Computers and Geotechnics 132 (2021) 103963

Contents lists available at ScienceDirect

Computers and Geotechnics

journal homepage: www.elsevier.com/locate/compgeo

Research Paper

Bayesian hierarchical and measurement uncertainty model building for liquefaction triggering assessment

Jonathan Schmidt [*], Robb Moss

1 Grand Avenue, San Luis Obispo, CA 93407, United States

ARTICLE INFO

Keywords:
Liquefaction
Bayesian statistics
Predictive modeling
Earthquake engineering
Natural hazards

ABSTRACT

This study examines the details of creating and validating an empirical liquefaction model, using a worldwide cone penetration test (CPT) liquefaction database with the intent of incorporating the rigor found in predictive modeling in other fields and addressing shortcomings of existing models. Our study implements a logistic regression within a Bayesian measurement error framework to incorporate uncertainty in predictor variables and allow for a probabilistic interpretation of model parameters when making future predictions. The model is built using a hierarchal approach to account for intra-event correlation in loading variables and differences in event sample sizes. The model is tested using an independent set of recent case histories.

We found that the Bayesian measurement error model considering two predictor variables, normalized CPT tip resistance and cyclic stress ratio decreased model uncertainty while maintaining predictive utility for new data. Hierarchical models revealed high model uncertainty potentially due to the database lacking in high loading non-liquefaction sites. Models considering friction ratio as a predictor variable performed worse than the two variable case and will require more data or informative priors to be adequately estimated. The framework developed is flexible and can be extended using different methods of predictor variable selection, model function forms, and validation processes.

1. Introduction

Seismic soil liquefaction is a major cause of earthquake damage to the built environment, second only to tsunamis in overall cost. For purposes of this paper, seismic soil liquefaction is defined as when a loose, saturated, granular soil loses shear strength due to dynamic earthquake loading (NAP, 2016).

Current practice relies on empirical liquefaction models (ELM's) to make predictions of potential liquefaction occurrence at future sites during engineering design and analysis (NAP, 2016). These models assess liquefaction potential using predictive models built on a database of observed case histories. As discussed at length by the Committee on State of the Art and Practice in Earthquake Induced Soil Liquefaction Assessment there are appreciable shortcomings of current ELM's (NAP, 2016).

Ideally, liquefaction assessment will eventually be conducted in a fully performance-based engineering (PBE) approach that evaluates engineered features over the entire range of possible loadings rather than a single or discrete group of seismic events. A PBE approach requires a probabilistic description of liquefaction potential; a prediction of the probability of failure rather than a factor of safety or yes/no output. Currently, only two models used in common practice (Moss et al., 2006; and Boulanger and Idriss, 2016) provide predictions of liquefaction probability.

However, a greater limitation of existing ELM's is a lack of openness regarding the model building process. Because these training methods and metrics are often not reported, practitioners cannot currently evaluate model biases when selecting which relationships to use or recommend in guidance documents.

Furthermore, there is a lack of rigorous model performance validation. Many popular studies simply do not report performance metrics (e.g. Moss et al., 2006, Boulanger and Idriss, 2016). Others use the same data to validate the model as was used to build the model (Juang et al., 2002, Yazdi and Moss, 2017, Lai et al., 2006, etc). This approach results in optimistically biased performance metrics because they measure the model fit to the training data, not necessarily how well it will perform for out-of-sample predictions. This is referred to as over-fitting in the predictive modeling world (Kuhn and Johnson, 2013). To date, only a

[*] Corresponding author at: 1 Grand Avenue, San Luis Obispo, CA 93407, United States.
E-mail addresses: jschmidt@yeh-eng.com (J. Schmidt), rmoss@calpoly.edu (R. Moss).

https://doi.org/10.1016/j.compgeo.2020.103963
Received 28 June 2020; Received in revised form 18 October 2020; Accepted 3 December 2020
Available online 12 January 2021

few relevant models (Oommen et al., 2010, Rezania et al., 2011, etc.) split their databases into training and testing sets through cross validation or other methods to develop relatively unbiased metrics of model performance.

Finally, existing models do not account for intra-event correlation between outcomes or sample size discrepancies between events. This may become problematic as the updated Next Generation Liquefaction (NGL) database introduces a large number case histories from a limited number of events (Brandenberg et al., 2020). The field of ground motion modeling has addressed these shortcomings using hierarchical approaches (e.g Abrahamson and Youngs, 1992 or Kuehn and Scherbaum, 2015), commonly referred as mixed or random effects models. No existing liquefaction triggering model has implemented such techniques.

This work described here presents an open and extensible modeling methodology for triggering models developed on the updated NGL cone penetration test (CPT) case history database to address these model shortcomings.

2. Modeling Framework

Our overall methodology includes three major steps: exploratory data analysis, model building (including data preprocessing) and model validation. These last two steps can be performed in an iterative fashion, using results from previous model fits to inform future work. Because the primary goal of this work is to focus on the statistical modeling process it is necessary to adopt a liquefaction framework and database as givens. For purposes of this study, we use CPT liquefaction database as developed in Moss et al. (2006) while recognizing that case history screening and processing of raw field measurements are an important ongoing field of research.

2.1. Exploratory Data Analysis

Understanding the distributions and interactions between predictors and outcomes in the dataset is a critical step in selecting the proper model. The database used from Moss et al. (2006) included 182 case

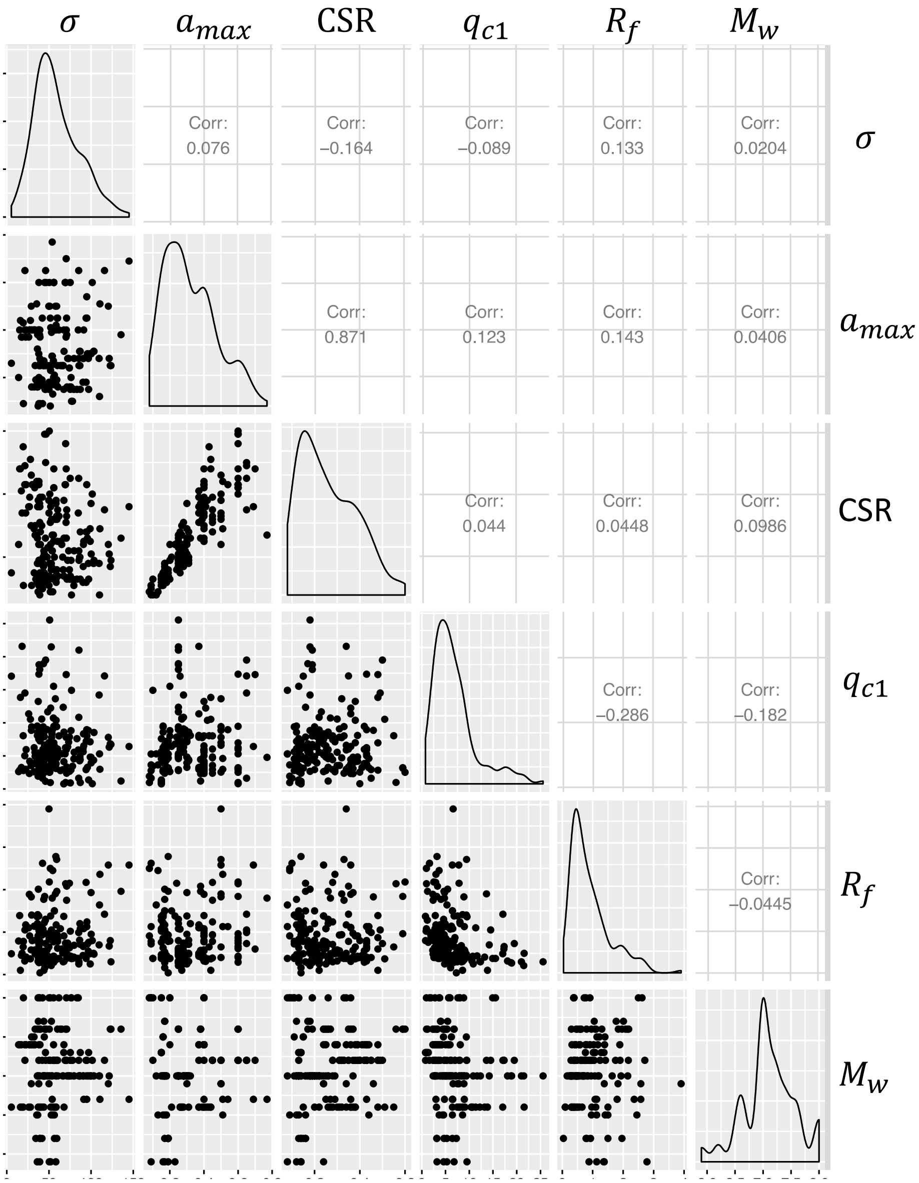

Fig. 1. Cross correlation of predictor variables used in the model building phase of this study. Scatter plots show trends in correlation between variables and distribution plots show the central tendency and dispersion of each variable.

histories from 18 events. These case histories contained 139 instances of liquefaction and 43 instances of nonliquefaction. These outcomes were associated with 12 predictor variables:

- Data class (A, B, or C), subjectively assigned based upon confidence in field data
- Critical depth: the depth range of the layer determined to have liquefied
- Groundwater table level: Depth below ground surface of the groundwater table
- Vertical total stress (σ_v)
- Vertical effective stress ($\sigma_v^{'}$)
- Peak ground acceleration (a_{max}), usually estimated indirectly from attenuation relationships
- Shear stress reduction coefficient (r_d) used to calculate CSR
- Cyclic stress ratio (CSR)
- CPT normalization exponent (c), an input to the equation for normalizing CPT measured tip resistance
- Normalized CPT tip resistance ($q_{c,1}$)
- Friction ratio (r_f): the CPT measured sleeve friction divided by the penetration resistance
- Moment magnitude (M_w)

Importantly many of these predictor variables are correlated or even functions of each other (Fig. 1). This can be problematic because many functional forms will have greater uncertainty in parameter estimates when predictors are correlated.

To illustrate the spread of data both as a whole and within the liquefaction/nonliquefaction classes, we first focus on a single load and resistance predictor; CSR and $q_{c,1}$. The database includes a reasonably wide range of CSR and $q_{c,1}$ mean values. In both cases, the data are left skewed and have a moderately high coefficient of variation ($\frac{\mu}{\sigma} \approx 0.5 - 0.7$). Additionally, Fig. 2 shows that each event has a slightly different distribution of load and resistance values. There is a noticeable association between event and CSR due to certain earthquakes having higher ground motions than others. Unlike CSR there is not a noticeable association between event and $q_{c,1}$.

Fig. 3 shows the separability between instances of liquefaction and nonliquefaction for the three predictor variables considered. Although there is no dramatic separation between the classes, liquefaction is generally associated with lower penetration resistance and higher CSR, and R_f does not show any clear separation.

2.2. Model Validation Framework

Modern predictive modeling techniques can learn complex relationships between predictors and outcomes (Kuhn and Johnson, 2013). However, if not supervised properly they may end up overemphasizing patterns that do not generalize to new data. In a sense, they have “memorized” the training data instead of learning how to predict future outcomes. This is further compounded when the model is validated using the same data as it was built on, because the apparent performance will be good despite the model making poor future predictions. With an appropriate data splitting strategy, e.g., k-fold cross validation or training/testing sets, data can be used independently for training and testing. This results in realistic measures of model performance on out of sample predictions.

Because the goal of this paper is the modeling process instead of a new triggering model we ultimately chose to use the entire Moss et al. (2006) data for training, and set of select case histories from the 2011 New Zealand Canterbury earthquake sequence summarized in Green et al. (2014) for testing. This choice was primarily made to make coding the models easier. Fig. 4 shows a scatterplot in $q_{c,1}$ and CSR space of the liquefied and non-liquefied case histories. The New Zealand case histories are indicated by open and closed triangles for nonliquefied and liquefied cases respectively and the Moss et al. (2006) case histories are indicated by open and closed circles. The mean values of predictors in the New Zealand testing set are generally similar to the Moss et al. training set, however the maximum values (i.e. data breadth) are lower. This is visualized by the clustering of the New Zealand cases in the bottom left corner of the chart. The class ratio of liquefaction to nonliquefaction is 49:15, which is nearly equal to the Moss et al. data.

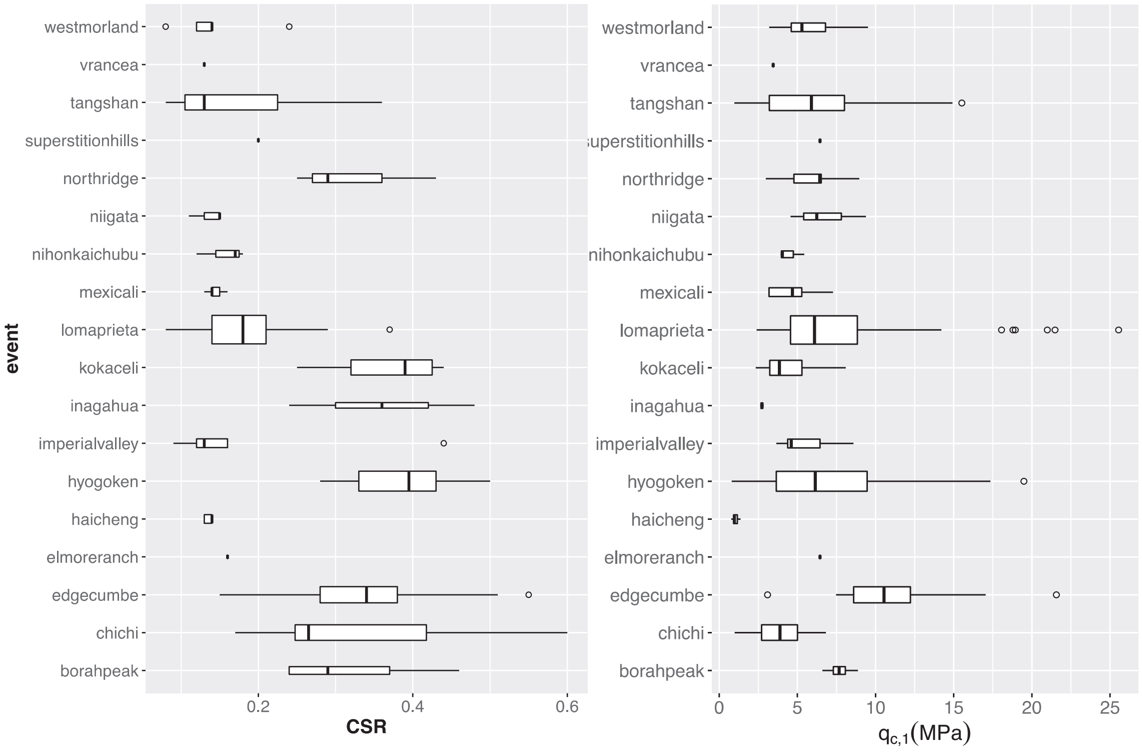

Fig. 2. Box plots of CSR and q_{c1} showing the central tendency and dispersion for each earthquake event.

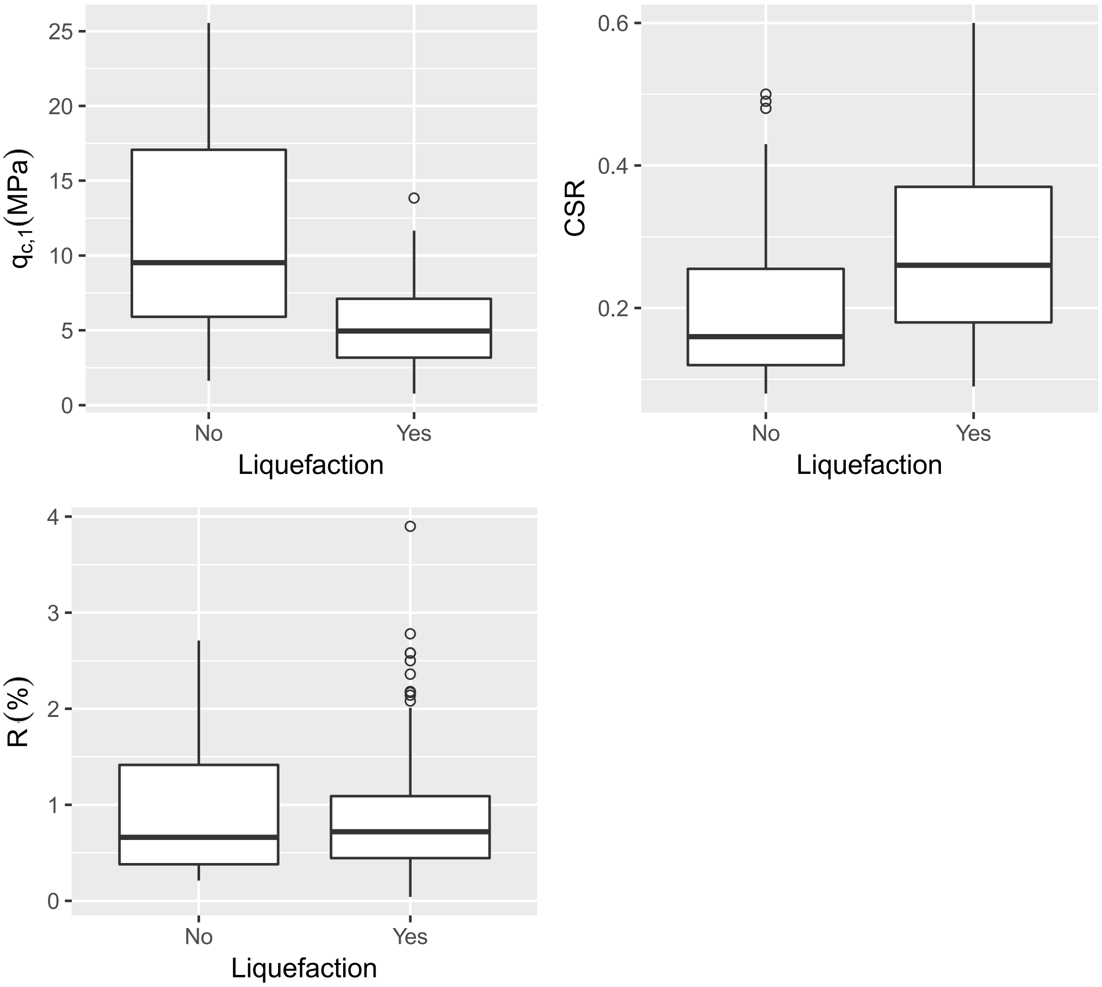

Fig. 3. Separability of liquefaction cases via the three primary predictors for the training data. Box plots show the central tendency and dispersion with respect to yes/no classification.

A binary classification model's performance on a training set is quantified by the confusion matrix, i.e., true positive, false positive, true negative, or false negative (Kuhn and Johnson, 2013). With these counts several metrics can be calculated, such as accuracy, precision, recall, and others. Most predictive modeling texts provide equations to calculate these metrics

and discuss their applicability. When selecting an appropriate metric, a modeler should be aware of the limitations of various performance metrics and select one that fits the purpose of the model. For example, if negative outcomes are infrequent then a model can achieve near perfect accuracy (the proportion of events labeled correctly) by only predicting positive outcomes (Kuhn and Johnson, 2013). If false positives are of interest, such as with unnecessary expensive ground improvements or cancer treatments, this is an inappropriate performance metric. A metric that is not affected by natural class frequencies would be more appropriate.

Models that produce outcome class probabilities of can be converted to binary classifiers by selecting a probability threshold at which all points below are considered “no” and all above are considered “yes”. This threshold between yes and no classifications can be thought of as of a threshold of acceptable liquefaction risk (TH_L). This can be selected based upon acceptable risk, on a project specific basis. Sites falling above that threshold are classified as liquefiable and those below are not and design proceeds accordingly. The selected threshold may differ appreciably from project to project so we want a model that performs well at all levels of TH_L.

Receiver operating characteristics (ROC) curves are a useful tool for evaluating a model's performance across all possible threshold values. They compare the true positive rate (after Fawcett, 2006):

$$TPR = \frac{\text{Postives Correctly Classified}}{\text{Total Positives}} = \frac{TP}{TP + FN} \qquad (1)$$

And the false positive rate (After Fawcett, 2006):

$$FPR = \frac{\text{Negatives Incorrectly Classified}}{\text{Total Negatives}} = \frac{FP}{FP + TN} \qquad (2)$$

ROC curves are plotted as FPR vs TPR, with each point corresponding to a specific threshold value. The area under the curve (AUC) is a useful scalar summary of performance, which will range from 0 0.5 to 1.0 with higher values indicating better model performance (Fawcett, 2006). The statistical interpretation of this value is the probability that a randomly chosen positive instance will have a computed higher probability of occurrence than a randomly chosen negative one (Fawcett, 2006). The benefit of using the TPR and FPR is that they are not sensitive to natural class frequencies because only the total positives or negatives are reflected in the denominators.

Precision-Recall (PR) curves, and their associated area under the curve (AUC-PR) are an alternative threshold independent metric used to evaluate binary classifiers. Recall is defined identically to TPR. Precision is defined as (After Davis and Goadrich, 2006):

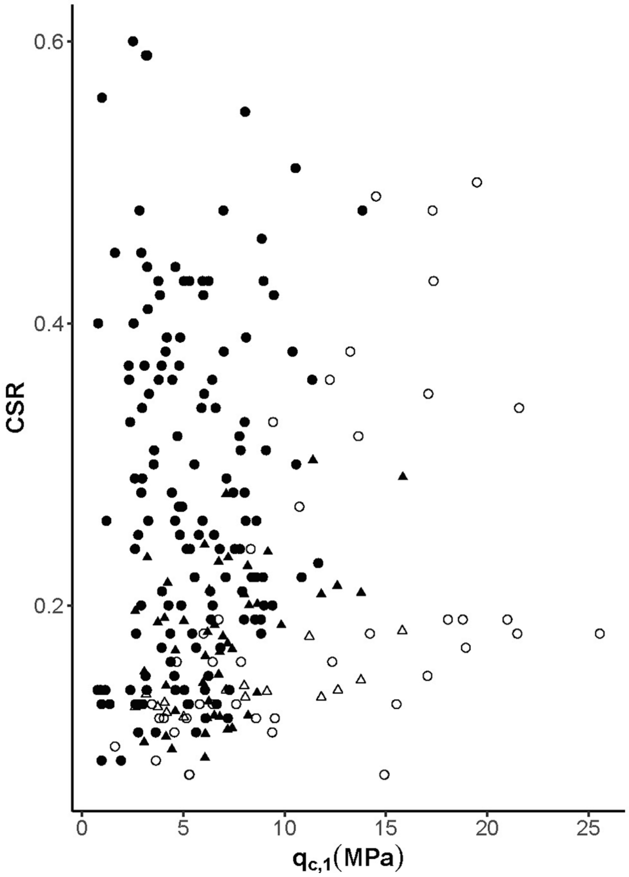

Fig. 4. Combined data sets from the training and testing data sets. Open and closed triangles from the New Zealand ((Green et al., 2014) data set, and open and closed circles from the worldwide (Moss et al., 2006) data set.

$$\text{Precision} = \frac{\text{Postives Correctly Classified}}{\text{Predicted Positive}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

PR curves are plotted as Recall vs Precision. PR curves are sometimes recommended over ROC curves by practitioners when evaluating performance on imbalanced data (e.g. Saito and Rehmsmeir, 2015, Davis and Goadrich, 2006, and others). The imbalanced datasets referred to in these cited works are dominated by negative cases, typically the number of negatives is hundreds to thousands of times more than the positives, and that the rare, positive outcome is the only outcome of interest. In this case, precision will vary considerably more between models than FPR and the algorithm that more accurately predicts positive outcomes will be more apparent. However, the PR curve does not consider true negatives and is unsuited for problems when both classes need to be identified correctly like liquefaction assessment. Although it is not without its flaws (see Cook, 2007) we adopt the AUC as our measure of model performance.

Reporting the metrics used to inform modeling choices is critical to understanding the unavoidable biases they introduce. For example, a model trained to maximize accuracy may have a high false positive rate. This can have positive or negative implications for the model's usage. Tosteson et al. (2014) discusses the implications of model training criteria in the context of breast cancer screening.

2.3. Basic Model Form

We now have a dataset with vectors of predictor variable values associated with binary outcomes coded 1 for liquefaction and 0 for nonliquefaction. The most common parametric assumption is that these observations are realizations of Bernoulli random variables. A Bernoulli random variable takes on a value of 1 with probability p and 0 with probability 1 - p. A logistic regression models the expected value of the i^{th} Bernoulli random variable given $\boldsymbol{\beta}$**,** a vector of model coefficients (with intercept β_0), and $\mathbf{x_i}$, an equal length vector of the corresponding predictor variables' values as (After Liao et al., 1988):

$$E(y_i) = \Pr(y_i = 1|\boldsymbol{\beta}, \boldsymbol{x}_i) = P_{L,i} = \frac{1}{1 + \exp\{-(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)\}} = \operatorname{sig}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \tag{4}$$

[Note the use of vector multiplication to simplify expressing a linear combination of model parameters and predictor variables.]

Assuming independent outcomes, we can use this formula to express the likelihood of observing the training dataset **D** (predictor variables and outcomes) for a fixed vector of parameters $\boldsymbol{\beta}$ for n_L liquefied cases and n_{NL} nonliquefied cases:

$$l(\boldsymbol{D}|\boldsymbol{\beta}) = \prod_{i=1}^{n_L} P_{L,i} \prod_{j=1}^{n_{NL}} (1 - P_{L,j}) \tag{5}$$

It is mathematically useful to deal instead with the logarithm of the above function. It can also be modified to include weights that account for the pre-existing class imbalances:

$$\ln[l(\boldsymbol{\beta})] = w_L \sum_{i=1}^{n_l} \ln(P_{L,i}) + w_{NL} \sum_{j=1}^{n_{NL}} \ln(1 - P_{L,j}) \tag{6}$$

Where w_L and w_{NL} are the weights assigned to liquefied and nonliquefied cases respectively. The maximization of the above function gives point estimates of model referred to as maximum likelihood estimates used in data pre-processing decision making.

2.4. Data Preprocessing

2.4.1. Predictor Variable Selection

Predictor variable selection is a fundamental step in any model development and can be thought of as a balance between under and over fitting (Kuhn and Johnson, 2013). Including predictors that have little effect on predicted outcomes will increase model uncertainty and make the model more difficult both to fit and interpret (Kuhn and Johnson, 2013). An overly complicated model may confuse users or be impractical for regular use.

As a preliminary tool, we used a stepwise selection process to determine which predictor variables were worth including based upon the Akaike Information Criterion (AIC) of the maximum likelihood fits. AIC is a metric for making relative comparisons about model utility that estimates the tradeoff between model goodness of fit and the simplicity of the model, essentially over versus under fitting (Burnham and Anderson, 2004). At each step in the process, the predictor variables are added or removed one by one from the model and the AIC calculated. The model with the lowest AIC is selected for the next step and the process continues until no proposed model outperforms the current.

Based upon these results there is justification for considering models of three predictor variables: $q_{c,1}$, CSR, and R_f. Qualitatively, these predictor variables cover several main factors affecting liquefaction: in-situ density, magnitude of cyclic shearing, and apparent fines content. Notably, including magnitude or stress did not produce a better ranking model.

2.4.2. Predictor variable transformations

A limitation of logistic regression is that it only allows for linear combinations of the predictor variables and model coefficients. However, performance is improved by instead dealing with transformations of predictor variables. We selected the Box-Cox family of monotonic transformations because of its flexibility and its ability to capture many common transformations such as powers and logarithms. A Box-Cox transformation of a predictor variable x, indexed by the parameter λ, is defined as (after Box and Cox, 1964):

$$x^{'} = \begin{cases} \ln(x) \text{if} \lambda = 0 \\ \dfrac{x^{\lambda} - 1}{\lambda} if \lambda \neq 0 \end{cases} \tag{7}$$

We used a simple grid search method to determine the group λ's that produced the best performing model, as measured by 5-fold cross validated AUC. Conceptually, this introduces a tuning parameter to the standard maximum likelihood logistic regression that allows for greater flexibility in the shape of the probability contours. The optimal Box-Cox parameters were $\lambda_{CSR} = -0.6$and $\lambda_{q_{c,1}} = 1.6$ for the two variable case and $\lambda_{CSR} = -0.6$, $\lambda_{q_{c,1}} = 1.0$, and $\lambda_{R_f} = 0.2$ for the three variable case.

Because only the mean and standard deviation of the predictor variables are included in the database, the transformed moments cannot be calculated directly for a nonlinear transformation without assuming a distributional form for each data point. Instead, we use a first order second moment (FOSM) approximation technique that calculates the moments of the Taylor series expansion of the transformation (see Moss, 2020 for a detailed derivation). The effects of these transformations are shown in Fig. 5, following.

2.4.3. *Accounting for sampling bias*

Because post-earthquake geotechnical reconnaissance is often focused on sites that have experienced ground failure (and subsequence impacts to engineered features) liquefaction databases contain more liquefied cases than non-liquefied cases. This is in contrast with the true class ratio, which likely contains more nonliquefied cases. For this study we considered up-sampling, randomly duplicating observations in the less frequent class to balance the dataset, and likelihood weighting to compensate for sampling bias. The weights used in likelihood function (equation (2)) were $w_L = 1.0$ and $w_{NL} = 1.5$ consistent with previous research (Cetin et al. 2004, Moss et al. 2006, Boulanger and Idriss, 2016, etc.).

In preliminary work, the coefficients estimated by weighting method were nearly identical to the upsampling method as was model predictive performance. Going forward, all models were fit on the up-sampled dataset.

2.5. *Model Fitting*

Bayes' Rule is the mathematical framework for updating our prior beliefs based upon observed evidence (Christensen et al., 2011). Bayes rule can be conceptualized as the posterior is proportional to the product of the prior and the likelihood (Posterior α Prior x Likelihood). There are three steps to performing Bayesian data analysis: specifying prior distributions for all model parameters, formulating a observational model for the training data to determine the appropriate likelihood function, and calculating or approximating the resulting posterior distributions for forward inference and model checking. In full generality, Bayes Rule is expressed mathematically as

$$f(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{l(\boldsymbol{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int l(\boldsymbol{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{8}$$

where $l(\boldsymbol{x}|\boldsymbol{\theta})$ is the joint likelihood of observing the data given fixed parameters, $f(\boldsymbol{\theta})$ is our prior beliefs about parameters expressed as a joint probability distribution, and $f(\boldsymbol{\theta}|\boldsymbol{x})$ is the joint posterior distribution of the parameters, given fixed data. The iterated integral of the product of the prior and likelihood over the support of all θ's in the denominator scales the resulting distribution so it obeys the axioms of probability. This Bayesian formula is usually difficult to compute exactly and is commonly approximated computationally.

2.5.1. *Prior selection*

We use weakly informative normal distributions to constrain the scale of the regression coefficients. A weakly informative prior contains enough information to limit the mode to realistic parameter values while remaining relatively uninformative over this range (Gelman, 2006). When data is relatively limited, such as this study, choosing uniform prior places far too much mass on unrealistic (or impossible) parameter values and prevents accurate inference (Gelman, 2006). To determine a realistic scale for our prior we follow the logic of Gelman et al. (2008) that a typical increase (i.e. one standard deviation) in a predictor variable would result in a jump from 1% probability to 99% probability. We default to a Normal (0,10) prior on regression slopes and use a slightly more diffuse distribution of Normal (0,25) for the intercept parameters. For a sensitivity study, we also run the models with prior standard deviations of 25 and 100 to examine the influence of the prior distribution

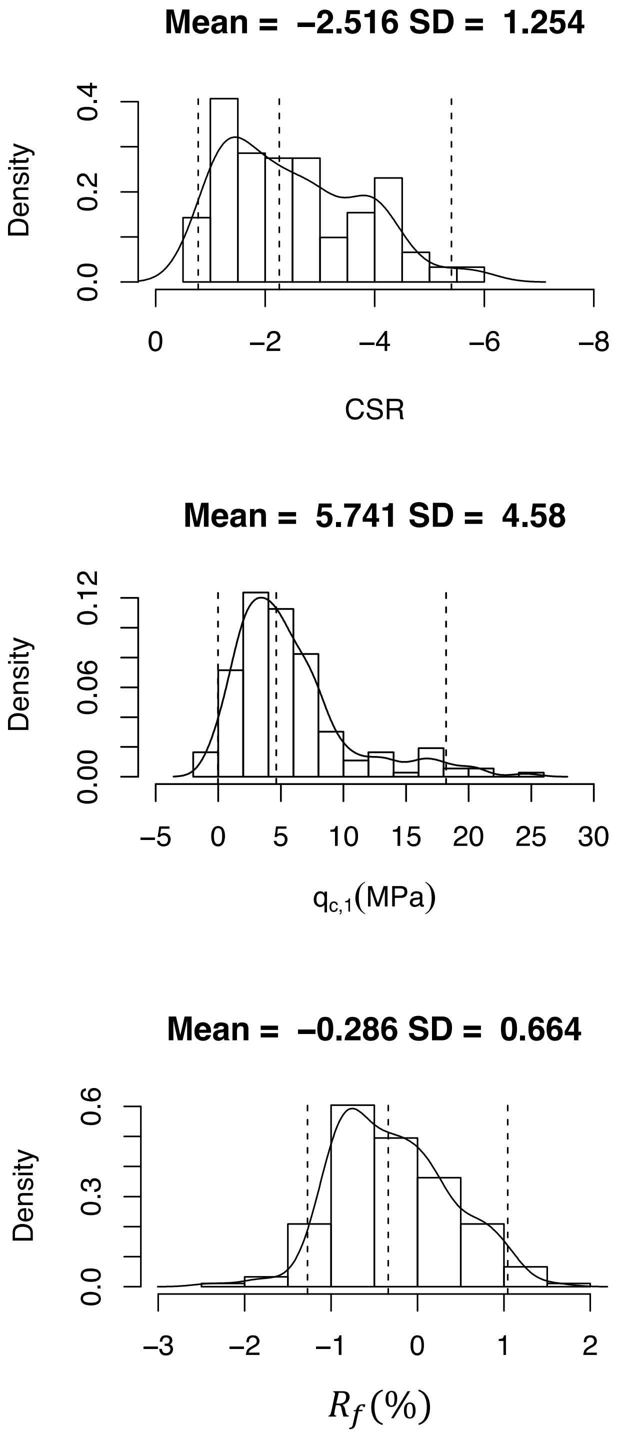

Fig. 5. Transformed distributions for the three primary variables with Box-Cox parameters of $\lambda_{CSR} = -0.6$, $\lambda_{q_{c,1}} = 1.0$, and $\lambda_{R_f} = 0.2$.

on the models' behavior.

2.5.2. Model forms

We developed 4 separate groups of models here to test out the influence of each step. Initially the model forms are expressed with only two predictors: $q_{c,1}$ and CSR. The subsequent extension to any arbitrary number of predictors is straightforward. Model 0 was a logistic regression fit using typical maximum likelihood methods. The baseline model, Model 1, was a Bayesian Logistic Regression with no parameter uncertainty to have something to compare subsequent models to. This produced nearly identical results to a logistic regression model (Model 0) which uses maximum likelihood to fit the functional form. Model 2 was a Bayesian Hierarchical model. Hierarchical models, referred to as mixed effects in ground motion attenuation relationship development (e.g Abrahamson and Youngs, 1992 or Kuehn and Scherbaum, 2015), allow model parameters (slopes and/or intercepts) to change between groups provided they are constrained by hyperparameters estimated from the data (Jiang, 2007). This can be thought of as a compromise between no pooling and complete pooling (Gelman and Hill, 2007). A no pooling model fits each group separately but is unfeasible for groups with limited data (Clark and Linzer, 2015). A completely pooled model fits a single set of parameters for all the data, ignoring group level variability (Gelman and Hill, 2007). Groups with more data points will "overwhelm" the smaller groups leading to "overconfident" out of sample predictions. This is of concern if events with limited data are the only ones available for data sparse regions (e.g. the upper right hand corner in $q_{c,1}$, CSR space).

Model 3 is a Bayesian Measurement Error model based off the work by Kuehn and Abrahamson (2018) and Moss et al. (2006). The measurement error model treats the true values of predictor variables as parameters to be estimated during the modeling process. The latent true values, x_{true}, are assumed to come from a normal distribution centered at the observed value, $x_{observed}$, from the database and with the corresponding estimated standard deviation, τ_x. The true values are given a hierarchical constraint that they come from the distribution of values observed in the database. This is assumed to be normal with database mean μ_x and standard deviation σ_x. The final model, Model 4, combines the measurement error and hierarchical functional forms from Models 2 and 3.

2.5.3. Posterior Solutions

In this study we used the program Stan (Carpenter et al., 2017) to perform Hamiltonian Markov Chain Monte Carlo (MCMC) simulation of the posteriors of interest. Convergence criteria were a targeted $\widehat{R}$ of 1.0 ± 0.1 and qualitative inspection of the trace and autocorrelation plots to verify independent sampling of the entire joint posterior distribution. $\widehat{R}$ is a commonly used measure that compares the between-chain and within-chain variances (Gelman and Rubin, 1992). Large differences between the two (high $\widehat{R}$) indicate nonconvergence (Brooks and Gelman 1997). We used 4 chains and selected an appropriate number of post-warmup iterations to satisfy convergence criteria. As necessary we modified sampler controls to ensure stability and efficiency. Schmidt (2020) describes the code use to fit the models and generate results in further detail.

2.6. Bayesian Posterior Predictive Inference

For a Bayesian analysis the result is a joint posterior distribution of model parameters (generically referred to as $\boldsymbol{\theta}$) conditioned on the observed data. In our case, the posteriors of interest are the regression coefficients ($\beta_0, \beta_1, \beta_2$, and β_3) or their population mean values (μ_β) for the hierarchical models. Technically, we do not have the analytic form of the posterior but rather a MCMC estimate $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(m)}\}$ where $\boldsymbol{\theta}^{(n)}$is the n$^{\text{th}}$ draw of a vector of model parameter values from the posterior. $\beta_0^{(n)}, \beta_1^{(n)}$,and so on, are elements of $\boldsymbol{\theta}^{(n)}$. Given new data, $\boldsymbol{x}$, we would like to predict the probability of liquefaction or $\Pr(y = 1||\boldsymbol{x})$. If we fix $\boldsymbol{\theta}$ to the maximum likelihood estimate or the posterior mean/median we can use equation (3) to compute the probability (referred to as the *maximum a posteriori* estimate). However, this process ignores the posterior variability in $\boldsymbol{\theta}$ we just estimated.

Instead we want to perform fully Bayesian inference from the posterior predictive distribution resulting from the observed new data $\mathrm{f}(y|\boldsymbol{x})$. From the assumptions of the logistic model we know the distribution of the new outcome conditional on fixed parameters $\mathrm{f}(y|\boldsymbol{x},\boldsymbol{\theta})$ is $\mathrm{Bernoulli}(p)$. To find the unconditional predictive distribution $\mathrm{f}(y|\boldsymbol{x})$ we marginalize with respect to the parameter posterior distributions by taking the integral:

$$\int \mathrm{f}(y|x, \boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{9}$$

This can be conceptualized as weighting our estimates for y by how likely their generating parameters values are, given the training data. Then, the probability of interest is computed as $\Pr(y = 1||\boldsymbol{x}) = \mathrm{E}[\mathrm{f}(y|\boldsymbol{x})]$. However, because we only have samples from the posterior $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(m)}\}$ we instead compute the MCMC estimate using draws of $\boldsymbol{\beta}^{(j)}$

$$\Pr(y = 1||\boldsymbol{x}) = \mathrm{E}[\mathrm{f}(y|\boldsymbol{x})] \approx \frac{1}{N}\sum_{j}^{N} \frac{1}{1 + \exp\left\{-\left(\beta_0^{(j)} + \boldsymbol{\beta}^{(j)T}\boldsymbol{x}\right)\right\}} \tag{10}$$

This implies that models with greater uncertainty in their parameter estimates will results in more uncertain estimates of probability.

3. Results

The following sections present the results of our modeling process. To visually compare model performance we present four key graphical summaries discussed in the sections following (Figs. 6 through 10). The remaining model visualizations are included in Schmidt (2020). To visualize the models, the probability contours of the resulting surface over $q_{c,1}$ and CSR are plotted. For the three variable models, this requires fixing R_f at 5, 50 (median) and 95 percentile values to visualize how the curves shift. The model summaries also include the ROC curve from testing the model on the training set and its AUC. In each graphic, the histograms summarize the posterior distributions for the regression intercept and the two slopes showing mean, standard deviation and the 5, 50 (median), and 95 percentiles indicated by dashed lines. The hierarchical models are generated using the group averaged coefficients, similar to the "ergodic" coefficents used in ground motion estimation.

We discuss the following models:

- Model 0 – Maximum likelihood model
- Model 1 – Baseline Bayesian model
- Model 2 – Hierarchal Bayesian model
- Model 3 – Bayesian measurement error model
- Model 4 – Combined measurement error and hierarchical model
- We will refer to models with three predictor variables with a −3 after the model number. For example, model 2–3 refers to the hierarchal Bayesian model with all three predictor variables.

3.1. Model Uncertainty

The uncertainty related to the outcome of a binary process, here liquefaction triggering, is sufficiently explained by its probability. The more certain about an event's occurrence the higher probability we assign to it or vice versa. It would not make sense to express a standard deviation to a probability – this would imply that the probability is itself a random variable and violate the fundamental axiom of probability

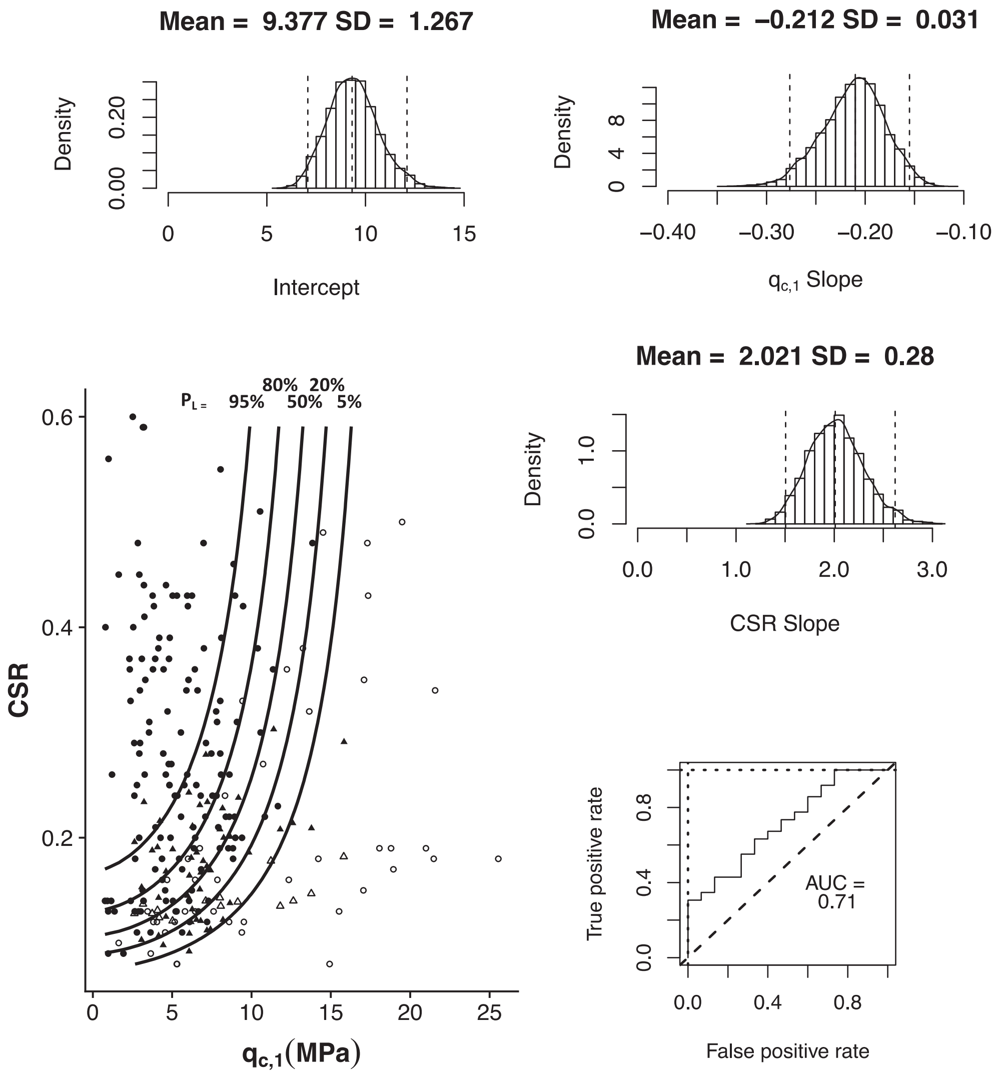

Fig. 6. The baseline Bayesian regression, Model 1–2, that mimics logistic regression using maximum likelihood to optimize the fit to the functional form.

requiring probabilities to be real numbers.

A binary random variable that takes on a value of 1 (yes) with probability p has variance $p(1-p)$ (DeGroot and Schervish, 2012). We define σ_{avg} as the mean computed standard deviation (square root of variance) of an evenly spaced grid of predictors on the domain covered by the database. This metric summarizes how confident the models estimates of probability are, on average. Visually, this appears as the probability contours shifting closer together.

4. Discussion

Table 1 compares the performance metrics between all models. The following discussions are qualitative in nature. Although statistical methods exist for comparing AUC values, selecting and justifying an appropriate one for discriminating between triggering model performance is out of the scope of this paper.

When considering all the models four major trends are apparent:

- The difference in AUC from the worst to best performing model is not large. However, this is expected given the limited number of case histories in the training and testing data.
- Two predictor variables generally outperformed three predictor variables
- The measurement error and hierarchal models outperform the baseline models, but not by a large margin.
- More complex models, both in terms of number of predictors and functional form, tend to have diffuse coefficient posterior distributions. This resulted in greater model uncertainty.

The following sections will discuss the final three trends in further detail.

4.1. *Comparing two and three predictor variables*

The lower than expected predictive performance of the three variable models can be interpreted in several ways. Because the testing set is mostly clean sand cases it is possible that it does not capture the model's overall performance. This is evidenced by the lack of separability between yes and no cases considering only R_f relative to the training set as shown in Fig. 11. In this case a testing set that includes a wider spread of R_f values may show improved predictive performance. Or, it can simply be that R_f as a predictor variable does not generalize well to new data.

Additionally, R_f and $q_{c,1}$ are correlated (Fig. 12) because R_f is a function of tip resistance and can be an input to the equation for the $q_{c},_{1}$ normalization exponent. When predictor variables are correlated the posterior will have a high spread because many logistic surfaces can fit the data leading to poor performance (Kruschke, 2015). Future models

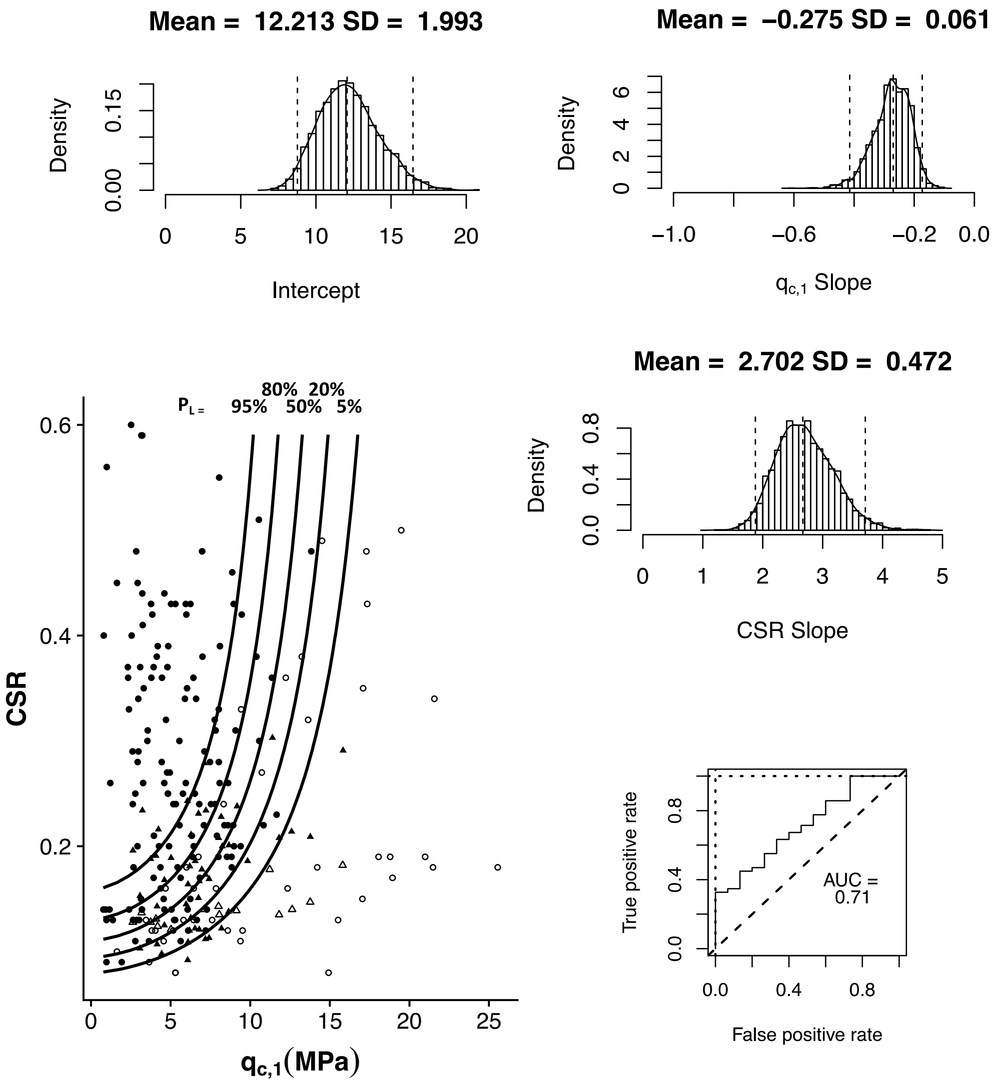

Fig. 7. The Bayesian hierarchical regression model, Model 2–2, that uses mixed or random effects to account for inter- and intra-event uncertainty.

might use sleeve friction directly instead of friction ratio to avoid this.

These results illustrate the importance of distinguishing between model fit to the training data and its actual predictive performance. The AIC based selection indicated that there is statistical utility in including R_f as a predictor variable. However, this metric is based off the model fit to the training data rather than performance on testing data. This may indicate that this predictor selection process may not identify the best candidate models if future predictive performance is the end goal.

4.2. *Comparing model complexity*

For the two variable case, all the non-baseline models showed only slight improvement in performance with the combined model having the highest AUC. The average model uncertainty followed a similar trend. For three variables the combined model showed a greater increase in performance relative to baseline. In general, the hierarchal models (models 2- and 4-) had larger uncertainty metrics than the fully pooled cases (models 1- and 3-). This is consistent with the notion that the pooled models underestimate the uncertainty associated with out of sample predictions but hierarchal models account for event-to-event variability in the coefficients. A result of this is that probability contours spread out in areas with fewer data points (the upper right of the scatter plot). This trend is generally considered one of the benefits of hierarchal models (Gelman and Hill, 2007). More data in this sparse region of “load-resistance” space are required to assess if this contributes or detracts from model predictive ability.

There are a few possible explanations for why the hierarchical and measurement models did not dramatically outperform the baseline. As discussed before, it could be that the testing set may not cover enough of the model space to adequately assess overall performance. A second explanation may be the diffuse posterior estimates result in greater model uncertainty which in turn produces lower model predictive ability. This can be partially explained by the correlation between R_f and $q_{c,1}$; correlated predictors leads to more diffuse posteriors. However, it also appears that particularly for models 3–3 and 4–3 that some or all posterior standard deviations are influenced by the prior. This inference is based on the rule of thumb proposed by the Stan development team that a posterior is “influenced” by a default prior if its standard deviation is greater than 10% of the prior’s (Gelman, 2019). That is, the data alone is not strong enough to constrain the posterior estimates for these parameters. However, the performance increase over baseline indicates these models have promise and can be improved with informative priors and/or more data. This is evidenced by the work of Kuehn and Abrahamson (2018), who used numerical simulation to justify informative priors on regression coefficients and had access to a larger dataset.

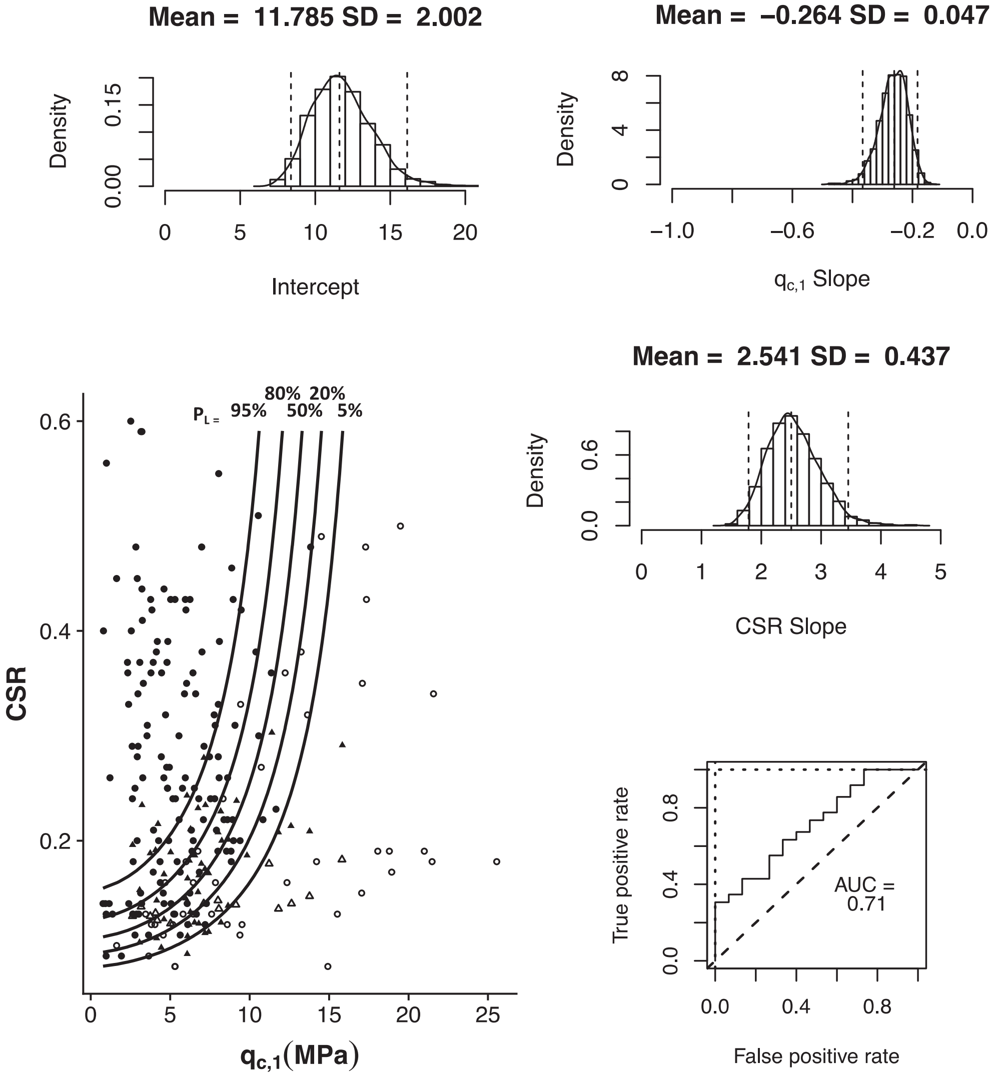

Fig. 8. The Bayesian measurement error model, Model 3–2, that includes parameter uncertainty for each variable.

4.3. *Model uncertainty*

While a reduction in model uncertainty usually leads to better predictive performance there will always be a certain amount of irreducible uncertainty in future observations. There will eventually be a point where models become “over-confident” and lose predictive ability even as the probability contours shrink closer and closer together. This can be seen by example with model 4–3. If we fix the population average parameters at their mean values and ignore the variability in their estimates the result is a model that appears very certain (Fig. 13). The average uncertainty in this case is only 0.046, considerably lower than what is produced by using the full posterior uncertainty. However, the AUC for this model is only 0.649, compared to 0.694 obtained by using the full posterior uncertainty. Thus, the model performs worse when making out of sample predictions despite having lower model uncertainty. The upshot is that future model development can balance seeking lower model uncertainty measures with making realistic out of sample predictions.

4.4. *A brief discussion on the dangers of overfitting*

Overfitting can be conceptually thought of as when a model has learned too much from training data. That is, it has learned overly complex patterns that do not generalize to new cases. Fig. 14 shows the predictive performance Model 4–2 but this time using the original training set for testing. Note that the AUC of Model 4–2 validated using the testing set is only 0.717. This approach, taken by most of the previous work in the field, dramatically overestimates the model’s predictive ability. This is further compounded with practitioners often having to rely on self-reported validation metrics when deciding which is a better model to use.

4.5. *Prior sensitivity study*

To further assess the influence of prior choice we performed a sensitivity study on our models. The default priors, used for the results reported above, are Normal (0,25) on the intercept parameter and Normal (0,10) on the slope parameters. We then tested a second case using a Normal(0,25) prior on the slopes and intercept. The final case considered a Normal(0,100) prior on the slopes and intercept. After recording the change in posterior distributions moments and model predictive performance we found for the 2 variable models the posterior distributions of model coefficients changed only slightly. Both AUC and σ_{50} remained relatively unchanged. The three variables showed a larger increase in the model coefficients posteriors standard deviations (and a shift of mean values). However, similar to the two variable case the predictive performance and model uncertainty remained relatively unchanged. Therefore, this indicates that model performance is not overly

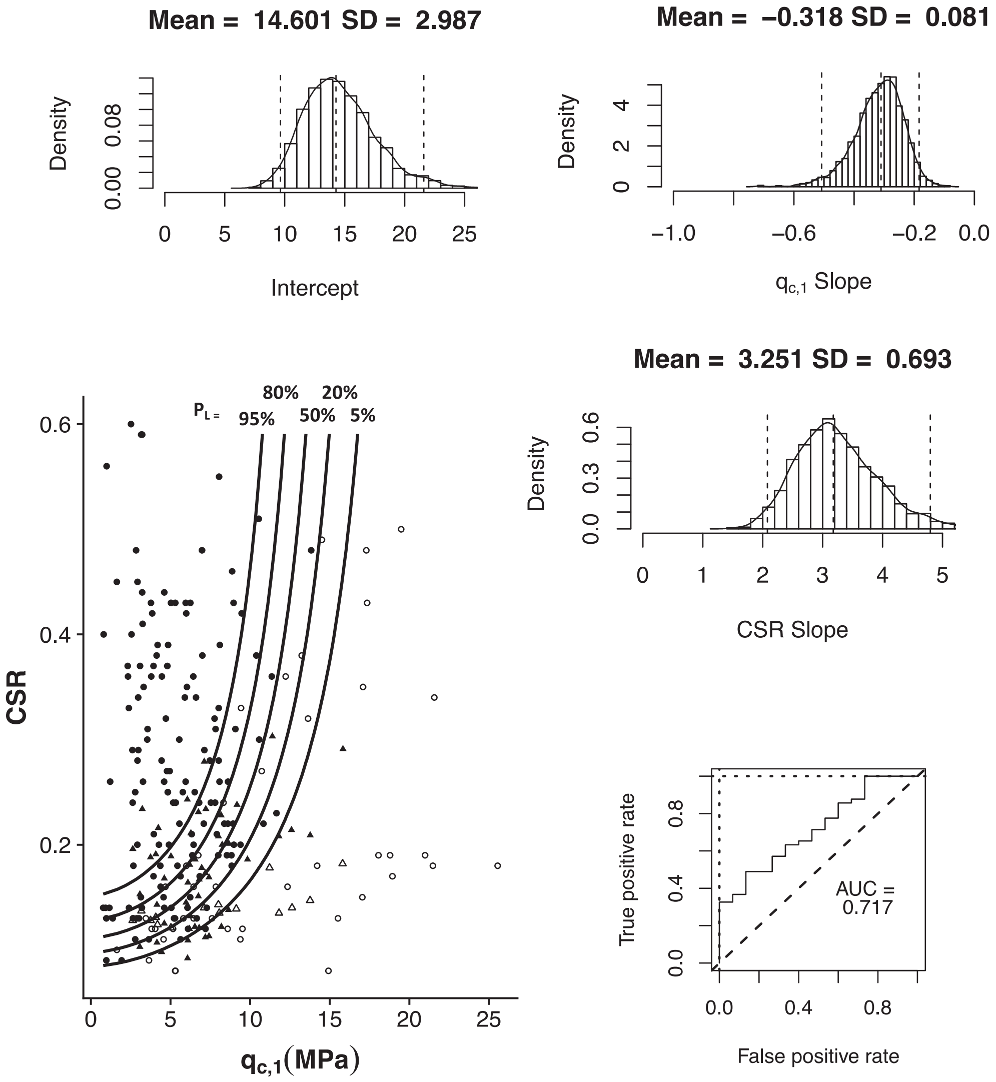

Fig. 9. The combined measurement and hierarchical model, Model 4–2, that utilizes all the features in the Models 2 and 3.

sensitive to more diffuse priors. This study did not consider the effects of tighter priors.

5. Conclusions and Recommendations for Future Work

This study built a predictive modeling workflow for developing empirical models for liquefaction triggering potential. The intent is that this principled statistical workflow can be used for developing new triggering models when data from the NGL project is available. Our workflow and novel functional form address several shortcomings of existing triggering models. Firstly, existing liquefaction models lack open and consistent model building and evaluation approaches. The performance, if assessed at all, is almost always estimated by refitting the model to training data. Our paper summarizes methods that modelers and guidance committees can use to develop principled new models and realistic comparisons of performance on out of sample data.

Secondly, this paper this paper demonstrates how a hierarchical model can be built using Bayesian methods that also accounts for uncertainty in measured or estimated predictors. This framework has not been applied in existing liquefaction triggering models. A hierarchical modeling approach, also called mixed effects, variance components, random effects, and others, will likely be necessary for NGL modeling efforts to account for intra and inter-event correlations and allow for a proportional weighting of the data from large and small sample events.

Finally, the likelihood forms implied by a hierarchical model are appreciably more complex than those used to develop current work (i.e. Boulanger and Idriss, 2016) and will require sophisticated fitting strategies. For example, even our relatively simple hierarchical model with 3 varying slopes and varying intercepts grouped by event cannot be solved by traditional maximum likelihood techniques. More complex models with additional predictor variables and models incorporating measurement error in the predictors will be similarly intractable without the inclusion of prior knowledge. A Bayesian approach can use weakly informative priors to sufficiently constrain the model and arrive at a solution as demonstrated in our work. Additionally, Bayesian priors can be tuned to ensure the model extrapolates according to principles of soil mechanics in regions of data sparsity. This is conceptually similar to Kuehn and Abrahamson (2018) used finite fault simulations to constrain coefficients for ground motion scaling.

We first selected the predictor variables that resulted in the model that best balanced over and underfitting. The best performing model included $q_{c,1}$, CSR, and R_f. Notably, this step indicated that models incorporating magnitude (M_w) and effective stress ($\sigma_v^{'}$), and others were not as effective as the simpler forms. This is seemingly at odds with soil mechanics principles and laboratory studies showing initial stress and number of shear reversals strongly influence liquefaction behavior.

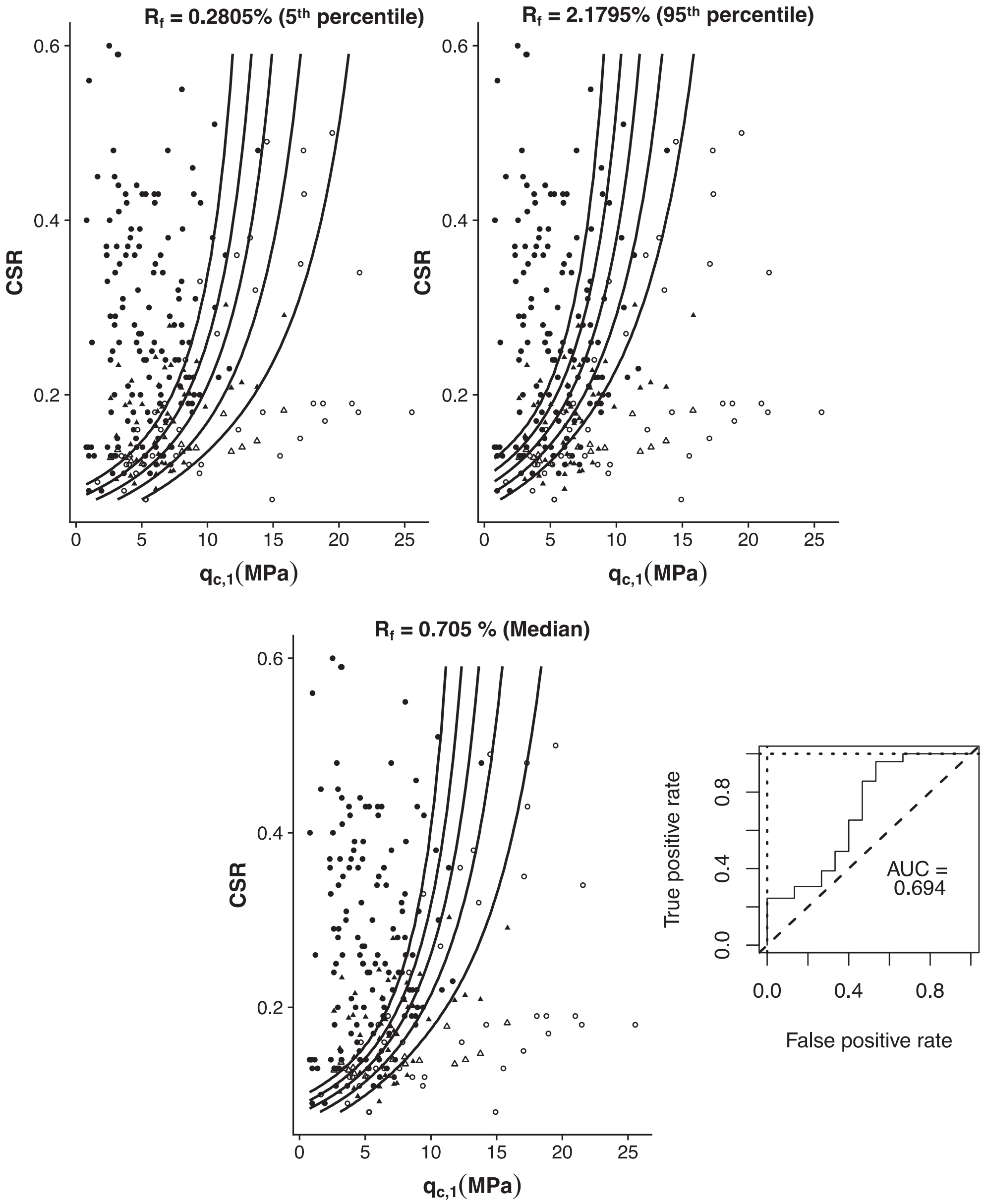

Fig. 10. Combined model with 3 predictor variables, Model 4–3. Probability curves are shown for three fixed values of R$_f$.

Table 1

Model	σ_{avg}	% Difference from baseline	AUC	% Difference from Baseline
0–2	0.12	–	0.699	–
0–3	0.13	8%	0.649	−7%
1–2	0.10	−17%	0.710	2%
1–3	0.13	8%	0.644	−8%
2–2	0.09	−25%	0.710	2%
2–3	0.14	17%	0.694	−1%
3–2	0.09	−25%	0.710	2%
3–3	0.07	−42%	0.644	−8%
4–2	0.09	−25%	0.717	3%
4–3	0.13	8%	0.694	−1%

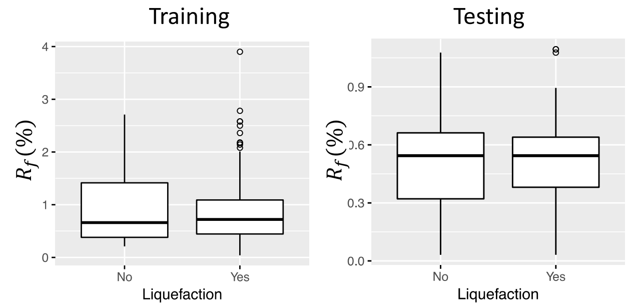

Fig. 11. Separability of Rf for the training (Moss et al., 2006) and testing data (Green et al., 2015).

However, the more useful interpretation is non-included predictors are either poor stand-ins for the underlying mechanical behavior (number of shear reversals, stress states during dynamic loading, etc.), the values included in the database are too similar to discriminate liquefaction occurrence, or the parameters are already embedded in the loading characterization using CSR.

This highlights an important theme that should be considered when developing empirical models. While soil mechanics and other knowledge certainly should guide efforts, it is misleading to view these models

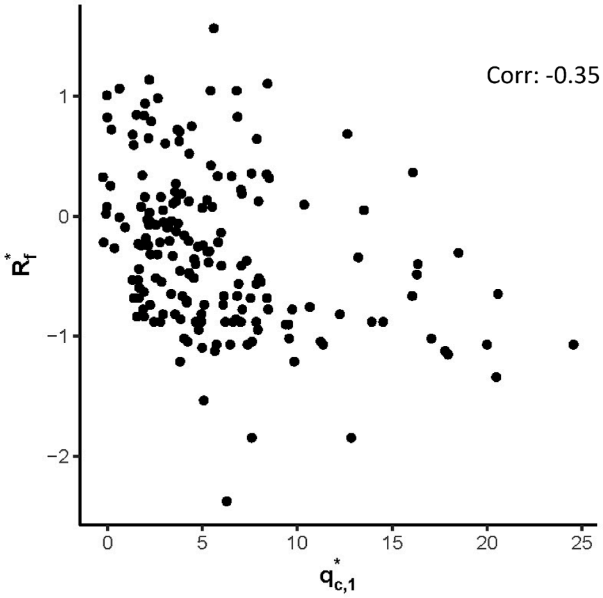

Fig. 12. Correlation between R_f and $q_{c,1}$ in the training data, shown in transformed space.

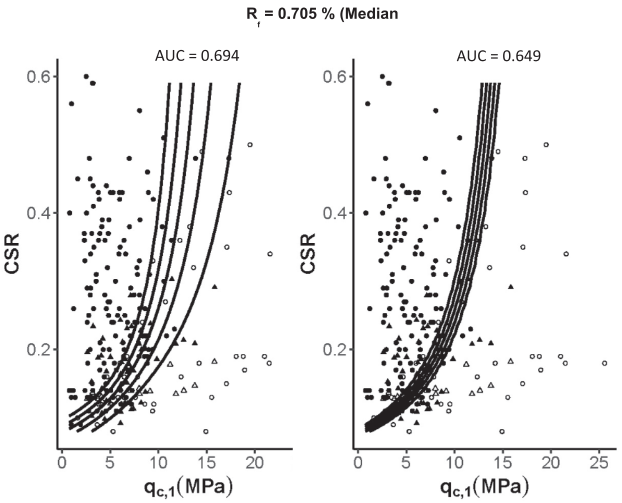

Fig. 13. Overconfident model predictions resulting from neglecting parameter uncertainty (right) versus more realistic predictions from using the full posterior distribution of model parameters for prediction (left).

as describing physical behavior. Empirical models learn the trends in the data using the mathematics of probability and statistics, not by computing stresses and strains. Too strong of an emphasis on “rational” physical interpretation may hamper model development by excluding otherwise strong predictors or parametric forms that lack physical interpretations.

We then selected Box-Cox transformations of these variables that produced the highest cross-validated AUC for both the two and three variable models. In this case, the separation of feature selection and transformation was for ease of computation. Future efforts can refine this approach by combining the two steps, i.e. selecting the best performing predictors and transformations at the same time with the same metric. More flexible combinations of predictors and transformations could also be considered. This will likely result in a large space of possible models and optimization techniques such that simulated annealing or genetic algorithms will be necessary to select promising candidates. After selecting the appropriate combinations and forms of predictors four main models were considered; a baseline logistic regression, a hierarchal (or mixed/random effects) model, a Bayesian measurement error model, and a combination of the last two. The performance of these models was assessed using a testing set of New Zealand case histories and reported as ROC curves and mean standard deviation of predictions (σ_{50}).

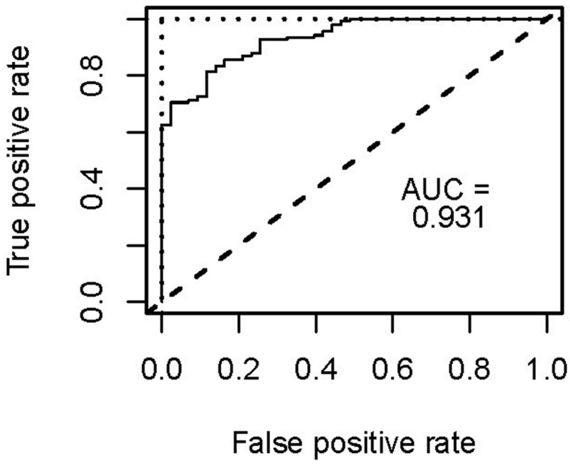

Fig. 14. The erroneous results when overfitting, that is in this case using the training dataset for a testing dataset.

Three main trends are apparent in the results. The two predictor variable models usually outperform those including R_f , though the magnitude of the difference varied. The hierarchical and measurement error models show improvements (both AUC and σ_{avg}) for both two and three predictor variables over the baseline and likely will benefit from informative priors and more data. Hierarchical models should be considered for future work because they can systematically account for inter- and intra- event variability and give improved estimates for imbalanced group sample sizes. They can also be used to develop region specific correlations that still learn from the global database. A Bayesian approach, such as the one described in this paper, will likely be necessary to fit these future models.

Finally, it appears that using a limited single testing set does not give the best view of overall model performance. To remedy this, a cross validation approach that uses all the data independently for training and testing or curation of a more representative test set should be considered. Regardless of the validation metrics used in future work assessing model performance on the data it was built on will lead to overfitting and should not be done.

Funding

Funding was provided from CalTrans (California Department of Transportation) and NRC (Nuclear Regulatory Commission) via PEER (Pacific Earthquake Engineering Research center) and SWRI (South West Research Institute) as part of the NGL (Next Generation Liquefaction) program. Additional funding was provided by California Polytechnic State University via the Graduate Assistant position. The opinions expressed here are the authors’ and are not necessarily those of the funding or research agencies.

CRediT authorship contribution statement

Jonathan Schmidt: Methodology, Software, Formal analysis, Writing - original draft, Data curation. **Robb Moss:** Conceptualization, Resources, Project administration, Funding acquisition, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research primarily used the software packages of Stan and R (R Core Team, 2020) to perform modeling and graphic generation. The non-base R packages used included caret (Kuhn, 2020), ggplot2 (Wickham et al., 2020), rstan (Guo et al., 2020), and ROCR (Sing et al., 2020). The authors would like to thank Dr. Gregg Fiegel, Dr. Kevin Ross, and Dr. Benjamin Turner for reviewing preliminary drafts of this work and the anonymous peer reviewers for providing feedback during the submission process. Several users of the Stan discussion boards (https://discourse.mc-stan.org/) also provided input for model building and interpretation.

Data Availability

The CPT database and the MCMC draws of model posterior solutions are available through Mendeley Data at https://doi.org/10.17632/8jnbnp5m8f.1.

References

Abrahamson, N.A., Youngs, R.R., 1992. A stable algorithm for regression analyses using the random effects model. Bull. Seismol. Soc. Am. 82 (1), 505–510.

Boulanger, R.W., Idriss, I.M., 2016. CPT-based liquefaction triggering procedure. J. Geotech. Geoenviron. Eng. 142 (2), 04015065. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001388.

Box, G.E.P., Cox, D.R., 1964. An Analysis of Transformations. Journal of the Royal Statistical Society 26 (2). https://www.jstor.org/stable/2984418.

Brandenberg, S. J., Zimmaro, P., Stewart, J. P., Kwak, D. P., Franke, K. W., Moss, R. E.S., Cetin, K. O., Can, G., Ilgac, M., Stamatakos, J., Weaver, T., and Kramer, S. L. "Next-generation liquefaction database". Earthquake Spectra, 36(2), 939-959. https://doi.org/10.1177%2F8755293020902477.

Burnham, K.P., Anderson, D.R., 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods and Research 33 (2) https://doi.org/10.1177%2F0049124104268644.

Brooks, S.P., Gelman, A., 1997. General Methods for Monitoring Convergence of Iterative Simulations. J. Comput. Graphical Statistics 7, 434–455.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: a probabilistic programming language. J. Stat. Softw. 76 (1) https://doi.org/10.18637/jss.v076.i01.

Cetin, K.O., Seed, R.B., Der Kiureghian, A., Tokimatsu, K., Harder, L.F., Kayen, R.E., Moss, R.E.S., 2004. Standard penetration test-based probabilistic and deterministic assessment of seismic soil liquefaction potential. J. Geotech. Geoenviron. Eng. 130 (12), 1314–1340. https://doi.org/10.1061/(ASCE)1090-0241(2004)130:12(1314).

Christensen, R., Johnson, W., Branscum, A., Hanson, T.E., 2011. Bayesian ideas and data analysis: an introduction for scientists and statisticians. Taylor and Francis, New York, NY.

Clark, T.S., Linzer, D.A., 2015. Should I use fixed or random effects? Political Sci. Res. Meth. 3 (02), 399–408. https://doi.org/10.1017/psrm.2014.32.

Davis, J., & Goadrich, M. (2006). "The relationship between precision-recall and ROC curves". Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

DeGroot, M.H., Schervish, M.J., 2012. Probability and Statistics. Pearson, Boston, MA.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Gelman, A. (2019). "Prior choice recommendations." Stan development wiki. < https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>. Accessed 31 May 2019.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 1 (3), 515–534. https://doi.org/10.1214/06-BA117A.

Gelman, A., Hill, J., 2007. Data analysis using regression and multilevel/hierarchical models. NY, Cambridge University Press, New York.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical Sci. 7 (4), 457–472. https://doi.org/10.1214/ss/1177011136.

Gelman, A., Jakulin, A., Pittau, M.G., Su, Y., 2008. A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Statistics 2 (4), 1360–1383. https://doi.org/10.1214/08-AOAS191.

Green, R.A., Cubrinovski, M., Cox, B., Wood, C., Wotherspoon, L., Bradley, B., Maurer, B., 2014. Select liquefaction case histories from the 2010–2011 Canterbury earthquake sequence. Earthquake Spectra 20 (1), 131–153. https://doi.org/10.1193/030713EQS066M.

Guo, J., Gabry, J., Goodrich, B., 2020. Rstan: R interface to Stan. R package version 2 (19), 3.

Jiang, J., 2007. Linear and generalized linear mixed models and their applications. Springer, New York, New York, NY.

Juang, C.H., Jiang, T., Andrus, R.D., 2002. Assessing probability-based methods for liquefaction potential evaluation. J. Geotech. Geoenviron. Eng. 128 (7), 580–589. https://doi.org/10.1061/(ASCE)1090-0241(2002)128:7(580).

Kruschke, J.K., 2015. Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan. Academic Press, Boston.

Kuehn, N.M., Scherbaum, F., 2015. Ground-motion prediction model building: a multilevel approach. Bull. Earthq. Eng. 13 (9), 2481–2491. https://doi.org/10.1007/s10518-015-9732-3.

Kuehn, N.M., Abrahamson, N.A., 2018. The effect of uncertainty in predictor variables on the estimation of ground-motion prediction equations. Bull. Seismol. Soc. Am. 108 (1), 358–370. https://doi.org/10.1785/0120170166.

Kuhn, M. (2020). Caret: Classification and Regression Training. R packages version 6.0-86.

Kuhn, M., Johnson, K., 2013. Applied predictive modeling. Springer, New York, New York, NY https://doi.org/10.1007/978-1-4614-6849-3.

Lai, S.Y., Chang, W.J., Lin, P.S., 2006. Logistic regression model for evaluating soil liquefaction probability using CPT data. J. Geotech. Geoenviron. Eng. 132 (6), 694–704. https://doi.org/10.1061/(ASCE)1090-0241(2006)132:6(694).

Liao, S.S.C., Veneziano, D., Whitman, R.V., 1988. Regression models for evaluating liquefaction probability. J. Geotech. Eng. 114 (4), 389–411. https://doi.org/10.1061/(ASCE)0733-9410(1988)114:4(389).

Moss, R.E.S., 2020. Applied Civil Engineering Risk Analysis, 2nd Edition. Springer.

Moss, R.E., Seed, R.B., Kayen, R.E., Stewart, J.P., Der Kiureghian, A., Cetin, K.O., 2006. CPT-Based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential. J. Geotech. Geoenviron. Eng. 132 (8), 1032–1051. https://doi.org/10.1061/(ASCE)1090-0241(2006)132:8(1032).

National Academies Press., 2016. State of the art and practice in the assessment of earthquake-induced soil liquefaction and its consequences, The National Academies Press, Washington, DC.

Oommen, T., Baise, L.G., Vogel, R., 2010. Validation and application of empirical liquefaction models. J. Geotech. Geoenviron. Eng. 136 (12), 1618–1633. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000395.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rezania, M., Faramarzi, A., Javadi, A.A., 2011. An evolutionary based approach for assessment of earthquake-induced liquefaction and lateral displacement. Eng. Appl. Artif. Intell. 24, 142–153. https://doi.org/10.1016/j.engappai.2010.09.010.

Saito, T., Rehmsmeir, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10 (3), e0118432. https://doi.org/10.1371/journal.pone.0118432.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., Unterthiner, T., and Ernst, F. G. M. (2020). ROCR: Visualizing the Performance of Scoring Classifiers. R package version 1.0-11.

Schmidt, J. (2020). "A Predictive Modeling Approach for Assessing Seismic Soil Liquefaction Potential Using CPT Data". Master's Thesis. https://digitalcommons.calpoly.edu/theses/2055.

Tosteson, N.A., Fryback, D.G., Hammond, C.S., Hanna, L.G., Grove, R.G., Brown, M., Wang, Q., Lindfors, K., Pisano, E., 2014. Consequences of false-positive screening mammograms. JAMA Internal Medicine 174 (6), 954–961. https://doi.org/10.1001/jamainternmed.2014.981.

Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahasi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., 2020. Ggplot2: Create Elegant Data Visualizations Using the Grammar of Graphics. R package version 3 (3), 2.

Yazdi, J.S., Moss, R.E.S., 2017. Nonparametric liquefaction triggering and postliquefaction deformations. J. Geotech. Geoenviron. Eng. 143 (3) https://doi.org/10.1061/(ASCE)GT.1943-5606.0001605.