Explainable Machine Learning Interpretations on New Zealand Random Forest Liquefaction Manifestation Predictions

Katherine Cheng, M.S. [i], Pablo Busch, M.S.i[ii] and Katerina Ziotopoulou, Ph.D., P.E. [iii]

i) Ph.D Candidate, Department of Civil and Environmental Engineering, University of California, Davis, Davis, CA 95616, USA.
ii) Ph.D Student, Energy and Efficiency Institute, University of California, Davis, Davis, CA 95616, USA.
iii) Associate Professor, Department of Civil and Environmental Engineering, University of California, Davis, Davis, CA 95616, USA.

ABSTRACT

The abundant post-earthquake data from the Canterbury, New Zealand (NZ) area is poised for use with machine learning (ML) to further advance our ability to better predict and understand the effects of liquefaction. Liquefaction manifestation is one of the identifiable effects of liquefaction, a nonlinear phenomenon that is still not well understood. ML algorithms are often termed as "black-box" models that have little to no explainability for the resultant predictions, making them difficult for use in practice. With the SHapley Additive exPlanations (SHAP) algorithm wrapper, mathematically backed explanations can be fit to the model to track input feature influences on the final prediction. In this paper, Random Forest (RF) is chosen as the ML model to be utilized as it is a powerful non-parametric classification model, then SHAP is applied to calculate explanations for the predictions at a global and local feature scale. The RF model hyperparameters are optimized with a two-step grid search and a five-fold cross-validation to avoid overfitting. The overall model accuracy is 71% over six ordinal categories predicting the Canterbury Earthquake Sequence measurements from 2010, 2011, and 2016. Insights from the SHAP application onto the RF model include the influences of PGA, GWT depths, and SBTs for each ordinal class prediction. This preliminary exploration using SHAP can pave the way for both reinforcing the performance of current ML models by comparing to previous knowledge and using it as a discovery tool for identifying which research areas are pertinent to unlocking more understanding of liquefaction mechanics.

Keywords: liquefaction, Canterbury earthquake sequence, Random Forest, explainable machine learning

1 INTRODUCTION

Liquefaction is a nonlinear phenomenon with many potential input parameters that vary across multiple scales, making it difficult to isolate. There is a fundamental understanding of liquefaction at the element level, however liquefaction at the system level introduces complexities that pose additional challenges. For example, studies of the response of a singular sand via an exhaustive direct simple shear (DSS) testing program are not straightforward to upscale to the field, since a liquefiable site operates as a system of multiple layers and/or inclusions that may affect the observed manifestations (e.g., Cubrinovski et al. 2019, Bassal et al. 2022). Therefore, case history studies become invaluable opportunities for studying multiple aspects of liquefaction.

New Zealand (NZ) is a seismically active island country and has experienced several major earthquakes. The 2010/2011 Canterbury earthquake sequence (CES) started with the M_w7.1 Darfield earthquake (September 4, 2010) and was followed by up to ten events which induced liquefaction in the affected region (Maurer et al. 2015). The depositional history of the soils was unable to be captured solely by traditional Cone Penetration Test (CPT) based liquefaction triggering assessments, which resulted in numerous false positive or false negative predictions of liquefaction manifestation occurrences (Beyzaei et al. 2018). The combination of available information on these complex deposits with well documented liquefaction manifestation data yielded a set of seminal case histories (Geyin et al. 2021).

This new dataset allows for the liquefaction phenomenon to be explored with more advanced techniques, such as machine learning (ML). A subset of Artificial Intelligence (AI), ML uses computer algorithms to make predictions on given data without explicit programming instructions. A ML algorithm can find patterns and relationships among the data of various distributions without explicit guidance and could potentially discover associations that may otherwise be missed. One of the drawbacks of utilizing ML is the inability to explain the reasoning behind the resultant predictions, thus termed a "black-box" model. By utilizing SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017), feature importance and contributions to individual predictions can be extracted

and identified from complex ML models such as Random Forest (RF). Through these explanations, insights can be gained from the effect on predicted liquefaction class of outliers or feature interactions in ML models.

This paper investigates the application of the RF ML algorithm coupled with explainable ML SHAP on an established large dataset from NZ. The feature engineering and hyperparameter tuning processes for building the ML model are demonstrated and established. The work presented uses post-processed data from the September 2010, February 2011, and February 2016 Canterbury, NZ Earthquakes (Geyin et al. 2021) to predict different severities of liquefaction manifestation in the form of multiple categories via the RF ML algorithm. The paper first introduces the dataset and feature engineering, then an overview of the RF algorithm, hyperparameters of RF, and the theory behind SHAP. The fitted RF model is then presented along with SHAP results, which are then evaluated and explained at the global and local level. The paper concludes with lessons learned through SHAP from the RF algorithm and the potential usage of SHAP in future geotechnical engineering ML application work.

2 DATASET AND FEATURE ENGINEERING

The dataset is composed of earthquake and CPT data from the September 2010, February 2011, and February 2016 earthquakes in Canterbury, NZ (Geyin et al. 2021). The number of CPT observations is around 5,600. It is a structured dataset containing typical CPT data, recorded peak ground acceleration (PGA) in *g*, and groundwater table (GWT) depth in *m* for each CPT location per earthquake, and the severity of liquefaction manifestation experienced at each CPT. The liquefaction manifestation class is the prediction target of the RF model. It can take the following values: 0 (no manifestation), 1 (minor), 2 (moderate), 3 (severe), 4 (lateral spreading), 5 (severe lateral spreading), and 10 (unknown). Further details of the separation of categories can be found in Geyin et al. (2021). All records corresponding to manifestation Class 10 were removed from the dataset, as they are essentially unknown values that took up only 5% of the total dataset.

Feature engineering is the practice of extracting features from raw data (such as CPT data) with the help of domain knowledge. Only the 4 m of soil below the GWT were considered as that is the depth of soil with the most impact on manifestations during the NZ earthquakes (Cubrinovski and Robinson, 2016). SBT was calculated via Robertson's (2010) cutoffs for each soil type with an assumption of 18 kN/m^3 for the saturated soil unit weight. The SBT categories consist of six types of soils (with label number): organic soils (2), clay (3), clay silt mixtures (4), sand mixtures (5), sands (6), and gravelly to dense sands (7). The layer of soil above the GWT was termed as the "crust" layer, and the average SBT was labeled for it. The CPT for 4 m beyond the GWT was discretized into 0.5m layers, with each layer's SBT Class corresponding to the majority SBT in the given layer. Layer 1 is the topmost layer, layer 8 is the bottommost, with 0.5m increments in between.

The corresponding surficial geology at each CPT was extracted via ArcGIS from the GNS Science (2020) New Zealand Geological Map as an additional feature. The geology covered (with label numbers): Middle Pleistocene loess deposits (0), active dune deposits (1), active riverbed deposits (2), Anthropic deposits (3), Holocene Estuary deposits (4), Holocene river deposits (5), Holocene stable dune deposits (6), Holocene swamp deposits (7). The GWT fluctuates per year and the PGA is different for each earthquake, therefore a different set of data for the same CPTs is created for each earthquake year due to the subsequent feature engineering, roughly tripling the base dataset. After this process of feature engineering the original dataset with all three earthquakes, the new dataset contained around 11,500 observations. The dataset is imbalanced, with the percentage counts: Class 0) 60%, Class 1) 18%, Class 2) 13%, Class 3) 3%, Class 4) 4%, Class 5) 2%, and was preserved as such, as attempts to balance the data by undersampling only served to worsen any interpretations and predictions. Upsampling was not used as there was no guarantee of physics being respected when artificial datapoints are generated. The features were all preserved in their original units and no feature transformation (e.g., squaring) was used as the intention was to keep consistent physical units throughout the whole process.

3 RANDOM FOREST ALGORITHM

The method chosen to predict liquefaction manifestation categories is the Random Forest (RF) algorithm implemented in Scikit- learn (Pedregosa et al. 2011). The RF algorithm is an extension of the decision tree algorithm, in that it fits multiple decision trees and uses a majority vote approach between the trained decision trees to make its predictions. The overall goal of the decision tree algorithm is to use the supplied predictor variables and recursively divide the dataset into homogeneous groups to accurately predict the response variable. In the fitting process of the RF algorithm, several hyperparameters are specified before training, such as the number of decision trees. Sampling with replacement (bootstrapping) at a certain percentage of the training set size is used to create new training datasets that are assigned to each of the individual decision trees, with each training set being unique from one another due to the bootstrapping method. These bootstrapped datasets contain only a subset of the overall predictors in the training set both from the size and replacement sampling restrictions. From there, the decision trees each begin a recursive split process on their assigned datasets to create a set of rules that determine a given observation's classification. Each

decision tree path to a split is called a branch, and where a branch splits is called a node, with a terminal node called a leaf node. In the training process for each decision tree, the algorithm first determines an optimal variable to split upon, and then an optimal value within the variable to separate the data. This process repeats recursively until various stop criteria have been reached. Some of the stop criteria for a decision tree are:

- **Minimum split size**: A threshold for how small the number of observations within each subgroup of data should be to continue splitting.
- **Minimum node size**: A threshold for how small the number of observations in a node should be to terminate the branch and convert the node into a leaf node.
- **Tree depth**: The number of recursive splits or nodes used to build the tree.

Once the controlling stop criterion is reached for each decision tree, each of the final splits within the decision trees are determined. For a new data point (observation), the RF algorithm checks the resulting classification from each decision tree in its "forest" and then uses a majority vote approach to predict the observation's final classification. RF was chosen as the ML model to use because: (i)Variable interactions are preserved in the RF algorithm process which is pertinent for liquefaction analysis; (ii) RF is a non-parametric method as it has no prior data distribution (e.g., Normal) requirements for the algorithm and is known to be able to fit nonlinear parameter relationships (Brieman, 2001); (iii) RF yields low bias with only moderate variance which indicates that the predictions will tend to center around the correct Class.

3.1 Hyperparameter Training

The dataset was split into two portions with 80% of it for training and 20% of it for testing. To improve the performance of the RF algorithm the hyperparameters were optimized in two stages. First, a random hyperparameter grid search with five- fold cross-validation (CV) was run on the training set over a wide range of hyperparameter values. Then, a local exhaustive grid search with five-fold CV was run with finer hyperparameter values in a close range around the previously found best hyperparameters. The hyperparameter search was restricted in range and subjected to five-fold cross-validation to prevent the RF algorithm from overfitting. The final best hyperparameters are: Max depth (90), Max samples (90%), Min samples per leaf (2), Min samples per split (5), Number of estimators (500). Bootstrapping was always kept as the default to reduce overly memorizing the training set since RF only uses a randomly drawn set with replacement from the training set datapoints. These measures lowered the accuracy from the training set but resulted in higher test set accuracies and a smaller gap between the two accuracies indicating an absence of overfitting.

4 SHAP

SHAP (Lundberg and Lee 2017) allows ML models to achieve both accuracy and interpretability. SHAP works as a wrapper around the original ML model, probing the ML model to calculate the contributions and interactions of the input features on the final prediction and can be applied to most ML models. The background of this technique originates from game theory as an additive feature attribution method, meaning the predicted output is a linear combination of the input features as showcased in Equation 1.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \quad (1)$$

where $z' \in 0,1^M$*, M is the number of simplified input features, and* $\phi_i \in R$ (Lundberg and Lee 2017)

SHAP values are to satisfy three criteria: 1) Local accuracy: The explanation model output should match that of the original ML model. 2) Missingness: If a feature value is 0, then it should reflect as such in the explanation model as a 0 for the influence value. 3) Consistency: The explanation model should consistently reflect any changes as the ML model changes. A unique solution has been proven to satisfy these three criteria as Equation 2.

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(z' \backslash i)] \quad (2)$$

where $|z'|$ *is the number of non-zero entries in* z'*, and* $z' \subseteq x'$ *represents all* z' *vectors where the non-zero entries are a subset of the non-zero entries in* x' (Lundberg and Lee 2017)

As Equation 2 is computationally intensive to solve, there are approximation SHAP algorithms, such as KernelSHAP and TreeSHAP. This paper utilizes TreeSHAP as it is designed for tree-based algorithms, such as RF, and efficiently calculates SHAP values for all input features. Utilizing SHAP allows us to explore model biases, outlier effects, and trends within input feature values that can hint towards their overall influence on the model.

5 RESULTS

The final RF model has an overall accuracy of 71% on the test set. Additionally, the dataset is imbalanced and therefore the per class accuracy is a more holistic view of model performance. The prediction accuracy per manifestation level was: 93.5 for Class 0, 39.4% for Class 1, 57.9% for Class 2, 14.1% for Class 3, 9.7% for Class 4 and 2.3% for Class 5, with the confusion matrix in Fig. 1. The latter classes 4 and 5 have poor accuracies, with many mispredictions to classes 1, 2, and 3 instead. This is due to the difficulty of solving two fundamental problems at once: liquefaction (class 0-3) or lateral spreading (class 4-5). The two are tied as lateral spreading is liquefaction but have been separated out in the original dataset into different classes. It is evident

that this RF model has learned liquefaction but cannot discern between lateral spreading yet with our limited set of data and input features.

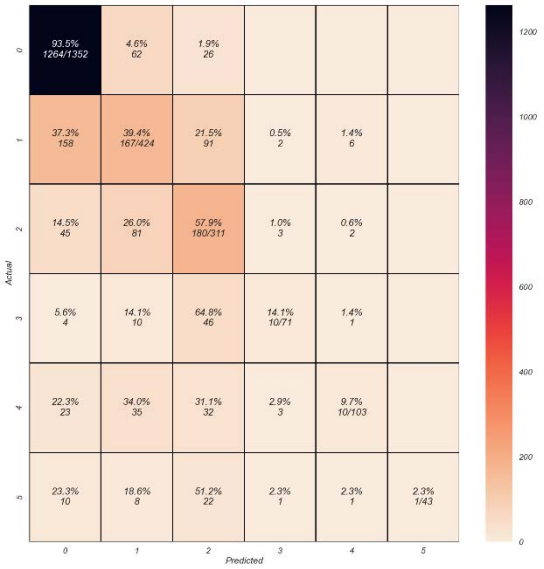

Fig. 1. Confusion matrix of final model, correct predictions are aligned along the diagonal. The percentage value in each box is the percentage predicted from the actual class (left) into the predicted class (bottom) out of the total actual class. The numerical value in each box is the number of datapoints predicted into the prediction categories.

The SHAP feature importance plot in Fig. 2 showcases the impact of each feature on the predicted class. Here, we can see that PGA has a large impact on Class 0 (no liquefaction) but not as much on Class 5 (severe lateral spreading), but it still has the greatest impact out of all the features overall. This is reasonable considering it is the imposed demand while all other features are essentially capacity. The per feature impact for the following other input features vary, with the overall size of the bar corresponding to how important the feature is for the final prediction with the segmented sections indicating its importance per class. GWT and geology are both more important than the SBT of individual layers, including the crust.

Delving into the individual class prediction SHAP summary plots in Fig. 3, previously hidden trends are uncovered. The color on the SHAP summary plot indicates the input feature value, with red being a high value (e.g., 5) and blue a low value (e.g., 0) relative to the feature's range. The x-axis indicates the impact in log-odds on the final prediction, with a high log-odds value (towards the right) indicating a bigger input feature impact on the probability of the given class being the correct final predicted class. The ordering of input features along the y-axis is based on the most impactful input feature on top and least impactful on bottom.

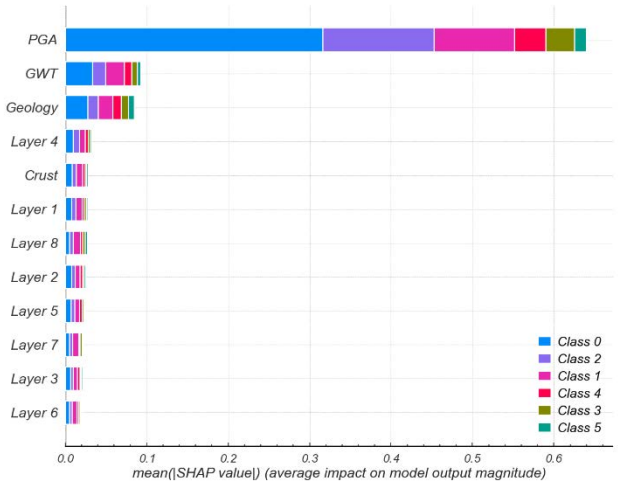

Fig. 2. SHAP feature importance plot, explanation in text.

From Fig. 3, for a high log-odds probability of Class 0 (no manifestation) the input features of a low PGA, deep GWT, and Holocene era geology are identified. The following summarizes the takeaways on the preferential input features for the rest of the classes: Class 1) high PGA, shallow GWT, and Holocene era geology. Class 2) high PGA, shallow GWT, Anthropic or active deposits. Class 3) high PGA, shallow to mid-depth GWT, active deposits. Class 4) high PGA, Anthropocene or fill geology, slight preference towards deep GWT. Class 5) high PGA, Anthropocene or fill geology, shallow to mid-depth GWT.

As the geology input feature becomes second most important for predicting Class 4 and 5, these classes rely more heavily on geology than classes 0-3. Low PGA only matters greatly for Class 0, while the rest have varying spreads of how impactful a high PGA is. Shallow GWT matters for classes 1,2,3, and less for classes 4,5, which hints that other unobserved features are at play for lateral spreading. The shift in geology importance notes that younger deposits (active or Anthropocene) tend to be looser and have a higher possibility of liquefaction than older deposits that have had time to consolidate and densify. These observations drawn from the SHAP values match with what is known in literature on liquefaction (Idriss and Boulanger 2008), showcasing the RF model's ability to place importance on input features without prior knowledge introduced.

Focusing on the SBTs of the crust and subsequent 0.5m layers, the preferences for which SBTs in each layers are influential towards the log-odds probability of the predicted Class occurring also change: Class 0) preference for sands at all layers, except layer 5 which has a clay preference. Class 1) preference for clays in all layers. Class 2) interlayering preference, with a preference for sands in the crust. Class 3) slight preference for clays in layer 1 and 2, with sands for the rest. Class 4) no discernable preference, with less emphasis on clays for layers 5, 6, 8. Class 5) no

discernable preference across layers. For Class 4 and 5 there is a narrower range for the x-axis, as the x-axis (range of log-odds) shrinks with increasing classes. This means the same input feature value has a smaller impact on the log-odds for predicting Class 5 than Class 0, which can be explored in detail in Fig. 4.

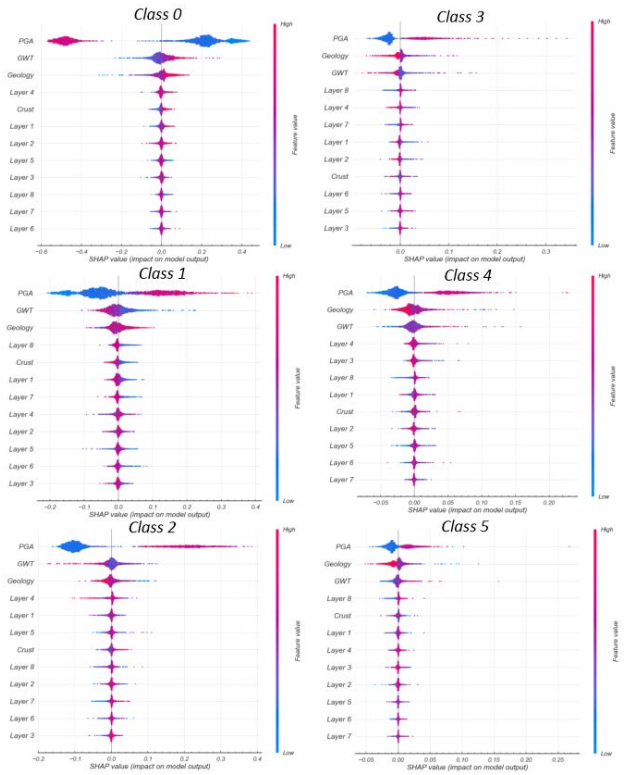

Fig. 3. SHAP summary plots per predictive Class, explanation in text.

Fig. 4. SHAP plots per predictive class for a correctly classified Class 0 datapoint. Explanation in text.

The x-axis in Fig. 4 is again the log-odds probability of the predicted class being the correct class. There is a base starting value for each class log-odds prediction which is equivalent to the expected value or bias of the class. The subsequent input feature values will then impact the log-odds probability by adding or subtracting from the base value. The impact per feature depends on the importance of the feature under the predicting class as seen in the overall Fig. 3 breakdown.

From the axis, the expected values per Class are: Class 0) 0.60, Class 1) 0.17, Class 2) 0.13, Class 3) 0.03, Class 4) 0.046, Class 5) 0.017. There is a large bias towards Class 0, which is expected as it was the overarching Class (60% of the dataset). The classes with lower accuracy (3,4,5) show lower bias, which makes it less likely for the model to predict these classes and explains the accuracy values.

The chosen individual datapoint in Fig. 4 to be inspected is a Class 0 that was correctly predicted as a Class 0. For this datapoint after the input features were added the log-odds per potential Class prediction settle and can be compared. The SHAP values push the log-odds per Class to strongly indicates a Class 0 prediction to be the most likely one, which has a final log-odds of 0.98 compared to that of the other classes that are all below 0.01. Based on the log-odds, Class 0 was the final predicted Class by the RF model, and thus this datapoint was correctly classified.

Fig. 5. SHAP plots per predictive class for an incorrectly classified as a Class 1 instead of Class 2 datapoint. Explanation in text.

SHAP not only helps track how the correct classification was achieved, but it can also be used to track how misclassifications happened or how close correct classifications were to being incorrect. Observing a different datapoint's SHAP individual plots in Fig. 5, one that was misclassified as a Class 1 rather than Class 2, reveals the misprediction path. Class 1 has a higher log-odds (0.4) than that of Class 2 (0.3) with the rest of the classes log-odds are all lower than 0.17. This indicates that the RF model was close to predicting Class 2, but the SHAP values show how the input features contributed to a greater log-odds for Class 1 with a greater impact from the same PGA of 0.582. This inspection of mispredictions through SHAP values leads

to informed next steps to course correct the RF model. The RF model overall does well at recognizing it is not any of the other classes, and the log-odds being close between Class 1 and 2 indicate that with some further improvements the RF model can possibly gain accuracy by bridging the log-odds impact gap.

6 CONCLUSION

Explainable ML (SHAP) was applied to a Random Forest (RF) classification algorithm to predict multiclass liquefaction from the Geyin et al. (2021) database of observations from three New Zealand earthquakes. The RF algorithm performed with an overall accuracy of 70.8%, while SHAP calculated a granular breakdown of input feature influences on final predictions for all potential predicted classes. The main observations from the SHAP breakdowns are that: low PGA is a large contributor to no liquefaction manifestation, increasingly shallow GWT levels mirrors increasing liquefaction manifestation, younger geological formations have a greater tendency to show liquefaction manifestation, and sand layers near the upper soil layers influence greater liquefaction manifestation.

This initial exploration with SHAP illustrates how previously black-box ML models can be transformed into white-box ones, with observations that match those known from literature. As data sharing and capture increase, ML models can act as a first-pass approach to handle large amounts of data with complex relationships. SHAP can then be applied to elucidate granular observations for which input features are impactful at what levels or understand prediction paths for both correctly predicted and mispredicted datapoints. Any new or unusual feature impacts revealed through SHAP from originally complex ML models can be used to identify areas of research to further pursue, whether it be physical experiments or constitutive modeling to parse out more insights through traditional geotechnical methods.

Future work will explore additional input features, such as slope or elevation, to increase accuracies and mitigate bias. There are also per-class accuracies that are lacking that will be addressed by returning to literature and reevaluating the misclassified points to see if it is an indication of a greater database misclassification. Finally, feature interactions will be explored with SHAP, to see if combining certain values of features (e.g., GWT and PGA) may lead to different effects than if independently input into the ML model.

ACKNOWLEDGEMENTS

The authors acknowledge Martin Yossifov for helping gather the data for geological units. This material is based upon work primarily supported by the National Science Foundation (NSF) under NSF Award Number CMMI-2047838. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the NSF.

REFERENCES

1) Bassal, P. C., Boulanger, R. W., and DeJong, J. T. (2022): System Response of an Interlayered Deposit with Spatially Distributed Ground Deformations in the Chi-Chi Earthquake. *Journal of Geotechnical and Geoenvironmental Engineering*, 148 (10), 05022004., doi: 10.1061/(ASCE)GT.1943-5606.0002869
2) Beyzaei, C. Z., Bray, J. D., van Ballegooy, S., Cubrinovski, M., and Bastin, S. (2018): Depositional environment effects on observed liquefaction performance in silt swamps during the Canterbury earthquake sequence. *Soil Dynamics and Earthquake Engineering*, 107, 303–321., doi: 10.1016/j.soildyn.2018.01.035
3) Breiman, L. (2001): Random forests. Machine Learning, 45(1), 5–32. , doi: 10.1023/A:1010933404324
4) Cubrinovski, M., and Robinson, K. (2016): Lateral spreading: Evidence and interpretation from the 2010–2011 Christchurch earthquakes. *Soil Dynamics and Earthquake Engineering*, 91, 187–201., doi: 10.1016/j.soildyn.2016.09.045
5) Cubrinovski, M., Rhodes, A., Ntritsos, N., and Van Ballegooy, S. (2019): System response of liquefiable deposits. *Soil Dynamics and Earthquake Engineering*, 124, 212–229. , doi: 10.1016/j.soildyn.2018.05.013
6) Geyin, M., Maurer, B. W., Bradley, B. A., Green, R. A., and van Ballegooy, S. (2021): CPT-based liquefaction case histories compiled from three earthquakes in Canterbury, New Zealand. *Earthquake Spectra,* 37(4), 2920–2945., doi:10.1177/8755293021996367
7) GNS Science (2020): New Zealand Geological Map. Website: https://www.gns.cri.nz/data-and-resources/geological-map-of-new-zealand/
8) Idriss, I. M., and Boulanger, R. W. (2008): *Soil liquefaction during earthquakes*. Earthquake Engineering Research Institute.
9) Lundberg, S. M., and Lee, S.-I. (2017): A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
10) Maurer, B. W., Green, R. A., Cubrinovski, M., and Bradley, B. A. (2015): Assessment of CPT-based methods for liquefaction evaluation in a liquefaction potential index framework. *Géotechnique*, 65(5), 328–336., doi:10.1680/geot.SIP.15.P.007
11) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011): Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830., doi: 10.5555/1953048.2078195
12) Robertson, P. K. (2010): Soil behaviour type from the CPT: An update. *2nd International Symposium on Cone Penetration Testing*, (pp. 56)
13) Van Ballegooy, S., Malan, P., Lacrosse, V., Jacka, M. E., Cubrinovski, M., Bray, J. D., O'Rourke, T. D., Crawford, S. A., and Cowan, H. (2014): Assessment of Liquefaction-Induced Land Damage for Residential Christchurch. *Earthquake Spectra*, 30(1), 31–55., doi: 10.1193/031813EQS070M