2012

SAMPLING BIAS IN EVALUATING THE PROBABILITY OF SEISMICALLY INDUCED SOIL LIQUEFACTION WITH SPT & CPT CASE HISTORIES

Abhishek Jain
Michigan Technological University

Follow this and additional works at: https://digitalcommons.mtu.edu/etds
Part of the Civil Engineering Commons, and the Geology Commons

SAMPLING BIAS IN EVALUATING THE PROBABILITY OF SEISMICALLY INDUCED SOIL LIQUEFACTION WITH SPT & CPT CASE HISTORIES

By

Abhishek Jain

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Geological Engineering

MICHIGAN TECHNOLOGICAL UNIVERSITY

2012

This thesis has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Geological Engineering

Department of Geological/Mining Engineering & Sciences

Thesis Advisor: *Dr. Thomas Oommen*

Committee Member: *Dr. Gregory Waite*

Committee Member: *Dr. Yue Li*

Department Chair: *Dr. Wayne Pennington*

Table of Contents

Preface

The Manuscript chapter "Epistemic uncertainty in evaluating the probability of seismically induced soil liquefaction" included in this thesis was written by me. The expertise and guidance of coauthor enhanced the quality and sharpness of this paper. All tables and figures used in this paper were made by me. In addition, the statistical modeling to develop figures and tables was coded and analyzed by me.

Acknowledgement

Firstly, I want to offer deepest gratitude and respect to my parents who supported my every decision, and are always behind me. I want to thank graduate school, department of geological and mining Eng. & Sci. and Dr. Thomas Oommen for offering me graduate research assistantship for last 4 semesters. I am grateful to Dr. Thomas Oommen for his tireless, kind, and friendly mentorship throughout this period. I want to acknowledge the kindness of committee members - Dr. Gregory Waite and Dr. Yue li to be in my thesis committee. At last, I want to thank every instructor- Dr. Thomas, Dr. Greg, Dr. Li, Dr. Vitton, and Dr. Dong to whom I have learned a lot from their classes. Without these classes, this thesis would not have been at this stage.

Apart from this, I want to thank Pearson Education Inc., Upper saddle river, NJ to give permission to reprint figures- 1.7, 1.8, 1.9, 1.12, and 1.13 in this thesis. I also want to acknowledge kindness of Dr. Boulanger, Director-Center of Geotechnical Modeling, UC Davis to grant permission to reprint figures-1.10 and 1.11 in this thesis. In addition, I am grateful to United States Geological Survey (USGS) to have an impressive archive of pictures in public domain; I selected three pictures to be republished as figures-1.1, 1.2, and 1.3 in this thesis.

Abstract

Several deterministic and probabilistic methods are used to evaluate the probability of seismically induced liquefaction of a soil. The probabilistic models usually possess some uncertainty in that model and uncertainties in the parameters used to develop that model. These model uncertainties vary from one statistical model to another. Most of the model uncertainties are epistemic, and can be addressed through appropriate knowledge of the statistical model. One such epistemic model uncertainty in evaluating liquefaction potential using a probabilistic model such as logistic regression is sampling bias. Sampling bias is the difference between the class distribution in the sample used for developing the statistical model and the true population distribution of liquefaction and non-liquefaction instances. Recent studies have shown that sampling bias can significantly affect the predicted probability using a statistical model. To address this epistemic uncertainty, a new approach was developed for evaluating the probability of seismically-induced soil liquefaction, in which a logistic regression model in combination with Hosmer-Lemeshow statistic was used. This approach was used to estimate the population (true) distribution of liquefaction to non-liquefaction instances of standard penetration test (SPT) and cone penetration test (CPT) based most updated case histories. Apart from this, other model uncertainties such as distribution of explanatory variables and significance of explanatory variables were also addressed using KS test and Wald statistic respectively. Moreover, based on estimated population distribution, logistic regression equations were proposed to calculate the probability of liquefaction for both SPT and CPT based case history. Additionally, the proposed probability curves were compared with existing probability curves based on SPT and CPT case histories.

Chapter 1 Introduction

1.1 Introduction

Earthquake induced liquefaction occurs when excess pore pressure is generated due to dynamic loading in saturated sandy soils. This phenomenon was observed after 1964 Niigata and Alaska earthquakes. Researchers have developed "simplified" procedure to assess the triggering of soil liquefaction. For the simplified procedure, two parameters Cyclic Stress Ratio (CSR) as seismic loading and Cyclic Resistance Ratio (CRR) as soil resistance are required to be estimated. The liquefaction potential is expressed in terms of Factor of Safety (FS=CRR/CSR).

The loading (CSR) can be indicated by either a full ground response analysis or a simplified approach. The soil resistance (CRR) can be calculated using either laboratory test or in-situ tests. The most widely in-situ tests used to measure soil resistance are Standard Penetration Test (SPT) and Cone Penetration Test (CPT). SPT measures number of blow counts (N-value) needed to penetrate defined depth intervals, and CPT measures cone resistance (q_c) and sleeve resistance (f_s) for a soil column. A case history corresponds to estimated value of loading parameter CSR and measured field N-value or q_c for a liquefied or a non-liquefied site. A triggering curve can be generated as shown in Figure 1.1 using field measured values N value or q_c and estimated CSR for a database of liquefied and non-liquefied sites after an earthquake. The deterministic curve or triggering curve developed using case histories cannot totally separate liquefaction and non-liquefaction instances, so the instances of liquefaction and non-liquefaction that gets misclassified on either side needs to be assigned a likelihood of failure. This problem could be addressed using probabilistic approach.

1.2 Motivation

The sampling bias is the difference between the class distribution of liquefaction and non-liquefaction instances in the sample used for developing the probabilistic model and the population distribution. There are different probabilistic model such as logistic regression, Bayesian analysis, and reliability method with Bayesian mapping have been used to evaluate probability of liquefaction. This problem exists because in post-earthquake reconnaissance, researchers often tend to collect more data from the liquefied site in comparison to non-liquefied sites, and these results in a biased sample database with more instances of liquefaction than in the population distribution. Probabilistic models developed using biased sample database instead of population distribution could give erroneous probability. The erroneous probability can impact hazard assessment and site characterization significantly.

The problem of sampling bias was first addressed by Cetin et al. [19] who developed a probabilistic model using SPT data with Bayesian updating. Cetin et al. [19] proposed a weighting ratio ($W_{non\text{-}liquefied}/W_{liquefied}$) of 1.5, or any value in the range of 1 to 3. This approach weights non-liquefaction instances more than liquefaction instances in the sample to overcome sampling bias. Cetin et al. [19] proposed this weighting ratio with expert consensus and based on the minimum variance for a given weighting ratio in their model. Subsequently, Moss et al. [18] used the same weighting ratio for CPT based case histories in Bayesian model. However, the population distribution based on weighting ratio of 1.5 for Moss et al. [18] is 64:36 (the reference for this distribution is liquefaction: non-liquefaction and it will remain same for this chapter) whereas for Cetin et al. [19], it is 45:55.

Seed [25] reviewed the 2008 Earthquake Engineering Research Institute (EERI) Monograph “Soil liquefaction during earthquakes” by I.M Idriss and R.W. Boulanger, and called the Idriss and Boulanger [32] liquefaction triggering curve highly unconservative based on the position of their curve relative to 15 % probability curve of Cetin et al. [19] (Cetin was a PhD student under Prof. Seed) curve as shown in Figure 1.1. This created havoc in geotechnical community and professionals. In Dec. 2010, Idriss and Boulanger [24] replied to the review of Seed [25] in their report “SPT-based liquefaction triggering procedures”. Idriss and Boulanger [24] reviewed the Cetin et al. [19] case histories, and evaluated 21 cases of liquefaction/non-liquefaction instances as misclassified. They also reviewed the Cetin et al. [19] curve as unnecessarily conservative, and advised geotechnical communities not to use that triggering curve. However, the Idriss and Boulanger [24] used the same weighting ratio of 1.5 recommended in Cetin et al. [19] to develop probabilistic model for population distribution of 40:60. But Idriss and Boulanger [24] acknowledge that the estimation of weighting ratio is unclear and subjective. On the other hand, Juang et al. [29] developed a probabilistic model using the same updated SPT database of Idriss and Boulanger [24]. Juang et al. [29]

used weighting ratio based on intuition and suggested no effect of weighting ratio on developed probabilistic model whereas Idriss and Boulanger [24] noted the shift in probability curve with change in weighting ratio.

Additionally, Ku et al. [28] used the population distribution of 45:55 given in Juang et al. [22] who developed their model using the same database of Cetin et al. [19]. However, using population distribution of 45:55, the weighting ratio for Ku et al. [28] was 3.75 which is outside the range of 1 to 3 suggested in Cetin et al. [19]. Apart from this, the researchers who used logistic regression model ([7]; [12]; [15]; [21]) did not account for

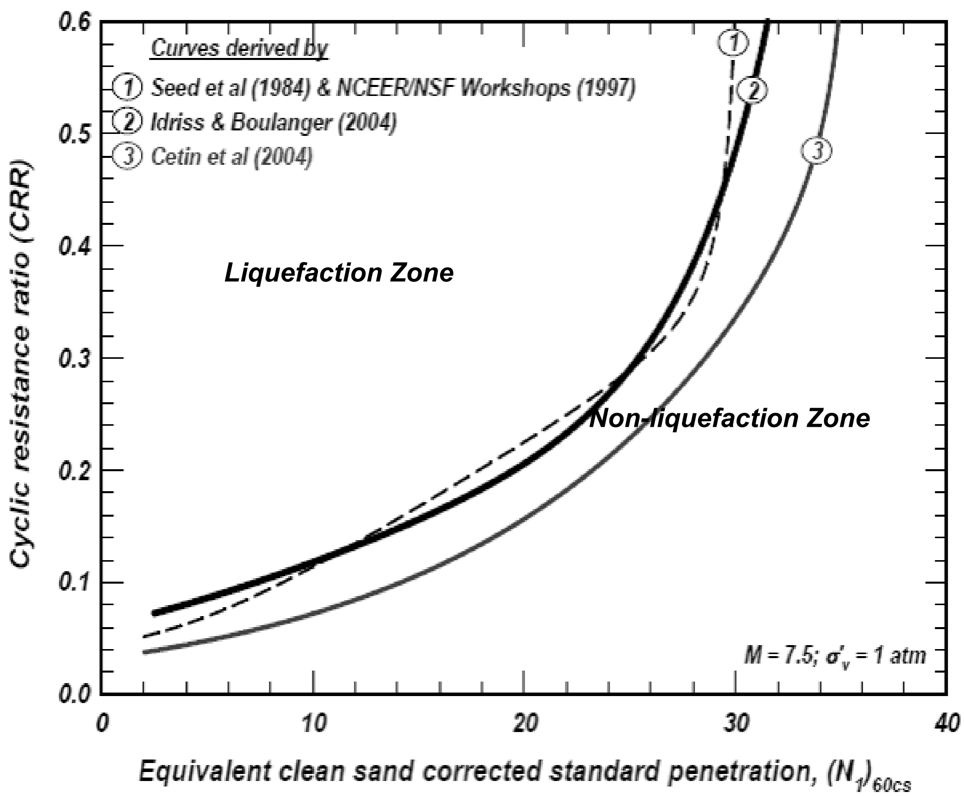

Figure 1.1: Triggering curve from different researchers Adapted from Idriss and Boulanger [24])

sampling bias in their probabilistic model. The combination of controversy, contradiction, and unclarity about the sampling bias issue and the estimation of weighting ratio provided the motivation for present work. In addition, logistic regression model for 50% probability curves for 5 different sample distributions are shown in Figure 1.2 for SPT case histories from Idriss and Boulanger [24]. It is clear from Figure 1.2 that the 5 distributions have five different probability curves, and which distribution represents the true probability is unclear. Figure 1.2 further strengthened the motivation for this work.

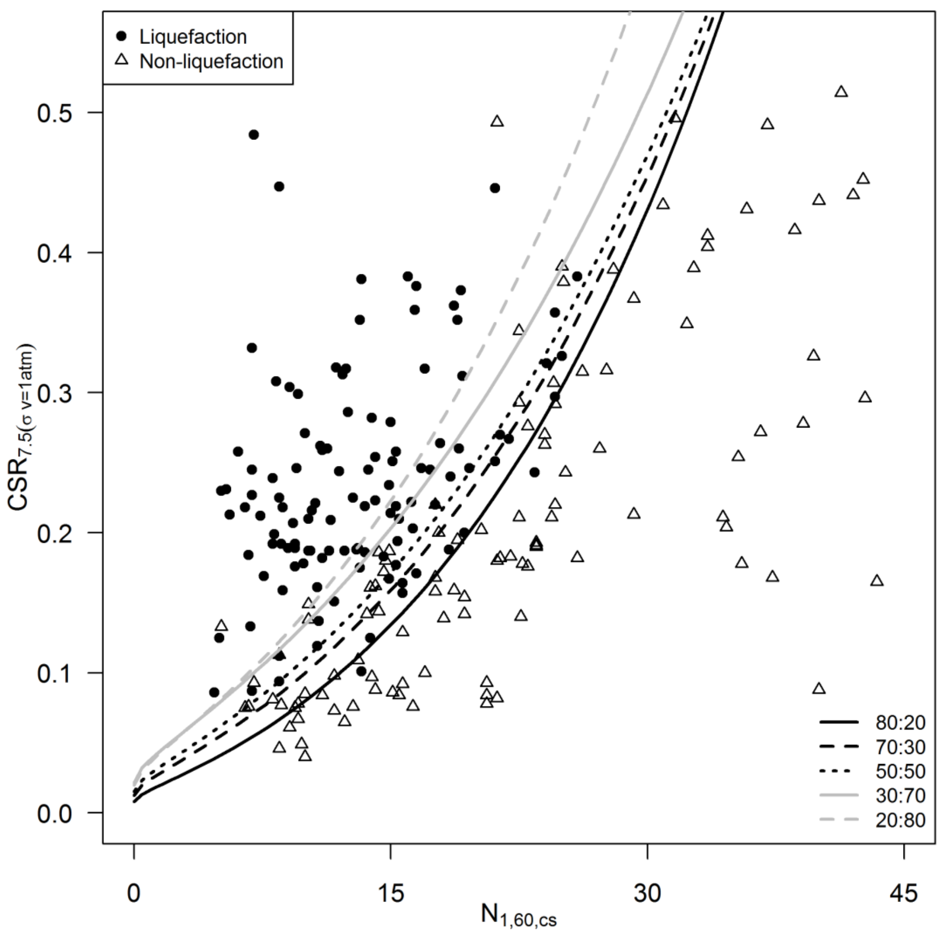

Figure 1.2: 50% probability curve for five different sample distributions

A previous work by Oommen et al. [27] used binary synthetic data to evaluate the influence of sampling bias on probabilistic modeling. Oommen et al. [27] demonstrated that the predicted probability matches the true probability only when the sample distribution is same as the population/true distribution (Figure 1.3). The conclusion of this work was based on the comparison of the predicted probability values estimated by logistic regression model with the true probability values

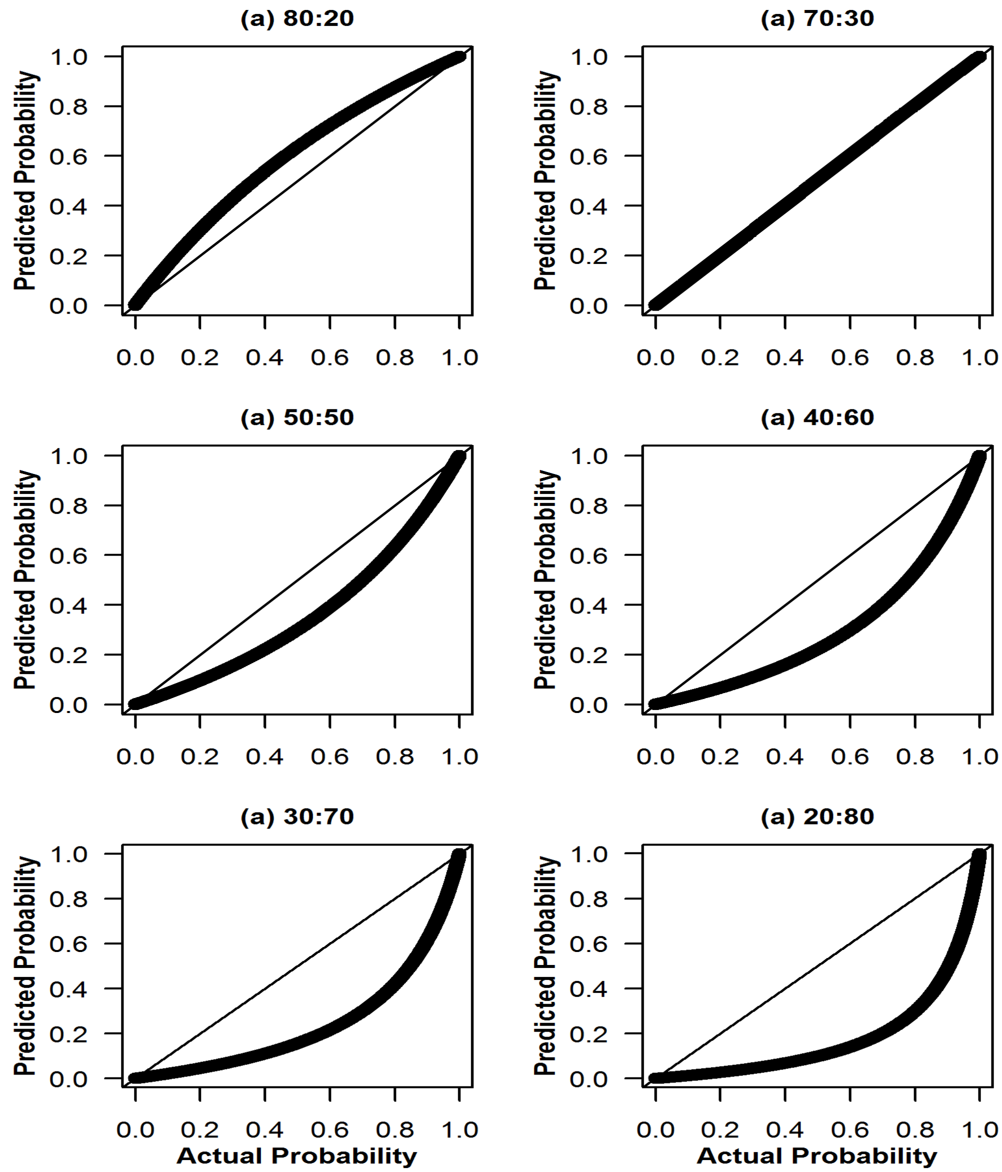

Figure 1.3: Scatter plot of predicted probabilities to actual probabilities for different sample distributions obtained from the population distribution of 70:30. (Adapted from Oommen et al. [27])

available for synthetic data. However, in reality, the true probabilities are not known. The present paper extended this information to real case histories of liquefaction/non-liquefaction and verified how the true population distributions can be determined from a sample.

1.3 Research Goals

The main goal of this study was to estimate population distribution for SPT based and CPT based case histories by addressing model uncertainty due to sampling bias. These case histories consist of biased class distribution of liquefaction and non-liquefaction cases. To address sampling bias, a logistic regression model in conjunction with Hosmer-Lemeshow statistic, was to be used. Hosmer-Lemeshow statistic is used to test the fitness of a logistic regression model when the explanatory variables are continuous

Apart from this, other model uncertainties like distribution of explanatory variables (these are the variables regressed with the response variable), and the importance of explanatory variables in model were to be addressed using KS test and Wald statistics respectively. Further, logistic regression probability curves developed based on population distribution of most updated SPT (Idriss and Boulager [24]) and of CPT (Ku et al. [28]) were to be proposed to evaluate liquefaction potential for a site. Furthermore, probabilistic curves developed using this population distribution with logistic regression model were to be compared with existing probabilistic and deterministic model for CPT based case histories ([18]; [20]; [13]; [28]) and SPT based case histories ([24]; [28]; [19]).

1.4 Background Information

1.4.1 Liquefaction

Earthquake induced liquefaction is a complex phenomenon and its evaluation has led to several controversies lately among researchers. The damage to roads, dams, foundations, and buried structures caused by the 1964 Nigata and Alaska earthquakes made researchers think about this phenomenon.

The basic definition of liquefaction is that deformation of soil caused by monotonic

, transient, and cyclic disturbance of saturated cohesionless soil under undrained conditions [11]. The generation of excess pore pressure is the key part of this phenomenon. When cohesionless soils are rapidly loaded under undrained conditions, so the densification of soils due to loading cause an increase in pore pressure, and simultaneously decrease in effective stresses in soil column. This phenomenon can be divided into two parts: 1) Flow liquefaction, 2) Cyclic mobility.

Flow liquefaction (Figure 1.4) produces the most dramatic failure of soil in its liquefied state. It occurs when the static shear stress of the soil mass is greater than the shear strength of the soil in liquefied state. The cyclic stresses (due to earthquake or due to any other induced shaking) may decrease the shear strength to a level which allows the static shear to produce failure. Here, cyclic stresses are secondary stresses responsible for flow liquefaction. Flow failure can move up to very large distance, and with very large speed.

On the other hand, cyclic mobility is produced by cyclic loading during earthquake shaking. The deformation occurred in cyclic mobility failure is incremental in contrast to the sudden origin of flow failure. The reason being the cyclic mobility failures are driven by cyclic as well as static shear stresses. These deformations are called lateral spreading (Figure 1.5), which occur on gently sloping ground or flat ground near to water bodies. Another case of cyclic mobility failure is level ground liquefaction in which the lateral deformation caused by horizontal static shear stresses does not exist. Level ground liquefaction failures are produced due to movement of water up in the soil for dissipation of excess pore pressure generations by earthquakes. The large vertical settlement, and development of sand boils (Figure 1.6) are characteristics of level ground liquefaction failure.

Liquefaction susceptibility is an important part in evaluating the liquefaction hazards because not all soils are susceptible to liquefaction. There are many criteria, such as historical, geological, compositional, and state (initial state of stress) by which liquefaction susceptibility can be judged. Youd [5] showed that the liquefaction often occurs on the same site if the soil and ground water conditions remain unchanged. Based on post-earthquake collection of case histories, the more general site conditions can be found which may be susceptible to liquefaction. Ambraseys [8] showed in his study that liquefaction effects are confined to a particular distance from the seismic source (epicenter). Ambraseys [8] plotted distances from epicenters after which liquefaction is not observed with different magnitudes as shown in Figure 1.7. This measure of distance from the epicenter can be helpful in liquefaction hazard mapping for a given earthquake scenario.

The depositional environment of soils, the age of soil deposits, and change in hydrological conditions all affect the liquefaction susceptibility.

Figure 1.4: Flow failure developed in highway fill in 1957 Daly City Earthquake in San Francisco. [33]

Geologic processes in which soil are deposited loosely and have uniform grain size distribution can be highly susceptible to liquefaction whereas the older and denser deposits are less susceptible to liquefaction [11]. For example the soils of Holocene age are more susceptible than the soil of Pleistocene age. Liquefaction only occurs in saturated soil so the groundwater depth is important for a soil to be liquefied. As the depth to the ground water table increases, the susceptibility of soil to liquefaction decreases.

Figure 1.5: Road damaged by lateral spread, near Pajaro River, 1989 Loma Prieta earthquake. [34]

Another important criterion for liquefaction susceptibility is the compositional criterion related to volume change behavior of a soil. The volume change behavior of a particular soil depends on soil particles size, shape and gradation. Fine grained soil that satisfies each of the following Chinese criteria [3] can be considered to liquefaction susceptible: 1. 15 percent or less soil should be finer than 5 micron size; 2. The liquid limit of soil should be less than or equal to 35%; 3. Moisture content should be greater than or equal to 90 percent of liquid limit; 4. Liquidity index should be less than 0.75. There have been some changes in these criteria to apply it in United States.

Even if these all preceding criteria are met, the state criteria must be met to start liquefaction. The initial stress from earthquake loading and density of soil is important to generate excess pore pressure which eventually produces liquefaction.

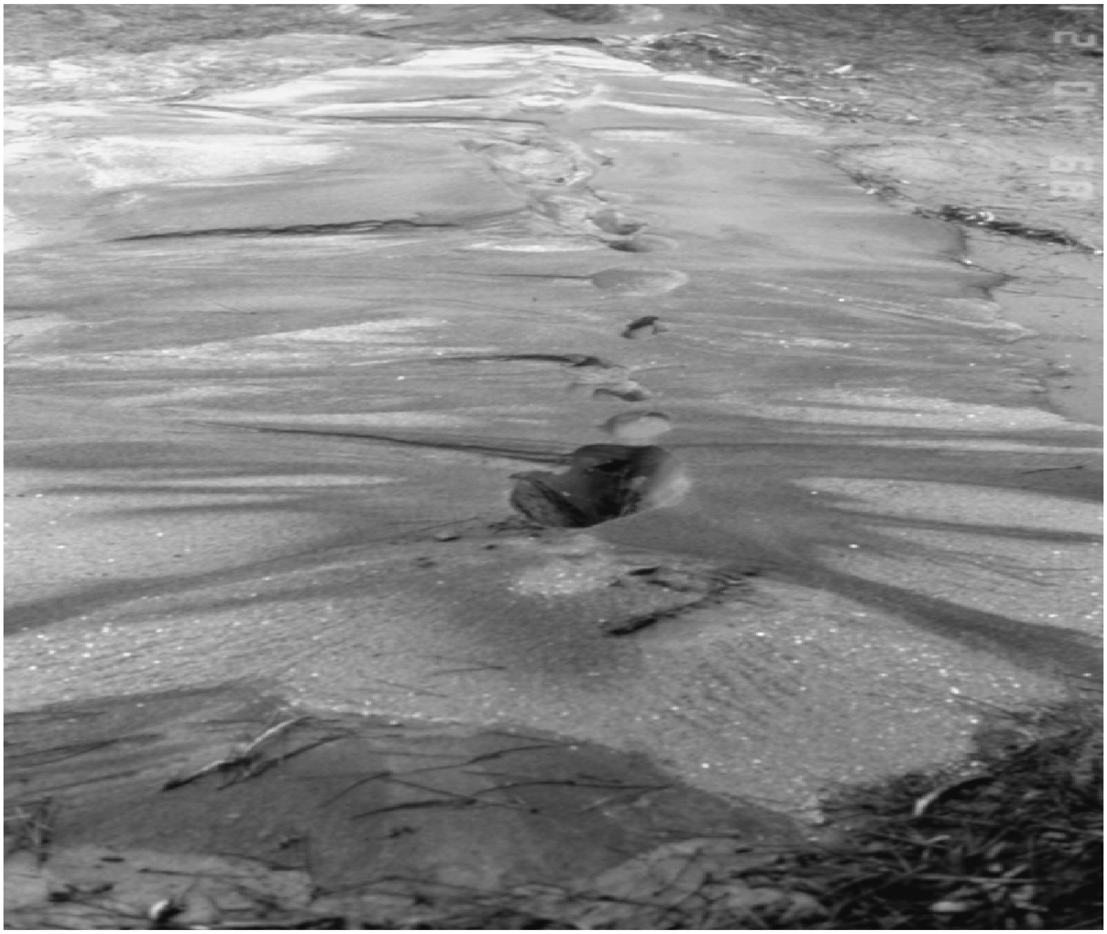

Figure 1.6: Sand boils along a fissure near the Pajaro River, 1989 Loma Prieta earthquake. [35]

1.4.2 Evaluation of Initiation of liquefaction

There are mainly two approaches developed to evaluate the initiation of liquefaction: 1) cyclic stress approach; 2) cyclic strain approach. The cyclic stress approach is widely used and most popular over the last few years. In the cyclic stress approach, the earthquake induced loading is compared with liquefaction resistance. If loading exceeds resistance then liquefaction occurs, and if loading does not exceed resistance then liquefaction does not occur.

1.4.2.1 Earthquake Loading

The excess pore pressure is the main cause of the liquefaction, and generation of excess pore pressure is related to amplitude and duration of earthquake induced loading. In cyclic stress approach, it is assumed that the excess pore pressure is related to cyclic shear stresses, and cyclic shear stress can be expressed in terms of seismic loading. The loading can be indicated in two ways: 1) full ground response analysis; 2) simplified approach.

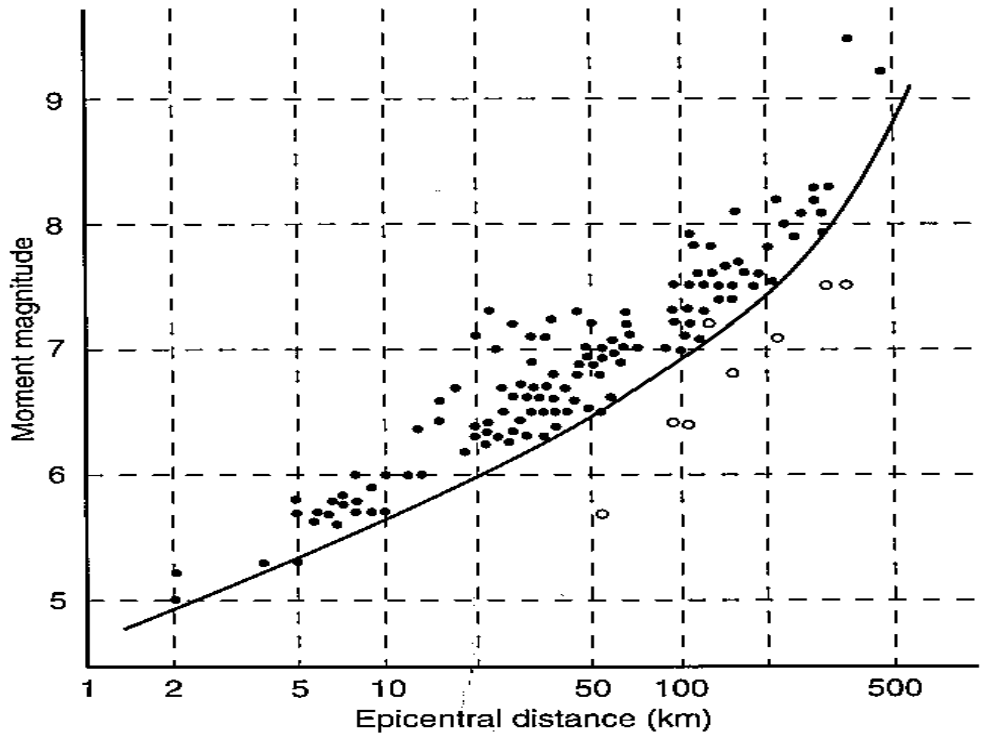

Figure 1.7: Relationship between limiting epicentral distances of sites where liquefaction has been observed and moment magnitude. (Kramer [11])

Ground response analysis is typically used to predict shear stress at various depths for a soil deposit. These analyses give time histories with the transient and irregular characteristics of actual earthquake motions as shown in Figure 1.8. In contrast to that, the liquefaction resistance of soil estimated in the laboratory is usually done with uniform cyclic shear stress time histories (uniform amplitude). So for comparison of earthquake induced loading to laboratory determined resistance, a conversion of irregular time histories of seismic loading to uniform time histories is required. Seed et al. [2] added a weighting procedure to find the number of uniform stress cycles from recorded strong ground motions. They used a 65% of peak cyclic shear stress to determine the equivalent number of cycles that would produce an equivalent pore pressure for irregular time histories ($\tau_{cyc} = 0.65\ \tau_{max}$).

Figure 1.9 shows that as the magnitude of an earthquake increases, the number of equivalent cycles also increases which is intuitive. In many cases, direct response analysis cannot be performed because of the constrain on time and budget, because of unavailability of all information for performing ground response analysis, and because of the compatibility of simplified procedure with procedure developed to estimate in-situ liquefaction resistance [19].

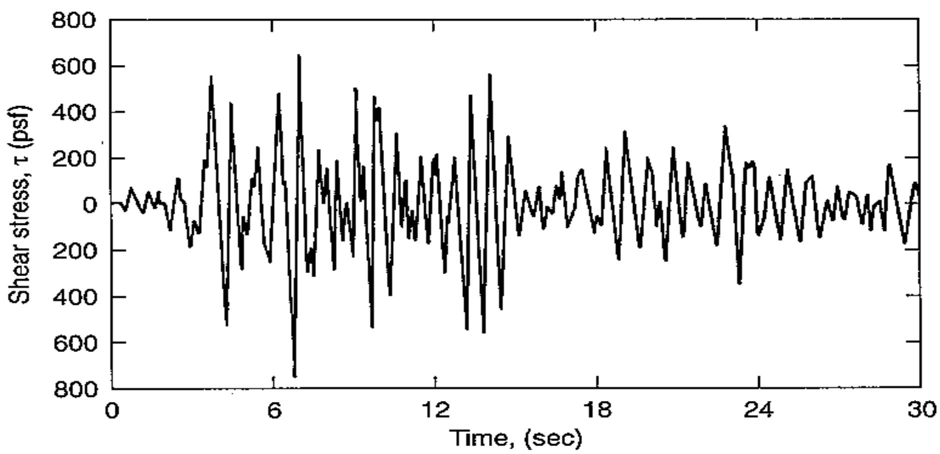

Figure 1.8 Typical irregular time history for shear stress. (Kramer [11])

The cyclic shear stresses acting on a horizontal plane are largely dominated by the cyclic shear stresses induced by the vertically or near vertically propagating shear waves. Using this as a base, the simplified procedure was developed to determine the induced cyclic shear stresses at particular depth as given in Seed & Idriss [1]. The Figure 1.10 shows an illustration of this scheme in which if a soil column above an element of soil above *h* would impose shear stress at depth *h*, then the maximum shear stresses on a horizontal plane at depth *h* would be as given in Cetin et al. [19].

$$\tau_{(max)rigidbody}= \Upsilon * h * \frac{a_{max}}{g} \qquad (1)$$

where Υ = total unit weight of the soil; *h*=height of soil column; a_{max}= maximum horizontal acceleration.

However, the soil does not respond as a rigid body; so the actual peak shear stress induced at *h* is less than the estimated by equation (1). In other words the deformable soil mass can induce less shear stress at depth *h* in comparison to shear stresses induced by a rigid body. In next step, the $\tau_{(max)rigidbody}$ is adjusted using r_d a nonlinear shear mass participation factor or stress reduction coefficient to get a practical or real $\tau_{(max)deformablesoil}$ shear stresses induced by deformable soil at depth h.

$$r_d = \frac{\tau_{(max)deformable\ soil}}{\tau_{(\max)\ rigid\ body}} \tag{2}$$

So using equation (1) and (2)

$$\tau_{(max)deformablesoil} = \Upsilon * h * \frac{a_{max}}{g} * r_d \tag{3}$$

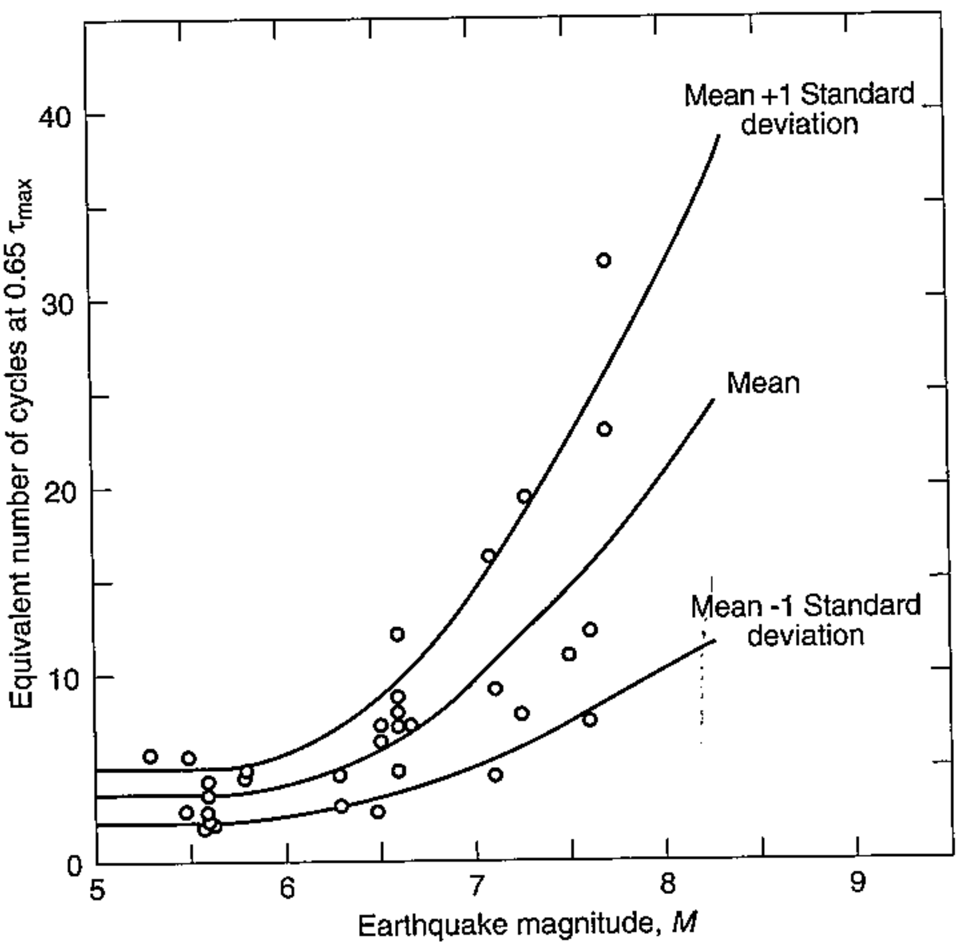

Figure 1.9: Number of equivalent uniform stress cycles for different earthquake magnitude. (Kramer [11])

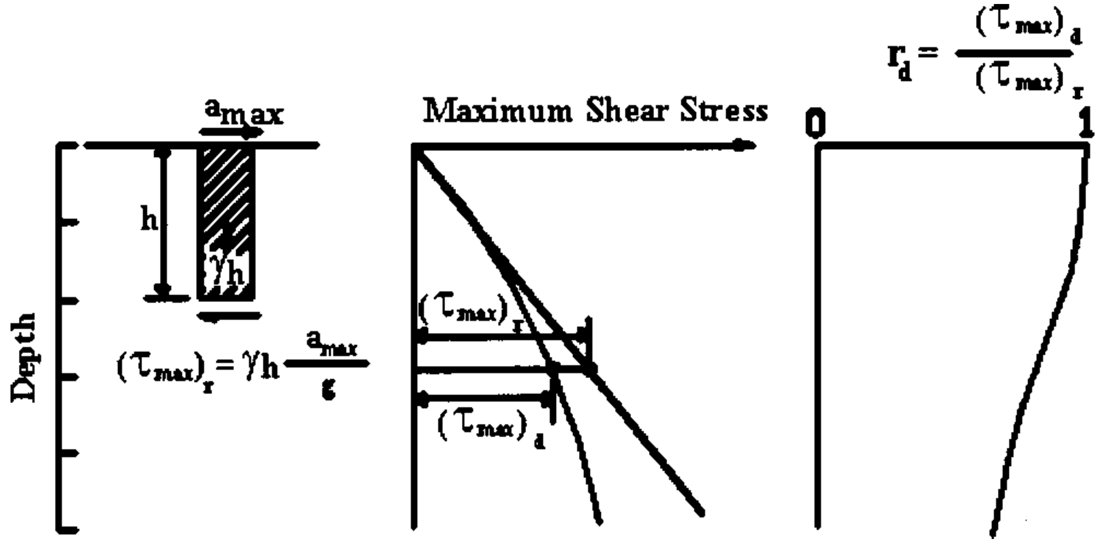

Figure 1.10: Schematic diagram of "Simplified Procedure".(After Seed and Idriss [1])

A factor of 0.65 is applied to reduce the peak cyclic stress to equivalent uniform cyclic shear stress so that it is compatible with procedure used to evaluate liquefaction resistance in laboratory.

$$\tau_{(equ.)}=0.65 * \Upsilon * h * \frac{a_{max}}{g} * r_d \tag{4}$$

when this equivalent uniform cyclic shear stress is normalized by effective overburden stresses, then it becomes the equivalent cyclic stress ratio (CSReq).

$$\text{CSReq}=0.65 * \frac{\sigma_v}{\sigma_v'} * \frac{a_{max}}{g} * r_d \tag{5}$$

Equation (5) is further normalized for different magnitude by using MSF (magnitude scaling factor), and for initial static stresses by using K_σ.

$$\text{CSR}_{7.5,1}=0.65 * \frac{\sigma_v}{\sigma_v'} * \frac{a_{max}}{g} * r_d * \frac{1}{MSF} * \frac{1}{K_\sigma} \tag{6}$$

For efficacy of this simplified procedure to evaluate CSR at any given depth, the proper estimation of r_d is required. The stress reduction coefficient (r_d) is dependent on site stratigraphy, soil properties, and characteristics of input ground motions. There have been many studies (Idriss and Boulanger [24]; Cetin et al. [19]; Kishida et al. [23]) which give different relationships to calculate r_d. (Figure 1.11)

1.4.2.2 Liquefaction resistance

Laboratory test: The cyclic stress approach initially emphasized to use laboratory testing to evaluate liquefaction resistance. Most laboratory tests were performed on an isotropically consolidated triaxial specimen or K_o consolidated simple shear specimen (Kramer [11]). In these tests, liquefaction failure was indicated as point where initial liquefaction was reached or some limiting cyclic strain amplitude was reached.

In-situ test:

Standard Penetration Test: SPT is the oldest in-situ test being used to evaluate liquefaction resistance based on some in-situ properties of soil deposits.

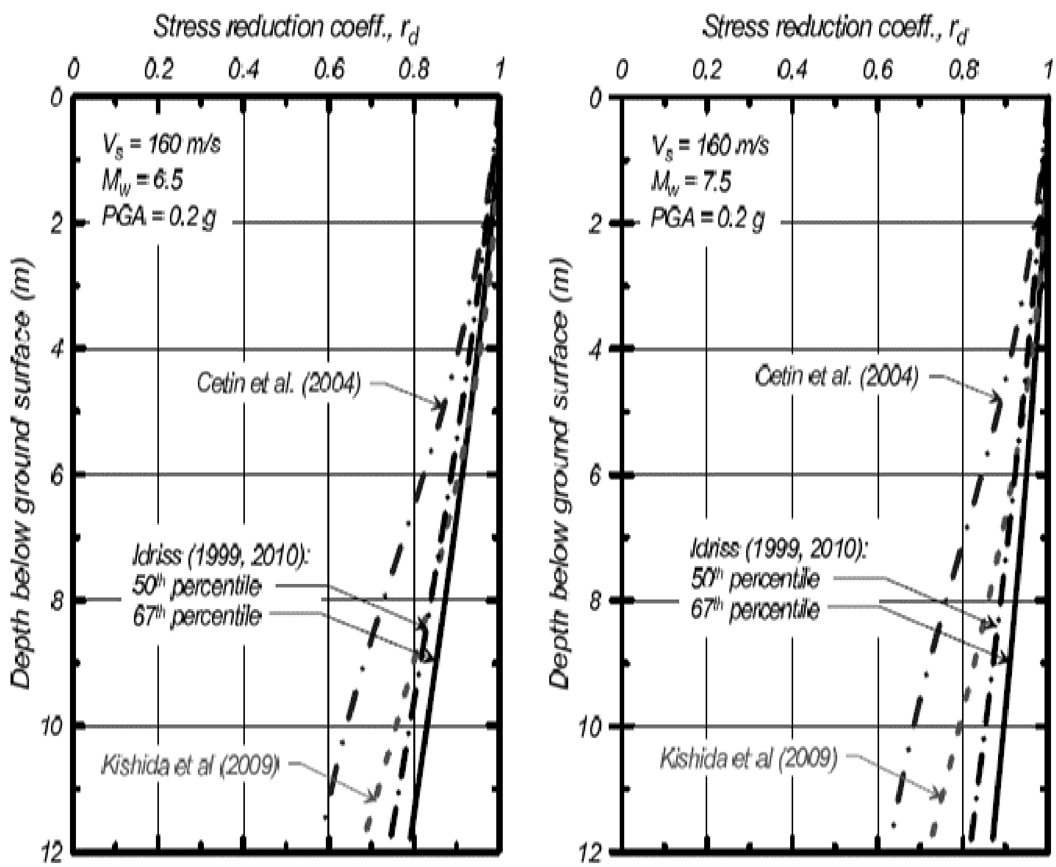

Figure 1.11: Variation of stress reduction coefficient with depth in a soil. (Idriss and Boulanger [24])

As shown in Figure 1.12, a hammer is dropped from a distance on 30 inch until it penetrated a distance of 18 inch in soil. The number of blows required to penetrate last 12 inch of soil deposit is referred as the N-value. The first six inches are not used because the bottom of exploratory boring is likely to be disturbed by the drilling process and may be covered with loose soil which may give erroneous N-value. Typically, the separation between two tests should be at least 1.5m [26].

Due to poor repeatability of SPT test, there have been some corrections proposed to N-value (Skempton [6]).

$$N_{60} = \frac{E_m C_B C_S C_R N}{0.60} \quad (7)$$

Where:

N_{60} = SPT N-value corrected for field procedure

E_m = hammer efficiency

C_B= borehole diameter correction

C_S= sampler correction

C_R= rod length correction

N = SPT N-value measured in field

The N_{60} obtained from equation (7) can be adjusted for overburden correction, this compensate the effect of test performed near the bottom of uniform soil result in higher N-value than those performed near the top.

The corrected N-value for overburden is

$$N_{1,60} = N_{60}\left(\frac{100\ KPa}{\sigma'_z}\right)^n \quad (8)$$

The exponent *n* in equation (7) is typically taken as half as proposed by Liao and Whitman [7] , however as it is dependent on the soil type so it can be regressed using the procedure given in Cetin et al. [19] or Idriss and Boulanger [24].

Where:

N_{60} = SPT N-value corrected for field procedure

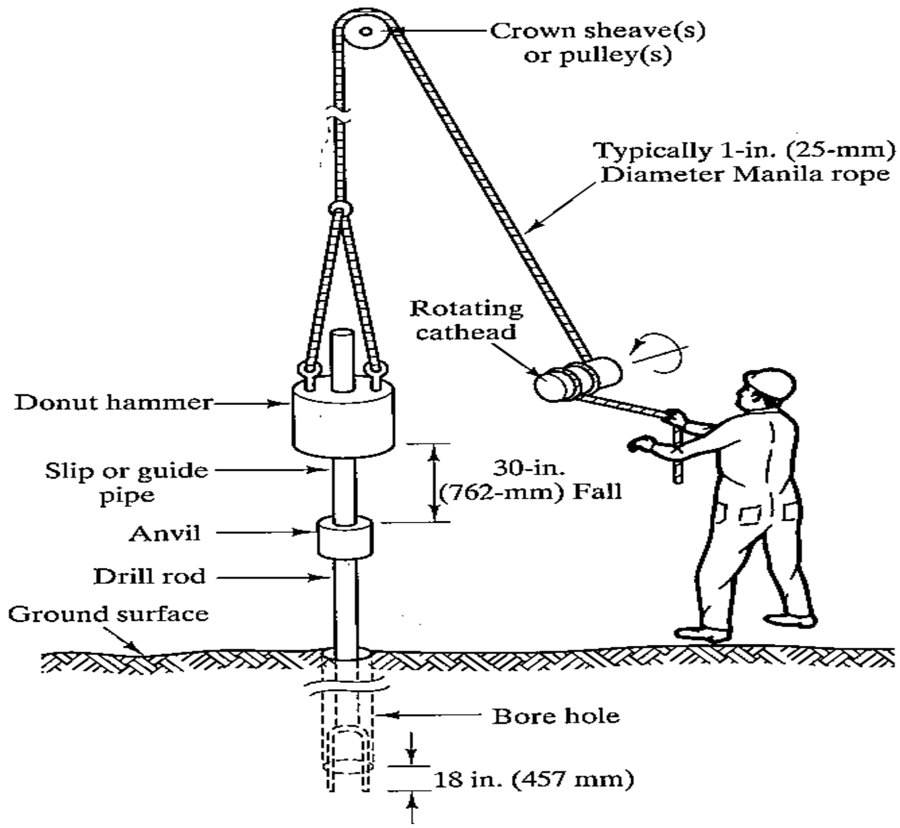

Figure 1.12: SPT equipment showing sampler in place. (Coduto et al. [26])

$N_{1,60}$ = SPT N-value corrected for field procedure and overburden stress

σ'_z = vertical effective stress at the testing depth

In recent years, researchers have corrected this $N_{1,60}$ for fines content to evaluate liquefaction potential, which is more convenient for professional and researcher to understand the liquefaction potential of a particular soil type. One of such correction proposed by Idriss and Boulanger [24] is given as

$$N_{1,60,cs}=N_{1,60}+\Delta N_{1,60} \tag{9}$$

$$\Delta N_{1,60}=\exp\left\{1.63+\frac{9.7}{FC+0.001}-\left(\frac{15.7}{FC+0.01}\right)^2\right\} \tag{10}$$

Where

FC= Fines content

$N_{1,60,cs}$ = SPT N-value corrected for field procedure, overburden stress, and fines content.

In spite of so much variability and concern of poor repeatability, SPT has some advantages over the other methods. The soil sample can be obtained, by which direct soil classification and fine contents can be determined. In case of probabilistic evaluation of liquefaction, SPT based case histories are numerous because this method has been in use for decades than any other in-situ method.

In evaluating the liquefaction potential, a critical stratum is selected based on minimum N value obtained from SPT test at a particular site. The representative N values for that site can be collected from available single SPT boring or multiple SPT boring. Subsequently, the N-values are corrected for all the corrections stated in above paragraph, and a mean $N_{1,60,cs}$ is selected for that particular site. The variability in $N_{1,60,cs}$ can be represented by the standard deviation or coefficient of variation (COV) of all measured values from single or multiple SPT borings. If there is only single N value in critical stratum available then the COV for $N_{1,60,cs}$ is considered to be 20% for that case (Cetin et al.[19]).

Cone Penetration Test:

The cone penetration test (CPT) is another common in-situ test. There are two major types of cone used for this: original mechanical cone and the electric cone. The mechanical cone is obsolete, and the electric cone is being widely used now. The electric cone has two parts; a 35.7 diameter cone shaped tip with apex angle of 60^o, and a 35.7 mm diameter & 133.7mm long cylindrical sleeve. This cone is pushed by hydraulic ram into the ground and the instrument attached to it measures the resistance to penetration. The cone resistance q_c is the total force acting on the cone divided by its projected area, and the cone side friction or sleeve friction f_{sc} is the total frictional force acting on the friction sleeve divided by its surface area (Coduto et al. [26]). Cone resistance is related to sleeve friction as

$$R_f = \frac{f_{sc}}{q_c} * 100 \quad (11)$$

where R_f = friction ratio

A typical soil profile generated from CPT test can be seen in Figure 1.13. CPT gives continuous soil profile and good repeatability in comparison to SPT; however, physically a soil sample cannot be obtained from CPT.

The thin layer correction is also applied when CPT test is performed in interbedded layers. The cone tip resistance is influenced by the soil ahead and behind the cone.

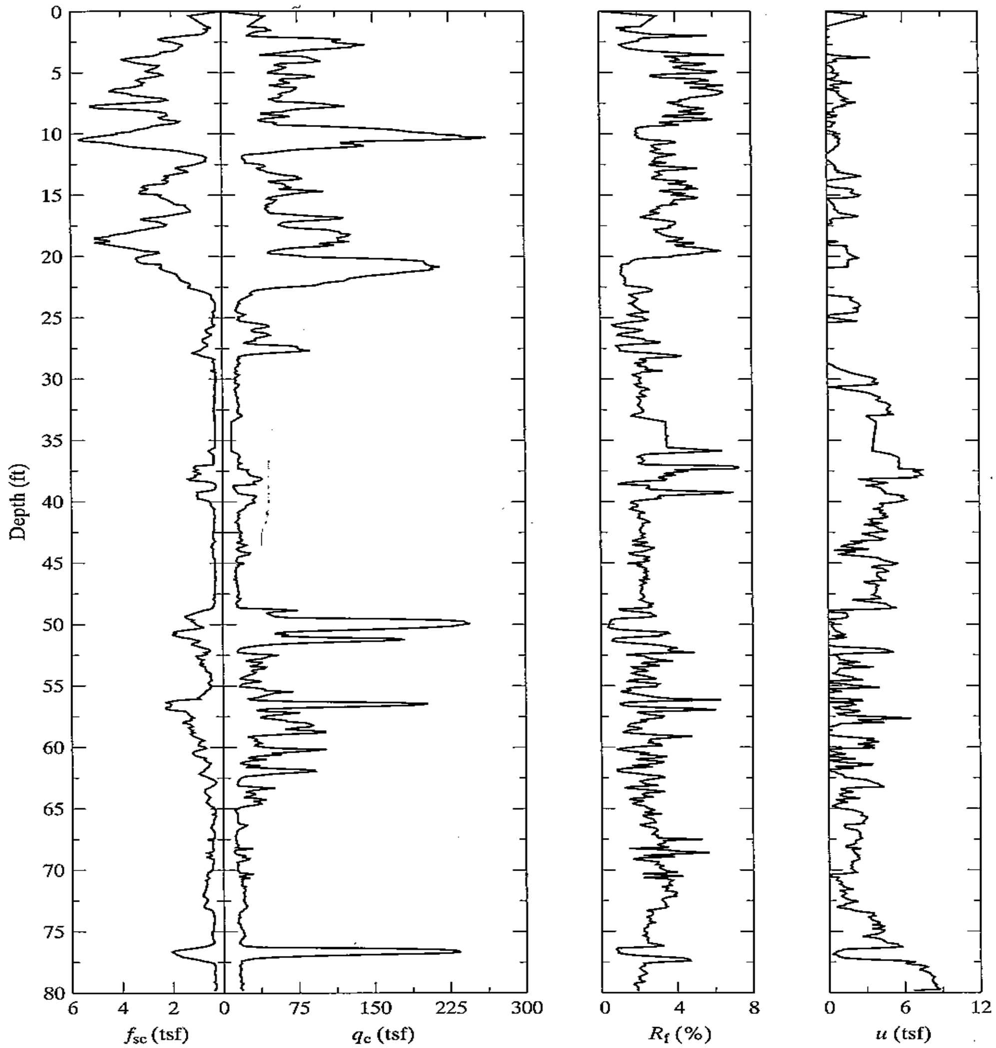

Figure 1.13: A Typical Modern Peizo CPT results with log for sleeve resistance, tip resistance, friction ratio, and pore pressure. (Coduto et al. [26])

The cone starts to sense a change in soil type before it reaches the new soil, and also cone continues to sense the original soil even it has entered in new soil (Robertson and Wride [13]). The distance over which the cone continues the interface depends on the stiffness of the soil. The soft soil has influence up to distance of 2-3 cone diameter whereas the stiff soils can have influence up to 20 cone diameter distance. This variability can affect the cone tip resistance when it penetrates in stiff thin layers situated in between soft layers. Researchers

Vreugdenhil et al. [10], Robertson and Wride [13], and Moss et al. [18] considered this variability and proposed some correction for raw tip resistance if a stiff thin layer is encountered while testing.

This raw cone resistance or corrected raw tip resistance on presence of stiff thin layer is required to be corrected for overburden stresses, and correction is same as applied for SPT.

$$q_{c1}=q_c\left(\frac{100\ KPa}{\sigma'_z}\right)^n \tag{12}$$

The exponent n in equation (11) is typically taken as half, however as it is dependent on the soil type so it can be regressed using the procedure given in Moss et al. [18] or Idriss and Boulanger [32].

where

q_c = raw cone tip resistance

q_{c1} = cone tip resistance corrected for overburden stresses

$$q_{c1n}=\frac{q_{c1}}{p_a} \tag{13}$$

where Pa = Atmospheric pressure in unit same as q_{c1}

This normalized correction factor for overburden stresses is then normalized to make it dimensionless, and subsequently it is corrected for fines content based on the procedure recommended in NCEER workshop documented in Youd et al. [31].

The normalized tip resistance for clean sand is given as

$$q_{c,1ncs}=k_c * q_{c1n} \tag{14}$$

k_c can be calculated based on procedure given in Youd et al. [31].

In evaluating the liquefaction potential, a critical stratum is selected based on lowest stretch of tip resistance with low friction ratio in a CPT log, and it can be also confirmed by nearby SPT log. The representative tip resistance and friction ratio values for that site can be collected from available single CPT boring or multiple CPT borings. Then these all raw tip resistance and friction ratio are corrected for all the corrections stated in above paragraph, and a mean $q_{c,1ncs}$ is calculated using above equation for that particular site.

1.5 Probabilistic modeling to evaluate liquefaction

The basic purpose of measuring $q_{c,1ncs}$, $N_{1,60,cs}$, and estimating $CSR_{7.5,1}$ from all around the world is to develop a relationship between these parameter using a regression analysis or a statistical model which can eventually be used for determining liquefaction potential

for a site. There have been many statistical models developed to evaluate liquefaction potential such as Bayesian mapping, reliability methods, and logistic regression method (Liao et al. [9]; Lai et al. [21]; Youd and Nobel [12]; Toprak et al. [14]; Juang et al. [15]; Moss et al. [18]; Juang et al. [20]; Cetin et al. [19]; Juang et al. [22]; Idriss and Boulanger [24]; Juang et al. [29]; Ku et al. [28]). Of these different methods, the logistic regression method is the oldest and most widely used to evaluate probability of liquefaction.

1.5.1 Logistic regression

A logistic regression model is used to develop a regression between categorical response variable with categorical or continuous explanatory (predictor) variables. Categorical variables are of two type; nominal and ordinal type. Ordinal categorical variable has order like for income (low, medium, high) whereas nominal categorical variable has no order like for primary transportation (bus, subway, and bicycle). If a categorical variable has two categories then that variable is a binary variable; for example liquefaction is a binary variable because it shows only two effects either it happens or does not happen.

Logistic regression model is the most popular method to regress binary data. Logistic regression binary response can be visualized as “Success” and “Failure”. In addition, the logistic regression models used in geosciences or in geotechniques are mostly comprised of binary response variable with continuous explanatory variables such assoil properties, temperature, etc.

The probability of liquefaction (P_L) in logistic regression framework can be given in terms of explanatory variables like $q_{c,1ncs}$, $N_{1,60,cs}$ and $CSR_{7.5,1}$ etc. which affects the occurrence of liquefaction for a site.

P_L can be defined as (Liao et al., [9])

$$log\left(\frac{P_L}{1-P_L}\right) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots\cdots\cdots\cdots\cdots \beta_n x_n \qquad (15)$$

$$logit(P_L) = log\left(\frac{P_L}{1-P_L}\right) \quad (16)$$

$$P_L = \frac{1}{1+exp[-(\beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots\cdots\cdots \beta_n x_n)]} \quad (17)$$

Where

P_L = Probability of liquefaction which varies between zero to one.

$x_1, x_2,$ ………… x_n = explanatory variables

$\beta_o, \beta_1, \beta_2$ ……… β_n = regression coefficient determined from binary regression

There are some conditions to be satisfied for a variable to quantify as an explanatory variable. Johnson and Wichern [16] described that firstly, all the explanatory variables used in logistic regression must be independent if more than one explanatory variable is used for model. Secondly, all explanatory variables must be normally distributed. Thirdly, the expected value (P_L) should be linearly dependent with explanatory variables. The third condition was modified by Menard [17] in which the *logit* (P_L) instead of P_L should be linearly dependent with explanatory variables.

1.6 References:

1. H.B. Seed, I.M. Idriss, “Simplified procedure for evaluating soil liquefaction potential”. J. Soil Mechanics and Foundation Division, ASCE, 97 (SM9) (1971), pp. 1249–1273.
2. H.B. Seed, I.M. Idriss, M. Makdisi, N. Banerjee, “Representation of irregular stress time histories by equivalent uniform stresses series in liquefaction analysis”. EERC 75-29 (1975), Earthquake engineering research centre, University of California, Berkeley.
3. W. Wang, “Some findings in soil liquefaction”. Water conservancy and hydroelectric power scientific research institute (1979), Beijing, China.
4. P.K. Robertson, R.G. Campanella, “Interpretation of cone penetration tests: Parts 1 and 2”. Canadian Geotechnical J., 20 (1983), pp. 718-745.
5. T.L. Youd, “Recurrence of liquefaction at the same site”. Preceedings of 8th world conference on earthquake engineering, 3 (1984a), pp. 231-238.
6. A.W. Skempton, “Standard penetration test procedures and the effects in overburden pressure, relative density, particle size, aging, and overconsolidation”. Geothechnique, 36 (3) (1986), pp. 425-447.
7. S.S.C. Liao, R.V. Whitman, “Overburden correction factors for SPT in sands”. J. of Geotechnical Engineering, ASCE, 112 (3) (1986), pp. 373-377.
8. N.N. Ambraseys, “Engineering seismology”. Earthquake engineering and structural dynamics, 17 (1988), pp. 1-105.
9. S.S.C. Liao, D. Veneziano, R.V. Whitman, “Regression model for evaluating liquefaction probability”. J. Geotechnical Engineering Division, ASCE, 114 (4) (1988), pp. 389-411.
10. R. Vreugdenhil, R. Devis, J. Berill, “Interpretation of cone penetration results in multilayered soils”. Int. J. for Numerical Methods in Geomechanics, 18 (1994), pp. 585-599.
11. S.L. Kramer, “Geotechincal earthquake engineering”. Prentice Hall (1996).
12. T.L. Youd, S.K. Noble, “Liquefaction criteria based on statistical and probabilistic analyses”. Proceedings of the NCEER workshop on evaluation of liquefaction resistance of soil, Technical Report NCEER-97-022(1997), pp.201-205.
13. P.K. Robertson, C.E. Wride, “Evaluating cyclic liquefaction potential using the cone penetration test”. Canadian Geotechnical J., Ottawa, 35(3) (1998), pp. 442-459.

14. S. Toprak, T.L. Holzer, M.J. Bannett, J.C. Tinslay, “CPT and SPT-based probabilistic assessment of liquefaction potential.” Proceedings of 7th U.S.–Japan Workshop on Earthquake Resistant Design of Lifeline Facilities and Countermeasures Against Liquefaction, Seattle, Technical Report MCEER-99-0019(1999).
15. C.H Juang, T. Jiang, R.D. Andrus, “Assessing probability-based methods for liquefaction evaluation”. J Geotechnical and Geoenvironmental Engineering, ASCE, 128(7) (2002), pp.580–589.
16. A.R. Jphnson, W.D. Wichern, “Applied multivariate statistical analysis”. 5th edition (2002), Prentice Hall.
17. S. Menard, “Applied logistic regression analysis”. 2nd edition (2002), Sage, London.
18. R.E.S. Moss, “ CPT-based Probabilistic Assessment of Seismic Soil liquefaction initiation”. PhD. Thesis (2003), University of California, Berkeley.
19. K.O. Cetin, R.B. Seed, A. Der Kiureghian, K. Tokimatsu, L.F. Harder Jr., R.E. Kayen, R.E.S. Moss, “ Standard penetration test-based probabilistic and deterministic assessment of seismic soil liquefaction potential”. J. Geotechnical and Geoenvironmental Engineering, ASCE, 130 (12) (2004), pp.1314–1340.
20. C.H. Juang, S.Y. Fang, E.H. Khor, “ First-order reliability method for probabilistic liquefaction triggering analysis using CPT”. J. Geotechnical and Geoenvironmental Engineering,ASCE, 132 (3) (2006), pp.337–350.
21. S.Y. Lai, W.J. Chang, P.S. Lin, “Logistic regression model for evaluating soil liquefaction probability using CPT data”. J. Geotechnical and Geoenvironmental Engineering, ASCE, 132 (6) (2006), pp.694–704.
22. C.H. Juang, S.Y. Fang, W.H. Tang, E.H. Khor, G.T.C. Kung, J. Zhang, “Evaluating model uncertainty of an SPT-based simplified method for reliability analysis for probability of liquefaction”. Soils and Foundations, 49 (1) (2009), pp.137–154.
23. T. Kishida, R.W. Boulnager, N.A. Abrahamson, M.W. Driller, T.M. Wehling, “Site effects for the Sacramento-San Joaquin Delta”. Earthquake Spectra, EERI 25 (3) (2009a), pp. 301.322.
24. I.M. Idriss, R.W. Boulanger, “SPT-based liquefaction triggering procedures”. Report No. UCD/CGM-10/02, Center for geotechnical modeling, Department of civil and environmental engineering (2010), University of California at Davis.

25. R.B. Seed, “Technical Review and commemts: 2008 EERI Monograph ‘Soil liquefaction during earhquakes ‘(by I.M. Idriss and R.W. Boulanger)”. Geotechnical Report No. UCB/GT – 2010/01, University of California at Berkeley (2010).
26. D.P. Coduto, M.R. Yeung, W.A. Kitch, “Geotechnical Engineering: Principal and Practices”, Second edition (2011), Pearson higher education Inc.
27. T. Oommen, L.G. Baise, R.M. Vogeal, “Sampling bias and class imbalance in maximum-likelihood logistic regression”. Mathematical Geosciences, 43(2011), pp.99-120.
28. C.S. Ku, C.H. Juang, C.W. Chang, J. Ching, “Probabilistic version of the Robertson and Wride method for liquefaction evaluation: development and application”. Canadian Geotechnical J., 49 (2012), pp.27–44.
29. C.H. Juang, J. Ching, Z. Luo, C.S. Ku, “New models for probability if liquefaction using standard penetration tests based on an updated database of case histories”. Engineering Geology, 133-134(2012), pp.85-93.
30. R. Boulanger, I. Idriss, “Probabilistic Standard Penetration Test–Based Liquefaction–Triggering Procedure.” J. Geotechnical and Geoenvironmental Engineering, ASCE,138 (10) (2012), pp.1185–1195.
31. T.L.Youd, I.M.Idriss, R.D. Andrus, I.Arango, G. Castro, J.T. Christian, R. Dobry, W.D.L. Finn, L.F. Harder, M.E. Hynes, K. Ishara, J.P. Koester, S.S.C. Liao, W.F. Marcuson, G.R. Martin, J.K. Mitchell, Y. Moriwaki, M.S. Power, P.K. Robertson, R.B. Seed, K.H. Stokoe, “Liquefaction resistance of soils: summary report from the 1996 NCEER and 1998NCEER/NSF workshops on evaluation of liquefaction resistance of soils”. J. Geotechnical and Geoenvironmental Engineering, ASCE, 127 (10) (2001), pp. 817–833.
32. I.M. Idriss, R.W. Boulanger, “Semi-empirical procedures for evaluating liquefaction potential during earthquakes”. Soil Dynamics and Earthquake Engineering, 26(2006), pp.115–130.
33. Photograph taken by M.G. Bonilla, Courtesy USGS.
34. Photograph taken by S.D. Ellen, Courtesy USGS.
35. Photograph taken by J.C. Tinsley, Courtesy USGS.

[1]Chapter 2 Manuscript

Epistemic Uncertainty in Evaluating the Probability of Seismically-induced Soil Liquefaction

Abhishek Jain and Thomas Oommen

Department of Geological and Mining Engineering and Sciences, Michigan Technological University, Houghton, MI, USA.

2.1 Abstract

Any statistical model usually possesses model uncertainties, and these vary from one statistical model to other. Most of the model uncertainties are epistemic, and can be addressed through appropriate knowledge of the statistical model. One such epistemic model uncertainty in evaluating liquefaction potential using a probabilistic model such as logistic regression is sampling bias. Sampling bias is the difference between the class distribution of liquefaction and non-liquefaction instances in the sample used for developing the statistical model and the true population distribution. Recent studies have shown that sampling bias can significantly affect the predicted probability using a statistical model. To address this epistemic uncertainty, a new approach was developed for evaluating the probability of seismically-induced soil liquefaction, in which logistic regression model in combination with Hosmer-Lemeshow statistic was used. This approach estimates the true distribution for liquefaction and non-liquefaction events for standard penetration test (SPT) and cone penetration test (CPT) based case histories. Moreover, based on estimated true distribution, logistic regression equations were proposed to calculate the probability of liquefaction for both SPT and CPT based analysis. Additionally, the proposed probability curves were compared with existing probability curves based on SPT and CPT case histories.

[1] The material contained in this chapter is formatted for submission to the Journal of Soil Dynamics and Earthquake Engineering.

2.2 Introduction

After the 1964 Nigata and Alaskan Earthquake, researchers started to analyze the loss of strength due to seismic loading in saturated sandy soils due to the build-up of pore water pressure commonly referred as liquefaction (Kramer [7]). Seed and Idriss [1] pioneered this analyses and developed the SPT based simplified procedure for evaluating liquefaction potential. For the simplified procedure, two parameters Cyclic Stress Ratio (CSR) as seismic loading and Cyclic Resistance Ratio (CRR) as soil resistance are required to be estimated. The standard penetration test (SPT) based simplified procedure given by Seed et al. [2], and cone penetration test (CPT) based method given by Robertson and Wride [9] to evaluate liquefaction potential were updated and incorporated respectively in 1998 National Center for Earthquake Engineering Research (NCEER) workshop which was documented by Youd et al. [12]. These methods are generally referred to as deterministic methods, where the liquefaction potential is expressed in terms of Factor of Safety (FS=CRR/CSR).

In the past few years, many probabilistic methods were developed for SPT ([4];[8];[10]; [14]; [17]; [24]; [26]; [28]; [31]), and for CPT based ([10]; [14]; [16]; [21]; [20]; [28]; [30]) case histories to evaluate soil liquefaction. These probabilistic methods are more meaningful because the curves from deterministic methods cannot separate liquefaction and non-liquefaction instances completely. Therefore, the incidences of liquefaction and non-liquefaction that get misclassified on either side needs to be assigned a likelihood of failure and this can be achieved by a probabilistic approach. Besides, the probability of liquefaction is more appropriate to define regional liquefaction hazard or in performance based earthquake engineering where at component level, estimation of unbiased probability is required (Ku et al. [30]) In addition, the variability in in-situ properties like number of blow counts (N-value), cone tip resistance (q_c), effective stress (σ') etc. or parameter calculated using in-situ properties like CSR and CRR can be accounted in a probabilistic model. Therefore, defining and addressing parameter and model uncertainty are important steps in building any probabilistic model.

2.2.1 Parameter Uncertainty

Parameter uncertainty in evaluating liquefaction probability can be divided into two parts. The first one is the uncertainty in input variables for a statistical model like variability in maximum horizontal acceleration (A_{max}), strength reduction factor (r_d), total stress (σ), Effective stress (σ'), water table level, depth of critical layer, normalized blow count for overburden stress and fines content ($N_{1,60,cs}$), and normalized tip resistance for overburden stress and fines content ($q_{c,1ncs}$). For the SPT case histories based probabilistic method, Cetin et al. [17], and Idriss and Boulanger [26] addressed the uncertainties in $N_{1,60,cs}$ and $CSR_{7.5,1}$ using the coefficient of variation or the standard deviation.

Juang et al. [24] used coefficient of variation of all input variables like for Amax, σ, σ', $N_{1,60,cs,}$ fines content (FC), Moment magnitude (M_w) in their probabilistic model to address the uncertainty in the parameters. For CPT case histories based probabilistic method, Moss et al. [16] used first order Taylor series to estimate the coefficient of variation for $CSR_{7.5.}$

The second type of parameter uncertainty exists due to quality and consistency of data used to develop a statistical model, because the predictive accuracy of any statistical model depends on the quality of data (Moss et al. [16]). In the past, due to scarcity of liquefaction and non-liquefaction instances, researchers did not realize that the data collected from different SPT and CPT equipment, and from different depositional environment (Holocene, Pleistocene etc.) could potentially add so much uncertainty in a statistical model. This problem was realized by Toprak et al. [10], and consequently they used liquefaction-nonliquefaction data collected by same equipment after 1989 Loma Preita Earthquake in logistic regression model. Furthermore, Cetin et al. [17] and Moss et al. [16] performed rigorous filtering criteria and classified the available SPT and CPT case histories into different classes. Based on criteria like coefficient of variation of CSR, thin layer correction, and the type of CPT equipment used Moss et al. [16] divided data in four classes named as A, B, C, and D. Also Cetin et al. [17] divided SPT case histories in five classes (A, B, C, D, and E) based on coefficient of variation in CSR and availability of information about variability in equipment and procedure.

Even though these case histories for SPT and CPT were analyzed very meticulously, there were subsequent retesting and reevaluation of liquefaction instances. Idriss and Boulanger [26] have reevaluated the Cetin et al. [17] SPT database, and removed 21 instances and added 70 new instances. Similarly, Moss et al. [18] retested the Imperial valley earthquake site, and also retested 1976 Tangchan earthquake site (Moss et al. [29]) to find some high quality liquefaction/non-liquefaction data which can be used for probabilistic analysis. In addition, Robertson [25] recommended the removal of class C data in Moss et al. [16] database with the argument that the class C data were collected using mechanical cone or non-standard cone, and had no friction sleeve data which made those instances less reliable. Based on the availability of new liquefaction/non-liquefaction instances and more research in this area, there can now be a considerable amount of reliability in SPT and CPT based case histories. Idriss and Boulanger [26] and Ku et al. [30] documented the latest updated data catalogue for SPT based and CPT based liquefaction and non-liquefaction instances respectively, which is used in this study.

2.2.2 Model Uncertainty

The model uncertainty is another important part which can affect evaluation of the probability of liquefaction. Model uncertainty basically depends on the specific model (e.g. Bayesian updating, logistic regression etc.) used to measure liquefaction potential. Cetin et al. [17], Moss et al. [16], and Idriss and Boulanger [26] used an error term in their Bayesian updating model to account for model uncertainty. The logistic regression model has been widely used to evaluate probability of liquefaction (Liao et al.. [4]; Youd and Noble [8]; Toprak et al. [10]; Juang et al. [14]; Lai et al. [21]; Juang et al. [31]; Ku et al. [30]). Most of the researchers who used logistic regression for determining the probability of liquefaction used modified likelihood ratio index (MLRI) to show the fitness of their model or in other words to quantify the predictive capability of their model.

All previous authors did not mention the uncertainty related to the distribution of explanatory variables in a logistic regression. Lai et al. [21] showed that the explanatory variables must be normally distributed, and they also used most appropriate explanatory variables log ($CSR_{7.5}$) and $\sqrt{q_{c1n}}$ in their model instead of $CSR_{7.5}$ and q_{c1n}. Furthermore, the importance of significance of model parameters which are coefficient of explanatory variables in logistic regression (Liao et al. [4]; Youd and Nobel [8]; Toprak et al. [10]; Juang et al. [14]) were not discussed.

2.2.2.1 Sampling Bias

Sampling bias is a model uncertainty due to the difference between the class distribution of liquefaction and non-liquefaction instances in the sample used for developing the statistical model and the true population distribution. In post-earthquake reconnaissance, researchers often tend to collect more data from the liquefied site in comparison to non-liquefied sites, this result in a biased sample database with more instances of liquefaction than the true population distribution. To address the problem of sampling bias Cetin et al. [13] proposed using a weighting ratio which weighs the non-liquefied instances more than the liquefied instances to represent the population distribution or the actual field occurrence. The weighting ratio ($W_{non\text{-}liquefied}/W_{liquefied}$) suggested by Cetin et al. [13] based on experts' advice and minimum variance achieved in the Bayesian modeling is 1.5 or any value between 1 and 3. Later, Cetin et al. [17], Juang et al. [24], and Idriss and Boulanger [26] used this weighting ratio of 1.5 recommended by Cetin et al. [13]. This resulted in population distributions of 45:55, 45:55, and 40:60 (the reference for the given distribution is liquefaction: non-liquefaction, and it will remain same for the rest of the text.) for the work by Cetin et al. [17], Juang et al. [24], and Idriss and Boulanger [26] respectively. Both Idriss and Boulanger [26] and Juang et al. [31] mentioned the influence of weighting factor on the resultant probability curve using the same SPT case histories. Interestingly, Idriss& Boulanger [26] showed that their probability curve changed based on the weighting factor that they used, whereas, Juang et al. [31] found no effect on their probability curve from changing the weighting factor.

For CPT based model to evaluate the probability of liquefaction, Moss et al. [16] used the same weighting ratio recommended for SPT based case histories by Cetin et al. [13], and the corresponding population distribution was 64:36. Ku et al. [30] used a population distribution of 45:55 (calculated weighting ratio of 3.73) recommended by Juang et al. [24]. However, assigning the same weighting factor used by Cetin et al. [17], which best fits to their database might not be appropriate or competent with other database. In addition, the use of weighting factor developed using SPT based case histories for CPT based case histories might be a cause of significant uncertainty in a statistical model, if the CPT and SPT case histories are not same. For example, Idriss and Boulanger [26] used the same weighting factor as Cetin et al. [17] even after removing 21 instances due to misclassification of liquefaction or non-liquefaction instances in the database. Idriss and Boulanger [32] acknowledge that the estimation of weighting ratiois unclear and subjective. Ku et al. [30] used a weighting ratio of 3.73 which resulted in a population distribution of 45:55. This weighting factor is outside the recommended range of 1 to 3 by Cetin et al. [17]. It is evident from these studies that the choice of weighting ratio and the determination of the population distribution for probabilistic modeling of liquefaction lacks objective guidelines. Oommen et al. [28] demonstrated using synthetic binary data that only when the sample distribution is same as the population/true distribution does the predicted probability match the true probability. The conclusion of that work was based on the comparison of the predicted probability values estimated by logistic regression model with the true probability values available for synthetic data. However, in reality, the true probabilities are not known. This work has been the motivation for this paper to extend this information to real case histories of liquefaction/non-liquefaction and verify how the true population distributions can be determined from a sample.

2.3 Research Objectives

The main objective of this study was to estimate population distribution for SPT based and CPT based case histories by addressing model uncertainty due to sampling bias. These case histories consist of biased class distribution of liquefaction and non-liquefaction cases. To address sampling bias, a logistic regression model in combination with Hosmer-Lemeshow statistic was to be used. Hosmer-Lemeshow statistic is used to test the fitness of a logistic regression model when the explanatory variables are continuous.

Apart from this, other model uncertainties like distribution of explanatory variables (these are the variables regressed with the response variable), and the importance of explanatory variables in model were to be addressed using KS test and Wald statistic respectively. Further, logistic regression probability curves developed based on population distribution of most updated SPT (Idriss and Boulager [26]) and of CPT (Ku et al. [30])were to be proposed to evaluate liquefaction potential for a site. Furthermore, probabilistic curves developed using this population distribution with logistic regression model were to be compared with existing probabilistic and deterministic model for CPT based case histories ([16]; [20]; [9]; [30]) and SPT based case histories ([26]; [31]; [17]).

2.4 Method

To achieve the research objectives, a hypothesis was given and subsequently verified using binary data. The hypothesis was that when the sample distribution is similar to population distribution, the Hosmer-Lemeshow statistic for logistic regression model for that sample distribution gives highest P-value (where the P value represents the probability obtained from a Chi-square distribution for the corresponding Hosmer-Lemeshow statistic, with P values < 0.05 call for the rejection of the hypothesis), and thus that sample distribution will give the true probability values. This hypothesis was proved using binary synthetic data (discussed in detail later).

2.4.1 Hosmer-Lemeshow Statistic

The Hosmer-Lemeshow statistic is used to test the fitness of a logistic regression model when the explanatory variables are continuous. This statistic was developed by Hosmer-Lemeshow [11]. Agresti [23] explained that when instances are ungrouped or highly sparse, the pearson chi-square statistic (X^2) and likelihood ratio chi square statistic (G^2) does not have approximate chisquare distribution, whereas the Hosmer-lemeshow statistic does follow the chi-square distribution even in these situations.

In Hosmer-Lemeshow statistic, the fitted probabilities of total number of n binary instances (0 or 1), estimated by logistic regression model is divided into 10 groups of equal size. The first group consists of n/10 size with highest estimated probabilities. The next group refers to n/10 size having the second decile of estimated probabilities, and so forth (Agresti [23]). Each group has some instances of zeros and some instances of ones and their respective estimated probabilities. The fitted or estimated probability for an outcome (0 or 1) is sum of the estimated probabilities for that outcome for all observation in that group. The observed values for an outcome (0 or 1) are sum of the instances of zeros or ones for all observation in that group. Eventually, for each group two observed values (one for each outcome) and two estimated values (one for each outcome) are available. The Hosmer-Lemeshow test uses a pearson test statistic to compare the observed and fitted counts for these 10 groups. So for 10 groups, there are 20 estimated and observed probabilities used to calculate pearson chi square statistic (X^2).

This test statistic can be approximated by chi-squared with degree of freedom = number of groups -2.

where $X^2 = \sum_{i=1}^{20} \frac{(observed-estimated)^2}{estimated}$

of freedom 8 for calculated X^2 statistcs.

2.4.2 Verifying the Hypothesis

To prove this hypothesis, 50,000 binary data were generated with the conditions fulfilling the discussion on explanatory variable in logistic regression by Lai et al. [21]. These 50,000 instances of binary (0 or 1) data were generated for population distribution of 50:50 and 70:30. These synthetic instances have true probability values which can be compared to the fitted probabilities. Further, these distributions of 50:50 and 70:30 were sampled to obtain seven samples of distributions 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, and 80:20 for each population distribution.

Table 2.1: P-value for Hosmer-Lemeshow test for different sample distribution for population distribution of 50:50 and 70:30

Sample Distribution	P-Value for (50:50) Hosmer-Lemeshow	P-Value for (70:30) Hosmer-Lemeshow
20:80	0.34	0.29
30:70	0.10	0.92
40:60	0.38	0.91
50:50	0.98	0.31
60:40	0.16	0.81
70:30	0.80	0.99
80:20	0.35	0.73

Then the logistic regression model was developed for each of these samples and their corresponding Hosmer-Lemeshow P-values were computed. The computed P value of the Hosmer-Lemeshow test for each sample is presented in Table 2.1. It is evident from this table that the highest P-value for the Hosmer-Lemeshow test is for the sample that has same distribution as the population. For the population distribution of 50:50, the sample distribution of 50:50 has the highest P-value (0.98). Similarly, for population distribution of 70:30, the highest P-value of 0.99 is for sample distribution of 70:30. This verifies our hypothesis that the Hosmer-Lemeshow P-value statistic is highest for the sample that has the same distribution as the population. In other words, the Hosmer-Lemeshow P-value statistic can be an indication of how close a sample distribution is to its population distribution.

A scatter plot between the actual and predicted probability for the samples of population distribution of 50:50 and 70:30 are presented in Figure 2.1 and Figure 2.2 respectively. It is evident from these figures that the difference in actual and predicted probabilities is minimal when the sample has the same distribution as the population.

There is one caveat while using Hosmer-Lemeshow test: (Oliver [15]) if two different combinations of liquefaction/ non-liquefaction instances are used to develop logistic regression model for a particular sample distribution, it gives slightly different P-values. This is important in the present study because the P-values are very high for all sample distributions, and a small change can affect the true distribution. To address this problem, a total number of instances for all seven distributions were made same for synthetic data as well as for SPT and CPT based case histories.

2.5 Liquefaction Data Catalogue

2.5.1 SPT Database

The SPT based liquefaction and non-liquefaction instances were obtained from Idriss& Boulanger [26]. This database contains 227 instances of which 115 instances are liquefied and 112 instances are non-liquefied. Idriss& Boulanger [26] compiled this dataset by removing 21 instances from Cetin et al. [17] due to misclassification and added some additional instances from the 1995 Kobe Earthquake and few others. As shown in Table 2.2, the 227 instances of liquefaction/non-liquefaction are from 25 earthquakes since 1944 Tohankai to 1995 Hyogenken –Nambu earthquake. The variables required for the liquefaction analysis using SPT data are M_w, A_{max}, CSR, depth of

critical layer, ground water depth, fines content, normalized blow count for overburden stress ($N_{1,60,cs}$), and fines content (FC), and these have range of values 5.9-8.3, 0.05-0.84, 0.04-0.49, 1.75-14.34m, 0-7.2m, 4.7-63.7, and 0-92% respectively. More detailed information about case histories can be found in Idriss and Boulanger [26].

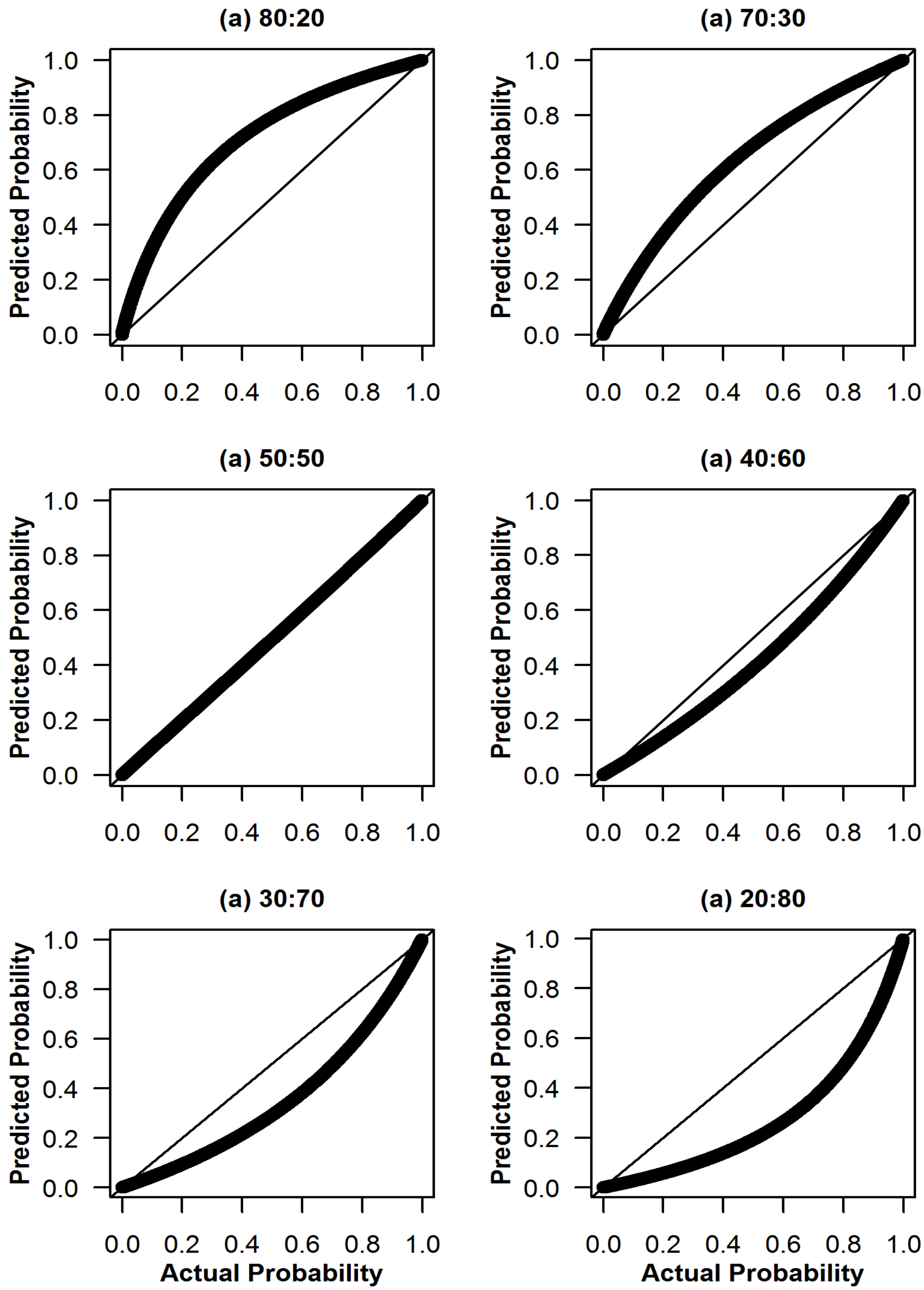

Figure 2.1: Scatter plot of predicted probabilities to actual probabilities for different sample distributions obtained from the population distribution of 50:50.

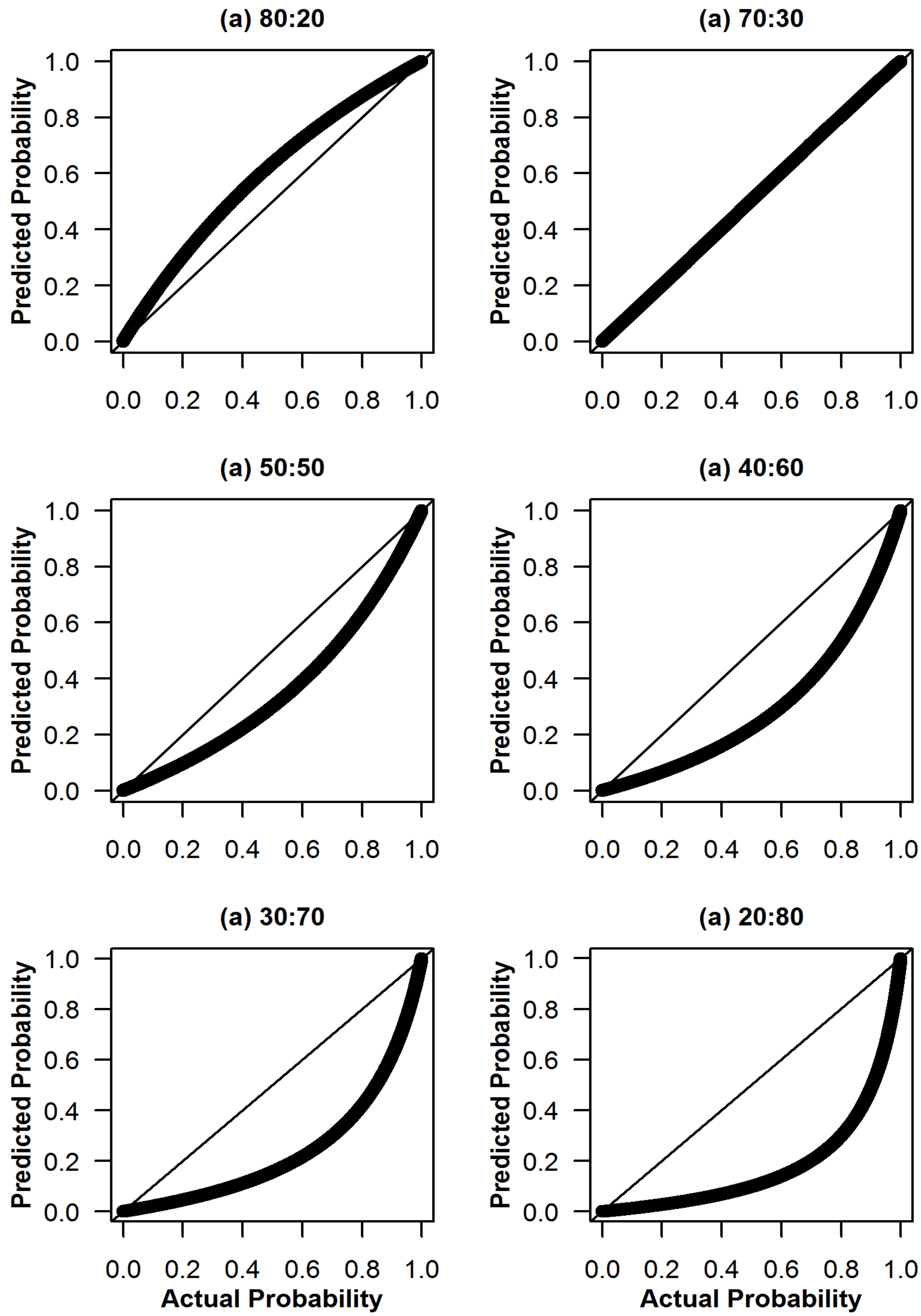

Figure 2.2: Scatter plot of predicted probabilities to actual probabilities for different sample distributions obtained from the population distribution of 70:30.

Table 2.2: SPT case histories information

Earthquake	No. of liquefaction instances	No.of non-liquefaction instances
1944 Tohankai	3	0
1948 Fukai	2	0
1964 Niigata	7	4
1968 Hososhima	0	1
1968 Tokachi-Oki	3	2
1971 San Fernendo	2	0
1975 Haicheng	3	1
1976 Guatemala	1	1
1976 Tangshan	5	2
1977 Argentina	3	2
1978 Miyagiken-oki	1	13
1978 Miyagiken-oki	14	6
1979 Imperial Valley	4	5
1980 Mid-Chiba	0	2
1981 Westmorland	3	4
1982 Urakwa-Oki	0	1
1983 Nihonkai-Chubu	1	2
1983 Nihonaki-Chubu	16	13
1984 Hososhima	0	1
1987 Superstition Hills	1	11
1989 Loma Prieta	15	9
1990 Luzon	1	1
1993 Kushiro-Oki	2	1
1994 Northridge	3	1
1995 Hyogoken-Nambu	25	29

2.5.2 CPT Database

The CPT based liquefaction and non-liquefaction instances were obtained from the dataset compiled by Ku et al. [30]. This dataset has a total of 165 instances of which 125 are from instances of liquefaction and 40 are non-liquefaction instances. Ku et al. [30] compiled this dataset using152 (116 liquefied and 36 non-liquefied) instances from Robertson [25] and 13 (9 liquefied and 4 non-liquefied) instances from Moss et al. [29]. Robertson [25] dataset was obtained by modifying the Moss et al. [16] dataset to exclude 30 (23 liquefaction and 7 non-liquefaction) instances that were classified as C. Robertson [25] described the class C data as unreliable because they were obtained using a mechanical cone. Previous studies by

Robertson & Chambella [3] have shown that mechanical cone friction sleeve value can be significantly different from electric cone value in same soil and hence not comparable.

As shown in Table 2.3, the database compiled by Ku et al. [30] contains information from 16 earthquakes starting from the 1964 Nigata earthquake to 1999 Kocali earthquake. In this 165 instances of liquefaction and non-liquefaction, the variables magnitude of earthquake (M_w), A_{max}, CSR, depth of critical layer, friction ratio(R_f), soil behavior index(I_c), and normalized tip resistance for overburden stress and fines content($q_{c,1ncs}$) ranges from 5.8-7.8, 0.08-0.70, 0.08-0.60, 1-12m, 0.04-3.65 %, 1.31-2.58, and 39.66-187.62 respectively. More detailed information about each case can be found in Ku et al. [30].

Table 2.3: CPT case histories information

Earthquake	No. of liquefaction instances	No.of non-liquefaction instances
1999 Kocali	15	0
1999 Chi Chi	14	0
1995 Kobe	17	6
1994 Northridge	5	0
1989 Loma Prieta	37	16
1987 Edgecumbe	10	5
1987 Elmore Ranch	0	1
1987 Superstition	1	0
1983 Nihonkai	1	1
1983 Borah Peak	4	0
1981 Westmorland	3	2
1980 Mexicali	4	1
1979 Imperial Valley	2	3
1976 Tangshan	9	4
1968 Inaguaha	1	0
1964 Niigata	2	1

2.6 Evaluation of SPT based case histories

In order to identify the population distribution of the SPT data, the verified hypothesis was applied on, the 227 instances of liquefaction/ non-liquefaction from

Idriss and Boulanger [26]. This dataset was sampled to obtain seven different samples of distribution 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, and 80:20. One of the uncertainties that can affect the logistic regression model is to select appropriate explanatory variables. These explanatory variables must follow the normal distribution. Kolmogorov-Smirnov (KS) test (Adapted from Ang et al., [23]) was performed on $N_{1,60,cs}$, $\sqrt{N_{1,60,cs}}$, CSR7.5.1, and $\log(CSR_{7.5.1})$ variables for each of the seven sampled distributions. KS test is used to compare the experimental cumulative distribution with the cumulative distribution

function of an assumed theoretical distribution. The best combination of explanatory variables based on P-value for each variable was log $(CSR_{7.5.1})$ and $\sqrt{N_{1,60,cs}}$ as shown in Table 2.4. A P-value of more than 0.05 was desired for an explanatory variable to have normal distribution. The variable $CSR_{7.5.1}$ has a P-value less than 0.05 for 20:80 sample distribution, and the variable $N_{1,60,cs}$ has P-value less than 0.05 for 60:40 and 70:30 sample distribution. Therefore, the variables $\log(CSR_{7.5})$ and $\sqrt{N_{1,60,cs}}$ are more appropriate than the variable $CSR_{7.5}$ and $N_{1,60,cs}$, and they were selected as explanatory variables in logistic regression.

.Table 2.4: P-value for KS test for explanatory variables for SPT based analysis

Sample Distribution	**P-value for KS-Test for $\sqrt{N_{1,60,cs}}$**	**P-value for KS-Test for log($CSR_{7.5,1}$)**	**P-value for KS-Test For $N_{,160,cs}$**	**P-value for KS-Test For $CSR_{7.5,1}$**
20:80	0.67	0.34	0.15	**0.04**
30:70	0.62	0.22	0.10	0.16
40:60	0.45	0.10	0.08	0.20
50:50	0.35	0.10	**0.05**	0.26
60:40	0.34	0.08	**0.04**	0.22
70:30	0.47	0.14	**0.04**	0.19
80:20	0.52	0.20	0.06	0.17

After selecting the most appropriate explanatory variables, logistic regression model was developed for seven sample distributions to estimate population distribution. Further P-value for Hosmer-Lemeshow statistic was estimated for all seven sample distributions. The highest P-value (0.99) as shown in Table 2.5 was obtained for sample distribution of 50:50. Therefore, based on our hypothesis, the sample distribution of 50:50 is closest to the population distribution for SPT based dataset by Idriss and Boulanger [26]. In other words, the probability estimates obtained from the logistic regression equation developed using this sample distribution (50:50) will have the minimal difference from the true probabilities.

The significance of explanatory variables were also checked using Wald statistic (Adapted from Agresti, [22]), and it was found that the coefficients of $\sqrt{N_{1,60,cs}}$ and $\log(CSR_{7.5,1})$ were highly significant. A P-value less than 0.05 was desirable for an explanatory variable to be significant in a logistic regression model. The P-value for Wald statistic for the coefficients of the two explanatory variables ranged from 0 to 0.001, which indicates high significance.

Table 2.5: P-value for Hosmer-Lemeshow for sample distributions for SPT based analysis.

Sample Distribution	P-Value for Hosmer-Lemeshow
20:80	0.85
30:70	0.54
40:60	0.82
50:50	**0.99**
60:40	0.93
70:30	0.91
80:20	0.85

2.7 Evaluation of CPT based case histories

In order to identify the population distribution of the CPT data, the verified hypothesis was applied on the 165 instances of liquefaction/ non-liquefaction from the Ku et al. [30]. This dataset was sampled to obtain seven different samples of distribution 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, and 80:20. The selections of appropriate explanatory variables are essential to reduce some of the model uncertainty. These appropriate explanatory variables must follow the normal distribution. KS test (Adapted from Ang et al., [23]) was performed on $q_{c,1ncs}$, $\sqrt{q_{c,1ncs}}$, CSR7.5, and log($CSR_{7.5}$) variables for each of the seven sampled distributions. The most suitable explanatory variables among four variables tested were log($CSR_{7.5}$) and $\sqrt{q_{c,1ncs}}$. A P-value of more than 0.05 was desired for an explanatory variable to have normal distribution. The variable $CSR_{7.5}$ has a P-value less than 0.05 for 20:80, 30:70 and 40:60 sample distributions (Table 2.7). However, the variable $q_{c,1ncs}$ has no P-value less than 0.05 but the P-value for each sample distribution is lower than the corresponding P-value for $\sqrt{q_{c,1ncs}}$ variable. Therefore, the variables log($CSR_{7.5}$) and $\sqrt{q_{c,1ncs}}$ are more appropriate than the variable $CSR_{7.5}$ and $q_{c,1ncs}$, and they were selected as explanatory variables in logistic regression.

After selecting the most appropriate explanatory variables, logistic regression model was developed for seven sample distributions to estimate population distribution. Further P-value for Hosmer-Lemeshow statistic was estimated for all seven sample distribution. The highest P-value (0.99) as shown in Table 2.7 was obtained for sample distribution of 40:60. Based on our hypothesis, it can be inferred from the results of Table 2.7 that the sample distribution of 40:60 is closest to the population distribution for CPT based case histories from Ku et al. [30]. Therefore, the probability estimates obtained from the logistic regression equation developed using this sample distribution (40:60) should have the minimal difference from the true probabilities.

Table 2.6: P-value for KS test for explanatory variables for CPT based analysis

Sample Distribution	**P-value for KS-Test for $\sqrt{q_{c,1ncs}}$**	**P-value for KS-Test for log(CSR $_{7.5}$)**	**P-value for KS-Test for $q_{c,1ncs}$**	**P-value for KS-Test for CSR $_{7.5}$**
20:80	0.86	0.31	0.82	**0.01**
30:70	0.82	0.49	0.67	**0.02**
40:60	0.67	0.57	0.42	**0.04**
50:50	0.70	0.87	0.40	0.21
60:40	0.75	0.78	0.46	0.42
70:30	0.46	0.91	0.45	0.45
80:20	0.56	0.81	0.29	0.29

The significance of explanatory variables were also tested using Wald statistic (Adapted from Agresti [22]), and it was found that the coefficients of $\sqrt{q_{c1ncs}}$ and $\log(CSR_{7.5})$ were highly significant. A P-value less than 0.05 was desirable for an explanatory variable to be significant in a logistic regression model. The P-value for Wald statistic for coefficients of two explanatory variables were between 0 to 0.01, which exhibits high significance.

Table 2.7: P-value for Hosmer-Lemeshow for sample distributions for CPT based analysis

Sample Distribution	P-Value for Hosmer-Lemeshow
20:80	0.80
30:70	0.72
40:60	**0.99**
50:50	0.97
60:40	0.89
70:30	0.81
80:20	0.62

2.8 Results and Discussion

2.8.1 Discussion for SPT based method

After applying Hosmer-Lemeshow test on seven sampled distributions, the population distribution for SPT case histories was estimated as 50:50, and it is similar to the distribution of liquefaction and non-liquefaction instances in 227 case histories from Idriss and Boulanger [26]. A probabilistic curve assigned to triggering curve or deterministic boundary as shown in Idriss and Boulanger [19] depends on consideration of model uncertainty and parameter uncertainty in a statistical model. For example, the triggering curve proposed by Idriss and Boulanger [19] lies nearby 15% probability curve of Idriss and Boulanger [26] when only model uncertainties were considered and lies nearby 35% when parameter uncertainty along with model uncertainties were also considered.

In this study, only model uncertainties were considered, and therefore, 15%

probability curve which is also in range of 10-15% proposed in Seed [27] is recommended as deterministic boundary.

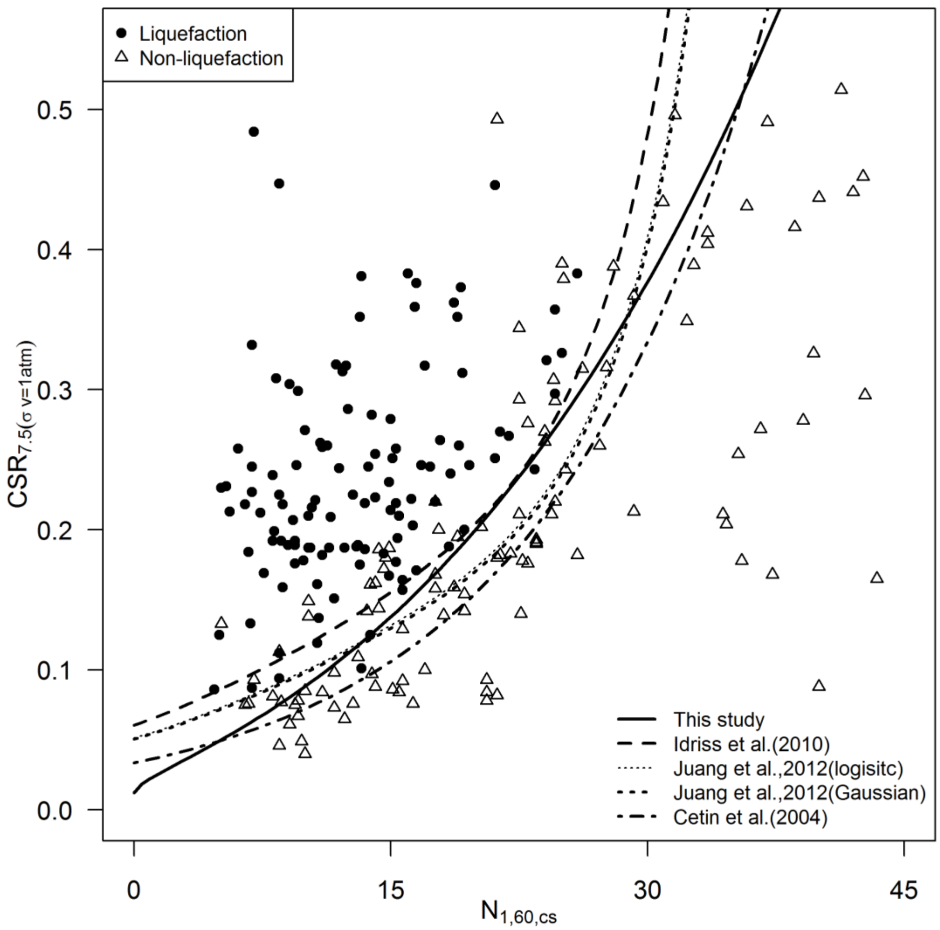

Figure 2.3: The comparison of population distribution 15% probability curve from this study with other researchers. (Note-Juang et al., 2012(logistic) and Juang et al., 2012(gaussian) overlaps.)

Moreover, this deterministic boundary placed most of the liquefied points on left side supports the needed conservatism in the absence of consideration of parameter uncertainty. The Figure 2.3 shows a comparison of 15% probability curve by different researchers (Cetin et al. [17]; Idriss and Boulanger [26]; Juang et al. [31])

with 15% curve developed from this study using logistic regression for population distribution of 50:50.

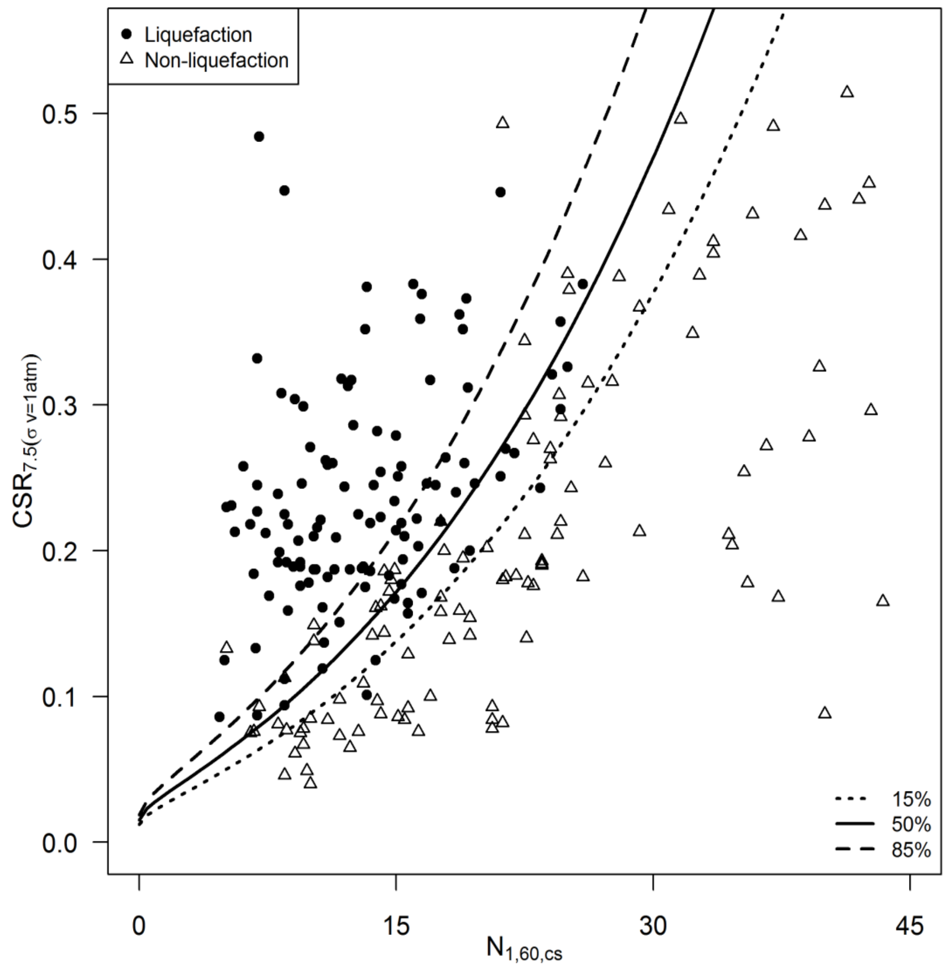

Figure 2.4: 15, 50, and 85 percent probability curve from logistic regression for population distribution (50:50) in this study.

It is observed from Figure 2.3 that for higher CSR values, the curve from this study matches the curve from Cetin et al. [17]. This is more acceptable considering the observation of Seed [27] that at higher CSR values, Idriss and Boulanger [26] curve is non-conservative. Seed [27] also noted that for higher CSR values, the Cetin et al. [17] curve matched with the Yoshimi et al. [6] curve developed from cyclic testing of high quality frozen samples. At lower CSR values (less than 0.15), many liquefied points positioned right side of (Idriss and Boulanger [26]) curve which were supposed to be on left side of the curve; this makes again Idriss and Boulanger [26] 15% curve non-conservative. However, for CSR values between 0.15 and 0.30, the population distribution (50:50) 15% probability curve from this study and Idriss and Boulanger [26] curve gives the same probability. This non-conservativeness in Idriss and Boulanger [26] curve could be partially due to the use of the same weighting ratio of 1.5 recommended by Cetin et al. [13], which could have induced significant uncertainty in their model. In addition, Idriss and Boulanger [32] also mentioned the lack of clarity in estimation of weighting ratio. If Idriss and Boulanger [26] used the weighting ratio of 1.02 based on the population distribution of 50:50, their 15% probability could have positioned lower. It is also supported from the result from Idriss and Boulanger [26] study that a weighting ratio of 1.2 lowered 15% probability curve in comparison to position of 15% probability curve when a weighting ratio of 1.5 was used.

Juang et al. [31] selected weighting factor by intuition, and have shown that weighted likelihood function generated the same result as without weights which is not in line with Idriss and Boulanger [26] and this study. This might be due to inability of their statistical model to consider sampling bias. However, the 15% curve of the two best models (gaussian and logistic) as shown in Figure 2.3of Juang et al. [31] was positioned lower relative to Idriss and Boulanger [26], but it is still higher than the population distribution curve from this study at low and high CSR values. Figure 2.4 shows the 15%, 50%, and 85% probability curve which can be considered as boundary between low, moderate, and high liquefaction zone respectively. Here, only model uncertainty is considered; No variability in parameters $CSR_{7.5.1}$ and $N_{1,60,cs}$ was considered.

The logistic regression equation developed in this study with population distribution of 50:50 to get probabilities values for mean values of $CSR_{7.5.1}$ and $N_{1,60,cs}$ for a critical layer is:

$$\text{Logit}(P_L) \leftrightarrow \quad \log(P_L/1-P_L)=32.89-4.92*\text{sqrt}(N_{160cs}) +7.84*\log(CSR_{7.5,1})$$

$$P_L=\frac{1}{1+\exp(-(32.89-4.92*\sqrt{N_{160cs}}+7.84*\log(CSR_{7.5.1})))} \tag{1}$$

Where P_L is the probability of liquefaction and the $CSR_{7.5.1}$ and $N_{1,60,cs}$ can be calculated using procedure given by Idriss and Boulanger [26].

2.8.2 Discussion for CPT based method

After applying Hosmer-Lemeshow test on seven sampled distributions, the population distribution for CPT case histories was estimated as 40:60. Here, the 10% probability curve is recommended as deterministic curve because it matches with the Robertson and Wride [9] triggering curve. Robertson and Wride [9] is the most agreed deterministic method to evaluate liquefaction potential using CPT and was maintained until in Robertson [25]. Moreover, 10% probability curve as deterministic curve is also in range of 10-15% suggested by Seed [27]. In addition, the 10% probability curve as deterministic boundary possess needed conservatism in absence of consideration of variability in $CSR_{7.5}$ and $q_{c,1ncs}$., and placed most of the liquefied instances on its left side. It can be seen from Figure 2.5 that at low CSR values, 15 % Juang et al. [20], deterministic Robertson and Wride [9], 15% Moss et al. [16], and 15% population distribution (40:60) from this study probability curves match each other considerably.

Juang et al. [20] did not include any weighting factor to consider sampling bias in first order reliability method to evaluate liquefaction potential. Moss et al. [16] used the same weighting factor of 1.5 used by Cetin et al. [13], and this factor observed a population distribution of 64:36 which is almost opposite to the population distribution (40:60) estimated in this study. This might be due to two reasons: firstly, they used the same weighting ratio estimated in Cetin et al. [13], and this weighting factor could have been different, if Cetin et al. [17] had removed the 21 misclassification case histories; Secondly, 30 case histories from Moss et al. [16] were also found of poor quality (Ku et al. [30]) which might affect the position of the probability curve. However, the 15% probability curve from Moss et al. [16] and 10% probability curve from this study fall near to each other for $CSR_{7.5}$ values less than 0.45.

Ku et al. [30] used the population distribution of 45:55 which was recommended by Juang et al. [24], and this population distribution provided a weighting ratio of 3.75 for their case histories which is not in the range of 1 to 3 given in Cetin et al. [13]. Ku et al. [30] proposed 35% probability curve for Robertson and Wride [9] deterministic curve which is too far from the 10-15% range proposed in Seed [27]. Furthermore, the Figure 2.6 shows that 20% probability curve developed with logistic regression model in this study for a sample distribution of 75:25 which is nearly the distribution of case histories from Ku et al. [30], and this matches with the 15% probability curve of Ku et al. [30] probabilistic model. It seems that the use of weighting ratio does little affect the position of 15% probability curve, therefore hints the inability of their model to consider sampling bias appropriately.

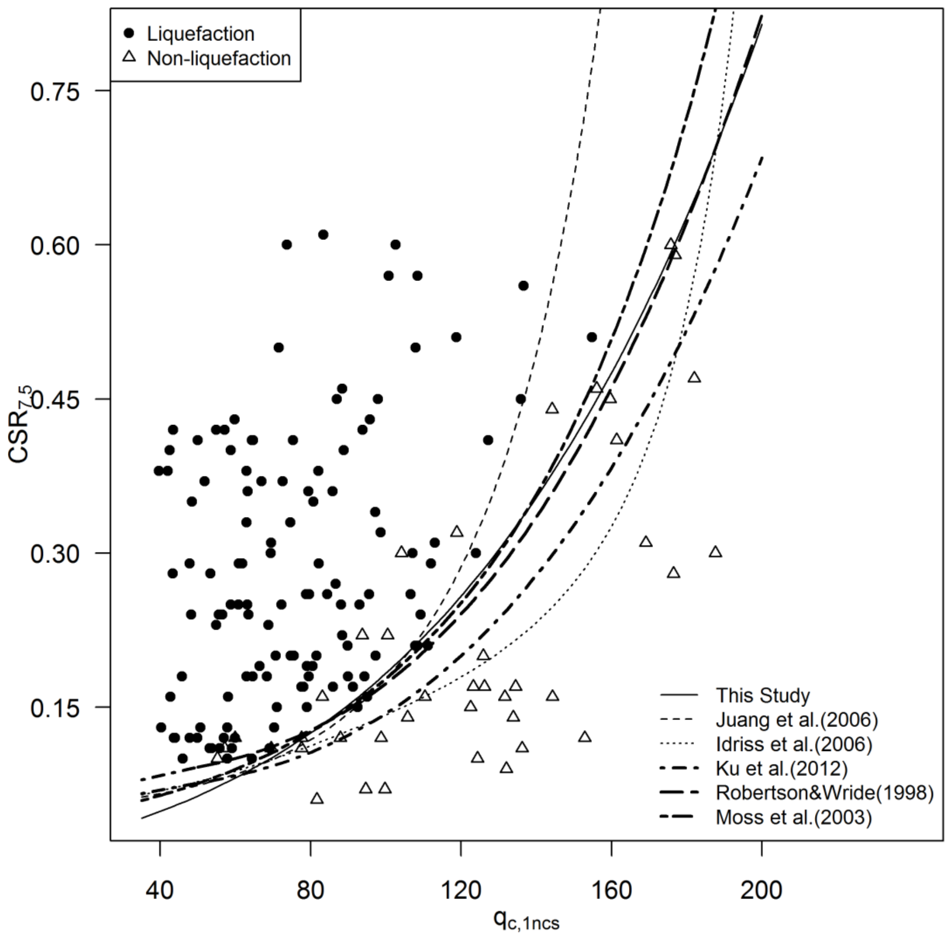

Figure 2.5: The comparison of population distribution 15% probability curve from this study with other researchers

Figure 2.7 shows the effect of inconsistency and quality of data on logistic regression model. This was plotted using case histories documented in Juang et al. [20] which also includes 182 case histories from Moss et al. [16], and case histories from Ku et al. [30] that excludes 30 inconsistent instances form Moss et al. [16]. It has been analyzed using logistic regression model for case histories from both Juang et al. [20] and Ku et al. [30] paper that even though the population distribution was the same as 40:60 in both case histories (estimated in this study), the difference in position of 15% probability curves was significant. It can be inferred that the presence of some poor quality data can undermine the consideration of all aspects of a statistical model uncertainty, and that model can give misleading probability values. The availability of more high quality data in the near future will be able to remove this uncertainty considerably.

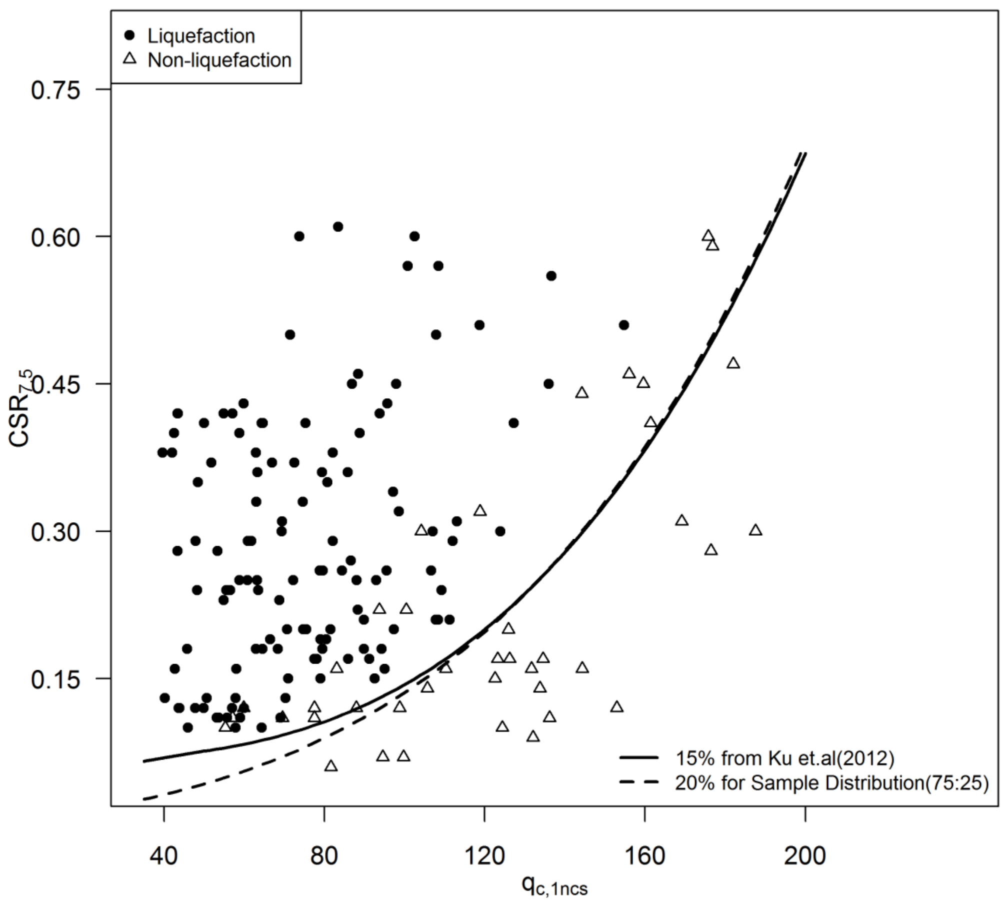

Figure 2.6: The comparison between probability curves of case histories distribution of 75:25

There is one higher P-value (0.97) for Hosmer-lemeshow test (Table 2.7) for 50:50 sample distributions. When the probability curves for this sample distribution plotted with the population distribution of 40:60, it matches entirely with population distribution (Figure 2.8). This may be due to either a relatively low number of non-liquefaction instances in comparison to non-liquefaction instances in SPT database (112 non-liquefied case histories for SPT in comparison to only 40 non-liquefied for CPT) or presence of some poor quality data. It suggests the need of more high quality non-liquefaction instances to distinguish as clearly as possible the difference between the sample distribution and population distribution

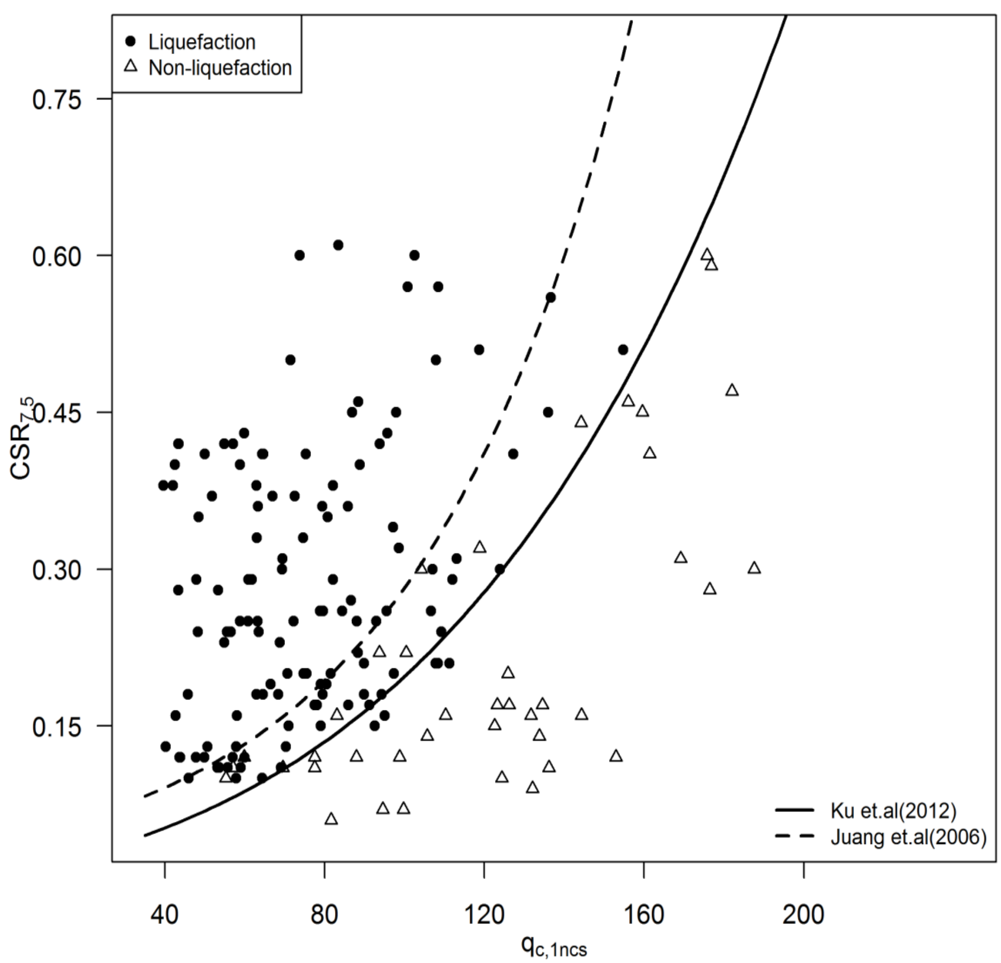

Figure 2.7: Impact of data quality on 15% probability curve developed using logistic regression model

The Figure 2.9 shows the 15%, 50%, and 85% probability curve which can be considered as boundary between low, moderate, and high liquefaction zone respectively. This model seems more appropriate to be a probabilistic version of Robertson [25] than the Ku et al. [30]. Again, here only model uncertainty is considered; no variability in $CSR_{7.5}$ and $q_{c,1ncs}$ was considered.

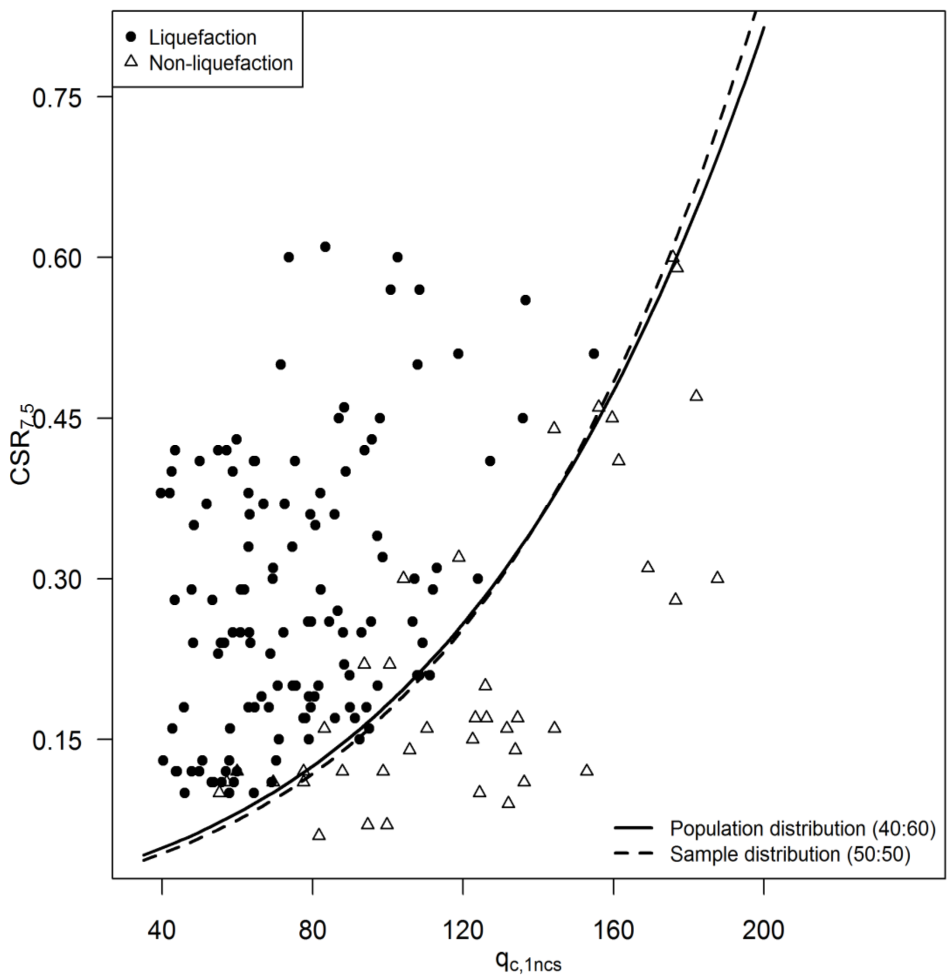

Figure 2.8: Comparison of high P-value (0.97) sample distribution with population distribution.

The logistic regression equation is given below to get probability value of a liquefaction incident for the mean values of $CSR_{7.5}$ and $q_{c,1ncs}$ for a critical layer :

$\text{Logit}(P_L) \leftrightarrow \log(P_L/1\text{-}P_L)=30.324\text{-}2.21*\text{sqrt}(q_{c,1ncs}) +6.138*\log(CSR_{7.5})$

$$P_L=\frac{1}{1+\exp(-(30.32-2.21*\sqrt{q_{c,1ncs}}+6.138*\log(CSR_{7.5})))} \tag{2}$$

The $CSR_{7.5}$ and $q_{c,1ncs}$ can be calculated using procedure given in Robertson and Wride [9].

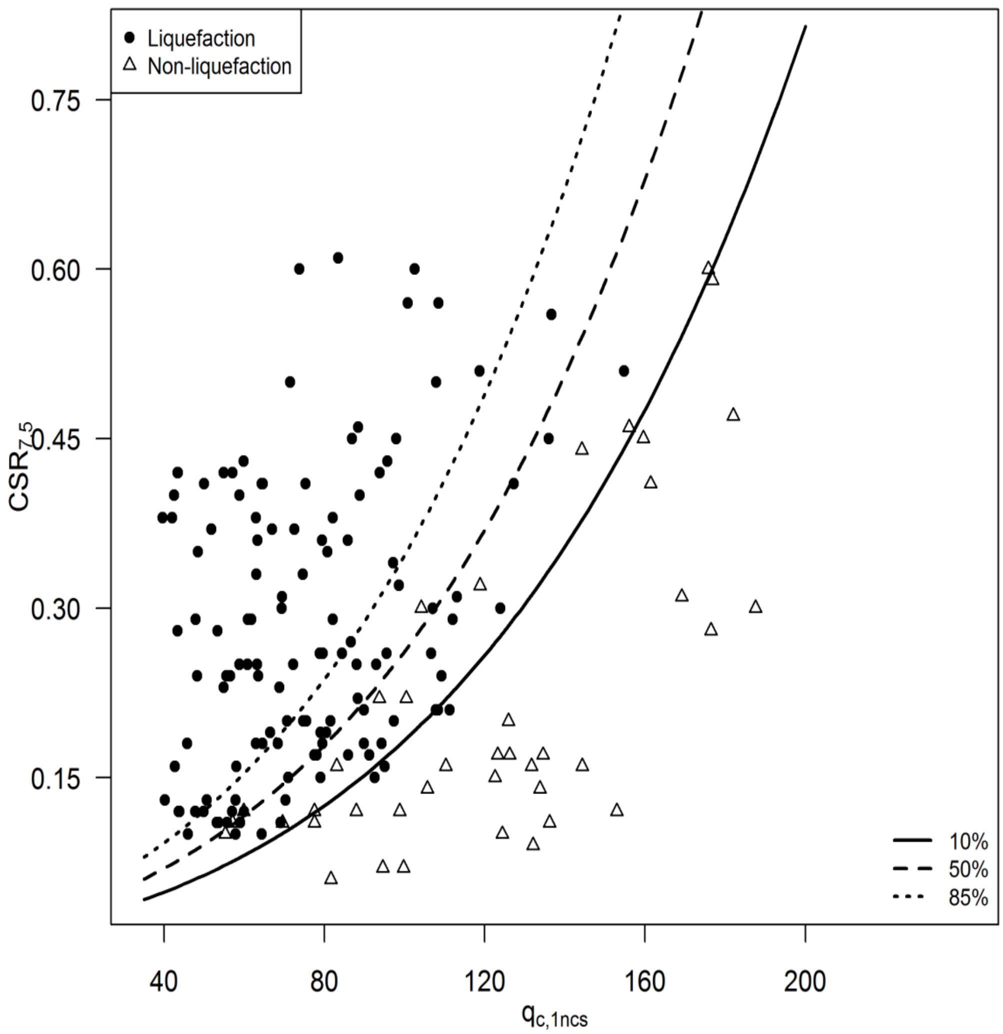

Figure 2.9: 10, 50, and 85 percent probability curves for population distribution.

2.9 Conclusion

Sampling bias is an important part of model uncertainty that can have significant impact on population distribution and thus, on predicted probabilities. This paper has shown a way to considerably remove this epistemic uncertainty by using Hosmer-lemeshow statistic in logistic regression model. Further, this paper presented a methodology to estimate population distribution for most updated SPT and CPT based case histories. The associated logistic regression model with population distribution can further be used to estimate near true probabilities as hypothesized in this study. The population distribution for Ku et al. [30] CPT case histories is 40:60 and the population distribution for Idriss and Boulanger [26] SPT case histories is 50:50.

In this study, the different model uncertainties such as distribution of explanatory variables, significance level of parameter (coefficients) estimated and sampling bias in liquefaction/non-liquefaction instances were addressed using different statistical tests. In addition, the logistic regression model can give misleading probability values if one of those uncertainties is not considered. The impact of data quality on logistic regression model is significant, and even undermines the consideration of all the model uncertainties stated above. The recent updated database and collection of high quality data in near future can enhance the predictive efficacy of logistic regression model.

The logistic regression equation (1) and (2) in this paper for SPT and CPT based analysis respectively can be used to find the liquefaction potential of a site having variables values in range of the latest SPT and CPT case histories.

Logistic regression is the most popular method to analyze binary problems. The combination of logistic regression with Hosmer-Lemeshow test can be a promising tool to analyze binary problems like liquefaction, landslides, etc., with continuous explanatory variables. The continuous explanatory variables are most common in geoengineering. Apart from the evaluation of population distribution for liquefaction, this methodology can also be used to achieve fitness of a logistic regression model with continuous explanatory variables.

2.10 References

1. H.B. Seed, I.M. Idriss, "Simplified procedure for evaluating soil liquefaction potential". J. Soil Mechanics and Foundation Division, ASCE, 97 (SM9) (1971), pp. 1249–1273.
2. H.B. Seed, K. Tokimatsu, L.F. Harder Jr,R. Chung, "Influence of SPT procedures in soil liquefaction resistance evaluations".J. Geotechnical Engineering, ASCE, 111(12) (1985), pp. 1425–1445.
3. P.K. Robertson, R.G. Campanella, "Liquefaction Potential of sands using the cone penetration test". J. Geotechnical Engineering Division, ASCE, 111(3) (1985), pp.384-406.
4. S.S.C. Liao, D. Veneziano, R.V. Whitman, "Regression model for evaluating liquefaction probability".J. Geotechnical Engineering Division, ASCE, 114(4) (1988), pp.389-411.
5. S. Lai, H. Tsuchida, K. Koizumi, "A liquefaction criterion based on field performances around seismographs stations". Soils and Foundations, 29 (2) (1989), pp.52-68.
6. Y. Yoshimi, K. Tokimatsu, J. Ohara, "In-situ liquefaction resistance of clean sands over a wide density range". Geotechnique, 44 (3) (1994), pp. 479-494.
7. S.L. Kramer, "Geotechincal earthquake engineering".Prentice Hall (1996).
8. T.L. Youd, S.K. Noble, "Liquefaction criteria based on statistical and probabilistic analyses". Proceedings of the NCEER workshop on evaluation of liquefaction resistance of soil, Technical Report NCEER-97-022(1997), pp.201-205.
9. P.K. Robertson, C.E. Wride, "Evaluating cyclic liquefaction potential using the cone penetration test". Canadian Geotechnical J., Ottawa, 35(3) (1998), pp. 442-459.
10. S. Toprak, T.L. Holzer, M.J. Bannett, J.C. Tinslay, "CPT and SPT-based probabilistic assessment of liquefaction potential." Proceedings of 7th U.S.–Japan Workshop on Earthquake Resistant Design of Lifeline Facilities and Countermeasures Against Liquefaction, Seattle, Technical Report MCEER-99-0019(1999).
11. D.W. Hosmer, S. Lemeshow, "Applied Logistic Regression". John Wiley and Sons (2000).
12. T.L.Youd, I.M.Idriss, R.D. Andrus, I.Arango, G. Castro, J.T. Christian, R. Dobry, W.D.L. Finn, L.F. Harder, M.E. Hynes, K. Ishara, J.P. Koester,

S.S.C. Liao, W.F. Marcuson, G.R. Martin, J.K. Mitchell, Y. Moriwaki, M.S. Power, P.K. Robertson, R.B. Seed, K.H. Stokoe, "Liquefaction resistance of soils: summary report from the 1996 NCEER and 1998NCEER/NSF workshops on evaluation of liquefaction resistance of soils". J. Geotechnical and Geoenvironmental Engineering, ASCE, 127 (10) (2001), pp. 817–833.

13. K.O. Cetin, A. Der Kiureghian, R. B. Seed, "Probabilistic models for the initiation of seismic soil liquefaction". Structural Safety, 24(1) (2002), pp. 67–82.
14. C.H Juang, T. Jiang, R.D. Andrus, "Assessing probability-based methods for liquefaction evaluation". J Geotechnical and Geoenvironmental Engineering, ASCE, 128(7) (2002), pp.580–589.
15. K. Oliver, " Global goodness of fit tests in logistic regression with sparse data." Statistics in Medicine, 21 (2002), pp. 3789-3801.
16. R.E.S. Moss, " CPT-based Probabilistic Assessment of Seismic Soil liquefaction initiation". PhD. Thesis (2003), Universityof California, Berkeley.
17. K.O. Cetin, R.B. Seed, A. Der Kiureghian, K. Tokimatsu, L.F. Harder Jr., R.E. Kayen, R.E.S. Moss, "Standard penetration test-based probabilistic and deterministic assessment of seismic soil liquefaction potential". J. Geotechnical and Geoenvironmental Engineering, ASCE, 130 (12) (2004), pp.1314–1340.
18. R.E.S. Moss, B.D. Collins, D.H. Whang, "Retesting of Liquefaction/Nonliquefaction Case Histories in the Imperial Valley". Earthquake Spectra, 21(1) (2005), pp.179-196.
19. I.M. Idriss, R.W. Boulanger, "Semi-empirical procedures for evaluating liquefaction potential during earthquakes". Soil Dynamics and Earthquake Engineering, 26(2006), pp.115–130.
20. C.H. Juang, S.Y. Fang, E.H. Khor, "First-order reliability method for probabilistic liquefaction triggering analysis using CPT". J. Geotechnical and Geoenvironmental Engineering, ASCE, 132 (3) (2006), pp.337–350.
21. S.Y. Lai, W.J. Chang, P.S. Lin, "Logistic regression model for evaluating soil liquefaction probability using CPT data".J. Geotechnical and Geoenvironmental Engineering, ASCE, 132 (6) (2006), pp.694–704.
22. A. Agresti, "An introduction to categorical data analysis". John Wiley and Sons (2007).

23. A.H-S. Ang, W.H. Tang, “Probability concepts in engineering: Emphasis on application in civil & environmental engineering”. John Wiley and Sons (2007).
24. C.H. Juang, S.Y. Fang, W.H. Tang, E.H. Khor, G.T.C. Kung, J. Zhang, “Evaluatingmodel uncertainty of an SPT-based simplified method for reliability analysis for probability of liquefaction”. Soils and Foundations, 49 (1) (2009), pp.137–154.
25. P.K. Robertson, “Performance based earthquake design using the CPT”. In proceedings of IS-Tokyo 2009: International conference on performance-based design inearthquake geotechnical engineering- From case history to practice, Tokyo, Japan, 15-18 June(2009).
26. I.M. Idriss, R.W. Boulanger, “SPT-based liquefaction triggering procedures”. Report No. UCD/CGM-10/02, Center for geotechnical modeling, Department of civil and environmental engineering (2010), University of California at Davis.
27. R.B. Seed, “Technical Review and commemts: 2008 EERI Monograph ‘Soil liquefaction during earhquakes ‘(by I.M. Idriss and R.W. Boulanger)”. Geotechnical Report No. UCB/GT – 2010/01, University of California at Berkeley (2010).
28. T. Oommen, L.G. Baise, R.M. Vogal, “Sampling bias and class imbalance in maximum-likelihood logistic regression”. Mathematical Geosciences, 43(2011), pp.99-120.
29. R.E.S. Moss, R.E. Kayen, L.Y. Tong, S.Y. Liu, G.J. Kai, J. Wu, “Retesting of Liquefaction and Nonliquefaction Case Histories from the 1976 Tangshan Earthquake.”. J. Geotechnical and Geoenvironmental Engineering, ASCE, 137 (2011), pp.334–343.
30. C.S. Ku, C.H. Juang, C.W. Chang, J. Ching, “Probabilistic version of the Robertson and Wride method for liquefaction evaluation: development and application”. Canadian Geotechnical J., 49 (2012), pp.27–44.
31. C.H. Juang, J. Ching, Z. Luo, C.S. Ku, “New models for probability if liquefaction using standard penetration tests based on an updated database of case histories”. Engineering Geology, 133-134(2012), pp.85-93.
32. R. Boulanger, I. Idriss, “Probabilistic Standard Penetration Test–Based Liquefaction–Triggering Procedure”. J. Geotechnical and Geoenvironmental Engineering, ASCE, 138(10) (2012), pp.1185–1195.

Chapter 3 Future Work

3.1 Future work

In present study, only model uncertainties in logistic regression were considered. The variability in explanatory variable can cause significant uncertainty in a statistical model. A probability curve developed in Idriss and Boulanger [4] for deterministic boundary shown in Idriss and Boulanger [3] for liquefaction/non-liquefaction instances was dependent on the consideration of either model uncertainties alone or model uncertainties with parameter uncertainties. Idriss and Boulanger [4] obtained a 15% probability curve matching with deterministic boundary (Idriss and Boulanger [3]) when only model uncertainties were used. On the other hand, Idriss and Boulanger [4] obtained a 35% probability curve matching with deterministic boundary (Idriss and Boulanger [3]) when parameter uncertainties along with model uncertainty were considered. A probability difference of 20% can be critical when it is used for high risk projects. However, it is reasonable to use only model uncertainties when the probability of liquefaction is evaluated for low risk project or for mapping liquefaction hazard for a region.

The main part of parameter uncertainty is caused in the estimation of $CSR_{7.5}$. The equation (6) in chapter 1 shows that A_{max} is required to calculate $CSR_{7.5}$. The Campbell [1] proposed an attenuation relationship to estimate A_{max} for a distance from the earthquake source and for a given magnitude. This attenuation relationship retains a standard deviation of 0.57 which is significantly high, and subsequently induces substantial uncertainty in estimation of $CSR_{7.5}$. Another parameter caused a significant uncertainty is ground water level, and Moss et al. [2] assigned a standard deviation of 0.3 for ground water level. Apart from uncertainties in $CSR_{7.5,}$ uncertainties in obtaining liquefaction resistance of soil by $N_{1,60,cs}$ (for SPT based analysis) and $q_{c,1ncs}$ (for CPT based analysis) can be important. Sometimes these parameter uncertainties can supersede the model uncertainties. Therefore, to make the methodology proposed in Chapter 2 to evaluate liquefaction potential by using logistic regression with Hosmer-Lemeshow statistic more efficient and applicable in most conditions, the inclusion of parameter uncertainties are required.

3.2 References

1. K.W. Campbell, “Near source attenuation of peak horizontal acceleration”. BSSA, 71 (6) (1981), pp. 2039-2070.
2. R.E.S. Moss, “CPT-based Probabilistic Assessment of Seismic Soil liquefaction initiation”. PhD. Thesis (2003), University of California, Berkeley.
3. I.M. Idriss, R.W. Boulanger, “Semi-empirical procedures for evaluating liquefaction potential during earthquakes”. Soil Dynamics and Earthquake Engineering, 26(2006), pp.115–130.
4. I.M. Idriss, R.W. Boulanger, “SPT-based liquefaction triggering procedures”. Report No. UCD/CGM-10/02, Center for geotechnical modeling, Department of civil and environmental engineering (2010), University of California at Davis.

Chapter 4 Conclusion

4.1 Conclusion

This paper made an attempt and provided a methodology to reduce one important epistemic uncertainty (sampling bias) in evaluating the probability of seismically induced liquefaction. The logistic regression model with Hosmer-Lemeshow test was used to identify the population distribution for most updated SPT and CPT based case histories. The associate logistic regression model with respect to estimated population distribution can further be used to calculate true (reliable) probabilities as described in given hypothesis. The population distribution for Ku et al. [3] CPT case histories is 40:60, and the population distribution for Idriss and Boulanger [2] SPT case histories is 50:50.

In this study, the different model uncertainties like distribution of explanatory variables, significance level of parameter (coefficients) estimated and sampling bias in liquefaction/non-liquefaction instances were addressed using different statistical test. A statistical test always reveals the uncertainty in a statistical model. Hosmer-lemeshow, KS test and Wald statistic removed most part of uncertainty from logistic regression model. Furthermore, maximum uncertainties in a statistical model should be addressed because not acknowledging one uncertainty can have minor or major impact on the outcomes of a statistical model even all other uncertainties were well addressed.

The impact of data quality on logistic regression model is significant, and even undermines the consideration of all the model uncertainties stated above. As described in this study, even addressing every single uncertainty by appropriate statistical tests, the poor quality of data could deviate probability curve significantly which can be misleading. The recent updated database for SPT (Idriss and Boulanger [2]) and CPT (Ku et al. [3]) were scrutinized in past, and can be now considered most reliable in present time. The more collection of high quality data in near future can enhance the predictive efficacy of logistic regression model. For practical purpose, this problem can also be ameliorated by using more than one statistical model to average the effect of this uncertainty.

The logistic regression equation (1) and (2) in Chapter 2 for SPT and CPT based analysis respectively can be used to find the liquefaction potential for a site. These equations are easy to estimate probability of liquefaction. For a given site, the variables $CSR_{7.5}$ and q_{c1ncs} can be calculated using procedure given in Robertson and Wride [1] for CPT base analysis. At other side, the variable $CSR_{7.5.1}$ and N_{160cs} can be calculated using procedure given by Idriss and Boulanger [2].

Logistic regression is the most popular method to analyze binary problems. The combination of logistic regression with Hosmer-Lemeshow test can be a capable tool to analyze binary problems like liquefaction, landslides etc. with continuous explanatory variables. The continuous explanatory variables are most common in geosciences and engineering like soil properties, rainfall, temperature etc. The combination of logistic regression and Hosmer-Lemeshow can be used for a region to map hazards like liquefaction and landslides. Apart from the evaluation of population distribution for liquefaction, this methodology can also be used to achieve fitness of a logistic regression model with continuous explanatory variables. A statistical model with higher P-value is considered better or more fit to data than a lower P-value.

4.2 References

1. P.K. Robertson, C.E. Wride, “Evaluating cyclic liquefaction potential using the cone penetration test”. Canadian Geotechnical J., Ottawa, 35(3) (1998), pp. 442-459.
2. I.M. Idriss, R.W. Boulanger, “SPT-based liquefaction triggering procedures”. Report No. UCD/CGM-10/02, Center for geotechnical modeling, Department of civil and environmental engineering (2010), University of California at Davis.
3. C.S. Ku, C.H. Juang, C.W. Chang, J. Ching, “Probabilistic version of the Robertson and Wride method for liquefaction evaluation: development and application”. Canadian Geotechnical J., 49 (2012), pp.27–44.

Appendix Copyright Permissions

Pearson Education

Permission 1

Permission belongs to figures 1.7, 1.8, and 1.9 of this thesis, and these figures appear in “Geotechnical Earthquake Engineering” as figure 9.4, E91, and 9.24 respectively.

PEARSON

Legal/Permissions
One Lake Street
Upper Saddle River, NJ 07458
Fax: 201-236-3290
Phone: 201-236-3281
Vineta.Lewis@Pearson.com

Dec 19, 2012 PE Ref # 174787

Abhishek Jain
MICHIGAN TECHNOLOGICAL UNIV
422, Department of Geological Engineering
1400 Townsend Drive
HOUGHTON, MI 49931-1295 Fax #: 906-487-3371

Dear Abhishek Jain:

You have our permission to include content from our text, ***GEOTECHNICAL EARTHQUAKE ENGINEERING, 1st Ed. by KRAMER, STEVEN L.,*** in your masters thesis at MICHIGAN TECHNOLOGICAL UNIV.

Content to be included is:
pp. 353, 370, 371 Figs. 9.4, 9.24, E91

Permission does not include any content that appearson by permission or courtesy of other sources.
If so, you may contact those sources for permission.

Please credit our material as follows:
KRAMER, STEVEN L., GEOTECHNICAL EARTHQUAKE ENGINEERING, 1st Edition, © 1996. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ

Sincerely,
Vineta Lewis, Permissions Supervisor

Permission 2

Permission belongs to figures 1.12 and 1.13 of this thesis, and these figures appear in "Geotechnical Engineering: Principles & Practices" as figure 3.24 and 3.28 respectively.

PEARSON

Legal/Permissions
One Lake Street
Upper Saddle River, NJ 07458
Fax: 201-236-3290
Phone: 201-236-3281
Vineta.Lewis@Pearson.com

Dec 19, 2012 PE Ref # 174788

Abhishek Jain
MICHIGAN TECHNOLOGICAL UNIV
422, Department of Geological Engineering
1400 Townsend Drive
HOUGHTON, MI 49931-1295 Fax #: 906-487-3371

Dear Abhishek Jain:

You have our permission to include content from our text, ***GEOTECHNICAL ENGINEERING: PRINCIPLES & PRACTICES, 2nd Ed. by CODUTO, DONALD P.; YEUNG, MAN-CHU RONALD; KITCH, WILLIAM A.,*** in your masters thesis at MICHIGAN TECHNOLOGICAL UNIV.

Content to be included is:
pp. 91, 99 Figs. 3.24, 3.28

Permission does not include any content that appears by permission of or courtesy of another source. If so, then you may contact that source for permission.

Please credit our material as follows:
CODUTO, DONALD P.; YEUNG, MAN-CHU RONALD; KITCH, WILLIAM A., GEOTECHNICAL ENGINEERING: PRINCIPLES & PRACTICES, 2nd Edition, © 2011. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ

Sincerely,
Vineta Lewis, Permissions Supervisor

Center of Geotechnical Modeling, UC Davis

Permission belongs to figures 1.10 and 1.11 of this thesis, and these figures appear in “Report No. UCD/CGM-10-02” as figure B1 (Appendix B) and 4.17 respectively.

MichiganTech Abhishek Jain <jabhishe@mtu.edu>

Request for permission to reprint figures
3 messages

Abhishek Jain <jabhishe@mtu.edu> Sun, Dec 2, 2012 at 5:07 PM
To: cgm@ucdavis.edu

Hi CGM,

I am a master student at Michigan Technological University. I want to use two of the figures (Fig 4.17- page 67 and Figure B1-page B10,Appendix B) from Report No. UCD/CGM-10-02 in my Master's Thesis. So,I am requesting you to permit me to republish these two figures.

Thanks

Regards,
Abhishek Jain
Graduate Research Assistant,
Secretary MTU AEG Student Chapter,
Dept of Geological and Mining Eng & Sciences
422,Dow Engineering
Michigan Technological University
Phone-906 487 3098

Ross W. Boulanger <rwboulanger@ucdavis.edu> Sun, Dec 2, 2012 at 6:33 PM
To: Abhishek Jain <jabhishe@mtu.edu>
Cc: "cgm@ucdavis.edu" <cgm@ucdavis.edu>

Dear Abhishek,

Permission granted. Best of luck with your studies.

Ross

On Sun, Dec 2, 2012 at 2:07 PM, Abhishek Jain <jabhishe@mtu.edu> wrote:

> Hi CGM,
>
> I am a master student at Michigan Technological University. I want to use two of the figures (Fig 4.17- page 67 and Figure B1-page B10,Appendix B) from Report No. UCD/CGM-10-02 in my Master's Thesis. So,I am requesting you to permit me to republish these two figures.
>
> Thanks
>
> Regards,
> Abhishek Jain
> Graduate Research Assistant,
> Secretary MTU AEG Student Chapter,
> Dept of Geological and Mining Eng & Sciences
> 422,Dow Engineering
> Michigan Technological University
> Phone-906 487 3098

--
Ross W. Boulanger, Professor
Director, Center for Geotechnical Modeling
Department of Civil & Environmental Engineering
University of California, Davis, CA 95616
http://cee.engr.ucdavis.edu/faculty/boulanger

Abhishek Jain <jabhishe@mtu.edu> Mon, Dec 3, 2012 at 1:50 PM
To: "Ross W. Boulanger" <rwboulanger@ucdavis.edu>

Thanks a lot CGM and Prof. Boulanger.

Regards

United States Geological Survey (USGS)

Permission belongs to figures 1.4, 1.5 and 1.6 of this thesis, and the pictures in these figures are in public domain.

USGS science for a changing world

USGS Visual Identity System

Acknowledging or Crediting USGS as Information Source

Revised May 10, 2006

Most U.S. Geological Survey (USGS) information resides in the public domain and may be used without restriction. When using information from USGS information products, publications, or Web sites, we ask that proper credit be given. Acknowledging or crediting the USGS as an information source can be provided by including a line of text citation such as those shown below. Note that some non USGS photographs, images, and/or graphics are used by the USGS with permission from the copyright holder. To use these copyrighted materials, you must obtain permission from the copyright holder under the copyright law.

If you are an outside organization that has a cooperative partnership or contractual agreement with the USGS that will result in the creation of various information products and you plan to use the trademarked USGS identifier, refer to the use of the trademarked USGS identifier by non-USGS organizations web page for guidance.

When Using USGS Information at No Cost

- (*Product or data name*) courtesy of the U.S. Geological Survey
- (*Photograph*) courtesy of the U.S. Geological Survey
  Example: dlvg32 software courtesy of the U.S. Geological Survey

When Purchasing and Reproducing USGS Information for Resale

- Data compiled by the U.S. Geological Survey
- (*Product or data name*) produced by the U.S. Geological Survey
- (*Data name*) data compiled by the U.S. Geological Survey

When Packaging USGS Information with Non-USGS Products for Resale

- (*Product or data name*) produced by the U.S. Geological Survey

When Value-Adding to USGS Information for Resale

- Source of (*product or data name*): U.S. Geological Survey
- U.S. Geological Survey (*product or data name*) used in compiling this information
  Example: Source of Parkfield seismic data: U.S. Geological Survey

Optional Statement

The following statement may be added after an acknowledgment or credit:

The USGS home page is http://www.usgs.gov.

Return to Visual Identity System home

U.S. Department of the Interior | U.S. Geological Survey
URL: http://www.usgs.gov/visual-id/credit_usgs.html
Page Contact Information: visual_id@usgs.gov
Page Last Modified: Tuesday, 03-Aug-2010 14:33:13 EDT