximity to unity. This proximity to one signifies that these interaction terms manifest a neutral role in the estimation of  $V_{s30}$  values.

Table 4.2 Model coefficients value

Model Coefficient	Model Coefficient's value
$a_0$	207.81
$a_1$	1.02
$a_2$	1.44
$a_3$	2.27
$a_4$	2.74
$a_5$	0.99
$a_6$	0.98
$a_7$	0.99

Aligning with theoretical fundamentals, it's anticipated that geological classes characterized by greater stiffness would correspondingly yield higher  $V_{s30}$  values, shown by the patterns evident in Figure 4.1. However, upon examining Figure 4.2, predicated on Equation 4.17, a discordant trend emerges. In contrast to the initial assumption, the model plotted in Figure 4.2 reveals that geological classes  $C_1$  and  $C_2$  exhibit elevated  $V_{s30}$  values compared to inherently stiffer geological classes, notably for certain higher gradient values. This juxtaposition between the theoretical expectation and the model's outcomes underscores that the utilization of interaction terms within Equation 4.17 does not effectively enhance the model's behavior.

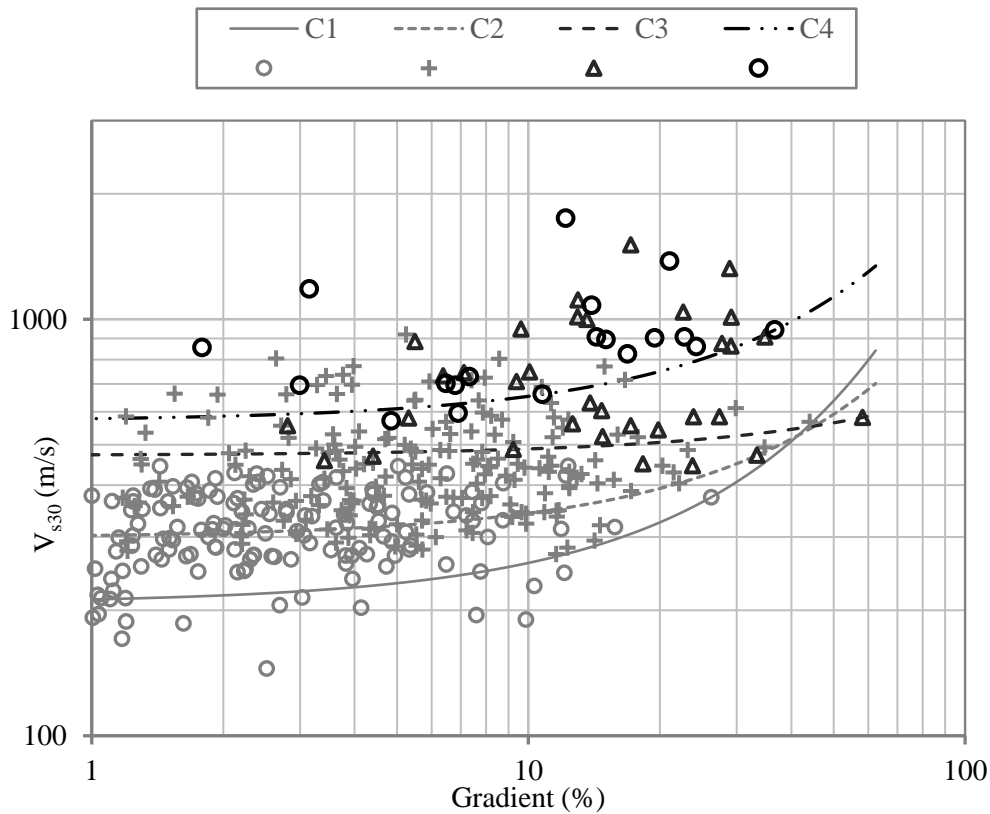


Figure 4.2  $V_{s30}$  (m/s) vs Gradient (%) for all geological classes according to Equation 4.17

Furthermore, an endeavor was undertaken to construct a nonlinear categorical quantile regression model fortified with interaction terms. This complex model, delineated in Equation 4.18, was constructed to encompass the joint impact of geological classes and terrain surface texture on the estimation of  $V_{s30}$  values. The aim was to transcend the isolated influences exerted by individual terrain surface texture and geological classes on the dependent variable,  $V_{s30}$  values. Despite the model's more flexible and complicated form, a notable enhancement in the goodness of fit criteria remained elusive,  $R^2 = 0.364$  and adjusted  $R^2 = 0.354$ . Also, the neutral effect of these interaction terms can be figured out considering the coefficients of these terms, which are proximate to unity (**Error! Reference source not found.**).

Table 4.3 Model Coefficient vs Estimation

Model Coefficient	Estimation
$a_0$	212.36
$a_1$	1.01
$a_2$	1.48
$a_3$	1.68
$a_4$	3.39
$a_5$	1.00
$a_6$	1.01
$a_7$	0.99

Furthermore, the exploration delved into the evaluation of weighted nonlinear categorical quantile regression models. Surprisingly, these weighed models failed to exhibit any noticeable improvement concerning the  $R^2$  and adjusted  $R^2$  values. The research extended its evaluation to encompass a meticulous comparison. A comparison between the developed NCQRMs and the nonlinear quantile models constructed individually for each geological class highlighted compelling outcomes. Specifically, the NCQRMs exhibited superior performance, emphasized by the attainment of higher values of pseudo  $R^2$  and adjusted  $R^2$ .

### 4.3.3 Discussion of Results

Quantile regression is useful in various scenarios, especially when the conditional distribution of the response variable is not symmetric or when the research is interested in understanding different parts of the distribution, such as the lower or upper quantiles (Davino et al., 2013) When the distribution of the response variable given the predictor variables, often referred to as the conditional distribution, is not symmetric, it means that the distribution may be skewed to the right, positively skewed, or to the left, negatively skewed. In such cases, the means might not fully capture the central tendency of the distribution. The quantile regression model doesn't assume that this distribution is symmetric around its mean, as ordinary least square regression models typically do. It allows for modeling different quantiles of the response variable, providing a more detailed view of how the response variable varies across its entire distribution with changes in the predictor variables. This makes quantile regression models more robust to outliers and a powerful tool when the relationship between predictors and the response variable is not adequately captured by just the conditional mean.

Simultaneously with the current study, another research was initiated to formulate nonlinear regression models employing the ordinary least squares method. The objective was to predict  $V_{s30}$  values using the same dataset (Equation 3.6, Equation 3.7, Equation 3.8 and Equation 3.9). By examining the residuals of these regression models and plotting the histogram of the residuals of each regression model to see their distribution, the skewness in the residual distribution was detected. The skewed distribution of residuals suggests the conditional distribution of the dependent variable is not symmetric, and quantile regression may be more appropriate than regression models developed according to OLS (Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6)

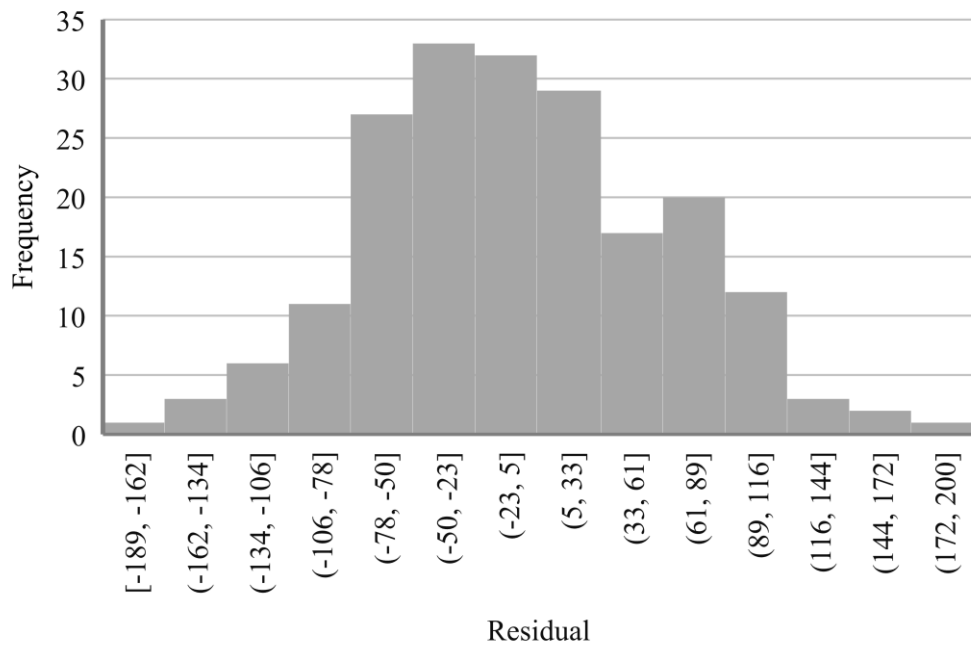


Figure 4.3 Histogram of residuals according to Equation 3.6

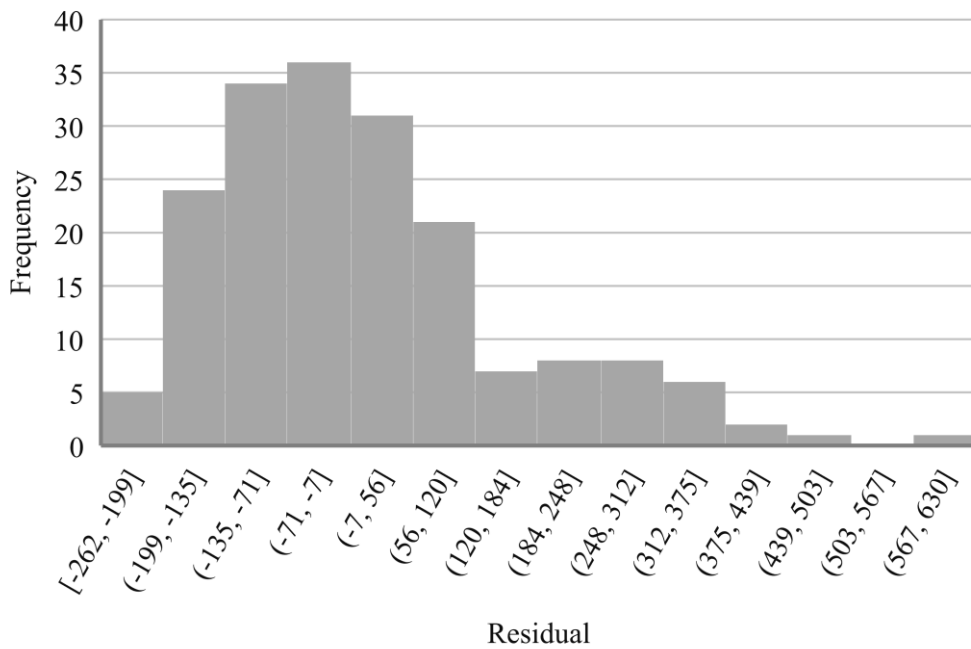


Figure 4.4 Histogram of residuals according to Equation 3.7

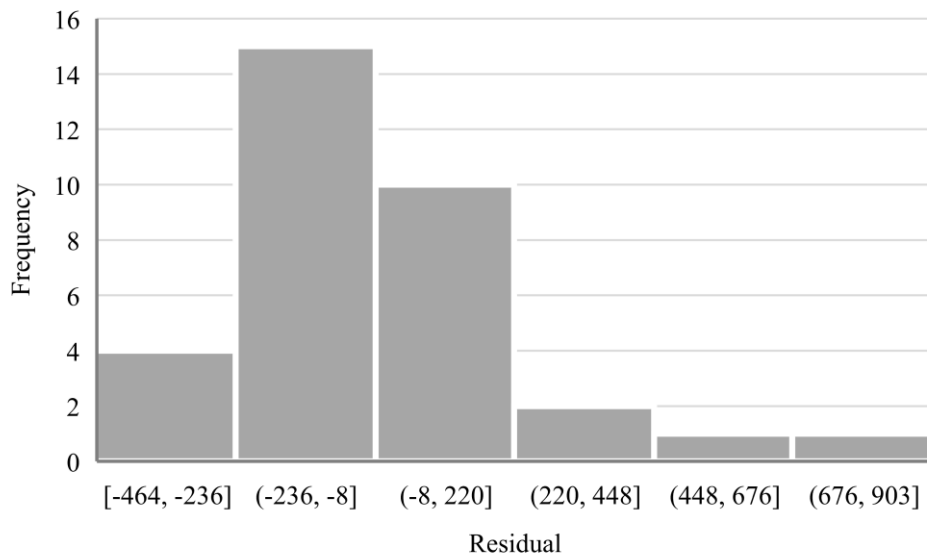


Figure 4.5 Histogram of residuals according to Equation 3.8

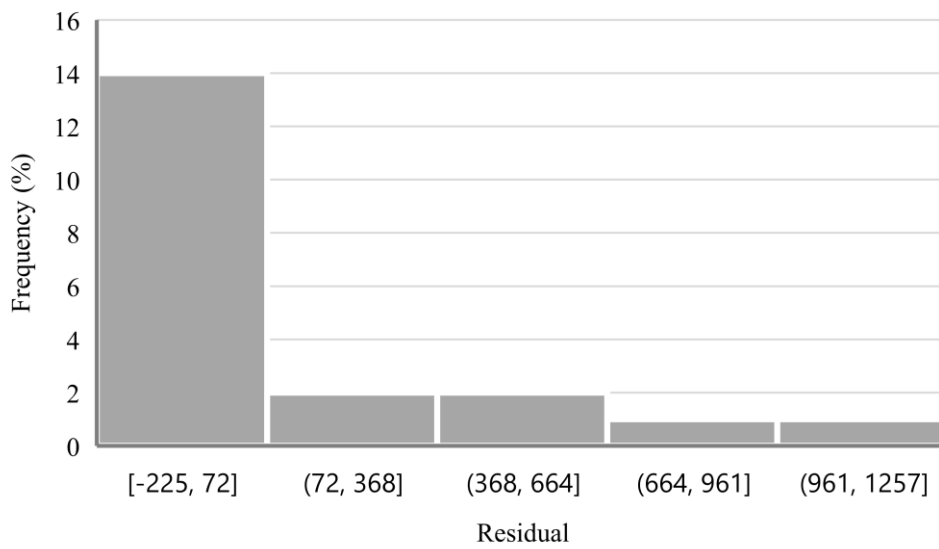


Figure 4.6 Histogram of residuals according to Equation 3.9

Furthermore, to assess the precision of the NCQR models, the estimations for the 16% quantile are compared with the observed proportions in the sample. This comparison involves dividing the topographical gradients into intervals with widths tailored to encompass at least 63  $V_{s30}$  values, with an anticipation of 10 extreme values in each interval, ensuring a consistent sample size across intervals ranging from 0 to the maximum gradient. For each gradient interval, the 90% confidence interval for the population proportion is calculated according to Equation 4.19, assuming a binomial distribution for exceeding 16% thresholds in each interval (Table 4.4). According to Table 4.4, NCQR models perform better for extreme  $V_{s30}$  values. However, by merging the first two intervals, a wider gradient interval, the quantile 16% falls inside the corresponding confidence interval. Furthermore, the 90% confidence interval for the population proportion was prepared according to nonlinear regression models, developed concurrently with the ongoing study. A comparison of the 90% confidence intervals for the population proportion of gradient intervals demonstrates that NCQR yields more acceptable results.

$$\tilde{p} \pm z_{\alpha/2} \sqrt{\frac{(\hat{p}\hat{q})/n + z_{\alpha/2}^2/4n^2}{1 + z_{\alpha/2}^2/n^2}} \quad \text{Equation 4.19}$$

where:

$$\tilde{p} = [\hat{p} + z_{\alpha/2}^2/2n] / [1 + z_{\alpha/2}^2/n]$$

$$\hat{q} = 1 - \hat{p}$$

$$\hat{p} = X/n$$

$n$ : sample number

$X$ : Sample fraction of successes

$z_{\alpha/2}$ : Z value for the corresponding confidence level

Table 4.4 90% The proportion of Vs30 observations that are lower than (a) 16% quantile, and (b)  $\mu$ - $\sigma$  estimation.

Slope (%) interval (a)	Percentage of sites with Vs30 < (a)	Percentage of sites with Vs30 < (b)	is 16% rejected for (a), and (b)?
<b>0.00 – 1.12</b>	27.40% (20/73) [90% CI: 19.7% - 36.7%]	4.17% (3/72) [90% CI: 1.7% - %10.0]	Yes, Yes
<b>1.12 – 2.21</b>	8.22% (6/73) [90% CI: 4.3% - 15.1%]	5.48% (4/73) [90% CI: 2.5% - 11.7%]	Yes, Yes
<b>2.21 – 3.78</b>	9.72% (7/72) [90% CI: 5.4% - 17.0%]	8.33% (6/72) [90% CI: 4.4% - 15.3%]	No, Yes
<b>3.78 – 5.97</b>	12.50% (9/72) [90% CI: 7.4% - 20.3%]	18.06% (13/72) [90% CI: 11.8% - 26.6%]	No, No
<b>5.97 – 11.08</b>	11.11% (8/72) [90% CI: 6.4% - 18.7%]	29.17% (21/72) [90% CI: 21.2% - 38.6%]	No, Yes
<b>11.08 – 58.18</b>	8.33% (6/72) [90% CI: 4.4% - 15.3%]	22.06% (15/68) [90% CI: 15.0% - 31.3%]	No, No

The lowest  $V_{s30}$  value predicted by NCQR for the 16% percentile is 212 m/s, which belongs to Qa/Q1c geological class, alluvium. Also, the lowest four  $V_{s30}$  values are 131, 145, 171, and 181 which belong to the same geological class. Three of these extremely low  $V_{s30}$  are near Izmir Bay, close to areas heavily affected by the 2020  $M_w$  6.6 Sisam (Aegean Sea) Earthquake. No statistically significant parameter could identify such extremely low  $V_{s30}$  ranges, primarily because of the limited data available from similar soft sites. This emphasizes the importance of identifying and incorporating more soft sites for strong-motion site installation to augment the database for extreme site amplifications.

Additionally, the currently developed NCQR model at 16% and 50% are compared with the previously constructed proxy-based models (Allen and Wald, 2007.; Iwahashi et al., 2010). Iwahashi et al. (2010) developed a model to predict  $\log(V_{s30})$  based on logarithm of elevation, slope gradient and surface texture (Equation 4.20).

$$Y = 2.28085 + (0.074 \times \log e) + (0.00331 * g) + (0.00305 * t) \quad \text{Equation 4.20}$$

where:

$e$ : Elevation

$g$ : Gradient at sites

*t*: Terrain surface texture

Also, Allen and Wald (2007) suggested a table which contains eight ranges of slopes in four categories to estimate range of  $V_{s30}$  values (Table 4.5).

Table 4.5 Summary of slope ranges for subdivided NEHRP  $V_{s30}$  categories.

Soil Type	$V_{s30}$ (m/s) range for each soil type	Slope ranges for active tectonics (m/m)
E	<180	$<1.0 \cdot 10^{-4}$
D	180–240	$1.0 \cdot 10^{-4} - 2.2 \cdot 10^{-3}$
	240–300	$2.2 \cdot 10^{-3} - 6.3 \cdot 10^{-3}$
	300–360	$6.3 \cdot 10^{-3} - 0.018$
C	360–490	0.018 – 0.050
	490–620	0.050 – 0.10
	620–760	0.10 – 0.138
B	>760	> 0.138

Through an evaluation, the performance of the model developed by Iwahashi et al. (2010) is compared with the performance of NCQR models 50%, in addition to the nonlinear regression models (Equation 3.6, Equation 3.7, Equation 3.8 and Equation 3.9) by considering their population proportion confidence interval (Equation 4.19) after normalization of predicted  $V_{s30}$  values (Table 4.6). By considering the confidence level's limits, it can be inferred that in addition to nonlinear regression models, NCQR at 50% exhibit better performance in comparison with the other models. Also the  $V_{s30}$  values of sites are compared with estimated  $V_{s30}$  values through Figure 4.7, Figure 4.8 and Figure 4.9.

Table 4.6 Statistical analysis of NCQR at 50%, Iwahasi et al. (2010) and Nonlinear regression models (n=434)

		Confidence level
Data statistics		90%
NCQR (50%)	Mean	1.02
	Lower limit	1.00
	Upper limit	1.04
Iwahashi (2010)	Mean	0.89
	Lower limit	0.86
	Upper limit	0.91
Nonlinear regression model	Mean	1.05
	Lower limit	1.03
	Upper limit	1.08

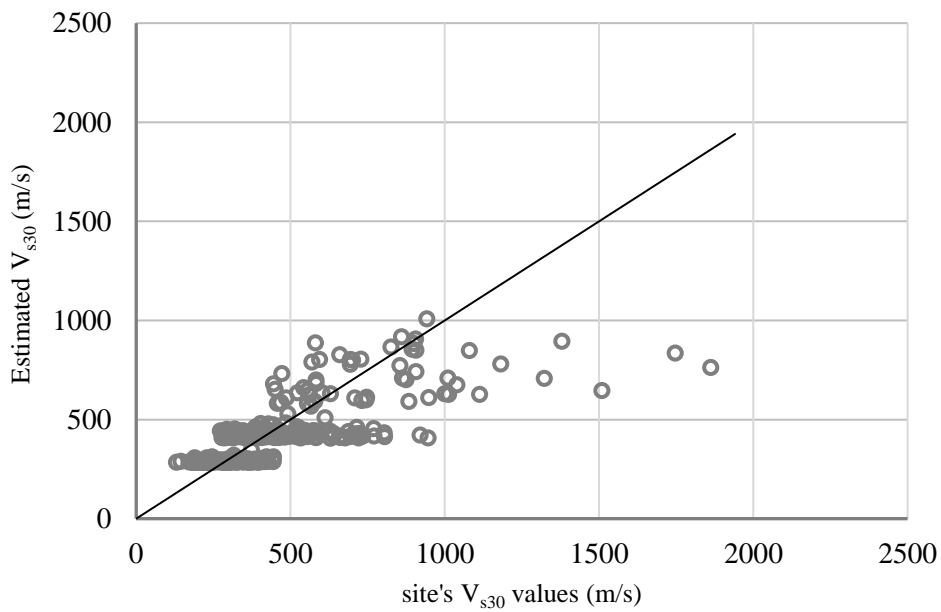


Figure 4.7 Estimated  $V_{s30}$  (m/s) vs site's  $V_{s30}$  values for NCQR at 50%

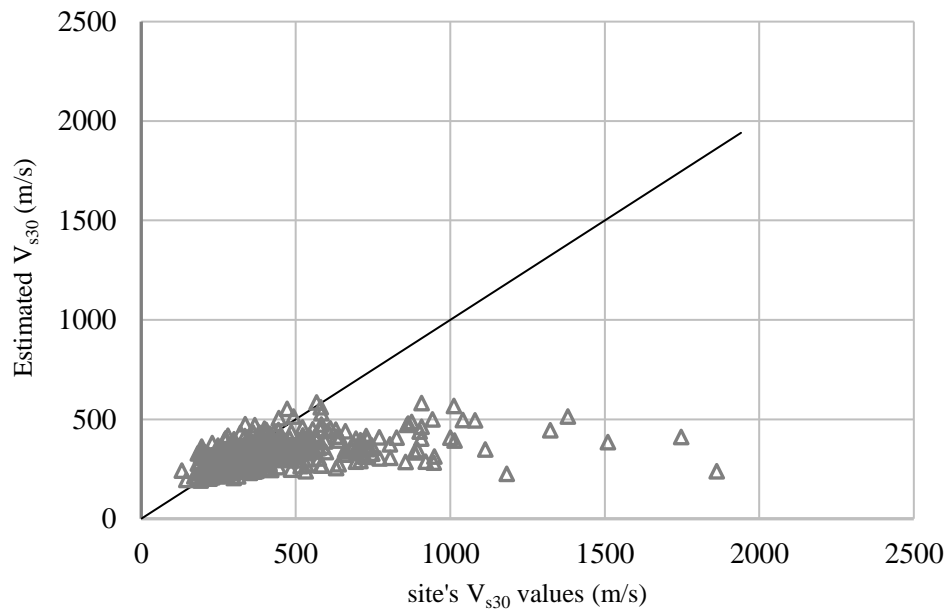


Figure 4.8 Estimated  $V_{s30}$  (m/s) vs site's  $V_{s30}$  values for model developed by Iwahashi et al. (2010)

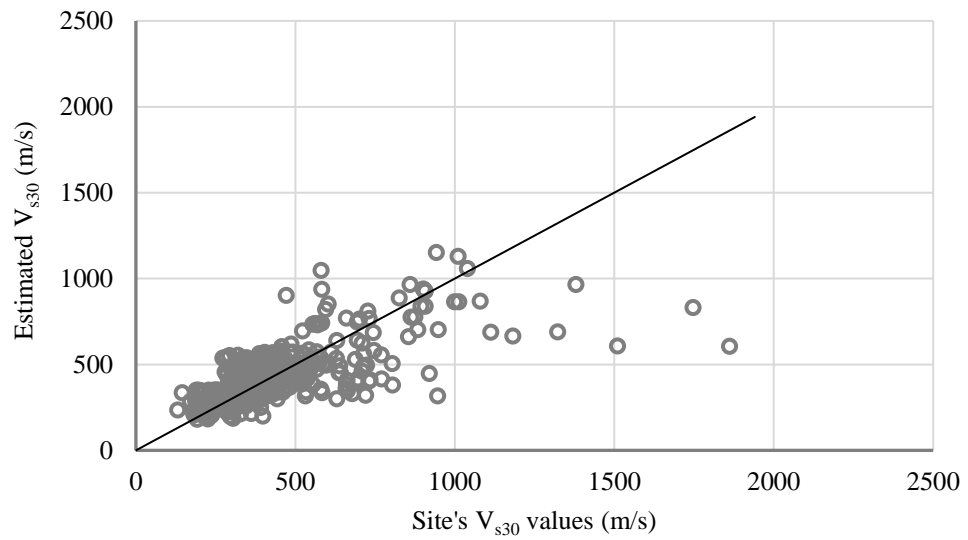


Figure 4.9 Estimated  $V_{s30}$  (m/s) vs. site's  $V_{s30}$  values for nonlinear regression models developed by Gokhan et al. (2024)

The performance of the method proposed by Allen and Wald (2007) is assessed through statistical analysis and comparison of estimated  $V_{s30}$  values and real  $V_{s30}$  values (Table 4.7). According to the Table 4.7 the performance of the method in determination of  $V_{s30}$  range is not acceptable for most of the ranges according to proposed slope ranges as reference. The main reason can be inattention of the method to the geological properties, which plays key role in estimating of  $V_{s30}$  values in addition to slope gradient.

The number of sites for which  $V_{s30}$  values are predicted correctly determined and population proportion confidence intervals are calculated for each  $V_{s30}$  range correspondingly according to estimations (Table 4.8-Table 4.15). Since the suggested method does not give importance to geological properties, population proportion confidence intervals of correctly predicted  $V_{s30}$  values express inefficiency in estimations of  $V_{s30}$  values referring to the proposed slope ranges as prediction parameters.

Table 4.7 Performance evaluation of method suggested by Allen and Wald (2007)

Soil Type	$V_{s30}$ values predicted correctly for each soil type (%)	$V_{s30}$ range for each soil type	No. of sites in each $V_{s30}$ range	No. of sites in each $V_{s30}$ range with correctly predicted $V_{s30}$ values	$V_{s30}$ values predicted correctly for each soil type range (%)
E	0.00	<180	3	0	0
D	8.20	180–240	50	2	4.00
		240–300	64	3	4.69
		300–360	81	12	14.81
C	29.75	360–490	131	45	34.35
		490–620	44	12	27.27
		620–760	30	4	13.33
B	54.83	>760	31	17	54.84

Table 4.8 Statistical analysis of method suggested by Allen and Wald (2007)

for  $V_{s30} > 180 \text{ m/s}$  – Slope:  $- < 1.0 \cdot 10^{-4}$  (n=3)

180 m/s < $V_{s30}$ – Slope: $- < 1.0 \cdot 10^{-4}$			
	Confidence level		
Data statistics	90%	80%	70%
Proportion of sites predicted correctly (%)	0.00	0.00	0.00
Lower limit	0.000	0.000	0.000
Upper limit	0.000	0.000	0.000

Table 4.9 Statistical analysis of method suggested by Allen and Wald (2007)

for  $180 \text{ m/s} < V_{s30} < 240 \text{ m/s}$  – Slope:  $1.0 \cdot 10^{-4} - 2.2 \cdot 10^{-3}$  (n=50)

180 m/s < $V_{s30} < 240 \text{ m/s}$ – Slope: $1.0 \cdot 10^{-4} - 2.2 \cdot 10^{-3}$			
	Confidence level		
Data statistics	90%	80%	70%
Proportion of sites predicted correctly (%)	4.00	4.00	4.00
Lower limit	-0.006	0.004	0.011
Upper limit	0.086	0.076	0.069

Table 4.10 Statistical analysis of method suggested by Allen and Wald (2007)

for  $240 \text{ m/s} < V_{s30} < 300 \text{ m/s}$  – Slope:  $2.2 \cdot 10^{-3} - 6.3 \cdot 10^{-3}$  (n=64)

240 m/s < $V_{s30} < 300 \text{ m/s}$ – Slope: $2.2 \cdot 10^{-3} - 6.3 \cdot 10^{-3}$			
	Confidence level		
Data statistics	90%	80%	70%
Proportion of sites predicted correctly (%)	4.69	4.69	4.69
Lower limit	0.005	0.015	0.022
Upper limit	0.095	0.085	0.078

Table 4.11 Statistical analysis of method suggested by Allen and Wald (2007)

for  $300 \text{ m/s} < V_{s30} < 360 \text{ m/s}$  – Slope:  $6.3 \cdot 10^{-3} - 0.018$  (n=81)

300 m/s < V <sub>s30</sub> < 360 m/s – Slope: 6.3 · 10 <sup>-3</sup> –0.018			
Data statistics	Confidence level		
	90%	80%	70%
Proportion of sites predicted correctly (%)	14.81	14.81	14.81
Lower limit	0.085	0.099	0.109
Upper limit	0.215	0.201	0.191

Table 4.12 Statistical analysis of method suggested by Allen and Wald (2007)

for  $360 \text{ m/s} < V_{s30} < 490 \text{ m/s}$  – Slope: 0.018 – 0.050 (n=131)

360 m/s < V <sub>s30</sub> < 490 m/s – Slope: 0.018–0.050			
Data statistics	Confidence level		
	90%	80%	70%
Proportion of sites predicted correctly (%)	34.35	34.35	34.35
Lower limit	0.272	0.287	0.297
Upper limit	0.408	0.393	0.383

Table 4.13 Statistical analysis of method suggested by Allen and Wald (2007)

for  $490 \text{ m/s} < V_{s30} < 620 \text{ m/s}$  – Slope: 0.050–0.10 (n=44)

490 m/s < V <sub>s30</sub> < 620 m/s – Slope: 0.050–0.10			
Data statistics	Confidence level		
	90%	80%	70%
Proportion of sites predicted correctly (%)	27.27	27.27	27.27
Lower limit	0.160	0.184	0.201
Upper limit	0.380	0.356	0.339

Table 4.14 Statistical analysis of method suggested by Allen and Wald (2007)

for  $620 \text{ m/s} < V_{s30} < 760 \text{ m/s}$  – Slope: 0.10 – 0.138 (n=30)

620 m/s < V <sub>s30</sub> < 760 m/s – Slope: 0.10–0.138				
Data statistics	Confidence level	90%	80%	70%
Proportion of sites predicted correctly (%)		13.33	13.33	13.33
Lower limit		0.029	0.051	0.066
Upper limit		0.231	0.209	0.194

Table 4.15 Statistical analysis of method suggested by Allen and Wald (2007)

for  $760 \text{ m/s} < V_{s30}$  – Slope: - >0.138 (n=31)

760 m/s < V <sub>s30</sub> – Slope: - >0.138				
Data statistics	Confidence level	90%	80%	70%
Proportion of sites predicted correctly (%)		54.84	54.84	54.84
Lower limit		0.403	0.435	0.457
Upper limit		0.697	0.665	0.643



## CHAPTER 5

### LOGISTIC REGRESSION

#### 5.1 Introduction

Logistic regression, also known as classifier, plays an important role in numerous fields where the analysis and prediction of categorical outcomes are essential. This method finds its application in diverse domains from medicine, engineering, and marketing to finance and beyond. At its core, logistic regression is a robust tool for modeling the probability of a binary or categorical event occurring based on one or more predictor variables. It's highly valued for its ability to handle complex relationships between input factors and the likelihood of a particular outcome. By providing insights into the probability of categorical events, logistic regression empowers decision-makers and analysts to make informed choices and predictions, in modern data-driven research. In this section, the diverse applications of logistic regression in addressing civil engineering challenges across various domains are considered.

Rafiee et al. (2013) evaluated tunnel stability in the face of potential rockfall hazards. They employed binary and multinomial logistic regression models, taking into account crucial factors such as rock mass rating, RMR, Q-rock mass classification, and hydraulic radius.

Roof fall hazards are significant challenges in underground coal mining, often caused because of the unpredictability of geological interpretation and varying mining factors. Employing binary logistic regression, the study aims to predict the severity of roof fall accidents by considering key contributing parameters (Palei and Das, 2009).

Using a binary logistic regression model, in addition to linear discriminant analysis, Ghasemi et al. (2019) developed empirical correlations for squeezing deformation

prediction in tunnels before starting the project construction. Employing a logistic regression model, besides to artificial neural network model, an automated strategy was developed for distinguishing deep and shallow microseismic events from each other based on the waveforms (Mousavi et al., 2016). Also Vallejos et al. (2013) proposed procedures based on logistic regression and neural network classification techniques to identify the seismic records in seismically active mines, which is useful to monitor the rock mass response around mining excavation. In Gyeongju, South Korea, the seismic vulnerability of the  $M_L=5.8$  Gyeongju Earthquake was studied through logistic regression to develop appropriate models for evaluating seismic susceptibilities (Han et al., 2019). In Indonesia, a probabilistic model grounded in the principles of logistic regression was formulated to monitor seismic building vulnerability, utilizing damage data resulting from significant earthquakes (Saputra et al., (2017). Zhu et al., (2023) employed a cost function logistic regression model, which is a variant of logistic regression incorporating cost factors for misclassification errors, to calculate seismic hazard levels. Subsequently, their study conducted a multi-criteria seismic risk assessment using the GIS platform. In order to conduct a seismic risk assessment for museum exhibition halls, on the basis of the multinomial logistic regression model, the seismic fragility analysis was performed to study the damages caused by earthquakes on museum buildings and artifacts (Yang et al., 2023).

Recognizing the critical role of fragility curves in seismic risk assessment within the performance-based earthquake engineering framework, Kiani et al. (2019) employed logistic regression alongside nine other classification methods to predict structural responses and construct fragility curves for earthquake scenarios. Automatic first arrival picking is a system designed to identify the initial occurrence of primary seismic waves within earthquake signals. Syabani et al. (2020) undertook the task of designing this system and assessing the efficacy of logistic regression in detecting primary waves upon their first arrival. The conclusive findings demonstrated the remarkable accuracy achieved by the logistic regression method in this context. Research conducted by Mori et al. (2020) introduced a methodology for the probabilistic assessment of the emergency structural system operation in the case of a

seismic event occurrence. To this end in addition to the estimating ground motion amplification and shaking scenarios in terms of PGA, co-seismic permanent failure and deformations, including phenomena such as landslides and liquefaction, were assessed using logistic regression methods. A hybrid logistic regression method was used as a classifier tool through the seismic assessment of the rocking response of structures to evaluate the finite rocking rotations and rocking overturn (Gkountakou et al., 2023). The stepwise logistic regression with a nonlinear logit function was utilized by Kameshwar et al. (2014) to develop parametric bridge fragility functions to study the risk analysis and quantify the vulnerability of bridges to seismic hazards.

Seismic soil liquefaction is a significant and devastating aspect of earthquakes, primarily occurring in loosely to moderately saturated sandy soils. In Duan's research (Duan et al., 2021) a state parameter, that accounts for both relative density and effective stress of soil was employed to establish a probabilistic liquefaction evaluation method using a logistic regression model. Yao et al. (2021) introduced an innovative liquefaction evaluation formula utilizing the logistic regression method relying on shear wave velocity data. Commonly cone penetration test, CPT, is used to assess the liquefaction potential of soils. To predict the probability of liquefaction, a logistic regression model was applied using comprehensive in-situ CPT test results (Jairi et al., 2021). Papathanassiou (2008) applied logistic regression analysis to introduce an LPI-based formula for predicting liquefaction occurrence. This formula was developed using datasets gathered from SPT tests conducted at both liquefied and non-liquefied sites in Taiwan, Türkiye, and Greece. A decision tree, DT, is a popular machine learning and data analysis technique used for both classification and regression tasks. It was compiled by 620 records of seismic parameters and soil properties for post-earthquake soil liquefaction assessment and results were compared with those concluded by the logistic regression method (Gandomi et al., 2013). Because of the nonlinear nature of liquefaction classification, it is challenging to develop a comprehensive model using a simplified technique based on in situ tests. Zhang et al. (2016) proposed a modified adaptive regression splines approach based on logistic regression to study seismic liquefaction potential. In addition to the Bayesian

approach, Juang et al. (2001) employed the logistic regression method to conduct an assessment of several probabilistic liquefaction evaluation techniques, which were developed according to SPT and CPT test results. Based on liquefaction occurrence and damage severity, a hybrid geotechnical-geospatial liquefaction assessment approach was introduced to classify sites, through which surface geospatial data was referred to as proxies for liquefaction-related parameters (Azul et al., 2023). A set of geospatial parameters including weighted-magnitude, peak ground acceleration, weighed  $V_{s30}$  values, and compound topographic index were coupled with the logistic regression method to evaluate the liquefaction hazard in continental Europe (Bozzoni et al., 2021). Using a dataset obtained from field experiments, specifically SPT, conducted on soils in the Adapazari region following the 1999 earthquake, Ozsagir et al. (2022) utilized various machine learning techniques, including the logistic regression method, to estimate liquefaction potential. Zhen et al. (2021) proposed a novel hybrid classifier that incorporates the logistic regression method, along with six other classifiers, to enhance the generalizability of previous models. This improvement was achieved using the weighted voting method, which utilizes the combined advantages of these classifiers for more robust predictions.

Landslides, which encompass various forms of mass wasting, and mass movement, occur when a slope, or a section of it, undergoes processes that transition it from a stable condition to an unstable one. This geotechnical disaster causes thousands of deaths and billions of dollars of damage each year. An increase in pore water pressure, earthquake shaking, and human activity are the main reasons for landslides.

Logistic regression is one of the valuable tools that can be employed to assess landslide susceptibility. Rare events logistic regression, a variation of the standard logistic regression, was used by Van et al. (2006) to prepare a landslide susceptibility map in a 200 km<sup>2</sup> study area located in the Flemish Ardennes, Belgium. The logistic regression method was utilized to evaluate landslide susceptibility in the Badulla district of Sri Lanka. In this assessment, factors such as slope, aspect, lithology, land cover, as well as proximity to rivers and roads were taken into account as contributors to landslide occurrence (Hemasinghe et al., 2018). Dai et al. (2001) conducted a

research study in Lantau Island, Hong Kong, to assess landslide susceptibility. This investigation utilized a Geographic Information System, GIS, in conjunction with a logistic regression model to analyze and predict landslide occurrences in the area.

In the landslide-prone national highway road section in the northern Himalayas, a logistic regression method was proposed to assess the landslide hazard. Not only the results were verified through the comparison with the geotechnical-based slope stability probability classification, but also the logistic regression model performance was assessed by the receiver operator characteristics curve (Das et al., 2010). A comparison was conducted to evaluate GIS-based landslide susceptibility mapping methods, including the application of logistic regression, in Koyulhisar (Sivas, Türkiye). This comparative analysis aimed to assess the performance of various methods in the preparation of landslide susceptibility maps (Yilmaz, 2010). Multivariate statistical analysis, in the form of logistic regression, was employed to generate a landslide susceptibility map in the Kakuda-Yahiko Mountains of Central Japan. This analysis considered various independent variables, including lithology, bedrock-slope relationship, lineaments, slope gradient, aspect, elevation, and road network, to assess landslide susceptibility in the region (Ayalew and Yamagishi, 2005). Recognizing the absence of standardized practices for presenting landslide evaluation results, Lombardo et al. (2018) proposed a comprehensive framework for harmonizing the way researchers share their findings about landslide susceptibility assessment using the logistic regression method. In the Selangor area of Malaysia, a comprehensive landslide hazard analysis and mapping project was conducted using Geographic Information Systems, GIS. This assessment utilized landslide-occurrence factors and employed both frequency ratio and logistic regression models to evaluate and map the landslide hazard in the region (S. Lee and Pradhan, 2007). Ohlmacher and colleagues (2003) used the multiple logistic regression method to assess the probability of future landslide occurrence and create a landslide hazard map for Atchison, Kansas, and its surrounding areas. To assess the frequency and distribution of landslides in the Zhongxian–Shizhu segment, a region known for its high susceptibility to landslides in China, researchers utilized the logistic regression method to evaluate the probability

of landslide hazards in this particular area (Bai et al., 2010). In the Hoa Binh province of Vietnam, an evaluation of landslide susceptibility was conducted, considering ten influencing factors for landslide occurrence. Also, the results were compared using two distinct methods: the logistic regression method and a statistical index approach (Bui et al., 2011).

In this section, the primary objective was to demonstrate the utility of logistic regression across various fields, with a particular focus on its applications in geotechnical engineering, through the presentation of previous research. The following section considers the fundamental mathematical principles of logistic regression.

## 5.2 Logistic Regression

As discussed previously, logistic regression is a statistical modeling technique widely used in various fields to analyze the relationship between one or more independent variables, predictors, and binary or categorical dependent variables. It is particularly suited for situations where the dependent variable is not continuous, but rather falls into distinct categories. It is crucial to emphasize that the objective of an analysis utilizing this model aligns with the fundamental goal of any regression model in statistics, to identify the most suitable and interpretable model that characterizes the relationship between a dependent variable and a set of independent variables. The distinctions between logistic and linear regressions become apparent through both the model's form and the underlying assumptions that underpin it (Hosmer et al., 2000).

The first difference concerns the nature of the relationship between the dependent and independent variables. In a common regression problem, a fundamental quantity of interest is the conditional mean of the dependent variable,  $E(Y|X)$ , which represents the expected value of the dependent variable,  $Y$ , given a specific value of the independent variable,  $X$ . In this concept, also referred to as the expected value of  $Y$ ,  $Y$  represents the outcome variable, and  $X$  represents a particular value of the independent variable. In linear regression, it is assumed that the expected value of  $Y$  may be expressed as an equation linear in  $X$  (Equation 5.1). This expression implies that it is possible for the expected value to take on any value as  $x$  ranges between  $-\infty$  and  $+\infty$ .

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i \quad \text{Equation 5.1}$$

where:

$Y_i$ : Dependent variables

$X_i$ : Independent variables

$\beta_0$  and  $\beta_1$ : Coefficients of the linear model

In contrast to common regression, model the expected value of regression with a dichotomous independent variable must be greater than or equal to zero and less than or equal to one (i.e.,  $0 \leq E(Y_i | X_i) \leq 1$ ). To this end, many distribution functions have been proposed for analyzing dichotomous outcome variables. The choice of logistic distribution is primarily motivated by two key factors. First of all, from a mathematical perspective, logistic distribution is highly flexible and makes it well-suited for a wide range of applications. Secondly, the parameters of the logistic distribution help to generate useful estimates of effect, influence of independent variables on the probability of a particular dependent variable, making it valuable for practical data analysis. When the logistic distribution is used, the expected value of  $Y$  given  $X$  is presented through equation 5.2.

$$E(Y_i | X_i) = \pi(X_i) = \frac{1}{1 + e^{[-(\beta_0 + \sum_{i=1}^k \beta_i X_i)]}} \quad \text{Equation 5.2}$$

where:

$Y_i$ : Dependent variables

$X_i$ : Independent variables

$\beta_0$  and  $\beta_1$ : Coefficients of the linear model

$\pi(X_i)$ : Expected value of  $Y$  given  $X$  for logistic distribution

Considering a dichotomous dependent variable with possible values one for success and zero for failure,  $p = P(S) = P(Y=1)$ . Since the value of  $p$  will depend on the value of some independent variable,  $X_i$ , instead of using just the symbol  $p$  for the success probability,  $p(X_i)$  is used to emphasize the dependence of this probability on the value of  $X$ . Consequently, the expected value of logistic distribution can be demonstrated as Equation 5.3.

$$\pi(X_i) = [1 \times p(X_i)] + [0 \times (1 - p(X_i))] = p(X_i) = P(Y = 1) \quad \text{Equation 5.3}$$

where:

$p(X_i)=P(Y=1)$ : Success Probability

Indeed, the logistic regression model is designed to estimate the probability of an event occurring based on a given dataset of independent variables. In this approach, since the outcome is a probability, the dependent variable is bounded between zero and one which demonstrates the probability of the occurring events.

Another significant distinction between linear and logistic regression models pertains to the conditional distribution of the dependent variable. In the linear regression model, the dependent variable is assumed to be expressed as  $Y = E(Y_i|X_i) + \varepsilon$ . According to assumptions in linear regression, the error term,  $\varepsilon$ , follows a normal distribution with zero expected value and constant variance across levels of the independent variable. Consequently, the conditional distribution of the dependent variable given  $X$  in linear regression follows a normal distribution with a mean equal to  $E(Y/X)$  and a constant variance. On the other hand, for the dichotomous dependent variable, the value of the dependent variable for the given independent variable  $X$  is  $Y = \pi(X_i) + \varepsilon$ . In this context, the quantity  $\varepsilon$  can take one of two potential values. If  $Y = 1$ , then  $\varepsilon = 1 - \pi(X)$  with a probability of  $\pi(X)$ , and if  $Y = 0$ , then  $\varepsilon = -\pi(X)$  with a probability of  $1 - \pi(X)$ . As a result, the error,  $\varepsilon$ , follows a distribution with an expected value, mean, of zero and a variance equal to  $\pi(X) * [1 - \pi(X)]$ . This signifies that the conditional distribution of the outcome variable adheres to a binomial distribution, with the probability determined by the conditional mean,  $\pi(X)$ .

In summary, for regression analysis with dichotomous dependent variables:

- a) The model for the conditional mean of the regression equation must be bounded between zero and one, which is satisfied by the logistic regression model,  $\pi(X)$ .
- b) The binomial, not the normal, distribution describes the distribution of the errors and is the statistical distribution on which the analysis is based.
- c) The principles that guide an analysis using linear regression also guide us in logistic regression.

In logistic regression, the method used to estimate the model parameters, and coefficients, is the maximum likelihood estimation, MLE, method. In a general sense,

the MLE method seeks to find values for unknown parameters that maximize the probability of obtaining the observed dataset. To apply this method, it is necessary to construct a function known as the likelihood function. This function expresses the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimators, MLEs, of these parameters, are the values that make this likelihood function as large as possible. Accordingly, the resulting estimators closely align with the observed data.

**5.2.1 Different Types of Logistic Regression**

In general, there are three main types of logistic regression models, which are the binary logistic regression model, the multinomial logistic regression model, and the ordinal logistic regression model. This section will provide a detailed explanation of each model.

The binary logistic regression model is the simplest form of logistic regression, used when the dependent variable is dichotomous (i.e., has two possible outcomes) while independent variables can be categorical or continuous. This model submits the probability of occurrence of an event according to the supplied dataset, dependent and independent variables. Therefore, the binary type of logistic regression model results in probability values between 0 and 1, which reveals the probability of the events that are prone to occur (Equation 5.4).

$$P (Y = 1) = \frac{1}{1 + \exp (-g(x))} \tag{Equation 5.4}$$

where:

$P(Y=1)$ : Success probability

$g(x)$ : The model to be used to predict the success probability

The multinomial logistic regression model, also known as the multiclass logistic regression model, is the second extension of the logistic regression model. In this type

of logistic regression model, the dependent variable has three or more classes of possible dependent variable; however, these dependent classes are not in specified order.

The ordinal logistic regression model, the third type, is used when the dependent variable has three or more possible outcomes, classes, with a defined order ordered (i.e., there is a meaningful sequence or rank order to the outcome classes). In this type of logistic regression, instead of considering the probability of an individual event, the probabilities of that event and all events that are ordered before it are considered, in other words, ordinal logistic regression models cumulative probabilities of the dependent variable being below or above certain thresholds ( Equation 5.5 and Equation 5.6). It is necessary to note that, since the dataset used in this research is in defined order, it is more consistent with data type to develop ordinal logistic regression to categorize the sites according to cumulative probability

$$P (R \leq L) = \frac{1}{1 + \exp(-Z(x)_l)}; l < L \quad \text{Equation 5.5}$$

$$Z (x)_l = D_l - h(x)_l \quad \text{Equation 5.6}$$

where:

*R*: The class number of the dependent variable of ordinal logistic regression model (taken as the site classes in Table 5.1)

*L*: Number of all dependent variable classes that have defined order

*l*: Number of each dependent variable class

*P (R ≤ L)*: The probability by which the dependent variable belongs to a specific class

*D<sub>l</sub>*: The cut-point value to be used for each class

*Z (x)<sub>l</sub>*: The equation to be used in the ordinal logistic regression model for each class

*h (x)*: The model to be used to predict the cumulative success probability

In summary, while binary logistic regression is used for predicting the probability of one of two possible outcomes, multinomial logistic regression, and ordinal logistic regression extend the concept to handle dependent variables with more than two categories, with the latter also considering the order of those categories.

### 5.2.2 Goodness of Fit

In the context of logistic regression, McFadden's  $R^2$  (Equation 5.7), also titled as pseudo  $R^2$  in the software STATA, is the most common statistical measure to evaluate the goodness of fit. To compare models with different numbers of parameters, it might be more appropriate to use McFadden's adjusted  $R^2$  (Equation 5.8), which adjusts for the number of parameters in the model (McFadden, 1973). This statistical measure is an analogous concept to  $R^2$  in linear regression, which represents the proportion of variance explained by the model. However, unlike the  $R^2$  in linear regression, McFadden's  $R^2$  does not represent the proportion of variance explained by the model because logistic regression models probabilities, not variances. McFadden's  $R^2$  has a range between 0 and 1, with higher values indicating a better fit of the model to the data.

$$R_{MF}^2 = 1 - \frac{LL_{Model}}{LL_{Null}} \quad \text{Equation 5.7}$$

$$R_{MFA}^2 = 1 - \frac{LL_{Model} - N}{LL_{Null}} \quad \text{Equation 5.8}$$

where:

$R_{MF}^2$ : McFadden's  $R^2$

$R_{MFA}^2$ : McFadden's adjusted  $R^2$

$LL_{Model}$ : Natural logarithm of the likelihood of the model with predictors

$LL_{Null}$ : Natural logarithm of the likelihood of the model without predictors

$N$ : Number of parameters, coefficients, in the logistic regression mode

### 5.3 Application of Logistic Regression Model in Site Classification

In this section, multinomial and ordinal logistic regression models are developed based on Türkiye's  $V_{s30}$  catalog (Table 5.1), which is used for site classification. In this context, the independent variables include four geological classes, representing the geological properties of sites, as well as terrain surface texture and gradient values, which represent the topographical properties of the sites, where  $V_{s30}$  values are submitted by AFAD. Since there are only three sites belonging to soil type  $Z_A$ ,  $V_{s30} > 1500$ , and soil type  $Z_E$ ,  $V_{s30} < 180$ , the process of model development is performed considering site zones  $Z_B$ ,  $Z_C$ , and  $Z_D$ .

Table 5.1 Türkiye's Soil Type Catalog

Soil Type	$V_{s30}$ (m/s)
$Z_A$	$> 1500$
$Z_B$	760 - 1500
$Z_C$	360 - 760
$Z_D$	180 - 360
$Z_E$	$< 180$

#### 5.3.1 Application of Ordinal Logistic Regression Model for Site Classification

To develop ordinal logistic regression classifiers, which model the cumulative probabilities of the categorical dependent variables being below or above certain thresholds, sites are arranged in a specified order according to Türkiye's soil type catalog (Table 5.1). Similar to the previous section, four distinct geological classes, as well as terrain surface texture and gradient values are utilized as geological and topographical attributes respectively. These attributes are incorporated into the model

$h(x)$ , which is used in ordinal logistic regression to predict the cumulative success probability.

$$\begin{aligned}
 h(x) = & (a_1 \cdot g) + (a_2 \cdot t) + (a_3 \cdot g \cdot t) + (a_4 \cdot C_1) + (a_5 \cdot C_2) && \text{Equation 5.9} \\
 & + (a_6 \cdot C_3) + (a_7 \cdot C_1 \cdot g) + (a_8 \cdot C_2 \cdot g) + (a_9 \cdot C_3 \cdot g) \\
 & + (a_{10} \cdot C_1 \cdot t) + (a_{11} \cdot C_2 \cdot t) + (a_{12} \cdot C_3 \cdot t) \\
 & + (a_{13} \cdot C_1 \cdot g \cdot t) + (a_{14} \cdot C_2 \cdot g \cdot t) + (a_{15} \cdot C_3 \cdot g \cdot t)
 \end{aligned}$$

where:

$h(x)$ : the model to be used to predict the probability of success (occurring event)

$g$ : Gradient at sites

$t$ : Terrain surface texture

$C_1, C_2, C_3$ : Geological classes

$a_0, a_1, a_2, \dots, a_{15}$ : Nonlinear categorical quantile regression model coefficient

Considering  $R_{MFA}^2$  and  $R_{MF}^2$  (Table 5.2) as criteria for evaluating goodness of fit, and removing terms with ineffective influence on these goodness of fit criteria, a simplified model is developed as Equation 5.10. The corresponding coefficients of Equation 5.10 are submitted in

Table 5.3.

$$h(x) = (a_1 \cdot g) + (a_2 \cdot C_1) + (a_3 \cdot C_2) + (a_4 \cdot C_3) \qquad \text{Equation 5.10}$$

Table 5.2 Models with different numbers of terms vs goodness of fit criteria

No. of terms in the model	McFadden's $R^2$	McFadden's adjusted $R^2$
15	0.341	0.29
14	0.329	0.288
13	0.329	0.290
...	...	...
6	0.317	0.296
5	0.317	0.299
4	0.317	0.301
3	0.304	0.291

Table 5.3 Coefficient's value of Equation 5.10

$a_1$	$a_2$	$a_3$	$a_4$	1 <sup>st</sup> $D_l$	2 <sup>nd</sup> $D_l$
-0.060	6.108	4.039	1.523	-0.064	4.743

As expected according to theoretical principles, sites with higher gradient are more likely to exhibit higher probability of belonging to soil types characterized by higher  $V_{s30}$  values (Figure 5.1, and Figure 5.2).

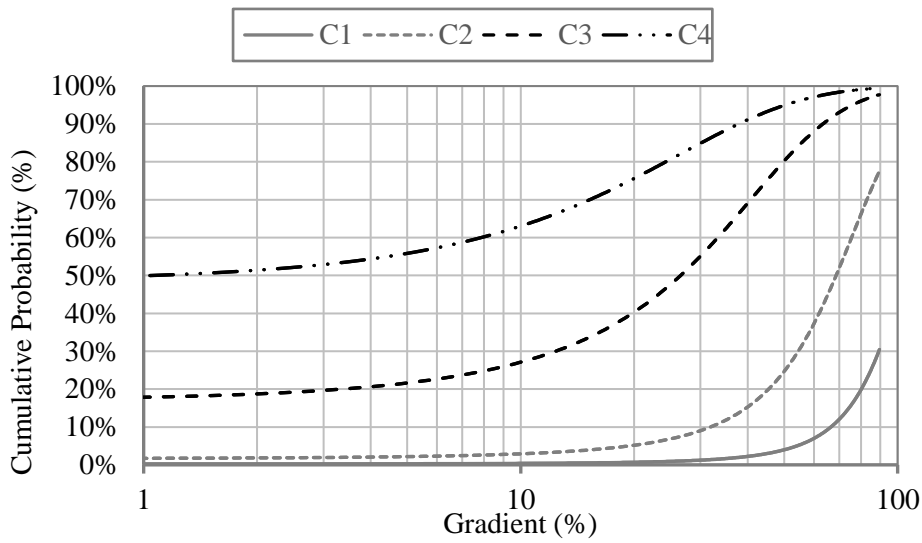


Figure 5.1 The probability of  $V_{s30} > 760$  m/s due to ordinal logistic regression model.

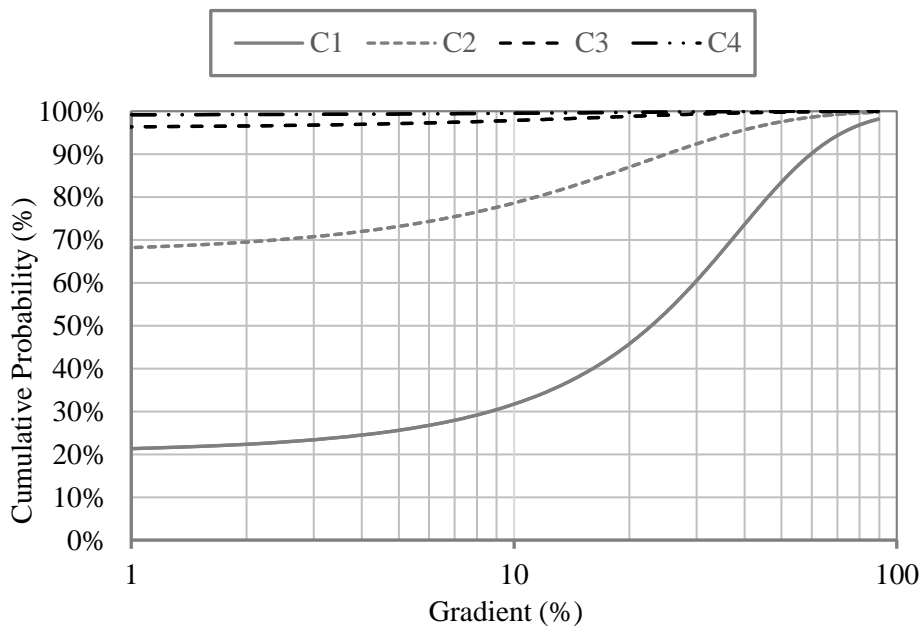


Figure 5.2 The gradient vs probability of  $V_{s30} > 360$  m/s due to ordinal logistic regression model.

Figure 5.3 through Figure 5.10 demonstrate the ordinal logistic regression models developed to estimate the likeliness (probability) of  $V_{s30} > 360$  m/s and  $V_{s30} > 760$  m/s

for each geological class. In these figures, the sample for  $V_{s30} > 360$  m/s and  $V_{s30} > 760$  m/s are demonstrated dichotomously; 1 representing the threshold  $V_{s30}$  is exceeded, and 0 otherwise. According to Figure 5.3 the  $V_{s30}$  values of all class  $C_1$  sites are less than 760 m/s. For geological classes  $C_2$  and  $C_3$ , a small proportion of data set has  $V_{s30} > 760$  m/s as shown in Figure 5.4 and in Figure 5.5 respectively. On the other hand, most of the sites belonging to geological class  $C_4$  have  $V_{s30} > 760$  m/s as shown in Figure 5.6. As it has been expected for stiffer geological classes,  $C_4$  and  $C_3$ , there are sites with higher  $V_{s30}$  values.

Figure 5.7 through Figure 5.10 show the data and the logistic models for the event  $V_{s30} > 360$  m/s. For geological class  $C_1$ , a proportion of sites satisfy the inequality  $V_{s30} > 360$  m/s, while most of the sites belonging to  $C_2$  possess  $V_{s30}$  values more than 360 m/s as shown in Figure 5.8 respectively. All sites belonging to stiffer geological classes  $C_3$  and  $C_4$ , have  $V_{s30}$  values more than 360 m/s as shown in Figure 5.9 and Figure 5.10.

Also, the performances of ordinary logistic models for  $V_{s30} > 360$  m/s and  $V_{s30} > 760$  m/s are evaluated by comparing the proportion of sites with  $V_{s30}$  values more than 360 m/s and 760 m/s. The proportion of successful observations in the sample, defined as  $V_{s30} > 360$  m/s or, 760 m/s, for a set of gradient intervals are shown in Figure 5.11- Figure 5.17. The sample is also presented by the tables following each figure. Comparing the proportion of sites with  $V_{s30} > 360$  m/s and those with  $V_{s30} > 760$  m/s, and the corresponding probabilities of success by the ordinary logistic models, it is seen that the agreement is generally acceptable for the geological classes  $C_1$  and  $C_2$ . In geological classes  $C_3$  and  $C_4$  all sites have  $V_{s30}$  values more than 360 m/s and this is consistent by the probability predictions of ordinary logistic regression models. As it is stated by Figure 5.3, since there is no site belonging to the geological class  $C_1$  and showing  $V_{s30} > 760$  m/s. However, exhibits some inconsistencies between the sample and the logistic regressions. For the gradients exceeding 30%, the probabilistic predictions increase up to 100% gradually, the sample's proportion stays around 50% for  $V_{s30} > 760$  m/s in the geological class  $C_3$ . However, the sample size is limited to 4 for this range, so a strong conclusion on the model's predictions is not possible. Also, this is supporting the necessity to use a single function to be fit on a complete sample

set, so that any inadequateness in a sample range can be compensated by the data existing in other ranges. Note that, a similar discrepancy exists for  $V_{s30} > 760$  m/s in class C<sub>4</sub>, but the sample size is limited to 20 for the higher gradient range showing the deviation of model from the data. It is obvious that more data from steep slopes is necessary, nonetheless the density of settlements is usually low on such grounds, so that the need for  $V_{s30}$  predictions will be limited.

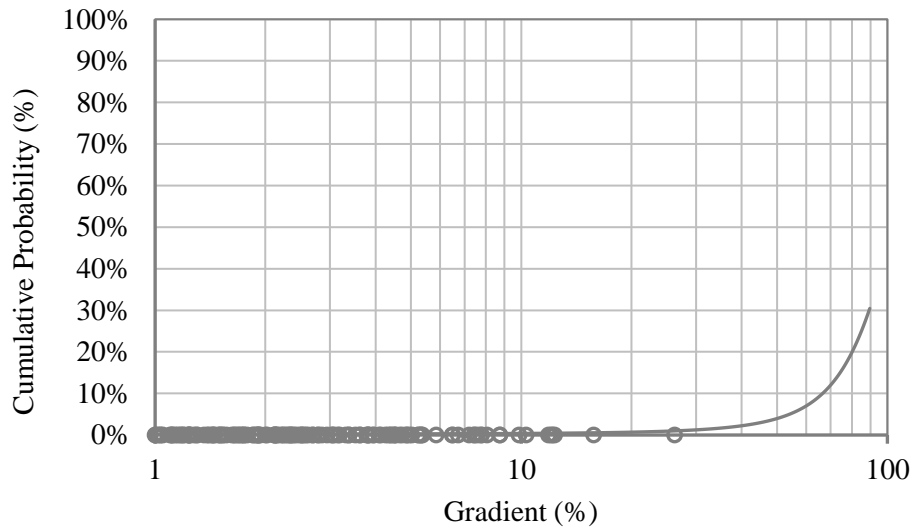


Figure 5.3 Gradient vs cumulative probability of ordinal logistic regression model to predict site with  $V_{s30} > 760$  m/s for geological class  $C_1$

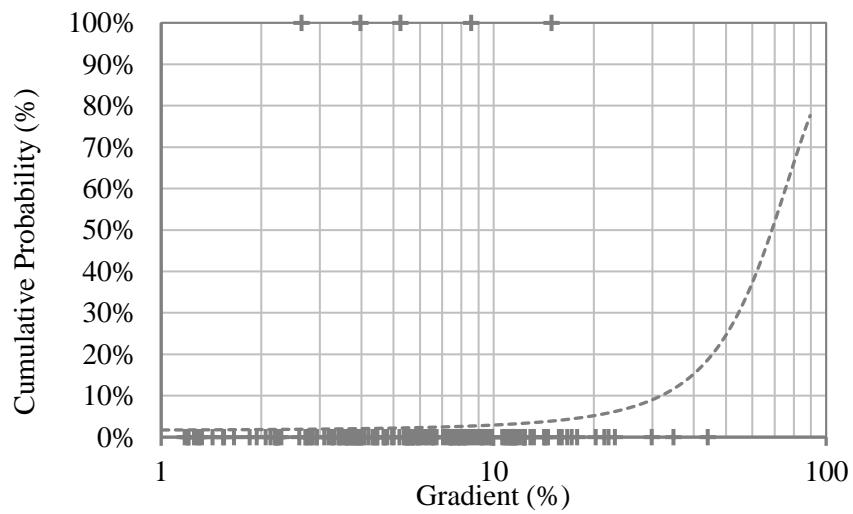


Figure 5.4 Gradient vs cumulative probability of ordinal logistic regression model to predict site with  $V_{s30} > 760$  m/s for geological class  $C_2$

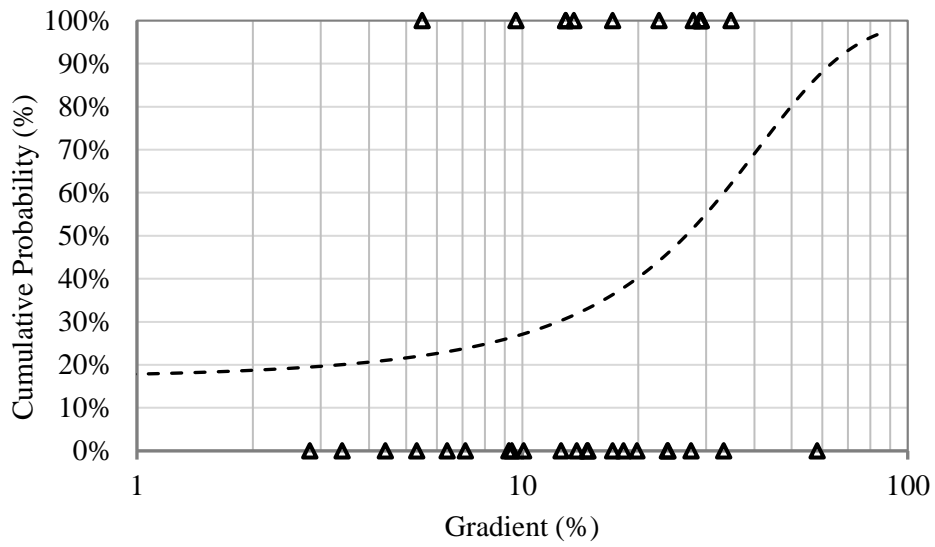


Figure 5.5 Gradient vs cumulative probability of ordinal logistic regression model to predict site with  $V_{s30} > 760$  m/s for geological class  $C_3$

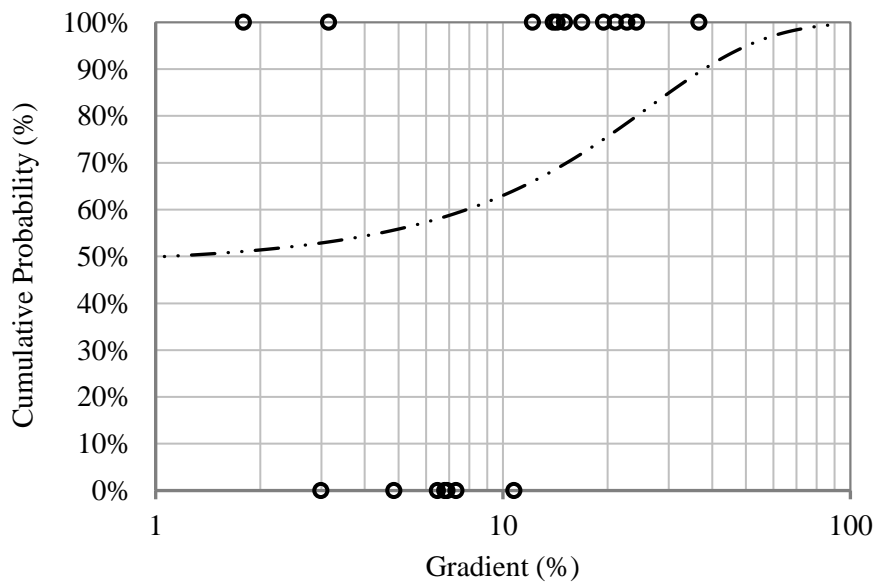


Figure 5.6 Gradient vs cumulative probability of ordinal logistic regression model to predict site with  $V_{s30} > 760$  m/s for geological class  $C_4$

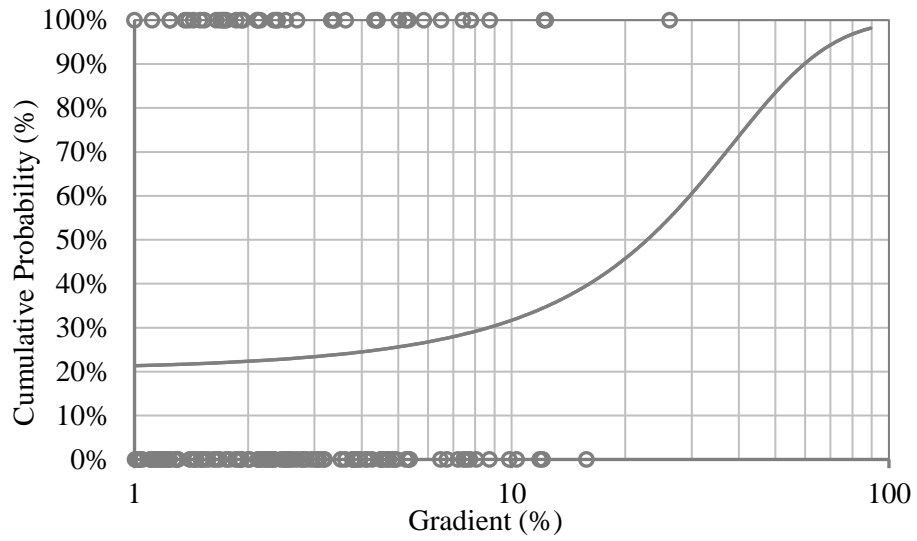


Figure 5.7 Gradient vs cumulative probability of ordinal logistic regression model to predict site with  $V_{s30} > 360$  m/s for geological class  $C_1$

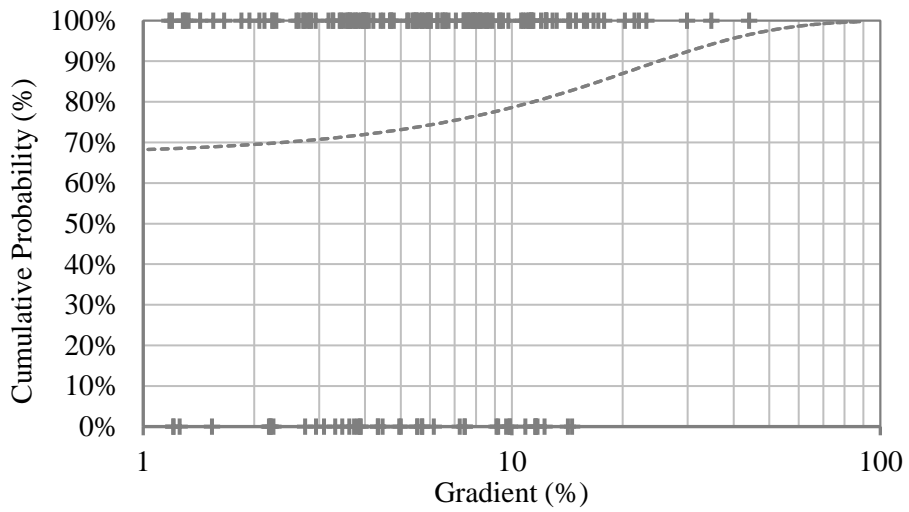


Figure 5.8 Gradient vs cumulative probability of ordinal logistic regression model to predict site with  $V_{s30} > 360$  m/s for geological class  $C_2$

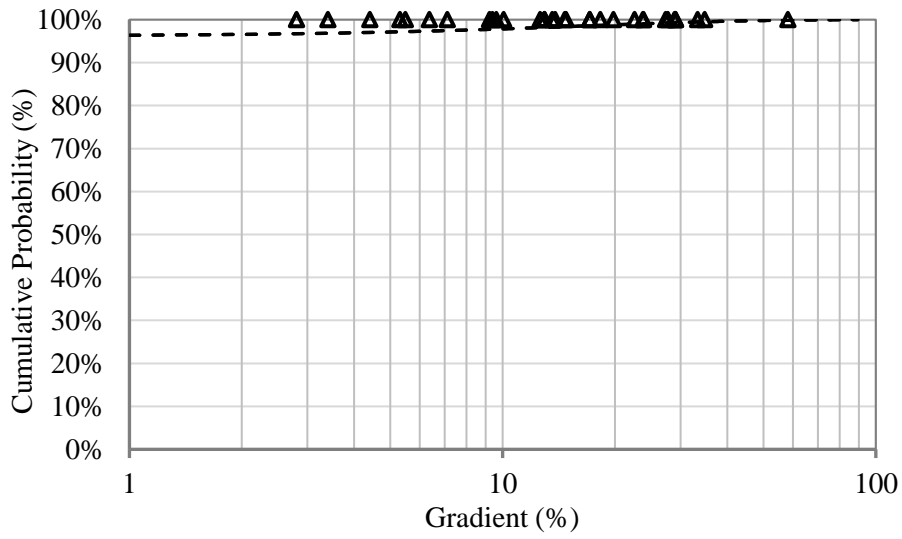


Figure 5.9 Gradient vs cumulative probability of ordinal logistic regression model to predict site with  $V_{s30} > 360$  m/s for geological class  $C_3$

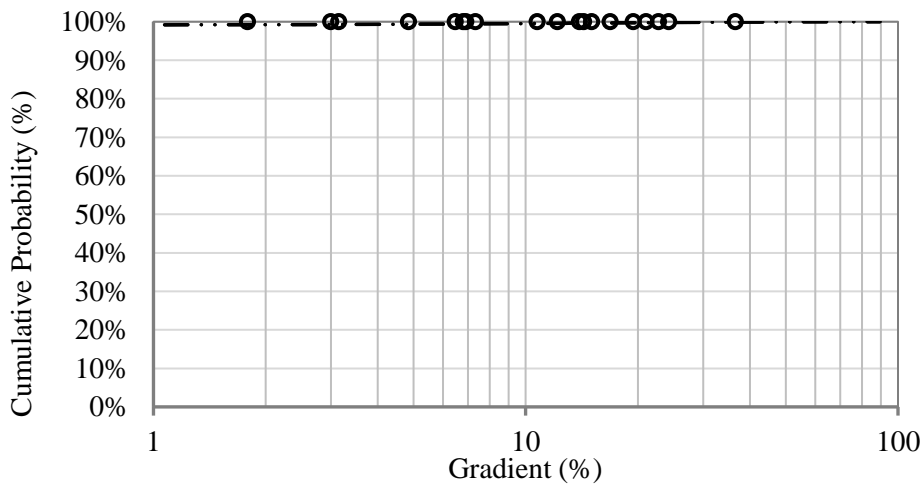


Figure 5.10 Gradient vs cumulative probability of ordinal logistic regression model to predict site with  $V_{s30} > 360$  m/s for geological class  $C_4$

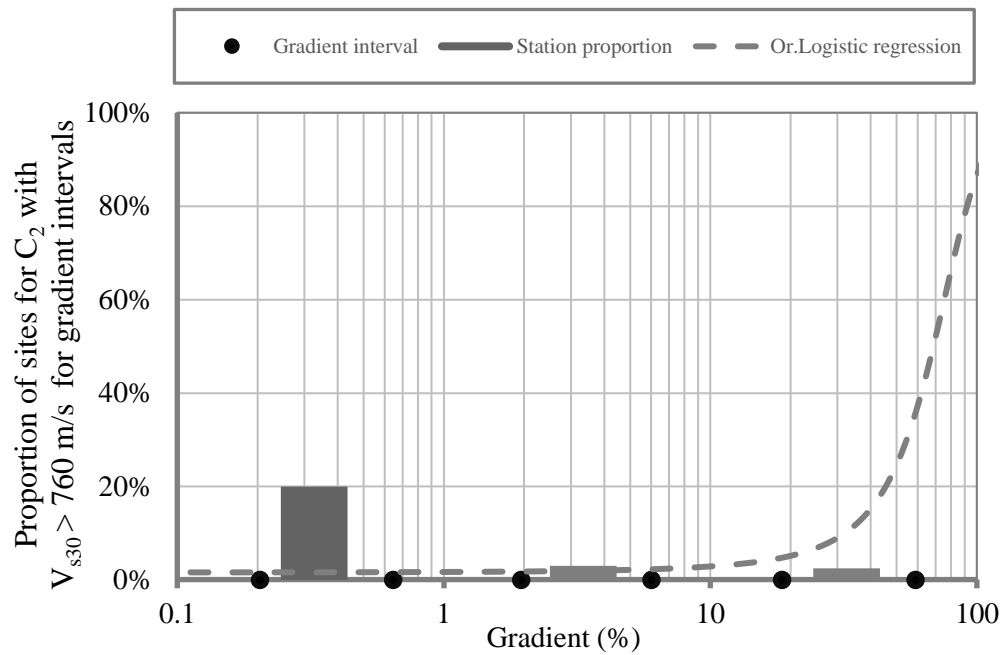


Figure 5.11 The change on proportion of sites with  $V_{s30} > 760$  m/s according to the gradient intervals for geological class  $C_2$

Table 5.4 The sample for sites with  $V_{s30} > 760$  m/s in Figure 5.11

Gradient intervals	No. of sites in each interval	No. of sites in each interval with $V_{s30} > 760$ m/s
[0.006 - 0.645]	5	1
(0.645 - 6.025]	99	3
(6.025 - 58.884]	80	2

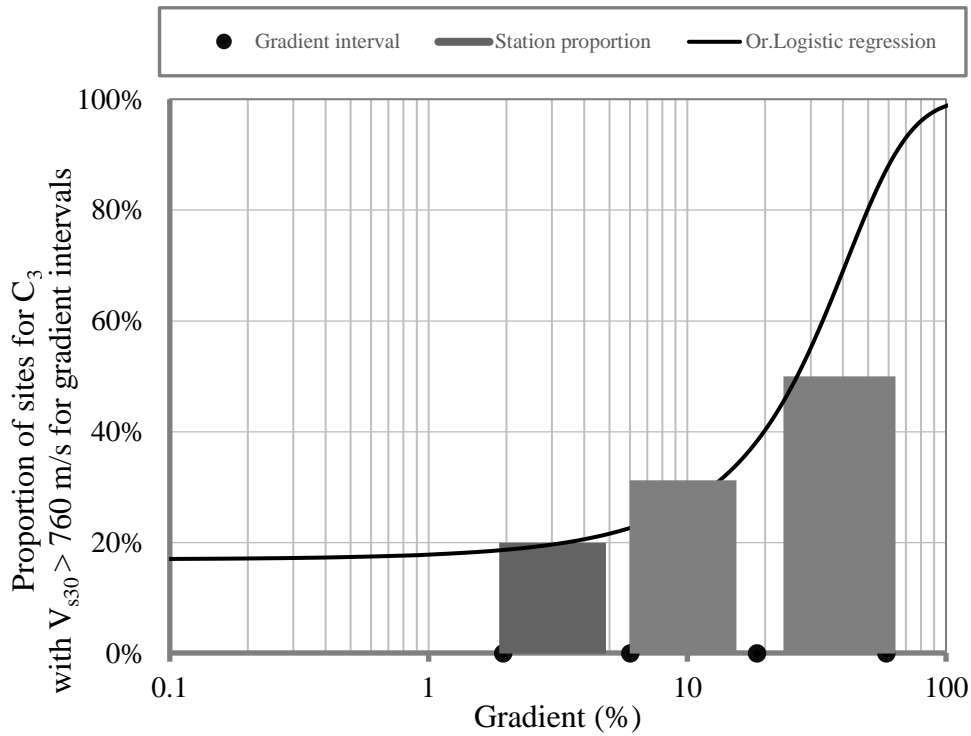


Figure 5.12 The change on proportion of sites with  $V_{s30} > 760$  m/s according to the gradient intervals for geological class  $C_3$

Table 5.5 The sample for sites with  $V_{s30} > 760$  m/s in Figure 5.20

Gradient intervals	No. of sites in each interval	No. of sites in each interval with $V_{s30} > 760$ m/s
[0.006 - 6.025]	5	1
(6.025 - 18.620]	16	5
(18.620 - 58.884]	12	6

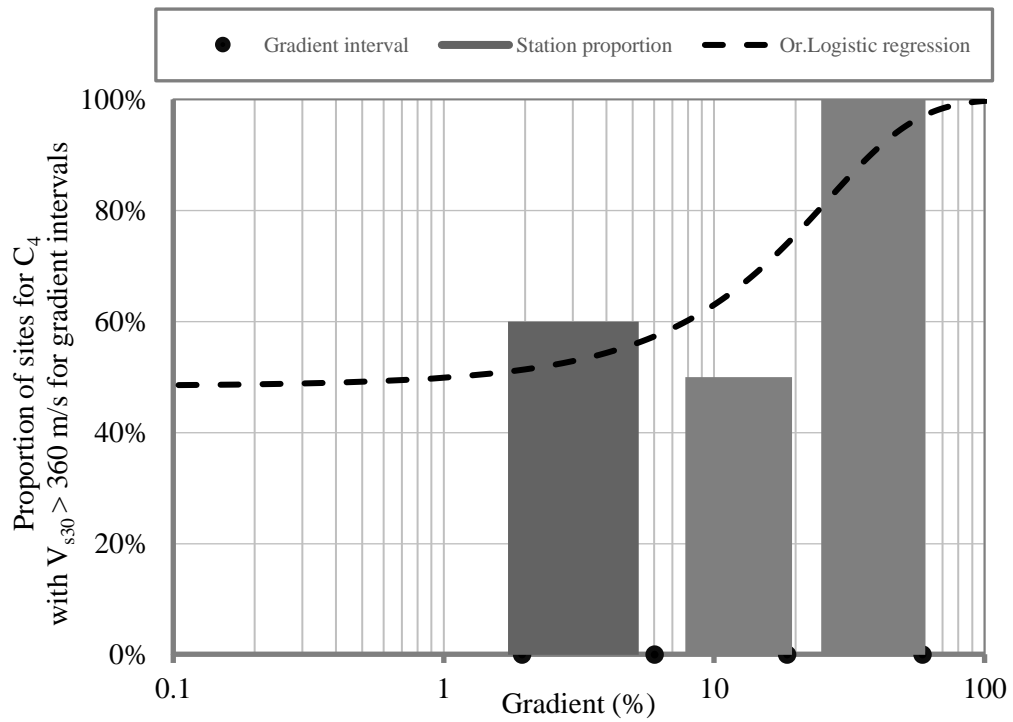


Figure 5.13 The change on proportion of sites with  $V_{s30} > 760$  m/s according to the gradient intervals for geological class  $C_4$

Table 5.6 The sample for sites with  $V_{s30} > 760$  m/s in Figure 5.13

Gradient intervals	No. of sites in each interval	No. of sites in each interval with $V_{s30} > 760$ m/s
[0.006 - 6.025]	5	3
(6.025 - 18.620]	10	5
(18.620 - 58.884]	5	5

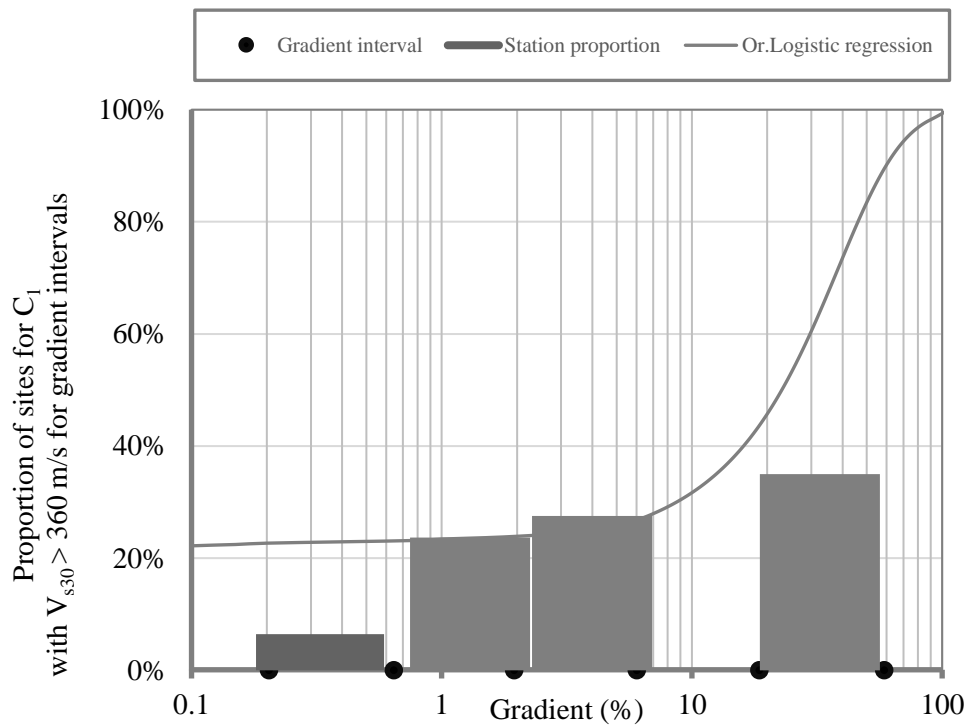


Figure 5.14 The change on proportion of sites with  $V_{s30} > 360$  m/s according to the gradient intervals for geological class  $C_1$

Table 5.7 The sample for sites with  $V_{s30} > 360$  m/s in Figure 5.14

Gradient intervals	No. of sites in each interval	No. of sites in each interval with $V_{s30} > 360$ m/s
[0.006 - 0.645]	31	2
(0.645 - 1.948]	76	18
(1.948,6.025]	69	19
(6.025,58.884]	20	7

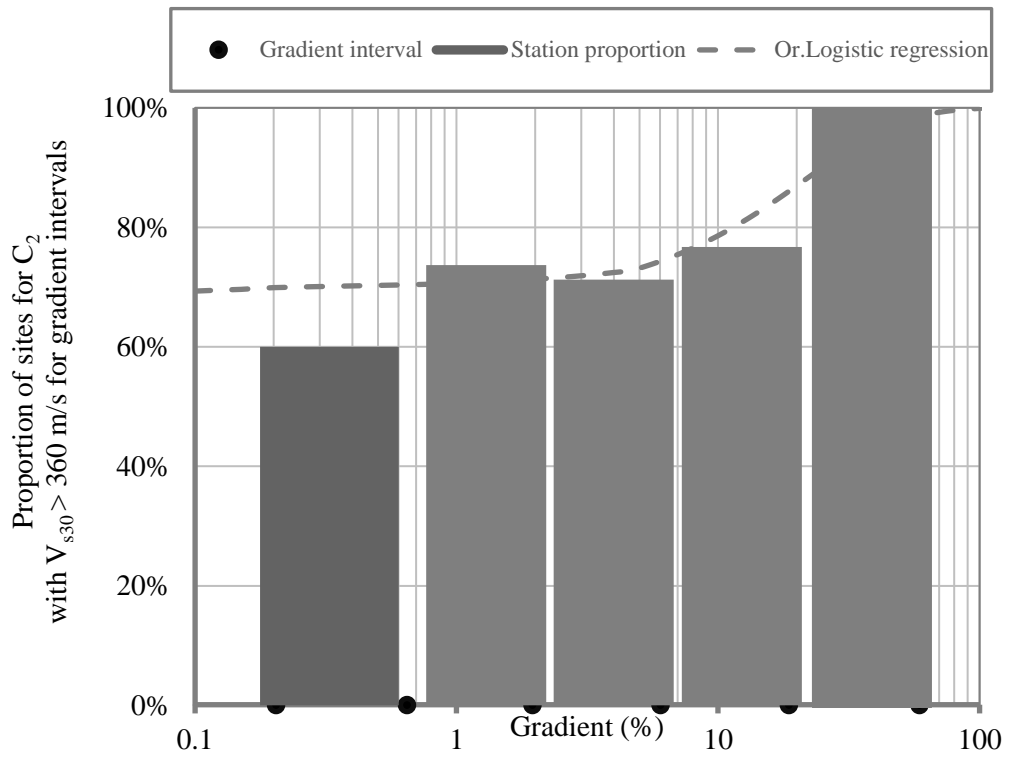


Figure 5.15 The change on proportion of sites with  $V_{s30} > 360$  m/s according to the gradient intervals for geological class  $C_2$

Table 5.8 The sample for sites with  $V_{s30} > 360$  m/s in Figure 5.15

Gradient intervals	No. of sites in each interval	No. of sites in each interval with $V_{s30} > 360$ m/s
[0.006 - 0.645]	5	3
(0.645 - 1.948]	14	19
(1.948,6.025]	57	80
(6.025,18.620]	56	73
(18.620 - 58.884]	7	7

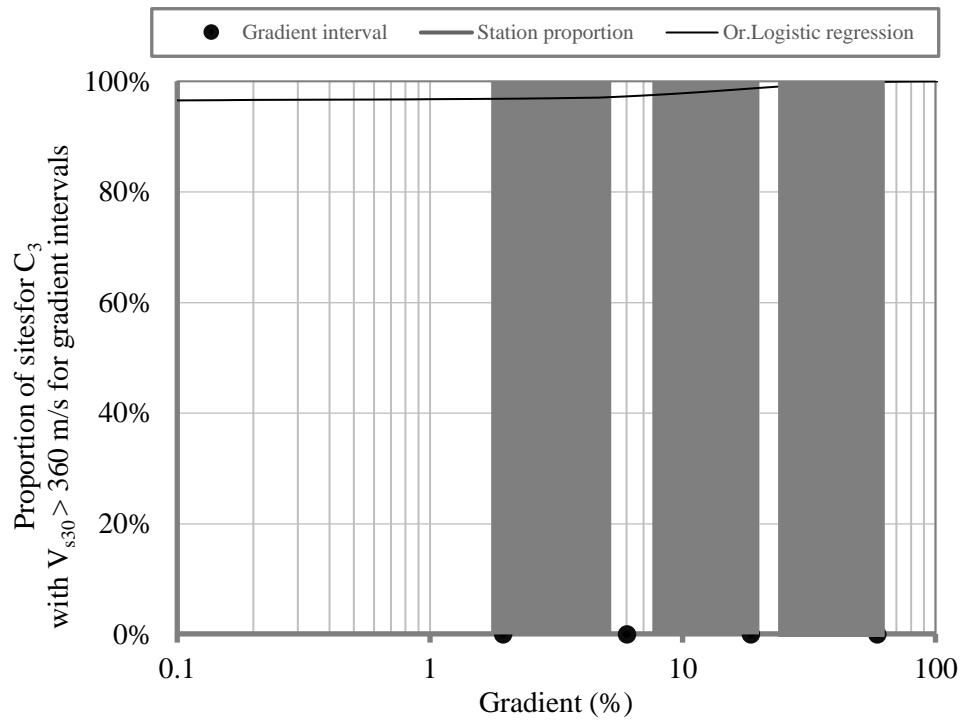


Figure 5.16 The change on proportion of sites with  $V_{s30} > 360$  m/s according to the gradient intervals for geological class  $C_3$

Table 5.9 The sample for sites with  $V_{s30} > 360$  m/s in Figure 5.16

Gradient intervals	No. of sites in each interval	No. of sites in each interval with $V_{s30} > 360$ m/s
[0.006 -6.025]	5	5
(6.025 - 18.620]	16	16
(18.62 - 58.884]	12	12

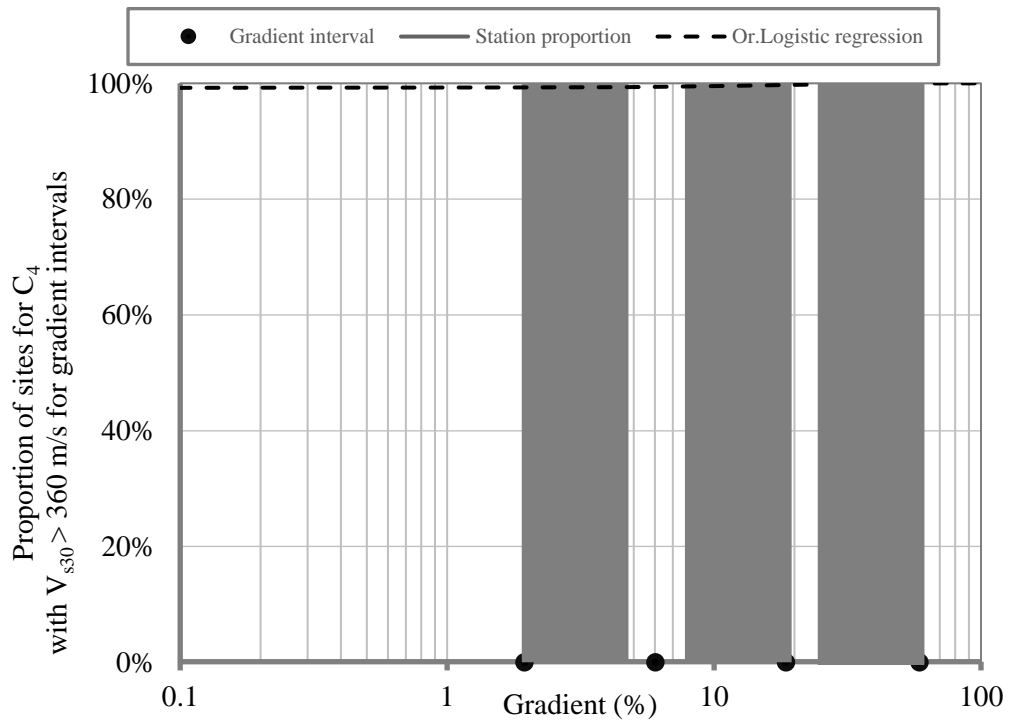


Figure 5.17 The change on proportion of sites with  $V_{s30} > 360$  m/s according to the gradient intervals for geological class  $C_4$

Table 5.10 The sample for sites with  $V_{s30} > 360$  m/s in Figure 5.17

Gradient intervals	No. of sites in each interval	No. of sites in each interval with $V_{s30} > 360$ m/s
[0.006 - 6.025]	5	5
(6.025 - 18.620]	10	10
(18.620 - 58.884]	5	5

The ordinal logistic regression models are compared with previously developed nonlinear categorical quantile regression models (Figure 5.18- Figure 5.23 ). Through this comparison for each geological class NCQRs are developed at different percentiles, 10%, 16%, 15%, 20%, 25%, 30%, 40% and 50%. For each percentile the gradient values at which the  $V_{s30}$  values are equal to 360 m/s or 760 m/s are found by backcalculation. In other words, given these gradient values, NCQRs present the probability,  $1-p$ , by which  $V_{s30}$  values are less than 360 m/s or 760 m/s for any specific geological class: 90%, 84%, 85%, 80%, 75%, 70%, 60%, and 50% respectively. Also, the ordinal logistic regression provides the estimation of the probability by which  $V_{s30}$  values will be more than 360 m/s or 760 m/s for a specific (conditional) gradient. So the conditional probability of exceedance,  $p$ , estimated by the ordinal logistic regression model is consistent with the percentiles of NCQR, but  $1-p$  is to be considered for the latter.

Considering figure 5.18 and Figure 5.19 for the geological class  $C_1$ , the results submitted by NCQR models match with those presented by ordinal logistic regression for  $V_{s30} > 360$  m/s and  $V_{s30} > 760$  m/s, except at 90 % and 50% percentiles. Figure 5.20 demonstrates that for geological class  $C_2$  for all percentiles, there is a good agreement between NCQR and ordinal logistic regression model particularly for the event  $V_{s30} > 360$  m/s. On the other hand, for the event  $V_{s30} > 760$  m/s, NCQR models are not in good agreement with ordinal logistic regression at 90% and 50% percentiles, (Figure 5.21 ). These observations may yield to following conclusions:

- i) At least one of these methods is not accurate for the extreme probability range,  $p \geq 90\%$ , possibly due to the limited data size, the final reliable probability/percentile limit is around 85% for extremes (i.e., minimum possible conditional  $V_{s30}$ ).
- ii) At least one of these methods cannot yield accurate prediction of the median value, possibly due to the functional form. A comparison with least-squares estimations may clarify this issue.

Since for geological classes  $C_3$  and  $C_4$  NCQR models for all percentiles present high  $V_{s30}$  values, it is not possible to obtain the gradient at which  $V_{s30}$  value is equal to 360 m/s. Consequently, for these geological classes the ordinal logistic regression models are compared with NCQR in the range  $V_{s30} > 760$  m/s. For geological class  $C_3$ , although the results of NCQR do not fit to those of the ordinal logistic regression at 50% and 90% conditional probabilities, but the agreement for other percentiles between these two (Figure 5.22 ) is satisfactory. This is consistent with the previous conclusions.

Considering Figure 5.23 , for most of the percentiles NCQR's results are not in good match with those obtained from ordinal logistic regression, but this can be explained the limitation of the sample size (see Table 5.6 ), such that the proportion of the successful observations ( $V_{s30}>760$  m/s) suddenly jumps to 100% when the gradients exceed 10% among class  $C_4$  sites. In that case, the functional forms used in regressions become critically important for probabilistic estimations. Further data will be necessary to solve this issue, but the practical consequences of is rather limited since  $V_{s30}$  is usually high for  $C_4$ -class sites no matter what topographical parameters are.

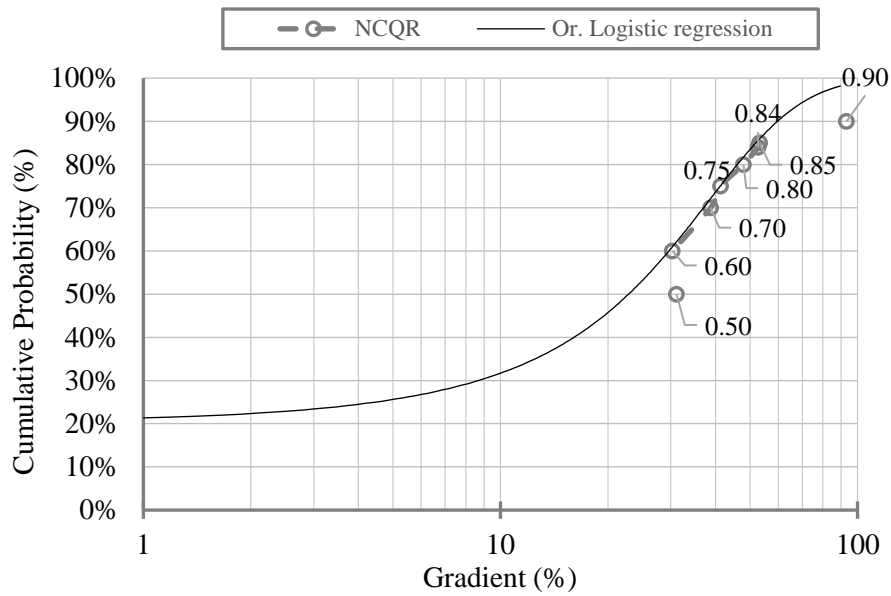


Figure 5.18 NCQR vs ordinal regression model for geological class C<sub>1</sub> with  $V_{s30} > 360$  m/s

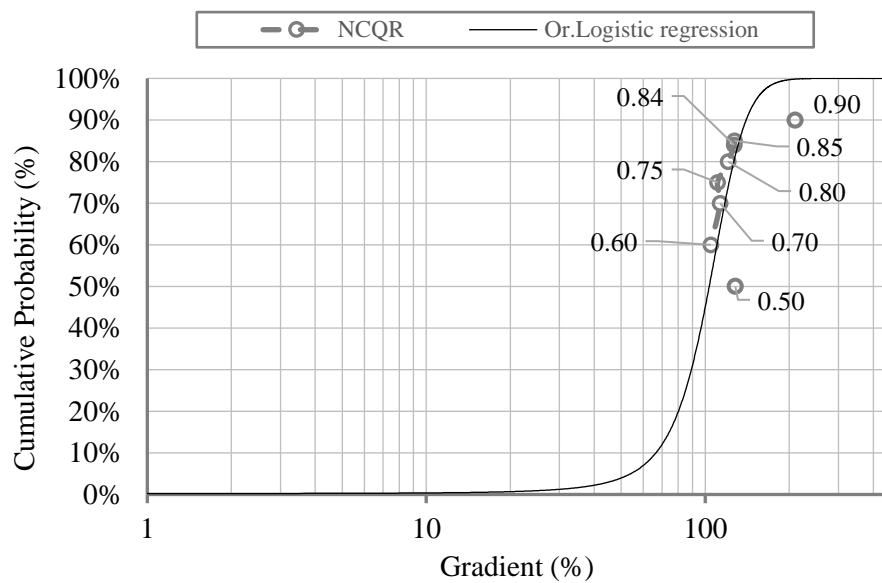


Figure 5.19 NCQR vs ordinal regression model for geological class C<sub>1</sub> with  $V_{s30} > 760$  m/s

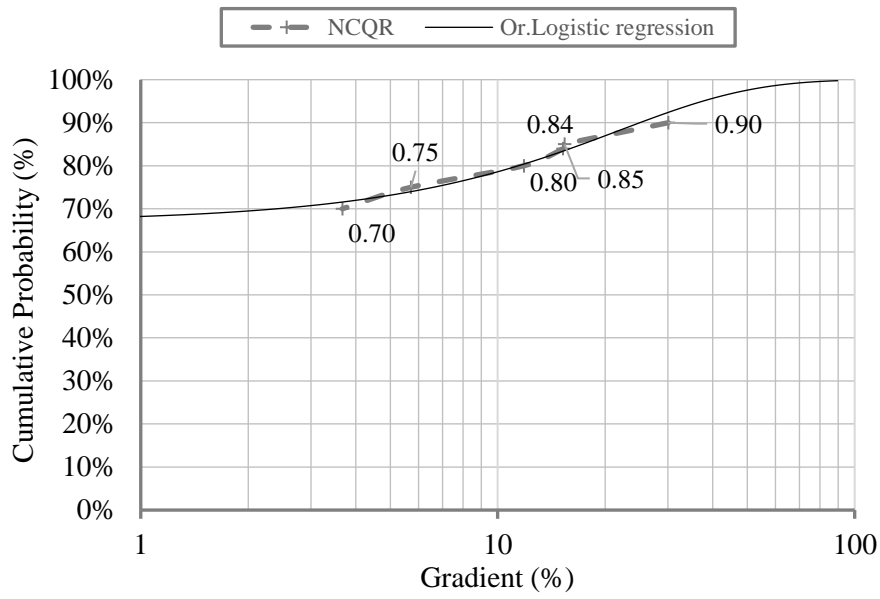


Figure 5.20 NCQR vs ordinal regression model for geological class  $C_2$  with  $V_{s30} > 360$  m/s

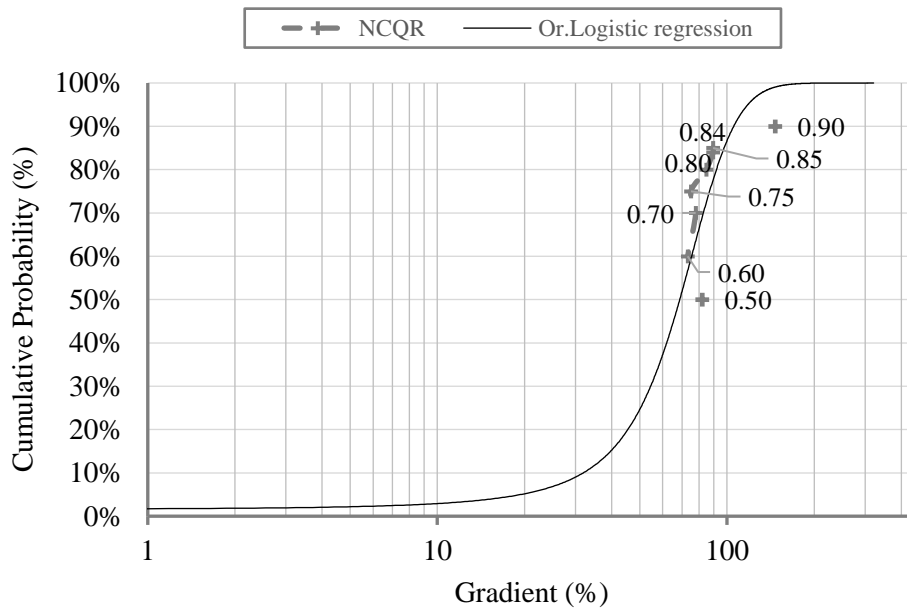


Figure 5.21 NCQR vs ordinal regression model for geological class  $C_2$  with  $V_{s30} > 760$  m/s

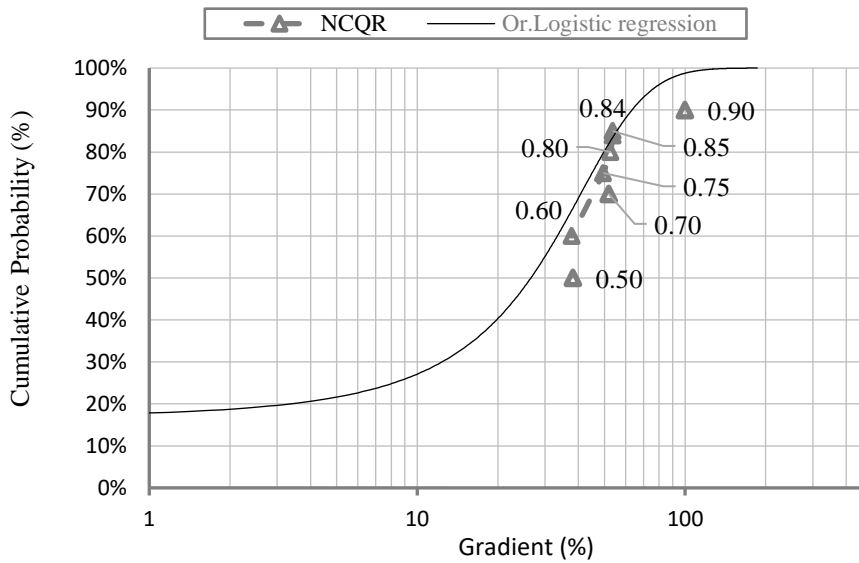


Figure 5.22 NCQR vs ordinal regression model for geological class C<sub>3</sub> with  $V_{s30} > 760$  m/s

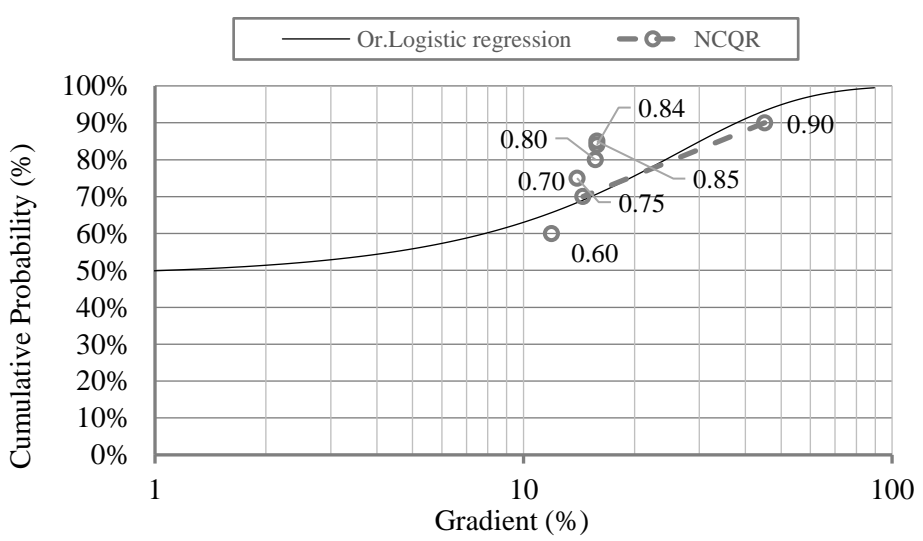


Figure 5.23 NCQR vs ordinal regression model for geological class C<sub>4</sub> with  $V_{s30} > 760$  m/s

## CHAPTER 6

### CONCLUSION

#### 6.1 General

The aim of this research was to develop a model to predict  $V_{s30}$  values for Türkiye according to topographic and geologic restrictions. To this end, nonlinear categorical quantile regression models were developed and evaluated according to corresponding goodness of fit criteria. Additionally, multinomial and ordinal logistic regression models were established, as classifiers, according to Türkiye's  $V_{s30}$  catalog using topographic and geological restrictions to categorize sites into different soil types. The ordinal logistic regression is more emphasized since  $V_{s30}$  is a positive real number, a continuous prediction, not a category. McFadden's adjusted  $R^2$  was referred as a criterion to select the most appropriate classifiers. Furthermore, the regression models developed through the current research are compared with those constructed previously.

This chapter of the thesis provides a summary of the research, conclusions, and contributions of the current study and offers recommendations for future studies.

## 6.2 Conclusions

The conclusions of the study are presented in the following list.

- When the conditional distribution of  $V_{s30}$  given the predictor variables deviates from being symmetrical, the probabilistic estimations of  $V_{s30}$  cannot be possible by least-squares estimations that are based on the assumption of normality. In such cases, it is recommended to employ quantile regression models to estimate conditional probabilistic limits for  $V_{s30}$ .
- Considering the  $R^2$  and adjusted  $R^2$  values, the nonlinear categorical quantile regression model involving five distinct estimator (condition) terms, 4 geological age classes and topographical gradient (slope), emerges as the most suitable option for estimating  $V_{s30}$  values along with the observed behavior of the model. This model is supported by its strong alignment with theoretical expectations, where sites on higher topographical gradients and older geological ages exhibit higher  $V_{s30}$  values.
- Comparing the  $R^2$  and adjusted  $R^2$  values, it can be determined that geological age class has predominant role in developing nonlinear categorical quantile regression to predict  $V_{s30}$  values, and topographical gradient has a secondary effect.
- Comparing the developed NCQR model (Equation 4.16) with the previously constructed models, (Allen and Wald., 2007.; Iwahashi et al., 2010, Sahin et al., 2024), the confidence interval for proportion of successful events shows that NCQR model developed at 50% percentile and nonlinear set of models prepared for each geological classes separately (Equation 3.6-Equation 3.9) depict more accurate predictions.
- Due to the lack of data in soil type classes  $Z_A$  ( $V_{s30} > 1500$  m/s), and  $Z_E$ , ( $V_{s30} < 180$  m/s), the regression models can be specifically developed for site classes  $Z_B$ ,  $Z_C$ , and  $Z_D$  only.
- Analysis of  $R_{MF}^2$  and  $R_{MFA}^2$  in ordinal regression models reveals that geological age and topographical gradient (slope) are the essential parameters for

developing country-wide site-class maps of Turkiye. The age is the predominant parameter, though. The sites with higher gradient and older geological age tend to have higher  $V_{s30}$  as expected.

- Evaluations of the ordinal logistic regression models within gradient intervals for four geological-age classes for the events  $V_{s30} > 360$  m/s and  $V_{s30} > 760$  m/s show that, the ordinal logistic regression models have acceptable performance in predicting the probability of the event  $V_{s30} > 360$  m/s for all geological classes. On the other hand, its performance is not satisfactory within high gradient intervals for gradients exceeding 10% for the probability of  $V_{s30} > 760$  m/s. This can be explained by the lack of sufficient data size for the steep ground conditions, though.
- Referring to the comparison of developed NCQR and ordinal logistic regression models, both models show consistent conditional probability estimations for percentiles in the range  $50\% < p < 90\%$ , when  $V_{s30} > 360$  m/s and  $V_{s30} > 760$  m/s) except for the oldest geological age, possibly due to limitations in the data. Nonetheless, the two models are not consistent for the median prediction probability 50%, and the extreme probability 90%. The difference for 50% can be explained by the differences in functional forms, but the extreme probability range can be explained by the lack of sufficient data.

### 6.3 Recommendations for Future Study

The following studies are deemed as necessary.

- Referring to the comparison of developed NCQR and ordinal logistic regression models, both models show consistent conditional probability estimations for  $V_{s30}$ . However, a risk analysis is necessary to determine a probabilistic or percentile limit for minimum  $V_{s30}$ , conditional to geological age and topographical gradient (slope).
- All samples are compiled from AFAD's strong motion sites, which are installed without considering site conditions. With the mean estimations, and likelihood of extreme ranges for  $V_{s30}$ , a method for selection of optimum locations for new site sites can be developed. Particularly, data for soil classes ZE ( $V_{s30} < 180$  m/s) and ZA ( $V_{s30} > 1500$  m/s) is necessary for completeness of the sample set. Besides, the sample should be enriched for steeper slopes, particularly for topographical gradients exceeding 10%.
- This study can be easily revised by further data from strong motion site sites, which provides a reasonably uniform sampling of  $V_{s30}$  over Turkiye. Improving the digital geological maps or providing borehole data that is identifying the geological formations according to their geological ages is apparently critical. Consequently, a visit to the sites for which  $V_{s30}$  could not be attributed to geological age will be necessary.

## REFERENCES

- Abrahamson, N. and Silva, W., 2008. Summary of the Abrahamson & Silva NGA ground-motion relations. *Earthquake spectra*, 24(1), pp.67-97.
- Akkar, S., Çağnan, Z., Yenier, E., Erdoğan, Ö., Sandikkaya, M.A. and Gülkan, P., 2010. The recently compiled Turkish strong motion database: preliminary investigation for seismological parameters. *Journal of Seismology*, 14, pp.457-479.
- Allen, T.I. and Wald, D.J., 2007. *Topographic slope as a proxy for seismic site-conditions (VS30) and amplification around the globe* (No. 2007-1357). Geological Survey (US).
- Allen, T.I. and Wald, D.J., 2009. On the use of high-resolution topographic data as a proxy for seismic site conditions (VS 30). *Bulletin of the Seismological Society of America*, 99(2A), pp.935-943.
- Ambrus, A., Alyaev, S., Jahani, N., Pacis, F.J. and Wiktorski, T., 2022, June. Rate of penetration prediction using quantile regression deep neural networks. In *International Conference on Offshore Mechanics and Arctic Engineering* (Vol. 85956, p. V010T11A010). American Society of Mechanical Engineers.
- Augusti, G. and Schuëller, G.I. eds., 2005. *Safety and Reliability of Engineering Systems and Structures*. Millpress.
- Ayalew, L. and Yamagishi, H., 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology*, 65(1-2), pp.15-31.
- Azul, K., Orense, R. and Wotherspoon, L., 2023. Framework Development for a Hybrid Geotechnical-Geospatial Liquefaction Assessment Model.
- Bai, S.B., Wang, J., Lü, G.N., Zhou, P.G., Hou, S.S. and Xu, S.N., 2010. GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. *Geomorphology*, 115(1-2), pp.23-31.

- Boore, D. M. (1997). Equations for Estimating Horizontal Response Spectra and Peak Ground Acceleration from Western North American Earthquakes: A Summary of Recent Work. *Seism. Res. Lett.*, 68, 154–179.
- Boore, D.M., 2004. Can site response be predicted? *Journal of earthquake Engineering*, (spec01), pp.1-41.
- Boore, D.M., 2006, August. Determining subsurface shear-wave velocities: a review. In *Third International Symposium on the Effects of Surface Geology on Seismic Motion* (Vol. 30, pp. 67-85). Grenoble, France.
- Boore, D.M., Joyner, W.B. and Fumal, T.E., 1993. *Estimation of response spectra and peak accelerations from western North American earthquakes; an interim report* (No. 93-509). US Geological Survey.
- Boore, D.M., Joyner, W.B. and Fumal, T.E., 1997. Equations for estimating horizontal response spectra and peak acceleration from western North American earthquakes: a summary of recent work. *Seismological research letters*, 68(1), pp.128-153.
- Borcherdt, R.D., 1994. Estimates of site-dependent response spectra for design (methodology and justification). *Earthquake spectra*, 10(4), pp.617-653.
- Bozorgzadeh, N. and Harrison, J.P., 2015, June. Characteristic triaxial strength of intact rock. In *ARMA US Rock Mechanics/Geomechanics Symposium* (pp. ARMA-2015). ARMA.
- Bozzoni, F., Bonì, R., Conca, D., Lai, C.G., Zuccolo, E. and Meisina, C., 2021. Megazonation of earthquake-induced soil liquefaction hazard in continental Europe. *Bulletin of Earthquake Engineering*, 19, pp.4059-4082.
- Briollais, L. and Durrieu, G., 2014. Application of quantile regression to recent genetic and-omic studies. *Human genetics*, 133(8), pp.951-966.
- Buhai, S., 2005. Quantile regression: overview and selected applications. *Ad Astra*, 4(4), pp.1-17.

- Bui, D. T., Lofman, O., Revhaug, I., and Dick, O. (2011). Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. *Natural Hazards*, 59, 1413–1444.
- Chen, M., Wei, W. and Jiang, Q., 2022. Use of Quantile Regression with Fukui–Okubo Model for Prediction and Early Warning of Reservoir Bank Slope Failure. *Rock Mechanics and Rock Engineering*, 55(11), pp.7145-7169.
- Castellaro, S., Mulargia, F. and Rossi, P.L., 2008. VS30: Proxy for seismic amplification. *seismological research letters*, 79(4), pp.540-543.
- Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V. and Böhner, J., 2015. System for automated geoscientific analyses (SAGA) v. 2.1. 4. *Geoscientific model development*, 8(7), pp.1991-2007.
- Crespo, M.J., Benjumea, B., Moratalla, J.M., Lacoma, L., Macau, A., González, Á., Gutiérrez, F. and Stafford, P.J., 2022. A proxy-based model for estimating VS30 in the Iberian Peninsula. *Soil Dynamics and Earthquake Engineering*, 155, p.107165.
- Dai, F.C. and Lee, C.F., 2001. Terrain-based mapping of landslide susceptibility using a geographical information system: a case study. *Canadian Geotechnical Journal*, 38(5), pp.911-923.
- Das, I., Sahoo, S., van Westen, C., Stein, A. and Hack, R., 2010. Landslide susceptibility assessment using logistic regression and its comparison with a rock mass classification system, along a road section in the northern Himalayas (India). *Geomorphology*, 114(4), pp.627-637.
- Davino, C., Furno, M. and Vistocco, D., 2013. *Quantile regression: theory and applications* (Vol. 988). John Wiley and Sons.
- Davis, J.C. and Sampson, R.J., 1986. *Statistics and data analysis in geology* (Vol. 646). New York: Wiley.

- Devore, J.L., Berk, K.N. and Carlton, M.A., 2012. *Modern mathematical statistics with applications* (Vol. 285). New York: Springer.
- Duan, W., Congress, S.S.C., Cai, G., Liu, S., Dong, X., Chen, R. and Liu, X., 2021. A hybrid GMDH neural network and logistic regression framework for state parameter-based liquefaction evaluation. *Canadian Geotechnical Journal*, 99(999), pp.1801-1811.
- Farr, T.G. and Kobrick, M., 2000. Shuttle Radar Topography Mission produces a wealth of data. *Eos, Transactions American Geophysical Union*, 81(48), pp.583-585.
- Feng, R., Grana, D. and Balling, N., 2021. Imputation of missing well log data by random forest and its uncertainty analysis. *Computers and Geosciences*, 152, p.104763.
- Field, E.H., 2000. A modified ground-motion attenuation relationship for southern California that accounts for detailed site classification and a basin-depth effect. *Bulletin of the Seismological Society of America*, 90(6B), pp. S209-S221.
- Fitzenberger, B., Koenker, R. and Machado, J.A. eds., 2013. *Economic applications of quantile regression*. Springer Science and Business Media.
- Frankel, A.D., Carver, D.L. and Williams, R.A., 2002. Nonlinear and linear site response and basin effects in Seattle for the M 6.8 Nisqually, Washington, earthquake. *Bulletin of the Seismological Society of America*, 92(6), pp.2090-2109.
- Fullerton, D.S., Bush, C.A. and Pennell, J.N., 2004. Map of surficial deposits and materials in the eastern and central United States (East of 102 West Longitude). *US Geological Survey Geologic Investigation Series I*, 2789.
- Fumal, T.E., 1978. *Correlations between seismic wave velocities and physical properties of near-surface geologic materials in the southern San Francisco Bay region, California* (No. 78-1067). US Geological Survey.

- Gandomi, A.H., Fridline, M.M. and Roke, D.A., 2013. Decision tree approach for soil liquefaction assessment. *The Scientific World Journal*, 2013(1), p.346285.
- Gavin, H.P., 2019. The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems. *Department of Civil and Environmental Engineering Duke University August*, 3.
- Ghasemi, E. and Gholizadeh, H., 2019. Development of two empirical correlations for tunnel squeezing prediction using binary logistic regression and linear discriminant analysis. *Geotechnical and Geological Engineering*, 37, pp.3435-3446.
- Gkoutakou, F.I., Bantilas, K.E., Kavvadias, I.E., Elenas, A. and Papadopoulos, B.K., 2023. Fuzzy Multivariate Regression Models for Seismic Assessment of Rocking Structures. *Applied Sciences*, 13(17), p.9602.
- Han, J., Park, S., Kim, S., Son, S., Lee, S. and Kim, J., 2019. Performance of logistic regression and support vector machines for seismic vulnerability assessment and mapping: a case study of the 12 September 2016 ML5. 8 Gyeongju Earthquake, South Korea. *Sustainability*, 11(24), p.7038.
- Hao, L. and Naiman, D.Q., 2007. *Quantile regression* (No. 149). Sage.
- Heath, D.C., Wald, D.J., Worden, C.B., Thompson, E.M. and Smoczyk, G.M., 2020. A global hybrid VS 30 map with a topographic slope-based default and regional map insets. *Earthquake Spectra*, 36(3), pp.1570-1584.
- Hemasinghe, H., Rangali, R.S.S., Deshapriya, N.L. and Samarakoon, L., 2018. Landslide susceptibility mapping using logistic regression model (a case study in Badulla District, Sri Lanka). *Procedia engineering*, 212, pp.1046-1053.
- Hoffmann, J.P., 2021. Linear regression models: applications in R. Chapman and Hall/CRC.

- Holzer, T.L., Bennett, M.J., Noce, T.E. and Tinsley III, J.C., 2005. Shear-wave velocity of surficial geologic sediments in northern California: statistical distributions and depth dependence. *Earthquake Spectra*, 21(1), pp.161-177.
- Holzer, T.L., Padovani, A.C., Bennett, M.J., Noce, T.E. and Tinsley III, J.C., 2005. Mapping NEHRP VS30 site classes. *Earthquake Spectra*, 21(2), pp.353-370.
- Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression*. John Wiley and Sons.
- Hunter, J.A., Crow, H.L., Brooks, G.R., Pyne, M., Motazedian, D., Lamontagne, M., Pugin, A.J.M., Pullan, S.E., Cartwright, T., Douma, M. and Burns, R.A., 2010. *Seismic site classification and site period mapping in the Ottawa area using geophysical methods*. Geological Survey of Canada= Commission géologique du Canada.
- Iwahashi, J., Kamiya, I. and Matsuoka, M., 2010. Regression analysis of Vs30 using topographic attributes from a 50-m DEM. *Geomorphology*, 117(1-2), pp.202-205.
- Iwahashi, J., Kamiya, I., Matsuoka, M. and Yamazaki, D., 2018. Global terrain classification using 280 m DEMs: segmentation, clustering, and reclassification. *Progress in Earth and Planetary Science*, 5, pp.1-31.
- Iwahashi, J. and Pike, R.J., 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology*, 86(3-4), pp.409-440.
- Jairi, I., Fang, Y. and Pirhadi, N., 2021. Application of logistic regression based on maximum likelihood estimation to predict seismic soil liquefaction occurrence. *Human-Centric Intelligent Systems*, 1(3), pp.98-104.
- Juang, C.H., Jiang, T., Andrus, R.D. and Lee, D.H., 2001. *Assessing Probabilistic Methods for Liquefaction Potential Evaluation—An Update*.

- Kameshwar, S. and Padgett, J.E., 2014. Multi-hazard risk assessment of highway bridges subjected to earthquake and hurricane hazards. *Engineering Structures*, 78, pp.154-166.
- Kang, H.-S., & Kim, Y.-T. (2016). A Study on Warning Level-based-Landslide Triggering Rainfall Criteria considering Weathered Soil Type and Landslide Type. *Journal of Korean Society of Hazard Mitigation*, 16, 341–350. <https://doi.org/10.9798/KOSHAM.2016.16.2.341>
- Karimzadeh, S., Feizizadeh, B. and Matsuoka, M., 2019. DEM-based Vs30 map and terrain surface classification in nationwide scale—A case study in Iran. *ISPRS International Journal of Geo-Information*, 8(12), p.537.
- Kelly, D., 2006. Seismic site classification for structural engineers. *Structure*, 21, pp.21-24.
- Kiani, J., Camp, C. and Pezeshk, S., 2019. On the application of machine learning techniques to derive seismic fragility curves. *Computers & Structures*, 218, pp.108-122.
- Kim, S.W., Chun, K.W., Kim, M., Catani, F., Choi, B. and Seo, J.I., 2021. Effect of antecedent rainfall conditions and their variations on shallow landslide-triggering rainfall thresholds in South Korea. *Landslides*, 18, pp.569-582.
- Koenker, R.W. and Bassett, J., Gilbert (1978):“Regression Quantiles,”*Econometrica*, 46(1).
- Koenker, R. and Machado, J.A., 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448), pp.1296-1310.
- Kramer, S.L., 1996. *Geotechnical earthquake engineering*. Pearson Education India.
- Kuo, C.H., Wen, K.L., Hsieh, H.H., Chang, T.M., Lin, C.M. and Chen, C.T., 2011. Evaluating empirical regression equations for Vs and estimating Vs30 in

- northeastern Taiwan. *Soil Dynamics and Earthquake Engineering*, 31(3), pp.431-439.
- Kurtuluş, C., Sertçelik, F., Sertcelik, I., Kuru, T., Tekin, K., Ates, E., Apak, A., Kokbudak, D., Sezer, S. and Yalcin, D., 2020. Determination of site characterization in Turkey strong motion recording stations. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 35(4), pp.1829-1846.
- Kurtulmus, T.O., Yerlikaya-Ozkurt, F. and Askan, A., 2023, November. Evaluation of Multivariate Adaptive Regression Splines for Prediction of Kappa Factor in Western Türkiye. In *International Conference on Energy and Environmental Science* (pp. 157-162). Cham: Springer Nature Switzerland.
- Lee, J.U., Cho, Y.C., Kim, M., Jang, S.J., Lee, J. and Kim, S., 2022. The effects of different geological conditions on landslide-triggering rainfall conditions in South Korea. *Water*, 14(13), p.2051.
- Lee, S. and Pradhan, B., 2007. Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. *Landslides*, 4(1), pp.33-41.
- Lemoine, A., Douglas, J. and Cotton, F., 2012. Testing the applicability of correlations between topographic slope and VS 30 for Europe. *Bulletin of the Seismological Society of America*, 102(6), pp.2585-2599.
- Li, B., Ling, Z., Zhang, J., Chen, J., Wu, Z., Ni, Y. and Zhao, H., 2015. Texture descriptions of lunar surface derived from LOLA data: Kilometer-scale roughness and entropy maps. *Planetary and Space Science*, 117, pp.303-311.
- Lin, J., Moon, S., Yong, A., Meng, L. and Davis, P., 2019. Length-Scale-Dependent Relationships between VS 30 and Topographic Slopes in Southern California. *Bulletin of the Seismological Society of America*, 109(6), pp.2614-2625.
- Lombardo, L. and Mai, P.M., 2018. Presenting logistic regression-based landslide susceptibility results. *Engineering geology*, 244, pp.14-24.

- Ma, J., Niu, X., Tang, H., Wang, Y., Wen, T. and Zhang, J., 2020. Displacement prediction of a complex landslide in the Three Gorges Reservoir Area (China) using a hybrid computational intelligence approach. *Complexity*, 2020(1), p.2624547.
- Magistrale, H., Rong, Y., Silva, W. and Thompson, E., 2012, December. A site response map of the continental US. In *Proc. of the Fifteenth World Conference on Earthquake Engineering*.
- Matsuoka, M. (1995). GIS-based integrated seismic hazard mapping for a large metropolitan area. *Proc. Fifth International Conference on Seismic Zonation*, 2, 1334–1341.
- Matsuoka, M., Wakamatsu, K., Fujimoto, K. and Midorikawa, S., 2005, June. Nationwide site amplification zoning using GIS-based Japan engineering geomorphologic classification map. In *Proc. 9th int. conf. on struct. Safety and reliability* (pp. 239-246).
- Maxwell, K., Rajabi, M. and Esterle, J., 2021. Spatial interpolation of coal properties using geographic quantile regression forest. *International Journal of Coal Geology*, 248, p.103869.
- McFadden, D., 1972. Conditional logit analysis of qualitative choice behavior.
- McGann, C.R., Bradley, B.A. and Cubrinovski, M., 2017. Development of a regional Vs30 model and typical Vs profiles for Christchurch, New Zealand from CPT data and region-specific CPT-Vs correlation. *Soil Dynamics and Earthquake Engineering*, 95, pp.48-60.
- Michelini, A., Faenza, L., Lauciani, V. and Malagnini, L., 2008. ShakeMap implementation in Italy. *Seismological Research Letters*, 79(5), pp.688-697.
- Mori, F., Gena, A., Mendicelli, A., Naso, G. and Spina, D., 2020. Seismic emergency system evaluation: The role of seismic hazard and local effects. *Engineering geology*, 270, p.105587.

- Mori, F., Mendicelli, A., Moscatelli, M., Romagnoli, G., Peronace, E. and Naso, G., 2020. A new Vs30 map for Italy based on the seismic microzonation dataset. *Engineering Geology*, 275, p.105745.
- Mousavi, S.M., Horton, S.P., Langston, C.A. and Samei, B., 2016. Seismic features and automatic discrimination of deep and shallow induced-microearthquakes using neural network and logistic regression. *Geophysical Journal International*, 207(1), pp.29-46.
- Ohlmacher, G.C. and Davis, J.C., 2003. Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. *Engineering geology*, 69(3-4), pp.331-343.
- Okalp, K., 2013. Landslide susceptibility assessment of Turkey using qualitative and semi-quantitative methods.
- Ozsagir, M., Erden, C., Bol, E., Sert, S. and Özocak, A., 2022. Machine learning approaches for prediction of fine-grained soils liquefaction. *Computers and Geotechnics*, 152, p.105014.
- Pain, C.F., 1985. Mapping of landforms from Landsat imagery: an example from eastern New South Wales, Australia. *Remote Sensing of environment*, 17(1), pp.55-65.
- Palei, S.K. and Das, S.K., 2009. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach. *Safety science*, 47(1), pp.88-96.
- Papathanassiou, G., 2008. LPI-based approach for calibrating the severity of liquefaction-induced failures and for assessing the probability of liquefaction surface evidence. *Engineering Geology*, 96(1-2), pp.94-104.
- Park, C.B., Miller, R.D. and Xia, J., 1999. Multichannel analysis of surface waves. *Geophysics*, 64(3), pp.800-808.

- Parker, G.A., Harmon, J.A., Stewart, J.P., Hashash, Y.M., Kottke, A.R., Rathje, E.M., Silva, W.J. and Campbell, K.W., 2017. Proxy-based VS 30 estimation in central and eastern North America. *Bulletin of the Seismological Society of America*, 107(1), pp.117-131.
- Powell, J.L., 1984. Least absolute deviations estimation for the censored regression model. *Journal of econometrics*, 25(3), pp.303-325.
- Powell, J.L., 1986. Censored regression quantiles. *Journal of econometrics*, 32(1), pp.143-155.
- Rafiee, R., Ataei, M. and Kamali, M., 2013. Tunnels stability analysis using binary and multinomial logistic regression (LR). *Journal of Geology and Mining Research*, 5(4), pp.97-107.
- Rawlings, J.O., Pantula, S.G. and Dickey, D.A. eds., 1998. *Applied regression analysis: a research tool*. New York, NY: Springer New York.
- Riley, S.J., DeGloria, S.D. and Elliot, R., 1999. Index that quantifies topographic heterogeneity. *intermountain Journal of sciences*, 5(1-4), pp.23-27.
- Romero, S.M., 2001. *Ground motion amplification of soils in the upper Mississippi embayment*. Georgia Institute of Technology.
- Sahin, G., Okalp, K., Kockar, M.K., Yilmaz, M.T., Jalehforouzan, A., Temiz, F.A., Askan, A., Akgun, H. and Erberik, M.A., 2024. Development of a GIS-based predicted-V s30 map of Türkiye by using geological and topographical parameters: Case study for the region affected by the 6 February 2023 Kahramanmaraş earthquakes. *Seismological Research Letters*.
- Saito, H., Nakayama, D. and Matsuyama, H., 2010. Relationship between the initiation of a shallow landslide and rainfall intensity—duration thresholds in Japan. *Geomorphology*, 118(1-2), pp.167-175.
- Salee, R., Chinkulkijniwat, A., Yubonchit, S., Horpibulsuk, S., Wangfaoklang, C. and Soisompong, S., 2022. New threshold for landslide warning in the southern part

- of Thailand integrates cumulative rainfall with event rainfall depth-duration. *Natural Hazards*, 113(1), pp.125-141.
- Sankaranarayanan, B., Abubakar, A., Allen, D.F. and Diaz Granados, I., 2021, September. Automating the log interpretation workflow using machine learning. In *SPE Annual Technical Conference and Exhibition* (p. D011S018R001). SPE.
- Saputra, A., Rahardianto, T., Revindo, M.D., Delikostidis, I., Hadmoko, D.S., Sartohadi, J. and Gomez, C., 2017. Seismic vulnerability assessment of residential buildings using logistic regression and geographic information system (GIS) in Pleret Sub District (Yogyakarta, Indonesia). *Geoenvironmental Disasters*, 4, pp.1-33.
- Schroeder, L.D., Sjoquist, D.L. and Stephan, P.E., 2016. Understanding regression analysis: An introductory guide (Vol. 57). Sage Publications.
- Shang, R., Peng, P., Shang, F., Jiao, L., Shen, Y. and Stolkin, R., 2020. Semantic segmentation for SAR image based on texture complexity analysis and key superpixels. *Remote Sensing*, 12(13), p.2141.
- Somers, M. and Whittaker, J., 2007. Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research*, 183(3), pp.1477-1487.
- Staffa, S.J., Kohane, D.S. and Zurakowski, D., 2019. Quantile regression and its applications: a primer for anesthesiologists. *Anesthesia & Analgesia*, 128(4), pp.820-830.
- Stewart, J.P., Klimis, N., Savvaidis, A., Theodoulidis, N., Zargli, E., Athanasopoulos, G., Pelekis, P., Mylonakis, G. and Margaris, B., 2014. Compilation of a local VS profile database and its application for inference of VS 30 from geologic-and terrain-based proxies. *Bulletin of the Seismological Society of America*, 104(6), pp.2827-2841.

- Sya'bani, Y.A., Novianty, A. and Prasasti, A.L., 2020, June. Implementation of automatic first arrival picking on P-Wave seismic signal using logistic regression method. In *2020 8th International Conference on Information and Communication Technology (ICoICT)* (pp. 1-5). IEEE.
- Thompson, E.M., Baise, L.G., Kayen, R.E., Morgan, E.C. and Kaklamanos, J., 2011. Multiscale site-response mapping: A case study of Parkfield, California. *Bulletin of the Seismological Society of America*, *101*(3), pp.1081-1100.
- Thompson, E.M., Baise, L.G., Kayen, R.E., Tanaka, Y. and Tanaka, H., 2010. A geostatistical approach to mapping site response spectral amplifications. *Engineering geology*, *114*(3-4), pp.330-342.
- Thompson, E.M., Wald, D.J. and Worden, C.B., 2014. A VS30 map for California with geologic and topographic constraints. *Bulletin of the Seismological Society of America*, *104*(5), pp.2313-2321.
- Tinsley, J. C., Fumal, T. E., & Ziony, J. I. (1985). Mapping Quaternary sedimentary deposits for areal variations in shaking response. *Evaluating Earthquake Hazards in the Los Angeles Region—An Earth Science Perspective*, *1360*, 101–126.
- Travis, M.R., 1975. *VIEWIT: computation of seen areas, slope, and aspect for land-use planning* (Vol. 11). Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station.
- Tufano, R., Annunziata, L., Di Clemente, E., Falgiano, G., Fusco, F. and De Vita, P., 2021. Analysis of Shear Strength Variability of Ash-Fall Pyroclastic Soils Involved in Flow-Like Landslides. *Understanding and Reducing Landslide Disaster Risk: Volume 4 Testing, Modeling and Risk Assessment 5th*, pp.329-334.
- Vallejos, J.A. and McKinnon, S.D., 2013. Logistic regression and neural network classification of seismic records. *International Journal of Rock Mechanics and Mining Sciences*, *62*, pp.86-95.

- Van Den Eeckhaut, M., Vanwalleghem, T., Poesen, J., Govers, G., Verstraeten, G. and Vandekerckhove, L., 2006. Prediction of landslide susceptibility using rare events logistic regression: a case-study in the Flemish Ardennes (Belgium). *Geomorphology*, 76(3-4), pp.392-410.
- Vilanova, S.P., Narciso, J., Carvalho, J.P., Lopes, I., Quinta-Ferreira, M., Pinto, C.C., Moura, R., Borges, J. and Nemser, E.S., 2018. Developing a geologically based VS30 site-condition model for Portugal: Methodology and assessment of the performance of proxies. *Bulletin of the Seismological Society of America*, 108(1), pp.322-337.
- Wald, D.J. and Allen, T.I., 2007. Topographic slope as a proxy for seismic site conditions and amplification. *Bulletin of the Seismological Society of America*, 97(5), pp.1379-1395.
- Wald, D.J., Lin, K.W. and Quitoriano, V., 2008. *Quantifying and qualifying USGS ShakeMap uncertainty* (p. 26). Reston, VA: US Geological Survey.
- Wald, D.J., McWhirter, L., Thompson, E. and Hering, A.S., 2011, August. A new strategy for developing Vs30 maps. In *Proceedings of the 4th IASPEI/IAEE International Symposium: Effects of Surface Geology on Seismic Motion, Santa Barbara, CA* (Vol. 1).
- Wang, C., Liang, F. and Li, J., 2021. Probabilistic Quantile Regression-Based Scour Estimation Considering Foundation Widths and Flood Conditions. *Journal of Harbin Institute of Technology (New Series)*, 28(1), pp.30-41.
- Wei, Y., Kehm, R.D., Goldberg, M. and Terry, M.B., 2019. Applications for quantile regression in epidemiology. *Current Epidemiology Reports*, 6, pp.191-199.
- Wills, C.J. and Clahan, K.B., 2006. Developing a map of geologically defined site-condition categories for California. *Bulletin of the Seismological Society of America*, 96(4A), pp.1483-1501.

- Wills, C.J., Gutierrez, C.I., Perez, F.G. and Branum, D.M., 2015. A next generation VS30 map for California based on geology and topography. *Bulletin of the Seismological Society of America*, 105(6), pp.3083-3091.
- Wills, C.J., Petersen, M., Bryant, W.A., Reichle, M., Saucedo, G.J., Tan, S., Taylor, G. and Treiman, J., 2000. A site-conditions map for California based on geology and shear-wave velocity. *Bulletin of the Seismological Society of America*, 90(6B), pp. S187-S208.
- Wills, C.J. and Silva, W., 1998. Shear-wave velocity characteristics of geologic units in California. *Earthquake Spectra*, 14(3), pp.533-556.
- Withers, M.M., 2007. Final Technical Report Mid-America ShakeMap: a Large Regional Implementation.
- Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., Sampson, C.C., Kanae, S. and Bates, P.D., 2017. A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44(11), pp.5844-5853.
- Yamazaki, F., Wakamatsu, K., Onishi, J. and Shabestari, K.T., 2000. Relationship between geomorphological land classification and site amplification ratio based on JMA strong motion records. *Soil Dynamics and Earthquake Engineering*, 19(1), pp.41-53.
- Yang, W., Zou, X., Wang, M. and Liu, P., 2023. A multinomial logistic regression model-based seismic risk assessment method for museum exhibition halls. *Journal of Building Engineering*, 69, p.106312.
- Yao, X., Liu, L., Wang, Z., Shen, Z. and Gao, H., 2021. A Vs-Based Logistic Regression Method for Liquefaction Evaluation. *Advances in Civil Engineering*, 2021(1), p.5535387.
- Ye, J., 1998. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441), pp.120-131.

- Yerlikaya-Özkurt, F., Askan, A. and Weber, G.W., 2014. An alternative approach to the ground motion prediction problem by a non-parametric adaptive regression method. *Engineering Optimization*, 46(12), pp.1651-1668.
- Yilmaz, I., 2010. Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. *Environmental Earth Sciences*, 61, pp.821-836.
- Yong, A., Hough, S.E., Abrams, M.J., Cox, H.M., Wills, C.J. and Simila, G.W., 2008. Site characterization using integrated imaging analysis methods on satellite data of the Islamabad, Pakistan, region. *Bulletin of the Seismological Society of America*, 98(6), pp.2679-2693.
- Yong, A., Hough, S.E., Iwahashi, J. and Braverman, A., 2012. A terrain-based site-conditions map of California with implications for the contiguous United States. *Bulletin of the Seismological Society of America*, 102(1), pp.114-128.
- Zhang, H., Zhou, C., Lv, G., Wu, Z., Lu, F., Wang, J., Yue, T., Luo, J., Ge, Y. and Qin, C., 2020. The connotation and inheritance of geo-information tupu. *J. Geo-Inf. Sci*, 22(4), pp.653-661.
- Zhang, J. and Wang, Y., 2021. An ensemble method to improve prediction of earthquake-induced soil liquefaction: a multi-dataset study. *Neural Computing and Applications*, 33(5), pp.1533-1546.
- Zhang, W. and Goh, A.T.C., 2016. Evaluating seismic liquefaction potential using multivariate adaptive regression splines and logistic regression.
- Zhu, J., Zhang, Y., Zhang, J., Chen, Y., Liu, Y. and Liu, H., 2023. Multi-Criteria Seismic Risk Assessment Based on Combined Weight-TOPSIS Model and CF-Logistic Regression Model—A Case Study of Songyuan City, China. *Sustainability*, 15(14), p.11216.

Zijl, G. M., Ellis, F., & Rozanov, A. (2014). Understanding the combined effect of soil properties on gully erosion using quantile regression. *South African Journal of Plant and Soil*, 31, 163–172. <https://doi.org/10.1080/02571862.2014.944228>



## APPENDICES

### A. Data transformation

The software STATA is proficient in generating linear quantile regression models (Equation A.). However, the central objective of the ongoing research is centered around crafting a nonlinear quantile regression model to effectively accomplish this goal, an essential requirement involves the application of specific transformations on the independent and dependent variables in addition to coefficients derived from the linear quantile regression model, as provided by STATA.

$$Y_i = a_0 \cdot a_1^{X_i} \quad \text{Equation A. 1}$$

$$\log Y_i = \log a_0 + X_i \log a_1 \quad \text{Equation A. 2}$$

$$Z_i = c_0 + X_i' c_1 \quad \text{Equation A. 3}$$

Therefore:

$$\log Y_i = Z_i \quad \text{Equation A. 4}$$

$$\log a_0 = c_0 \quad \text{Equation A. 5}$$

$$X_i = X_i' \quad \text{Equation A. 6}$$

$$\log a_1 = c_1 \quad \text{Equation A. 7}$$

Where:

$Y_i$ : Dependent variable of nonlinear model

$X_i$ : Independent variable of nonlinear model

$a_0, a_1$ : Coefficients of nonlinear model

$Z_i$ : Dependent variable of the linear model

$X_i'$ : Independent variable of the linear model

$c_0, c_1$ : Coefficients of the linear model

## B. Tables

### B. 1 Statistics of Geological Classes

Through the current study, all the sites were categorized into four distinct geological classes “P/PM (C<sub>4</sub>)”, “M/MT (C<sub>3</sub>)”, “T/TQ/Q1vm (C<sub>2</sub>)”, and “Qa/Q1c (C<sub>1</sub>)”. presents statistical data of all geological classes.

Table B. 1 Mean, quartile, minimum, maximum, variance, and standard deviation of V<sub>s30</sub> data classified by geological age

Statistic	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>
No. of observations	197	184	33	20
Minimum	131	273	445	570
Maximum	445	946	1510	1862
Range	314	673	1065	1292
1st Quartile	237	358	549	696
Median	293	431	630	877
3rd Quartile	351	528	928	1046
Mean	296	458	742	949
Variance (n)	5110	18180	71292	124281

## B. 2 Quantile Regression Model Coefficient

Through the current study, different nonlinear categorical quantile regression models with different numbers of terms are evaluated. Taking into account the goodness of fit criteria and model behavior, the nonlinear categorical quantile regression model characterized by five distinct terms (Equation 4.16) emerges as the most fitting and suitable option for estimating  $V_{s30}$  values. The coefficients of this model for different percentiles are presented in Table B. 2.

Table B. 2 Coefficient and goodness of fit values of Equation 4.16 for different percentiles

Percentile	$a_0$	$a_1$	$a_2$	$a_3$	$a_4$	Pseudo $R^2$	Adj. Pseudo $R^2$
0.05	190.351	1.003	1.511	2.231	3.058	0.364	0.358
0.10	198.224	1.006	1.498	2.019	2.872	0.362	0.356
0.15	210.653	1.010	1.464	2.101	3.076	0.352	0.346
0.16	211.747	1.010	1.458	2.090	3.060	0.350	0.343
0.20	220.276	1.010	1.448	2.011	2.938	0.338	0.331
0.25	230.635	1.011	1.468	1.934	2.835	0.331	0.325
0.30	243.673	1.010	1.424	1.850	2.696	0.330	0.324
0.40	265.380	1.010	1.372	1.960	2.540	0.328	0.321
0.50	283.294	1.008	1.426	1.999	2.684	0.330	0.324

## C. Figures

### C.1 Scatter Plots of Geological Classes vs Topographic Attributes

Through the current study, all the sites were systematically classified into four distinct geological classes, namely "P/PM (C<sub>4</sub>)", "M/MT (C<sub>3</sub>)", "T/TQ/Q1vm (C<sub>2</sub>)", and "Qa/Q1c (C<sub>1</sub>)". Furthermore, various topographic attributes were taken into account to assess the regression models. These attributes encompass slope, which is equivalently referred to as gradient, texture, and convexity. In this section, scatter plots are presented for each of the four geological classes, illustrating their relationships with the aforementioned topographic attributes.

#### C.1.1 Scatter Plots of Geological Class C<sub>1</sub> vs Topographic Attributes

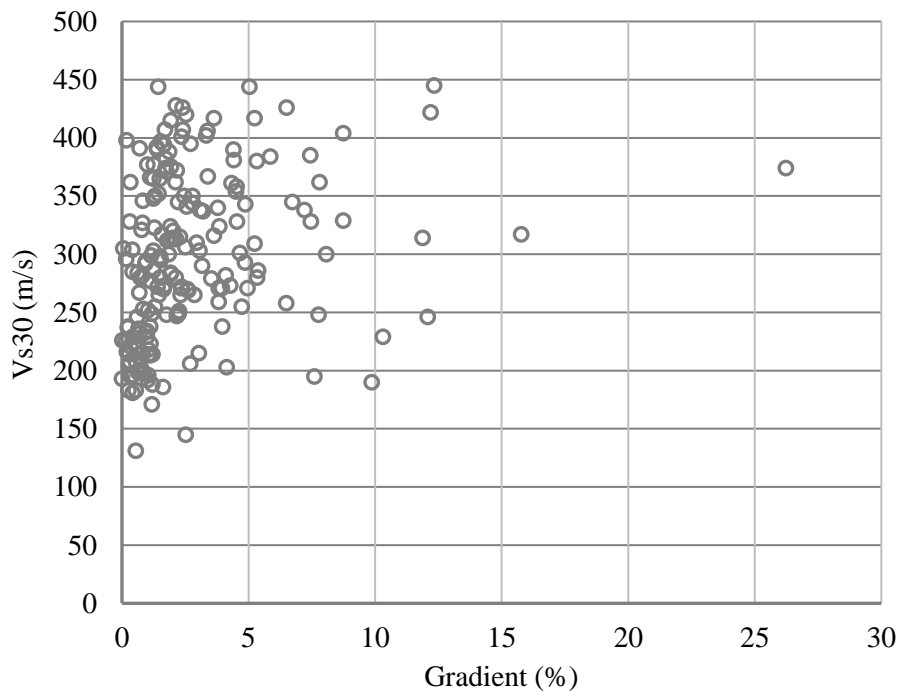


Figure C. 1 Gradient vs  $V_{s30}$  for geological class C<sub>1</sub>

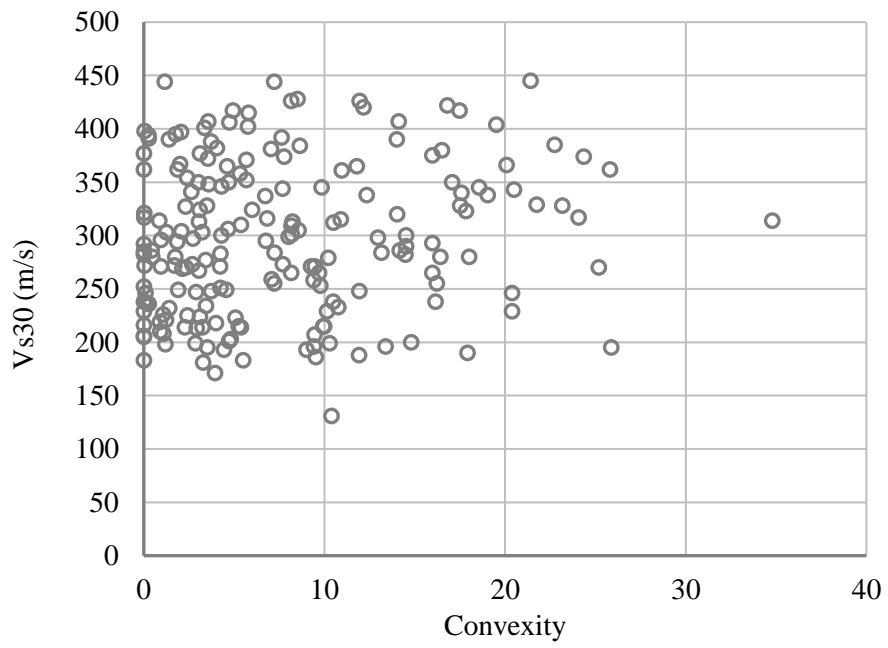


Figure C. 2 Convexity vs  $V_{s30}$  for geological class  $C_1$

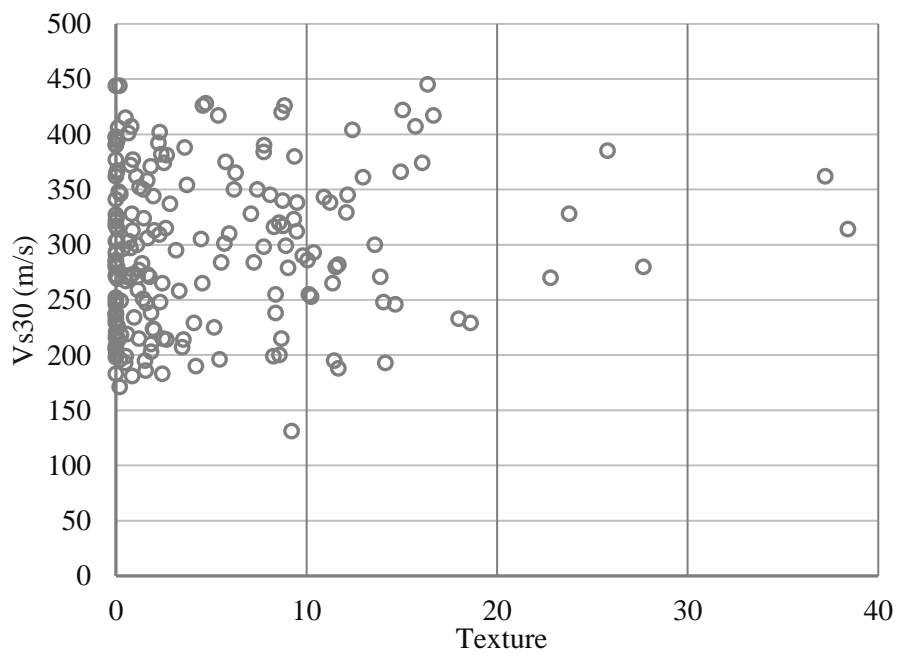


Figure C. 3 Texture vs  $V_{s30}$  for geological class  $C_1$

**C.1.2 Scatter Plots of Geological Class C<sub>2</sub> vs Topographic Attributes**

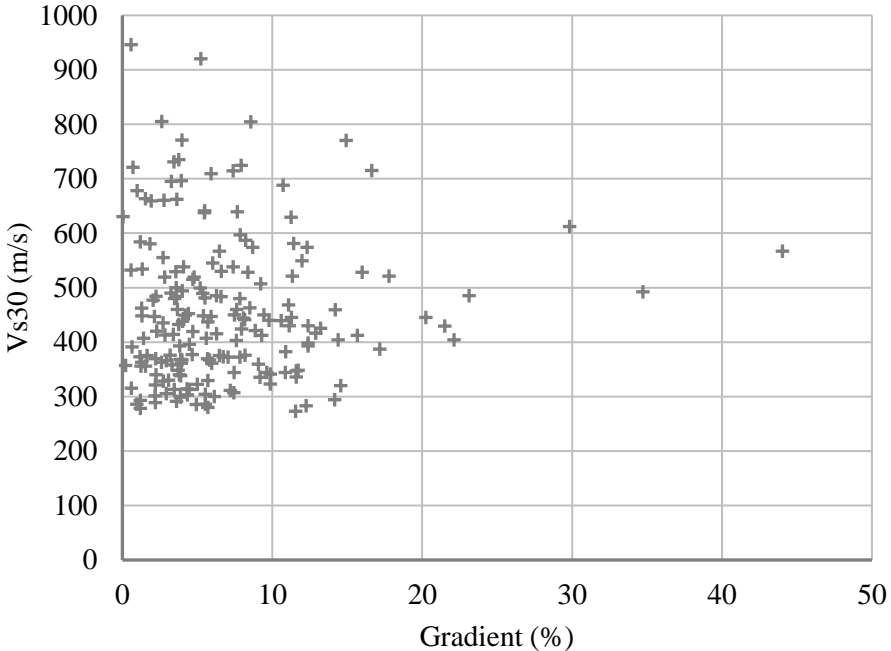


Figure C. 4 Gradient vs  $V_{s30}$  for geological class C<sub>2</sub>

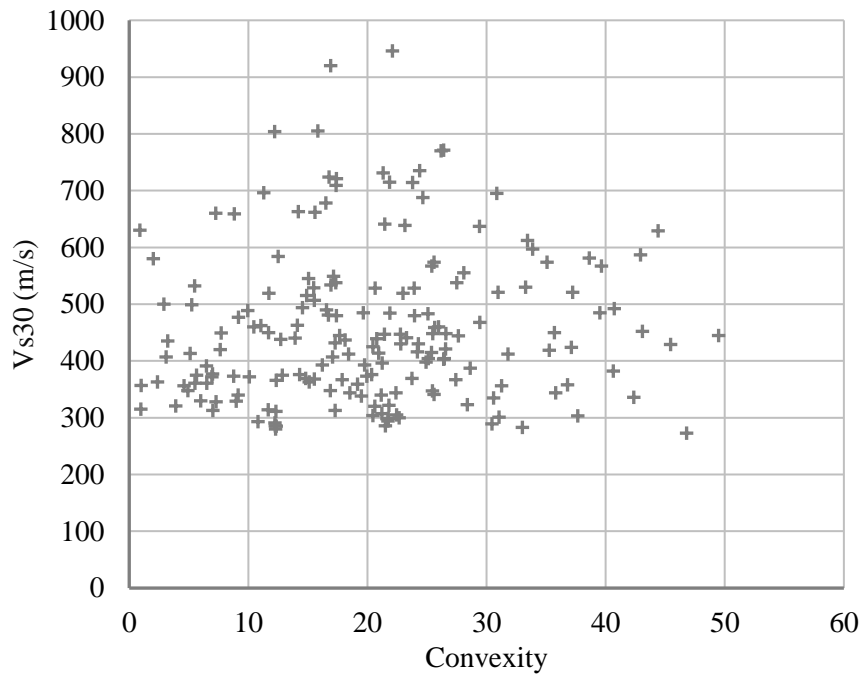


Figure C. 5 Convexity vs  $V_{s30}$  for geological class  $C_2$

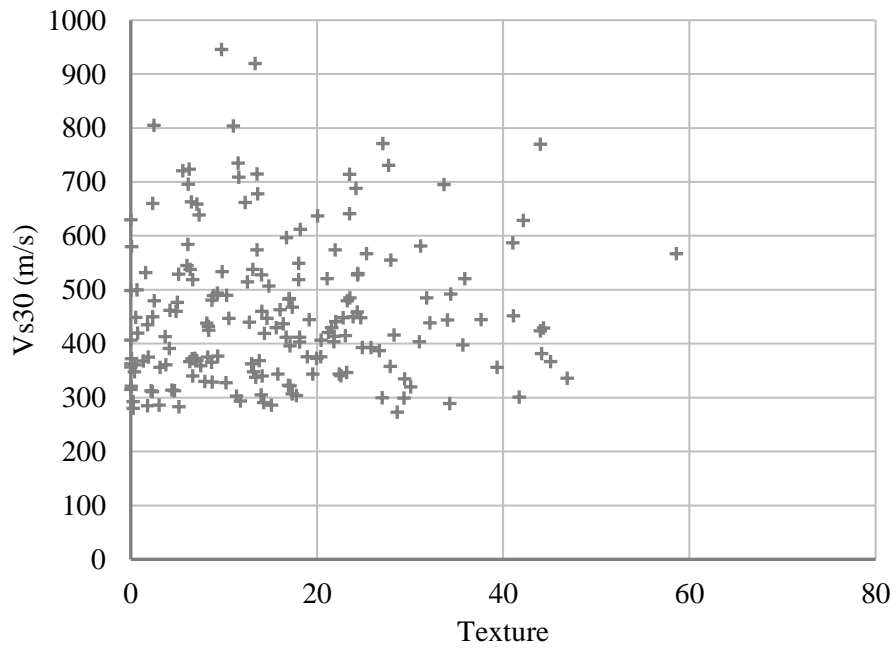


Figure C. 6 Convexity vs  $V_{s30}$  for geological class  $C_2$

C.1.3 Scatter Plots of Geological Class C3 vs Topographic Attributes

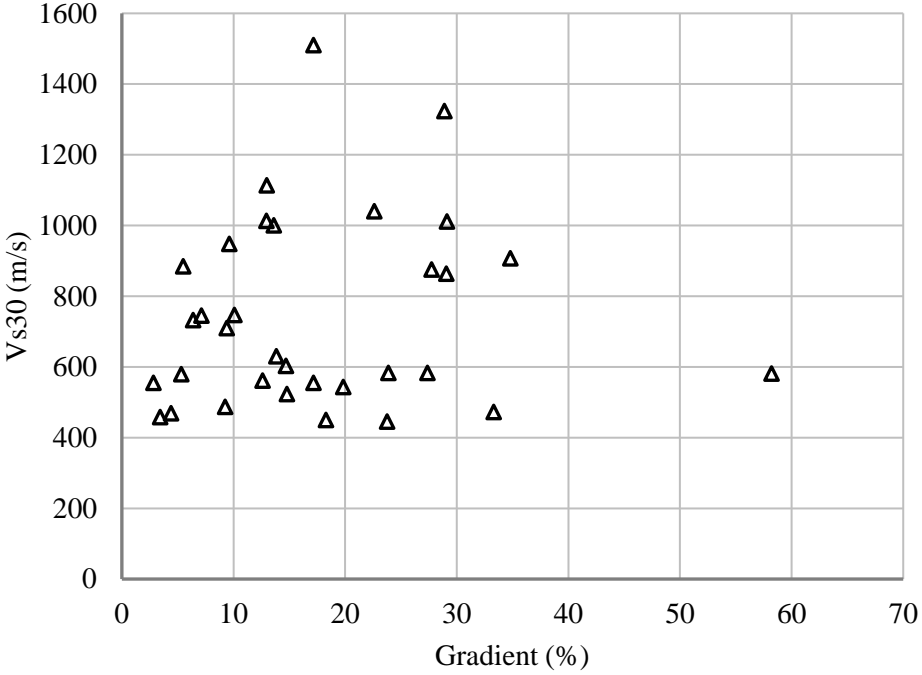


Figure C. 7 Gradient vs Vs30 for geological class C3

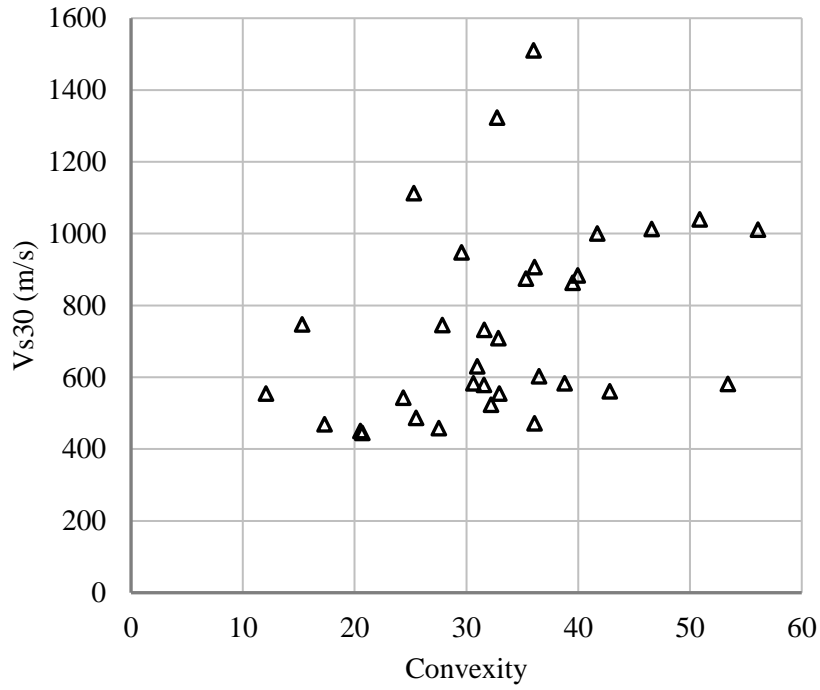


Figure C. 8 Convexity vs  $V_{s30}$  for geological class  $C_3$

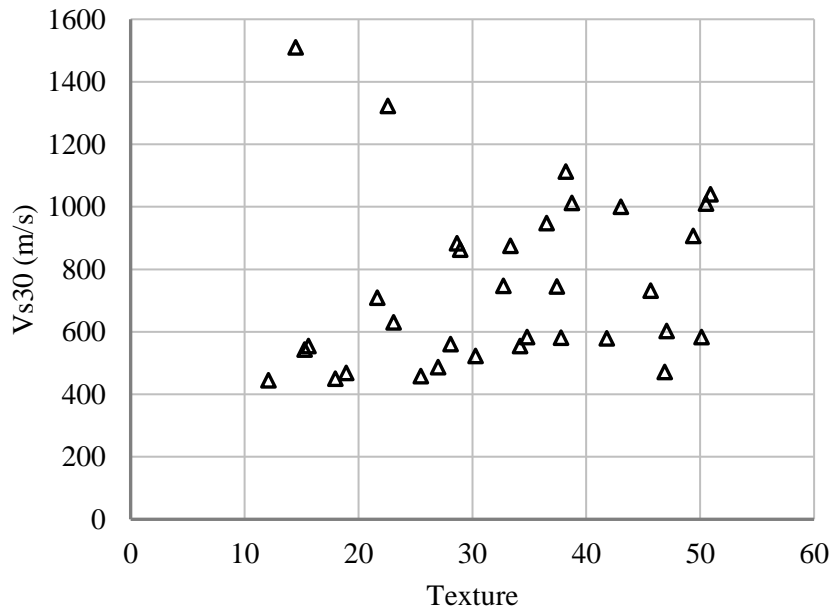


Figure C. 9 Convexity vs  $V_{s30}$  for geological class  $C_3$

**C.1.4 Scatter Plots of Geological Class C4 vs Topographic Attributes**

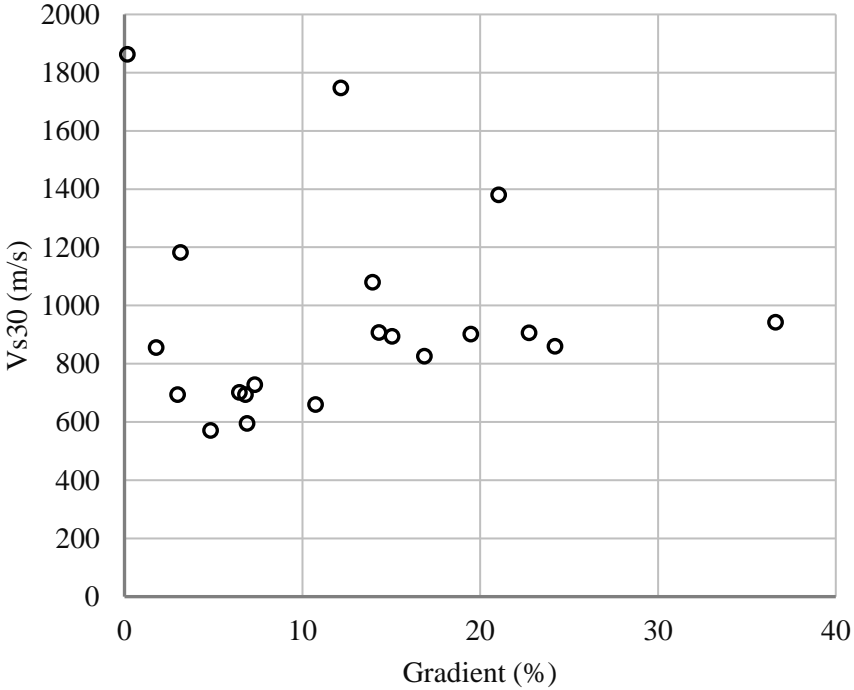


Figure C. 10 Gradient vs Vs30 for geological class C4

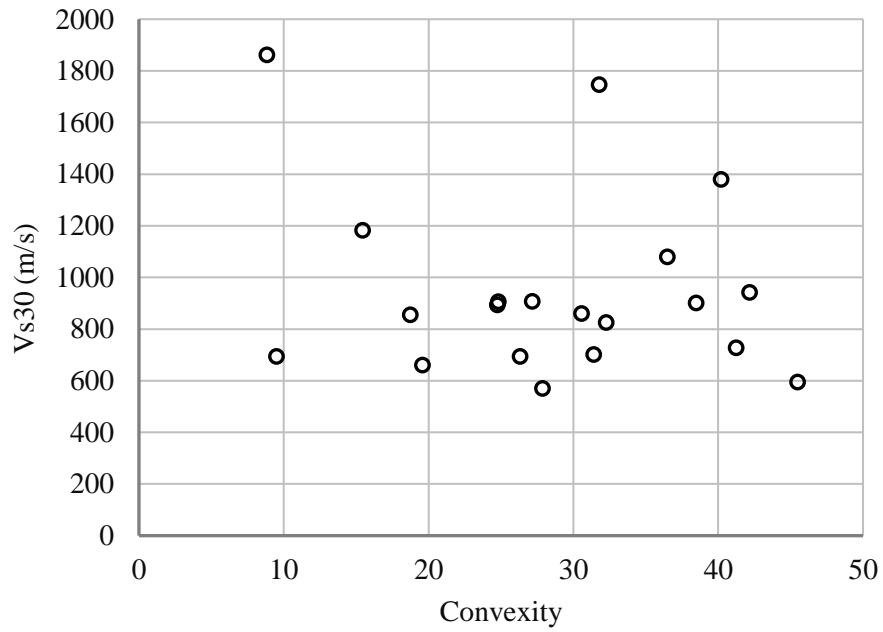


Figure C. 11 Convexity vs  $V_{s30}$  for geological class  $C_4$

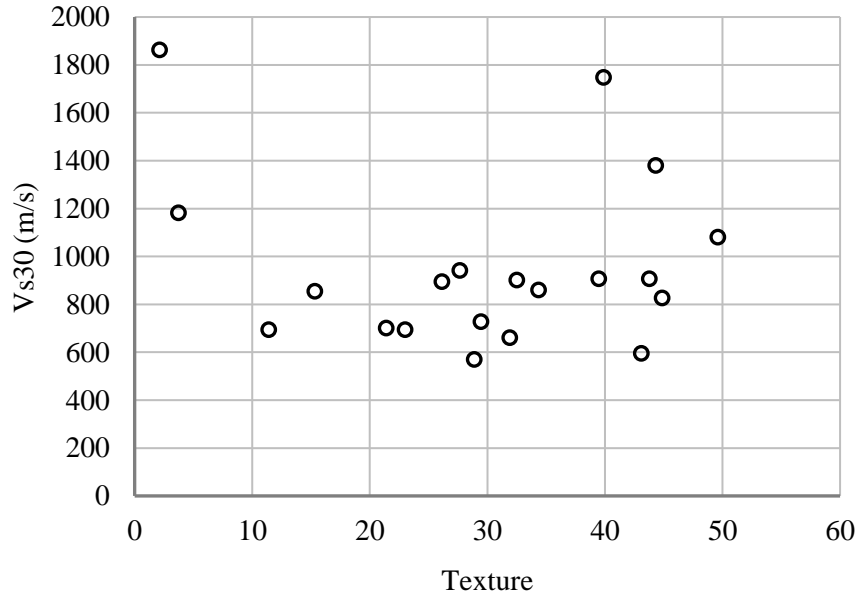


Figure C. 12 Texture vs  $V_{s30}$  for geological class  $C_4$



## CURRICULUM VITAE

Surname, Name: Jalehforouzan, Amir

### EDUCATION

Degree	Institution	Year of Graduation
MS	METU Civil Engineering	2016
BS	IAU - Mahabad Civil Engineering	2008
High School	Dehkhoda High School, Orumiyeh	2002

### FOREIGN LANGUAGES

Advanced English, Fluent Persian

### PUBLICATIONS

Sahin, Gokhan, Kivanc Okalp, Mustafa K. Kockar, Mustafa T. Yilmaz, Amir Jalehforouzan, Faik A. Temiz, Aysegul Askan, Haluk Akgun, and Murat A. Erberik. "Development of a GIS-based predicted-V s30 map of Türkiye by using geological and topographical parameters: Case study for the region affected by the 6 February 2023 Kahramanmaraş earthquakes." *Seismological Research Letters* (2024).