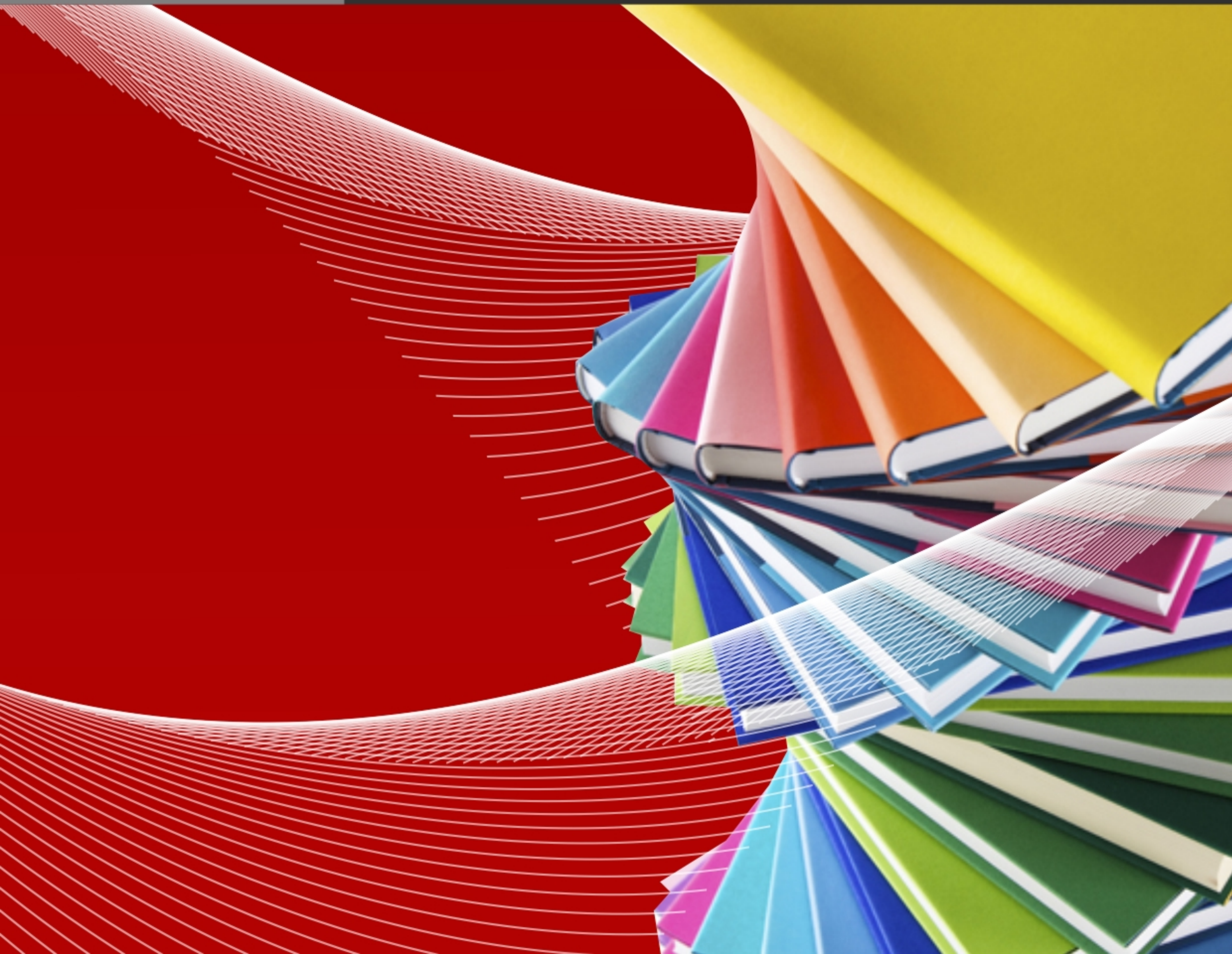
澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Outstanding Academic Papers by Students
學生優秀作品



Correlation Analysis of Shear Wave Velocity of Macau Soils using Nonlinear Curve Fitting and Genetic Programming

by

NG KA FAI

Final Year Project Report submitted in partial fulfillment of the requirement of the Degree of

Bachelor of Science in Civil Engineering

2014/2015

Faculty of Science and Technology
University of Macau

DECLARATION

I declare that the project report here submitted is original except for the source materials explicitly acknowledged and that this report as a whole, or any part of this report has not been previously and concurrently submitted for any other degree or award at the University of Macau or other institutions.

I also acknowledge that I am aware of the Rules on Handling Student Academic Dishonesty and the Regulations of the Student Discipline of the University of Macau.

Signature : 伍 家 輝

Name : NG KA FAI

Student ID : DB127257

Date : 18th May. 2015

APPROVAL FOR SUBMISSION

This project report entitled "**Correlation analysis of Shear Wave Velocity of Macau soil using Nonlinear Curve Fitting and Genetic Programming**" was prepared by NG KA FAI in partial fulfillment of the requirements for the degree of Bachelor of Science in Civil Engineering at the University of Macau.

Endorsed by,

Signature :

Supervisor : Dr. Lok, Man Hoi

ACKNOWLEDGEMENTS

I would like to express my most sincere appreciation to my supervisor, Dr. Lok Man Hoi, for his keen guidance and encouragement throughout my final year project.

I would like to sincere thanks to NGAI SHUN engineering firm, LECM, and Dr. Lee, Hay Yip who provided the information of LRT project and Central Part of Taipa. On the other hand, I would like to sincere thanks to SaeTeng, Pun Hou Kun, Chao Sai Choi and Kuan Wai Kin for their organization of the significant amount of data to allow me to complete the final year project conveniently.

ABSTRACT

In the past half century, a number of researchers studied the correlation between shear wave velocity and SPT-N value based on the simple power-law regression model. However, the prediction performance is not good enough to apply in practice with confidence. It may be caused by using a single measurement of SPT-N and neglecting the influence of vertical effective stress and cone resistance. Therefore, those parameters are combined with SPT-N to establish new correlation models for an existing database in Macau in this study.

In this project, nonlinear curve fitting algorithm and genetic programming in Matlab software were used for the correlation analysis between the shear wave velocity with SPT-N value, vertical effective stress and cone resistance, for database obtained from the LRT project. Correlation models with different format and different soil parameters are compared in order to obtain the improved performance of model prediction.

Based on the results of this study, the prediction performance is improved based on the additional soil parameters combined to form correlation models with SPT-N. Comparing between vertical effective stress and cone resistance, cone resistance has better performance when combined with SPT-N to form power law regression model than vertical effective stress. In addition, the prediction performance has successfully improved when genetic programming was applied to establish correlation of shear wave velocity. However, the complexity of the correlation models from genetic programming, the format of which cannot be assigned directly, is much more complex than the simple power law function.

TABLE OF CONTENTS

CHAPTER 3 FIELD TESTING

CHAPTER 4 RESULT OF PREVIOUS STUDY

CHAPTER 5 METHODOLOGY

CHAPTER 7 CONCLUSION

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

LIST OF TABLES

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Background

Shear wave velocity(V_s) is a fundamental geotechnical parameter that acts as the main input of site response analysis in geotechnical earthquake engineering.

Standard Penetration Test (SPT-N) is basic soil parameter which is commonly used to indicate the density and compressibility of granular soils. It also commonly applied to estimate the liquefaction potential of saturated granular soils for earthquake design. In the past of half century, numbers of researchers trend to use regression statistical analysis to establish relationship between the shear wave velocity and SPT-N value in the worldwide. Macau had already been working this research in recent years. The motivation of this research is because the direct measurement of shear wave velocity need to deal with very high cost and lacked in workers with official qualification and skill to perform the shear wave velocity field test, numbers of research trends to applied the statistical regression method to modify correlation equations which can predicted accuracy data of shear wave velocity in a specific site condition without spend the extra money and time.

Instead of SPT-N, some literatures published that the effective stress and cone penetration of soil are also necessary to take into account to reduce the statistically errors from neglecting the effect of the other representative parameters. On the other hand, numbers of the advanced computing software is widely applied in this geotechnical field research. Since the correlations between shear wave velocity and SPT-N values have considerable dispersions, it may due to the different condition of measurement of shear wave velocity and SPT-N values, geotechnical and geological conditions and the model type of regression statistical analysis.

For this study, the Matlab software was used to perform the correlation analysis. Regarding to get a higher prediction performance, Matlab software would be applied which contains powerful functions, such as nonlinear curve fitting. Moreover, a previous study suggested that Genetic Programming can perform a sufficient improvement for the correlation performance. In order to achieve those purposes, the Macau soil data set are selected from the final year projects of previous studies in University of Macau.

1.2 Scope of Work

There are three main parts in this Final Year Project. First part is to establish the new correlation between shear wave velocity and SPT-N value based on the same database of Macau soil. Second part is to establish the correlation of shear wave velocity with other parameters besides of SPT-N value. The last part is to establish the correlation of shear wave velocity by using genetic programming toolbox to establish the correlation.

The quality improvement of shear wave velocity prediction can be compared for the correlation with using different method. The prediction performance can be determined by different statistical index, such as coefficient of determination (R^2), root mean square error (RMSE).

Finally, based on the existing database of Macau soils, the correlation model with the higher degree of accuracy can be obtained by comparing different correlation models.

There are total 7 chapters in this study. Chapter 1 is included the introduction and scopes of this project. Chapter 2 is introduced some concepts, theories and data analysis and observation of prediction performance from the existing literatures.

Chapter 3 is introduced the geological background of Macau soil and measurement method of field tests. Chapter 4 is summarized the data analysis and observation of result from the final year projects of previous studies in University of Macau. Chapter 5 is introduced the methodology of linear curve fitting and nonlinear curve fitting modeling, the corresponding models between this two different modeling are compared graphically. Chapter 6 is introduced data analysis and evaluation of new correlation models in this study. Finally, the conclusion of this final year project is summarized in Chapter 7.

CHAPTER 2LITERATURE REVIEW

2.1 Introduction

In this chapter, the concept of correlation analysis based on the simple power-law regression and genetic programming (symbolic regression) will be introduced. However, the predication performance is not only based on the formation of regression models, it also affected by the unpredictable geological uncertainty, the influence of vertical effective stress, and the elimination consideration of the cone resistance. Those contents would be introduced in this chapter.

2.2 Power-law regression

There are many correlation models of shear wave velocity with SPT-N values established in the past various studies. Different correlation models were established by different location or period. Based on the literature that published from Electronic Journal of Geotechnical Engineering (EJGE), they established the new correlation equation between shear wave velocity and SPT-N value based on the 27 various locations with various years so that such correlation model can be applied to all soil types for everywhere.

There are many different opinions may also take as the influence of the correlation equation between shear wave velocity and SPT-N value according to observation from various researchers. It manly included the soil type (Jafari 2002), depth (Holzer et al. 2005), overburden pressure (Brandenberg 2010), Geological Age (Andrus et al. 2009) and corrected SPT N (Anbazhagan et al. 2012).

Since the database was provided by 27 different researchers at different country and different years, they simplicity use the simple power-law correlation analysis to establish the correlation model between shear wave velocity and SPT-N value, which express as:

$$V_s = A \times (N_{60})^B \tag{2.1}$$

where $V_s = shear\ wave\ velocity, m/s$; $N_{60} = penetration\ resistance$; A, B = the regression analysis coefficients.

Since soil behavior is not homogenous in actual environment, the established correlation models only can apply to all soil type in order to reduce the complexity of data analysis. The comparison between 27 correlation models is illustrated in figure 2.1.

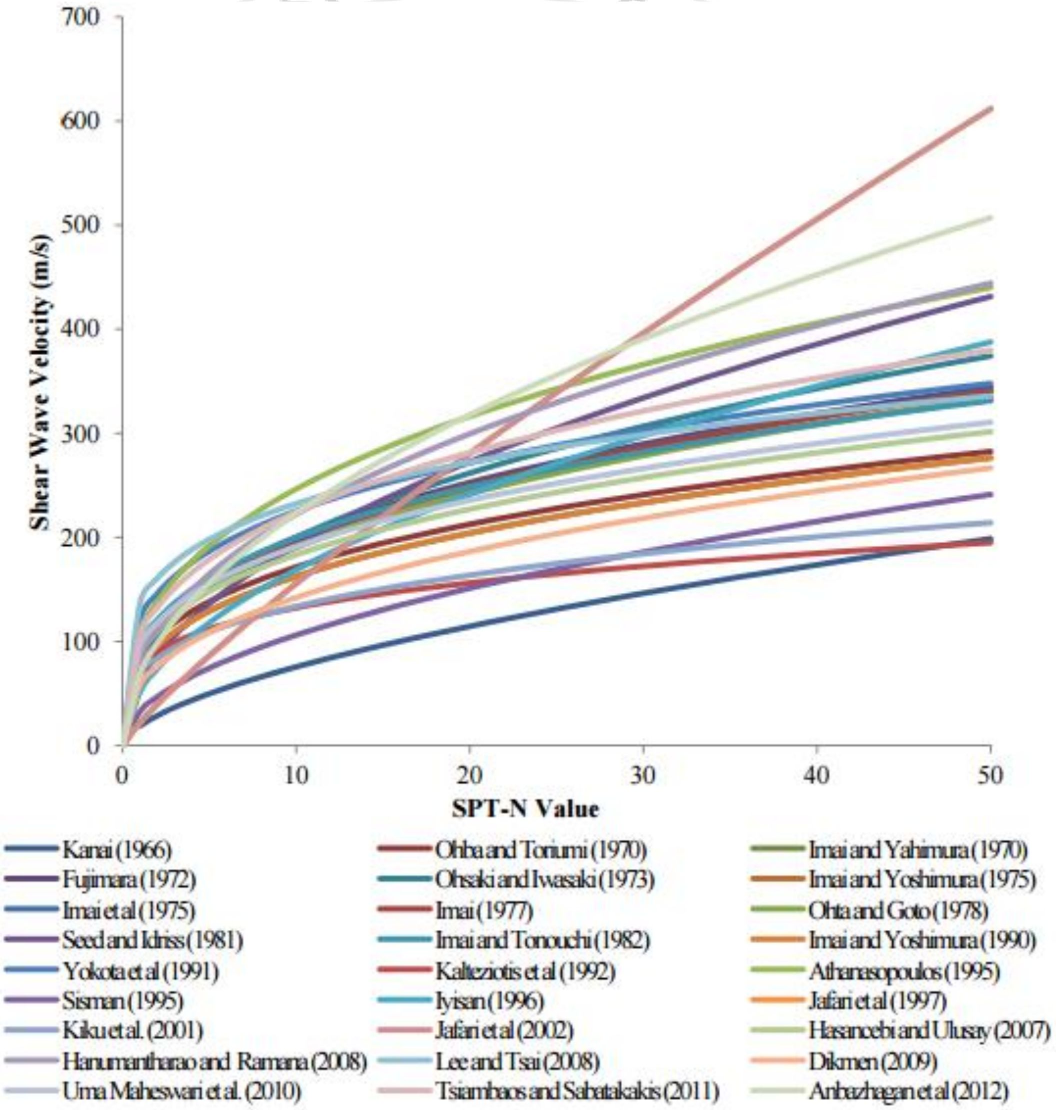


Figure 2.1 Established correlation models for all soil type. (Marto et al. 2013)

Since simple power-law statistical analysis was applied to establish the correlation between shear wave velocity and SPT-N value, the result of correlation model can be established based on two different methods, without outliner and with outliner. The accuracy predication performance is expressed by coefficient of determination (r^2). First, the database of 27 correlations were established by different researchers and combined to establish the new correlation equations:

$$V_s = 69.76N^{0.401}, r^2 = 0.624. \tag{2.2}$$

where $V_s = shear\ wave\ velocity, m/s$; $N_{60} = penetration\ resistance$;

The correlation model was illustrated with other 27 correlation models in figure 2.2.

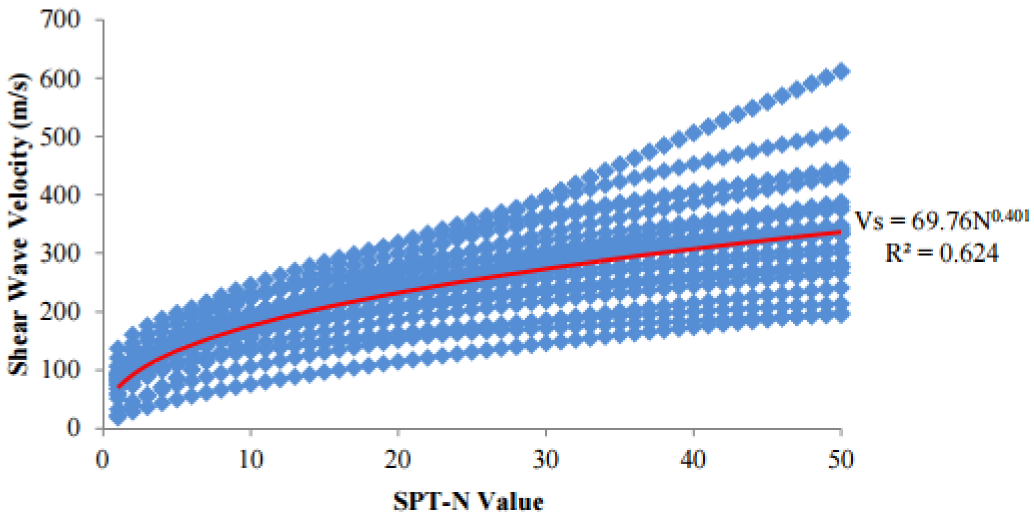


Figure 2.2 Correlation models based on previous 27 correlations for all soil types (Marto et al. 2013)

The low value of coefficient of determination indicated the correlation model provided poor prediction performance. It might cause by the different practice of site investigation works and variation of geological conditions. Therefore, in order to improve the prediction performance, the distribution of the data was assumed normally distributed and identified the data located within the range of mean value$\pm$ standard deviation were selected and acted as upper boundary and lower boundary, respectively. In this case, the mean value of correlation was established based on 68.2% of whole data.

Finally, the mean value of correlation with high boundary and low boundary correlation were established, this correlation model (with outliner) with previous correlation models were illustrated in figure 2.3.

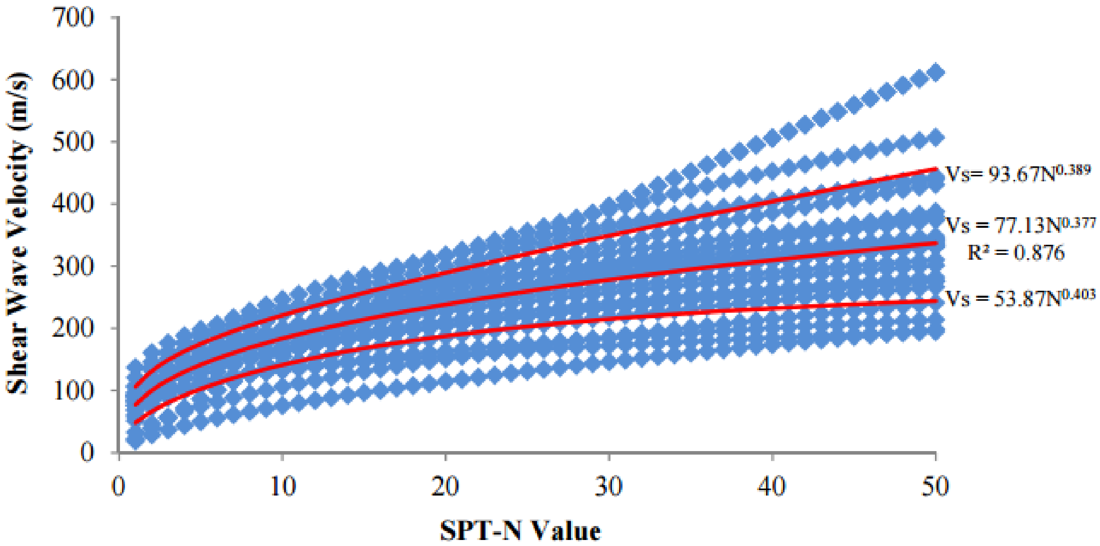


Figure 2.3 Correlation model with higher and lower boundary (Marto et al. 2013)

Table 2.1the comparison between without outliner and with outliner (Marto et al. 2013)

Dataset	correlations	r^2	Remark
Without outliner	$V_s = 93.67N^{0.389}$	0.624	
With outliner	$V_s = 93.67N^{0.389}$		Lower boundary
	$V_s = 77.13N^{0.377}$	0.876	
	$V_s = 53.87N^{0.407}$		Higher boundary

According to table 2.1, although the prediction performance have got significantly improved after outliner was excluded, however, there are no any convincing evidence to plot the regression correlations by lower boundary and higher boundary, so that it is not commonly recommend to perform the correlation equation in this way.

2.3 Geotechnical variability

In the previous topic, single simple power-law regression model of shear wave velocity with SPT-N value was discussed. However, there are various types of geotechnical uncertainty influence the reliability of database and prediction performance of correlation equation. The influences cannot be evaluated directly based on the result of correlation model. For the source of those uncertainly was introduced in the chapter 2.3.1.

2.3.1 Geotechnical uncertainty

Geotechnical variability commonly contains in the natural soil environment, which is a complex behavior that observed from many unpredictable uncertainties. There are mainly three sources of geotechnical uncertainties were classified, which include inherent variability of soil, measurement error of operation, and transformation uncertainty of correlation model. The first type of uncertainty is mainly caused by the effects of natural environment. The second is caused by equipment quality, precision of operation, and random effects due to testing. In other words, inherent soil variability and effect of measurement error can be classified as data scatter. (Phoon & Kulhawy 1999).

The third type of uncertainty is obtained when field test and laboratory test measurements are used to interpret design value of soil properties based on various empirical or other correlation models. Therefore, the contribution between those three primary uncertainties in the design of soil property clearly depends on the site conditions, degree of equipment and procedural operation control, and precision of the correlation model. In that case, the design value of soil properties can only be determined based on total variability analysis. (Kulhawy 1992)

The inherent soil variability which can be interpreted based on Coefficient of Variance (COV). However, there were facing some problems when started to calculate the COVs of inherent soil variability since most of the database reported based on total variability analysis, so that the reported COVs might result a relatively larger than the actual conditions. Four potential problems can be explained why this happened,

1. soil data set was observed from uncertainty of geological;
2. insufficient equipment quality and precision of procedural;
3. deterministic trends in the soil data are not removed;
4. soil dataset were taken over a long time period.

Therefore, the results were already examined critically based on the consideration given above. (Phoon & Kulhawy 1999).

After the available data organized for the comparison purpose, it can summarize the general soil type, the number of data groups and tests per group, and the mean and COV of the soil property with corresponding soil parameters. Since the description of soil type is useful because the site-specific COVs are applicable to other locations and the soil type can be provided to similar soil profile. The number of tests is also a very useful to indicate the accuracy of the mean and COV estimates. If large number of tests per group can be performed, the corresponding errors can be minimized in the statistical point of view. (Phoon & Kulhawy 1999).

According to their summary result, the COVs of inherent soil variability for standard penetration test is around 19 to 62% (from the best to the worst case). The reason why this large amount of range occurred is because of the number of tests per group, which around 2 to 300 times of tests per group. (Phoon & Kulhawy 1999).

For coefficient of variance of measurement error, since measurement error is highly depend on the equipment quality, precision control for procedural and random testing effects. In order to calculating the COVs of measurement error, those three factors were described as the COVs also and take root of sum of square to calculate the COVs of total measurement error and corresponding result is around 15 to 45 percents. (Orchant et al. 1988; Kulhway 1996)

2.3.2 Transformation Uncertainty

Transformation uncertainty is the last primary source of geotechnical uncertainty. Since the inherent soil variability and measurement error have already been introduced in pervious section, so that transformation uncertainty is introduced herein to extend the analysis. After that, the second-moment probabilistic approach was applied to combine with those three primary types of uncertainties based on which design soil properties are derived.

The format of the transformation model is related to the computation method of the corresponding deign soil properties based on corresponding measurement data. However, the transformation models were commonly not applied with confidence because of the low quality of prediction performance. It is because most of transformation models are obtained based on standard power law regression function and the data scatter can only be quantified using probabilistic methods. In this method, they used regression analyses approach to establish the transformation model, and the data from the regression curve were modeled as zero-mean random variable. The standard deviation of this variable can be used to indicate the magnitude of transformation uncertainty, as illustrated in figure 2.4.

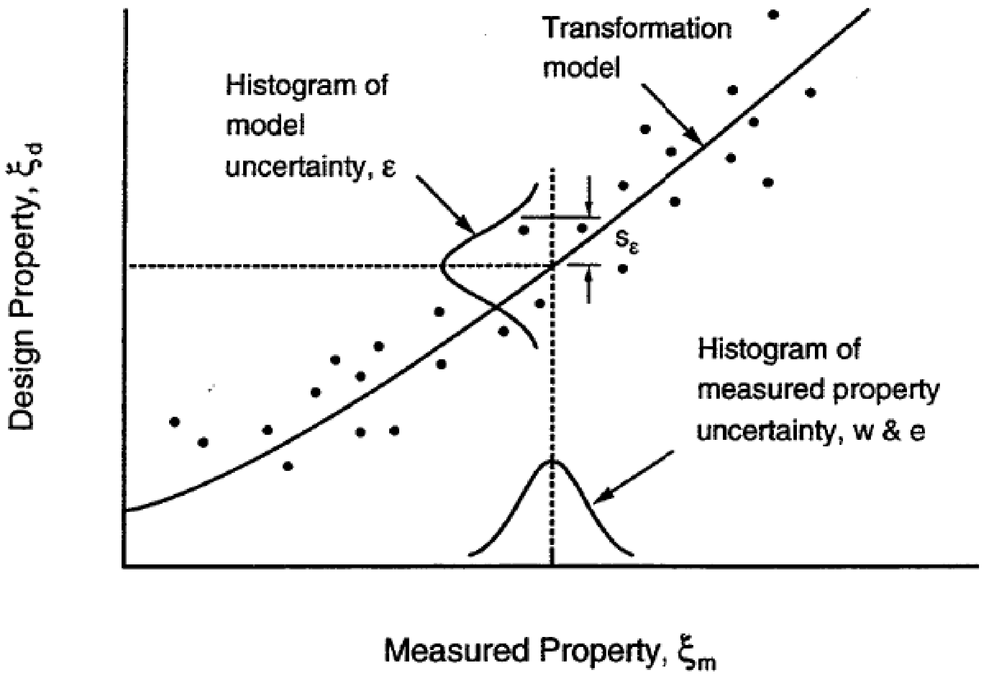


Figure 2.4 the probabilistic characterization of transformation model. (Phoon & Kulhawy 1999).

Transformation uncertainty should be evaluated separately with the other two types of uncertainties. In order to achieve this result, the selection of the database should be satisfied by using least square approach to minimize the effect of inherent soil variability. Moreover, the data set should be measured by same standard approach in order to reduce the systematic measurement errors. (Kulhawy et al. 1992)

Some of transformation models for the corresponding design soil properties and test measurements were summarized based on geotechnical literature. The availability of these models for a soil properties have been determined from various theory, laboratory test or field test, which also included standard penetration test. However, most of the transformation models were based on the empirical method or without sufficient theory to support the model evaluation. Therefore, the total transformation uncertainty for those empirical model cannot be determined directly, but it found that they were likely to be as large as the result for the transformation model when the second-moment statistics method was applied, especially for the correlation model for standard penetration test N value with stress strength properties, liquefaction resistance or the other parameter which is not directly related to standard penetration test N value. However, it still without convincing evidence to say that it can be predicted without considering the significant uncertainties. (Kulhawy & Mayne 1990)

2.4 Influence of vertical effective stress

The correlation relation between shear wave velocity and SPT-N value were mainly established by the simple power-law regression. However, the prediction performance was commonly not good. According to the previous sections, it mentioned that increase the number of test per group or using the site-specific condition database can get an improvement of prediction performance. On the other hand, the influence of effective stress was not commonly considered in the simple power-law regression analysis. (Brandenberg et al. 2010)

2.4.1 Physical relationship between SPT-N and vertical effective stress

In order to evaluate the influence of effective stress, the existing database was collected at various California Bridge sites between 1993 and 2001. It was applied to establish the correlation model between shear wave velocity, SPT-N value and vertical effective stress. The common overburden correction equations of N_{60} and V_s were applied in equation, which can be represented by:

$$(\mathrm{N}_1)_{60} = \left(\frac{\mathrm{P_a}}{\sigma'_\mathrm{v}}\right)^\mathrm{n} \mathrm{N}_{60} \tag{2.3}$$

$$\mathrm{V}_{s1} = \left(\frac{\mathrm{P_a}}{\sigma'_\mathrm{v}}\right)^\mathrm{m} \mathrm{V_s} \tag{2.4}$$

where N_{60} = SPT blow count; $(N_1)_{60}$ = stress-normalized SPT blow count; σ'_v = effective vertical stress, kN/m^2; P_a = atmosphere pressure, kN/m^2; V_s = shear wave velocity, m/s; V_{s1} = stress-normalized shear wave velocity, m/s;

m, n = empirical constants

2.4.2 Application of Least Square approach

Based on the substitution from equation (2.3) and (2.4), the correlation equation can be shown as:

$$V_s = 87.8{N_{60}}^{0.253}\left(\frac{P_a}{\sigma'_v}\right)^{0.253n-m} \tag{2.5}$$

where n, m = empirical constants; N_{60} = SPT blow count; P_a = atmosphere pressure, kN/m^2; V_s = shear wave velocity, m/s; σ'_v = effective vertical stress, kN/m^2;

The magnitude of exponent n and m depend on soil type, cementation, and plasticity index, however, this is commonly hard to estimate the value of m and n, especially when the directly geophysical measurement is not available. Therefore, the least-squares approach was applied to solve this problem. The corresponding statistical random regression model had been linearization and can be shown:

$$\ln(\overline{V}_s)_{ij} = \beta_0 + \beta_1 \ln(N_{60})_{ij} + \beta_2 \ln(\sigma'_v)_{ij} + \eta_i + \varepsilon_{ij} \tag{2.6}$$

where N_{60} = SPT blow count; V_s = shear wave velocity, m/s; σ'_v = effective vertical stress, kN/m^2;η_i = random effect for i^{th} boring (i.e. inter-boring effect); ε_{ij} = variation for the j^{th} measurement from i^{th} boring (i.e. intra-boring effect); $\beta_0, \beta_1, \beta_2$ = regression constants; η_i = inter-boring variation; ε_{ij} = intra-boring variation

The random effect and variation of measurement for the each boring are assumed to be independent and normally distributed with two different terms of standard deviations. It permitted the possibility that might over-predict the measurement of shear wave velocity for some location and under-predict for others. (Brandenberg et al. 2010) The simple power-law regression between shear wave velocity and SPT-N cannot provide this feature.

The regression results based on equation (2.6) are represented for sand, silt and clay correspondingly by using the lmer function in R, which is the open-source software environment for statistical computing. Since it involved both N_{60} and σ'_v, the multi-variable regression model are required. In order to have a fair comparison of the relative influence of N_{60} and σ'_v between each soil type, the models contain with the mean value and plus and minus one standard deviation for N_{60} and σ'_v.

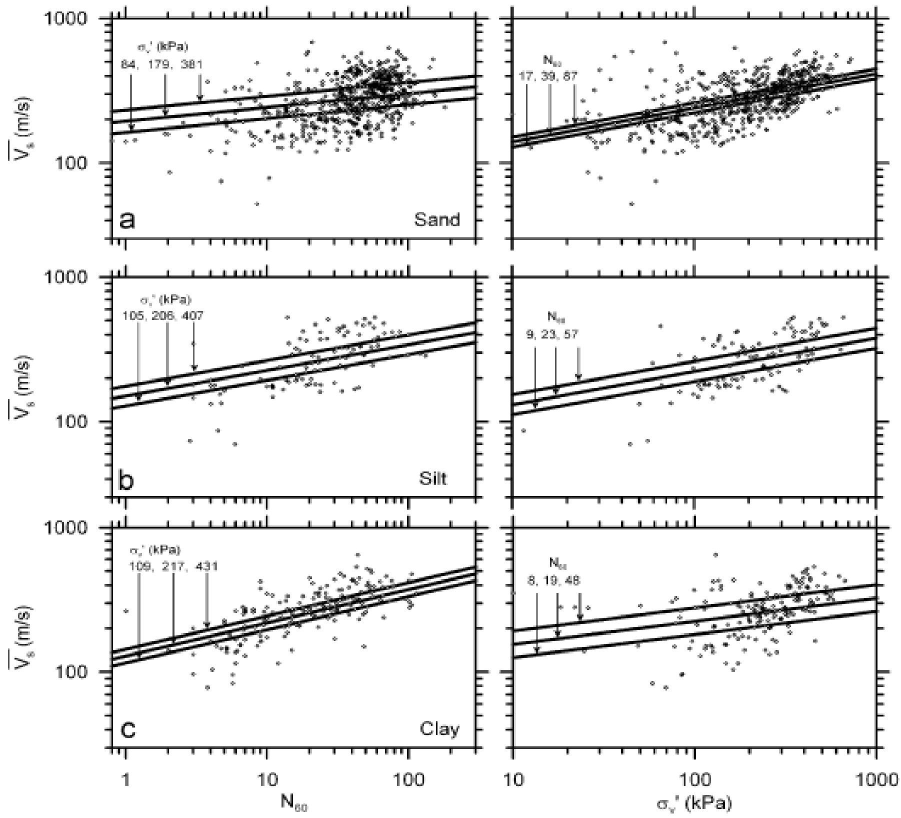


Figure 2.5 Results of regression model for sand, silt and clay with trend lines corresponding to the mean and $\pm 1\sigma$ for N_{60} and σ'_v (Brandenberg et al. 2010)

According to comparison between those trend line function, it can be obtained that the influence of effective stress are highest at the sand fill layer, but this influence becomes weaker for silt and clay soil layer. (Brandenberg et al. 2010)

2.4.3 Intra-boring effect

Based on the equation (2.6), the intra-boring residuals defined in equation 2.7 and illustrated for each soil type in figure 2.6.

$$\varepsilon_{ij} = \ln(\overline{V}_s)_{ij} - \left[\beta_0 + \beta_1 \ln(N_{60})_{ij} + \beta_2 \ln(\sigma'_v)_{ij} + \eta_i\right] \tag{2.7}$$

where N_{60} = SPT blow count; V_s = shear wave velocity, m/s; σ'_v = effective vertical stress, kN/m^2;η_i = random effect for i^{th} boring (i.e. inter-boring effect); ε_{ij} = variation for the j^{th} measurement from i^{th} boring (i.e. intra-boring effect); $\beta_0, \beta_1, \beta_2$ = regression constants; η_i = inter-boring variation; ε_{ij} = intra-boring variation

The mean of the residuals equals to zero, and no trend is obtained in the residuals with either N_{60} and σ'_v, which indicates that there is no bias respect to those input variables and the standard deviation of the intra-boring residuals decreases as vertical effective stress increases, it indicated the model would provide poor relation at low depth. The reason of that correlation decrease at low depth is not clear, but it could be caused by the field test method, such as suspension logging.

The standard deviations of subsample were plotted versus the logarithm of the subsample mean σ'_v values and apply the linear least square to fit based the measurement data. The result are commonly indicates silt and clay depend only weakly on σ'_v. (Brandenberg et al. 2010)

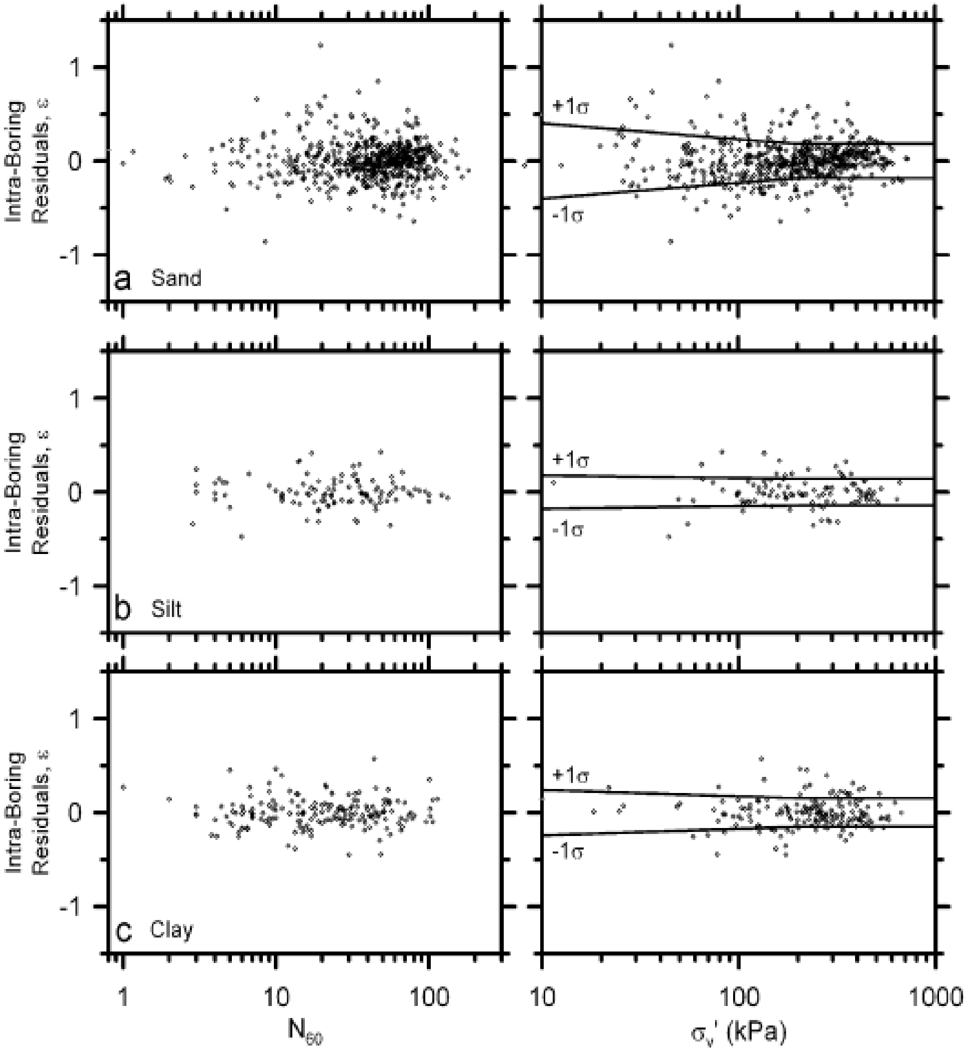


Figure 2.6 Intra-boring residuals versus N_{60} and σ'_v for sand, silt and clay. (Brandenberg et al. 2010)

2.4.4 Inter-boring effect

For evaluate effect of inter-boring, the residuals of inter-boring were plotted as surface geologic epoch function with respect the Holocene, Pleistocene, and Pre-quaternary surface geology epochs. (Knudsen et al, 2009)

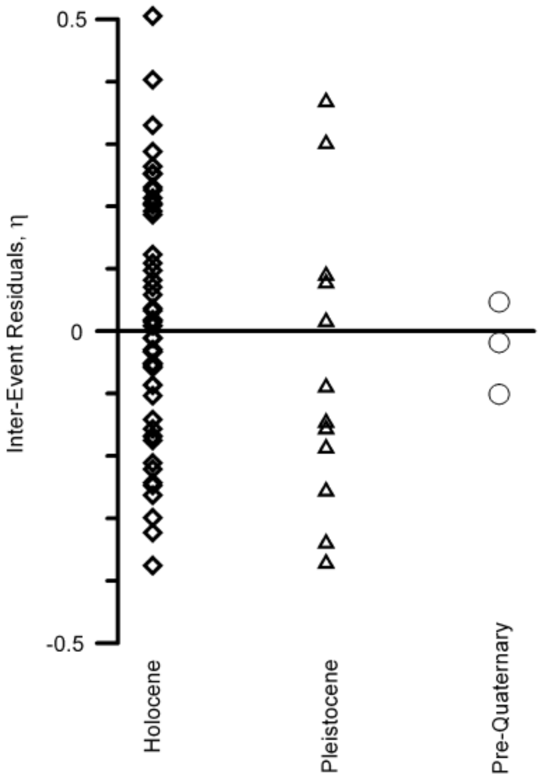


Figure 2.7 Inter-boring residuals as functions of surface geologic epoch. (Brandenberg et al. 2010)

A positive value of inter-boring residual indicates the median value of $\overline{V}_s$ is predicted by equation 2.6 would under-predict the measured value. It can be easily found that a weak trend occurred when the residual of inter-boring decrease with geologic age. It indicates when N_{60} value is given, the corresponding Vs value decreases as age increases. The result indicates little effect of geologic age contained on the correlation between Vs and N_{60}. (Sykora & Koester, 1988)

Since the correlation equation of shear wave velocity has been established based on standard penetration resistance and vertical effective stress, the statistical error from the most of geotechnical literatures are assumed that the effect of effective stress is neglected. However, this kind of assumption or correlation functions cannot be applied commonly or only suitable for rare cases. In order to eliminate this kind of error, it is necessary to establish the new correlation functions for predict the shear wave velocity.

Although the prediction performance had been improved, it is still contain some uncertainly which increased dispersion in ground motion predicted from shear wave velocity. Therefore, it is also suggested that if the directly measurement is available, the proposed correlations should be avoided.

2.5 Influence of Cone Resistance

The relationships between cone resistance (q_c) and shear wave velocity (V_s)were investigated since the early 1980s. These investigations have shown that cone tip resistance, cone sleeve friction, confining stress, depth soil type, and geologic age are factors that influencing the relationship. There are often obtained that the correlation relationship between q_c and V_s were mostly developed for either sand or clays, without intermediate range of soil types.

2.5.1 Prediction performance between shear wave velocity and cone resistance

According to the literature "Prediction of the shear wave velocity Vs from CPT and DMT at research sites", it demonstrated that the non-seismic dilatometer tests (DMT) have more reliable and consistent prediction performance compare with Cone Penetration Test (CPT). The reason is cone resistance parameter scarcely detected the pre-straining or aging of soil structure. (Amoroso 2013)

2.5.2 Bustamante and Gianeselli method (1982)

Since the cone resistance is commonly contained some significant variation in a certain depth. In order to apply this soil parameter to estimate shear wave velocity, Bustamante and Giasenelli method is used in this study.

This method was designed according to 197 pile load (and extraction) tests analysis at a wide area of foundation with each specific soil types.

The equivalent average cone resistance, q_{ca}, at the base of the pile used to compute the pile unit end bearing, q_p, is the mean q_c value measured along two fixed distances a ($a = 1.5D$, where D is the pile diameter) above ($-a$) and below ($+a$) the pile tip. It suggested that in order to calculate q_{ca} in three steps, as shown in Figure.2.8.

1. Calculate mean value of cone resistance between $-a$ and $+a$.
2. For the data between $-a$ to $+a$ which is higher than 1.3 time the mean value of cone resistance, and the data between mid-depth to -a that lower than 0.7 time the mean cone resistance would be eliminated.
3. Calculate the corresponding mean value of cone resistance again to equal equivalent average cone resistance.

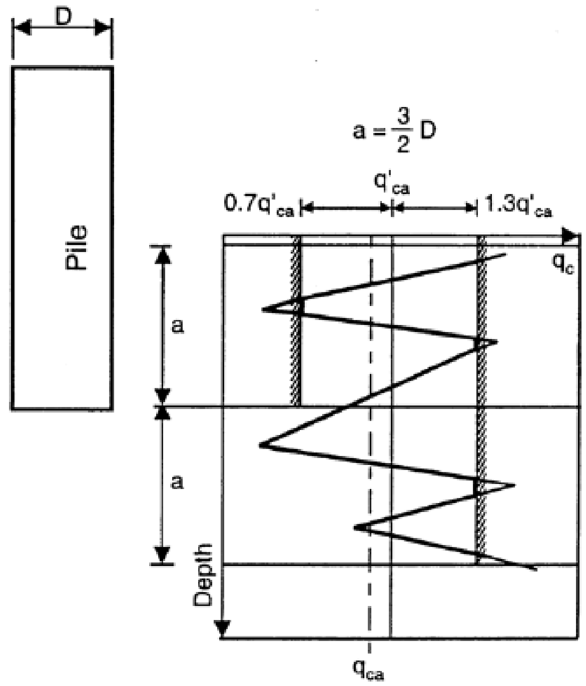


Figure 2.8 Elimninationr equirement of equivalent average cone resistance (Lunne et al. 1997)

2.6 Genetic Programming

Since simple power-law regression analysis commonly applied to establish the correlation between shear wave velocity and SPT-N value based on same research literatures, however, the prediction performance is not good enough so that it cannot be applied to most of sites. After that, a new statistical correlation approach was published to establish those two measured parameters, which is Genetic Programming computing algorithm. It can available to achieve more accuracy correlation model.

Regarding to various literatures about the correlation between shear wave velocity and SPT-N values, it also have different approaches and suggestion to analysis the correlation relations and it can be generally summarizes to two reasons, one is the correlations between shear wave velocity and SPT-N values have considerable dispersals. This may be due to various measurement methods used to measure the shear wave velocity and SPT-N value, geotechnical and geological conditions and so on. Another reason is the low quality of prediction performance for this type of regression analysis to establish the correlation relation between those two parameters.

2.6.1 General introduction and setting of Genetic Programming

Genetic Programming (GP) is one of advanced computing approach which can suitably be used for pattern recognition purposes. Some of the parameters are necessary to define first before run to computing software program.

Population of GP is represented by parse trees, including non-linear individual entities of different sizes and shapes. Individuals are structures consisting of functions and terminals, which both are selected out from larger set of functions and terminals. Each function contains basic arithmetic operators (plus, minus, cross, and so on.) and Boolean logic functions (AND, NOT, and so on.) and also user defined functions. However, terminals contain constant parameters that form the whole individual structure together with functions. In each generation, these functions and terminals are selected randomly and can form many individuals with different fitness values. The fitness value strongly depends on the definition of fitness function, which is the most influential element in GP. In the current study, the fitness function is root mean square error. At each step, by selecting the best individuals and then breeding them together using GP operations, such as cross-over and mutation, a new generation will be created. (Searson 2009)

The procedure could be continued until the predefined fitness value was obtained or reached the specified number of generations. Therefore, the best individual will be proposed or specified number of generations was reached. At the end, the best individual will be proposed as the best model between inputs and output data. The GPTIPs 1.0 toolbox for MATLAB software was used in this literature analysis. (Searson 2009)

2.6.2 Evaluation between Simple Regression and Genetic Programming

According to the literature "A New Statistical Correlation between Shear Wave Velocity and Penetration Resistance of Soils Using Genetic Programming", it provided a database which contained 613 pairs of shear wave velocity and SPT blow count, which measured in Iran. The data of SPT-N was mainly measured by down-hole method, and also seismic refraction and Spectral Analysis of Surface Waves method (SASW). These measurement data are carried out by different soil types, such as sand, silt and clay.

In this literature, they applied the simple power-law regression equation which shown in Equation. The corresponding correlation model can be obtained in equation 2.8.

$$\mathrm{V_s} = 129.4\mathrm{N}^{0.336} \tag{2.8}$$

where $\mathrm{V_s}$ = shear wave velocity, m/s; N = uncorrected SPT blow count

For Genetic Programming method, it commonly separate 80% of all data were set in to training sub-set and remaining part of data were included into testing sub-set , so that the maximum, minimum, mean and standard deviation of Vs and SPT-N in two sub-sets were as equal to each other as possible.

In order to evaluate the prediction performance of the correlation models, coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), which the corresponding calculation equations are:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(m_i - p_i)^2}{\sum_{i=1}^{n}(m_i - \bar{m})^2} \tag{2.9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(m_i - p_i)^2}{n}} \tag{2.10}$$

$$MAE = \frac{\sum_{i=1}^{n}|m_i - p_i|}{n} \tag{2.11}$$

where m = the values of measured Vs in the field; p = predicted Vs by the proposed model; $\bar{m}$ = mean measured value; n = the number of data presented in database

After that, the best fitting function was selected by comparing the values of R^2, RMSE, and MAE, which is expressed as following equation form:

$$V_s = 136.581 + \frac{0.0576N}{A} + \frac{10.88527N}{A+2.91052} + \frac{1.35276}{A+1.18224} + \frac{6.39661}{A-0.4613} - \frac{1.4529}{A+0.17753} - \frac{2.9866}{A+1.03558} \tag{2.12}$$

where $A = 0.01918N - 2.3876$; $V_s = \text{shear wave velocity}, m/s$; $N =$ uncorrected SPT blow count;

The corresponding prediction performance of the simple regression correlation model and Genetic Programming model summarized based on the value of the R^2,RMSE, and MAE, which also shown in table 2.2.

Table 2.2 Summarization of prediction performance between GP based model and simple regression analysis based model (Barkhordari et al. 2013)

Parameters	GP based model		Simple regression analysis based model		Improvement Ratio (%)
	training	testing	all data	all data	
R^2	0.49	0.39	0.47	0.25	88
RMSE (m/sec)	171.1	185.3	174.0	207.0	15.9
MAE (m/sec)	123.7	136.0	126.1	147.5	14.5

According to table 2.2, l a significant better performance was created by GP based model compare with simple regression analysis based model. In that case, it should be mention that the use of soft computing methods, such as evolutionary computing and artificial neural network can considerably improve the accuracy of Vs estimation using penetration resistance of soils.

CHAPTER 3 FIELD TESTING

3.1 Introduction

Macau is a Special Administrative Region of China, which is situated on the southeast coast of China and the west shore of Pearl River. There are three parts of Macau area, namely the Macau peninsula, the Taipa, and the Coloane Islands. The total amount of area equals to 30.3 km^2 which contains a large portion of the region is from land reclamation. The soil profile of the reclamation area is a layer of fill covering the marine deposit, followed by alluvium of alternating sand and silty clay which overlies the completely decomposed granite and the bedrock.

Due to the rapid economic growth, there are many infrastructures under construction in Macau, such as Hong Kong Zhuhai Macau Bridge project, Macau public housing and LRT, etc. Although there had many tests performed on the engineering properties of the Macau soil, but it can only refer to the static loading condition. It is not enough since the reclamation area is usually due with very large dynamic soil behavior, such as deformation and settlement. Therefore, the investigation of the dynamic soil behavior is important.

Shear wave velocity is a soil parameter commonly used for study of the dynamic behavior research. Although it preferable to determine shear wave velocity directly by field test, however, it deal with very high cost, space constraints and lacked in workers with official qualification and skill to do the shear wave velocity field test. On the other hand, SPT-N is significant soil parameter along with shear wave velocity. These two parameters are also widely used to describe in soil characteristic. Therefore, the determination of shear wave velocity by using the correlation with SPT-N value is one of the important studies in geotechnical engineering research. In order to improve from the previous correlations, vertical effective stress and cone resistance are also be considered.

Based on the database information which provided by the same research of Final Year Project, the field tests which performed in this research included down-hole seismic test and cross-hole seismic test, standard penetration test and cone penetration test. The detail description of each test is discussed in the following.

3.2 Standard Penetration Test

Standard Penetration Test is a field testing commonly used in worldwide for indicates the density and compressibility of granular soils. It is can be used to check the consistency of stiff or stony cohesive soils and weak rocks. On the other hand, it can also be applied to estimate the liquefaction potential of saturated granular soils which relative to earthquake design and decision of foundation design for both shallow and deep foundation.

According to the ASTM D1586, the test procedures of SPT are described in the following, and the visually test performance is shown in figure 3.1.

- Rest the split spoon on the bottom of the borehole.
- Drive the split spoon with the 140-lb. hammer falling 30 (760 mm) in and count the number of blows applied in each increment until either a total of 50 blows in any one increment, or a total of 100 blows is reached, or there is no observable penetration, or the full penetration of 18 in (450 mm) is achieved.
- Repeat the previous two steps for every 5 ft (1.5 m), homogeneous strata reached, and at every stratum change.
- Measure the blow courts for each 6 in (150 mm) penetration. If a full drive is achieved, add the blows for the last 12 in (300 mm) of penetration to measure

SPT-N. On the contrary, the blow courts record for each measurement or part increment (precision as 1 in) should record on the boring log.

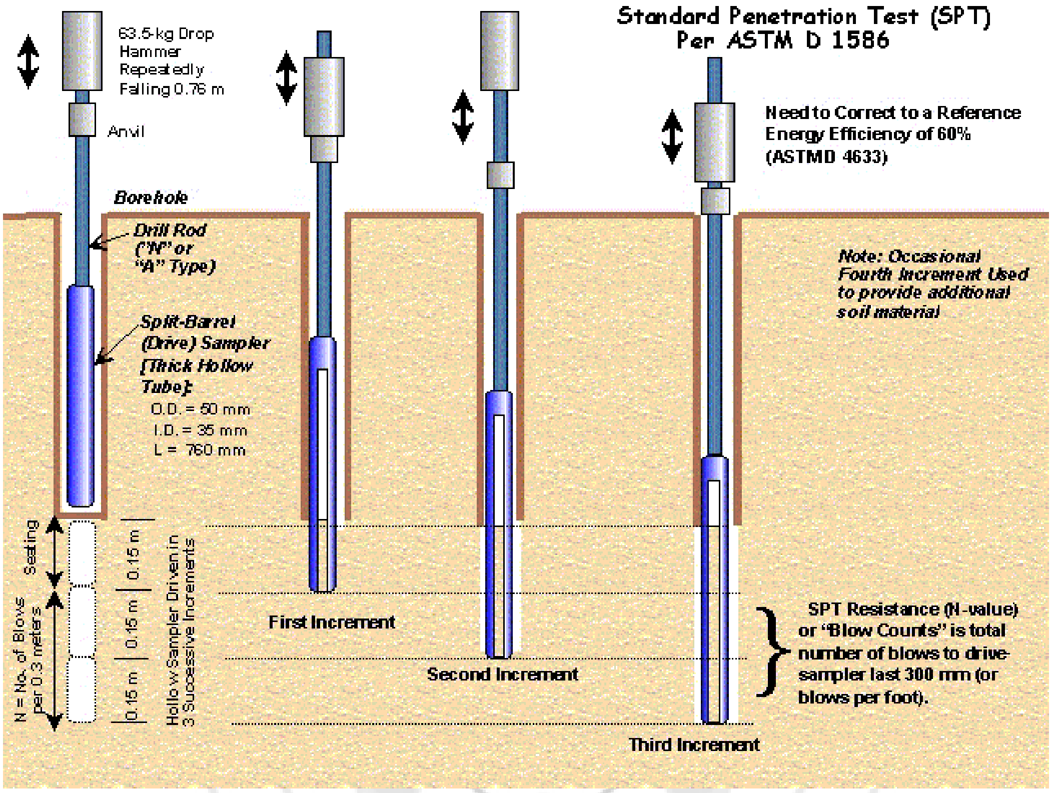


Figure 3.1 Standard Penetration Test (ASTM D 1586)

3.3Cone Penetration Test

Cone Penetration Test is commonly used for the situ investigation of soil for engineering purposes. In this test, a cone on the end of rods is pushed into the ground constantly and continuous measure the resistance to penetration of the cone defined as cone resistance, q_c, and of a surface sleeve defined as local side sleeve friction, f_s. There are several of size and shape of penetrometers being used for corresponding site investigation. The standard size of the penetrometers for most countries is the cone with an apex angle of 60° and a base area of 10 cm^2 with a friction sleeve having an area of 150 cm^2, which the shape of cone is shown in figure 3.2. For mechanical cone, it can be advanced separately by means of sounding rods pushed vertically into the soil at a constant rate of 2 cm/sec. Initially, the cone is pushed through a distance of 5 cm to measure q_c, it can also measure both q_c and fs with further advance type of cone. For the electric cone type, it can be measured independently and continuously with penetration by means of load cells installed the body of the probe. It could also measure pore water pressure with the filter element placed close behind the cone. The result of cone resistance can be used as input parameter and applied into the corresponding correlation equation to obtain the soil parameters, such as bearing capacity, Young's modulus of elasticity, compression index, etc.

By comparing to the other field test techniques, CPT has some advantages for the purpose of soil investigation which include to the following:

- The test equipment can be easily and quickly mobilized to the site.
- provide information on soils under undisturbed or natural conditions
- provides a continuous record of data measurement for investigated soil depth
- provides repeatable and reliable data which not depend on operator
- Lack of theoretical support for CPT data interpretation.

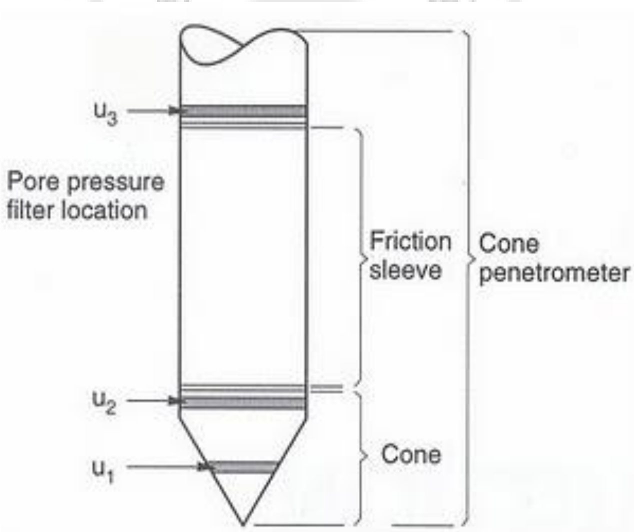


Figure 3.2 Cone penetration tip

3.4 Measurement method of Shear Wave Velocity

3.4.1 Seismic Test

Since the dynamic geotechnical properties of soil are very important for the site investigation purpose or decision of foundation design, the corresponding field test are carried out in the field need to be high accuracy. Therefore, seismic test is commonly applied for the measurement of shear wave velocity, which mainly included Down-hole seismic test and Cross-hole seismic test.

The actual parameters that directly measured in seismic test are Compression wave (P-wave) and Shear wave (S-waves). Compression wave can travel through solid and fluid but often limited to unsaturated soils. Shear wave can travel through soil structure but not for solid and fluid, it also used to determine the elastic shear modulus in low strain level.

Although there are various types of field test are capable for the measurement of low strain shear wave velocity profile, however, the considerable reliability can be obtained from the seismic test. The detail information for these two types of seismic tests are introduced in section 3.4.2 and 3.4.3 respectively.

3.4.2 Down Hole Seismic Test (DS)

The objective of seismic down-hole tests is to measure travel times of P and S-waves from the energy source to the receiver. These tests can be performed in single borehole. A hole is drilled to the required depth at the testing site and a vibrating source is created to determine shear wave velocity for various soil layers. In this case, the waves travelling in vertical direction either down or up depending on the location of the source of impulse.

In down-hole seismic test, the generated waves propagate through the soil layers in downward direction. A single wave source is located on the ground surface adjacent to the borehole. In the edge of the borehole, either a string of multiple receivers at known depths is fixed or a single receiver is moved to different depths as the test advances. All receivers are connected to high speed recording system Freedom Data PC and the output is measured as a function of time. The configuration of the Down hole Seismic Test as shown in figure 3.3.

3.4.3 Cross Hole Seismic Test

The simplest Cross-hole test consists of 3or more boreholes with 3 meter spacing, one for a P-SV source and another for a receiver. By fixing both the source and the receiver at the same depth, the wave velocity through soil/rock material between the holes is measured for the depth. By testing at various depths a velocity profile against depth can be obtained.

The boreholes are typically 76mm to 100mm in diameter cased with PVC casing. Formulated mixture grouting around the casing should be performed to fill up the spacing between casing and soil. All cables should be connected and checked properly by running the checking function of system controlling program Win-GEO before putting down the P-SV source and triaxial geophones into boreholes.

The source and geophones are clamped firmly into place at decided depth with the rope cleat, and then air is pumped to the bladders attached at the source and the geophones, The impulse source is activated and both captured signals from geophones are displayed simultaneously on the Freedom Data PC. The signal amplitude and duration should be adjusted such that shear wave traces are displayed entirely on the monitor, same procedure is repeated at every 1.5 m interval (in some cases of this project is reduced to 1.0 m interval) until the final depth. The configuration of the cross hole Seismic Test as shown in figure 3.4.

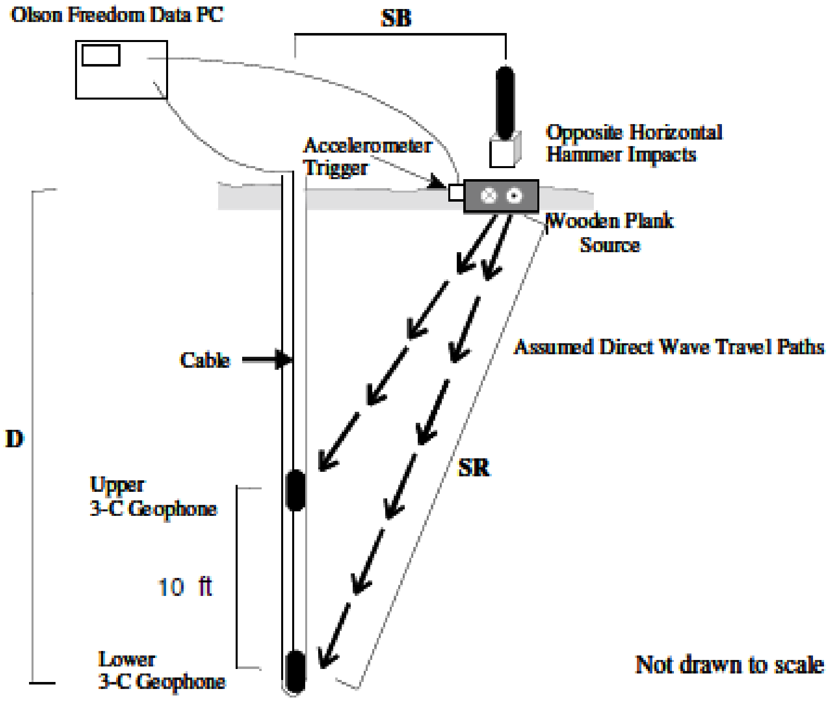


Figure 3.3 Downhole Seismic Test

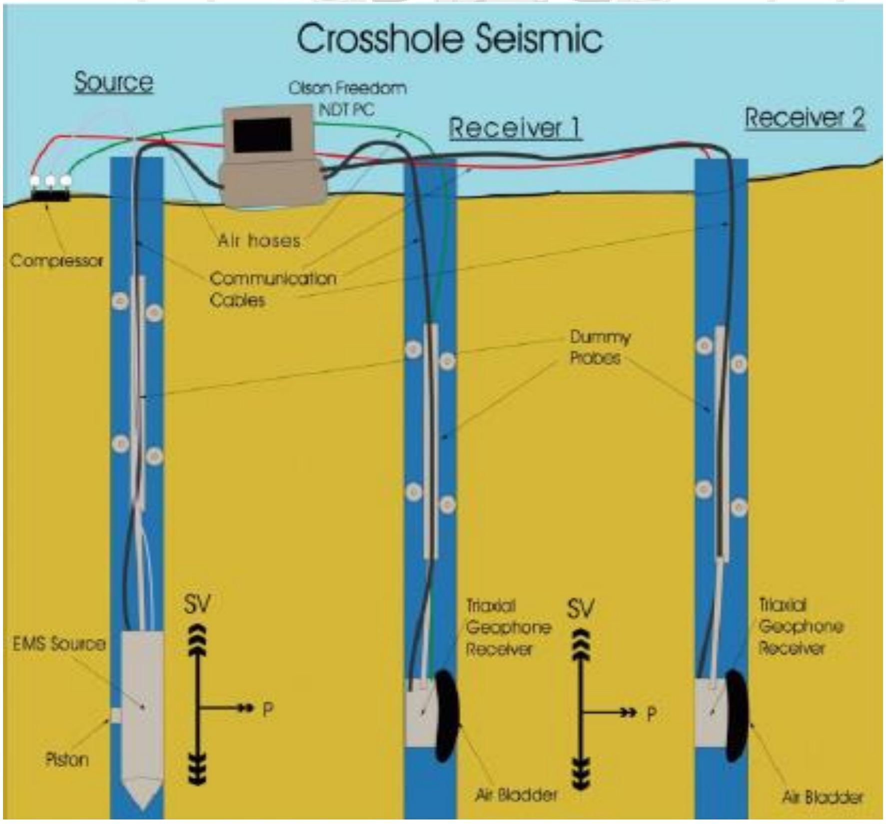


Figure 3.4 Cross hole Seismic Test

3.4.4 Seismic Cone penetration test

The seismic cone penetration test are used to measures the shear wave velocity. It allows the small strain shear modulus (G_0) and the constrained modulus (M_0) to be evaluated. The small strain shear modulus is an important parameter for reduction of ground surface motions from earthquake excitation, evaluation vibrating of foundations, offshore structures behaviour due to wave loading, and the prediction of deformations at surrounding area of excavations.

Seismic cone penetrometer is a memory oscilloscope and an impulse source with a trigger for the oscilloscope. The source can consist of a steel beam for the generation of shear (*S*) wave or a flat plate for the generation of compression (*P)* wave.

The shear wave source usually perpendicular to cone and pressed against to ground by the weight of the hammer and CPT vehicle and penetration is stopped for each 1 m intervals. During the rest in penetration process, shear waves is generated at the ground surface and send to reach the seismometer and the corresponding travel time required for the shear wave is measured. The shear wave is generated by using hammer to hit the beam end horizontally. The computer in the CPT rig collects and processes all the data from the CPT or CPTU.

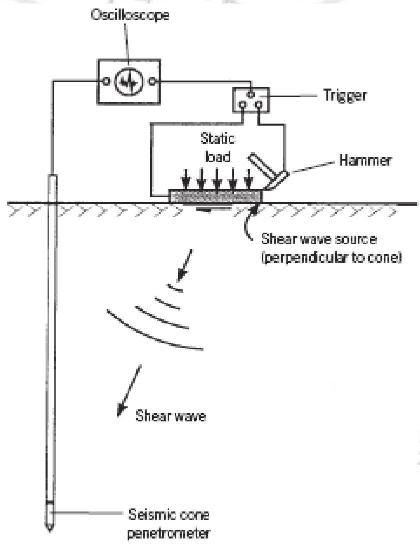


Figure 3.5 Seismic Cone Penetration Test

CHAPTER 4 RESULT OF PREVIOUS STUDY

4.1 General

The establishment researches of the correlation between shear wave velocity and standard penetration blow count for Macau soil have been working in recent year, the databases which obtained from SaeTeng (2009), Pun Hou Kun and Chao Sai Choi (2010), and Kuan Wai Kin (2011) are selected.

For SaeTeng (2009), there are five site investigation project databases were carried, which are University of Macau, Ocean World, Chun Su Mei, Oriental Golf Course and Vehicle Detention Centre and Rua de Seng Tou. Cross-hole seismic test and down-hole seismic test were applied at those site locations to measure the shear wave velocity of the soil with corresponding depths, in order to establish the correlations equation with SPT-N value.

For Pun (2010), the location for site measurement of the shear wave velocity is mainly was the site of Rua de Seng Tou, Taipa. Cross-hole seismic test, down-hole seismic test and cross-hole seismic SPT (CS-SPT) were performed to measure the shear wave velocity. The correlation relation between shear wave velocity and SPT-N values were established based on the database for Rua de Seng Tou only. Moreover, the combined correlation relations between the Pun (2010) and SaeTeng (2009) was established. The corresponding test location for Pun (2010) and SaeTeng (2009) is shown in figure 4.1.

Figure 4.1 all test location for Pun (2010) and SaeTeng (2009)

For the Kuan Wai Kin (2011), the corresponding site project investigation for establish the new correlation relations between shear wave velocity and SPT-N values was provided from the LRT C-250 project. There are five boreholes were contained in this project, the corresponding locations were shown in figure. 4.2.

Figure 4.2 the corresponding locations of five boreholes for LRT C-250 project

The depths of five boreholes are around 18m to 66m. Down-hole seismic test and seismic-cone penetration test (S-CPT) were performed for the measurement of shear wave velocity measurement. The new correlation relation between shear wave velocity and SPT-N values are also obtained based on the LRT C-250 project only. It also combined with previous two correlations relation results to establish the new correlation relation in order to fit in those three project cases.

4.2 Study of SaeTeng (2009)

According to the research result from SaeTeng (2009), it contains 102 data points which obtained based on the representative strata profiles and dynamic properties. There are 3 correlation equations between shear wave velocity and SPT-N value based on the soil type, which mathematic models express in 4.1 to 4.3.

$$V_s = 82.14N^{0.337} \quad \text{(all soil with } 0 < N \leqq 100) \qquad (4.1)$$

$$V_s = 65.95N^{0.4} \quad \text{(granular soil with } N \geqq 10) \qquad (4.2)$$

$$V_s = 71.12N^{0.522} \quad \text{(clayey soil with } N \leqq 10) \qquad (4.3)$$

where V_s = shear wave velocity, m/s; N = standard penetration test N value

Notes that the uncorrected blow-counts applied in these correlations for the initial result exhibits the energy ratio of SPT hammer in Macau is around 80 to 90 %.

In order to evaluate the correlation models visually, the corresponding graphical models with log scale express in figure 4.3 to 4.5 respect to those 3 models, and the coefficients of determination (R^2) also determined to evaluate the prediction performance.

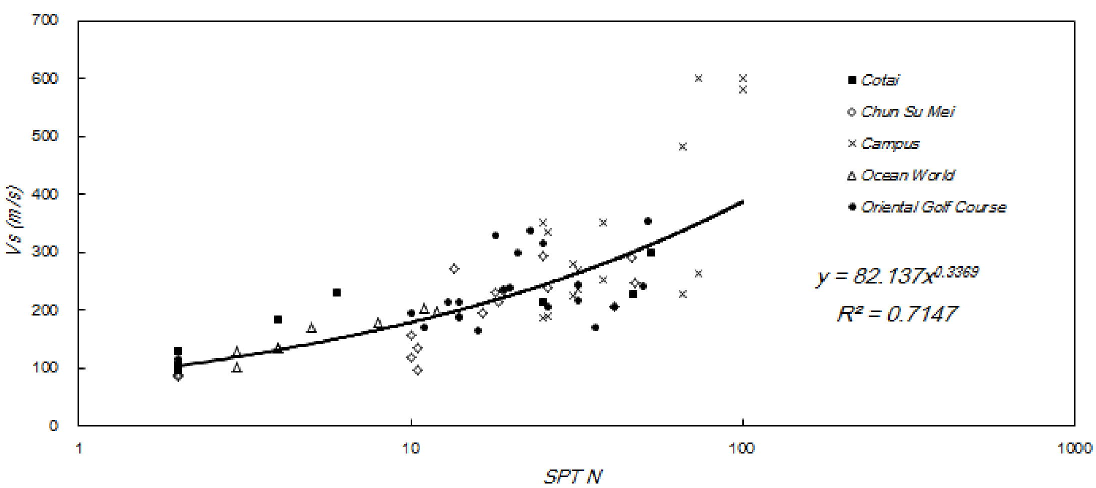


Figure 4.3 Correlation between Vs and SPT-N value for all soils (equation 4.1)

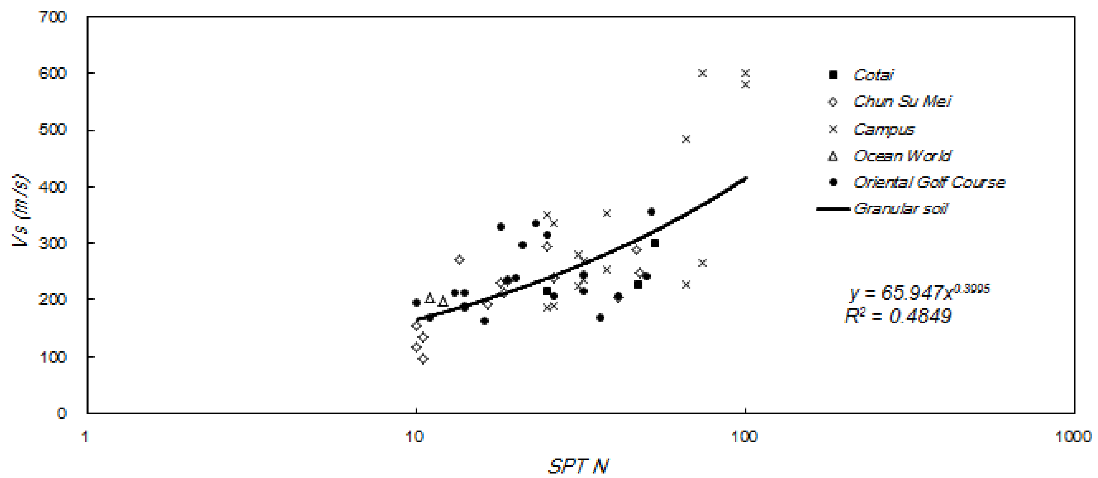


Figure 4.4 Correlation between Vs and SPT-N value for granular soil (equation 4.2)

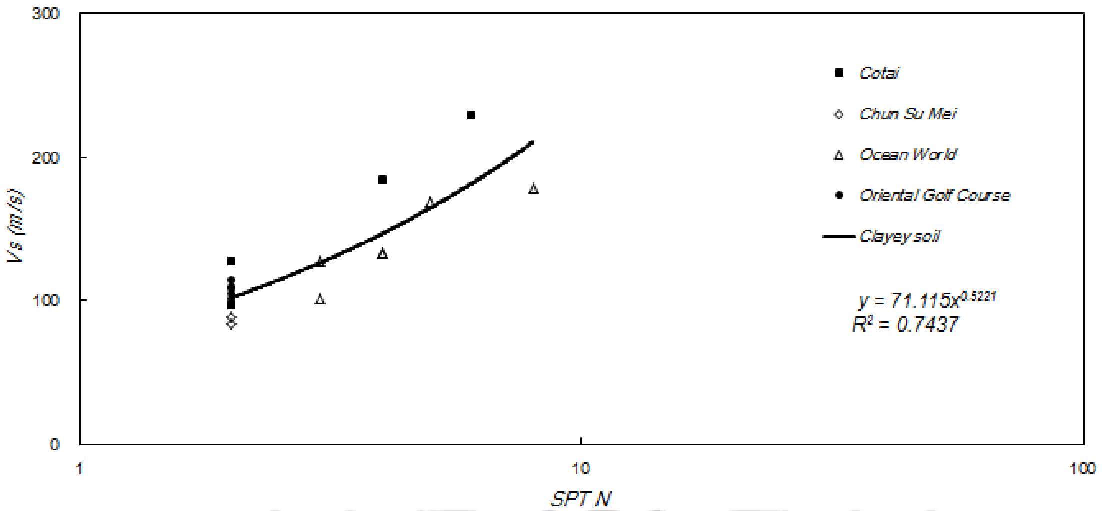


Figure 4.5 Correlation between Vs and SPT-N value for clayey soil (equation 4.3)

Based on the observation from those three different correlation models, it can be demonstrated that the result are compromised with the inherent disadvantage of SPT method where it does not estimate well for clayey soils. In contrast, the better prediction of Vs is obtained for granular soils can be reached by elimination of clayey factor.

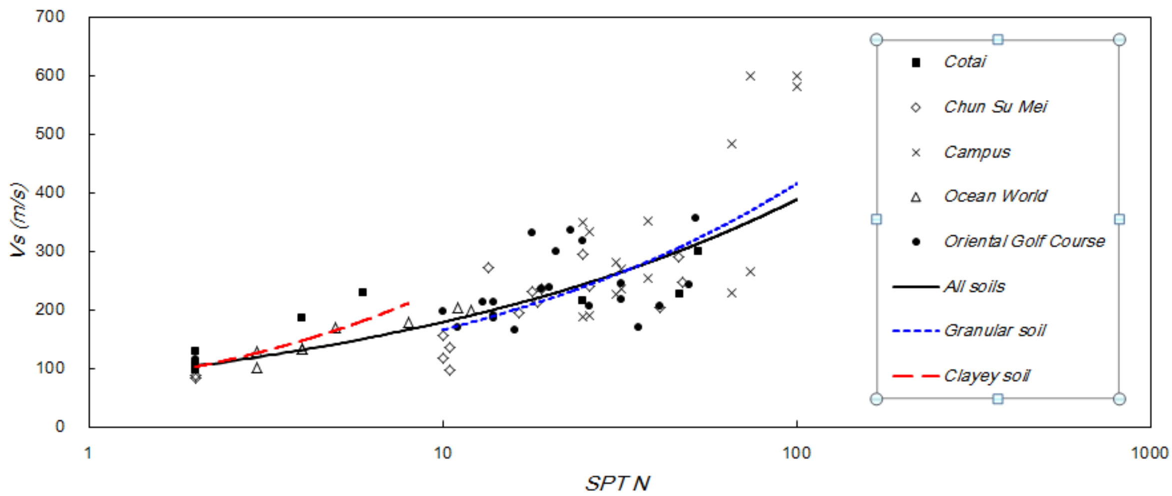


Figure 4.6 Regression lines of all soils, granular soil and clayey soil

After the three different models were plotted in figure 4.6, it can be obtained that the difference between the regression models of those three is very small. Therefore, it preferred to applied regression models of all soils to estimate Vs without concern the miscellaneous process in ground response analysis.

4.3 Study of Pun and Chao (2010)

According to the research result from Pun and Chao (2010), it contains 48 data points based on the investigation at Rua de Seng Tou, Taipa and calibrated with the correlation equations of SaeTeng (2009). The correlation model between shear wave velocity and SPT-N based on the dataset from Pun and Chao is shown in equation 4.4. The same correlation model combined with SaeTeng (2009) is shown in equation 4.5. Moreover, the graphical expression (with log scale of SPT-N value) for equation 4.4 is shown in figure 4.7. It also made the comparison between SaeTeng (2009) with Pun and Chao (2010) in figure 4.8, and the comparison between SaeTeng (2009) and combination between SaeTeng (2009) and Pun and Chao (2010).

$$V_s = 114N^{0.214} \tag{4.4}$$

$$V_s = 94.7N^{0.296} \tag{4.5}$$

where V_s = shear wave velocity, m/s; N = standard penetration test N value

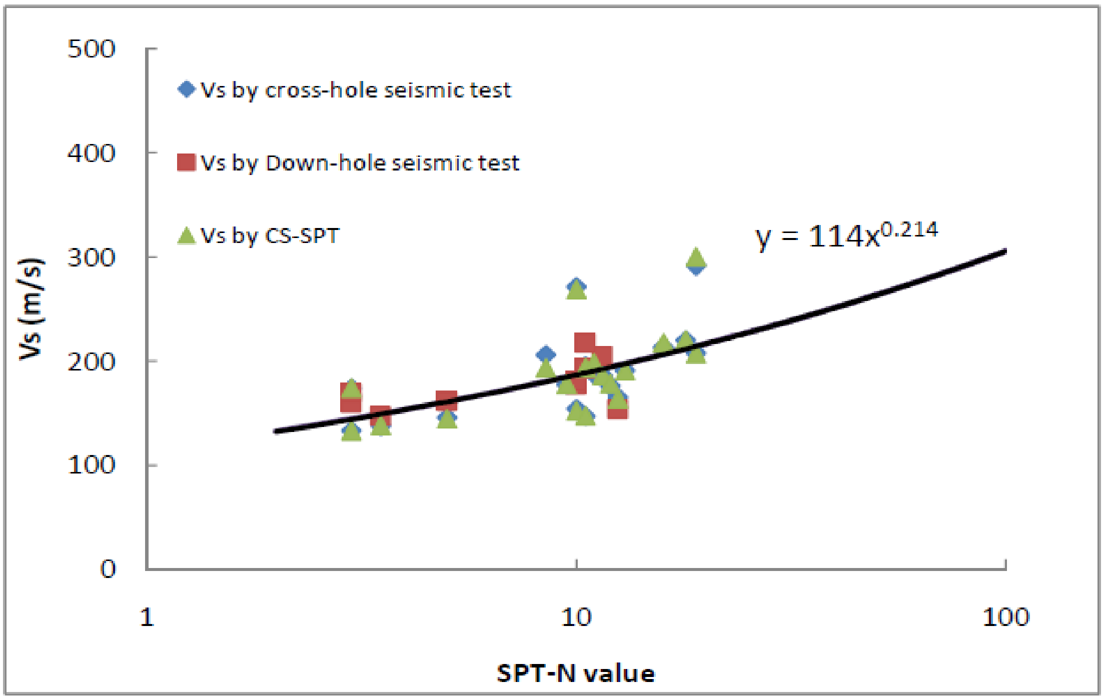


Figure 4.7 Correlation model for Pun and Chao (2010)

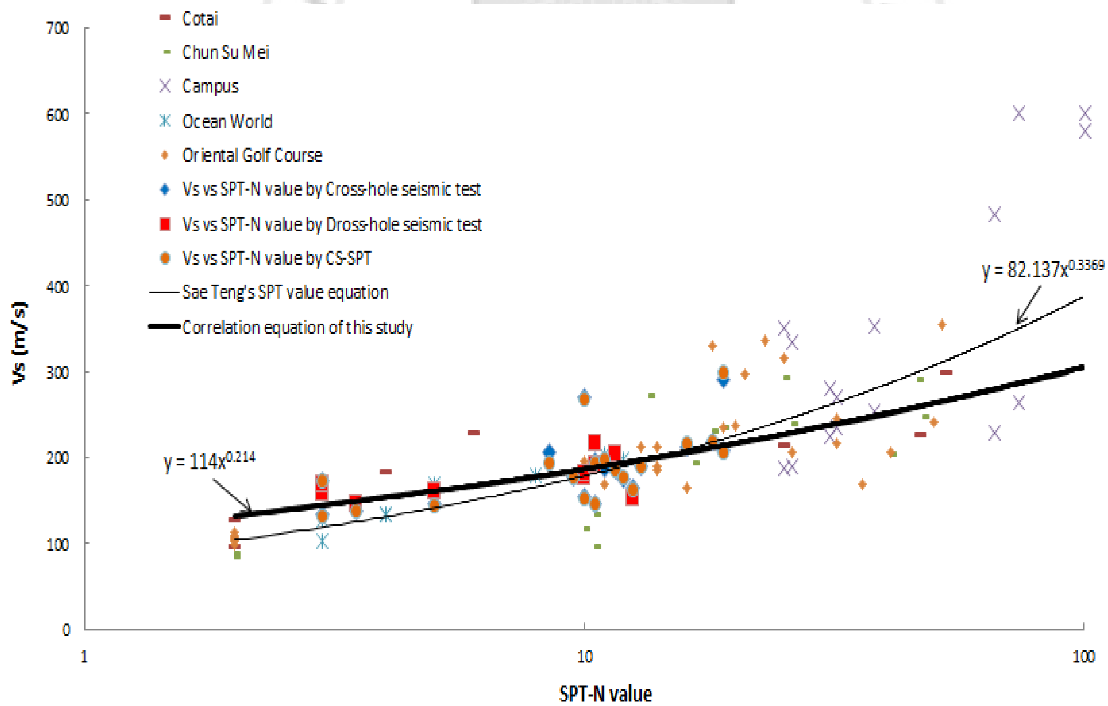


Figure 4.8 Comparisons between SaeTeng (2009) and Pun and Chao (2010)

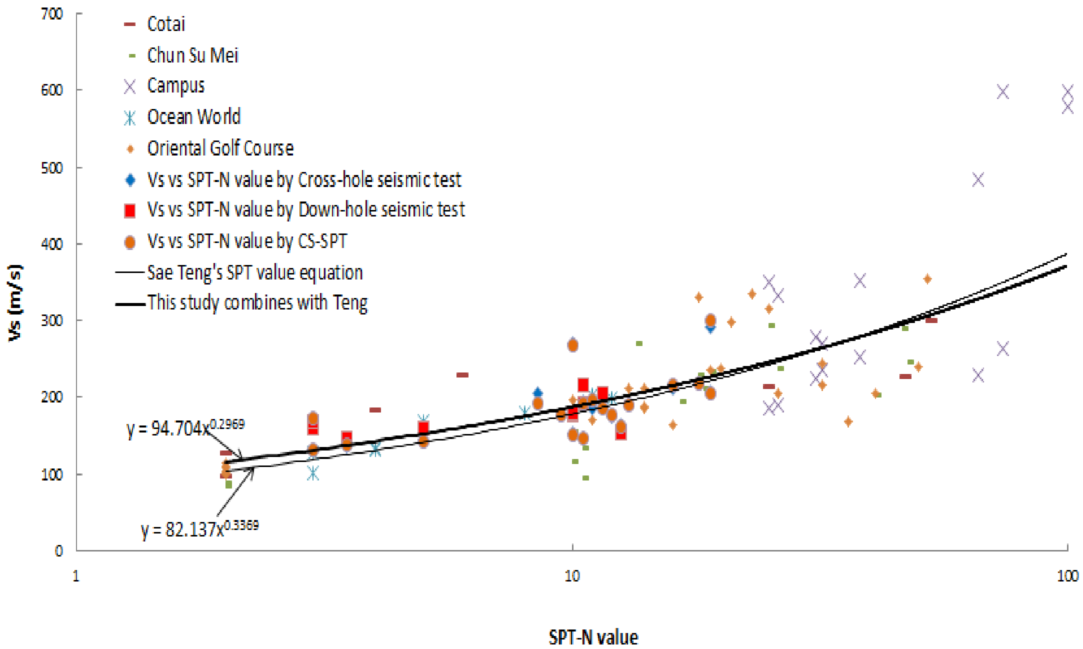


Figure 4.9 Comparison between SaeTeng (2009) and combination between SaeTeng (2009) and Pun and Chao (2010)

Based the comparison in figure 4.8, it can be obtained that the correlation model for Pun and Chao (2010) determine the larger value of Vs when SPT-N less than 100, and determine smaller value of Vs when SPT-N bigger than 100.

For the comparison in figure 4.9, although the database from the SaeTeng (2009) and Pun and Chao (2010) were combined to examine the new correlation model, however, the different is not significant larger by comparing with correlation model for SaeTeng (2009).

4.4 Study of Kuan WK (2011)

For Kuan WK (2011), there are two different correlation models were applied based on the database of LRT C-250 project.

4.4.1 CPT-qc model

It first applied the data of cone resistance qc and Vs to establish the corresponding prediction performance. The establishment of the database is based on the following assumption:

- a corresponding average qc was calculated for the same depth interval (below 1 meter) of the Vs record;
- each pair of q_c and V_s was referred to the mid-depth of the corresponding measurement;
- the specific gravity for the clay layer is assumed to be 2.75;
- The prediction is not made at those depths where they types of soil is unknown;
- The prediction is not made at those intervals where the value of q_c is not completely measured.

For modification method of the first correlation model in Kuan WK (2011), the correlation models between q_c and Maximum Shear Modulus would be applied. The selection of correlation model would be different depends on the soil types.

$$G_{max} = 1634(q_c)^{0.250}(\sigma'_v)^{0.375} \quad \text{(sand)} \qquad (4.6)$$

$$G_{max} = 406(q_c)^{0.695}(e)^{-1.130} \quad \text{(clay)} \qquad (4.7)$$

where G_{max} = maximum shear modulus, kPa; q_c = cone resistance, kPa; σ'_v = effective stress, kPa; e = void ratio

. In order to calculate the Maximum Shear Modulus based on those two equations, the corresponding effective stress and void ratio are necessary to calculate.

$$\sigma'_v = \sigma_v - u \qquad (4.8)$$

$$e = \frac{G_s\gamma_w - \gamma_{sat}}{\gamma_{sat} - \gamma_w} \qquad (4.9)$$

where σ'_v = effective vertical stress, kPa; σ_v = total vertical stress, kPa; G_s = specific gravity of soil; γ_{sat} = saturated unit weight of soil, kN/m^3; γ_w = unit weight of water, 9.81 kN/m^3

After the Maximum Shear Modulus determined based on the equations that provided, the corresponding shear wave velocity can also be determined.

$$V_s = \sqrt{\frac{G_{max}}{\rho}} \tag{4.10}$$

where V_s = shear wave velocity, m/s; G_{max} = Maximum Shear Modulus, Pa; ρ = density of soil, kN/m^3

In order to evaluate the prediction performance, the comparison result with respect to actual measured Vs were shown.

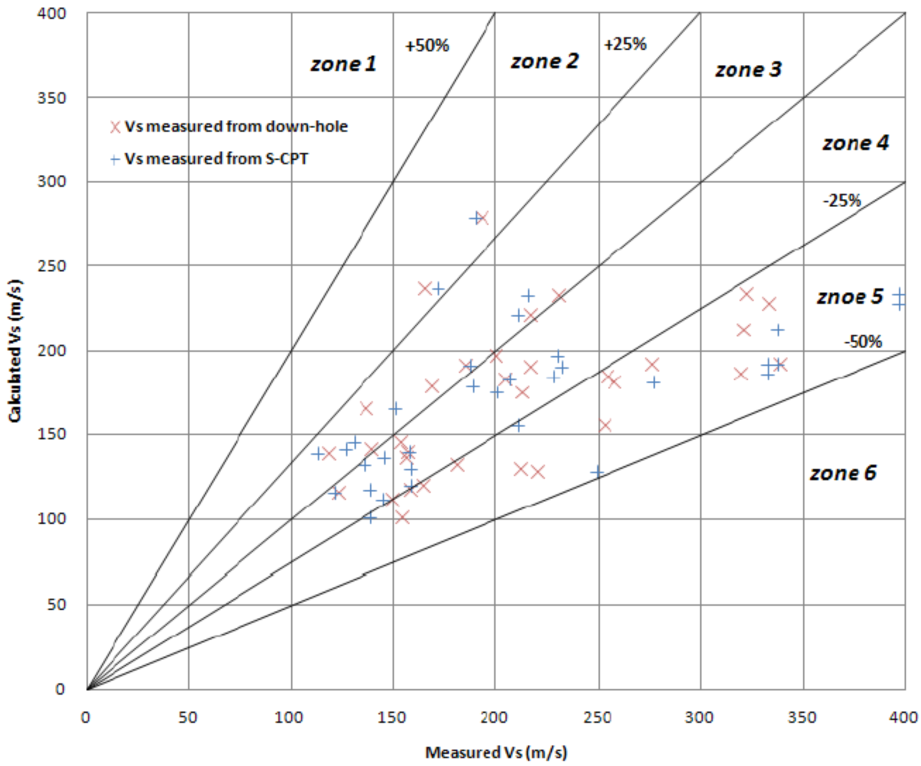


Figure 4.10 Comparison of calculate Vs (CPT-qc model) with respect to measured Vs

Table 4.1 Data distribution for each zone (CPT-qc model)

Zone	Number of data fall in	Percentage of data fall in (%)
1	0	0
2	4	6.06
3	14	21.21
4	22	33.33
5	26	39.40
6	0	0
Percentage of over-predicted, %	27.27	
Percentage of under-predicted, %	72.73	
Percentage of $\pm 25\%$	54.54	

4.4.2 SPT-N model

For SPT-N model, there are two correlation model were examined by power-law regression model, first is based on the database which obtained from LRT C250 (Kuan 2011) project. Second is based on the database which combined by SaeTeng (2009), Pun and Chao (2010), and Kuan (2011). These two correlations model are expressed in equation 4.11 and 4.12, respectively.

$$V_s = 139.4N^{0.216} \quad (4.11)$$

$$V_s = 118.1N^{0.229} \quad (4.12)$$

where V_s = shear wave velocity, m/s; N = standard penetration test N value

In order to evaluate those two SPT-N models, the measured database from the LRT C-250 project would be acting as Validation data set to compare with the calculated Vs which predicted from those two correlation models. The corresponding prediction evaluation result is shown in below.

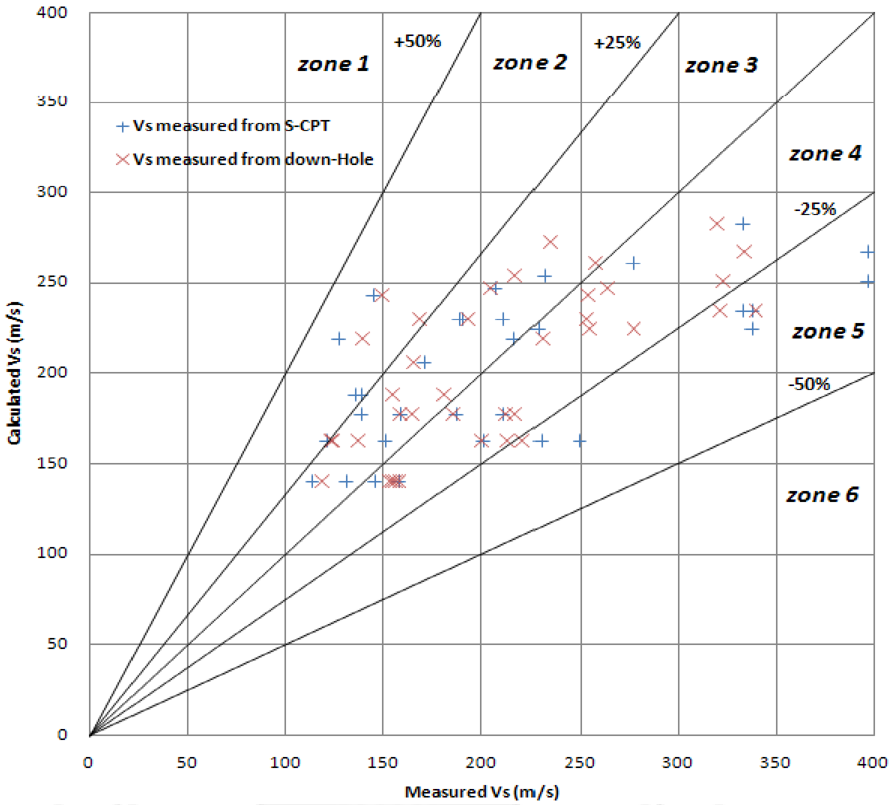


Figure 4.11 Comparison of calculated Vs based on the correlation equation 4.11 with measured Vs based on LRT C250

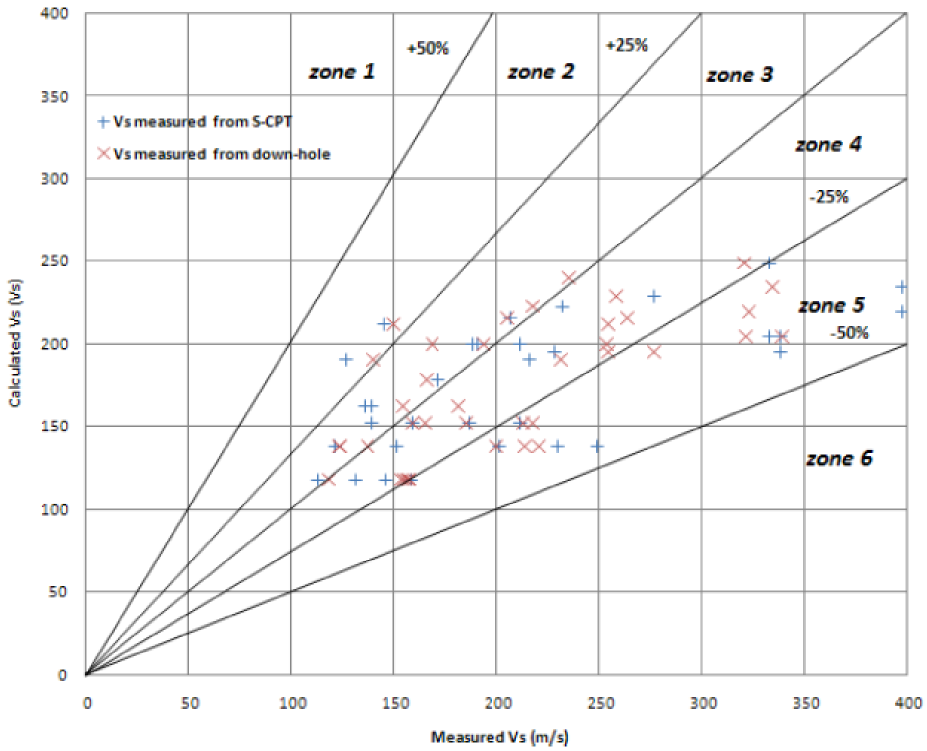


Figure 4.12 Comparison of calculated Vs based on the correlation equation 4.12 with measured Vs based on LRT C250

Table 4.2 Data distribution for each zone (equation 4.11 and 4.12)

	Equation 4.11		Equation 4.12	
Zone	Number of data fall in	Percentage of data fall in (%)	Number of data fall in	Percentage of data fall in (%)
1	0	0	0	0
2	8	11.27	4	5.63
3	27	38.03	19	26.76
4	26	36.62	26	36.62
5	10	14.08	28	30.99
6	0	0	0	0
Percentage of over-predicted, %	49.3		32.39	
Percentage of under-predicted, %	50.7		67.61	
Percentage of $\pm 25\%$	74.65		63.38	

Based on the prediction performance for those three models respect to LRT C-250 project, the correlation equation 4.11 would provide the best prediction performance, simply because the correlation model is based on the same dataset. For the comparison between the CPT-qc model and SPT-N model (equation 4.12), the SPT-N model would perform the better prediction performance compare with the CPT-qc model. The prediction performance for each model can also be evaluated by following indexes, as shown in table 4.3.

Table 4.3 Evaluation index of comparison of proposed models

Model Type	Sum of square error	Data points	RMSE	R^2
CPT-qc	329258.16	66	71.73	0.328
SPT-N (4.11)	202788.50	71	53.82	0.448
SPT-N (4.12)	285022.70	71	40.83	0.446

CHAPTER 5 METHODOLOGY

5.1 Current Database

According to the database of LRT-C250 which was concluded by previous final year project, there may contain some outliers which are highly concerned the prediction performance of correlation models in this study. In order to identify the outlier, the correlation model for all soil which was established by Kuan Wan Kin (2011) was applied to plot with the database of LRT project in figure 5.1.

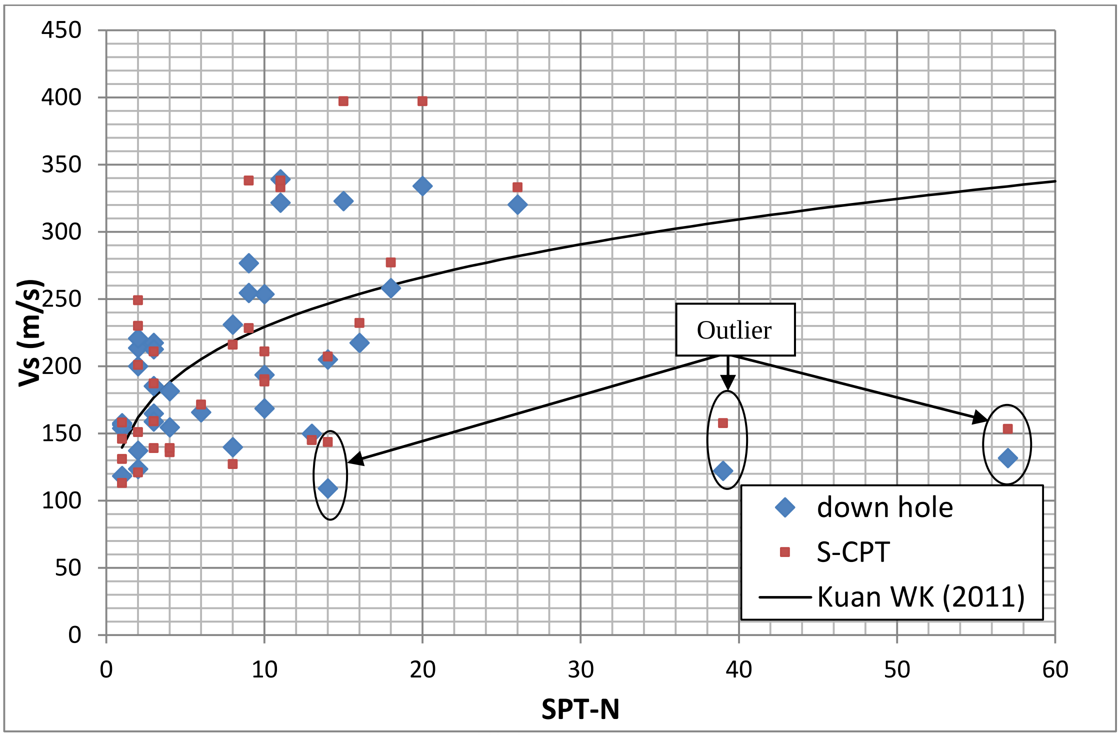


Figure 5.1 the correlation model of Kuan WK (2011) for all soil with 72 data points

According to figure 3.6, there are 6 data points (down hole: 3, cross hole: 3) were identified as outliers because they were generally under-predicted regarding to the correlation model. On the other hand, the measurement of low value of shear wave velocity with high SPT-N value is not respect to the actual physical meaning. The reason to cause these unreliable measurements is due to the complexity of soil structure and the precision of measurement. Therefore, the prediction performance of correlation model is affected if these outliers are selected into correlation analysis.

5.2 Linear Regression Analysis (EXCEL)

According to the Final Year Project of Kwan WK in 2011, The simple power-law regression analysis method were applied to identify the line or curve functions that are able to provide the best function based a set of data points. The reason to apply this method is because the curves functions can be identified a trend relationship by using linear, parabolic, or some other formation. There are various methods in regression analysis can be performed by using different computing software. Microsoft Excel software was applied into the research of establishment of correlation equation between shear wave velocity and SPT N value in 2011.The formation of the correlation equation is form by power law function, as shown in equation 5.1.

$$V_s = A \times (N)^B \tag{5.1}$$

where $V_s =$ shear wave velocity, m/s; $N =$ standard penetration blow count;

A, B = the regression analysis coefficients

The method of how to interpret the regression analysis coefficient is based on the linear Least-Square method. It is fair to mention that a rigorous functional dependence from the experimental data is seldom observed, since each of the involved magnitudes can depend on many random or unpredictable factors. Although the result that determine from the correlation equation is not equal to the measurement data, this approach is still be able to justify due to the practical utility of the obtained correlation equation.

5.2.1 Root mean square error (RMSE)

The demand of the proximity of the measured values from the original database with the predicted values by using correlation equation is based on examine the minimum root mean square error (RMSE), which the formula is shown in equation 5.2.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(m_i - p_i)^2}{n}} \tag{5.2}$$

where $\overline{m}$ = mean of measured values; n = the number of data presented in database

5.2.2 Linear Least Square method

Since power-law function is assigned in this research, in order to find the best fit of this function types so that the RMSE is minimum, the detail steps is show in following. First, the power law function with two parameters is shown in equation 5.3.

$$y = F(x, a, b) = ax^b \tag{5.3}$$

where: a,b = regression parameter

However, power law function is not a linear function, therefore, it is necessary to extract the natural logarithm which respect to equation 5.3 with assuming a>0, then the formula is reformed and shown in equation 5.4.

$$\ln F = \ln a + b \ln x \tag{5.4}$$

In order to make the equation look easier, $\ln a$, $\ln x$ and $\ln F$ are make the substitution to u, v and w, respectively. The formula express as a linear function:

$$w = \emptyset(v, u, b) = u + bv \tag{5.5}$$

where $u = \ln a$; $v = \ln x$; $w = \ln F$

It is necessary to take the derivative for equation 5.5, respect to coefficients u and b.

$$\frac{\partial \emptyset}{\partial u} = 1, \qquad \frac{\partial \emptyset}{\partial b} = v \tag{5.6}$$

Using linear least square method to solve the coefficients,

$$\begin{cases} \sum_{i=1}^{n}[m_i - \emptyset(v_i, u, b)]\, \frac{\partial \emptyset}{\partial u} = 0 \\ \sum_{i=1}^{n}[m_i - \emptyset(v_i, u, b)]\, \frac{\partial \emptyset}{\partial b} = 0 \end{cases} \tag{5.7}$$

Substitute the $\frac{\partial \emptyset}{\partial u}$ and $\frac{\partial \emptyset}{\partial b}$

$$\begin{cases} \sum(m_i - u - bv_i) = 0 \\ \sum(m_i - u - bv_i)v_i = 0 \end{cases} \tag{5.8}$$

Calculate the summation of v_i, m_i, ${v_i}^2$ and $m_i v_i$ for solving the equation 5.9.

$$\begin{cases} u + \frac{b}{n}\sum v_i = \frac{1}{n}\sum m_i \\ \frac{u}{n}\sum {v_i}^2 + \frac{b}{n}\sum {v_i}^2 = \frac{1}{n}\sum m_i v_i \end{cases} \tag{5.9}$$

where n is the number of database.

After the coefficients u and b solved, it should be cautioned that the coefficient u need to transform to an exponential function to compute the coefficient a, which is respect to the equation 5.4.

$$a = e^u \tag{5.10}$$

5.2.3 Coefficient of determination (R^2)

In order to evaluate the prediction quality of the corresponding correlation functions, Coefficient of Determination (R^2) equals to the square of the Pearson function moment correlation coefficient r in Excel. The correlation coefficient can be used to identify the extent of a linear relationship between two data sets.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} \tag{5.11}$$

For the correlation functions is power law functions, $Vs = AN^B$ ($\ln Vs = B\ln N + A$).

where $x = \ln N$; $y = \ln Vs$.

The reason of these two parameters need to take the logarithmic is because the power law function is nonlinear function. Since the correlation coefficient is only based on the linear relationship of the function and the model transformation between nonlinear function to linear function, as shown in equation, the identification of corresponding dataset is necessary to transform.

After correlation coefficient is determined, the coefficient of determination can be calculated by square of the correlation coefficient. It can be interpreted as the relationship of the variance proportion in y dataset to the variance in x dataset.

In this study, power law regression model are applied to establish shear wave velocity. However, the power law function is nonlinear function. It is necessary to transform to linear function by take logarithm for each parameter in order to fit in the linear least square approach, as mention in section 5.2.2. Therefore, this process should be avoided.

5.3 Nonlinear curve fitting

5.3.1 Curve fitting tool

In order to solve the coefficients for the power law function by using nonlinear least square method, the Matlab software is applied in this research. Curve fitting toolboxes is a very powerful and convenience functions, which are available to solve various 3 dimensional curves or surface fitting problem.

Curve Fitting Toolbox are allowed user to set an option of the fitting model if the model is not performed the expectation of user, the fitting option are included Robust, Algorithm, Finite Differencing Parameters, Fit Convergence Criteria and Coefficient Parameter. The concept and corresponding setting would be described in following.

5.3.2 Robust

Robust : which is used to specify whether use robust least-squares fitting method.

It can be set Off, On or LAR.

- Off — robust fitting do not applied (default).
- LAR — using minimize the least absolute residuals method to fit.
- Bisquare— using minimize the residuals and reduce the weight of outliers using bisquare weights. It is common the best choice for robust fitting.

5.3.3 Algorithm

- Trust-Region—which is the default algorithm in Matlab curve fitting tool for the nonlinear curve fitting purpose, and must be applied in the situation that if the lower or Upper coefficient constraints are specified.
- Levenberg-Marquardt—it is not commonly applied except if the trust-region algorithm cannot provide a reasonable fitting curve or surface, or the Lower or Upper coefficients are not specified.
-

5.3.4 Finite Differencing Parameters

- DiffMinChange— Minimum change in coefficients for finite difference Jacobians.
- DiffMaxChange— Maximum change in coefficients for finite difference Jacobians.

5.3.5 Fit Convergence Criteria

- MaxFunEvals— Maximum number of function (model) evaluations.
- MaxIter— Maximum number of fit iterations.
- TolFun— Termination tolerance used on stopping conditions involving the function (model) value.
- TolX— Termination tolerance used on stopping conditions involving the coefficients.

5.3.6 Coefficient Parameters

• StartPoint— The starting values of regression coefficient depends on the format of model.

• Lower — Lower limit on the regression coefficients. To indicates the unconstrained coefficients, -Inf is defaulted for most library models.

• Upper — Upper bounds on the fitted coefficients. The tool only uses the bounds with the trust region fitting algorithm.

5.3.7 Optimized Starting Points and Default Constraints

If the starting points are optimized, the model was fitted heuristically based on the existing data set in Matlab.

If constraints were not applied to model, the starting points and constraints of the coefficient can be overridden by values setting in the Fit Options dialog box.

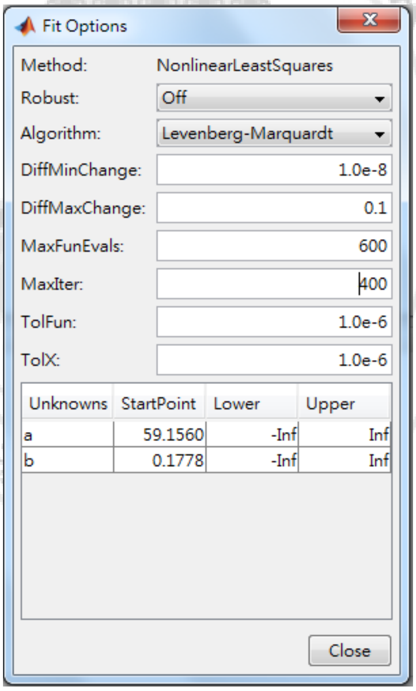


Figure 5.2 the Fit Options dialog box

Notes that StartPoint for each fitting models would be different since the Curve Fitting Toolbox can automatically optimize the starting point for power-law regression model.

5.3.8 NonLinearModel.fit

There had some limitation about the number of parameters in Curve Fitting Toolbox, it only provided 2 or 3 parameters for the curve fitting analysis. Regarding to correlation between shear wave velocity and SPT-N value, there are many unpredictable uncertainly in actual site situation. Therefore, effective stress and cone resistance are taken into the correlation analysis of shear wave velocity. The influence of neglect the effective stress and cone resistance have introduced in Character 2.4 and Character 2.5, respectively.

For the influence of the cone resistance, based on the correlation regression functions between the Maximum Shear Modulus and cone resistance is same to the correlation between shear wave velocity and SPT-N value.

$$\mathrm{G}_{max} = A\ (q_c)^B \tag{5.12}$$

where G_{max} = maximum shear modulus, kPa ; q_c = cone resistance, kPa ; A, B = correlation coefficients

Moreover, the physical relation between Maximum Shear Modulus and shear wave velocity.

$$\mathrm{V}_s = \sqrt{\frac{\mathrm{G}_{max}}{\rho}} \tag{5.13}$$

where Vs = shear wave velocity, m/s; G_{max} = maximum shear modulus, kPa; ρ = density of soil, kg/m^3

Regarding to equation 5.12 and 5.13, it has been considering that the cone resistance also contain the similar power-law correlation functions.

$$\mathrm{Vs} = A'\ (q_c)^{B'} \tag{5.14}$$

where Vs = shear wave velocity, m/s; q_c = cone resistance, kPa; A', B' = correlation coefficients

Based on the SPT-N, effective stress and cone resistance has the similar correlation relation with shear wave velocity. It can be estimated the new formation of correlation functions in order to take all those parameters into one equation.

$$\mathrm{Vs} = \mathrm{A}'' \, (\mathrm{N})^{B''} (\sigma'_{v})^{C''} (q_c)^{D''} \quad (5.15)$$

where Vs = shear wave velocity, m/s; N = standard penetration resistance; σ'_{v} = effective stress, kPa; q_c = cone resistance, kPa; A″, B″, C″, D″ = correlation coefficients

However, since this regression analysis is involved four parameters, Curve Fitting Toolbox cannot be applied in this regression analysis. Therefore, in order to solve this nonlinear regression functions, the “NonLinearModel.fit” functions which is provided in Matlab would be applied. The operation of this function is not convenience compare with Curve Fitting Toolbox. An extra function file necessary to be assigned in Matlab software in order to generate the “NonLinearModel.fit” function based on the regression model. On the other hand, each parameter is identified into a matrix and the initial (guess) value is necessary to assign for the curve fitting operation purpose. The advantage of this method there has not the limitation of the number of parameters. It is helpful if there have any addition parameters or data are available to be taken into this research analysis.

5.3.9 Levenberg-Marquardt algorithm

The concept of this function is same as the Curve Fitting Toolbox, it also applied the Levenberg-Marquardt algorithm to solve the type of nonlinear least square problem. It is clear that the purpose of curve fitting is to minimize the value of sum of square for corresponding fitting function f(x).

$$\underset{\mathrm{x}}{\mathrm{minimize}}\, \mathrm{f(x)} = \frac{1}{2}\|F(x)\|_2^2 = \sum_i {F_i}^2(x) \quad (5.16)$$

where F is the vector-valued functions $F(x) = (f_1(x), f_2(x) \cdots f_m(x))^T$; f(x) is the residual for the corresponding data point

The reason why the least square function has been derivatives by $\frac{1}{2}$ is to make the derivatives of the functions become less cluttered. In this curve fitting method, it can control where the output, y(x,t), minus the continuous model trajectory, $\varphi(t)$, for vector x and scalar t. This problem can be expressed as:

$$\min_{x \in R} \int_{t_1}^{t_2} (y(x,t) - \varphi(t))^2 dt \qquad (5.17)$$

where $y(x,t)$ are measured sclar functions; $\varphi(t)$ are predicted scalar functions.

When the integral is discretized using a suitable quadrature formula, it can be formulated as a least-squares problem:

$$\min_{x \in R} f(x) = \sum_{i=1}^{m} (\bar{y}(x,t_i) - \bar{\varphi}(t_i))^2 \qquad (5.18)$$

where $\bar{y}$ and $\bar{\varphi}$ include the weights of the quadrature scheme.

Therefore, the problem respect to the vector function$F(x)$ can be express as:

$$F(x) = \begin{bmatrix} \bar{y}(x,t_1) - \bar{\varphi}(t_1) \\ \bar{y}(x,t_2) - \bar{\varphi}(t_2) \\ \cdots \\ \bar{y}(x,t_m) - \bar{\varphi}(t_m) \end{bmatrix} \qquad (5.19)$$

In this least square method, the residual $\|F(x)\|$ is reduced to the smallest value (0) as possible. It is general practice based on the setting of initial value. Although the function in Least Square can be minimized using unconstrained minimization technique, certain characteristics of the problem is still contained and exploited to improve the iterative efficiency of the minimization process of least square.

It is necessary to examine the gradient vector, Jacobian matrix (first derivative of F(x) respect to regression coefficient) and Hessian matrix of F(x) for solving the Least Square problem.

$$G(x) = J(x)^T F(x) \tag{5.20}$$

$$H(x) = 2J(x)^T J(x) + 2Q(x) \tag{5.21}$$

$$Q(x) = \sum_{i=1}^{m} F_i(x) H_i(x) \tag{5.22}$$

where G(x) = the gradient vector of f(x); H(x) = the Hessian matrix of f(x)

Let x_k be the solution of this Least Square problem, which indicates the residual $\|F(x)\|$ and Q(x) tend to zero. The Gauss-Newton direction can be used as a basis for an optimization procedure when $\|F(x)\|$ is small at the solution. However, it is often encounter problems when the Q(x) is significant. Therefore, the Levenberg-Marquardt method is applied, because it can overcome this problem. The search direction of this method can be expressed as a linear set of equations.

$$[J(x_k)^T J(x_k) + \lambda_k I] d_k = -J(x_k)^T F(x_k) \tag{5.23}$$

where λ_k = Damping factor; I = Identity matrix; d_k = search direction which obtained at each major iteration

The Damping factor λ can control both the magnitude and direction of d_k, which the initial damping factor λ_0 can be set by own evaluation. Since the damping factor is adjusted in each iteration, if the damping factor is close zero when the reduction of $\|F(x)\|$ is large, the search direction of d_k is close to Gauss-Newton algorithm. On the other hand, if reduction $\|F(x)\|$ is small, the damping factor will increase and the search directiontends to the steepest descent direction with the magnitude tend to equal to zero. It indicates for some sufficiently large value of λ_k, and $F(x_k + d_k)$still less than $F(x_k)$. Therefore, the damping factor λ_k is decreasing even when second-order terms restrict the efficiency of the Gauss-Newton method is encountered.

5.3.10 Operation

In order to applied the "NonLinearModel.fit" function in Matlab, First, a new function file which contain formation of fitting model. Second, insert the relative database and set the initial (guess) value for each regression parameter (1 is commonly set as initial value in this research). Finally, input the "NonLinearModel.fit" function code as shown in below.

betahat = NonLinearModel.fit(unnamed4,unnamed,@curvefittingqc,unnamed5)

where betahat = regression parameter result; unnamed4 = predictor variable (i.e. SPT-N, qc, σ'_v); unnamed = response variable (i.e. Vs); curvefittingqc = fitting model; unnamed5= initial (start) guess value for each parameter.

Instead of the Matlab function can calculate the fitting models without spend significant time, it also calculate the corresponding evaluate index at the same time, i.e. R^2, RMSE, etc. The example of the result after the "NonLinearModel.fit" functions run would be shown in Figure 5.3.

```
>> load('databaseqcclay.mat')
>> betahat=NonLinearModel.fit(unnamed4,unnamed,@curvefittingqc,unnamed5)

betahat =


Nonlinear regression model:
    y ~ curvefittingqc(b,X)

Estimated Coefficients:
          Estimate     SE          tStat     pValue
    b1     32.651       13.732     2.3778      0.022285
    b2    0.27563      0.06145     4.4855    6.0003e-05


Number of observations: 42, Error degrees of freedom: 40
Root Mean Squared Error: 55.6
R-Squared: 0.343,  Adjusted R-Squared 0.326
F-statistic vs. zero model: 289, p-value = 1.71e-24
```

Figure 5.3 the result for the power-law regression between V_s and q_c (5.12) calculated by Matlab "NonLinearModel.fit" function.

Notes the calculation equation of the R-squared and Root Mean Squared Error is shown in equation 5.24 and 5.25, respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(m_i - p_i)^2}{\sum_{i=1}^{n}(m_i - \bar{m})^2} \quad (5.24)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(m_i - p_i)^2}{n}} \quad (5.25)$$

where m = the values of measured Vs in the field; p = predicted Vs by the proposed model; $\bar{m}$ = mean of measured values; n = the number of data presented in database

5.4 Genetic Programming

According to the literatures from the EJGE and the correlation analysis result which respect to Macau soil, the simple power law regression is not available to provide a good enough prediction performance for the SPT-N and Shear Wave velocity. Regarding to improve the prediction performance, EJGE was published that Genetic Programming software (GP) can achieve the higher prediction performance, therefore, it is decided that the genetic programming software is be applied to Macau soils in order to obtain the higher prediction quality. The concept of this computing software is introduced in this chapter.

5.4.1 General

GP is an advanced computer programs that used to perform the significant prediction potential regression model with numbers of tree structure. It can be modified after a population of computer programs (represented by tree structures) is generated. A new population is created by the best performing trees though the mutating and crossing over process. This process is iterated until the population contains programs that provided a desirable solution.

When GP is going to build an empirical model based on existing database, it often classify as symbolic regression. It is not similar to the standard regression analysis that the structure of the mathematical model must be specified, GP can evolve both the structure and the parameters of the model automatically.

5.4.2 GPTIPS1.0

GPTIPS1.0 is a Matlab toolbox software which is widely applied to various discipline research in recent year. Multigene symbolic regression formed a unique type of symbolic regression which is applied in GPTIPS1.0. It evolves linear combinations of non-linear transformation of the input variables. GPTIPS1.0 can provide a number of convenient functions which can use to explore the population of evolved models, investigating model behavior, post-run model simplification and export to different formats. One of the main features of GPTIPS is that it can be configured to evolve Multigene individuals.

5.4.3 Multigene individuals

A Multigene individual model can be assembled with one or more genes. Each of genes expresses as a "traditional" GP tree and acquired incrementally by individuals in order to improve fitness. The fitness approach is commonly similar to least square method on a data set that respect to the corresponding model.

5.4.4 Symbolic Regression

Symbolic regression is assembled by a population of trees which evolve by genetic programming. A $(N \times 1)$ vector of outputs $\mathbf{y}$ is predicted by symbolic regressions which contain one variable. For a multi-variable model, the outputs of y is formed a

$(N \times M)$ matrix. Where N indicates the number of measurement data and M indicates the number of input (predictor) variables.

5.4.5 Multigene Symbolic Regression

For Multigene Symbolic Regression, the regression model is formed by linear combination of each gene, and the gene is formed nonlinear combination. In GPTIPS, the least square approach is applied to obtain the optimal solution for the automatically. The prediction $\hat{y}$ of the each output variable y is constructed by each of the trees or genes weights in the multigene individual (multi variable model) plus a bias term (constant). Each tree is a function of zero or more of the N input variables $x_1, \cdots x_N$.

Mathematically, a Multigene regression model can be written as:

$$\hat{y} = d_0 + d_1 \times \mathrm{tree1} + \cdots + d_M \times \mathrm{treeM} \qquad (5.26)$$

where d_0 = bias (offset) term; $d_1, \cdots, d_M$ = the gene weights; M = the number of genes (trees);

This model structure contains non-linear terms but is linear in the parameters with respect to the coefficient (gene weights). In GPTIPS, the user need to defined constraints of the maximum number of genes G_{max} and maximum tree depth D_{max} before it is evolved automatically during a run based on training data and evaluate the evolved model by using testing data.

Training Data – a set of inputs values and output values which is determined by directly measurement.

Testing Data – another set of input and output values which are used to evaluate the prediction potential.

Since the linear coefficients ($d_0, d_1, \cdots$) are estimated based on the training data by using least squares approach. Therefore, Multigene GP combined with traditional linear regression approach and contained non-linear term without specify the structure of the nonlinear regression models before the establishment the regression model.

5.4.6 Crossover and Mutation

For genetic programming, there are numbers of individuals (models) are created by generated trees with between 1 and G_{max} genes, where G_{max} represent the maximum number of genes contain in one single individual. The initial populations is constructed after each individuals are generated. Two point high level crossover operator is applied to exchange or eliminate between the genes during the GPTIPS is running. If the ith gene in an individual is represent G_i, then this crossover operator is performed in the following.

Let the 3 genes ($G_1 G_2 G_3$) is contained by first individual (model) and the other 4 genes contained by second individual (model). The genes are selected randomly between each individual by two crossover point. The genes enclosed by the crossover points are denoted by $\langle \cdots \rangle$.

Individual 1: $(G_1 \langle G_2 \rangle G_3)$ Individual 2: $(G_4 \langle G_5 G_6 G_7 \rangle)$

If the maximum number of genes equals to 5, the gens enclosed by the crossover points and exchanged between these two genes and formed two new individuals, as shown in below.

New individual 1: $(G_1 G_5 G_6 G_7 G_3)$ New individual 2: $(G_4 G_2)$

New genes for both individuals are produced or removed after two point high level crossover applied. If an individual is contained more genes than G_{max} after the gens exchange, the additional genes are randomly selected and deleted until the individual contains G_{max} genes. The process is operating continuously until the assigned number of generation is reached.

5.4.7 General setting

In GPTIPS1.0, the regression model is not concrete format as simple power-law regression. The created models are represented in tree-based structures as express in 5.26. GPTIPS1.0 uses the lexicographic tournament selection approach to control the complexity of the model, the corresponding setting in this research shown as Table 5.1.

Table 5.1 Setting of initially defined parameters in GPTIPS1.0

Parameter	Value
Population size	100
Number of generations	100
Tournament size	1
Maximum depth of trees	5
Function set	+ , – , × , ÷
Maximum number of genes	2
Constants range	[-10, 10]

5.5 Evaluation

Based on the Chapter 5.1 and 5.3, it introduced that the power-law regression approach are different between Excel and Matlab. For Excel, the corresponding nonlinear regression model is necessary to transform to linear regression, and apply the linear least square method to compute the regression constants. For Matlab, the nonlinear curve fitting is directly applied to regression model and compute the regression constants. Since the regression approach is different between Excel and Matlab, the corresponding regression model is also different. In order to compare the difference, the correlation models are calculated again by using Matlab based on the same database which obtained from previous study. The correlation models are organized and shown in table 5.2. On the other hand, in order to improve the prediction performance of same database, Genetic Programming toolbox (GPTIPS1.0) is also applied based on the setting which is shown in Table 5.1, and the corresponding GP-model is shown in equation 5.27.

Table 5.2 Comparison of the correlation model between linear curve fitting and nonlinear curve fitting

Database	Data point	Correlation Model	
		Linear curve fitting	Nonlinear curve fitting
SaeTeng, 2009	73	$\mathrm{Vs} = 82.13N^{0.336}$	$\mathrm{Vs} = 68.1N^{0.406}$
Pun and Chao, 2010	48	$\mathrm{Vs} = 114N^{0.214}$	$\mathrm{Vs} = 109.1N^{0.239}$
KuanWai Kin,2011	71	$\mathrm{Vs} = 139.4N^{0.216}$	$\mathrm{Vs} = 137N^{0.243}$
Combined	192	$\mathrm{Vs} = 118.1N^{0.229}$	$\mathrm{Vs} = 109.5N^{0.277}$

According to table 5.2, the results of correlation models between Excel and Curve Fitting Tool are different for the same database. In order to compare the prediction performance between linear curve fitting and nonlinear curve fitting, the correlation models by using linear curve fitting and the models by using nonlinear curve fitting are plotted in the following.

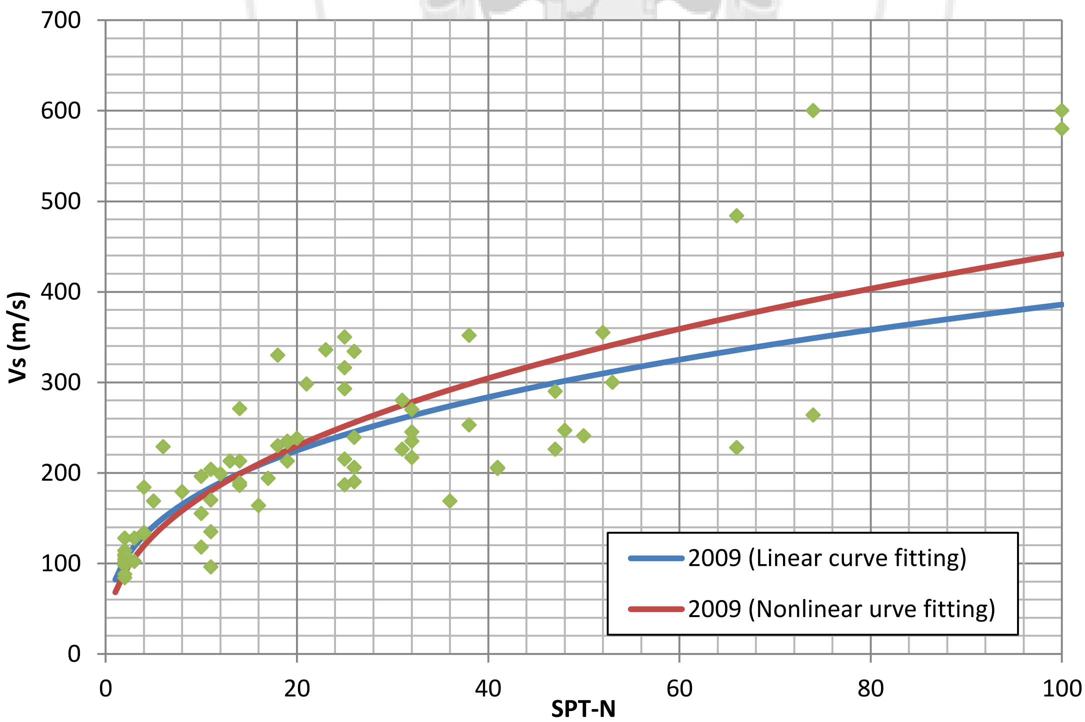


Figure 5.4 Correlation models for linear curve fitting and nonlinear curve fitting based on database of SaeTeng

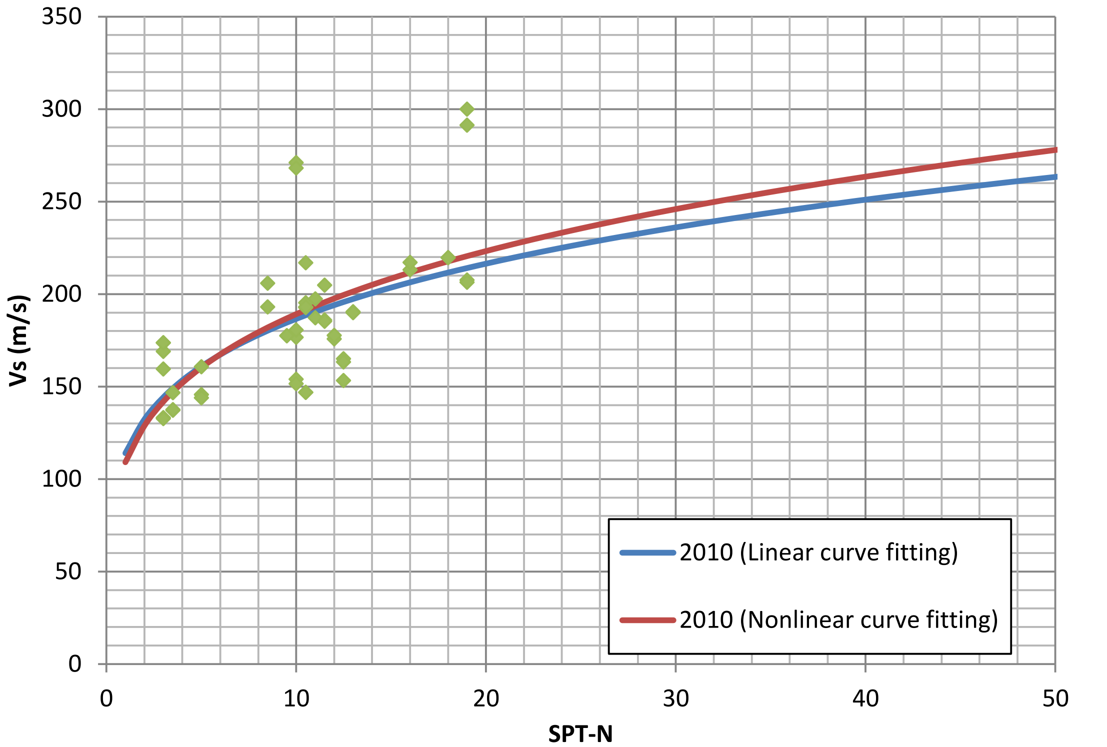


Figure 5.5 Correlation models for linear curve fitting and nonlinear curve fitting based on database of Pun and Chao

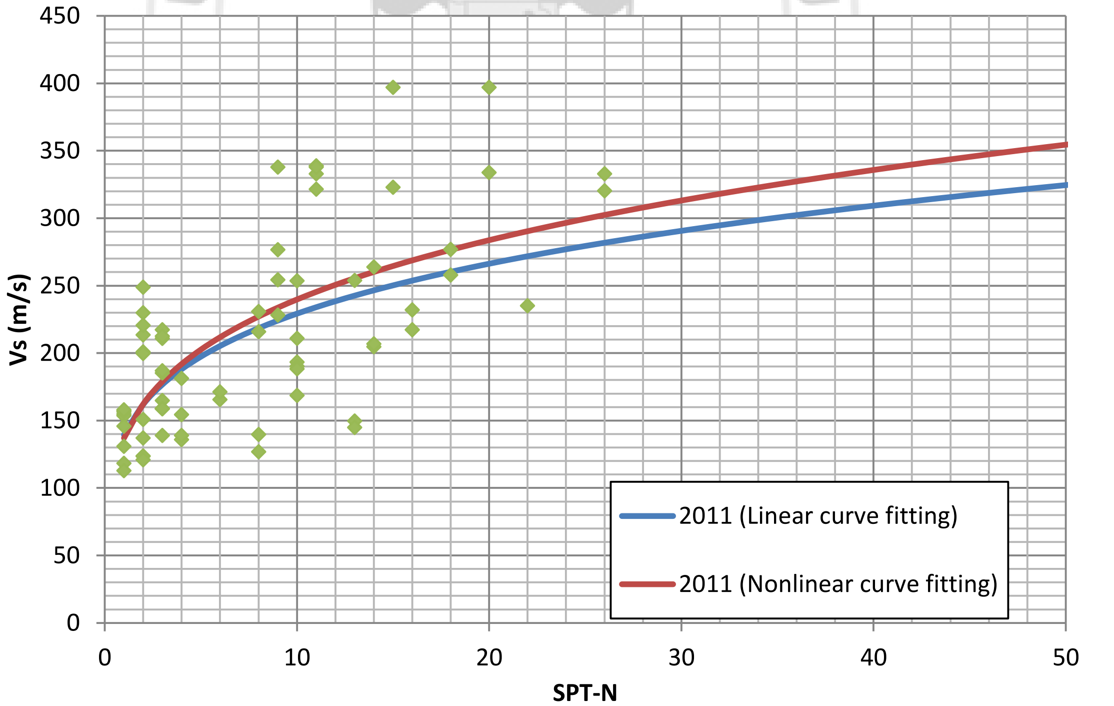


Figure 5.6 Correlation models for linear curve fitting and nonlinear curve fitting based on database of Kuan WK

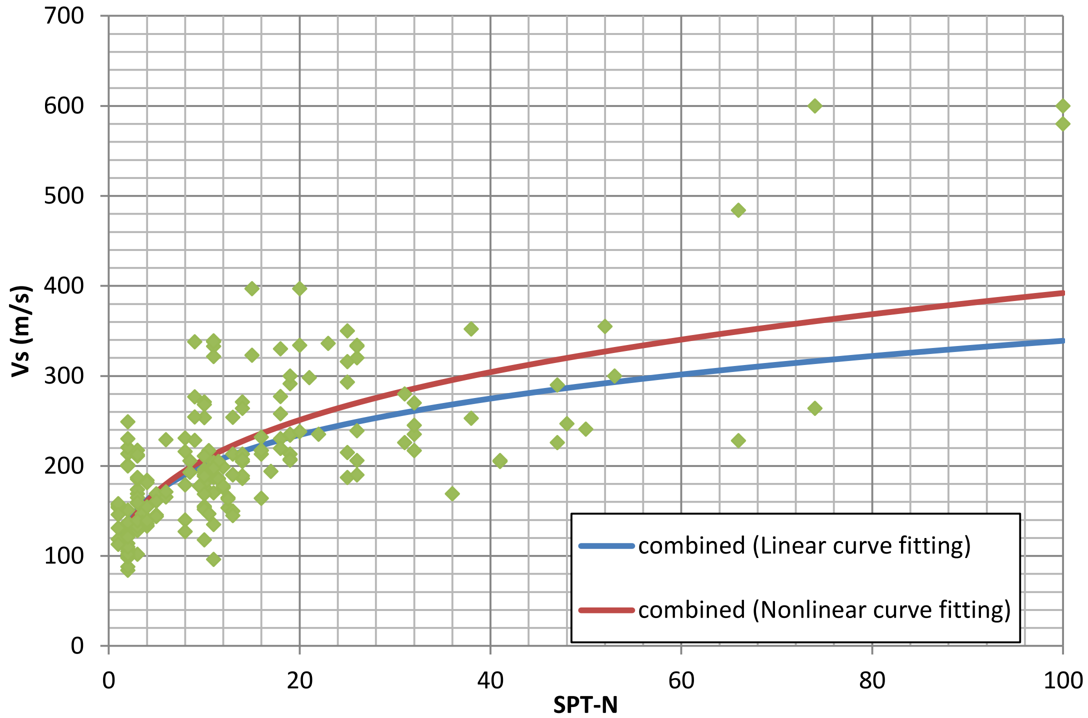


Figure 5.7 Correlation models for linear curve fitting and nonlinear curve fitting based on combined database

According to figure 5.4, 5.5, 5.6 and 5.7, the correlation models of nonlinear curve fitting provide higher value of shear wave velocity compared with linear curve fitting when SPT-N is large, as shown in figures 5.4 to 5.7. For large SPT-N values, the measurements of shear wave velocity are generally larger than the predictions of correlation model. Therefore, correlation models established by Matlab can reduce the difference for the case of large shear wave velocity. It indicates Matlab correlation models have higher prediction performance for when SPT-N value is large, but it cannot be identified which method provides better prediction performance when SPT-N is small.

In order to evaluate the prediction performance between those models, root mean square error (RMSE) is used to identify and the calculation equation is shown in 5.5. The lower value of RMSE is represented the better prediction performance. The corresponding results of RMSE for each model from table 5.2 are shown in table 5.3.

Table 5.3 Comparison of RMSE between linear curve fitting and nonlinear curve fitting

Validation Database	RMSE	
	Linear curve fitting	Nonlinear curve fitting
2009	67.94	65.72
2010	30.17	29.96
2011	53.44	52.63
combined	61.94	60.08

Based on the table 5.3, the correlation models computed from Curve Fitting Tool is better than Excel. Therefore, Matlab software was applied in this study in order to obtain the better correlation models.

GP model based on the combination of SaeTeng (2009), Pun and Chao (2010) and Kuan Wai Kin (2011):

$$Vs = 0.04691N - \frac{3421}{N+15.11} - 0.02125N^2 + 0.0004948N^3 + \frac{4.383}{N-0.6444} + 341.6 \quad (5.27)$$

where V_s = shear wave velocity, m/s; N = standard penetration test N value

For GP model, the highest prediction performance is provided compare with linear curve fitting and nonlinear curve fitting, the corresponding evaluation of this model is shown in table 5.4.

Table 5.4 the RMSE of GP model based on the corresponding Validation Database

Validation Database	RMSE
	GP
2009	59.89
2010	33.71
2011	60.04
combined	54.60

Based on table 5.4, GP model provided the best prediction performance for SaeTeng (2009), KuanWai Kin (2011) and combined database. On the other hand, it provided the worst prediction performance for Pun and Chao (2010). Although the perfect correlation model cannot be established, however, the prediction performance is successfully improved after genetic programming was applied.

CHAPTER 6 DATA ANALYSIS

6.1Introduction

In this chapter, there are total 28 correlation models of shear wave velocity and mainly separate into 4 sections to introduce. In section 6.2, the correlation models of shear wave velocity with SPT-N value are introduced. The correlation models of shear wave velocity with SPT-N and vertical effective stress are introduced in section 6.3.In section 6.4, the cone resistance is combined with SPT-N to establish correlation model of shear wave velocity, however, cone resistance commonly contained large variation of soil behavior, the appropriate calculation method of the average cone resistance is hard to identify. Therefore, four different calculation methods of average cone resistance were applied in this study in order to identify the best calculation based on the prediction performance of corresponding models. In section 6.5, SPT-N value, vertical effective stress and cone resistance are combined to establish the correlation models of shear wave velocity. Finally, the prediction performance of the correlation models are summarized and compared for each specific soil layer in order to identify the best correlation models in section 6.6.

In order to establish the new correlation models of shear wave velocity of Macau soil, the databases of LRT-C250 project were selected in this study and shown in Appendix A. After the outliers are eliminated, There are total 66 datasets (Marine Deposit: 32, Alluvium: 34) were selected.

Root mean square error (RMSE) is applied to identify the prediction performance based on the validation database of marine deposit, alluvium and the combination of these two soil type. Moreover, the best correlation model which obtains from either Curve Fitting Tool or “NonLinearModel.fit” function is compared with the GP model by using predicted Vs verses measured Vs plot, the higher percentage with Region 3 and 4 indicates higher chance to obtain reasonable prediction.

6.2 Correlation Model of Shear Wave Velocity with SPT-N

6.2.1 Models

In this section, Curve Fitting Tool and Genetic Programming (GPTIPS1.0) were applied to establish the correlation models of shear wave velocity with SPT-N.

SPT-N model:

(Marine Deposit) $$V_s = 155.1N^{0.0525} \tag{6.1}$$

(Alluvium) $$V_s = 150.2N^{0.2259} \tag{6.2}$$

(All soil) $$V_s = 136.2N^{0.2487} \tag{6.3}$$

GP SPT-N model:

(All soil) $$Vs = 5.199N - \frac{5.133N}{N-11.65} + 0.169N^2 + 151.3 \tag{6.4}$$

where V_s = shear wave velocity, m/s; N = standard penetration test N value

For equation 6.1, the lower curvature performance can be obtained for equation 6.1 because the power number of SPT-N is significant smaller compare with equation 6.2 and 6.3. Moreover, equation 6.4 (GP model) performed more complex format of model than power law functions. In order to compare these four different models, there are plotted in the same figure as shown in figure 6.1.

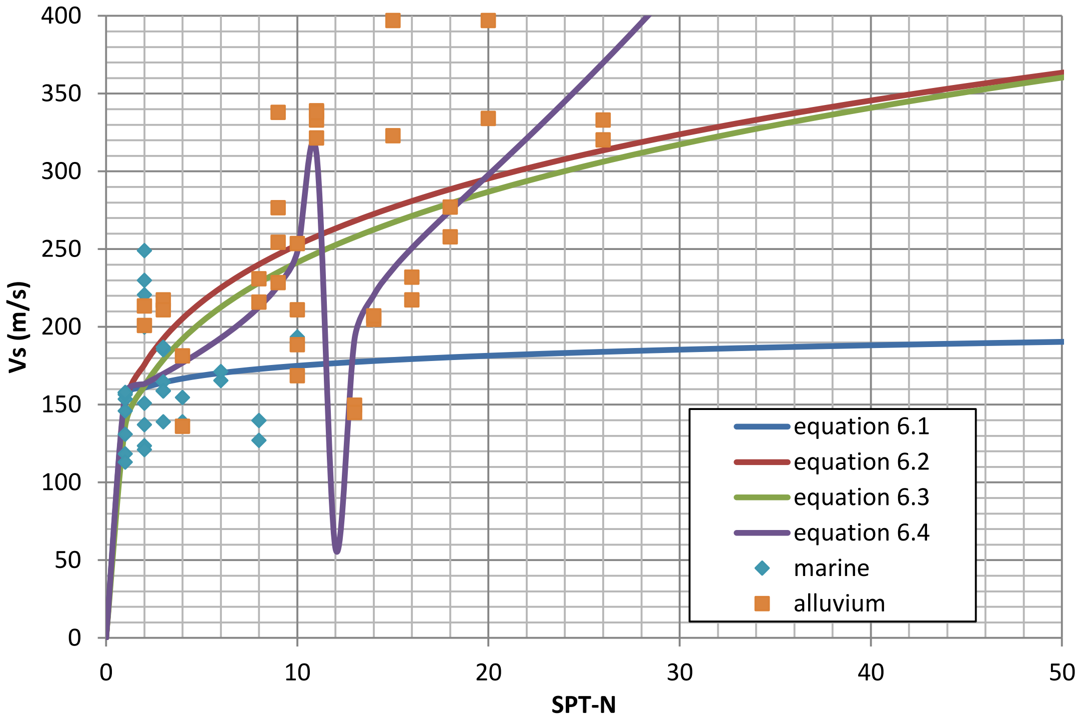


Figure 6.1 Correlation models of Shear Wave Velocity and SPT-N by Curve Fitting Tool and genetic programming

6.2.2 Model Evaluation

According to figure 6.1, the shape of the equation 6.1 is performed as a straight line function. This performance is not respect to the general correlation relationship between shear wave velocity and SPT-N. The largest value is determined by equation 6.2 and it contains less difference with equation 6.3. In this case, the measurement of alluvium is more reliable than marine deposit, because the outliers are generally obtained from the database of marine deposit. Equation 6.4 is provided a unique correlation model between shear wave velocity and SPT-N. A significant increment is contained when SPT-N equals to 11, and suddenly drops down when SPT-N equals to 12. The correlation relationship between shear wave velocity and SPT-N contained large variation when SPT-N is less 20. Finally, the model keeps increasing and never decrease when N is larger than 20. The reason of this behavior provided by equation 6.4 is the limitation of SPT-N database from Appendix A, because the maximum value of SPT-N is only up to 26, so that the genetic programming generate high shear wave velocity when the value of SPT-N larger the range of existing database.

For the prediction evaluation, equation 6.1 and 6.2 are not included in the comparison with others, it is because the calibration data set of these two models only based on marine deposit and alluvium, respectively. Therefore, equation 6.3 and 6.4 are compared by using root mean square error (RMSE) which respect to marine deposit, alluvium and all soil accordingly, the corresponding result are provided in table 6.1.

Table 6.1 Result of RMSE for Equation 6.3 and 6.4

Validation data set	Number of dataset	SPT-N model	GP SPT-N model
		Equation 6.3	Equation 6.4
Marine Deposit	32	42.996	39.848
Alluvium	34	62.429	52.289
All soil	66	53.889	46.673

According to table 6.1, the GP model provided better predication performance compare with equation 6.3 respect to those 3 different soil types because the result of RMSE of equation 6.4 (GP SPT-N model) is less than equation 6.3. In order to investigate the prediction performance statistically, the figures of predicted Vs versus measured Vs are plotted based on SPT-N model (equation 6.3) and the corresponding GP model (equation 6.4).

6.2.3 Result

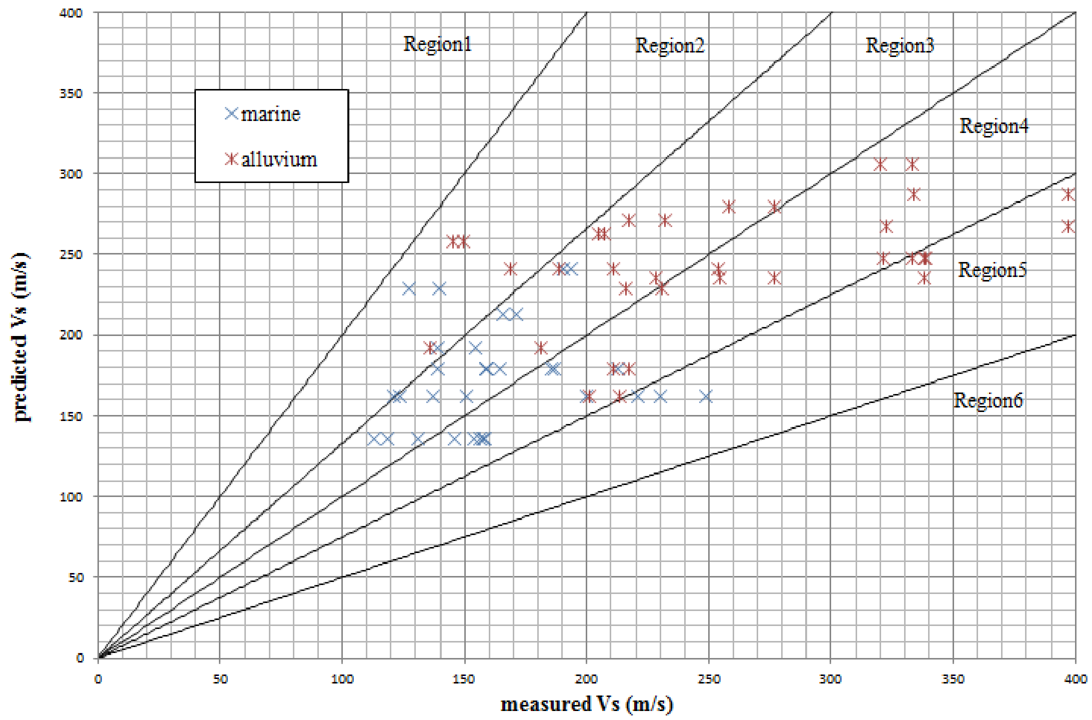


Figure 6.2 predicted Vs verses measured Vs for equation 6.3

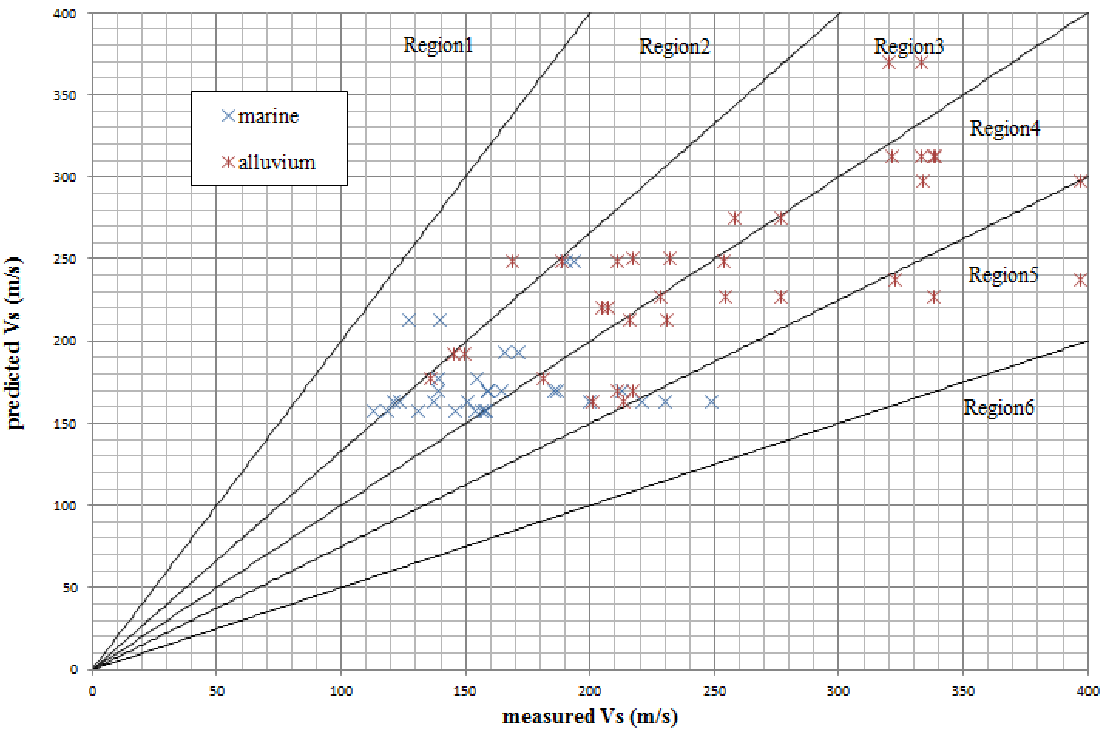


Figure 6.3 predicted Vs verses measured Vs for equation 6.4

Table 6.2 Distribution of data for equation 6.3

SPT-N model Equation 6.3	Marine Deposit		Alluvium	
Region	Number of data	Percentage of data (%)	Number of data	Percentage of data (%)
1	0	0.00	0	0.00
2	4	12.50	4	11.76
3	16	50.00	11	32.35
4	9	28.13	13	38.24
5	3	9.38	6	17.65
6	0	0.00	0	0.00
Percentage of over-predicted, %	62.50		44.12	
Percentage of under-predicted, %	37.50		55.88	
Percentage of within Region 3 and 4, %	78.13		70.59	

Table 6.3 Distribution of data for equation 6.4

GP SPT-N model Equation 6.4	Marine Deposit		Alluvium	
Region	Number of data	Percentage of data (%)	Number of data	Percentage of data (%)
1	0	0.00	0	0.00
2	4	12.50	1	2.94
3	19	59.38	12	35.29
4	6	18.75	18	52.94
5	3	9.38	3	8.82
6	0	0.00	0	0.00
Percentage of over-predicted, %	71.88		38.24	
Percentage of under-predicted, %	28.13		61.76	
Percentage of within Region 3 and 4, %	78.13		88.24	

For equation 6.3, the percentage of over-predicted for marine deposit equals to 62.50%, and 44.12% for alluvium. Therefore, this model is over-predicted to marine deposit and under-predicted to alluvium. Moreover, this model has higher chance to obtain reasonable prediction for marine deposit than alluvium because the percentage of within region 3 and 4 for marine deposit equals to 78.13%, and 70.59% for alluvium.

For equation 6.4, the percentage of over-predicted for marine deposit equals to 71.88%, and 38.24% for alluvium. Therefore, this model is also over-predicted to marine deposit and under-predicted for alluvium. Moreover, this model has higher chance to obtain reasonable prediction for alluvium than marine deposit because the percentage of within region 3 and 4 equals to 78.13% for marine deposit, and 88.24% for alluvium.

In comparing this 2 equations based on table 6.2 and table 6.3, the result represented that equation 6.4 (GP SPT-N model) has higher prediction performance. However, the overall prediction performances for these two models are not good although a significant improvement can be obtained for alluvium after the genetic programming applied. Therefore, the additional representative soil parameters are necessary to take into consideration in order to improve the prediction performance for marine deposit and alluvium.

6.3 Correlation Model of Shear Wave Velocity with SPT-N and vertical effective stress

Based on prediction performance of correlation model of shear wave velocity with SPT-N, the correlation model still cannot apply with confidence. Therefore, the dataset of vertical effective stress is applied to combine with the database of SPT-N to establish the correlation models of shear wave velocity in order to improve the prediction performance.

6.3.1 Models

After the database of SPT-N and vertical effective were combined, Curve Fitting Tool and Genetic Programming (GPTIPS1.0) were applied to establish the correlation models of shear wave velocity with SPT-N and vertical effective stress.

SPT-N,σ'_v model

Marine Deposit

$$V_s = 457.4(N)^{0.1226}(\sigma'_v)^{-0.2543} \tag{6.5}$$

Alluvium

$$V_s = 13.69(N)^{0.1276}(\sigma'_v)^{0.523} \tag{6.6}$$

All soil

$$V_s = 43.41(N)^{0.176}(\sigma'_v)^{0.2639} \tag{6.7}$$

GP SPT-N,σ'_v model:

All soil

$$V_s = 30.83N - \frac{4.019\sigma'_v}{N^2+N-\sigma'_v} + \frac{2336N-8008}{3N-\sigma'_v} + 52.59 \tag{6.8}$$

where V_s = shear wave velocity, m/s; N = standard penetration test N value; σ'_v = vertical effective stress, kPa

6.3.2 Model Evaluation

In comparing between equations 6.5 to 6.7, the difference of equation 6.5 is obtained by comparing with equation 6.6 and 6.7.A negative power number for the vertical effective stress parameter and the apparent large regression coefficient are obtained from equation 6.5. It is caused by the unreliable measurement and the soil complexity for marine deposit, as described in section 6.2.2. In order to identify the prediction performance for each correlation models, coefficient of determination (R^2) is selected to identify the prediction performance for corresponding validation data set, as shown in table 6.4.

Table 6.4 Result of R^2 for equation 6.5, 6.6, 6.7 and 6.8

Model	Equation	Calibration data set	Validation data set	Number of data point	R^2
SPT-N,σ'_v	6.5	Marine Deposit	Marine Deposit	32	0.120
	6.6	Alluvium	Alluvium	34	0.358
	6.7	All soil	All soil	66	0.464
GP SPT-N,σ'_v	6.8	All soil	All soil	66	0.651

According to table 6.4, equation 6.5 provided a poor prediction performance of shear wave velocity because the lowest value of R^2 is obtained compare with other models. On the other hand, equation 6.8 provided the best prediction performance after the Genetic Programming applied.

In order to identify the improvement of GP model, equation 6.7 and equation 6.8 are selected in order to evaluate the prediction performance for each soil type by using RMSE, as shown in table 6.7.

Table 6.5 Result of RMSE for equation 6.7 and 6.8

Validation data set	Number of data point	SPT-N,σ'_v model	GP SPT-N,σ'_v model
		Equation 6.7	Equation 6.8
Marine Deposit	32	45.81	36.24
Alluvium	34	58.37	47.58
All soil	66	52.66	42.46

According to table 6.5, equation 6.8 (GP SPT-N,σ'_v model) provides lower value of RMSE than equation 6.7 for those 3 different soil types. Therefore, the better predication performance is provided by equation 6.7 (GP SPT-N, σ'_v model) compare with other 3 SPT-N, σ'_v models.

In order to investigate the prediction performance statistically, the figures of predicted Vs versus measured Vs were plotted based on the SPT-N, σ'_v model (equation 6.7) and the GP SPT-N, σ'_v model (equation 6.8).

6.3.3 Result

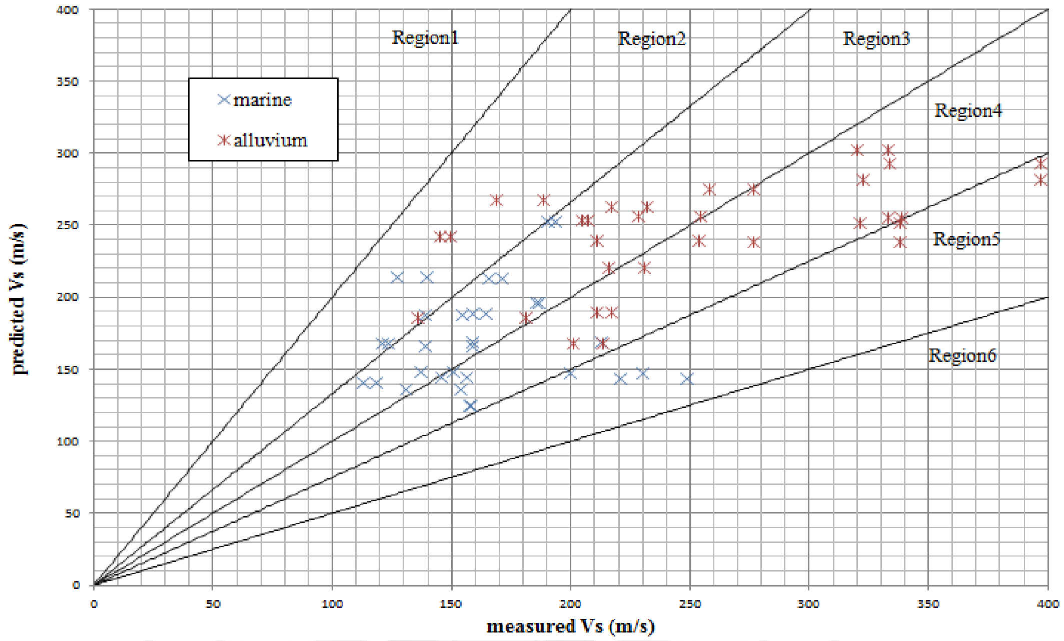


Figure 6.4 predicted Vs verses measured Vs for equation 6.7

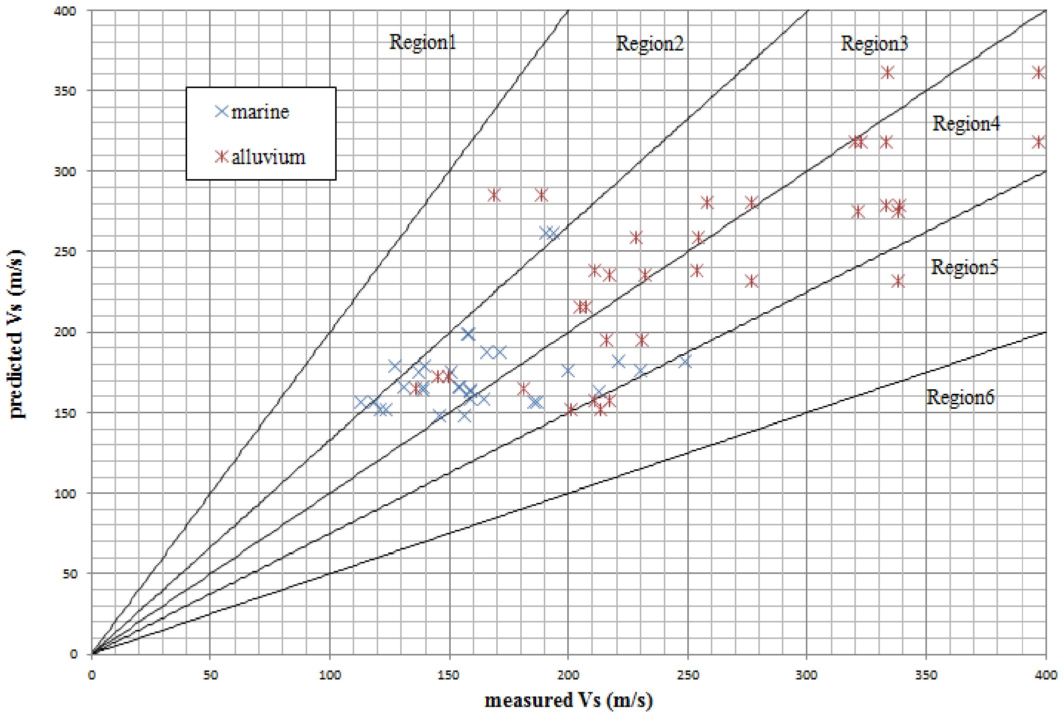


Figure 6.5 predicted Vs verses measured Vs for equation 6.8

Table 6.6 Distribution of data for equation 6.7

SPT-N,σ'_v model Equation 6.7	Marine Deposit		Alluvium	
Region	Number of data	Percentage of data (%)	Number of data	Percentage of data (%)
1	0	0.00	0	0.00
2	5	15.63	5	14.71
3	16	50.00	10	29.41
4	7	21.88	15	44.12
5	4	12.50	4	11.76
6	0	0.00	0	0.00
Percentage of over-predicted, %	65.63		44.12	
Percentage of under-predicted, %	34.38		55.88	
Percentage of within Region 3 and 4, %	71.88		73.53	

Table 6.7 Distribution of data for equation 6.8

GP SPT-N,σ'_v model Equation 6.8	Marine Deposit		Alluvium	
Region	Number of data	Percentage of data (%)	Number of data	Percentage of data (%)
1	0	0.00	0	0.00
2	4	12.50	2	5.88
3	18	56.25	13	38.24
4	9	28.13	15	44.12
5	1	3.13	4	11.76
6	0	0.00	0	0.00
Percentage of over-predicted, %	68.75		44.12	
Percentage of under-predicted, %	31.25		55.88	
Percentage of within Region 3 and 4, %	84.38		82.35	

For equation 6.7, the percentage of over-predicted for marine deposit equals to 65.63%, and 44.12% for alluvium. Therefore, this model is over-predicted for marine deposit and under-predicted for alluvium. Moreover, this model provided higher chance to obtain reasonable prediction for alluvium than marine deposit because the percentage of within region 3 and 4 for marine deposit equals to 71.88%, and 73.53% for alluvium.

For equation 6.8, the percentage of over-predicted for marine deposit equals to 68.75%, and 44.12% for alluvium. Therefore, this model is also over-predicted for marine deposit and under-predicted for alluvium. Moreover, this model provided higher chance to obtain reasonable prediction for marine deposit than alluvium because the percentage of within region 3 and 4 for marine deposit equals to 84.38%, and 82.35% for alluvium.

In comparing this 2 models based on table 6.6 and table 6.7, the result represented that equation 6.8 (GP model) provided higher prediction performance. In order to identify the improvement after the database of vertical effective stress combined with SPT-N to establish shear wave velocity, table 6.2, 6.3, 6.6 and 6.7 are taken into comparison. For equation 6.3 and 6.7, the effect of vertical effective stress cause bad prediction performance for marine deposit. For equation 6.4 and 6.8, the vertical effective stress cause bad prediction performance for alluvium. The prediction improvement is not provided significantly after the vertical effective stress is combined. In this case, the database of cone resistance is selected to compare the prediction improvement with vertical effective stress.

6.4 Correlation model of Shear Wave Velocity with SPT-N and cone resistance

Based on the section 6.3, the prediction performance is improved but no significant after the database of vertical effective stress combined with SPT-N to establish the new correlation models. In this section, the database of cone resistance is applied. It is used to combine with the database of SPT-N to establish the correlation models of shear wave velocity with SPT-N and cone resistance.

However, some significant variations were commonly obtained in the measurement of cone resistance, as mentioned in Chapter 2.5. Therefore, 4 different calculation methods of the equivalent average of cone resistance are introduced in below.

Method 1: the mean value of cone resistance between current depth and 1 m below;

Method 2: the mean value of cone resistance between 1 m above and 1 m below of current depth;

Method 3: the mean value of cone resistance between current depth and 0.5 m below;

Method 4: refer to Bustamante and Gianeselli method (Chapter 2.5.2) based on the pile diameter is equal to standard diameter of cone (3.57 cm).

6.4.1 Models

According to Chapter 5.3, the cone resistance can also form as a power-law regression to establish shear wave velocity instead of SPT-N. The corresponding correlation models based on different calculation method of cone resistance are shown in following.

q_c models:

Method 1:

$$V_s = 49.19q_c^{0.2053} \tag{6.9}$$

Method 2:

$$V_s = 39.4q_c^{0.2352} \tag{6.10}$$

Method 3:

$$V_s = 51.42q_c^{0.2003} \tag{6.11}$$

Method 4:

$$V_s = 54.63q_c^{0.1956} \tag{6.12}$$

where V_s = shear wave velocity, m/s; q_c = cone resistance, kPa

After the dataset of SPT-N and cone resistance are combined, Curve Fitting tool and Genetic Programming (GPTIPS1.0) are applied to establish the correlation models of shear wave velocity with SPT-N and cone resistance.

SPT-N,$q_c$model:

Method 1:

$$V_s = 53.81(N)^{0.2036}(q_c)^{0.1419} \tag{6.13}$$

Method 2:

$$V_s = 41.87(N)^{0.207}(q_c)^{0.1755} \tag{6.14}$$

Method 3:

$$V_s = 38.35(N)^{0.2211}(q_c)^{0.1866} \tag{6.15}$$

Method 4:

$$V_s = 41.48(N)^{0.2199}(q_c)^{0.1792} \tag{6.16}$$

GP SPT-N,$q_c$model:

Method 1:

$$V_s = 5.92N - \frac{3509(0.1389N^2+N+q_c+8.373)}{2(N+q_c)} + \frac{5.92N^2}{N+7.535} + 1931 \tag{6.17}$$

Method 2:

$$V_s = 8.313N + \frac{839.9q_c-4958}{4N+q_c+7.467} - \frac{28.41}{N^2} - 653.6 \tag{6.18}$$

Method 3:

$$V_s = 0.0226q_c - \frac{1.202q_c^2}{100000N} + 0.003765Nq_c + 148.3 \tag{6.19}$$

Method 4:

$$V_s = 0.03465q_c - \frac{1.507q_c^2}{100000N} + N(0.004133q_c - 0.00003013) + 139.8 \qquad (6.20)$$

The correlation models of shear wave velocity with cone resistance are expressed in equation 6.9 to 6.12. In order to compare the shape of the curve between these four different models, there were plotted in figure 6.6.

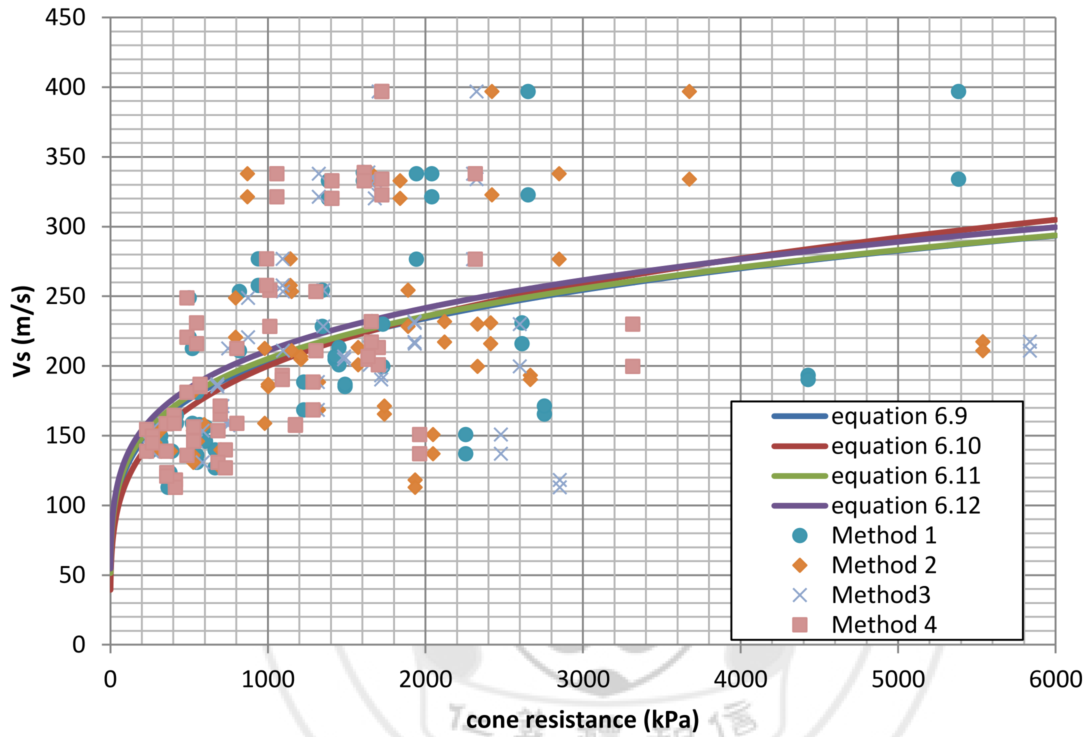


Figure 6.6 Correlation models of Shear Wave Velocity and cone resistance using Curve Fitting Tool

6.4.2 Model Evaluation

According to figure 6.6, the larger value of shear wave velocity is established by equation 6.10 compare with the other 3 correlation models, and the lower value of shear wave velocity is established by equation 6.9.

The database of cone resistance was calculated by 4 different methods. Each method is also possible to calculate outlier, as shown in figure 6.6. However, the outlier cannot be identified directly because the prediction influence is not significant as SPT-N database. Therefore, the consideration of outlier is not applied in database of cone resistance.

The prediction performance is hard to compare directly between these 4 different methods because the correlation models are similar. Therefore, coefficient of determination (R^2) is applied to identify which calculation of cone resistance can provide the best prediction performance. After that, these 4 different databases of cone resistance are combined with SPT-N to establish shear wave velocity. Finally, Genetic Programming is applied to establish the shear wave velocity in order to obtain the improvement of the correlation models. The comparisons between different formats of modeling with different calculation method of cone resistance are shown in table 6.8.

Table 6.8 Result of R^2 for equation q_c Models, SPT-N,q_c Models and GP SPT-N,q_c Models

Model	Equation	Calibration data set	Validation data set	R^2
q_c	6.9	All soil (Method 1)	All soil	0.2974
	6.10	All soil (Method 2)	All soil	0.276
	6.11	All soil (Method 3)	All soil	0.2002
	6.12	All soil (Method 4)	All soil	0.2263
SPT-N,q_c Model	6.13	All soil (Method 1)	All soil	0.5548
	6.14	All soil (Method 2)	All soil	0.5704
	6.15	All soil (Method 3)	All soil	0.583
	6.16	All soil (Method 4)	All soil	0.5812
GP SPT-N,q_c	6.17	All soil (Method 1)	All soil	0.630
	6.18	All soil (Method 2)	All soil	0.606
	6.19	All soil (Method 3)	All soil	0.641
	6.20	All soil (Method 4)	All soil	0.628

In comparing between q_c models, the highest value of R^2 is provided by equation 6.9. Therefore, this correlation model provided based on the database of cone resistance is calculated by method 1. On the other hand, equation 6.11 provided the lowest prediction performance because the lowest value of R^2 is provided based on the database of cone resistance was calculated by method 3.

In comparing between SPT-N,q_c models, the highest value of R^2 is provided by equation 6.15. Therefore, this correlation model provided the highest prediction performance based on the database of cone resistance was calculated by method 3. On the other hand, the lowest prediction performance is provided by equation 6.13 because the lowest R^2 is provided based on the cone resistance was calculated by method 1.

In comparing between GP SPT-N,q_c models, the highest value of R^2 is provided by equation 6.19. Therefore, this correlation model provided the highest prediction performance based on the cone resistance was calculated by method 3. On the other hand, the lowest prediction performance is provided by equation 6.18 because the lowest R^2 is provided based on the cone resistance was calculated by method 2.

In comparing between the overall models in table 6.8, the prediction performance is successfully improved after the genetic programming was applied because the highest value of R^2 is provided by equation 6.19 (GP SPT-N,q_c model). Moreover, the database of cone resistance was calculated by method 3 the database is suitable to combined with the database of SPT-N value to establish the shear wave velocity based on it commonly performed the better prediction performance of shear wave velocity.

In order to identify the improvement of GP model, equation 6.15 and equation 6.19 are selected to compare the prediction performance for each soil type by using RMSE, as shown in table 6.9.

Table 6.9 Result of RMSE for Equation 6.15 and 6.19

Validation data set	Number of dataset	SPT-N,q_c Model	GP SPT-N,q_c Model
		Equation 6.15	Equation 6.19
Marine Deposit	32	40.402	33.197
Alluvium	34	51.481	51.938
All soil	66	46.440	43.863

According to table 6.9, the RMSE of equation 6.19 (GP SPT-N,q_c model) is less than equation 6.15 (SPT-N,q_c Model) for marine deposit and all soil, but not for alluvium. Therefore, the better predication performance is provided by equation 6.19 (GP SPT-N, q_c Model) for marine deposit and all soil, but not for alluvium. In comparing the overall prediction performances for equation 6.15 and 6.19, cone resistance is suitable applied with SPT-N value to form as power-law regression model to establish shear wave velocity.

On the other hand, the better prediction performance is provided by equation 6.15 (SPT-N, q_c model) compare with equation 6.7 (SPT-N, σ'_v model) because the corresponding RMSE of equation 6.7 is larger than equation 6.15. It implies SPT-N contain higher correlation relationship with cone resistance rather than vertical effective stress.

In order to investigate the prediction performance statistically, predicted Vs versus measured Vs is plotted based on the SPT-N, q_c Model (equation 6.15) and the corresponding GP SPT-N,q_c model (equation 6.19).

6.4.3 Result

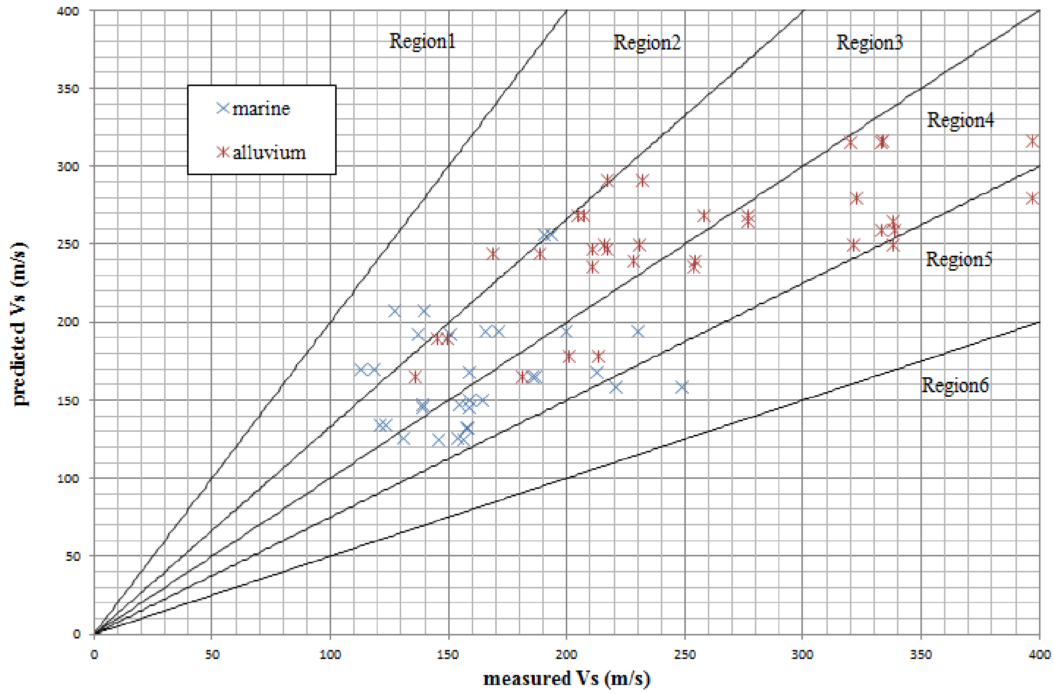


Figure 6.7 predicted Vs verses measured Vs for equation 6.15

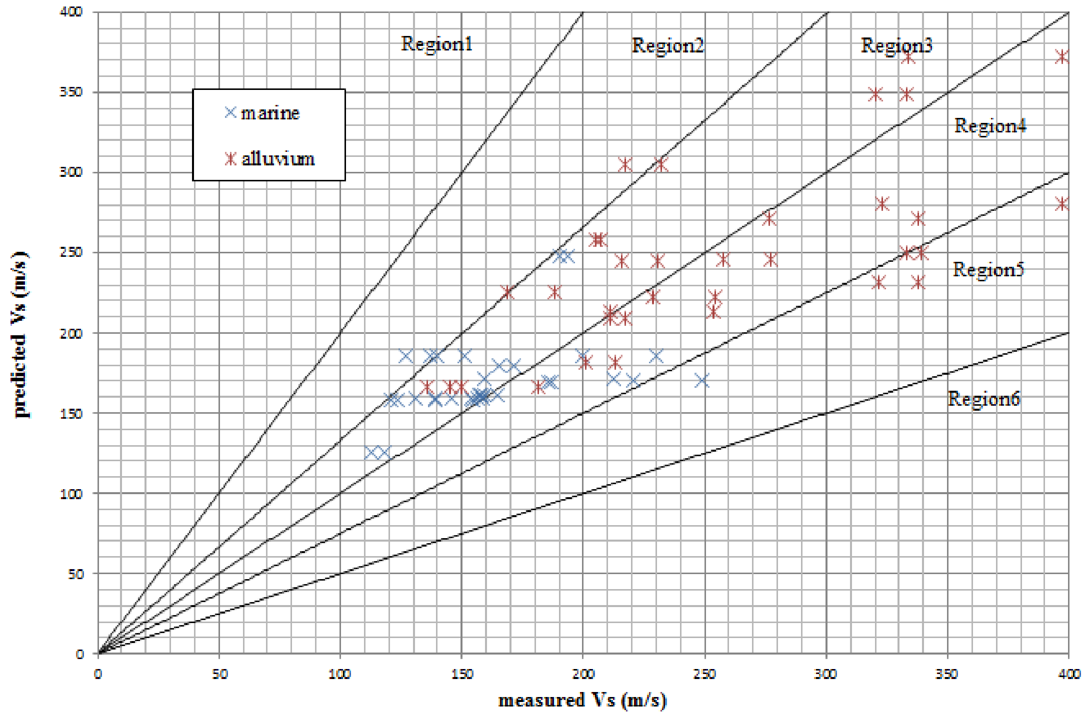


Figure 6.8 predicted Vs verses measured Vs for equation 6.19

Table 6.10 Distribution of data for equation 6.15

SPT-N,q_c Model Equation 6.15	Marine Deposit		Alluvium	
Region	Number of data	Percentage of data (%)	Number of data	Percentage of data (%)
1	0	0.00	0	0.00
2	6	18.75	2	5.88
3	9	28.13	14	41.18
4	15	46.88	16	47.06
5	2	6.25	2	5.88
6	0	0.00	0	0.00
Percentage of over-predicted, %	46.88		47.06	
Percentage of under-predicted, %	53.13		52.94	
Percentage of within Region 3 and 4, %	75.00		88.24	

Table 6.11 Distribution of data for equation 6.19

GP SPT-N,q_c Model Equation 6.19	Marine Deposit		Alluvium	
Region	Number of data	Percentage of data (%)	Number of data	Percentage of data (%)
1	0	0.00	0	0.00
2	2	6.25	2	5.88
3	22	68.75	13	38.24
4	7	21.88	15	44.12
5	1	3.13	4	11.76
6	0	0.00	0	0.00
Percentage of over-predicted, %	75.00		44.12	
Percentage of under-predicted, %	25.00		55.88	
Percentage of within Region 3 and 4, %	90.63		82.35	

For equation 6.15, the percentage of over-predicted for marine deposit equals to 46.88%, and 47.06% for alluvium. Therefore, this model is over-predicted for alluvium and under-predicted for marine. Moreover, this model provided higher chance to obtain reasonable prediction for alluvium than marine deposit because the percentage of within region 3 and 4 equals to 75.00% for marine deposit, and 88.24% for alluvium.

For equation 6.19, the percentage of over-predicted for marine deposit equals to 75.00%, and 44.12% for alluvium. Therefore, this model is over-predicted for marine deposit and under-predicted for alluvium. Moreover, this model provided higher chance to obtain reasonable prediction for marine deposit than alluvium because the percentage of within region 3 and 4 equals to 90.63%for marine deposit, and 82.35% for alluvium.

To compare this 2 equations based on table 6.10 and table 6.11, the result indicate that equation 6.19 (GP model) has higher prediction performance for marine deposit but weaker performance for alluvium by compare with equation 6.15. For all soil points of view, improvement of prediction is provided by equation 6.19 based on higher percentage are provided to establish the reasonable result of shear wave velocity compare with equation 6.15.

The effect of vertical effective stress and cone resistance are identified based on comparison between table 6.6, 6.7, 6.10 and 6.11. In comparing between table 6.6 and 6.10, equation 6.15 provided improvement of prediction performance for marine deposit and alluvium based on the power law functions. For table 6.7 and 6.11, equation 6.19 provided improvement of prediction performance for marine deposit only based on the GP model. Therefore, cone resistances provide higher correlation relationship between SPT-N and shear wave velocity than vertical effective stress.

6.5 Correlation model of Shear Wave Velocity with SPT-N, vertical effective stress and cone resistance

6.5.1 Models

Based on the previous section, it mentions that the cone resistance provided the improvement for the prediction of shear wave velocity. Since both vertical effective stress and cone resistance available to provide improvement for prediction of shear wave velocity, database of these two soil parameters were combined at the same time with SPT-N value to establish new correlation models. "NonLinearModel.fit" function and Genetic Programming (GPTIPS1.0) were applied in this study. Those 4 calculation methods of cone resistance are evaluate again based on the new format of correlation models.

SPT-N,σ'_v, q_c Model:

Method 1:

$$V_s = 40.744(N)^{0.184}(\sigma'_v)^{0.074}(q_c)^{0.136} \tag{6.21}$$

Method 2:

$$V_s = 31.357(N)^{0.186}(\sigma'_v)^{0.078}(q_c)^{0.16855} \tag{6.22}$$

Method 3:

$$V_s = 23.902(N)^{0.187}(\sigma'_v)^{0.122}(q_c)^{0.122} \tag{6.23}$$

Method 4:

$$V_s = 23.413(N)^{0.17824}(\sigma'_v)^{0.145}(q_c)^{0.172} \tag{6.24}$$

GP SPT-N,σ'_v, q_cModel :

Method 1:

$$V_s = 0.007279\sigma'^2_v - \frac{0.00652(\sigma'_v - 5.631)^2}{N} - \frac{0.0001349\sigma'^4_v}{q_c - 1.845} + 140.7 \tag{6.25}$$

Method 2:

$$V_s = \frac{0.06422q_c}{\sigma'_v} - 0.06422N(2N - \sigma'_v + 7.314) - \frac{926(N+\sigma'_v+q_c-0.9382)}{2q_c-N+\sigma'_v+7.314} + 644 \quad (6.26)$$

Method 3:

$$V_s = N\left(\frac{3.112q_c\sigma'_v}{100000} - 1.094\right) - 0.00003112(\sigma'_v + N^2q_c) + \frac{3.017}{N-10.97} + 162 \quad (6.27)$$

Method 4:

$$V_s = 0.002148q_c + \frac{5.279Nq_c\sigma'_v}{100000} - \frac{2.169N(Nq_c-24.56)(N+q_c+20.34)}{100000\sigma'_v} + 154.5 \quad (6.28)$$

6.5.2 Model Evaluation

In order to compare the difference and prediction performance between correlation models, coefficient of determination (R^2) is applied to identify which calculation of cone resistance can provided the best correlation model. After the database of SPT-N, vertical effective stress and cone resistance were combined to form as power-law correlation functions. Moreover, new correlation models are established after genetic programming was applied. The result of R^2 for those is compared in table 6.12.

Table 6.12 Result of R^2 for SPT-N,σ'_v, q_c models and GP SPT-N,σ'_v, q_c models

Model	Equation	Calibration data set	Validation data set	R^2
SPT-N,σ'_v, q_c	6.21	All soil (Method 1)	All soil	0.557
	6.22	All soil (Method 2)	All soil	0.572
	6.23	All soil (Method 3)	All soil	0.588
	6.24	All soil (Method 4)	All soil	0589
GP SPT-N,σ'_v, q_c	6.25	All soil (Method 1)	All soil	0.631
	6.26	All soil (Method 2)	All soil	0.631
	6.27	All soil (Method 3)	All soil	0.741
	6.28	All soil (Method 4)	All soil	0.665

In comparing between SPT-N,σ'_v, q_c models, the highest value of R^2 is provided by equation 6.24. Therefore, this correlation model provided based on the database of cone resistance is calculated by method 4. On the other hand, equation 6.21 provided the lowest prediction performance because the lowest value of R^2 is provided based on the database of cone resistance was calculated by method 1.

In comparing between GP SPT-N, σ'_v, q_c models, the highest value of R^2 is provided by equation 6.27.Therefore, this correlation model provided based on the database of cone resistance is calculated by method 3. On the other hand, equation 6.14 provided the lowest prediction performance because the lowest R^2 is provided based on the database of cone resistance was calculated by method 1.

Based on the result of R^2 from table 6.12, the prediction performance can be improved significantly if the correlation model combined with SPT-N, vertical effective stress and cone resistance at the same time. Moreover, the highest prediction performance after the genetic programming was applied.

In order to identify the improvement of GP model, equation 6.23 and equation 6.27 are selected to compare the prediction performance for each soil type by using RMSE, as shown in table 6.16.

Table 6.13 Result of RMSE for Equation 6.24 and 6.28

Validation data set	Number of dataset	SPT-N, σ'_v, q_c Model	GP SPT-N, σ'_v, q_c Model
		Equation 6.23	Equation 6.27
Marine Deposit	32	41.11	34.14
Alluvium	34	50.44	38.72
All soil	66	46.15	36.57

Based on table 6.13, equation 6.27 (GP SPT-N, σ'_v, q_c model) provided less value of RMSE than equation 6.23 (SPT-N, σ'_v, q_c model) respect to those 3 different validation data set. Therefore, GP model provided higher predication performance compare with the power-law correlation model.

After vertical effective stress and cone resistance were combined with SPT-N to establish shear wave velocity, the prediction performance for corresponding models become higher if more reliable soil parameter are combined to form a correlation function. In comparing between the table 6.1, 6.5, 6.8, and 13, the lowest RMSE is calculated based on GP SPT-N,σ'_v, q_c model (equation 6.27). Therefore, either the more reliable soil parameters applied or Genetic Programming apply, the higher prediction performance is also obtained.

In order to investigate the prediction performance statistically, the figure of predicted Vs versus measured Vs is plotted based on the equation 6.23 (SPT-N,σ'_v, q_c Model) and the equation 6.27 (GP SPT-N,σ'_v, q_c Model).

6.5.3 Result

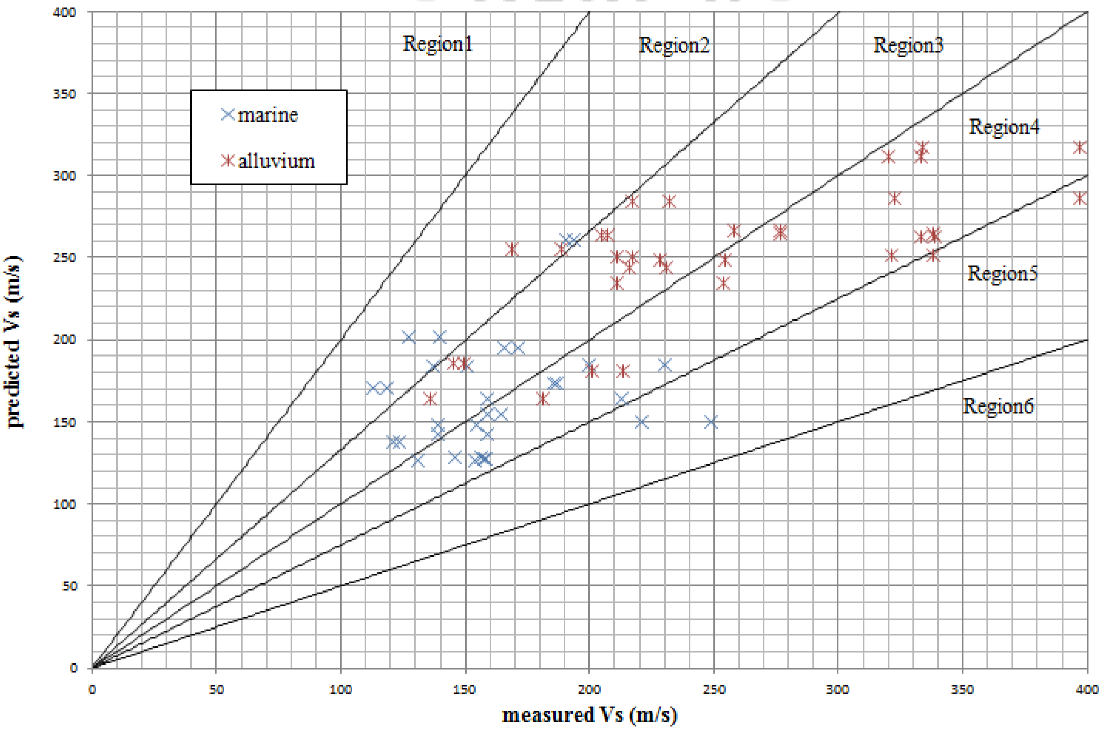


Figure 6.9 predicted Vs verses measured Vs for equation 6.23

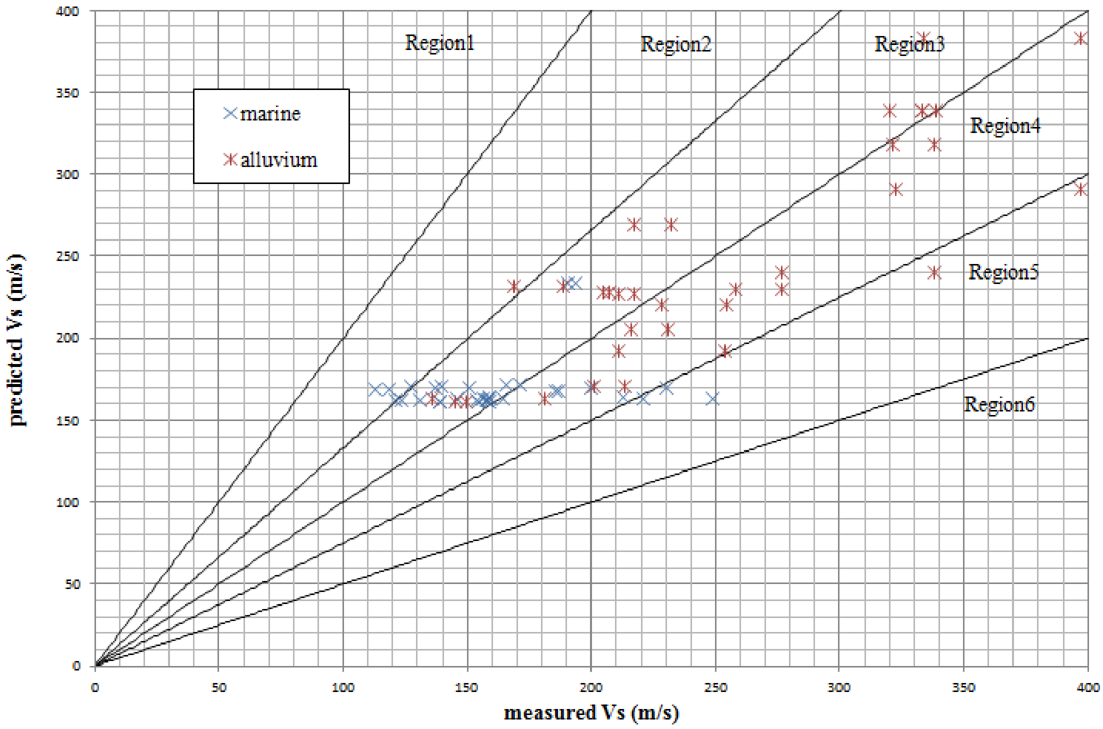


Figure 6.10 predicted Vs verses measured Vs for equation 6.27

Table 6.14 Distribution of data for equation 6.23

SPT-N,σ'_{v}, q_c Model Equation 6.23	Marine Deposit		Alluvium	
Region	Number of data	Percentage of data (%)	Number of data	Percentage of data (%)
1	0	0.00	0	0.00
2	7	21.88	2	5.88
3	8	25.00	14	41.18
4	15	46.88	16	47.06
5	2	6.25	2	5.88
6	0	0.00	0	0.00
Percentage of over-predicted, %	46.88		47.06	
Percentage of under-predicted, %	53.13		52.94	
Percentage of within Region 3 and 4, %	71.88		88.24	

Table 6.15 Distribution of data for equation 6.27

GP SPT-N, σ'_v, q_c Model Equation 6.27	Marine Deposit		Alluvium	
Region	Number of data	Percentage of data (%)	Number of data	Percentage of data (%)
1	0	0.00	0	0.00
2	4	12.50	1	2.94
3	19	59.38	14	41.18
4	6	18.75	17	50.00
5	3	9.38	2	5.88
6	0	0.00	0	0.00
Percentage of over-predicted, %	71.88		44.12	
Percentage of under-predicted, %	28.13		55.88	
Percentage of within Region 3 and 4, %	78.13		91.18	

For equation 6.23, the percentage of over-predicted for marine deposit equals to 46.88%, and 47.06% for alluvium. Therefore, this model is over-predicted for alluvium and under-predicted for marine deposit. Moreover, this model provided higher chance to obtain reasonable prediction for alluvium than marine deposit because the percentage of within region 3 and 4 for marine deposit equals to 71.88%, and 88.24% for alluvium.

For equation 6.27, the percentage of over-predicted for marine deposit equals to 71.88%, and 44.12% for alluvium. Therefore, this model is over-predicted for marine deposit and under-predicted for alluvium. Moreover, this model provided higher chance to obtain reasonable prediction for alluvium than marine deposit because the percentage of within region 3 and 4 for marine deposit equals to 78.13%, and 91.18% for alluvium.

In comparing this 2 equations based on table 6.14 and table 6.15, the highest prediction performance is provided by equation 6.27 (GP model) based on higher percentage to establish the reasonable result of shear wave velocity is provided.

To compare the result with overall correlation models, the prediction performance is improved based on the additional soil parameters combined to form correlation models. However, the highest prediction model may only provide best prediction performance for all soil, but it cannot be standardized as the best fit for each specific soil layer.

6.6 Summary

The database of LRT-C250 project is selected to establish the correlation models of shear wave velocity in this study. In order to establish various categorizes of correlation models based on this database, curve fitting tool, "NonLinearModel.fit" function and genetic programming (GPTIPS1.0) were applied. On the other hand, Coefficient of determination (R^2), Root Mean Square Error (RMSE) and the plots between predicted Vs verses measured Vs were applied to evaluate prediction performance of the correlation models.

There are total 28 correlation models were established and separated into 4 different categorizes, which are represented to section 6.2 to 6.5, respectively.

In section 6.2, the shear wave velocity for marine deposit contains poor relation of power law regression with SPT-N. The prediction performance was improved after the genetic programming applied, however, this correlation model (equation 6.4) still cannot be applied with confidence. It may be caused by the limitation of the database, unnatural soil behavior of marine deposit and the influence of other soil parameters.

In section 6.3, the database of vertical effective stress was combined with SPT-N to establish new correlation model of shear wave velocity. The prediction performances were not improved significantly. The vertical effective stress caused bad effect for marine deposit based on power-law correlation model (equation 6.7).The prediction performances also improved after the genetic programming was applied.

In section 6.4, the database of cone resistance was combined with SPT-N to establish new correlation model of shear wave velocity. Since the database of cone resistance contains large variation in field measurement, there are 4 different calculation methods provided to calculate the equivalent average cone resistance in order to obtain the best correlation model. Based on the result, the database of cone resistance calculated by method 3 was commonly provided the best fit correlation model. However, Method 3 is not represented the best database for any format of correlation model. Therefore, it cannot be standardized which method can provide the best correlation model. The result was highly affected by the formation of model, site condition, soil classification and reliability of database. Moreover, the prediction improvement for combine with cone resistance was significant compare with vertical effective stress.

In section 6.5, vertical effective stress and cone resistance can provide improvement of prediction of shear wave velocity. Therefore, the database of these two soil parameters are combined with SPT-N at the same time in order to establish new correlation model so that the highest prediction performance can be achieved. To compare the result with overall correlation models, the prediction performance was improved based on the vertical effective stress and cone resistance combined to form correlation models.

Table 6.16 the result of RMSE of the best correlation models of each section

Section	Model	Equation	RMSE		
			Marine Deposit	Alluvium	All soil
6.2	SPT-N	6.3	42.30	62.43	53.89
	GP SPT-N	6.4	39.85	52.29	46.67
6.3	SPT-N,σ'_v	6.7	45.81	58.37	52.66
	GP SPT-N,σ'_v	6.8	36.24	47.58	42.46
6.4	SPT-N,q_c	6.15	40.40	51.48	46.44
	GP SPT-N,q_c	6.19	31.12	51.94	43.86
6.5	SPT-N,σ'_v, q_c	6.23	41.11	50.44	46.15
	GP SPT-N,σ'_v, q_c	6.27	34.58	47.35	41.65

According to table 6.16, GP SPT-N,q_c model have highest prediction performance for marine deposit because the value of RMSE is lowest for Marine Deposit, on the other hand, the highest prediction performance for alluvium and all soil were established by GP SPT-N,σ'_v, q_c Model. As result, one single correlation model provides the best fit of correlation model for all soil, but it cannot provide the best fitting for one specific soil layer. Therefore, one single correlation model might not apply with confidence in every specific soil type.

CHAPTER 7 CONCLUSION

7.1 Conclusion and Findings

In this study, nonlinear curve fitting and genetic programming were applied to establish the correlation models of shear wave velocity with SPT-N, vertical effective stress and cone resistance.

The new correlation models were established in this study based on the LRT C-250 project. After the outliers were eliminated, the newest existing database in this study contains total 66 data points (Marine Deposit: 32, Alluvium: 34). In order to compare the prediction performance for each models, Coefficient of determination (R^2), root mean square error (RMSE) and the plot of predicted versus measured were applied to identify the which model provide best fit function and evaluate the corresponding prediction performance.

Matlab software was applied in this study instead of using Excel software. The reason is power-law regression is classified as nonlinear curve function, since Excel is using linear curve fitting method to establish the nonlinear correlation model, that means the power-law function need to transform to linear function so that it can be solved by using linear curve fitting. Based on the model comparison between linear curve fitting and nonlinear curve fitting, the higher value of shear wave velocity was generally performed by nonlinear curve fitting than linear curve fitting when SPT-N is large, that means Matlab correlation model have higher prediction performance when SPT-N value is large, however, it may not be identified which method provided better prediction performance when SPT-N is small. Therefore, the result of RMSE is applied to evaluate the correlation models between linear curve fitting and nonlinear curve fitting. After that, the lowest value of RMSE was performed by nonlinear curve fitting model so that Matlab software was selected to apply in this study.

After each correlation models with different format or different soil parameter were established in this study, there are some finding can be observed:

- The improvement of prediction performance can be observed after the reliable soil parameters combined with SPT-N to establish new correlation models.
- To comparing between vertical effective stress and cone resistance, the higher prediction performance is provided by cone resistance rather than vertical effective stress for combined with SPT-N to establish shear wave velocity.
- GP SPT-N, qc Model indicated highest prediction performance for Marine Deposit because the value of RMSE is lowest for Marine Deposit.
- GP SPT-N, σ'_v, q_c model provided the highest prediction performance for alluvium and all soil, however, it may not provide the best prediction for marine deposit. It indicates one single correlation model cannot apply to practice with confidence for each specific soil type.

7.2 Recommendation and Future work

The prediction performance was successfully improved after genetic programming applied to establish shear wave velocity. At the same time, the complexity of the correlation models also increased relatively and the format of correlation models cannot be assigned directly as simple power law function. Since this type of model is not based on any theoretical relationship, it is highly depend on the dataset is reliable or not. Therefore, the precision of measurement is necessary to be guarantee and make the correlation model more convincing by providing the sufficient number of reliable database. In actual situation, there are contain many uncertainly for each soil types, 66 data points may not provide representative and applicable correlation model in Macau. The additional field testing is required in order to gain more dataset so that the correlation models can keep updating and become more applicable in Macau.

REFERENCES

Amoroso, S. (2013), “Prediction of the shear wave velocity Vs from CPT and DMT”, In: Cui Y J, eds, Proceedings of the 5th International Young Geotechnical Engineers’ Conference-5th iYGEC 2013.

Anbazhagan, P., Kumar, A., Sitharam, T. G. (2012), “Seismic Site Classification and Correlation between Standard Penetration Test N Value and Shear Wave Velocity for Lucknow City”, Indo-Gangetic Basin, Pure and Applied Geophysics.

Andrus, R. D., Hayati, H., Mohanan, N. P. (2009), “Correlation Liquefaction Resistance for Aged Sands Using Measured to Estimated Velocity Ratio”, Journal of Geotechnical and Geoenvironmental Engineering, Vol. 135(6), pp. 735-744.

Barkhordari, K., Nayeri, G. D., Nayeri, D. D. (2013), “*A New Statistical Correlation between Shear Wave Velocity and Penetration Resistance of Soils Using Genetic Programming*”, Electronic Journal of Geotechnical Engineering, Vol.18, pp. 2071-2078.

Brandenberg, S. J., Bellana, N., Shantz, T. (2010), “Shear Wave Velocity as Function of Standard Penetration Test Resistance and Vertical Effective Stress at California Bridge Sites”, Soil Dynamics and Earthquake Engineering, Vol.30, pp.1026-1035.

Chao, S.C., (2010). *Investigation of Shear Wave Velocity around an Excavation-Correlation with Void Ratio and Effective Stress. Final Year Report*, University of Macau.

Civil Engineering Consultants Co. Ltd. (2010). *Soil Investigation Report S-CPT*, NGAI SHUN ENGINEERING CO. LTD

Holzer, T. L., Bennett, M. J., Noce, T. E., Padovani, A. C., Tinsley III, J. C. (2005), "Shear Wave Velocity of Surficial Geologic Sediments: Statistical Distributions and Depth Dependence", Earthquake Spectra, Vol. 21(1), pp. 61-77.

Jafari, M. K., Shafiee, A., Razmkhah, A. (2002), "Dynamic Properties of Fine Grained Soils in South of Tehran", Journal of Seismological Earthquake Engineering, Vol. 4, pp. 25-35.

Kuan, W.K. (2011), *Correlation between shear wave velocity with penetration tests, Finial Year Report*, University of Macau, Macau.

Kulhawy, F. H., Orchant, C. J., Trautmann, C. H. (1988), *Reliability-based foundation design for transmission line structures: critical evaluation of in situ test methods,* Electric Power Research Institute, Palo Alto, California.

Lunne, T., Robertson, P. K., Powell, J. J. M. (1997), *Cone Penetration Testing in Geotechnical Practice*, BLAACKIE ACADEMIC & PROFESSIONAL, pp.153

Marto, A., Soon, T. C., Kasim, F. (2013), "*A Correlation of Shear Wave Velocity and Standard Penetration Resistance*", Electronic Journal of Geotechnical Engineering, Vol.13, pp. 463-471.

Phoon, K. K. (2008), *Reliability-Based Design in Geotechnical Engineering Computations and Applications*, London and New York.

Phoon, K. K., Kulhawy, F. H. (1999)[a], "*Characterization of Geotechnical variability*", Canadian Geotechnical Journal, pp.612-624.

Phoon, K. K., Kulhawy, F. H. (1999)[b], "*Evaluation of Geotechnical Property variability*", Canadian Geotechnical Journal, pp.625-639.

Pun, H. K. (2010). *Investigation of Shear Wave Velocity around an Excavation-Correlation with SPT-N value. Final Year Report,* University of Macau.

Sae, T. (2009). *In-situ shear wave measurement and ground response analysis for developing site-dependent response spectra in Macau*, M.S. thesis, University of Macau, pp.35-80.

Searson, D., (2009). *GPTIPS-Genetic Programming & Symbolic Regression for MATLAB User Guide.*

Sykora, D. W., Koester, J. P. (1988), "Review of existing correlations between shear wave velocity or shear modulus and standard penetration resistance in soils", In: Proceedings of the earthquake engineering and soil dynamics II conference, Park City, UT, pp. 389-404.

APPENDIX A. Existing Database of LRT-C250

Borehole	Soil Type	depth(m)	SPT-N value	Shear wave velocity (m/s)	
				down-hole	S-CPT
DH1	Marine Deposit	4	2	137.08	151.00
	Marine Deposit	6	1	118.41	113.00
	Marine Deposit	8	2	123.49	121.00
	Marine Deposit	10	3	164.79	159.00
	Marine Deposit	12	3	185.28	187.00
DH2	Marine Deposit	4	1	157.55	158.00
	Marine Deposit	6	1	153.68	131.00
	Marine Deposit	8	1	156.51	146.00
	Alluvium	10	2	213.45	201.00
	Alluvium	12	3	217.41	211.00
	Alluvium	14	9	276.65	338.00
	Alluvium	16	11	338.89	333.00
	Alluvium	18	20	334.00	397.00
DH3	Marine Deposit	5	2	220.74	249.00
	Marine Deposit	7	3	158.78	139.00
	Alluvium	9	4	181.29	136.00
	Alluvium	11	8	230.97	216.00
	Alluvium	13	10	253.57	211.00
	Alluvium	15	11	321.51	338.00
	Alluvium	17	26	320.28	333.00
	Alluvium	19	15	322.92	397.00
DH4	Marine Deposit	5	14	108.93	143.60
	Marine Deposit	7	57	131.73	153.20
	Marine Deposit	9	8	139.85	127.00
	Marine Deposit	11	6	165.59	171.30
	Marine Deposit	13	39	122.18	157.60
	Marine Deposit	15	10	193.43	190.30
	Alluvium	17	9	254.44	228.40
	Alluvium	19	10	168.60	188.50
DH5	Marine Deposit	5	2	199.91	230.00
	Marine Deposit	7	3	212.57	159.00
	Marine Deposit	9	4	154.55	139.00
	Alluvium	11	13	149.72	145.00
	Alluvium	13	14	204.86	207.00
	Alluvium	15	16	217.33	232.00
	Alluvium	17	18	257.98	277.00

Table A.1 database of SPT-N value and shear wave velocity in LRT-C250

Borehole	Soil Type	depth(m)	Vertical effective stress, kPa
DH1	Marine Deposit	4	65.31
	Marine Deposit	6	85.75
	Marine Deposit	8	105.76
	Marine Deposit	10	125.77
	Marine Deposit	12	145.79
DH2	Marine Deposit	4	53.86
	Marine Deposit	6	75.07
	Marine Deposit	8	95.99
	Alluvium	10	106.22
	Alluvium	12	127.94
	Alluvium	14	146.53
	Alluvium	16	166.93
	Alluvium	18	187.12
DH3	Marine Deposit	5	58.72
	Marine Deposit	7	78.01
	Alluvium	9	96.87
	Alluvium	11	117.09
	Alluvium	13	137.72
	Alluvium	15	157.65
	Alluvium	17	176.68
	Alluvium	19	196.46
DH4	Marine Deposit	5	64.19
	Marine Deposit	7	83.51
	Marine Deposit	9	104.39
	Marine Deposit	11	125.41
	Marine Deposit	13	147.70
	Marine Deposit	15	169.29
	Alluvium	17	191.50
	Alluvium	19	211.87
DH5	Marine Deposit	5	63.67
	Marine Deposit	7	82.70
	Marine Deposit	9	102.33
	Alluvium	11	120.79
	Alluvium	13	136.53
	Alluvium	15	144.13
	Alluvium	17	159.23

Table A.2 database of vertical effective stress in LRT-C250

Borehole	Soil Type	depth(m)	Average cone resistance kPa			
			Method 1	Method 2	Method 3	Method
DH1	Marine Deposit	4	2255.19	2047.98	2479.86	1963.10
	Marine Deposit	6	365.53	1935.72	2852.04	412.71
	Marine Deposit	8	378.31	362.05	356.19	358.19
	Marine Deposit	10	419.72	405.29	401.78	405.24
	Marine Deposit	12	1490.83	1001.99	679.44	569.35
DH2	Marine Deposit	4	566.15	596.33	747.50	1175.34
	Marine Deposit	6	549.22	526.37	586.49	683.90
	Marine Deposit	8	603.24	553.16	542.53	529.28
	Alluvium	10	1451.60	1573.42	1645.45	1702.07
	Alluvium	12	7209.78	5539.18	5836.55	7348.11
	Alluvium	14	1942.05	2847.93	2302.48	2316.88
	Alluvium	16	1605.42	1650.26	1637.51	1611.00
	Alluvium	18	5383.99	3675.03	2324.21	1724.47
DH3	Marine Deposit	5	499.73	795.82	874.46	486.65
	Marine Deposit	7	389.90	383.17	344.43	354.71
	Alluvium	9	549.14	525.56	485.10	485.62
	Alluvium	11	2613.73	2414.89	1932.11	546.34
	Alluvium	13	819.36	1150.53	1096.62	1304.34
	Alluvium	15	2040.05	870.57	1324.05	1056.29
	Alluvium	17	1384.73	1839.44	1678.99	1406.66
	Alluvium	19	2651.22	2422.55	1702.58	1724.21
DH4	Marine Deposit	5.00	3303.88	1911.15	745.49	512.07
	Marine Deposit	7.00	860.64	868.34	788.55	897.80
	Marine Deposit	9	664.49	702.42	726.22	730.37
	Marine Deposit	11	2755.90	1739.83	711.66	698.25
	Marine Deposit	13.00	787.29	1302.77	799.86	737.38
	Marine Deposit	15	4429.92	2665.98	1720.30	1092.46
	Alluvium	17	1345.87	1890.29	1352.66	1012.92

	Alluvium	19	1228.04	1321.65	1316.33	1286.98
DH5	Marine Deposit	5	1729.75	2330.61	2599.29	3316.25
	Marine Deposit	7	520.16	980.25	747.57	804.28
	Marine Deposit	9	253.17	324.57	264.76	229.45
	Alluvium	11	319.07	282.65	250.68	266.52
	Alluvium	13	1426.93	1210.36	1482.94	1635.95
	Alluvium	15	1661.15	2121.24	1930.15	1657.31
	Alluvium	17	940.36	1143.36	1093.55	993.01

Table A.3 Database of average cone resistance for LRT-C250