Comparison of different prediction methods to derive synthetic CPT profiles – an offshore wind farm case study from the German North Sea

Lennart Siemann[1#], *Pedram* Masoudi[2], *Rajeswar* Reddy Maraka[1], *Raluca* Opris[1], *Yashwardhan* Pande[1], *Nikolas* Römer-Stange[3], *Natasha* Morales[1] and *Tobias* Mörz[3]

[1]*Fraunhofer Institute for Wind Energy Systems (IWES), Department of Subsurface Investigation, Am Fallturm 1, 28359 Bremen, Germany*
[2]*Geovariances, 44 avenue de Valvins, 77210 Avon, France*
[3]*University of Bremen, Faculty of Geosciences, Klagenfurter Straße 4, 28359 Bremen, Germany*
[#]*Corresponding author: lennart.siemann@iwes.fraunhofer.de*

ABSTRACT

The further development of offshore windfarm areas in various countries plays a key role in the transition of energy production towards renewable sources. As offshore windfarm areas tend to expand and the amount of ground truth data is limited, the estimation of geotechnical parameters at unknown locations integrating other site investigation data becomes a necessary tool. This is especially relevant for cost efficient area wide site characterization. Here, the proper integration and correlation of geotechnical and geophysical data is a key factor for reliable ground model building. This study investigates different prediction methods, while presenting a modelling framework which incorporates geological, geotechnical, and geophysical information to derive synthetic Cone Penetration Testing (CPT) profiles using offshore windfarm site investigation data from the German North Sea. We combine geological interpretation, CPT data and 2D ultra high-resolution seismic reflection data. The geophysical and geological information are used to guide geotechnical parameter prediction. Additionally, seismic horizons constrain the prediction as structural information. For evaluation, we test and compare several prediction techniques, with different level of complexity, from geostatistical methods to machine learning. Seismic attributes are used as auxiliary information to improve CPT parameter prediction. To validate the results, CPT parameters are predicted onto a representative 2D seismic line and a leave-one-out cross-validation (blindtest) is performed. Though all methods struggle to replicate local extremes, results indicate a reduction of prediction uncertainty when implementing seismic attributes.

Keywords: synthetic CPT; site characterization; offshore wind; data integration.

1. Introduction

Planning and development of offshore wind farms requires a good understanding of the sub-seafloor geologic conditions to design the foundations of offshore wind turbines (OWT) and decide on installation procedures. Especially the complexity of shallow deposits in Northern European waters, dominated by glacial geological processes, poses considerable challenges in understanding small-scale variability of geotechnical parameters within spatially heterogenous deposits (e.g. Cartelle et al. 2022, Emery et al. 2019). Standard procedures for wind farm site characterization include a dense grid of 2D ultra-high-resolution (UHR) multichannel seismic profiles and a geotechnical campaign including varying numbers of Cone Penetration Tests (CPT) and borehole investigations. Paricularly the amount of available geotechnical data depends on the stage of the project and preliminary ground models are often based on a very limited number of ground truth locations. However, a consistent understanding of the major geological units, especially at an early stage is essential for the planning of subsequent geotechnical campaigns and also for subsequent foundation design.

In recent years, an increasing interest for predicting synthetic CPTs at any point of interest for a windfarm areas in the framework of Integrated or Quantitative Ground Models seen in the industry, to aid the understanding of the subsurface. Such approaches commonly involve the integration of geophysical and geotechnical data based on identified geological units and some form of estimation of sediment geotechnical parameters between CPT/borehole locations (e.g. Sauvin et al. 2018, 2022; Vanneste et al. 2022). A crucial aspect of such models is the deduction of geotechnical parameters laterally within identified units between ground truth locations, which may be implemented using inversion workflows (e.g. Vardy 2015) of the seismic data and correlations to geotechnical parameters using available CPT data or through geostatistical or machine learning approaches (e.g. Sauvin et al. 2019; Siemann et al., 2022). Especially in cases with very limited numbers of geotechnically explored locations, the ability of interpolation workflows to reconstruct the geotechnical

character of individual units exhibits an improvement of the quality of the resulting ground model.

In this study different prediction methods to derive synthetic CPTs within an exemplary wind farm area are tested and evaluated, showcasing the differences of the results and the limitations in comparison to real measured data. This work does not aim to judge on the most suitable methods, since this is very site and data dependent, but to summarize various methods as basis for deciding on the most appropiate tool for future investigations. Nevertheless, results indicate that although all methods perform comparably good due to the high amount of ground thruth data, the prediction uncertainty is reduced by including seismic atttributes as auxillary information

2. Study area and available data

In this study, geotechnical and seismic data from an offshore windfarm area from the German North Sea is used. The site is located north of Heligoland. The windfarm is fully in operation and underwent a complete site investigation ahead including a 3D seismic survey and a full geotechnical campaign. Additionally, the research cruise He569 acquired additional seismic profiles in this area from which seismic data is used in this study. Figure 1A illustrates the location of used CPTs with respect to the analysed seismic line.

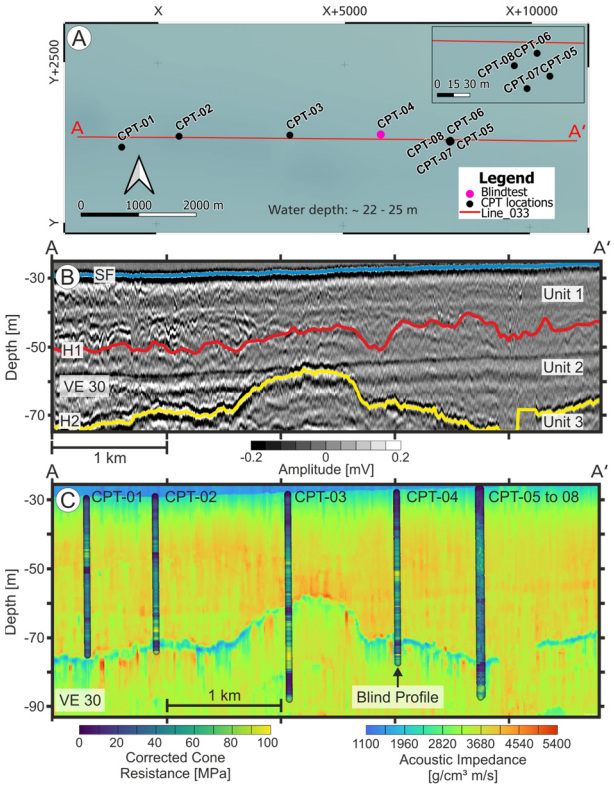

Figure 1*. Overview of the study area and the data being used in this study. A) Location of the seismic line and CPTs. B) UHR seismic data with the CPT logs. C) Absolute acoustic impedance results.*

2.1. Data description

A representative 2D Ultra High Resolution (UHR) line was selected for this study (Fig. 1B). The total length of the profile is approx. 8300 m and the depth of consideration is 50 mbsf for the synthetic CPT prediction. Here, depth-migrated data is available, with an original vertical resolution of 0.25 m. For the purpose of this study three main units have been identified based on the interpreted horizons as shown in Figure 1B. Unit I is located between the seafloor reflector (SF) and H1. Unit II lies between H1 and H2. H2 is a very pronounced reflector indicating a prominent clay unit as seen in the borehole data which is covering the whole line. Deposits below reflector H2 is considered as Unit III. The seafloor multiple reflection has not been removed with additional processing efforts from the data and is therefore visible in the seismic line as well as in the acoustic impedance (Z_p) results and is located at 29 m to the West and 26 mbsf to the East of the profile.

Z_p (Fig. 1C) was calculated for the post-stack seismic data according to the description in section 3.1. Generally, Z_p increases with depth until appr. 20 mbsf where it stays mostly constant until a distinct clay unit which top is represented by H2. In the inversion results, the deposits appear as a thin layer of low Z_p, which can be traced throughout the whole section.

Along the 2D line, 8 CPTs have been conducted with a varying distance between only 2 m (CPT-06) and maximum 172 m (CPT-01) to the seismic line and penetration depths ranging from 45 m to 60 m. A cluster of four CPTs can be found east of the profile with a maximum distance of 24 m to each other. Before using the CPT data for further studies, the raw data was cleaned from outliers and artifacts mostly caused by the form of the downhole CPT data acquisition. The first ten data points of each push as well as data with no related sleeve friction value was removed. By doing so, we ensure to not use erroneous data in further prediction. In this study only the corrected cone resistance (q_t) is considered and predicted for the 2D seismic line.

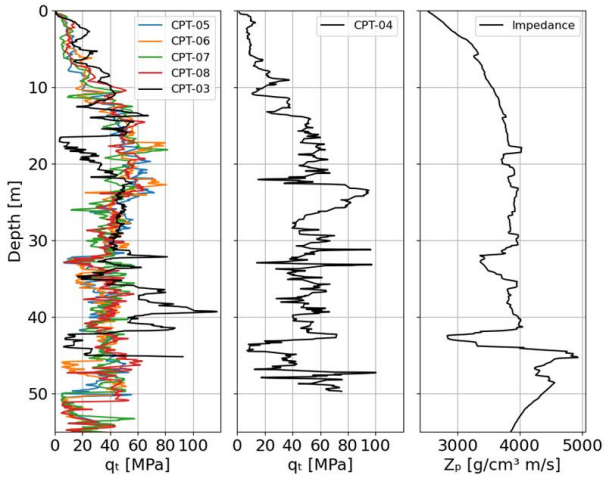

Figure 2*. Plotted q_t for the blindtest profile CPT-04, together with the acoustic impedance at the CPT position and the adjacent CPT profiles for comparison.*

For the evaluation and validation of the different prediction methods, CPT-04 is used as blindtest profile. This profile is removed for modeling to be later compared to the predicted data at the exact same location. Figure 2 shows the blind profile together with the

acoustic impedance at the same position which is used for parts of the method as auxiliary information. In Unit I there is a positive trend of increasing q_t values until H1. In contrast, Unit II shows a slight negative trend of the values with some extremes between approximately 20 and 35 m. Unit II has the highest variation of q_t for CPT-04, containing the lower clay unit, as well as a very dense layer at 48 m. Z_p follows the general trend of increasing values of the CPT profile, with a defined low impedance corresponding to low q_t values at the lower clay unit. The neighboring CPTs of the blind profiles including CPT-03 and the CPT cluster 05-08 are illustrated as well to visualize the closest profiles being weighted stronger for the distance-weighted methods.

2.2. Local geology

The geology in the southeastern North Sea is dominated by Neogene Eridanos topset delta deposits below a prominent glacial unconformity. These deposits are incised by deep tunnel valleys attributed to the Elsterian glaciation (Lutz et al. 2009), which are filled by glacial, lacustrine and marine deposits (Hepp et al. 2012, Coughlan et al. 2018, Fleischer et al. 2022). These Elsterian deposits are overlain by a discontinuous intercalation of glacial and interglacial deposits, poorly preserved due to extensive erosion processes attributed to a late Saalian ice advance. The upper sequence comprises Weichselian deposits, which are characterized by extensive cut-and-fill structures correlating to a non-glaciated lowland throughout the last glaciation. Holocene deposits consist of a sand unit of few meters thickness. Drowned and infilled river valleys are common, originating from the early Holocene transgression in the area (Özmaral et al. 2022).

3. Methodology

3.1. Acoustic impedance

A band-limited acoustic impedance section has been generated based on post-stack seismic data and merged with a low-frequency model derived from interval velocities to generate an absolute impedance estimate. In the first step for the low-frequency model generation, seismic reflection events have been picked. As an adoption of Barros et al. (2015) those picks have been inverted for an interval velocity model with a differential evolution genetic algorithm and a second order move-out equation.

Secondly, band-limited impedance was determined based on the post-stack seismic image with a genetic algorithm as described in Vardy (2015). In this global search and stochastic algorithm (Sen and Stoffa 1992), forward modeling was performed with a convolution of randomly initialized reflectivity models with a wavelet. The band-limited impedance is consecutively derived from the reflectivity. For this approach, the wavelet was extracted from the post-stack seismic image by stacking the tapered seafloor reflection along the profile. Further details of the reproducible implementation are described in Römer-Stange et al. (in prep. 2024).

Finally, to merge the band-limited impedance inversion results with the low-frequency model and thus generate an absolute impedance estimate, the BLIMP algorithm described in Ferguson and Margrave (1996) was extended and applied. In this method, the band-limited impedance was merged with the low-frequency model in the frequency domain with a Linkwitz-Riley crossover filter after scaling the band-limited impedance. Using this procedure, the wavelet and post-stack seismic image did not need calibration.

3.2. Interpolation methods

A variety of prediction methods to estimate geotechnical target parameters, focusing here on q_t, along the seismic line are tested and evaluated. A short description of the different methods is given in the following sub-chapters.

3.2.1. Inverse distance weighting (IDW)

In IDW the values at unexplored positions are calculated by the weighted average of the input data within a search radius. Thereby the weight is estimated by the inverse of a power of the distance between the data points and the one to be calculated. The higher the power, the more weight is given to points which are closer to the target location.

3.2.2. Kriging variants

Kriging is a linear geostatistical method of interpolation. The interpolated value is a linear combination of measurements in the neighborhood, while the weights are optimized to minimize the interpolation error. The Kriging weights are calculated during a matrix inversion, and they depend on distances between the positions of measurements and target, distances between the positions of measurements among themselves, and a model of spatial variability (variogram or covariance). Depending on the stationarity condition and number of variables, different variants of Kriging could be applied. Here, three variants are used:

Ordinary kriging (OK) is a variant of kriging for univariate prediction of synthetic CPTs. It can be used for modelling a stationary variable. The term “ordinary” indicates that it is applicable for modelling a non-stationary variable if local stationarity is established by the neighborhood (Chilès and Delfiner 2012).

Ordinary collocated cokriging (CoK) is a variant of OK for multivariate modelling of the principal variable (e.g. q_t), considering an auxiliary variable from seismic data (e.g. Z_p), which has a better or even area-wide geographical coverage. The term “collocated” refers to the situation when the auxiliary variable is known at all the target points (Wackernagel 2003, Masoudi et al. 2023). Cokriging makes use of the cross-correlation between a primary and secondary variable to minimize the estimation error variance (Minnitt and Deutsch 2014).

Kriging with external drift (KED) is a variant of kriging for modelling a non-stationary variable. In this method, the q_t drift is calculated as a function of Z_p from seismic data, here polynomial equation of second degree is used. Then, the q_t residual is calculated by subtracting the drift. The q_t residual is considered as a stationary

variable, so simple kriging is used for modelling it without applying any neighborhood limit (Pyrcz and Deutsch 2014).

For each interpolated position, the kriging variants enable the quantification of uncertainty by means of the standard deviation of the interpolation error. It tends to be zero close to the measurements and increases as getting far from the measurements. The uncertainty measure is controlled by the variogram model. It means that it is possible to compare the standard deviation of error of interpolated value of two different q_t models if the variogram models are identical, otherwise the comparison must be done with caution.

3.2.3. Turning band collocated co-simulation

Geostatistical simulations are developed to perform stochastic analysis based on the interpolated value by kriging and the associated uncertainty (standard deviation of error of interpolated value; Chilès and Delfiner 2012). In this study, the method of turning band simulations (TBS) is used to generate 100 realizations of q_t conditioned to the CPT measurements and Z_p. The mean of the realizations is considered to be the best estimate for the prediction.

3.2.4. Random forest regression (RF)

In this study, we employ random forest regression (RF) to produce synthetic q_t values at designated locations. RF has previously demonstrated efficacy in analyzing offshore windfarm data, as evidenced by studies such as Vanneste et al. (2022). RF operates by constructing a substantial ensemble of decision trees, with each tree serving as an autonomous regression model (Breiman 2001). The combined prediction of the RF regression is obtained through averaging the outputs of all constituent decision trees. This approach enables to include multiple additional attributes in the prediction process.

3.2.5. Feed forward neural network (FFNN)

The architecture of feed forward neural networks (FFNN) encompasses several essential components: hidden layers, the number of neurons in each hidden layer, the activation function for each neuron, and the training algorithm for determining collective weights and biases. While the input and output layers fulfill distinct roles in the presentation and extraction of data, the allocation of neuron quantities within these layers is governed by the characteristics of the input and output features. Positioned between the input and output layers, hidden layers possess designated weights and biases, steering input data through activation functions to undergo nonlinear transformations (Sharma et al. 2017). The input data is normalized using min-max normalization which helps faster convergence with high learning rates. Research, such as that conducted by Shoukat et al. (2023), used FFNN for forecasting synthetic CPTs within offshore environments.

3.3. Modelling workflow

3.3.1. Geostatistical methods

The CPTs are not exactly located on the seismic line. Therefore, as a first step, the CPT information is projected onto its trajectory. A semi three-dimensional grid is created as basis for the calculations covering the extent of the seismic line used for this study, with a vertical resolution of 0.1 m and a horizontal spacing of 10 m leading to a total amount of 640 800 samples. The grid is used to extract synthetic CPT profiles at any point of interest. The seismic information which acts as guided information for the prediction is migrated onto that grid. At the CPT locations, the geotechnical parameters are as well projected onto the grid and are upscaled in the same step to the resolution given by the seismic data. The model is subdivided into three sub-domains corresponding to the main units shown in Figure 1. Modeling is performed separately in these sub-domains, which ensures more accurate prediction results, by not mixing up deposits not belonging to each other. For creating sub-domains, the spatial information from the interpreted and gridded seismic horizons is used. The model sub-domains are thus defined by the gridded horizons. The corresponding CPT data is assigned to the respective sub-domains as well.

A search neighborhood must be defined which has a strong influence on the estimation result. Only data points being within the search volume are considered for interpolation. For the present data set, a relatively small neighborhood of 2.5 m was chosen for the vertical direction since data is expected to change more rapidly across geological strata. For the horizontal direction a comparably large neighborhood of 3000 m was selected due to the large distance between the ground truth locations and in order to include a sufficient amount of information for each interpolation point. This large lateral neighborhood assumes a relatively homogenous geology of the identified units.

3.3.2. Space deformation by unfolding

To improve lateral spatial continuity, modelling is performed in flattened space. A simplified illustration of the concept is seen in Figure 3. Tectonic forces applied to the sedimentary deposits or dynamic sedimentary systems, impose structural modifications and consequently, make the space more heterogenous by reducing spatial continuity of the variables. Unfolding is a technique of space deformation that reduces space heterogeneities by restoring the sedimentary deposits to the condition prior to the application of tectonic forces or by accounting for non-horizontal and heterogenous layering (Caixeta and Costa 2021, Chautru et al. 2021).

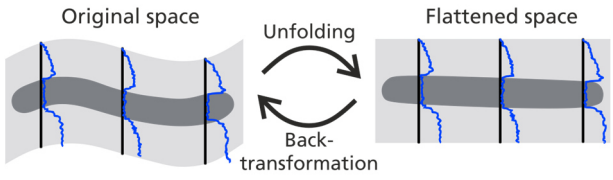

Figure 3. *Simplified representation of the space deformation by unfolding approach for improved modeling.*

In this work, all the geostatistical methods and IDW are applied in the flattened space to improve the spatial continuity, but also to consider the structural geology in the models. The models created in the unfolded space, are then transferred to the real position in the folded space. Among several interpreted horizons, three were chosen to do unfolding separately in three intervals according to the interpretation given in Figure 1. The interpolation is performed within the different units separately to account for the interpretation and geological changes. The unfolding and geostatistical computations were performed by making use of functions exposed by the Isatis.neo software package.

3.3.3. Considerations for machine learning

For the machine learning models Rf and FFNN the procedure to work in flattened space is not necessary because both methods will make use of multi attribute regression which is not restricted by geometrical constraints. For the calculation, the same grid as described in section 3.3.1 is used. Seismic and geotechnical data from the study area are included for training both machine learning models. Aside from q_t, additional information, including seismic attributes, water depth, geographical coordinates and the identification of geological units was utilized in the training process. Besides the Z_p, eight additional post-stack attributes are added to the training process: Instantaneous Amplitude including first and second derivatives, Energy, Instantaneous Bandwidth, Instantaneous Frequency, Instantaneous Q-Factor and Semblance. The incorporation of more attributes and therefore also more information is a clear advantage of this methodology compared to the others.

4. Results and discussion

The target value for predicting along the seismic line is q_t, which is calculated for every grid cell. Fig. 4 shows the results of the predicted q_t, together with the model uncertainty in the form of the standard deviation for all tested methods, except IDW and FFNN. It is not a common output of the latter and is therefore not discussed.

Due to the modeling procedure of space deformation as described in section 3.3.2, q_t follows the seismic reflectors for all methods, even when applying more simple methods as OK and IDW that only use the geotechnical information as input. Generally, the standard deviation increases with depth and with distance to the CPT locations, which is valid for all the tested methods. The vertical lines represented as areas with low standard deviation indicate the positions of the CPT profiles used in the prediction. There is in general not a significant difference in q_t prediction for IDW and the geostatistical methods. This is partly caused by the larger amount of CPT profiles acting as fix points of information along the seismic line thereby limiting variability.

IDW is a rather simple method which can be used to get results fast and does not need a certain experience as for the geostatistical methods and machine learning. In the results shown a power of two was chosen which led to the most reliable results for this particular data set. IDW generally generates rather very smooth predictions, leading to a loss of detail in some regions. For instance, the thin layer located at the left of the profile cannot be robustly delineated.

Generally, the predicted q_t using OK shows a very similar appearance to IDW, while estimating much clearer boundaries between the different layers. The standard deviation for OK increases with increasing distance to the CPT location reaching its highest values at maximum distance between two actual measured locations. The standard deviation differs within the separate units due to different inputs and variograms in the corresponding units. The highest standard deviation can be found in the lower units, which can be explained by a decreasing amount of ground truth information with depth.

Compared to the other methods, KED consistently leads to inaccurate and poorly resolved predictions of q_t, which becomes most apparent in Units I and III. In contrast to IDW and the other geostatistical methods, KED is able to capture the position and structure multiple from the seismic data, by considering the Z_p as drift.

CoK makes use of the acoustic impedance results as co-variable along the grid where no CPT information is available. This is seen in the way that it adds more local variations in the prediction, making it less smoothed as for kriging. Looking at the statistics of the output it strikes that the results obtain negative values, partly even below - 20 MPa within the clay units, which is far from typically expected values. The reason why CoK has negative values in Unit III is that there exist many blocks along the seismic profile with collocated values (Z_p), inferior to the minimum of Z_p at the CPT positions. In other words, the modelling is valid for the range of acoustic impedance observed at the measured locations, leading to an unrealistic negative extrapolation of q_t in areas where Z_p is lower. In direct comparison with OK, the overall standard deviation decreases, especially away from the CPT locations. Nevertheless, close to the measured location the standard deviation is higher compared to OK because of the chosen variogram model for this particular data set, which imposes higher variability at shorter distances compared to the OK model.

In the case of TBS different equally possible realizations have been derived. The displayed q_t is the average of all the results. Simulation does not suffer from negative values as CoK, since the input data has been normalized prior to modelling. TBS honours the input statistics for the prediction, leading to the minimum and maximum q_t for each unit being the same as for the input, although the variation within those boundaries might vary. Close to the lower q_t layer, areas of high q_t are encountered. They correspond to zones where high Z_p values can be observed, which could not be replicated by the other methods. The standard deviation shows a more dynamic appearance, which is governed by the Z_p. Areas with high variability of Z_p obtain higher uncertainties and vice versa. The overall standard deviation is again reduced, especially in Unit III.

The RF q_t prediction does not replicate the layering structure as the other methods described before, which is due to the different modeling approach chosen for RF and FFNN. Since the acoustic impedance does not show a clear layering, it cannot be replicated with the RF approach. The lower clay unit is not as properly captured as in the other methods, and predominantly too high values are obtained in the East of the profile. In some areas the magnitude of the predicted q_t suddenly changes at sharp boundaries. Same artifacts can be seen in the standard deviation which is overall the lowest for all tested methods. The appearance makes it more difficult to track the pre-defined horizons. This is caused by a limitation of the method when using limited amounts of data. For this particular input data set the subsets in the decision trees become too large leading to a blocky appearance in the seismic profile.

The overall structural similarity of the FFNN results with the inverted acoustic impedance suggests that the former is strongly influenced by the latter. FFNN can capture the lower clay unit, although it predicts an area of relatively high q_t directly above, which is mostly not seen for the other methods. Also, the multiple leaves an undesired but strong imprint on the reconstruction. FFNN suffers the same problem as CoK and KED, predicting negative values within the lower clay unit. This is as well explainable by the fact that away from the training data, lower AI values are encountered, which have not been considered for training. This leads again to a negative extrapolation. This issue can be overcome by adjusting the normalizing range of the input data for training.

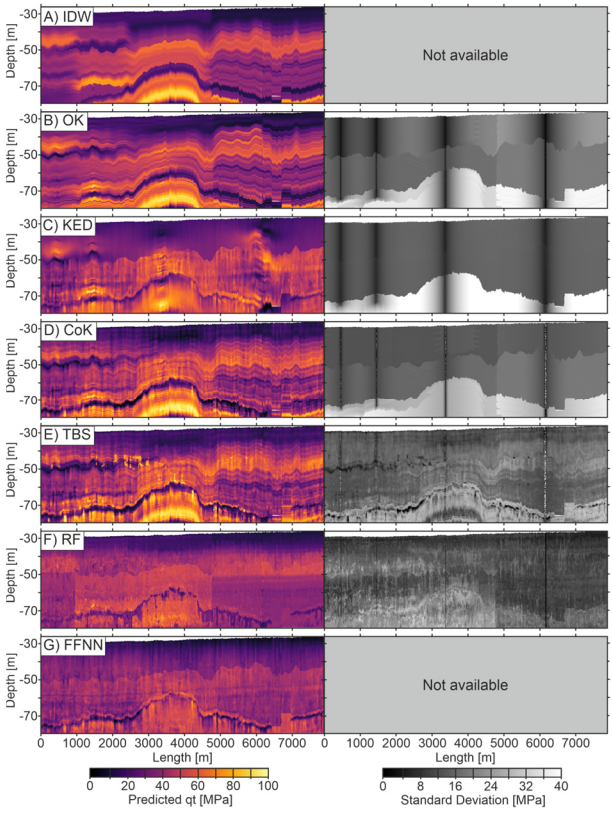

Figure 4. *Left: Best estimate prediction of qt - Right: Standard deviation of the corresponding methods. From top to bottom. A) IDW, B) OK, C) KED, D) CoK, E) TBS, F) RF, G) FFNN.*

In addition to the q_t prediction on the seismic line, Figure 5 illustrates the results of the different tested methods (blue curve) at the position of the blind profile CPT-04 (black curve). We have refrained from smoothing the measured data to properly display and address limitation of synthetic profiles compared to real data. The shaded blue area indicates the standard deviation of the corresponding method, which is as stated before not a common by-product of IDW and the FFNN approach.

All methods capture the general trend of q_t in Unit I, though KED and RF tend to oversimplify while KED obtains the highest variation of the tested methods. In Unit II all methods have a higher deviation from the measured data, compared to Unit I. They are especially not able to properly predict the local extremes. This is due to the fact, that e.g. the very dense layer at around 25 m depth is not present in the adjacent profiles as seen in Figure 2 and is therefore not predictable. Also, this layer is not indicated in the acoustic impedance results making it invisible for TBS, CoK, KED, RF and FFNN, which use Z_p as auxiliary information. Smaller extremes (e.g. at approx. 31 m and 33 m) cannot be seen in the prediction because their vertical extent is smaller than what the seismic data can resolve. This information gets lost in the upscaling process and is therefore not considered for prediction. All methods tend to underpredict the measured profile except for the low resistances at appr. 22 m and 33 m. This is more critical since especially low resistance layers represent a potential a hazard in pile installation and are of particular interest in design processes and therefore relevant to be captured in synthetic CPT profiles.

For Unit III, all methods capture the general course of the blind profile except for KED, TBS and RF. While IDW and OK tend to underestimate the original data, CoK, KED, TBS and FFNN are overestimating q_t, specifically below the clay unit. A slight vertical shift between the lower clay unit and low q_t predictions using the approaches which consider acoustic impedance as secondary information can be observed, which is not seen in the univariate methods. This is explainable on the one hand by vertical uncertainty introduced by the process of migration as well as by upscaling processes. On the other hand, uncertainty might be caused by the projection of the CPT data onto the seismic line. CPT-04 lies 88 m apart from the seismic line making vertical changes in the position of the clay deposit very likely. For engineering purposes, it is specifically important that low resistances areas are captured which is valid for most of the methods except KED and RF. The univariate methods (IDW and OK) led to improved predictions of the clay deposits because they are not influenced by the acoustic impedance and do not suffer from any potential mismatches.

In this study IDW and the geostatistical variants perform comparably well to the tested machine learning methods, because many CPTs are available along the tested seismic line. They are able to capture geological layering imposed by the seismic horizons using the unfolding technique, though it needs to be considered carefully, since it does not account for major geological changes between the reference horizons. Generally, IDW

and the geostatistical variants begin to struggle more in situations where there is less data and when the target points are far away from measured data because of its distance-based weighting. Here, the tested machine learning approaches obtain advantages because they are regression based and therefore independent of the distance between measured data and target location. In addition, machine learning can make use of multiple attributes, although their choice has to be made carefully to ensure that they favour improved reconstructions.

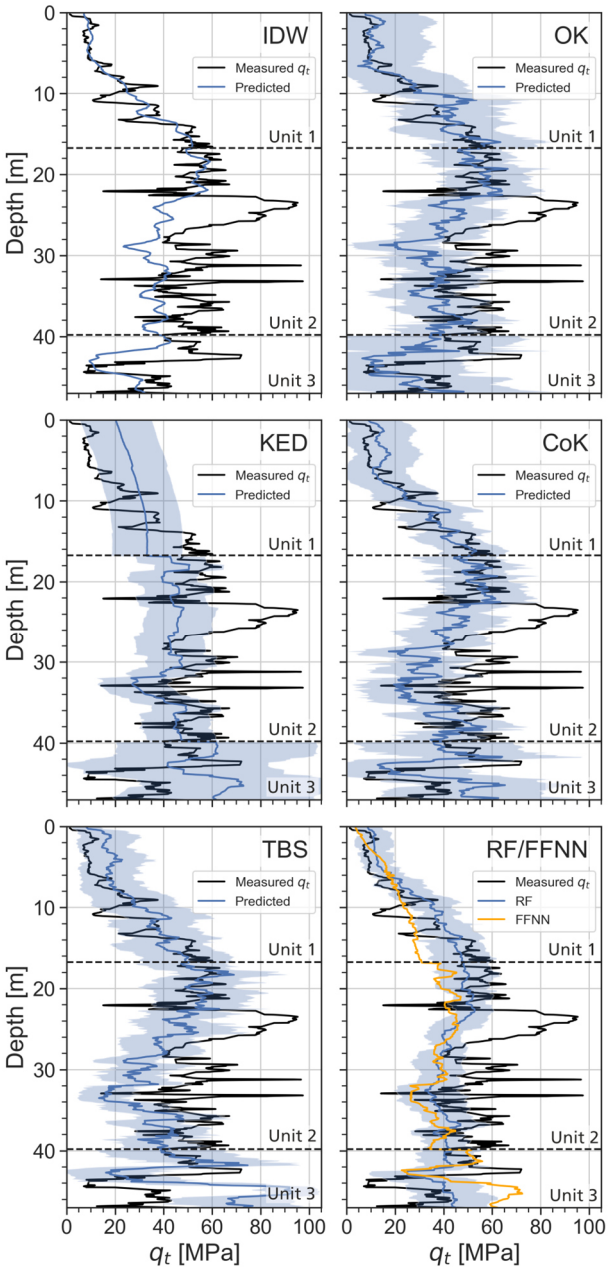

Figure 5. *Comparison of the different tested prediction methods using CPT-04 as blindtest. The Black line indicates the measured data and blue the predicted profile. The standard deviation is indicated by the shaded blue areas.*

5. Conclusions

A series of methods has been tested to predict q_t along a representative reflection seismic line from the German North Sea. The compared methods do not show a significant difference in the overall trend between the prediction results, though especially CoK, TBS and FFNN can account for local changes when making use of seismic attributes as secondary variable. The implementation of seismic attributes as auxiliary information reduced the overall uncertainty of the prediction along the seismic line. All methods can replicate the trend of the measured data, rather than changes on the smaller scale, especially when the information is neither carried by the primary (q_t) nor secondary (Z_p) information. The univariate methods perform well, even without auxiliary information, when having a high amount of CPT information as in this case study. Additionally, IDW and the geostatistical variants better replicate layering structure along the seismic line using the introduced unfolding workflow. In general, depending on the data density and the scope, it needs to be decided individually if the extra effort needed for the more advanced method is worth the outcome. Nevertheless, it has to be pointed out that the machine learning approaches obtain clear advantages by the straightforward implementation of various information.

Finally, it needs to be emphasized, that the outcome of every method depends on various circumstances like the amount of data, input spatial distribution and the quality of the data, therefore this study does not aim to judge on which method performs best but rather to give an overview and to address differences and limitations.

Future studies will focus on the investigation of the contribution of different seismic attributes on the prediction of geotechnical parameters, by more extensive sensitivity analysis, specifically for the machine learning approaches. Also, more data will be integrated by using site investigation data from other windfarm areas.

Acknowledgements

This study was funded though the projects SynCore (Project 03EE3020C) and ProbPerModel (Project 03EE2050A) by the German Federal Ministry for Economic Affairs and Climate Action. The authors are grateful for the provision of the CPT data by RWE (former Innogy) and the encouraging feedback from Benjamin Schwarz and Stefan Wenau.

References

Barros, T., Ferrari, R., Krummenauer, R. and Lopes, R. 2015. "Differential evolution-based optimization procedure for automatic estimation of the common-reflection surface traveltime parameters". *Geophysics* 80 (6): WD189–200. https://doi.org/10.1190/geo2015-0032.1.

Breiman, L. 2001. "Random Forests". *Machine Learning* 45 (1): pp. 5–32. https://doi.org/10.1023/A:1010933404324

Masoudi, P., Binet, H., Simon, C., Pelletier, B., Rambert, F., and Assy, Y. 2023. "Improving Seismic Velocity Mapping Using Standard Penetration Test Data in a Cokriging Interpolation". In: NSG2023 29th European Meeting of Environmental and Engineering Geophysics 2023 (1), pp. 1-5. European Association of Geoscientists & Engineers.

Caixeta, R. M. and Costa, J. F. C. L. 2021. "A robust unfolding approach for 3-D domains". *Computers & Geosciences* 155, 104844.

Cartelle, V., Barlow, N. L. M., Hodgson, D. M., Busschers, F. S., Cohen, K. M., Meijninger, B. M. L. and van Kesteren, W. P. 2021. "Sedimentary architecture and landforms of the late

Saalian (MIS 6) ice sheet margin offshore of the Netherlands". *Earth Surface Dynamics* 9, pp. 1399-1421.

Chautru, J., Binet, H., Masoudi, P., Geffroy, F. and Renard, D. 2021. "Modeling Complex Tectonic Structures in any Kind of Grid without Space Deformation". In: 82nd EAGE Annual Conference & Exhibition 2021 (1), pp. 1-5. European Association of Geoscientists & Engineers.

Chilès, J. P. and Delfiner, P. 2012. "Geostatistics: modeling spatial uncertainty" (Vol. 713). John Wiley & Sons.

Coughlan, M., Fleischer, M., Wheeler, A. J., Hepp, D. A., Hebbeln, D. and Mörz, T. 2018. "A revised stratigraphical framework for the Quaternary deposits of the German North Sea sector: a geological-geotechnical approach". *Boreas* 47, pp. 80-105.

Emery, A. R., Hodgson, D. M., Barlow, N. L. M., Carrivick, J. L., Cotterill, C. J. and Phillips, E. 2019. "Left High and Dry: Deglaciation of Dogger Bank, North Sea, Recorded in Pro-glacial Lake Evolution". *Frontiers in Earth Science* 7, 234.

Ferguson, R. J. and Margrave, G. F. 1996. "A simple algorithm for band-limited impedance inversion". *CREWES Research Report* 8, p. 10.

Fleischer, M., Abegunrin, A., Hepp, D. A., Kreiter, S., Coughlan, M. and Mörz, T. 2022. "Stratigraphic and geotechnical characterization of regionally extensive and highly competent shallow sand units in the southern North Sea". *Boreas* 52, pp. 78-98. https://doi.org/10.1111/bor.12595

Hepp, D. A., Hebbeln, D., Kreiter, S., Keil, H., Bathmann, C., Ehlers, J. and Mörz, T. 2012. "An east–west-trending Quaternary tunnel valley in the south-eastern North Sea and its seismic–sedimentological interpretation". *Journal of Quaternary Science* 27, pp. 844-853.

Lutz, R., Kalka, S., Gaedicke, C., Reinhardt, L. and Winsemann, J. 2009. " Pleistocene tunnel valleys in the German North Sea: spatial distribution and morphology" (English translation). *Zeitschrift Der Deutschen Gesellschaft Für Geowissenschaften* 160, pp. 225-235.

Minnitt, R. and Deutsch, C. V. 2014. "Cokriging for optimal mineral resource estimates in mining operations". *Journal of the Southern African Institute of Mining and Metallurgy* 114 (3), pp. 189-189.

Özmaral, A., Abegunrin, A., Keil, H., Hepp, D. A., Schwenk, T., Lantzsch, H., Mörz, T. and Spiess, V. 2022. "The Elbe Palaeovalley: Evolution from an ice-marginal valley to a sedimentary trap (SE North Sea)". *Quaternary Science Reviews* 282, 107453.

Pyrcz, M. J. and Deutsch, C. V. 2014. "Geostatistical reservoir modeling". Oxford University Press, USA.

Sauvin, G., Vanneste, M. and Madshus, C. 2018. "High-resolution quantitative ground-model for shallow subsurface". In: 3rd Applied Shallow Marine Geophysics Conference, European Association of Geoscientists & Engineers, pp. 1-5.

Sauvin, G., Vanneste, M., Vardy, M. E., Klinkvort, R. T. and Forsberg, C. F. 2019. "Machine Learning and Quantitative Ground Models for Improving Offshore Wind Site Characterization". In: Offshore Technology Conference, Offshore Technology Conference Day 2, 17

Sauvin, G., Vanneste, M., Klinkvort, R. T., Dujardin J. and Forsberg C. F. 2022. "Towards integrated ground models - an example from TNW Offshore windfarm". In: 83rd EAGE Conference & Exibition, Extended Abstracts, pp. 1-5.

Sen, M. and Stoffa, P. 1992. "Rapid sampling of model space using genetic algorithms: Examples from seismic waveform inversion". *Geophysical Journal International* 108, pp. 281–292. https://doi.org/10.1111/j.1365-246X.1992.tb00857.x

Sharma, S., Sharma, S. and Athaiya, A. 2017. "Activation functions in neural networks". *Towards Data Sci* 6 (12), pp. 310-316. https://api.semanticscholar.org/CorpusID:225922639

Shoukat, G., Michel, G., Coughlan, M., Malekjafarian, A., Thusyanthan, I., Desmond, C. and Pakrashi, V. 2023. "Generation of Synthetic CPTs with Access to Limited Geotechnical Data for Offshore Sites". *Energies* 16. https://doi.org/10.3390/en16093817

Siemann, L., Morales, N., Chautru, J.M., Pein, T., 2022. "Prediction and Probability Estimation of Geotechnical Parameters for Offshore Windfarm Site Characterization Using Geostatistical Simulation", pp. 1–5. https://doi.org/10.3997/2214-4609.202221076

Vanneste, M., Sauvin, G., Dujardin, J. R., Forsberg, C. F., Klinkvort, R. T., Forsberg, C. S. and Hansen, R. C. 2022. "Data-Driven Ground Models: The Road to Fully-Integrated Site Characterization and Design". In: Proceedings of the 2nd Vietnam Symposium on Advances in Offshore Engineering. Springer Singapore, Singapore, pp. 3-21.

Vardy, M. E. 2015. "Deriving shallow-water sediment properties using post-stack acoustic impedance inversion". *Near Surface Geophysics*, 13 (2), pp. 143–154. https://doi.org/10.3997/1873-0604.2014045

Wackernagel, H. 2003. "Multivariate geostatistics: an introduction with applications". *Springer Science & Business Media*.