Coelho, M.R.F.; Sena-Cruz, J.; Neves, L.A.C.; Pereira, M.; Cortez, P.; Miranda, T. (2016) "Using data mining algorithms to predict the bond strength of NSM FRP systems in concrete." Construction and Building Materials, 126: 484–495.

USING DATA MINING ALGORITHMS TO PREDICT THE BOND STRENGTH OF NSM FRP SYSTEMS IN CONCRETE

Mário R. F. Coelho[1,2], José M. Sena-Cruz[1,3], Luís A. C. Neves[4,5], Marta Pereira[6,7], Paulo Cortez[8,9], Tiago Miranda[1,10]

[1] ISISE, University of Minho, Department of Civil Engineering, Campus de Azurém, 4810-058 Guimarães, Portugal

[2] E-mail: mcoelho@civil.uminho.pt

[3] E-mail: jsena@civil.uminho.pt; *Corresponding author*

[4] University of Nottingham, Department of Civil Engineering, Nottingham, United Kingdom

[5] E-mail: luis.neves@nottingham.ac.uk

[6] University of Minho, Department of Information Systems, Campus de Azurém, 4810-058 Guimarães, Portugal

[7] E-mail: martamacedopereira@gmail.com

[8] ALGORITMI Centre, University of Minho, Department of Information Systems, Campus de Azurém, 4810-058 Guimarães, Portugal

[9] E-mail: pcortez@dsi.uminho.pt

[10] E-mail: tmiranda@civil.uminho.pt

Abstract: This paper presents the effectiveness of soft computing algorithms in analyzing the bond behavior of fiber reinforced polymer (FRP) systems inserted in the cover of concrete elements, commonly known as the near-surface mounted (NSM) technique. It focuses on the use of Data Mining (DM) algorithms as an alternative to the existing guidelines' models to predict the bond strength of NSM FRP systems. To ease and spread the use of DM algorithms, a web-based tool is presented. This tool was developed to allow an easy use of the DM prediction models presented in this work, where the user simply provides the values of the input variables, the same as those used by the guidelines, in order to get the predictions. The results presented herein show that the DM based models are robust and more accurate than the guidelines' models and can be considered as a relevant alternative to those analytical methods.

Keywords: FRP; NSM; Bond; Guidelines; Data Mining

1. Introduction

The strengthening technique that uses fiber reinforced polymers (FRP) inserted in the concrete cover of the element to be strengthened is known as near-surface mounted (NSM) technique. In the last 15 years intensive research has been devoted to the NSM technique, becoming a widespread technique in practical applications in the last years [1, 2].

Nevertheless, the NSM technique presents many challenges to overcome. In particular, the characterization of the transfer of stresses between the FRP system and the surrounding concrete, i.e. the bond behavior of NSM FRP systems, is not yet completely understood. The bond behavior has been studied through direct pullout tests (DPT) and/or beam pullout tests (BPT). Figure 1 presents a generic example of both tests including some of the parameters used to quantify the bond strength discussed later in this paper.

In Coelho et al. [2], a review on these bond tests was presented and two databases collecting a wide range of DPT and BPT results were presented and used for better understand key parameters affecting the bond performance of the NSM system. These databases were also used to evaluate the accuracy and limitations of two of the most relevant guidelines for predicting the bond strength of NSM FRP systems in concrete. The first formulation is included in the "Guide for the Design and Construction of Externally Bonded FRP Systems for Strengthening Concrete Structures" from the American Concrete Institute [3]. The second guideline is the "Design handbook for reinforced concrete structures retrofitted with FRP and metal plates: beams and slabs" from Standards Australia [4]. In this paper, those guidelines will be referred to as ACI and SA, respectively.

The difficulties in modeling the bond performance arise from the high complexity of the NSM technique which involves three different materials (FRP, adhesive and concrete) and two different interfaces (FRP/adhesive and adhesive/concrete). The variety of properties (physical and mechanical) of each material and interface leads to the existence of several failure modes. However, ACI and SA guidelines are not able to capture explicitly all of them. On the other hand, that large variety of properties is associated to a large number of variables and their influence on the bond behavior of NSM FRP is far from being completely understood [2].

In an attempt to provide an alternative to the referred guidelines, this paper introduces the use of prediction models based on Data Mining (DM) algorithms. In order to provide some insights on the use of DM in structural engineering, the following section presents a brief overview on DM, focusing on its use

in the context of this work. However, contrarily to what is common in the literature, no theoretical or mathematical formulations will be provided herein. Alternatively, basic concepts will be presented since, once the fundamental concepts are perceived, extensive existing literature exists on the mathematical background and implementation of these algorithms. Some examples will be provided latter.

Finally, the results of a comparison between the accuracy of the existing guidelines (ACI and SA) and DM models is presented. This comparison was made using the same databases of pullout tests as used before to assess the accuracy of the guidelines models [2].

2. Data Mining

Traditionally, the procedure adopted to achieve any design model is fundamentally empirical and ends up being a trial and error process. Three generic main steps can be outlined: (i) identify the problem, define an initial hypothesis and define a method for testing; (ii) run the test; (iii) analyze the test results and try to infer them to identical situations. In the present work the problem to be studied is the estimation of bond strength between FRP and concrete, which is believed to be assessable by bond tests. The traditional procedure is to perform a large set of bond tests, analyze the results and extrapolate them to identical situations. For instance, the guidelines presented later in this work were developed in this way. Regarding the third step, the most common procedure is a trial and error fitting of a mathematical expression (normally chosen in order to have physical significance in that context) to the results obtained in the tests using a set of previously chosen input parameters and regression analysis. If the tests are representative of the phenomenon being studied and if the obtained expression fits well the tests results, then it would be possible to use that expression in identical scenarios. All these steps are iteratively run until an acceptable solution is found for the model describing the phenomenon being studied.

Data Mining (DM) [5], which aims at the semi-automatic extraction of useful knowledge from raw data, is an interesting alternative tool to ease and speed up the last step of the process described above. In fact, one of the several tasks that DM algorithms are capable of performing is regression, i.e. finding a data-driven model that is capable of predicting the real value of some (dependent) variable when some (independent) input variable(s) is(are) provided. The main drawback of using DM rather than traditional data analysis procedure is that the former, depending on the algorithm used, might not allow obtaining a closed form expression (easy to understand) for the prediction model. Instead, several DM models are based in terms of complex mathematical functions or rules, thus the user can only see the

input and output variables, in what is often termed as "black-box" models [6]. As an advantage, the DM approach simplifies the data analysis process [7]. In effect, DM models tend to be more flexible, being capable of predicting complex nonlinear mappings and dealing with large amounts of data or noise. Such model learning flexibility often leads to higher predictive performances when compared with classical statistical models (e.g., multiple regression).

DM algorithms have been successfully used in regression tasks in many areas, including Civil Engineering [8-11]. More specifically, in the field of concrete structures strengthened with FRP systems there are examples where DM algorithms have been used to predict the lateral confinement coefficient for reinforced concrete columns wrapped with CFRP [12], the strength of FRP confined concrete cylinders [13], the shear strength of reinforced concrete beams reinforced with FRP systems using either the externally bonded (EBR) [14] or the NSM [15] techniques or even the bond strength of FRP EBR systems in concrete [16]. According to the author's best knowledge, only one work of their authorship is available where DM algorithms were applied to predict the bond strength of NSM FRP systems in concrete [17].

In this work, two DM algorithms were used: the Artificial Neural Networks (ANN) and the Support Vector Machines (SVM). These DM algorithms are briefly presented in the following sections.

2.1. Artificial Neural Networks

The Artificial Neural Network (ANN) is an algorithm that is inspired in the behavior of the human central nervous system. Hence, the learning ANN algorithm aims at finding the best connection weights in which a set of artificial neurons should communicate with each other in order to attain a certain target [18].

Figure 2 presents two ANN examples: (i) Figure 2(a) corresponds to a multiple linear regression, which is a widely known and commonly accepted type of regression model. This is an example of the simplest ANN, without hidden nodes; (ii) Figure 2(b) corresponds to a more complex ANN with one hidden layer and two hidden neurons (HN). As it can be seen, the only difference between them is the existence or not of an intermediate layer of hidden neurons.

In the multiple linear regression, several input variables (x) affected by different weights are combined and an output variable (y) is obtained. In the ANN with one hidden layer intermediary weights are also introduced thus a nonlinear relation between x and y can be obtained. The number of hidden

layers and neurons can be different from this example and, by increasing them, the degree of nonlinearity increases.

If the value of *y* is known *a priori*, then the multiple linear regression model is an expression identic to expression (1), where the only unknown is the set of weights (*w*) that make the equality true. In the case of ANN with hidden layers, such an expression is no longer straightforward to obtain. However, a similar procedure minimizing the difference between the predicted and observed values can be used to find the optimal weights, in a process called training.

The type of ANN adopted in this work uses only one hidden layer since this is the simplest nonlinear ANN and was found to attain good results. The number of hidden neurons determined during the analysis by comparing the quality of fit with increasing number of neurons (between 0 and 9) and selecting the one which presents lower prediction errors (when considering training data).

$$y = w_0 + \sum_{i=1}^{n} w_i x_i \tag{1}$$

2.2. Support Vector Machines

Support vector machines (SVM) can be seen as an upgrade to the ANN and were initially developed for classification tasks [19]. Considering the classification purpose, the basic concept of SVM is finding an optimal hyperplane for linearly separate patterns, i.e., finding the plane which maximizes the separation between the different patterns that exist in the analyzed data. To ease the understanding of SVM functioning in a classification task, Figure 3 presents an example of a database with two input variables (x_1 and x_2) divided in two patterns (circles and squares). In the database real space (middle chart in Figure 3) those patterns can only be separated using a curved line. However, it can be found a function ϕ_1 which, applied to the original data, can transform it into a new high dimensional space where the two patterns can actually be separated by a straight line. SVM algorithm optimizes the position of that single line such that it maximizes the separation of the two patterns. Several division lines can exist and are represented as full lines in the left side of Figure 3. However, in this case, the line that maximizes the separation of the patterns is the thicker one represented in that figure. Remark that, in more complex examples (with several variables), the lines would be actually hyperplanes, as referred before. In the end, since the optimal hyperplane is known, the relative position of all the data points, especially those passed by the dashed lines (designated by support vectors) is also known. Hence, a model traducing the separation of

the patterns can be defined which corresponds to the classification model that was sought in the beginning.

SVM were latter extended to also perform regression tasks, which are the important ones in the scope of this work, being its functioning similar to the classification case. However, in regression, another function ϕ_2 will transform the original data in order to find a line that passes through all data points (right chart in Figure 3). That line is the regression function which allows predict the value of each data point. Since finding such a line is quite complex, there are two new important parameters in the SVM for regression, namely the regularization parameter (C) and a loss function that in this work is the ε-insensitive (ε). The first defines the tradeoff between complexity and accuracy of the model to be found, while the second defines the width of a region in which the data points inside it are assumed to be on the regression line, thus an insensitive region. The data points outside this region are the support vectors in the regression SVM.

Besides these two parameters, the success of SVM for regression tasks is influenced by a kernel function. In this work, the Gaussian radial basis kernel function was adopted (2). This has only one hyperparameter, γ, which was adjusted using a greedy search (between 2^{-15} and 2^{3}). Similar procedure was also adopted for parameter ε, while parameter C was considered equal to 3 [20].

$$k(x,x') = exp\left(-\gamma\left\|x-x'\right\|^2\right), \gamma > 0 \quad (2)$$

2.3. Rminer tool

Nowadays, there are several tools that allow an easy application of DM algorithms with a limited knowledge of the mathematical background required for implementation. In this work, the rminer library [20] of the R Statistical Environment [21] was adopted, since it is particularly suited for generating ANN and SVM data-driven models.

Among the several features included in rminer, in this work the functions *mining*, *fit* and *predict* were used. For simplicity, the functions will be described using a parallel with a simple regression model.

The function *fit* allows finding an analytical expression in the form $y = mx + b$ with m and b adjusted to the database in analysis. Having the expression calibrated, *predict* gives the results (y) for new values of the independent variable (x) by replacing it in the expression found by *fit*. The function *mining* is a more sophisticated function. It performs several runs (i.e., sequences of *fit* and *predict* executions) under a user selected validation method. It is important to emphasize that, while *fit* uses the entire

database to adjust a model, *mining* only uses part of it, being the fitted model tested in unseen data (i.e., test set). This aspect is very important since it allows evaluating the performance of the adjusted model when applied to new data (depending on the validation method), thus measuring the true generalization capacity of the DM model. In this work, a holdout split validation method was adopted, in which 2/3 of the data entries were randomly selected as training data and the remaining 1/3 samples were used as test data. Another important difference is that only *fit* function allows storing a model that can be then used, like an analytical expression, to perform new predictions. In fact, depending on the chosen division of sets and number of runs, for example, *mining* function can produce a huge number of models. For practical reasons, the rminer library does not store any of these models.

3. Tests and analyses

The following paragraphs present the databases of tests used in this work. Then the ACI and SA analytical formulations, used as reference, are presented. Finally, the DM analyses carried out in this work are detailed.

3.1. Databases of pullout tests

As referred, two databases of pullout tests were built, one including 363 direct pullout tests (DPT) and other with 68 beam pullout tests (BPT). In the context of the present work, it was decided to build up a webpage to store the referred databases (www.frpbondata.civil.uminho.pt). It is believed that providing the scientific community free access to the vast majority of pullout tests available in the literature makes the process of continuously improving the existing prediction models faster and easier. It is expected that, with the contribution of all the researchers working in this field, this website will be continuously updated.

The referred website includes, besides the databases, a page to perform predictions of the maximum pullout force (F_{fmax}) using different formulations. It includes ACI and SA guidelines and the DM models developed herein. It is also believed that, by providing in the website an easy way of using and testing DM models, the acceptance and use of such powerful tools will increase. Hence, providing the required input variables, results obtained using all the prediction formulations described herein will be readily available.

To help the community in improving prediction models for NSM bond behavior a detailed and comprehensive data visualization tool is also included in the webpage. In addition, a Forum is also available to ease the interaction between all the researchers contributing for the website.

Regarding the details of the databases used in this work, for the sake of brevity, Table 1 presents an overview of the variables available in each database included in the final models only. This table also shows the range of values used for each parameter, identifying bounds of application of the proposed models.

More detailed information regarding the other variables included in the databases, as well as an overview of some of the main conclusions that can be drawn from these databases can be found in the webpage referred above and in [2].

3.2. Analytical formulations

ACI and SA formulations are summarized in Table 2. A detailed description of the guidelines and of their application to the databases presented can be found in [2].

Since in all the analyses, mean values for the mechanical properties and no additional safety factors or strength reductions were considered, a slight modification was made to ACI and SA formulations. This included the use of FRP ultimate tensile strength (f_{fu}) and concrete mean compressive strength (f_{cm}), instead of using their design values as defined in the guidelines. Those values (f_{fu} and f_{cm}) were estimated experimentally by the authors of the experimental works included in the databases.

3.3. Data mining analyses

A total of eight DM analyses were performed for each database, as shown in Table 3. Firstly, two types of analyses, denoted A and B, were considered. In the first, the input variables were defined based on the guidelines (ACI and SA). In the second type the input variables were estimated during the analysis by using an automatic selection process (RM) or by combining that with expert judgment (User). This resulted in 4 sets of input variables. For each set of input variables, models using both ANN and SVM algorithms were generated. The next paragraphs detail each of these analyses.

Analyses Type A were conducted assuming for DM models the same input variables as used by the guidelines' models. Hence, one analysis used the input variables considered by ACI (L_b, p_f, A_f, f_{fu})

while the other used those from SA (L_b, A_f, f_{fu}, d_g, b_g, E_f, f_{cm}). This allowed the direct comparison between the performance of DM and guidelines' models.

Each analysis of Type A consisted on running *mining* function over each database. A total of 20 runs were imposed being the database divided in four random sets of equal size (3 for training and 1 for testing). Then the prediction error metrics fluctuation was analyzed in order to check generalization capacity of each DM algorithm. To this purpose, the 95% t-student confidence interval was adopted. Finally, the error metrics obtained in all the 20 runs were averaged to allow comparisons between model's accuracy.

In analyses Type B, it was assumed that the input variables were not known *a priori*. Hence, besides the four and seven variables used by ACI and SA, respectively, all the numeric variables present in more than 2/3 of the records in each database were also included. This resulted in more than 20 input variables available on each database at the beginning of the calibration process.

The same procedure used in the analyses Type A was used for these new and larger databases. In the end of the *mining* sequence, a sensitivity analysis was performed in order to identify the most important variables in a backward selection procedure. After identifying the most important variables, the procedure was repeated with the limited input variables. This process was carried out several times, being the number of input variables successively reduced. In the end, a final set of input variables could be proposed as well as the DM models using those input variables.

Since this sensitivity analysis is influenced by the representativeness of each variable in the database, in some cases the final set of variables was found to be meaningless for design purposes. Hence, a different type of models were generated, taking into account the evolution of the variable's importance in the sensitivity analyses and also including all the variables thought meaningful for design.

Since in the first case the variables were chosen taking into account only the rminer sensitivity analysis, these were designated by RM. In the second case, since the choice was made by the user, the designation User was adopted instead.

Finally, it should be emphasized that all the analyses carried out used the maximum pullout force (F_{fmax}) as the only output variable. Also, in all the analyses, variables normalization was considered using a zero mean and a one standard deviation transformation for all input and output variables (-1 to 1 scale). Then, the inverse procedure was performed for the output variable in order to export it in its original scale.

4. Results

For each analysis three error metrics were calculated, namely, the mean absolute error (*MAE*), the root mean squared error (*RMSE*) and squared correlation coefficient (R^2). Those are defined in the equations (3) to (5), respectively. In these equations, the error e_i for the i^{th} specimen of the total N, is the difference between the numerical prediction of the maximum pullout force ($F_{fmax,Num}$) and its experimental value ($F_{fmax,Exp}$), as illustrated in equation (6). In equation (5), the parameters with an upper bar, represent the average value of the corresponding parameter.

$$MAE = \sum_{i=1}^{N} |e_i| \Big/ N \tag{3}$$

$$RMSE = \sqrt{\sum_{i=1}^{N} e_i^2 \Big/ N} \tag{4}$$

$$R^2 = \left[\frac{\sum_{i=1}^{N} \left(\left(F_{f\max,Exp} \right)_i - \overline{\left(F_{f\max,Exp} \right)_i} \right) \left(\left(F_{f\max,Num} \right)_i - \overline{\left(F_{f\max,Num} \right)_i} \right)}{\sqrt{\sum_{i=1}^{N} \left(\left(F_{f\max,Exp} \right)_i - \overline{\left(F_{f\max,Exp} \right)_i} \right)^2 \times \sum_{i=1}^{N} \left(\left(F_{f,\max,Num} \right)_i - \overline{\left(F_{f\max,Num} \right)_i} \right)^2}} \right]^2 \tag{5}$$

$$e_i = \left(F_{f\max,Num} \right)_i - \left(F_{f\max,Exp} \right)_i \tag{6}$$

Analyses Type A

Tables 4 and 5 present the average error metrics (*MAE*, *RMSE* and R^2) obtained in the 20 runs of *mining* function performed for all the analyses with DPT and BPT databases, respectively. Those metrics include, in parenthesis, the correspondent 95% t-student confidence intervals to allow verifying the stability of the predictions. For all the analyses presented, it was found that they are quite stable and capable of being used in unseen data since they presented simultaneously low errors and low dispersion values along the 20 runs performed on different data sets as shown by the low values of 95% t-student confidence intervals obtained.

Additionally, in these tables are also included the same error metrics obtained when applying to each database ACI and SA formulations, as defined in each guideline. Note that the number of specimens considered was not the same in all the analyses. This number depends on the input variables required in each analysis, which were not always available in the databases because the authors of the corresponding

experimental tests did not provide them. Nevertheless, the analyses can still be compared since the same number of specimens was used for each group of analyses using the same input variables.

Comparing the analyses Type A (using ACI and SA input variables) it can be seen that, for both databases, the worst results (higher *MAE* and *RMSE* and lower R^2) were obtained by the guidelines. As already verified in a previous work, SA presents better performance than ACI even though its R^2 value is lower [2].

In terms of DM models, for both databases, using SA input variables attained better results. Regarding DPT database, when ACI input variables are used, *MAE* and *RMSE* of both DM models (ANN and SVM) are at least 20% lower while R^2 is at least 40% bigger. When SA input variables are used, *MAE* and *RMSE* of both DM models (ANN and SVM) are at least 24% lower while R^2 is at least 50% bigger. In the case of BPT database, the improvement in the results is even bigger. The major difference when compared with the results of DPT database, is the fact that the error metrics are almost the same in both analyses Type A and B. This means that the improvements achieved with the DM models obtained in analyses Type B were lower for BPT database.

Analyses Type B

For analyses Type B, the first result to be considered is the importance of each variable in the prediction of the bond strength. In Tables 4 and 5, these variables are presented by decreasing order of importance. Further discussion about this subject will be given in following paragraphs.

A common aspect for both databases is that all analyses Type B presented better results than those from the guidelines, being the best results obtained using SVM and ANN algorithms for DPT and BPT databases, respectively. When compared with ACI results, the three metrics of the all four DM models are at least 50% better. When compared with SA results, the three metrics of the all four DM models are at least 40% better. In both cases, better means that *MAE* and *RMSE* are lower while R^2 is bigger.

In the case of DPT database, the RM input variable's selection, lead to include, as input variable, the concrete block length (L_c – see Figure 1). However L_c is not relevant from design viewpoint. On the other hand, RM selection did not included any input variable related with concrete nor adhesive mechanical properties. Hence, User selection process, which took into account both importance and relevance of each variable, proposes a different set of input variables where adhesive and concrete

mechanical properties are also represented. Analyzing the error metrics, it can be seen that RM analyses are slightly better. However, taking into account that User input variables are more reasonable to be used, the error metrics are still acceptable.

In the case of BPT database, the major difference between RM and User input variables is related with the removal of FRP modulus of elasticity (E_f), since there was already a more important variable related with FRP mechanical properties, and the inclusion of adhesive compressive strength (f_{ac}), in order to have the adhesive mechanical properties represented. Regarding the error metrics, User analyses attained better results.

The relative importance of each input variable obtained in all analyses Type B is summarized in Figure 4. Comparing the relative importance of each variable when the RM input variable's selection is used, the results differ between DPT and BPT databases. In DPT (Figure 4a), since the geometric variables appear in larger number, it seems that the geometry of specimen and the configuration of the strengthening have more impact in the predictions than the mechanical properties of the involved materials. In BPT (Figure 4c), both geometric and mechanical parameters appear in the same number.

Another interesting aspect is related with the variables' interaction that was found during the process of selecting the input variables. For example, considering the importance ranks depicted in Figure 4a and b it can be seen that, besides L_b, there is no other common variable in the two figures. However, as referred above, the only actions taken when moving from RM to User analysis, were the removal of L_c and the addition of f_{at} and f_{cm}. But when the sensitivity analysis was re-run, using the new set of variables, it was found that p_f, p_g and a_e were more important than their equivalents in RM set, i.e. A_f, d_g and b_c, respectively a variable referring to FRP geometry, groove geometry and location of the NSM FRP system in the concrete element. This suggests that there is interaction between variables which is the reason why the final set of variables suggested by the User (Figure 4b) is completely different from RM final set (Figure 4a).

5. Using DM models

As referred before, the analyses carried out using *mining* function do not allow storing a prediction model. Hence, the final DM models to be proposed were obtained by running *fit* function over each entire database. Since those final models were intended to be made available in the website that stores the databases (see section 3.1), where also guideline formulations can be easily applied, only DM models

using guideline input variables were generated. Hence, those willing to compare the maximum pullout force (F_{fmax}) obtained in their pullout tests, just need to provide the guidelines input variables and specify the type of test they are comparing with. Then, by clicking the "Calculate" button available in the website's page, six values of F_{fmax} prediction are obtained. The first two correspond to the guidelines ACI and SA (the step-by-step calculation procedure can also be seen). The remaining four predictions correspond to those obtained by DM models. Two correspond to the two DM models based on ANN algorithm using either ACI or SA input variables. The last two predictions are identic to the former two, but are based on SVM algorithm instead. Figure 5 presents an example of a prediction run in the website. Remark that the experimental value of that example was 20.4 kN.

Table 6 presents the error metrics for these final four models for both DPT and BPT databases. As it can be seen, the error metrics of these models are even lower than all the corresponding analyses presented so far. This is mainly related with the fact that *fit* function uses the entire database to adjust a model while in all the analyses with *mining* function only 3/4 of each database were being used for models adjustment.

To ease the comparison between guidelines and DM models prediction capability, Figure 6 presents the relationship between experimental and predicted pullout force obtained when each DM model included in Table 6 is applied and also when ACI and SA guidelines are applied. As it can be seen the clouds of points related with the guidelines models are larger than those of the DM models, revealing higher dispersion of the predictions.

In the importance charts presented in Figure 4 the bonded length (L_b) was always found to be the most important variable in the prediction of the maximum pullout force. Hence, L_b was selected to access the stability of the predictions obtained by each model. Figure 7 presents the relationship of the ratio between the quantities plotted in Figure 6, i.e. maximum pullout force predicted by each model ($F_{fmax,Num}$) and that obtained in the experimental tests ($F_{fmax,Exp}$), *versus* the bonded length. This figure allows to see that the guidelines' models performance is influenced by L_b, producing safe results for lower values of L_b and results successively more unsafe as L_b increases. Contrarily, this ratio for DM models is almost constant, revealing that the performance of DM models is not influenced by the variation of L_b.

Nevertheless, it is interesting to verify that, for both guidelines (ACI and SA) and with both databases (DPT and BPT), the amount of data points below the 45º line in Figure 6 or below the line

where the ratio $F_{fmax,Num}/F_{fmax,Exp}$ is 1 in Figure 7, is in general greater than above these lines. This means that the guidelines' predictions tend to be conservative, as already verified in a previous work [2].

In order to show the generalization capability of the proposed DM models, Figure 8 presents variable effect characteristic (VEC) curves [6,22] for bonded length, L_b. These curves reproduce the influence of L_b in the predictions, as it changes from its minimum to its maximum value in each database. The VEC curves were obtained by dividing, in each database, the range of L_b into several parts. Then, all the final DM models included in Table 6, as well as both ACI and SA guidelines, were applied using each value of L_b and the average values of all the remaining variables required by each model.

As can be seen, in terms of ACI and SA guidelines predictions, Figure 8b is just a zoom of Figure 8a, due to the smaller range of L_b values available in BPT database. If ACI and SA curves in both figures were overlapped, they will coincide, since the same variables were applied for both DPT and BPT databases. These curves show that using SA guideline the "average specimen" (fictitious specimen with all parameters on their average value) has a development length, L_d, of about 270 mm. Such threshold L_d was not predicted by ACI guideline.

Analyzing DM models predictions, two main conclusions can be drawn. The first is that, except for the model SVM_ACI (using SVM algorithm and ACI input variables) in Figure 8a, all other DM models in both figures present either ACI or SA guidelines' trends. Secondly, for those DM models that captured L_d, for values of L_b greater than L_d, the maximum pullout force (F_{fmax}) remained almost constant, as it should be. These two conclusions show that the DM models developed herein have the required generalization capacity.

6. Conclusions

In this work, a better understanding of the bond performance of NSM FRP systems was achieved by using data mining (DM) as an alternative to the existing analytical formulations (ACI and SA guidelines models) to predict the bond strength of such strengthening systems. All the analyses presented in this work were based on two large databases of direct and beam pullout tests with NSM FRP systems.

Regarding analyses Type A (using the input variables suggested by ACI and SA guidelines):

- they showed a direct comparison between the predictive capacity of guidelines models and DM models using the same input variables. In the end, all DM models performed better than the equivalent guidelines models;

- the DM models were find to be stable since the fluctuation of the error metrics was found to be quite low along the 20 runs conducted for each DM model.

Regarding analyses Type B (using sets of input variables suggested in this work):

- they showed that the maximum pullout force in NSM FRP bond tests could be better predicted if a set of input variables different from those adopted by guidelines is used;

- the sensitivity analyses conducted to choose the new input variables can lead to include variables that are not relevant for design, thus it was necessary to replace some input variables by other thought more significant. However, the impact in the predictive capacity of the DM models with this new set of input variables was quite low, thus there can be obtained DM models suitable for design and maintaining high accuracy.

Regarding the database website:

- in order to spread and encourage the use of DM in this field, the best DM models obtained herein were made available in a website built for that purpose. Only DM models using the same input variables as used in the analyzed guidelines were considered;

- the guidelines models predictive capacity seems to be influenced by the value of the bonded length. Contrarily, the predictive capacity of the final DM models were found to be independent from this important variable.

Finally, the generalization capacity of the proposed DM models was demonstrated. For this purpose, the bonded length was selected to conduct the parametric studies. These studies proved that the DM models are in agreement with the guidelines, thus they have the required generalization capacity.

Acknowledgements

This work was supported by FEDER funds through the Operational Program for Competitiveness Factors - COMPETE and National Funds through FCT (Portuguese Foundation for Science and Technology) under the project CutInDur PTDC/ECM/112396/2009 (Ref. PTDC/ECM/112396/2009) and partly financed by the project POCI-01-0145-FEDER-007633. The first author wishes also to acknowledge the Grant No. SFRH/BD/87443/2012 provided by FCT.

References

[1] De Lorenzis L, Teng JG. Near-surface mounted FRP reinforcement: An emerging technique for strengthening structures. Composites Part B: Engineering. 2007;38:119-43.

[2] Coelho M, Sena Cruz J, Neves L. A review on the bond behavior of FRP NSM systems in concrete. Construction and Building Materials. 2015;93:1157–1169.

[3] ACI. Guide for the Design and Construction of Externally Bonded FRP Systems for Strengthening Concrete Structures. Report by ACI Committee 4402R-08. American Concrete Institute, Farmington Hills, MI, USA. 2008. p. 76.

[4] SA. Design handbook for RC structures retrofitted with FRP and metal plates: beams and slabs. HB 305 - 2008. Standards Australia GPO Box 476, Sydney, NSW 2001, Australia. 2008. p. 76.

[5] Fayyad U, Piatesky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery: an overview. In: ed. F, editor. Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Cambridge MA. 1996. p. 471-93.

[6] Cortez P, Embrechts MJ. Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences. 2013;225:1-17.

[7] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine. 1996;17:37-54.

[8] Martins FF, Miranda TFS. Estimation of the Rock Deformation Modulus and RMR Based on Data Mining Techniques. Geotechnical and Geological Engineering. 2012;30:787-801.

[9] Tinoco J, Gomes Correia A, Cortez P. Application of data mining techniques in the estimation of the uniaxial compressive strength of jet grouting columns over time. Construction and Building Materials. 2011;25:1257-62.

[10] Chojaczyk AA, Teixeira AP, Neves LC, Cardoso JB, Guedes Soares C. Review and application of Artificial Neural Networks models in reliability analysis of steel structures. Structural Safety. 2015;52, Part A:78-89.

[11] Garzón-Roca J, Marco CO, Adam JM. Compressive strength of masonry made of clay bricks and cement mortar: Estimation based on Neural Networks and Fuzzy Logic. Engineering Structures. 2013;48:21-7.

[12] Doran B, Yetilmezsoy K, Murtazaoglu S. Application of fuzzy logic approach in predicting the lateral confinement coefficient for RC columns wrapped with CFRP. Engineering Structures. 2015;88:74-91.

[13] Cevik A. Modeling strength enhancement of FRP confined concrete cylinders using soft computing. Expert Systems with Applications. 2011;38:5662-73.

[14] Lee S, Lee C. Prediction of shear strength of FRP-reinforced concrete flexural members without stirrups using artificial neural networks. Engineering Structures. 2014;61:99-112.

[15] Perera R, Tarazona D, Ruiz A, Martín A. Application of artificial intelligence techniques to predict the performance of RC beams shear strengthened with NSM FRP rods. Formulation of design equations. Composites Part B: Engineering. 2014;66:162-73.

[16] Mashrei MA, Seracino R, Rahman MS. Application of artificial neural networks to predict the bond strength of FRP-to-concrete joints. Construction and Building Materials. 2013;40:812-21.

[17] Coelho M, Sena Cruz J, Dias S, Miranda T. Evaluation of code formulations for NSM CFRP bond strength of RC elements. FRPRCS-11. Guimarães, Portugal. 2013. p. 10.

[18] Haykin S. Neural networks and learning machines. 2009.

[19] Cortes C, Vapnik V. Support vector networks. Machine Learning. 1995;20:273-97.

[20] Cortez P. Data Mining with Neural Networks and Support Vector Machines Using the R/rminer Tool. In: Perner P, editor. Advances in Data Mining Applications and Theoretical Aspects: Springer Berlin Heidelberg; 2010. p. 572-83.

[21] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org. 2012. p.

[22] Çevik A, Kurtoğlu AE, Bilgehan M, Gülşan ME, Albegmprli HM. Support vector machines in structural engineering: a review. Journal of Civil Engineering and Management. 2015;21:261-81.

Table Captions

Table 1 – Range of the variables used in the prediction models.

Table 2 – Summary of ACI and SA guidelines' formulations.

Table 3 – Summary of the analyses performed.

Table 4 – Average error metrics obtained after 20 runs of *mining* function in the DPT database (best values in **bold**).

Table 5 – Average error metrics obtained after 20 runs of *mining* function in the BPT database (best values in **bold**).

Table 6 – Error metrics for the final models obtained by fitting DM algorithms to each entire database (best values in **bold**).

Table 1 – Range of the variables used in the prediction models.

Direct pullout tests database			Beam pullout tests database		
Variable	Number of records[1]	Range	Variable	Number of records[2]	Range
b_c [mm]	325	[90-300]	L_{arm} [mm]	56	[67-212.4]
L_c [mm]	361	[152-1000]	L_b [mm]	68	[40-304.8]
b_g [mm]	340	[3-50]	b_g [mm]	68	[3.3-25.4]
d_g [mm]	359	[5-60]	d_g [mm]	68	[7-26]
p_g [mm]	340	[27.2-100]	f_{cm} [MPa]	68	[26.7-73.5]
a_e [mm]	325	[11.5-150]	f_{ct} [MPa]	68	[2.47-6.01]
L_b [mm]	363	[30-510]	Ec [GPa]	68	[29.54-47.88]
f_{cm} [MPa]	309	[18.4-65.7]	d_f [mm]	68	[4.55-20]
E_f [GPa]	361	[37.17-273]	E_f [GPa]	68	[33.93-171]
f_{fu} [MPa]	363	[512-3100]	f_{fu} [MPa]	68	[773-2833]
ε_{fu} [‰]	363	[7.4-30]	ε_{fu} [‰]	68	[11.21-32.72]
A_f [mm^2]	363	[12-201.06]	A_f [mm^2]	68	[12.65-143.14]
p_f [mm]	363	[18.85-84.8]	p_f [mm]	68	[15.1-45]
f_{at} [MPa]	307	[8-62.05]	f_{ac} [MPa]	56	[44.4-87.7]

Note: [1]from a total of 363 specimens; [2]from a total of 68 specimens.

Table 2 – Summary of ACI and SA guidelines' formulations.

Parameter	ACI guideline	SA guideline
Development length [L_d]	$\frac{A_f f_{fd}}{p_f \tau_{avg}}$	$\frac{\pi}{2\sqrt{\frac{\tau_{\max} L_{per}}{\delta_{\max} (EA)_f}}}$
Maximum pullout force [F_{fmax}]	$\begin{cases} A_f f_{fd} & if\ L_b \geq L_d \\ A_f f_{fd} \frac{L_b}{L_d} & if\ L_b < L_d \end{cases}$	$\begin{cases} \sqrt{\tau_{\max} \delta_{\max}} \sqrt{L_{per} (EA)_f} \leq A_f f_{fd} & if\ L_b \geq L_d \\ \sqrt{\tau_{\max} \delta_{\max}} \sqrt{L_{per} (EA)_f} \frac{L_b}{L_d} \leq A_f f_{fd} & if\ L_b < L_d \end{cases}$
Comments	$\tau_{avg} = 6.9$ MPa	$\tau_{\max} = (0.8 + 0.078 \varphi_{per}) f_c^{0.6}$ $\delta_{\max} = \left(0.73 \varphi_{per}^{0.5} f_c^{0.67}\right) / \tau_{\max}$ $\varphi_{per} = \left(d_g + 1\right) / \left(b_g + 2\right)$ $L_{per} = 2\left(d_g + 1\right) + b_g + 2$

Table 3 – Summary of the analyses performed.

Database	Type A (Input variables known *a priori*)		Type B (Input variables unknown *a priori*)	
	Input variables	DM algorithm	Input variables	DM algorithm
DPT	ACI	ANN	RM	ANN
		SVM		SVM
	SA	ANN	User	ANN
		SVM		SVM
BPT	ACI	ANN	RM	ANN
		SVM		SVM
	SA	ANN	User	ANN
		SVM		SVM

Table 4 – Average error metrics obtained after 20 runs of *mining* function in the DPT database (best values in **bold**).

Inputs origin	*Type A*						*Type B*			
	ACI			*SA*			*RM*		*User*	
Input variables	L_b, p_f, A_f, f_{fu}			L_b, A_f, f_{fu}, d_g, b_g, E_f, f_{cm}			L_b, A_f, b_c, L_c, d_g, ε_{fu}		L_b, p_f, f_{at}, ε_{fu}, p_g, a_e, f_{cm}	
Model	ACI*	ANN	SVM	SA*	ANN	SVM	ANN	SVM	ANN	SVM
MAE [kN]	14.85	10.10 (±0.14)	**9.82 (±0.12)**	11.56	7.92 (±0.16)	**7.07 (±0.11)**	**5.64**	5.75	6.14	**5.70**
RMSE [kN]	19.34	15.38 (±0.29)	**14.93 (±0.2)**	15.16	11.52 (±0.27)	**10.67 (±0.16)**	8.60	**8.17**	8.71	**8.22**
R^2 [-]	0.58	0.82 (±0.01)	**0.83 (±0.01)**	0.53	0.80 (±0.01)	**0.83 (±0.01)**	0.89	**0.90**	0.88	**0.89**
Specimens [-]	363			286			208			

Note: The values in parenthesis are the correspondent 95% t-student confidence intervals. *Analysis according to the guideline.

Table 5 – Average error metrics obtained after 20 runs of *mining* function in the BPT database (best values in **bold**).

Inputs origin	*Type A*						*Type B*			
	ACI			*SA*			*RM*		*User*	
Input Variables	L_b, p_f, A_f, f_{fu}			L_b, A_f, f_{fu}, d_g, b_g, E_f, f_{cm}			L_b, ε_{fu}, L_{arm}, f_{ctm}, E_{cm}, E_f, d_f		L_b, ε_{fu}, L_{arm}, f_{ctm}, d_f, f_{ac}	
Model	ACI*	ANN	SVM	SA*	ANN	SVM	ANN	SVM	ANN	SVM
MAE [kN]	10.65	**3.98 (±0.17)**	4.63 (±0.31)	7.18	**3.18 (±0.26)**	3.62 (±0.16)	**3.62**	3.67	3.56	3.56
RMSE [kN]	13.56	**5.51 (±0.26)**	6.94 (±0.39)	8.90	**4.42 (±0.56)**	5.54 (±0.23)	**4.86**	5.10	**4.76**	4.97
R^2 [-]	0.43	**0.88 (±0.01)**	0.80 (±0.03)	0.62	**0.92 (±0.02)**	0.88 (±0.01)	0.88	0.88	**0.89**	0.88
Specimens	68						56			

Note: The values in parenthesis are the correspondent 95% t-student confidence intervals. *Analysis according to the guideline.

Table 6 – Error metrics for the final models obtained by fitting DM algorithms to each entire database (best values in **bold**).

Inputs origin	*ACI*				*SA*			
Input variables	L_b, p_f, A_f, f_{fu}				L_b, A_f, f_{fu}, d_g, b_g, E_f, f_{cm}			
Database	DPT		BPT		DPT		BPT	
Model	ANN	SVM	ANN	SVM	ANN	SVM	ANN	SVM
MAE [kN]	7.36	**6.93**	1.87	**1.53**	**3.78**	3.87	1.10	**0.62**
RMSE [kN]	10.77	**10.26**	2.50	**2.49**	**5.61**	5.77	1.48	**1.12**
R^2 [-]	0.84	**0.86**	0.95	0.95	0.91	0.91	0.98	**0.99**
Specimens [-]	363		68		286		68	

Figure Captions

Figure 1 – Direct (left) and beam (right) pullout tests for NSM FRP in concrete.

Figure 2 – Example of ANN: (a) without hidden layers; (b) with one hidden layer.

Figure 3 – Example of SVM classification (left) and regression (right) of non-linear data (middle).

Figure 4 – Relative importance of each input variable in the analyses Type B (*database* and *input variables*): (a) *DPT* and *RM*; (b) *DPT* and *User*; (c) *BPT* and *RM*; (d) *BPT* and *User*.

Figure 5 – Maximum pullout force prediction calculated in the website developed.

Figure 6 – Experimental *versus* predicted pullout force for the final models obtained by fitting DM algorithms (*database* and *input variables*): (a) *DPT* and *ACI*; (b) *DPT* and *SA*; (c) *BPT* and *ACI*; (d) *BPT* and *SA*.

Figure 7 – Variation of the ratio between experimental and predicted pullout force with the bonded length (*database* and *input variables*): (a) *DPT* and *ACI*; (b) *DPT* and *SA*; (c) *BPT* and *ACI*; (d) *BPT* and *SA*.

Figure 8 – VEC curves for L_b considering (a) *DPT* or (b) *BPT* databases. Note: composite designations include the DM model and the type of input variables, as defined in Table 6.

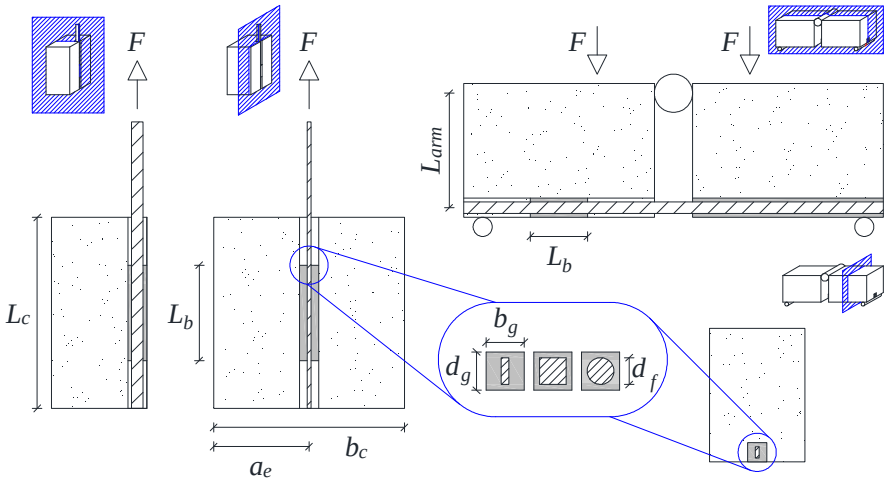


Figure 1 – Direct (left) and beam (right) pullout tests for NSM FRP in concrete.

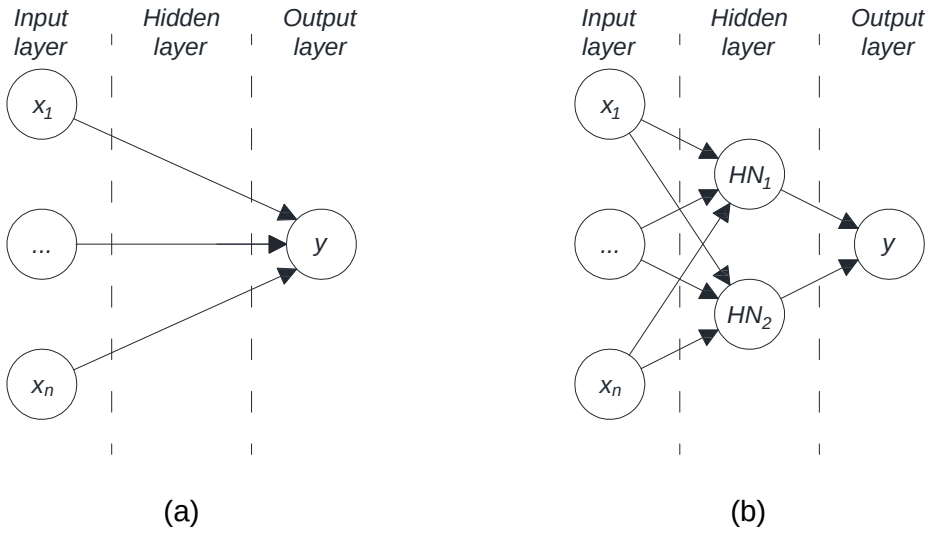


Figure 2 – Example of ANN: (a) without hidden layers; (b) with one hidden layer.

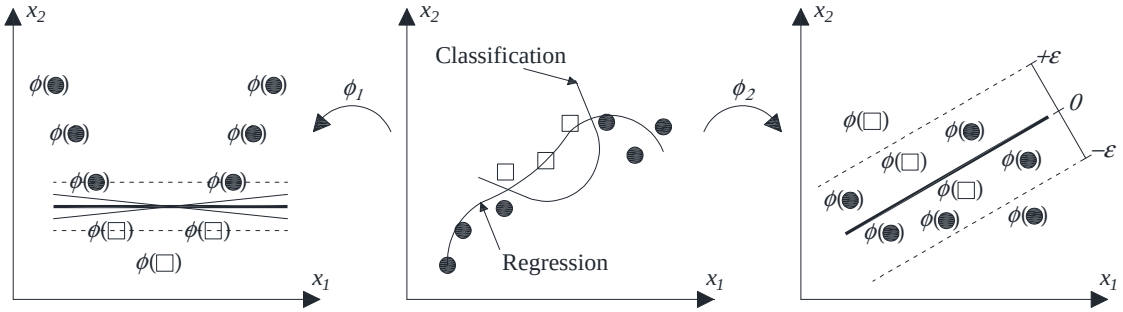


Figure 3 – Example of SVM classification (left) and regression (right) of non-linear data (middle).

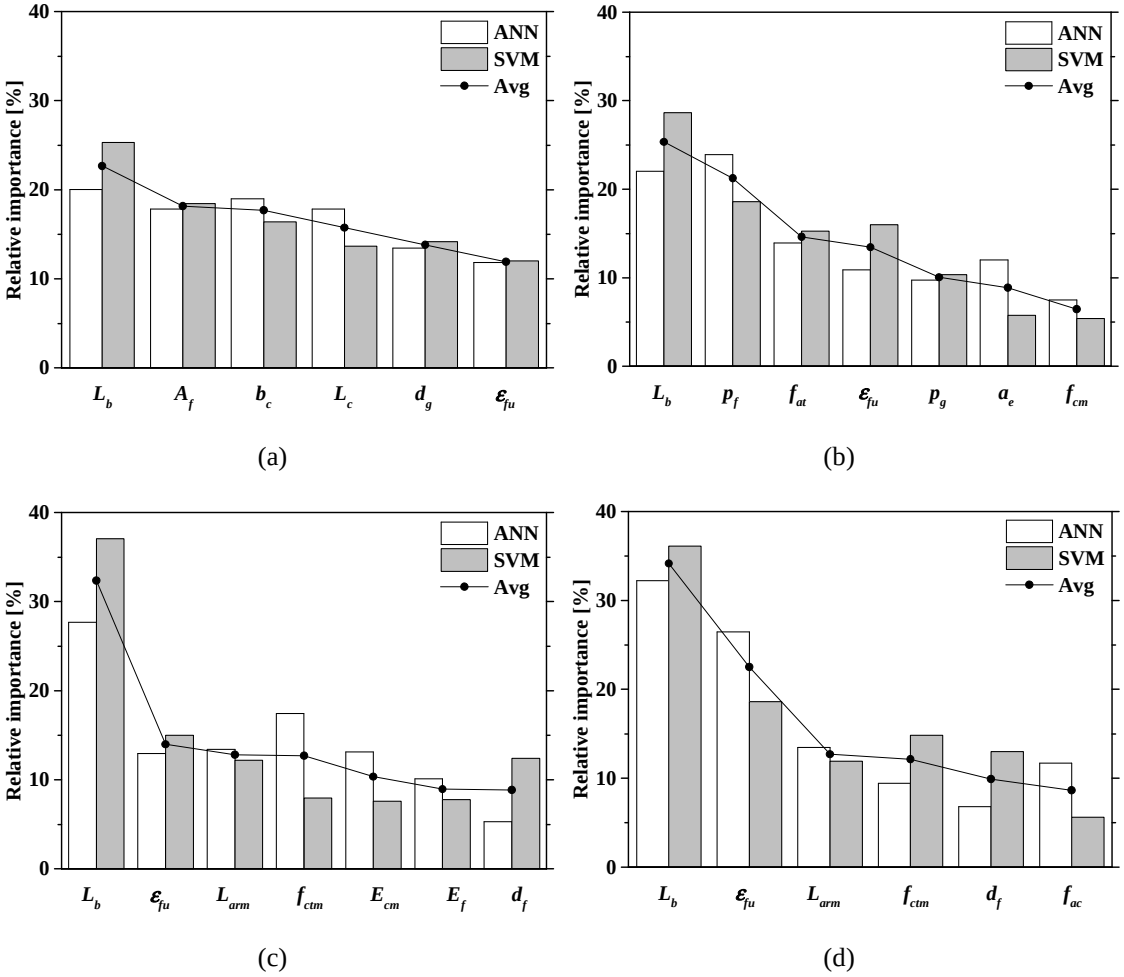


Figure 4 – Relative importance of each input variable in the analyses Type B (*database* and *input variables*): (a) *DPT* and *RM*; (b) *DPT* and *User*; (c) *BPT* and *RM*; (d) *BPT* and *User*.

Coelho, M.R.F.; Sena-Cruz, J.; Neves, L.A.C.; Pereira, M.; Cortez, P.; Miranda, T. (2016) "Using data mining algorithms to predict the bond strength of NSM FRP systems in concrete." Construction and Building Materials, 126: 484–495.

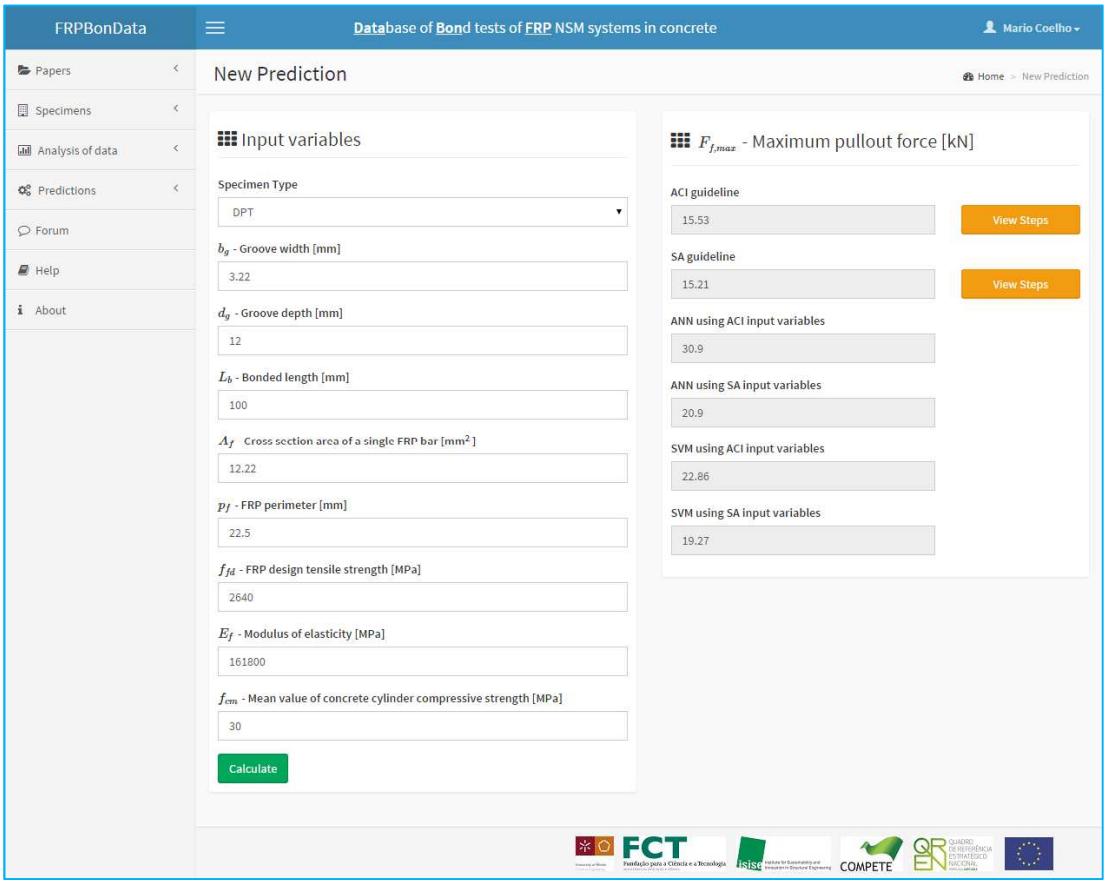


Figure 5 – Maximum pullout force prediction calculated in the website developed.

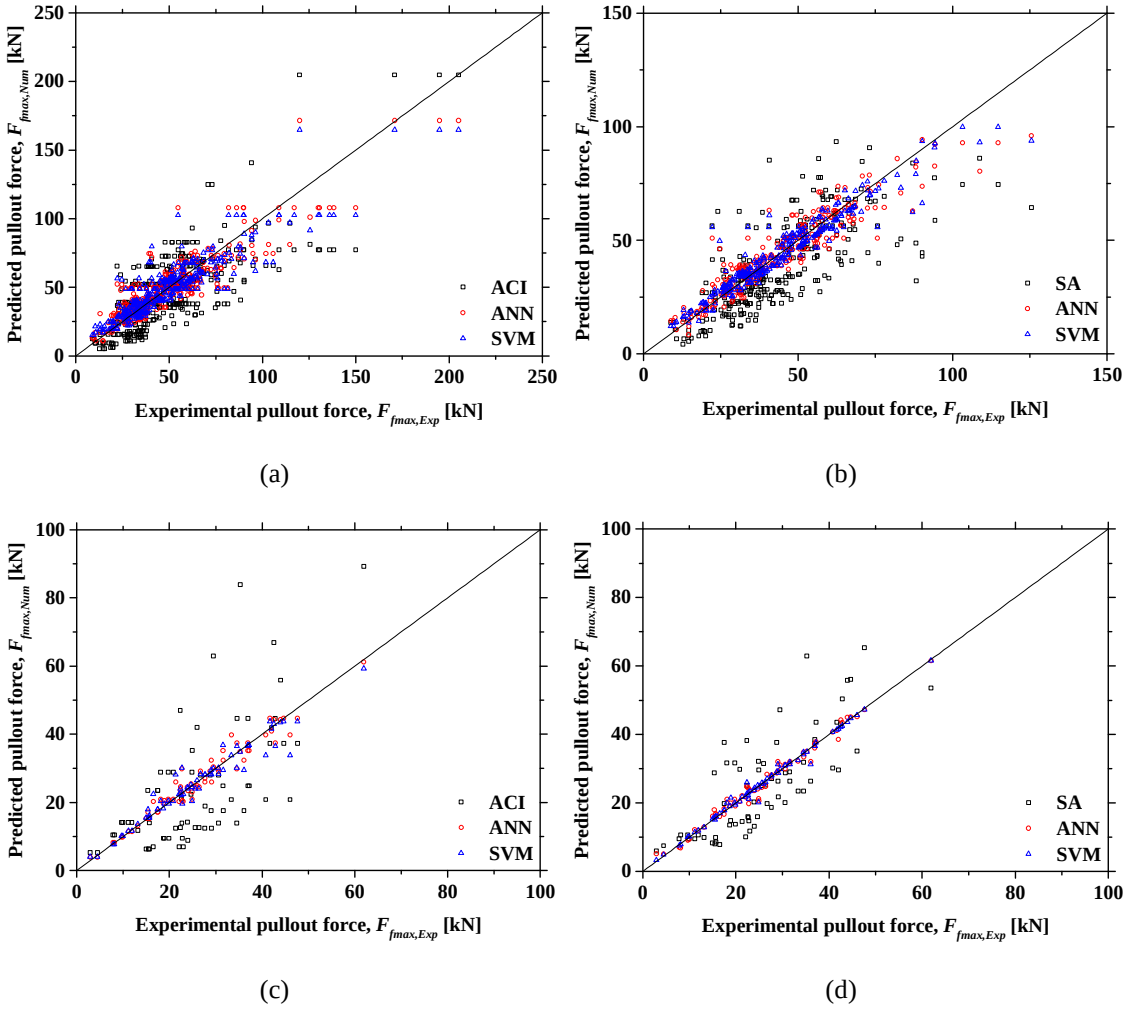


Figure 6 – Experimental *versus* predicted pullout force for the final models obtained by fitting DM algorithms (*database* and *input variables*): (a) *DPT* and *ACI*; (b) *DPT* and *SA*; (c) *BPT* and *ACI*; (d) *BPT* and *SA*.

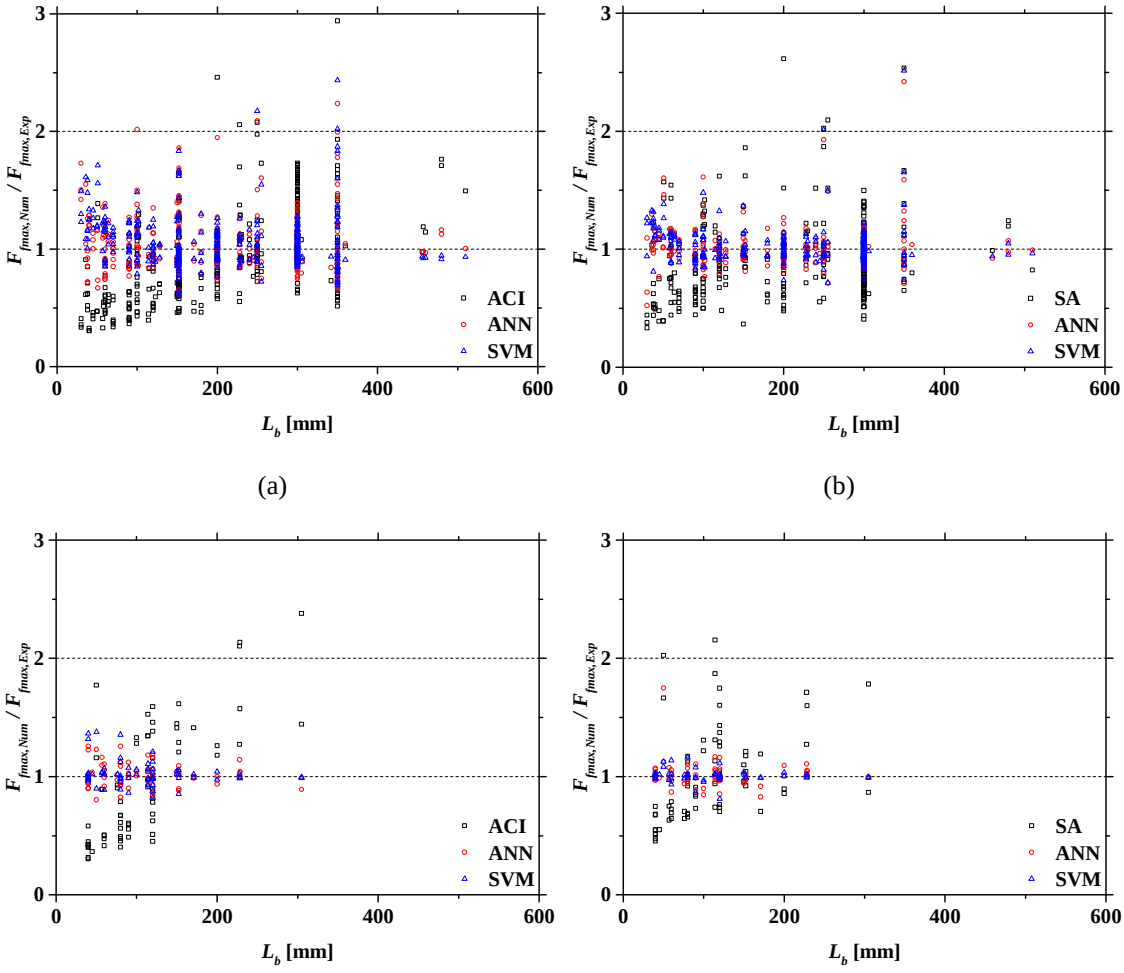
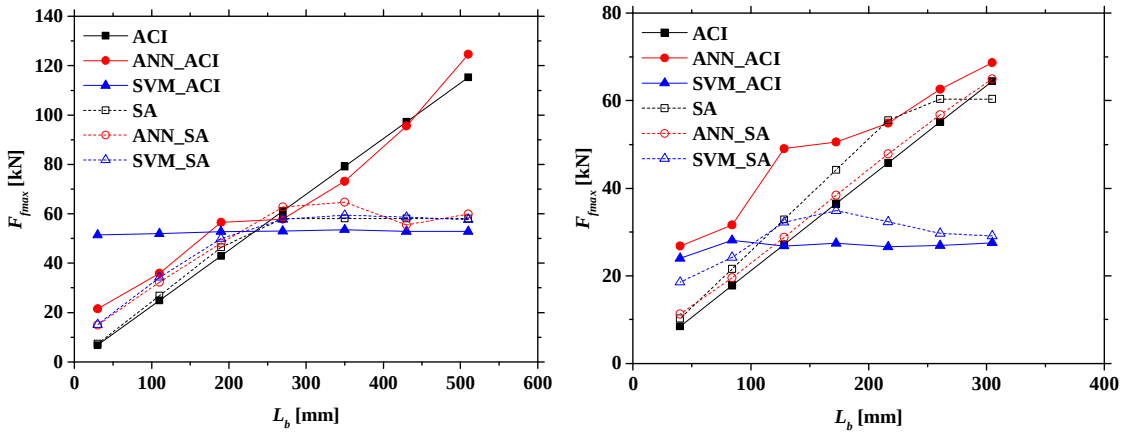


Figure 7 – Variation of the ratio between experimental and predicted pullout force with the bonded length (*database* and *input variables*): (a) *DPT* and *ACI*; (b) *DPT* and *SA*; (c) *BPT* and *ACI*; (d) *BPT* and *SA*.



(a) (b)

Figure 8 – VEC curves for L_b considering (a) *DPT* or (b) *BPT* databases. Note: composite designations include the DM model and the type of input variables, as defined in Table 6.

Notation

The following acronyms /symbols are used in this paper:

Acronyms

ACI	American Concrete Institute guideline
ANN	Artificial Neural Network
BPT	beam pullout tests
DM	Data mining
DPT	direct pullout tests
FRP	fiber reinforced polymer
NSM	near-surface mounted technique
SA	Standards Australia guideline
SVM	Support Vector Machine

Symbols

a_e	Distance from FRP to closest concrete block edge
A_f	FRP cross-section area
b_c	Concrete block width
b_g	Groove width
d_f	FRP width or diameter in quadrangular or round bars, respectively
d_g	Groove depth
Ec	Concrete modulus of elasticity
E_f	FRP modulus of elasticity
ε_{fu}	FRP ultimate strain
f_{ac}	Adhesive compressive strength
f_{at}	Adhesive tensile strength
f_c	Concrete design compression strength
f_{cm}	Concrete cylinder mean compressive strength
f_{ct}	Concrete tensile strength
f_{fd}	FRP design tensile strength
f_{fu}	FRP ultimate tensile strength
F_{fmax}	Maximum pullout force
L_{arm}	Vertical distance from the centroid of the center hinge to FRP centroid
L_b	Bonded length
L_c	Concrete block length
p_f	FRP perimeter
p_g	Groove perimeter