Proceedings of ISFOG 2025
5TH INTERNATIONAL SYMPOSIUM ON
FRONTIERS IN OFFSHORE GEOTECHNICS
Nantes, France | June 9-13 2025

ISBN 978-2-85782-758-0

Advancing offshore site characterisation practices through intelligent data utilisation

T. Zheng*, R. Buckley, E. Febrianto
University of Glasgow, Glasgow, UK

A. Lowther, P. Stock
Scottish Power Renewables, London, UK

M. Rattley, S. Whyte
Geowynd, London, UK

t.zheng.1@research.gla.ac.uk

ABSTRACT: Site investigation (SI) campaigns for offshore infrastructure require substantial vessel time and resource, invoking £10s of millions and large carbon footprints. Investigations typically cover large areas, utilise a wide range of testing methods, and involve multiple phases. Despite these complexities, the planning, interpretation, and integration of datasets is still relatively manual. Furthermore, characterisation methods (e.g., parameter correlations) can vary significantly, and there is a general lack of consensus in the choice and suitability of methods for a given site. The resources required to develop rigorous characterisation can be reduced through intelligent SI campaigns and interpretation methods which: leverage existing datasets and prior knowledge, maximise value from low-cost forms of acquisition (e.g., in situ testing), and handle uncertainty rigorously and correctly. This paper presents a new framework for the development and propagation of characterisation methods for both research and industry purposes. This follows the curation of a large database of site characterisation data at over 20 offshore sites, facilitating an initial base of prior knowledge. In particular, automated methods for cleaning, pre-processing, interpretation, and spatial pairing of data (where typically, manual processes are required) are detailed, enabling seamless preparation of reliable datasets for model training. Such procedures allow site characterisation models to be continually refined in a straightforward, standardised manner, while maintaining robustness, transparency, and objectivity. From these prepared datasets, following Bayesian principles, initial probabilistic (prior) models can be built from existing data or updated using site-specific data resulting in tailored (posterior) models.

Keywords: Site characterisation; pre-processing; data-driven approaches; Bayesian statistics; uncertainty quantification

1 INTRODUCTION

Site investigation (SI) campaigns to support the design of offshore structures require substantial vessel time and resource, often invoking £10s of millions and large carbon footprints. Investigations typically cover large areas and combine low-fidelity but spatially continuous geophysical surveying with intrusive geotechnical testing at targeted locations, both in situ and in the laboratory via sampling. Despite the scale of such campaigns, the planning, processing, interpretation, and integration of data are most often performed manually, and there is a considerable lack of standardisation over the most appropriate characterisation models to use for a given site or set of soil types. Furthermore, many available models stem from studies on small or specific datasets, and when extrapolating to new sites and soil types, their suitability can become uncertain. As a result of these factors, site characterisation approaches tend to vary significantly between projects, and often involve collections of models underpinned by different philosophies, held together and made site-specific through various forms of empiricism and individuals' engineering judgement. While empiricism and judgement are undoubtedly important tools within geotechnical engineering, a more structured framework for site characterisation would allow characterisation models to be developed, refined, and optimised more effectively, and enable their applicability and uncertainty across different sites and soil types to be better understood.

This paper is part of a wider research framework aimed at establishing intelligent and more optimised site characterisation methods for the offshore wind industry. In doing so, it aims to improve the efficiency of data collection and interpretation (i.e., reduce data wastage), facilitate a more considered and robust approach to uncertainty, and streamline the development and use of site characterisation

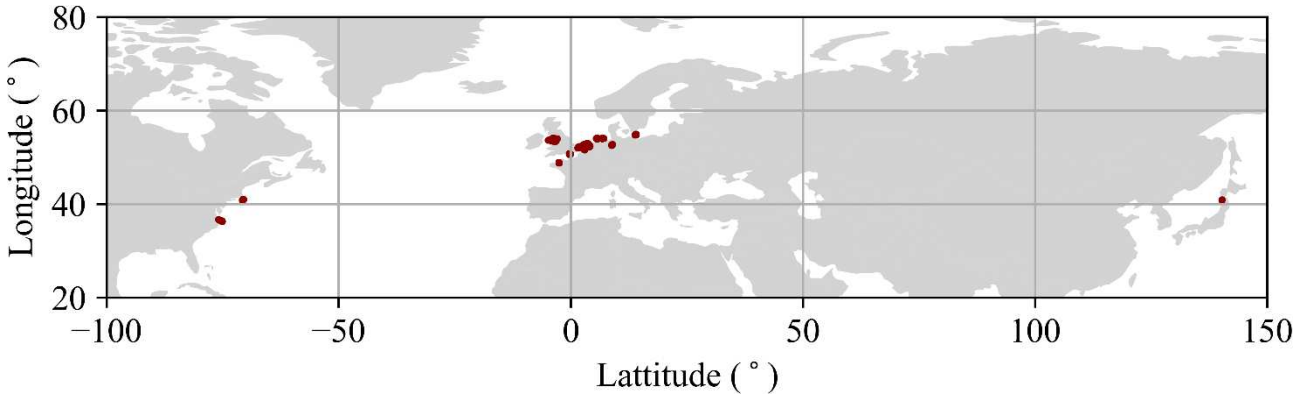

Figure 1 – Map of over 5000 borehole and in situ investigation locations from over 20 OWF sites included in database.

methods in future, considering how collective knowledge across the industry can be progressed. The project, named *Smart characterisation of geomaterials using database methods*, is a joint industry-academic project funded by the University of Glasgow, Scottish Power Renewables (SPR), and Geowynd. It involves the curation of a large, worldwide database of geotechnical data from existing offshore wind projects, drawing from SPR's data archives as well as a variety of public domain sources (e.g., RVO, Marine Data Exchange). Using this, characterisation models suitable for a wide range of soil types can be developed, and more general frameworks can be investigated and proposed. This paper specifically presents a framework for streamlined development and site-specific calibration of characterisation models for CPT or SCPT – i.e., combined CPT/Seismic Velocity Testing (SVT) – data, suitable for both research and industry purposes. The framework leverages existing data and knowledge when dealing with a new site; maximises value from low-cost, in situ testing methods; and handles uncertainty rigorously and appropriately.

2 INITIAL DATABASE

As foundation for the project, a large database containing a range of geotechnical data from over 20 offshore sites has been constructed, comprising over 5000 investigation locations worldwide – a map summarising the locations is shown in Figure 1. The database contains of raw data tables for each test type and/or aspect of metadata, with data from all projects extracted and converted into a consistent format.

The set of data tables includes those for: investigative locations (coordinates and EPSG system); factual descriptions of borehole layers (layer depths, visual descriptions, and primary soil types, e.g., SAND); CPT data (measurements, cone diameter, α factor); and any other relevant tests (e.g., SVT, classification, triaxial). By indexing every entry with the OWF name, location name, and depth, any given test measurement can be cross-referenced to its precise spatial location, and where available, the reported layer it resides in. For any given test type, by ensuring all data is in the same format and accessed via the same tables, data from various sites can be combined and queried together (or in any combination) seamlessly for model development or assessment. Note that while all analysis described herein was undertaken using this database, all procedures have been designed to be applicable to any dataset, regardless of the number of sites or locations, provided at least the location, layer description, and CPT data tables are available.

3 FRAMEWORK

(S)CPT-based characterisation models, often in the form of direct empirical correlations (e.g., Robertson and Cabal, 2022), are widely used within the offshore geotechnical industry. (S)CPTs are low-cost and offer continuous profiling, making them ideal for parameterisation across a site, provided the models used are appropriate. This latter point is generally difficult to verify, as typical models tend to be deterministic, and their uncertainty and generalisability is not well understood. This new framework describes a standardised, mathematically rigorous approach through which any (S)CPT-based models can be developed, refined, or calibrated to site-specific data, while handling uncertainty appropriately.

Adjacent in situ and borehole pairs offer a means of training or calibrating models, with the sample/ laboratory data acting as 'ground truth' for a parameter that is a target of a given model. However, to ensure the trained model is reliable, the analyst must be certain that all data is valid and representative. For CPT data, the correct layers should be paired, no thin laminations of alternative soils should be present, and no operational quirks should be seen in the data. This non-trivial process is typically performed manually in industry, and within academia, the majority of studies focus on very simple soil profiles in the field

or artificial profiles in the calibration chamber to avoid such complexities. For SVT data, there is no consensus on uncertainty quantification or the most appropriate interpretation method (Stolte and Cox, 2019; Parasie et al., 2022), and data can sometimes be compromised by measurement errors (Zheng et al., 2024a).

The new framework first describes automated methods for pre-processing of CPT and SVT data and for pairing of in situ and borehole data, which is shown to be robust in producing reliable training data across a wide range of cases. Through these automated methods, high-quality training data can be produced rapidly from any given set of (S)CPT-borehole pairs, allowing potential characterisation models to draw from a huge array of existing and future data. Research studies can curate large training sets from various data sources, and individual OWF surveys can be used to generate smaller, site-specific training sets for calibration (a small number of co-located borehole-in situ pairs can be used to calibrate a model which is subsequently used at lone in situ positions across the site).

The second phase of the framework involves applying a Bayesian treatment for all characterisation models. Models should be probabilistic, with uncertain model parameters as random variables described by appropriate probability densities. In the first instance, models are initialised through training, ideally using a wide array of data as per the *Smart* database. These initial models represent *prior* models, which can be used directly, further refined with additional data, or calibrated specifically with site-specific data. To perform the refinements/ calibrations, a process known as Bayesian inference is utilised, which involves updating the parameter densities in accordance with new data – resulting in *posterior* models. Thus, the framework offers a robust, standardised means of industry-wide continual improvement of models as more data is collected. Any research or industry environment can perform the above steps using only raw, factual data.

4 IN SITU DATA PROCESSING

4.1 CPT processing

Raw CPT data can contain features which do not accurately reflect the true behaviour of the tested soil. In particular (see Figure 2):

- A CPT can encounter *refusal* on meeting a hard layer. Since the hard layer influences the measurements but is not fully mobilised, the measurements are often unrepresentative.
- Following a "drillout" during downhole acquisition, the measurements increase rapidly from zero at the new free surface. These values are not representative of the real soil behaviour.
- At the boundary between soil layers, nearby measurements exhibit a transition zone which is not fully representative of either material.

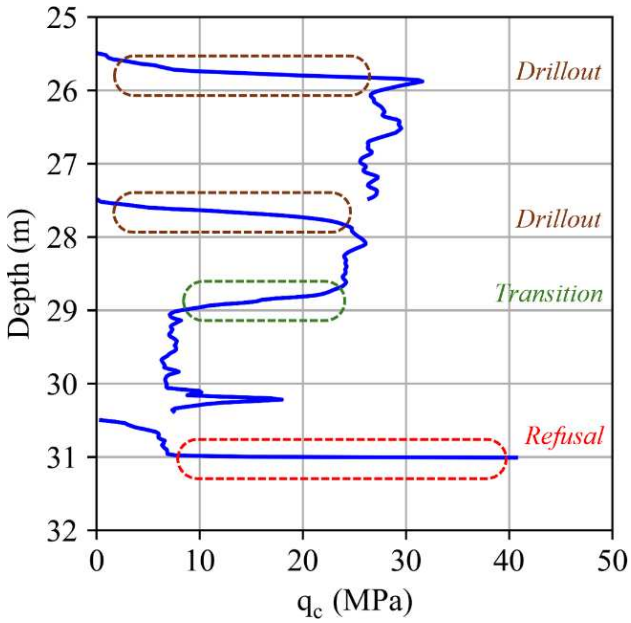

Figure 2 – Example of unrepresentative features in CPT.

Pre-processing algorithms to remove refusals and correct post-drillout measurement reductions have been developed and are described below. Transition zones are accounted for separately during the pairing process, described in Section 5.

4.1.1 Refusal Identification

Through manual observation of many refusal cases and trial and error of various approaches, the following procedure, involving two refusal criteria, has been developed (see Figure 3):

1. **Identify Type 1 Transition Zones:**
   Calculate the gradient of q_c with respect to depth z, and smooth via a moving average function (herein, averaging over 5 data points). Define type 1 transition zones as regions where the smoothed gradient profile continuously remains above a critical threshold (herein, 50 MPa/m).
2. **Identify Type 2 Transition Zones, following Boulanger and De Jong (2018):**
   Calculate gradient m of ln q_c with respect to normalised depth z/d_c (cone diameter), and smooth via a moving average function. As per Boulanger and De Jong (2018), define type 2 transition zones as regions where m remains above a minor threshold and contains a peak above a major threshold.

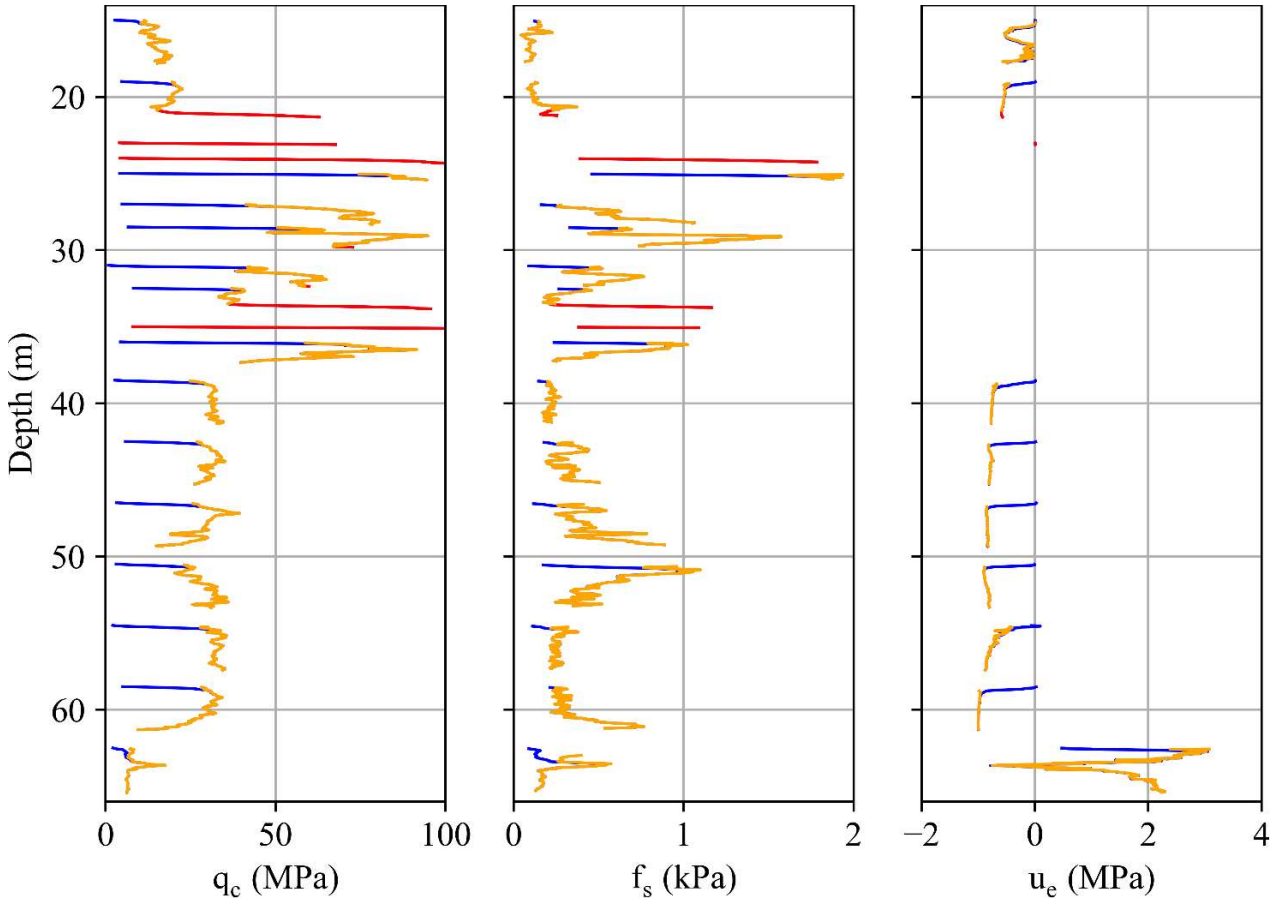

Figure 3 – Example of fully automated pre-processing CPT data showing raw profiles (blue) and new profiles after correcting for downhole effects (yellow) and removing refusals (red).

3. **If a transition zone of either type containing the final data point exists, flag the full zone as a refusal and remove it from the dataset.**

4.1.2 Downhole Correction

The proposed downhole correction approach comprises three algorithms for q_c, f_s, and u_2,respectively. Each relies on a semi-analytical model relating the measured CPT response to what would be the true response. They are informed by physical principles accounting for the reduction in q_c, f_s, u_2, due to the drillout, but contain a series of tuning parameters which can vary to account for differences in the precise nature of the mechanism for different soils. For each CPT push, an optimisation scheme is executed in order to find the set of tuning parameters which leads to the greatest consistency in near-surface data (using a pseudo-variance term as an objective function to be minimised). Finally, the true profiles are delivered using the optimum set of parameters. The models and precise algorithms are complex, so cannot be detailed comprehensively herein, but an overview of each is provided here.

For q_c, it is posited that the reduction is due to (i) a decrease in stress beneath the newly excavated surface, and (ii) a decrease in bearing factor due to shallow bearing replacing the deep penetration mechanism. The former is accounted for by calculating the new, post-drillout in situ stresses according to the solution for a circular stress acting on an elastic half plane (Boussinesq, 1885), and the latter is accounted for with an empirical function f describing the evolution of bearing factor with depth:

$$q_t - \sigma_{vd} = \left(\frac{\sigma'_{vd}}{\sigma'_{v0}}\right)^{\mu} \cdot f\left(\frac{z}{d_c}, \boldsymbol{w}\right) \cdot (q_{t,true} - \sigma_{v0}) \quad (1)$$

where q_t is the measured total cone resistance, $q_{t,true}$ is the true, corrected cone resistance (i.e., if no drillout were present), σ_{v0} the in situ vertical stress, σ_{vd} the new vertical stress post-drillout, and μ and $\boldsymbol{w}$ the tuning parameters.

Literature regarding analytical solutions for f_s are relatively sparse, and as such, the level of reduction is assumed to be similar to that of cone resistance:

$$\frac{f_{s,true}}{f_s} = A\left[\overline{\left(\frac{\sigma'_{v0}}{\sigma'_{vd}}\right)^{\mu}} \frac{1}{f} - 1\right] + 1 \quad (2)$$

where f_s is the measured sleeve friction, $f_{s,true}$ is the corrected sleeve friction, μ and f are equal to those used for Equation 1, A is the sole flexible tuning parameter, and the reduction terms are averaged along the full length of the sleeve.

It is assumed that the excess pore pressure $u_e = u_2 - u_0$ is reduced from its true, steady-state value $u_{e,true}$

to the measured value, which increases from 0 at the surface until $u_{e,true}$ is reached. To define the model, it is assumed that the rate of increase du_e/dz is equal to the rate of generation (due to expansion/shear) minus the rate of dissipation (due to drainage). Thus, u_e and $u_{e,true}$ are related via (i) a decrease in pore pressure generation resulting from reduced stress, and (ii) an increase in dissipation due to the new, local free surface. The former is modelled by assuming the rate of generation is proportional to the stress, and the latter by assuming the reciprocal of dissipation rate increases with depth in a similar form to the function for bearing capacity. These conditions lead to:

$$u_{\mathrm{e,true}} \cdot k f_{\mathrm{u}}\left(\frac{z}{d_{\mathrm{c}}}, \boldsymbol{w}_{\mathrm{u}}\right) = \left(\frac{\mathrm{d}u_e}{\mathrm{d}z} + k u_e\right) \cdot \frac{\sigma'_{\mathrm{v0}}}{\sigma'_{\mathrm{vd}}} \qquad (3)$$

where k is a tuning parameter, and f_u is a similar function to f, but subject to different tuning parameters $\boldsymbol{w}_u$.

Using the optimisation scheme, the most likely tuning parameters for Equations 1–3 can be ascertained and hence the *true* profiles calculated. Example results for a downhole acquisition covering several soil types are shown in Figure 3. This process is entirely automatic. It can be run once initially for any collection of CPTs to develop representative data suitable for general use, or for pairing with sample/ laboratory data and subsequently model training.

4.2 SVT Pre-processing

As part of the *Smart* project, a new method has been developed for the interpretation of SVT data, involving combined Bayesian inference of V_s and key measurement errors during acquisition from raw data. The method is described by Zheng et al. (2024a, b). It delivers probabilistic profiles of V_s conveying uncertainty (significant as uncertainty in acquisition and interpretation is known to be high) and is able to deduce and remove the influence of small timing, depth, or other measurement errors by assessing inherent inconsistencies in a given dataset. Traditional methods are unable to detect such errors, and their derived results can be affected significantly. Application of this method can therefore be used to deliver robust V_s values, which are representative of the soil, for direct input into characterisation models or in order to generate reliable training datasets. For examples of processed SVT data using this method, refer to Zheng et al. (2024b).

5 DATA PAIRING

With all relevant in situ data pre-processed, 'ground truth' labels from sample/laboratory data can be paired with respective in situ data from adjacent test locations. This process follows a layer-by-layer approach, in which the factually reported layers are used to collate sets of laboratory tests, such that for each layer the relevant parameters can be assigned. Then, all representative in situ data associated with that layer is extracted. This approach increases the reliability of the ground truthing by reducing dependence on individual tests and maximises the available CPT data extracted for training by considering the full layer thickness.

For example, suppose a model correlating UU-derived s_u with CPT measurements is to be trained. For each layer, s_u from all available UU tests are collated, and the mean and variance evaluated. For a more complex model accounting for linear changes in strength, a Bayesian linear regression on (s_u, z) data points could be performed to evaluate the mean and covariance of ($s_{u,0}$, k) for each layer, where $s_u = s_{u,0} + kz$, $s_{u,0}$ is the strength at the layer surface, k is the strength gradient, and z is the depth beneath the layer surface. The relevant parameters for each layer are then assigned as the 'targets' or 'labels' for the training set. Alternatively, labels can be inferred indirectly from tests. For a soil type classification model, a sand could be defined as a layer where the average fines content is lower than 35%.

Then, for each layer with a 'target' assigned, the following algorithm is used to extract the associated in situ data from the full pre-processed dataset.

1. Find the subset of CPT and SVT data from all investigated locations within some lateral distance x_{lat} (herein, 15 m) of the borehole.
2. Find the subset of CPT and SVT data between the top and base depths of the layer, subject to a buffer distance x_{vert} (herein, 1 m) at both ends to account for slopes in the layer surface.
3. Assess the above CPT data for both types of transition zone, and remove data points which are classed as such, as these are not representative of the primary material behaviour.
4. Within the above CPT data, remove data points for which q_{net} is more than N_{out} (herein, 2) standard deviations from the mean, removing laminations.

An example of this pairing process, collating key CPT parameters for layers determined to be 'clean sand' and 'clay' (via descriptions, Particle Size Distributions and Atterberg Limits), respectively, is shown in Figure 4. The Figure demonstrates that large, reliable training datasets can be feasibly prepared via this fully automated approach. Further manual refinement of the combined dataset can then be imposed easily.

6 MODEL DEVELOPMENT

Any Bayesian model is suitable for the framework presented, allowing for a wide range of options depending on the particular problem. These include: Bayesian regression/curve-fitting models, Bayesian Network models, and Bayesian formulations of advanced ML models (see Bishop, 2006). The key components of any such model are as follows:

- Definition of the model: The mathematical relationships between all variables should be defined and the model parameters listed.
- Solution procedure: The most appropriate methods for inference and prediction (either exact or approximate) should be defined and followed consistently for each update.
- Scaling parameters: Before model training or use, all values within a dataset should be scaled
- for computational stability. For consistency and continuity, initialisation, general use, and model updates should use an identical scaling regime, with values defined at initialisation.
- Parameter probability densities: Probability distributions (joint, where applicable) for all model parameters.

The above represent the primary elements of any model within the framework. With these, any party is able to update or apply the model directly. For best practice, the initial prior probability densities and each update should be fully documented, forming a record of the evolution of probability densities alongside the datasets used.

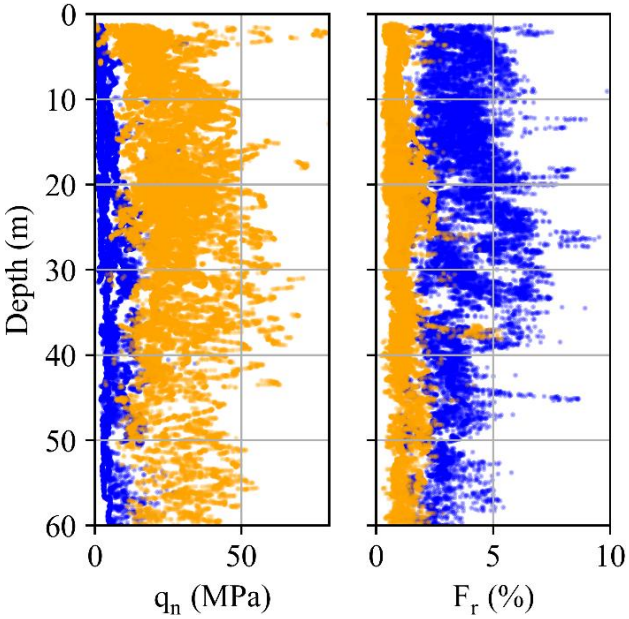

Figure 4 – Example of paired CPT data for adjacent borehole layers designated 'clean sand' (yellow) and 'clay' (blue) via laboratory tests.

7 CONCLUSION

This paper has presented a rigorous, standardised framework for developing and refining (S)CPT-based characterisation models, which can be followed within both academia and industry to progress collective knowledge effectively. The framework adopts a Bayesian probabilistic approach in order to ensure that uncertainty (typically high in geotechnical engineering) is managed appropriately and robustly. The approach enables new models to be developed, or existing models to be improved and/or calibrated to site-specific conditions, all through a unified and consistent mathematical framework. In order to support this, procedures for automatic generation of reliable training data, requiring only raw, factual data (i.e., as received from industry geotechnical surveys), have been presented. These procedures enable the utilisation of a wealth of industry data (both existing and to be acquired in future) not typically considered for academic research to be used to develop and continually progress the industry-wide collective knowledge of site characterisation techniques.

AUTHOR CONTRIBUTION STATEMENT

Author 1: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Software, Visualisation, Writing – original draft. **Authors 2-3**: Conceptualisation, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Writing – review and editing. **Authors 4-5**: Conceptualisation, Data curation, Funding acquisition, Writing – review and editing. **Authors 6-7**: Conceptualisation, Funding acquisition, Writing – review and editing.

ACKNOWLEDGEMENTS

The first Author is supported by Scottish Power Renewables, Geowynd, and the University of Glasgow through the latter's industrial doctorate programme. RVO, Marine Data Exchange, and BSH own, publish, and license the offshore data shown in this paper, the former under a Creative Commons 4.0 CC BY SA license.

REFERENCES

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*, Springer, NY, USA.

Boulanger, R.W. and DeJong, J.T. (2018). Inverse filtering procedure to correct cone penetration data for thin-layer and transition effects, In: *Proc. 4th Intl. Symp. Cone Penetration Testing (CPT'18)*, Delft, Netherlands, pp. 25-44.

Boussinesq, J. (1885). *Application des potentiels à l'étude de l'équilibre et du mouvement des solides élastiques*, Gauthier-Villars, Paris, France.

Parasie, N., Franken, T. and Peuchen, J. (2022). Assessment of Seismic Cone Penetration Testing for Small Strain Shear Modulus, In: *Proc. 5th Intl. Symp. Cone Penetration Testing (CPT'22)*, Bologna, Italy, pp. 203-208. https://doi.org/10.1201/9781003308829

Robertson, P.K and Cabal, K. (2022). *Guide to Cone Penetration Testing*, 7th ed., Gregg Drilling LLC, CA, USA.

Stolte, A.C. and Cox, B.R. (2019). Towards Consideration of Epistemic Uncertainty in Shear Wave Velocity Measurements Obtained via SCPT, *Canadian Geotech. J.*, 57(1), pp. 48-60, https://doi.org/10.1139/cgj-2018-0689.

Zheng, T., Buckley, R. and Febrianto, F. (2024a). Development of a framework for automatic quantification of uncertainty in seismic cone penetration testing, In: *Proc. 7th Intl. Conf. Geotechnical and Geophysical Site Characterization (ISC'7)*, Barcelona, Spain. https://doi.org/10.23967/isc.2024.152

Zheng, T., Buckley, R. and Febrianto, F. (2024b). Automatic interpretation of SCPT and DST data accounting for uncertainty and measurement error, *J. of Geotech. and Geoenv. Eng.*, in review.