Geotech Geol Eng (2012) 30:1085–1095
DOI 10.1007/s10706-012-9506-5

ORIGINAL PAPER

Validating and Improving Models for Vibratory Installation of Steel Sheet Piles with Field Observations

A. M. J. Mens · M. Korff · A. F. van Tol

Received: 21 July 2011 / Accepted: 11 March 2012 / Published online: 5 July 2012

Abstract Vibratory driving is the most common installation technique for steel sheet pile walls. In practice, the assessment of the feasibility of this installation process is mainly based on rules of thumb, on numerical and empirical models or on experts opinions. In order to improve these prediction methods and formulas, 252 observations from the Dutch engineering practice have been compared with six different types of models. This comparison has been carried out applying the receiver operating characteristic (ROC) curve technique, which is new in geotechnical engineering. This paper introduces the ROC-curve technique to estimate mainly the quality of a model and to be able to optimize parameters and variables in the model. 252 field observations were used to re-examine prediction methods for the minimum required vibration force and to prove the ROC method works. The paper shows this technique is suitable for three purposes: (1) determining the quality of a model, (2) objectively comparing several models to each other, given certain assumptions and (3) for optimizing thresholds within a model. The model with added professionals' experience proves to perform equally well as the numerical model Hypervib-I.

Keywords Design model · Sheet pile · Field observation · Vibratory driving · ROC-curve

A. M. J. Mens · M. Korff (✉) · A. F. van Tol
Deltares, Unit Geo-Engineering, PO Box 177, 2600 MH Delft, The Netherlands
e-mail: mandy.korff@deltares.nl
URL: www.deltares.nl

A. M. J. Mens · A. F. van Tol
Faculty of Civil Engineering and Earth Sciences, Delft University of Technology, Delft, The Netherlands

M. Korff
Cambridge University, Cambridge, UK

1 Introduction

Steel sheet pile walls often support deep excavations in urban areas. The most common installation technique for steel sheet pile walls is vibratory driving. Most projects are carried out in areas where the subsoil consist of several soft clay and peat layers on top of a medium or loosely compacted sand layer. Vibratory driving is attractive in such sub-soils, because of the relative straightforward technique and the high production rates. In practice, the feasibility of the installation process of these sheet piles is mainly based on rules of thumb, on empirical models or experts opinions. It is to be expected that the more experience is added to the rules and models, the more reliable prediction models become.

In 2004 Van Baars used the results of 18 observations to show the inferior quality of several models that predict the minimum required vibration force in order to determine the best vibrator for pile installations (Van Baars 2004). This current paper will show a

single error is not sufficient to determine the quality of a model. The method of 'Receiver Operating Characteristic' (ROC) Curves (Metz 1978) is introduced to overcome this problem. Instead of 18 observations, 252 observations were used to re-examine the methods described by Van Baars. A positive side effect of the ROC method comprises the possibility of threshold and parameter optimization, which will be shown as well.

This paper first describes the origin of the (Dutch) field observations that were used for the comparison of the design codes (Sect. 2). Subsequently, Sect. 3 shortly describes the prediction models and the parameter choices. Section 4 introduces the essentials of the ROC-curve technique and finally the results-section will compare the models, based on the field observations, using the ROC-curve technique.

2 Data–Observations

In order to validate the prediction models for vibratory driving, 252 field cases from the Dutch experience database 'GeoBrain' have been used (www.geobrain.nl). The GeoBrain experience database (Barends 2005; Hemmen 2005) contains case histories of foundation techniques. As most cases are unique, the experience gathered focuses on the generic techniques applied in these cases, such as piling. Since 2005, different contractors have been filling this database with their up to date experiences in the Netherlands. The total number of entries counted 2900 projects by the end of 2011. At the time of this evaluation in February 2009, 364 entries concerned the vibratory installation of steel sheet piles. An 'experience' or 'observation' is uniquely defined by the type of element (for example sheet-pile or prefabricated concrete pile), the type of equipment used and the soil conditions present. Additionally to this digitalized data, also details concerning the building pit, the crew and the surroundings have been included.

Although the database comprised 364 observations for vibratory driving at the time, only 252 of them have been used for this evaluation. An observation was discarded when

- essential data was lacking (like a Cone Penetration Test);
- a combination of installation techniques was used (both hammering and driving);
- unexpected obstacles were expected or detected in the subsoil;
- erroneous data was recorded, e.g. large differences (>1.5 m) between the entered length of the sheet pile and the difference between the head and the toe of the pile
- the head of the pile was deeper than 1.5 m below the surface.

A detailed example of one observation has been described in Mens and van Tol (2010).

In general, the projects from the GeoBrain database comprise, amongst others, the following features: the type of the vibration equipment, the type of pile that has been used, the results of one cone penetration test that reflects the mean circumstances and the number of piles that was used. "Appendix 1" shows eight boxplots with the frequency, displacement amplitude, eccentric moment, mass of the sheet pile, dynamical mass of sheet pile and vibrator, the pile length, the pile cross sectional area and the number of piles that were used for this investigation. More information about the projects and the geological area is available at geobrain.nl.

An observation is defined to be 'positive' whenever within the project 100 % of the piles reach the predetermined depth. This 100 % avoids major subjectivity, but is of course quite conservative. A short description of each design tool follows, together with a transformation to one single criterion that makes comparison between the methods possible.

3 Prediction Models

Current (European) practice uses four categories of models to predict the vibratory driving equipment for successful installation of a steel sheet pile, depending on the scope and the complexity of the project. Category one comprises (numerical) computer models, such as Hypervib-I (Holeyman and Legrand 1994; Holeyman et al. 1996; Holeyman et al. 2002; Gonin et al. 2006), the Karlsruhe model (Dierssens 1994) and Vipere (Vanden Berghe 2001). Viking (2002) presents an extensive explanation and comparison of these methods. This study regards Hypervib-I, since this computer model is in use in this region and it has been the basis for a simplified prediction equation used in the Netherlands. The second category comprises design

equations and empirical rules. Some of these are in fact simplifications of more complex computer models, adapted to specific circumstances. This study discusses a rule of thumb from the CUR [the Dutch centre for research, rules and regulations in the civil engineering practice (CUR166 2005)], the EAU (the German equivalent) (EAU 1990) and two design rules based on Hypervib-I (Azzouzi 2003; Van Baars 2004). (The term Vibrive is an equivalent to Hypervib-I, Viking (2002) mistakenly added a 'd' to Vibrive, making it Vib**d**rive, which has been quoted as such by Van Baars (2004)). The third category comprises design charts, developed by the NVAF (the Dutch federation of foundation contractors (Van Baars 2004; CUR166 2005). The fourth and last category contains a Bayesian Belief Model, based on expert knowledge (Bles et al. 2003; Hemmen 2005) Since this is neither a numerical model, nor a simple design equation or chart, it does not belong in one of the previous categories.

3.1 Objective Criteria

Engineers need an objective criterion to compare predictions to real results for each observation. In this research the percentage of refusals, piles in a project that did not reach the predetermined depth, is taken as criterion. A project has been defined as a (part of a) construction work that uses the same equipment and the same pile-type, in an area that can be described by one 'representative' Cone Penetration Test (CPT). All prediction tools can be 're-arranged' to predict whether the combination of soil, equipment and piles is sufficient to reach the pre-determined depth. In this paper, a 'positive' prediction is equivalent to 'technically being able to reach the pre-determined depth with the chosen equipment'.

3.2 Methods

3.2.1 Method 1 (Hypervib-I)

The computer model Hypervib-I (Holeyman et al. 1999) regards the sheet pile as a rigid body and models the vibratory installation of the pile, making use of four forces. (1) the vertical vibration force from the vibrator on top of the pile, (2) the resisting force on the shaft from the soil friction (3) the resisting force on the tip of the pile caused by the soil in the downward motion and (4) the gravitation forces on the total mass of the system. The model includes a strength reduction by degradation or liquefaction, both at the shaft and the tip of the pile. Using Newton's second law of motion, the model provides a velocity profile, based on a cone penetration test (CPT) and the specifications of both the sheet pile and the driving equipment. The occurrence of zero velocity equals refusal and is the criterion for not reaching the pre-determined depth. The original computer program bases its prediction on the time to penetrate 1 m of soil (1/velocity). Exceeding 999 s (>16 min.) means refusal and leads to a 'negative' prediction. Therefore, this study uses a threshold value of Vt = 1e−3 m/s (6 cm/min.) for the velocity profile. The parameters in the Hypervib-I code have been determined, based on Belgium engineering practice.

3.2.2 Method 2 (CUR)

The 'CUR-rule' calculates the free displacement amplitude (d), that is used to determine the appropriate vibration equipment (CUR166 2005):

$$d = \frac{M_e}{m_v + m_p} \tag{1}$$

where d = the displacement amplitude (m); M_e = the eccentric moment (kg m); m_v = the vibrating mass of the vibrator (kg) and m_p = the mass of the sheet pile (kg).

The vibrator to be chosen must have M_e large enough to fulfill the required displacement amplitude (larger than 0.005 m). If so the sheet piles will reach the pre-determined depth. To obtain the least required cyclic force, the eccentric moment (M_e) is multiplied with $(2\pi f)^2$, where f denotes the frequency of the equipment in Herz.

Remarkably, there is no soil involved in this equation. The idea behind this simple rule of thumb is that with an amplitude of 5 mm the sheet pile is able to degenerate the strength of the soil to relatively low values and as consequence the original strength of the soil is not involved any more.

3.2.3 Method 3 (Azzouzi)

Based on the Hypervib-I model (Holeyman et al. 2002), Azzouzi (2003) developed a formula that calculates the required vertical cyclic force (Fc) to

be able to determine the most suitable vibrator. Azzouzi used 180 calculations with the Hypervib1 model for the development of the formula that uses the mean cone resistance over the considered sand layers, taken from a cone penetration test (CPT):

$$F_{c,\text{Azzouzi}} = \alpha_A \cdot L \cdot \chi \cdot f_A(q_c) + \beta_A \cdot A_t \cdot g_A(q_c) \qquad (2)$$

where $F_{c,\text{Azzouzi}}$ = the required vertical cyclic force from the vibrator that should be used (N); L = pile penetration length (m); χ = the perimeter of the sheet pile (m); $f_A(q_c) = q_{c,\text{mean}}$ = the mean cone resistance over the sand layers (N/m^2); $g_A(q_c) = q_{c,\text{tip}}$ = the cone resistance at the tip of the pile (N/m^2); A_t = the cross-sectional area of the toe of the sheet pile in m^2; $\alpha_A = 1.92 \times 10^3$(–) and $\beta_A = 1.2 \times 10^{-2}$(–).

(Formula and unities adapted to the standard SI system).

When the used force is larger than the required force, the prediction counts for 'positive'; and otherwise it is a 'negative' prediction.

3.2.4 Method 4 (Van Baars)

Van Baars (2004) developed a slightly different formula, based on the same 180 calculations:

$$F_{c,\text{van Baars}} = \alpha_B \cdot L \cdot \chi \cdot f_B(q_c) + \beta_B \cdot L \cdot A_t \cdot g_B(q_c) \qquad (3)$$

where $F_{c,\text{van Baars}}$ = the required vertical cyclic force from the vibrator that should be used in N; $f_B(q_c) = g_A(q_c) = q_{c,\text{tip}}$; $g_B(q_c) = \exp(\gamma \cdot q_{c,\text{tip}})$; $\alpha_B = 1.2$(–); $\beta_B = 26.4 \times 10^{-3}$(–), and $\gamma = q_{c,\text{ref}} = 8.7$ N/m^2.

Again, formula and unities adapted to the SI-system and when the used force is larger than the required force, the prediction counts for 'positive'; and otherwise it is a 'negative' prediction.

3.2.5 Method 5 (EAU)

The German design rule (EAU 1990) calculates the minimum required vertical force using only the length of the pile and the dynamical mass m_d

$$F_{c,\text{EAU}} = \alpha_E \cdot L + \beta_E \cdot m_d \qquad (4)$$

where $F_{c,\text{EAU}}$ = the required vertical cyclic force from the vibrator that should be used (N); m_d = the dynamical part of the vibrator and the pile (kg); $\alpha_E = 15 \times 10^{-3}$(N/m) and $\beta_E = 3 \times 10^{-4}$(N/kg).

This rule quantifies the German criteria for their choice of vibratory equipment. For each meter of the pile one requires at least 15 kN vertical cyclic force and additionally for each 100 kg dynamical mass one needs 30 kN vertical cyclic force.

When the used force is larger than the required force, the prediction counts for 'positive'; and otherwise it is a 'negative' prediction.

3.2.6 Method 6 (Bayesian Belief Network)

Bles et al. (2003) developed a Bayesian Belief Network (BBN), based on professionals' experience (also abbreviated as 'experts'), to model the risks during installation of foundations. In general, BBNs use probabilistic theory for reasoning under uncertainty and risk in expert systems. Bayes' theorem is the cornerstone in this way of reasoning, because it provides a way to calculate the posterior probability. Bayes calculates the probability on some hypothesis h, given condition D: $P(h|D)$. This conditional probability of h, given D, is calculated using the prior probability $P(h)$, together with the probability on the (data-based) evidence $P(D)$ and the probability on the data, given the stated hypothesis $P(D|h)$ (Mitchell 1997):

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \qquad (5)$$

The method transforms joint probability functions to a set of stochastic variables, ordered in a network. The network itself consists of two parts. The qualitative part shows the relations between the variables in a graphical representation (the network). The quantitative part assigns conditional probabilities to all variables, using likelihood-tables, which describe the effect of preceding variables on the underlying ones.

The input variables include information about the subsurface (cone penetration test, presence of stiff clay or gravel, ground water level, etc.), the sheet pile (length, type, profile, mass, shape, etc.) and the method of installation (equipment, force, etc.). Experts from the Dutch Association for Contractors in Foundation Engineering (NVAF) supplied the necessary information for the likelihood tables, describing the qualitative part of the BBN. Finally, the BBN provides the user with a number between 0 and 100, describing the expected amount of risk. The lower the number, the smaller the expected problems,

involving not reaching the pre-determined depth. Another study (Mens et al. 2008) suggests a threshold value of 38 %, above which to start getting worried about the risks. Above this number the prediction is considered to be 'negative' and below or equal to, it is considered to be 'positive'.

4 ROC-Theory

How to measure the quality of a prediction? The diagnostic 'accuracy'—the fraction of cases for which the prediction appeared to be correct- can be very misleading. In case of binary predictions there are four possibilities: a positive prediction, that in reality fails (1) is a false positive (FP). A positive prediction that in reality is a success (2) is a true positive (TP). On the other hand, a negative prediction, carried out anyway and leading to a positive observation is false negative (FN), where as it fails as predicted it is called a true negative (TN). Now suppose only 5 % of the cases will not reach the predetermined depth and a model predicts the drivability to be always possible, its accuracy will be 95 %. But, this is based on 0 true negatives ánd 0 false negatives.

This paper introduces the 'receiver operating characteristic' (ROC)-curve technique within foundation engineering to provide a better criterion. The technique itself is not new and has been used extensively in other scientific areas, such as medical science (Metz 1978) and ecological engineering (Fawcett 2006).

Basically, this technique labels a method, based on both the sensitivity (the number of true positive predictions/the total number of positive predictions, see Eq. 6) and the specificity (the true negative predictions/the total number of negative predictions). Both of these performance indicators make up a 'sensitivity-pair', which can be plotted in a so called 'ROC-space', with the sensitivity on the vertical axis and (1 − specificity) on the horizontal axis (Figs. 1, 2, 3, 4, 5). By visualizing the sensitivity-pairs for all the mentioned design methods in one graph, an objective comparison between the tools is possible. Metz (1978) and Fawcett (2006) explain this theory in more detail. The sensitivity-pair (0,1) (Figs. 1, 2, 3, 4, 5) describes the 'perfect' model. The closer a random sensitivity pair is to this perfect point, the better the model or design rule is.

4.1 ROC Graph and Contingency Table

The input for the ROC graph is a given prediction model and a set with N observations (or field experiences). For these N observations, the binary result (positive or negative) is known. Using the information from these observations, it is possible to calculate the binary prediction results. These can be summarized in a two-by-two contingency table (or 'confusion'-matrix), which serves as the base for a point in the ROC-space. Table 1 provides an example of this matrix. O(−) represents the total number of

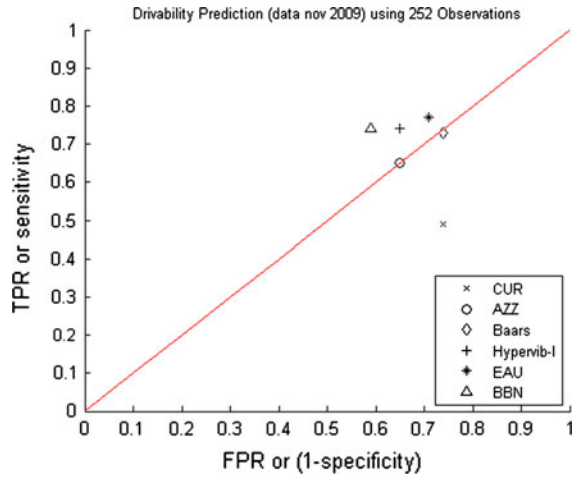

Fig. 1 Drivability prediction as a sensitivity-pair for 6 design codes, using 252 field observations. *TPR* true positive ratio, *FPR* false positive ratio

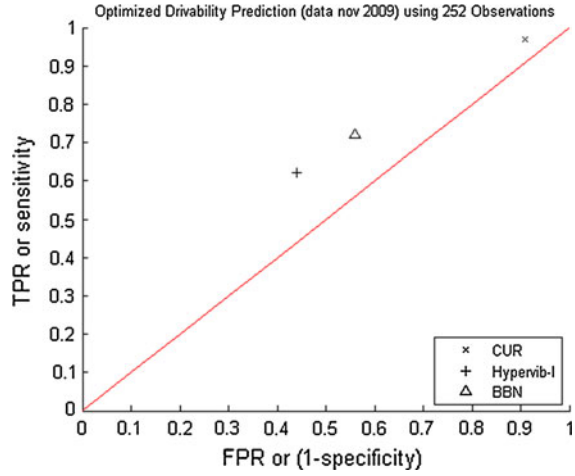

Fig. 2 Optimized driveability prediction for BBN, CUR and Hypervib-I code, changing the threshold values as indicated in Sect. 5.2

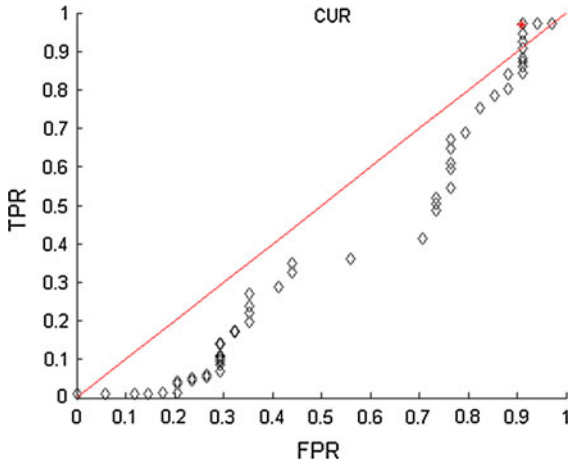

Fig. 3 ROC-curve for CUR rule, where displacement amplitude d varies between 0 and 0.01 m. The *red star* shows the 'best' sensitivity pair, for $d = 0.0029$, assuming both the TPR and the FPR are equally important

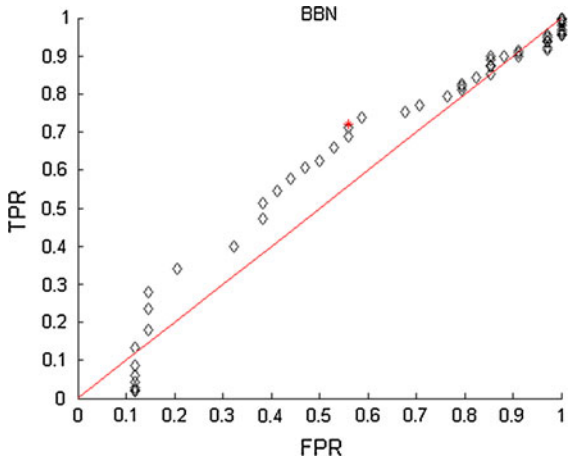

Fig. 4 ROC-curve for the BBN model, where displacement amplitude d varies between 0 and 100 %. The *red star* shows the 'best' sensitivity pair, for threshold value = 36 %, assuming both the TPR and the FPR are equally important

negative observations and O(+) the positive ones. P(−) and P(+) represent the total number of negative and positive predictions respectively.

The numbers from Table 1 enable calculating the following characteristics (amongst others):

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{O}(+)} \tag{6}$$

$$1 - \text{Specificity} = \text{FPR} = \frac{\text{FP}}{\text{O}(-)} \tag{7}$$

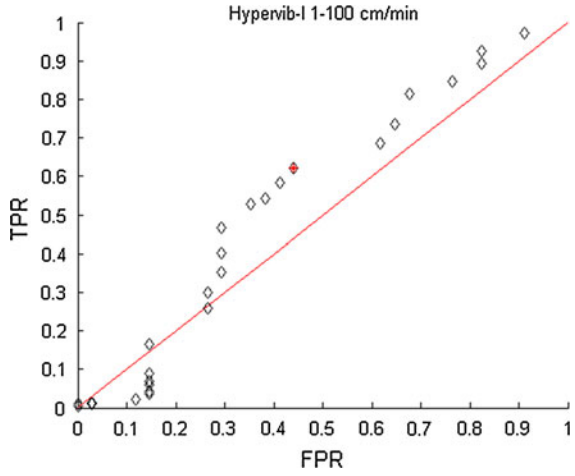

Fig. 5 ROC-curve for the Hypervib-I model, where velocity threshold Vt varies between 1 and 100 cm/min (see Sect. 3.2.1). The *red star* shows the 'best' sensitivity pair, for threshold value = 8 cm/min, assuming both the TPR and the FPR are equally important

Table 1 Example of a contingency table, or confusion matrix

	Predictions		
	−	+	Total
Obs.			
−	TN	FP	O(−)
+	FN	TP	O(+)
Total	P(−)	P(+)	N

These characteristics are depend on the threshold value used in the model. Take the Bayesian Belief Network (BBN) prediction model as an example. The BBN predicts a project to contain more unwanted events, if the resulting number (on a scale from 0 to 100 %) exceeds the threshold value, 38 % in this case. This 38 % determines more or less the result of the contingency table. Obviously, if we take 38 % as a threshold value, our contingency table will be different than if we take 50 % for a threshold. Table 2 shows the contingency table for the CUR-model, using the previously mentioned 252 observations from the GeoBrain experiences database.

4.2 Sensitivity Pair

Fortunately, a point in the ROC-space incorporates this threshold in its graph and therefore it is in fact an

Table 2 Example of a completed contingency table, or confusion matrix for the CUR rule, using threshold $d = 0.005$ m

	Predictions		
	−	+	Total
Obs.			
−	TN = 9	FP = 25	O(−) = 34
+	FN = 111	TP = 107	O(+) = 218
Total	P(−) = 120	P(+) = 132	N = 252

elaboration on the contingency table. The ROC point uses the fact that the true negative ratio (TNR(=TN/O(−))) plus the FPR equals 1, just like the TPR plus the false negative ratio (FNR(=FN/O(+))) equals 1. See also Eqs. 6 and 7. For different threshold values it is now possible to calculate the so-called sensitivity-pair (TPR,FPR). Fawcett (2006) explains this in more detail. The smaller the metric distance to the 'perfect model' (coordinates (0,1) in the ROC-space), the better the model is, assuming the TPR and the FPR are equally important. In practice, this is not correct because the costs of FP's are higher than for FN's. The effect of assigning different weights to FP's and FN's is studied in Sect. 5.2.1.

Using Table 2, the TPR equals 107/218 = 0.49 and the FPR equals 25/34 = 0.74 for the CUR rule. Therefore the sensitivity-pair (FPR, TPR) reads (0.74,0.49). Figure 1 shows this as a cross in the lower right corner. Under the assumptions that (1) the probability of a positive observation is approximately 95 % (P(O+) = 0.95) and that (2) the model predicted a positive result (P(P+) = 1), Bayes' rule (see Sect. 3.2.6) translates the true positive ratio (TPR = P(P(+)|O(+))) into a probability of failure[1] P(O(−)|P(+)) = 0.53.

4.3 Extension to ROC Curves

An additional advantage of this method is the possibility of creating ROC-curves. By changing the threshold value in a design rule, the sensitivity-pair will change as well. So, for a range of threshold values, a range of sensitivity-pairs can be constructed, resulting in a ROC –curve. One point of this curve will be closest to the perfect model and this leads to the optimal threshold value. Figures 3, 4, 5 show examples for such curves, demonstrating the behavior of the CUR-rule, the BBN prediction model and the Hypervib-I model respectively.

4.4 Conservative Predictions

Whenever a sensitivity pair is located at the lower left corner of the ROC-space, one can call the corresponding model 'conservative'. An example contingency table explains why (Table 3). This table provides a FPR of 0.12 and a TPR of 0.40, creating a sensitivity pair in the lower left corner of the ROC-space. The underlying model is better than a random guess, because TPR > FPR. Furthermore, the table shows a large amount of negative predictions (161), although in reality there were 218 positive observations. The model seems to perform very well, after all, given a positive prediction 96 % of the cases provides a positive observation. You might argue that a negative prediction will lead to a new design and therefore the project will not be carried out. Reality though proves otherwise: 161 negative predictions were carried out and 81 % of them in fact proved to be possible in contrast to the prediction. This is called conservativeness: the threshold that distinguishes between a positive and a negative prediction more often than necessary warns the professional the chosen equipment is due to fail. In 81 % of these negative predictions though, practice shows this was not necessary. If the contractor would have chosen more powerful equipment, this probably would have meant higher costs for no reason.

What is the tolerance on the difference between the prediction and the observation to define a 'positive' observation? One sensitivity-pair shows little to zero tolerance on the difference between the prediction and the observation. In practice, the contractor will always incorporate a certain tolerance to count for all

Table 3 Imaginary completed contingency table to illustrate a 'conservative' sensitivity-pair

	Predictions		
	−	+	Total
Obs.			
−	TN = 30	FP = 4	O(−) = 34
+	FN = 131	TP = 87	O(+) = 218
Total	P(−) = 161	P(+) = 91	N = 252

[1] $(O^-|P^+) = 1 - \frac{P(P^+|O^+)P(O^+)}{P(P^+)} = 1 - 0.49 \cdot 0.95 = 0.53$

uncertainties in the building pit. This means that for a positive model prediction the contractor will always investigate the available equipment at that time and use this information in I the decision. Therefore a ROC-curve probably says more about a prediction model than a single sensitivity-pair. This paper however aims to introduce the concept of sensitivity-pairs and ROC-curves. The reader is challenged to elaborate on the matter described.

A contractor will not choose a solution that is likely to fail. The probability of failure in this type of work is large and usually it is hard to calculate the financial consequences. Instead the contractor will choose slightly over dimensioned equipment, rather than the 'quick and dirty' solution. The prize for over dimensioned equipment is expected to be much lower than the costs of delay due to malfunctioning equipment and an unsafe working environment.

5 Results

All 252 projects in the GeoBrain experience database have been 'post'dicted, using the prediction tools described above. This resulted in six sensitivity-pairs, one for each tool with the standard threshold values, which have been plotted in the ROC-space below (Fig. 1). The diagonal straight line indicates the line of no discrimination. A design method with a marker below or at this line is practically worthless: one may as well 'throw a coin'. ["When throwing a coin several times, it can be expected to get half the positives and half the negatives correct; this yields the point (0.5, 0.5) in ROC space. If it guesses the positive class 90 % of the time, it can be expected to get 90 % of the positives correct, but its false positive rate will increase to 90 % as well, yielding (0.9,0.90) in ROC space", according to Fawcett (2006)].

5.1 Model Comparison in Current Situation

The ROC plot in Figs. 1 and 2 with the corresponding sensitivity pairs and metric distances in Table 4 show that the BBN (Δ) and Hypervib-I (+) score better than the EAU rule (*) and the other three design tools (o, x and ◇). This means adding professionals' experience to empirical rules improves those predictions. This is interesting, especially in those cases where there is no time for time-consuming numerical calculations. Remarkably, the EAU-rule is better than the CUR-rule and both Hypervib-I derivatives (van Baars' method and Azzouzis method, see also Sects. 3.2.3 and 3.2.4), although it does not contain any soil related parameters. Both Hypervib-I derivatives end up at the line of no discrimination. Perhaps the rules only are applicable to a specific subset of projects, because they were originally set up within a subset of the currently used variable space.

Table 4 Summary of the current and optimized threshold values for all models described here

Model	Current threshold	Current sensitivity pair	Current metric distance[a]	Optimized threshold	Optimized sensitivity pair	Optimized metric distance[a]
CUR	0.005 m	(0.74;0.49)	0.90[b]	0.0029 m	(0.91;0.97)	0.91[c]
AZZ	$F_{used}/F_c > 1$	(0.65;0.65)	0.74	[d]	[d]	[d]
Baars	$F_{used}/F_c > 1$	(0.74;0.73)	0.79	[d]	[d]	[d]
Hypervib-I	0.06 m/min	(0.65;0.74)	0.70	0.08 m/min	(0.44;0.62)	0.58
EAU	$F_{used}/F_c > 1$	(0.71;0.74)	0.76	[d]	[d]	[d]
BBN	38 %	(0.59;0.74)	0.64	36 %	(0.56;0.72)	0.63

If as stated in Sect. 4.2 the TPR and FPR are not equally important, the metric distance should be determined by factoring the vertical and horizontal distance to the 'perfect model'. Since the costs of FP's are higher than for FN's, the horizontal distance to FPR = 0 is much more important than the vertical distance to the value of TPR = 1. In the ultimate case, the FPR is the dominating aspect and the first coordinate of the sensitivity pair determines the quality of the model, with low values for the best models. As shown in Table 4 this results in only minor changes in the ranking of the models.

[a] The metric distance to the 'perfect model' is calculated by $\sqrt{(FPR^2 + (1 - TPR)^2}$

[b] Below the line of no discrimination

[c] Above the line of no discrimination

[d] Not applied yet

5.2 Improving Threshold Values

Every design rule has its own threshold value to distinguish between positive and negative predictions. Making a ROC-curve for each rule provides the best threshold for each model. Figures 3, 4 and 5 show three ROC curves for the CUR, the BBN and the Hypervib-I model respectively.

5.2.1 CUR

The threshold value for the CUR prediction varied between 0 and 0.01 m, using steps of 1e−4 m. In the current situation the threshold is 0.005 m which leads to a sensitivity pair of (0.74,0.49). This pair ends up below the 'line of no discrimination', which indicates that currently throwing a coin might provide better results than using the CUR-rule for the prediction. One might suggest to use a 'reversed' version of the rule: choose your equipment such as to stay under the required 0.005 m of displacement amplitude. Physically however, this would not make sense. In the extreme 0 displacement amplitude could be obtained and then the whole idea of vibratory driving is gone.

The optimized sensitivity pair reads (0.91,0.97), for a threshold value of 0.0029 m. Now the pair end up at the better side of the line. In the physical context a threshold of 0.0029 m is less conservative than the current threshold. This corresponds to the expert opinion that 0.005 m is quite conservative.

5.2.2 BBN

The threshold value for the BBN prediction varied between 0 and 100 %, using steps of 1 %. In the current situation the threshold is 38 % which leads to a sensitivity pair of (0.59,0.74). The optimized sensitivity pair reads (0.56,0.72), for a threshold value of 36 %. This pair ends up marginally closer to the 'perfect situation' (0,1). Practically there is no reason to change the threshold value. The metric distances column in Table 4 shows this with distance 0.64 that changes into 0.63.

5.2.3 Hypervib-I

The threshold value for the Hypervib-I prediction varied between 0.01 m and 1.00 m per 60 s, using steps of 0.01 m per 60 s. In the current situation the threshold is 0.06 m/min which leads to a sensitivity pair of (0.65,0.74). The optimized sensitivity pair reads (0.44,0.62), for a threshold value of 0.08 m/min. This pair ends up closer to the 'perfect situation' as well. The metric distances column in Table 4 shows this with the distance 0.70 that changes into 0.58.

5.2.4 Van Baars, Azzouzi and EAU Model

The Van Baars-model, the Azzouzi-model and the EAU-model have not been optimized. Both the Van Baars-model and the Azzouzi-model are derivatives from the Hypervib I-model and therefore it seems more logical to extract new derivatives from the updated original.

5.3 Model Comparison in the Improved Situation

Table 4 provides an overview of the current and the optimized threshold values. In the current situation the Hypervib-I -model and the BBN-model perform the best, compared by the others. The prediction value of the Hypervib-I model improved, using a slightly different threshold. Using the improved threshold, the Hypervib-I model performs better than the BBN-model. The van Baars model, the Azzouzi model and the EAU model have not been optimized. Remarkable is the fact that the optimized CUR-rule, the Dutch rule of thumb provides a threshold value (0.0029 m) that is comparable with the German rule of thumb that each 30 kN vertical cyclic force is needed per 100 kg sheet pile. Since in the EAU model is directly related to the frequency, this might be a coincidence. Nevertheless it is recommended to include in further investigations, keeping in mind that both the EAU model and the CUR model might have a shared origin.

6 Conclusions

This paper introduces the method of 'Receiver Operating Characteristic' to determine the quality of a model and to be able to optimize parameters and variables in the model. 252 field observations were used to re-examine prediction methods for the minimum required vibration force, based on a selection of Dutch cases and to prove the ROC method works. The Operating Characteristic (ROC)-space is suitable for three purposes: 1) determining the quality of a model,

2) objectively comparing several models to each other, given certain assumptions and 3) for threshold optimization within a model.

A sensitivity-pair, created from a confusion matrix that was filled by 252 comparisons between observations and predictions provides a position in the ROC-space that indicates the quality of the model used to make the predictions. Figure 1 shows six sensitivity-pairs, providing a quality label for the six models described in this paper, relative to the 'perfect model'. The ROC-space from Fig. 1 also enables a comparison between the six models, using the metric distance to the perfect model as a ranking order. In this ranking the numerical Hypervib-I model performs the best, closely followed by the model with added expert knowledge. A positive side effect of the ROC method comprises the possibility of threshold and parameter optimization. Figures 3,4 and 5 show the performance of three models in a ROC-curve, providing sensitivity pairs that indicate the best threshold value for each model.

Conclusively, the Receiver Operating Characteristic (ROC)-space is suitable for the objective comparison of several design models. Using project information from the GeoBrain observations database it is possible to validate the codes and to attach a performance label to them, making it much easier for an engineer or designer to choose the right code. Depending on the position in the ROC-space a design code can be labeled 'conservative' or not. Furthermore, the ROC-curve technique enables engineers to optimize threshold values in their codes, that in turn leads to better predictions and thus safer and cheaper projects. It was to be expected that the more experience is added to the rules and models, the more reliable prediction models become. Figures 1 and 2 prove this is true. The model with added professionals' experience currently performs better than all other models and after improving the numerical model it performs nearly equally well.

Appendix 1

See Fig. 6.

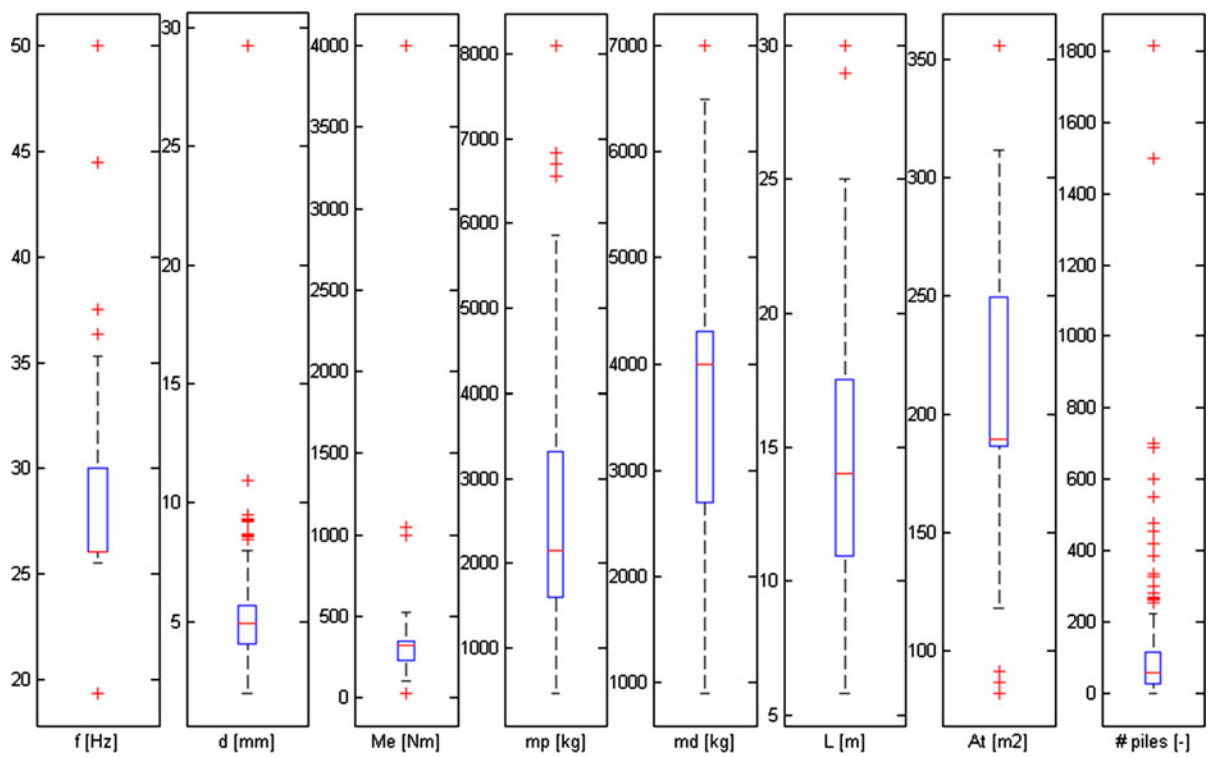

Fig. 6 *Boxplot* of the main features from the Dutch projects used for the calculations

References

Azzouzi S (2003) Intrillen van stalen damwanden in niet-cohesieve gronden—welke predictie is (on)juist? (in Dutch). Masters thesis, Delft University of Technology

Barends FBJ (2005) Associating with advancing insight—Terzaghi Oration 2005. XVI international conference on soil mechanics and geotechnical engineering. Osaka, pp 217–248

Bles T, Al-Jibouri S, van den Adel J (2003) A risk model for pile foundations. ISARC 2003, 20th international symposium on automation and robotics in construction

CUR166 (2005) Damwandconstructies, 4e druk (in Dutch). (Civieltechnisch Centrum Uitvoering Research en Regelgeving)

Dierssens G (1994) Ein bodenmechanisches Modell zur Beschreibung des Vibratios-rammens in körnigen Böden. University of Karlsruhe

EAU (1990) Empfehlungen des Arbeitsausschusses 'Ufereinfassungen'—Häfen und Wasserstrassen. Ernst und Sohn

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(Special Issue on 'ROC' Analysis in Pattern Recognition):861–874

Gonin H, Holeyman A, Rocher-Lacoste F (eds) (2006) TRANSVIB 2006: vibratory pile driving and deep soil compaction. Laboratoire Central des Ponts et Chaussées, Paris. ISBN 2-7208-2466-6

Hemmen BR (2005) The synergy between theory and practice in geo-engineering. XVI international conference on soil mechanics and geotechnical engineering. Osaka, pp 2809–2811

Holeyman A, Legrand C (1994) Soil modeling for pile vibratory driving. U.S. FHWA "International Conference on Design and Construction of Deep Foundations", Orlando, Florida, December 1994, vol II, pp 1165–1178

Holeyman A, Legrand C, Van Rompaey D (1996) A method to predict the driveability of vibratory driven piles. 3rd international conference on the application of stress-wave theory to piles. Orlando, USA, pp 1101–1112

Holeyman A, Vanden Berghe J-F, De Cock S (1999) Model testing of vibratory driven piles, vol. 2. In: Proceedings of the XIth ECSMFE, Amsterdam, June 1999, pp 769–776

Holeyman A, Vanden Berghe J-F, Charue N (2002) Vibratory pile driving and deep soil compaction. Zwets & Zeitlinger, Lisse. ISBN 90 5809 521 5

Mens AMJ, van Tol AF (2010) Validating models against experience in foundation engineering, using the ROC curve. NUMGE 2010. Trondheim

Mens AMJ, van Tol AF, Koelewijn AR (2008) Optimizing foundation engineering, validating models against experience using artificial intelligence. In: Singh DN (ed) IACMAG. Mumbai, India, pp 3384–3391

Metz CE (1978) Basic principles of ROC analysis. Seminars in nuclear medicine, vol. VIII, No. 4 (October), pp 283–298

Mitchell TM (1997) Machine learning. McGraw-Hill, Singapore

Van Baars S (2004) Design of sheet pile installation by vibration. Geotech Geol Eng 22:391–400

Vanden Berghe J-F (2001) Sand strength degradation within the framework of vibratory pile driving. Faculty of Applied Science. Louvain: université catolique de Louvain

Viking (2002) Vibrodriveability—a field study of vibratory driven sheet piles in non-cohesive soils. Division of Soil and Rock Mechanics. Stockholm, Sweden: Royal Institute of Technology