Developing SPT-CPT Correlation Models Using Hierarchical Bayesian Approach

Sara Khoshnevisan(✉) and Laith Sadik

University of Cincinnati, Cincinnati, OH, USA
sara.khoshnevisan@uc.edu

Abstract. In geotechnical practice, engineers often perform only one type of in-situ testing. However, under some circumstances, there might be a need for different testing for additional analysis. Having a correlation model in such cases eliminates the need for performing additional testing and thus, saving time and the associated costs. Over years, many researchers have developed models between different in-situ testing methods, most of which are based on regression analysis using data from different regions. However, these developed models do not account for the potential spatial variability between the regions data is taken from. Moreover, the applicability of these models is questionable in a new region; especially where no data or limited data is available. In addition, most of the developed models do not account for uncertainties. In this paper, the Hierarchical Bayesian Modeling approach is adopted to develop region-specific correlation models between two popular in-situ testing methods: The Standard Penetration Test and Cone Penetration Test. 220 high-quality data pairs of $N_{1,60\,cs}$ and $q_{t1N,cs}$ from six regions in Taiwan are used for illustration purposes. The developed model is validated using a new region that was not used for model development. Two $N_{1,60\,cs}$ - $q_{t1N,cs}$ existing correlation models are also adopted for comparison purposes. The models developed using the Hierarchical Bayesian Modeling approach are shown to perform better compared to these existing correlation models. In addition, the hierarchical proposed approach shows a strong prediction capability and high reliability in regions with limited or no data. With the proposed approach, region-specific $N_{1,60\,cs}$ - $q_{t1N,cs}$ correlation models can be developed for regions with limited or limited data.

Keywords: Hierarchical Bayesian Modeling · Standard Penetration Test · Cone Penetration Test

1 Introduction

In geotechnical practice, it is common for engineers to only perform one type of in-situ testing. However, in some situations, it may be necessary to perform additional testing to gather more information. Having a correlation model eliminates the need for additional testing. Accordingly, a lot of research effort has been put towards development of reliable correlation models; many of which are based on frequentist analysis using data from different regions.

R. E. Hammah et al. (Eds.): RIC 2023, AHE 19, pp. 655–665, 2023.
https://doi.org/10.2991/978-94-6463-258-3_61

However, frequentist statistics does not consider the uncertainties, which can be an issue especially when limited data is available and uncertainties are involved. Moreover, the developed models are not considering the effect of spatial variability in soil properties.

Bayesian modeling (BM) overcomes one shortcoming of frequentist analysis by incorporating the prior knowledge about the unknown parameters into the analysis. As shown in Eq. 1, in Bayesian Modeling, the prior knowledge assumed for the unknown parameters is combined with the likelihood function from observed data to produce a posterior probability distribution.

$$p(\theta|y) = \frac{p(y|\theta) * p(\theta)}{p(y)} \tag{1}$$

where $p(\theta|y)$ is the posterior distribution of the parameter or the set of parameters given the data; $p(y|\theta)$ is the likelihood function which represent the probability distribution of the data given the model parameters; $p(\theta)$ is the prior distribution of the model parameters; and $p(y)$ is the marginalization term which represents the probability of observing the data.

While BM approach has the advantage of incorporating the prior knowledge into the analysis, the developed model lacks robustness and the selected prior distribution can highly affect the posterior distribution. Hierarchical Bayesian Modeling (HBM) resolves the robustness issue by decomposing the prior knowledge into several levels.

In HBM approach, the data is grouped within multiple clusters, and the uncertainty at each level is described by a separate set of parameters, developing a hierarchy of probability distributions that describe the different levels of the data. The layers in the hierarchical analysis reduce the effect of assumed prior distribution less prone to the posterior calculations. In addition, HBM has the ability to account for unbalanced sampling in the data and incorporate spatial heterogeneities in model parameters.

Because of the capabilities of HBM, over the past few years, there has been a significant emphasis on studying the application of the HBM approach in various geotechnical engineering aspects (e.g., Bozorgzadeh et al. 2019; Bozorgzadeh and Bathurst 2020; Xiao et al. 2021). In this paper, HBM is adopted to develop region-specific correlation models between the two commonly used in-situ tests in geotechnical site investigation, namely Standard Penetration Test (SPT) and Cone Penetration Test (CPT). For illustration purposes, 200 high-quality data pairs of $N_{1,60\,cs}$ and $q_{t1N\,cs}$ are adopted from Lu et al. (2022). Figure 1 shows the regional distribution of the adopted dataset.

As shown in Fig. 1, the dataset is compiled from testing results in distinct locations. Using all data as one group and developing one model for all is not ideal as it will neglect the spatial variability in soil and the potential affect it has on the correlation model. Developing a separate model for each region will also make the developed models suffer from sparse data. HBM approach addresses these issues by acknowledging differences and similarities among these regions. Once the region-specific models are developed using HBM, they will be compared to the regression model developed by Lu et al. (2022) for the same dataset.

This paper is organized as follows. First, the derivation and structure of the SPT-CPT HBM model are introduced. Second, the suggested region-specific correlation relationships are presented, and their performance is compared with the regression model

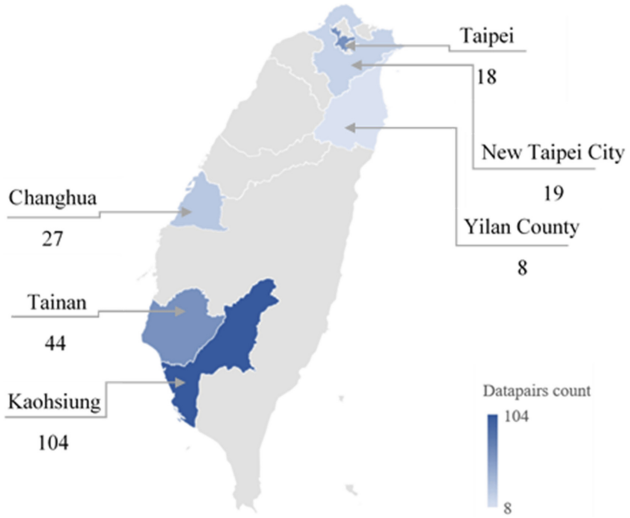

Fig. 1. Data counts and locations across regions for the adopted dataset (adopted from Lu et al. 2022)

developed for the same dataset. Third, the performance of the proposed HBM model is assessed in regions with limited data. And finally, the obtained results are discussed, and conclusions are drawn based on the study's outcomes.

2 Structuring the Hierarchical Bayesian Model (HBM)

Because most previous studies have adopted a linear relationship between $N_{1,60cs}$ and q_{t1cs}, a linear correlation is assumed for this analysis:

$$N_{1,60cs} = a + b(q_{t1Ncs}) \tag{2}$$

where a and b are the intercept and the slope of the linear model to be determined, respectively.

In order to account for region-specific $N_{1,60cs} - q_{t1Ncs}$ relationships, a_i and b_i are defined to represent the model parameters for the i^{th} region. Furthermore, the model uncertainty is considered by adding an error parameter, ε, to the correlation model:

$$N_{1,60cs} = a_i + b_i(q_{t1Ncs}) + \varepsilon \tag{3}$$

where ε represents the model error. ε is assumed to follow a normal distribution with a mean of zero and a standard deviation (σ_ε), herein.

In the Bayesian Modeling, it is crucial to ensure that the likelihood function captures the distribution of the data. Because the dataset shows a lognormal distribution for both $N_{1,60\ cs}$ and $q_{t1N\ cs}$ (see Fig. 2), the likelihood function is assumed to follow a lognormal distribution.

However, because a likelihood function with normal distribution is more efficient, $N_{1,60\ cs}$ and $q_{t1N\ cs}$ are transformed to normal distribution by taking the natural log of each

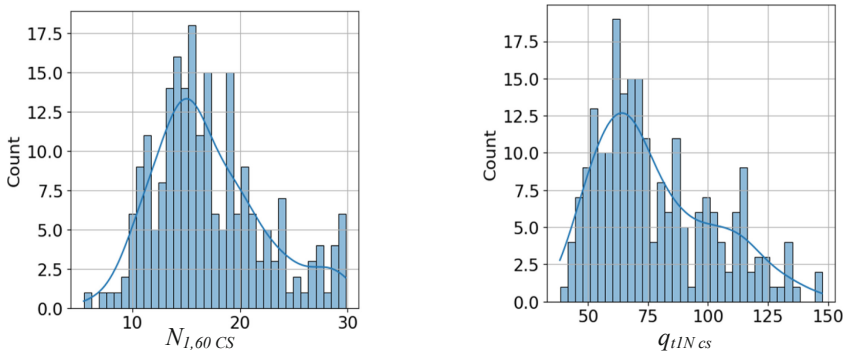

Fig. 2. Distribution of the adopted dataset

value; and thus, the likelihood function is also assumed to follow a normal distribution. Therefore, the correlation model becomes as follows:

$$lnN_{1,60cs} = a_i + b_i\left(lnq_{t1Ncs}\right) + \varepsilon \tag{4}$$

Although the $N_{1,60cs} - q_{t1Ncs}$ relationship is not the same in all regions, there might be similarities between the slope and intercept terms. To account for these similarities in different regions, a_i and b_i of each region are assumed to follow a normal distribution:

$$a_i \sim N(\mu_a, \sigma_a^2) \tag{5}$$

$$b_i \sim N(\mu_b, \sigma_b^2) \tag{6}$$

where μ_a and σ_a are the mean and standard deviation of the intercept a_i in different regions; and μ_b and σ_b are the mean and standard deviation of the model slope b_i in different regions.

Adopting Bayesian Modeling approach, the likelihood of observing the data pairs in a given region can be shown as follows:

$$f(d|a_i, b_i, \sigma_\varepsilon) = \prod_{i=1}^{R}\left[\prod_{j=1}^{ni} f\left(N_{ij}, qc_{ij}\middle|a_i, b_i, \sigma_\varepsilon\right)\right] \tag{7}$$

where $N_{1,60cs_{ij}}$, $q_{t1Ncs_{ij}}$ represents the j^{th} data pair from the i^{th} region; d_i represents the observed data in i^{th} region; R is the number of regions in the observed data points set d; and $f(d|a_i, b_i, \sigma_\varepsilon)$ is the chance to observe d.

Using HBM, the posterior distribution can be written as follows:

$$f(\theta|d) \propto f(\mu_a) f(\mu_b) f(\sigma_a) f(\sigma_b) f(\sigma_\varepsilon) \prod_{i=1}^{R} f(a_i|\mu_a, \sigma_a)$$
$$\times \prod_{i=1}^{R} f(b_i|\mu_b, \sigma_b) \prod_{i=1}^{R}\left[\prod_{j=1}^{ni} f\left(N_{ij}, qc_{ij}\middle|a_i, b_i, \sigma_\varepsilon\right)\right] \tag{8}$$

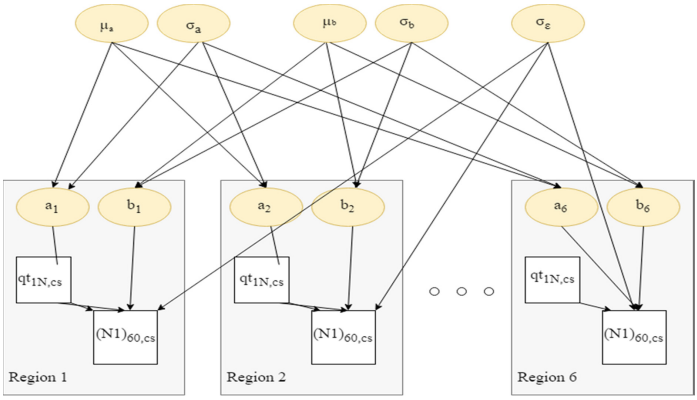

Fig. 3. Structure of the proposed Hierarchical Bayesian model (HBM)

where θ denotes the set of random variables of the HBM model to be determined ($\theta = \{\mu_a, \mu_b, \sigma_a, \sigma_b, \sigma_\varepsilon, a, b\}$, herein); and $f(\mu_a), f(\mu_b), f(\sigma_a), f(\sigma_b)$, and $f(\sigma_\varepsilon)$ represent the prior probability density functions of the model parameters.

To obtain the posterior distribution from Eq. 8, it is necessary to specify a prior distribution for each parameter in the model. Herein, vague priors (Eqs. 9–13) are implemented which will give more weight to the observed data when defining the posterior distributions.

$$\mu_a \sim N(0, 100) \tag{9}$$

$$\mu_b \sim N(0, 100) \tag{10}$$

$$\sigma_a \sim IG(0.001, 0.001) \tag{11}$$

$$\sigma_b \sim IG(0.001, 0.001) \tag{12}$$

$$\sigma_\varepsilon \sim IG(0.001, 0.001) \tag{13}$$

where N and IG represent normal and inverse-gamma distribution, respectively.

The proposed HBM is shown in Fig. 3.

3 The Proposed SPT-CPT Correlation Relationships

In contrast to single-level models, for which analytical solutions may be obtained using conjugate priors, it is not possible to obtain a direct analytical solution for a HBM model. Thus, Markov Chain Monte Carlo (MCMC) simulation should be adopted to sample from the posterior distribution of the model parameters. The JAGS library in the R programming language is adopted for MCMC simulation in this study.

Table 1. The proposed region-specific CPT-SPT correlation models

Region Number	Region	Relationship
1	Taipei	$N_{1,60cs} = \exp(0.1064 + 0.6082\ln q_{t1Ncs}) + \varepsilon$
2	Changhua	$N_{1,60cs} = \exp(0.1262 + 0.6137\ln q_{t1Ncs}) + \varepsilon$
3	Kaohsiung	$N_{1,60cs} = \exp(0.1393 + 0.613\ln q_{t1Ncs}) + \varepsilon$
4	New Taipei City	$N_{1,60cs} = \exp(0.1426 + 0.6209\ln q_{t1Ncs}) + \varepsilon$
5	Tainan	$N_{1,60cs} = \exp(0.1469 + 0.6265\ln q_{t1Ncs}) + \varepsilon$
6	Yilan County	$N_{1,60cs} = \exp(0.1524 + 0.6302\ln q_{t1Ncs}) + \varepsilon$
where $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(0, 0.2277)$ for all regions		

The MCMC algorithm was run for 100,000 iterations after defining the likelihood and prior functions for each parameter. To assess the convergence of the algorithm, autocorrelation plots and posterior predictive checks were adopted, as recommended by Gelman et al. (2013).

The developed region-specific correlation models are listed in Table 1. As shown in the Table 1, the model parameters for the different regions are similar, but the minor differences between parameters in different regions are what make the proposed HMB model perform better than conventional regression models that treat all data from different regions as a single entity. The HMB model can capture the unique characteristics of each region and make more accurate predictions.

4 HBM Model Performance

The proposed HMB model for each region is compared to the original correlation model by Lu et al. (2022):

$$N_{1,60cs} = 1.9 + 0.18q_{t1N,cs} + \varepsilon \tag{14}$$

Regional scatter plots and Root Mean Squared Error (RMSE) are used for model evaluations and comparison purposes.

Figure 4 shows the scatter plots for each region along with the model proposed by Lu et al. (2022), and also the proposed HBM models with their 95% confidence intervals. The RMSE comparison between HBM and the model proposed by Lu et al. (2022) is shown in Fig. 5.

As shown in Fig. 5, for the majority of the regions, the model developed using HBM has a smaller RMSE value compared to the model proposed by Lu et al. (2022) in which all data from distinct regions are considered as one group. However, the model is not performing well in Taipei region. This could be due to the shrinkage effect of the HBM, pulling the model prediction towards the group mean.

Table 2 shows the statistical summary of the dataset in each region. Because Taipei has the lowest $N_{1,60cs}$ mean value among all regions, the HBM model pulls the predictions towards the higher group mean values.

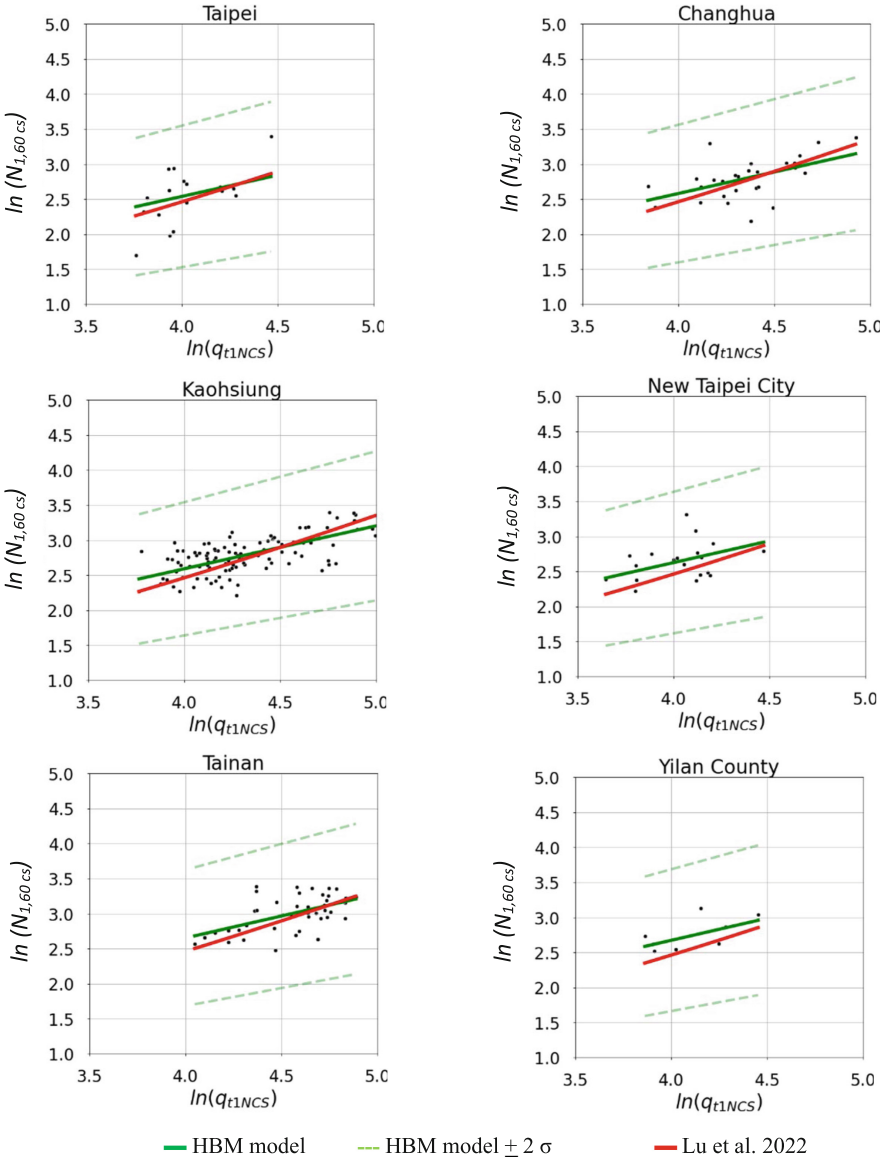

Fig. 4. Performance of the proposed model in the region-level

The amount of shrinkage that occurs in a hierarchical Bayesian model depends on the strength of the prior distribution used for the region-level parameters and the amount of data available for each region. If the prior distribution is informative, the estimate of the individual-level parameters will be less shrunk towards the overall mean. On the other hand, if the prior distribution is non-informative, or if there is relatively little data

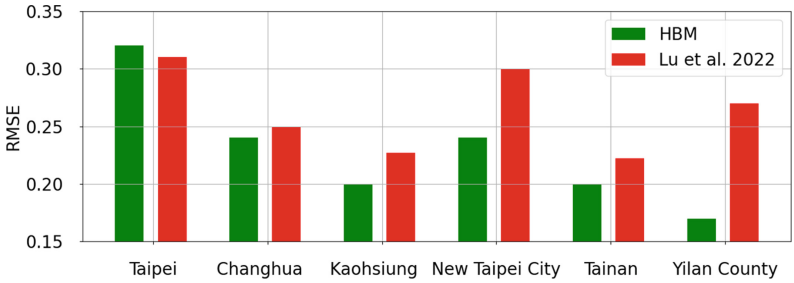

Fig. 5. RMSE comparison for each region

Table 2. Statistical summary of the dataset

Region	$N_{1,60cs}$			q_{t1Ncs}		
	Min	Max	Mean	Min	Max	Mean
Taipei	5.45	29.79	13.78	42.94	86.91	58.42
Changhua	8.92	29.31	17.07	46.60	137.56	80.28
Kaohsiung	9.11	29.83	16.69	43.11	147.55	77.76
New Taipei City	9.18	27.52	14.63	38.32	87.36	57.24
Tainan	11.90	29.56	20.71	57.32	132.69	96.85
Yilan County	12.41	22.88	16.18	47.60	85.92	61.87

available for each region, then the estimates of the individual-level parameters will be more shrunk towards the overall mean (Griffin and Brown 2017).

Shrinkage towards the group mean can be beneficial in cases where there is relatively little data available in the region, as it can help to regularize the estimates of the region-level parameters and reduce the variance of the estimates. However, shrinkage may also lead to a loss of power if the true values of the region-level parameters are very different from the group mean.

5 HBM Model Performance on New Regions and Limited Data Availability

To evaluate the performance of the HBM model in new regions, a new $N_{1,60cs}$-q_{t1Ncs} dataset is adopted from Andrus et al. (2004), which consists of 23 data pairs of $N_{1,60cs}$-q_{t1Ncs} from California, USA.

In addition, the robustness of the HMB model is tested on this new region by trying the model on different data sizes. For this reason, sets of five, ten, fifteen, and the full number of data points are adopted from this dataset. The HBM model is compared to the model proposed by Lu et al. (2022) and the model proposed by Andrus et al. (2004).

The models' error metrics are shown in Fig. 6; and the models' scatter plots comparison are shown in Fig. 7.

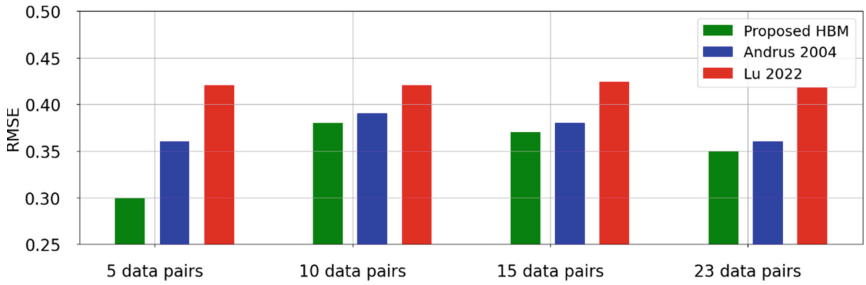

Fig. 6. RMSE of the three adopted models in the new region with data sparsity

HBM is shown to perform significantly better than other models in limited data availability. This could be due to the fact that HBM allows for "borrowing strength" across regions. In a hierarchical model, the region-level parameters are informed not only by the data within each region, but also by the overall distribution of the data across all regions.

6 Conclusions

In this research, a region-specific Hierarchical Bayesian Modeling approach is adopted for developing a SPT-CPT correlation model.

The comparison of the developed model with the model developed for the same dataset using frequentist statistics, indicated strong prediction capability of HBM.

In addition, the proposed HBM model demonstrated strong performance in regions with sparse and limited data, making this approach well-suited for correlation model developments in regions with limited data availability.

It should also be noted that limited data availability in one region, or non-informative prior distributions could lead to shrinkage toward the group mean in the model developed using HBM. While shrinkage towards the group mean can be beneficial in cases where there is relatively little data available in the region, it can also be detrimental if the true values of the region-level parameters are very different from the group mean.

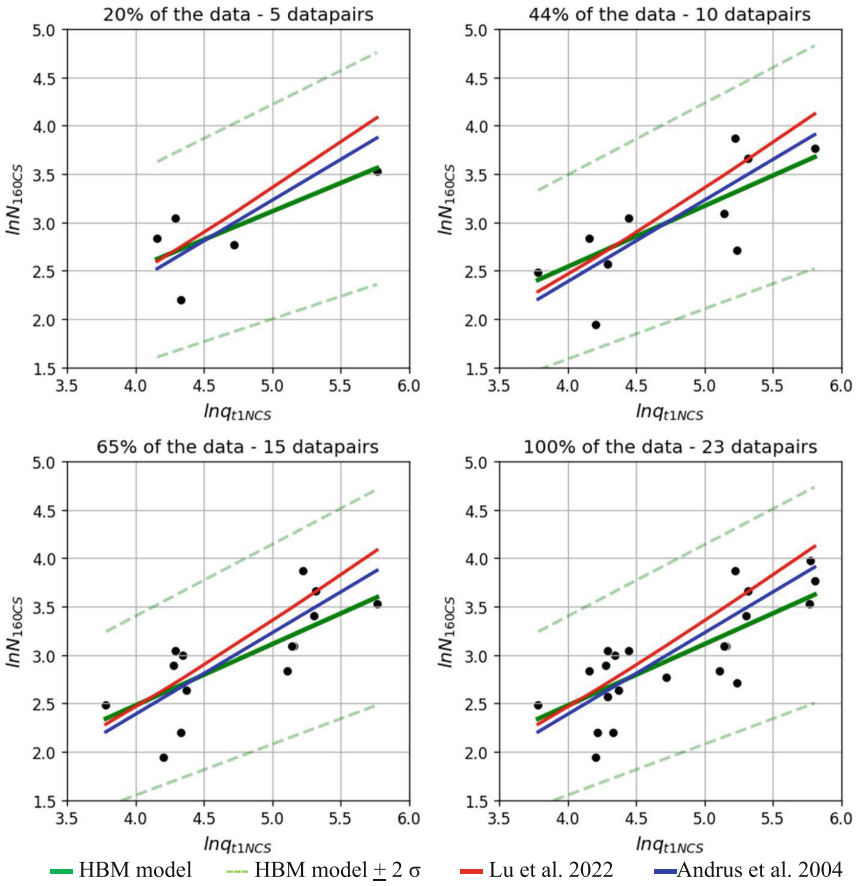

Fig. 7. Models' performance in new regions with data sparsity

References

Albert, J. & Hu, J., 2019. *Probability and Bayesian Modeling.* s.l.:CRC Press.

Andrus, R. D. et al., 2004. Comparing liquefaction evaluation methods using penetration-VS relationships. *Soil Dynamics and Earthquake Engineering,* pp. 713–721.

Bozorgzadeh, N. & Bathurst, R., 2020. Hierarchical Bayesian approaches to statistical modelling of geotechnical data. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards,* pp. 452–469.

Bozorgzadeh, N., Harrison, J. & Escobar, M., 2019. Hierarchical Bayesian modelling of geotechnical data: application to rock strength. *Géotechnique*, pp. 1056–1070.

Ebrahimian, B. & Movahed, V., 2017. Application of an evolutionary-based approach in evaluating pile bearing capacity using CPT results. *Ships and Offshore Structures,* pp. 937–953.

Gelman, A. et al., 2013. *Bayesian Data Analysis.* s.l.:s.n.

Griffin, J. & Brown, P., 2017. Hierarchical shrinkage priors for regression models. *Bayesian Analysis,* pp. 135–159.

Hatta, K. A. B. K. & Osman, S. B. A. S., 2015. Correlation of Electrical Resistivity and SPT-N Value from Standard Penetration Test (SPT) of Sandy Soil. *Applied Mechanics and Materials.*

Kumar, R., Bhargava, K. & Choudhury, D., 2016. Estimation of Engineering Properties of Soils from Field SPT Using Random Number Generation. *INAE Letters*, pp. 77–84.

Lu, Y.-C.et al., 2022. A new approach to constructing SPT-CPT correlation for sandy soils. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, pp. 1–17.

Robertson, P. K., Campanella, R. G. & Wightman, A., 1983. SPT-CPT Correlations. *Journal of Geotechnical Engineering*, pp. 1449–1459.

Shahri, A., Juhlin, C. & Malemir, A., 2014. A reliable correlation of SPT-CPT data for southwest of Sweden. *Electronic Journal of Geotechnical Engineering,* pp. 1013–1032.

Xiao, S., Zhang, J., Ye, J. & Zheng, J., 2021. Establishing region-specific N – Vs relationships through hierarchical Bayesian modeling. *Engineering Geology.*