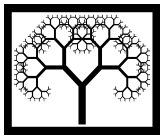


©Civil-Comp Press, 2009
Proceedings of the First International Conference on
Soft Computing Technology in
Civil, Structural and Environmental Engineering,
B.H.V. Topping and Y. Tsompanakis, (Editors),
Civil-Comp Press, Stirlingshire, Scotland

Paper 46

Data Mining Techniques and Ultrasonic Pulse Velocity Tests for the Assessment of Damage Levels in Concrete exposed to High Temperatures and subject to Compression

R. Marques, L. Lourenço and J. Barros
Institute for Sustainability and Innovation in Structural Engineering
Department of Civil Engineering, University of Minho, Portugal

Abstract

A Data Mining (DM) process was applied, aiming to develop a numerical tool for the assessment of the level of damage (measured from a strain parameter) of concrete columns subject to high temperature exposure (or fire). The database used was obtained from an experimental program carried out with measurements of Ultrasonic Pulse Velocity (UPV) through concrete specimens of various steel reinforcement arrangements after exposure to different levels of high temperature (reference, 250 ºC, 500 ºC and 750 ºC). The UPV tests were performed during the execution of the compression tests. Different DM techniques were tested and a sensitivity analysis was made. The obtained results show that DM models, particularly the k-nearest neighbours model, allow an acceptable prediction of the axial strain of concrete elements exposed to high temperatures using UPV data.

Keywords: Data Mining, k-nearest neighbours, experimental database, ultrasonic pulse velocity, concrete, high temperatures.

1 Introduction

The evaluation of the structural stability of a concrete structure that was subject to fire, as well as the consequent decision that should be taken about its upgrading or demolishing are tasks of high complexity and responsibility that require the use of appropriate equipments of inspection, and the application of sophisticated numerical models [1]. Due to the scarcity of this type of models, as well as the material constitutive laws able of simulating, with enough accuracy, the residual strength of concrete structures submitted to fire, research in this domain is quite necessary and relevant.

The inspection and the assessment of the residual strength of structural concrete elements affected by high temperature exposure are usually executed by the use of destructive tests (extraction of core cylinders) to evaluate modulus of elasticity and

compressive strength. However, this procedure only provides a local assessment, which can lead to unsafe predictions of the structural stability of the damaged building. Furthermore, destructive tests introduce extra damages in the structure and are, in certain cases, too time consuming. Therefore, the use of non-destructive tests for the assessment of concrete elements presents some benefits such as: low cost of inspection; ability to re-test; fast procedure; high number of measurements without introducing damages in the structure [2].

In the last years several studies were performed to predict the concrete strength by using artificial intelligence techniques, particularly the modelling by neural networks [3-7]. In the case of concrete elements affected by high temperature exposure, the temperature level of exposure plays an important role. In Chen et al. [7], the study was based in support vector machines technique, which is considered as a neural networks technique improvement. In fact, although the artificial neural networks are one of the most versatile (by analogy with the human brain), and widely used, the evolution of the branch of computation sciences conducted to the appearance of other techniques based on natural processes, which must be considered by engineers as an alternative to neural networks. Examples of these techniques are nonlinear methods of support vector machines, as well as nonparametric methods of decision trees and k-nearest neighbours.

The experimental procedures in Civil Engineering are favourable to the creation of large data repositories. Data Mining (DM) is a privileged process to extract high-level knowledge from raw data [8]. Large data with several variables and the development of methods capable to find invisible relations for traditional methods are the background to the diffusion of DM techniques, as a strong branch of the Artificial Intelligence field.

This paper presents the application of DM techniques as a tool to assess the level of damage in concrete elements subject to high temperature exposure. For this purpose, an experimental database of UPV measurements was used.

2 Data Mining

2.1 Techniques

In this study several DM techniques were tested, from the traditional technique of multiple regression and the non-parametric methods of decision trees and k-nearest neighbours to the nonlinear techniques based on neural networks and support vector machines.

The statistical technique of Multiple Regression (MR) appears in DM studies primarily as a baseline of comparison for nonlinear techniques. This technique represents a generalization of the linear regression for a model with various independent variables (called inputs).

The decision trees technique [9] is developed through the creation and training of subsets of information for which is inferred one or more rules. According to a tree structure, each node establishes a test based on attributes, and each descending branch of this node is one of the possible values for that attribute. These trees are

called *Regression Trees (RTs)* when they perform the prediction for a continuous variable.

The *k-Nearest Neighbours (k-NN)* method bases its predictions on the location of the k observations that are closest to the item being predicted (Figure 1). The determination of this similarity is normally based on distance measures. In this study, a formulation based on the transformation of distances in weights proposed by Hechenbichler and Schliep [10] was used. The regression task of this method is supported on the main following steps:

---

1. Let $L = \{(y_i, x_i), i = 1, \ldots, n_L\}$ be a learning set of observations x_i with given output y_i and let x be a new observation, whose output value y has to be predicted.
2. Find the $k + 1$ nearest neighbors to x according to a distance function $d(x, x_i)$.
3. The $(k + 1)$th neighbor is used for standardization of the k smallest distances via
$$D_{(i)} = D(x, x_{(i)}) = \frac{d(x, x_{(i)})}{d(x, x_{(k+1)})}. \quad (1)$$
4. Transform the normalized distances $D_{(i)}$ with any kernel function $K(.)$ into weights $w_{(i)} = K(D_{(i)})$. In this work, the triangular kernel was used.
5. As prediction for the output y of observation x compute value, which represents a weighted majority of the k nearest neighbors
$$\hat{y} = \max_r \left(\sum_{i=1}^{k} w_{(i)} I(y_{(i)} = r) \right) \quad (2)$$
where I is a window width function [10].

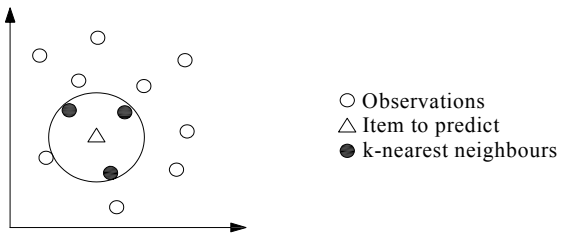


Figure 1: Illustration of the basic principle of the k-nearest neighbours model.

The *Artificial Neural Networks (ANNs)* are motivated by the performance of human brain. These networks consist of processing units (nodes) interconnected according a given configuration, where the *Multilayer Perceptron* is the most popular type [11]. The nodes are constituted by: a *set of connections* (w_{ij}), each labelled by a weight, which has an excitatory effect for positive values and inhibitory for negative ones; an *integrator* (g), which reduces the n entry arguments (stimulus) to a single value; and by a *activation function* (f) that can affect the output signal by introducing a component of non-linearity in the computational process. In

this study, the weights of the network are initially generated randomly in the range [-0.7, +0.7] and the activation function used is the *logistic* (1/(1+exp(-*x*))). Then, the training algorithm is applied by adjusting successively the weights until the error slope approaches zero or until a maximum of epochs. The prediction model for a neural network is given by the sum of all activated connections.

Support Vector Machines (SVMs) [12] are learning systems that use a space of hypothesis of linear functions in a wide space of features, which are trained with an optimisation algorithm that implements a statistical trend of learning. The basic idea is transform the input $x \in \Re^I$ into a high *m*-dimensional feature space by using a nonlinear mapping. Then, the *SVM* finds the best linear separating hyperplane, related to a set of support vector points, in the feature space. The transformation depends on the *kernel function* ($k(x, y) = \sum \phi_i(x)\ \phi_i(y)$) adopted. In this work, the popular *gaussian* kernel was used, which presents less hyperparameters and numerical difficulties than other kernels, as the polynomial or sigmoid [13].

2.2 Process

2.2.1 Database

The raw database used in this study was collected from an experimental program of non-destructive tests that was carried out measuring UPV through concrete column specimens of various steel reinforcement arrangements (Figure 2) and exposed to several levels of high temperature (reference, 250 ºC, 500 ºC and 750 ºC). In Figures 3 and 4 for each axial strain the corresponding UPV and vertical load is plotted for the specimens subject to the distinct levels of temperature exposure considered. Similar trend was observed in the other reinforced specimens [14]. These Figures evince UPV measures are capable of capturing the favourable effect of the reinforcement arrangements in terms of compressive residual strength.

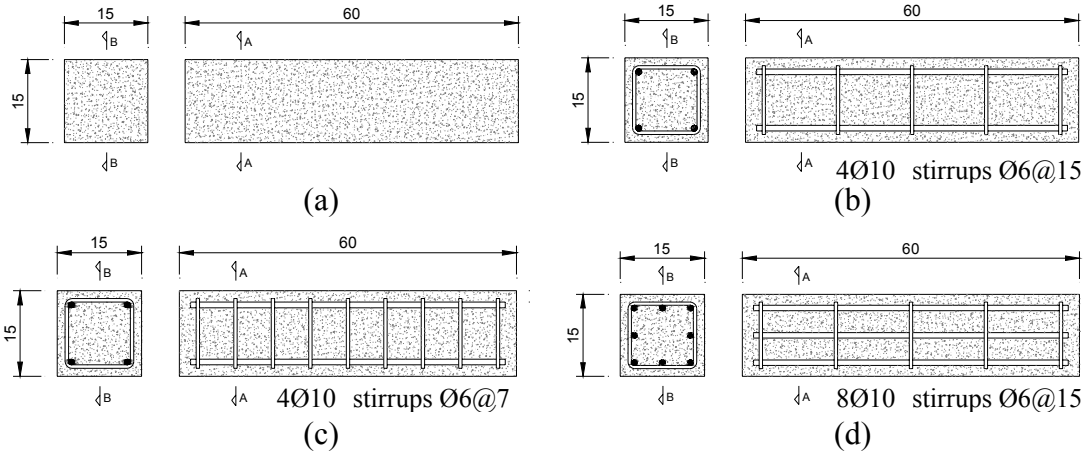


Figure 2: Concrete column specimens: (a) unreinforced; (b) reference reinforcement arrangement; (c) over-reinforcement in terms of steel hoops; (d) over-reinforcement in terms of longitudinal bars.

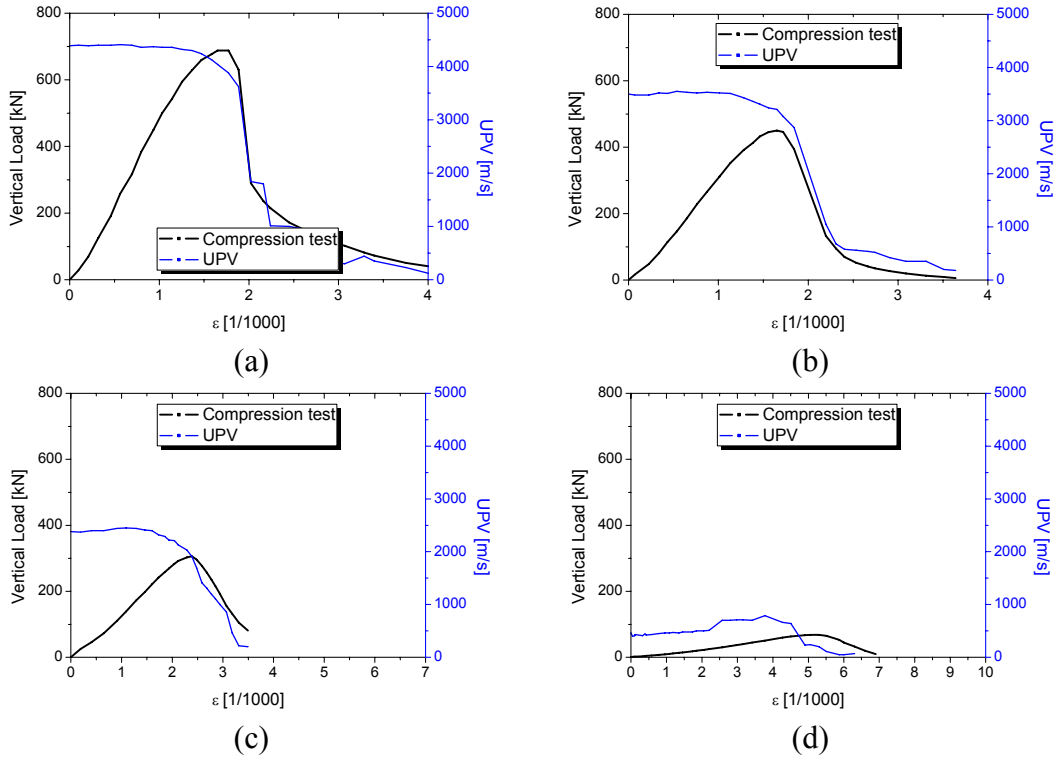


Figure 3: Vertical load and UPV versus axial strain (unreinforced): (a) reference; (b) 250 ºC; (c) 500 ºC; (d) 750 ºC.

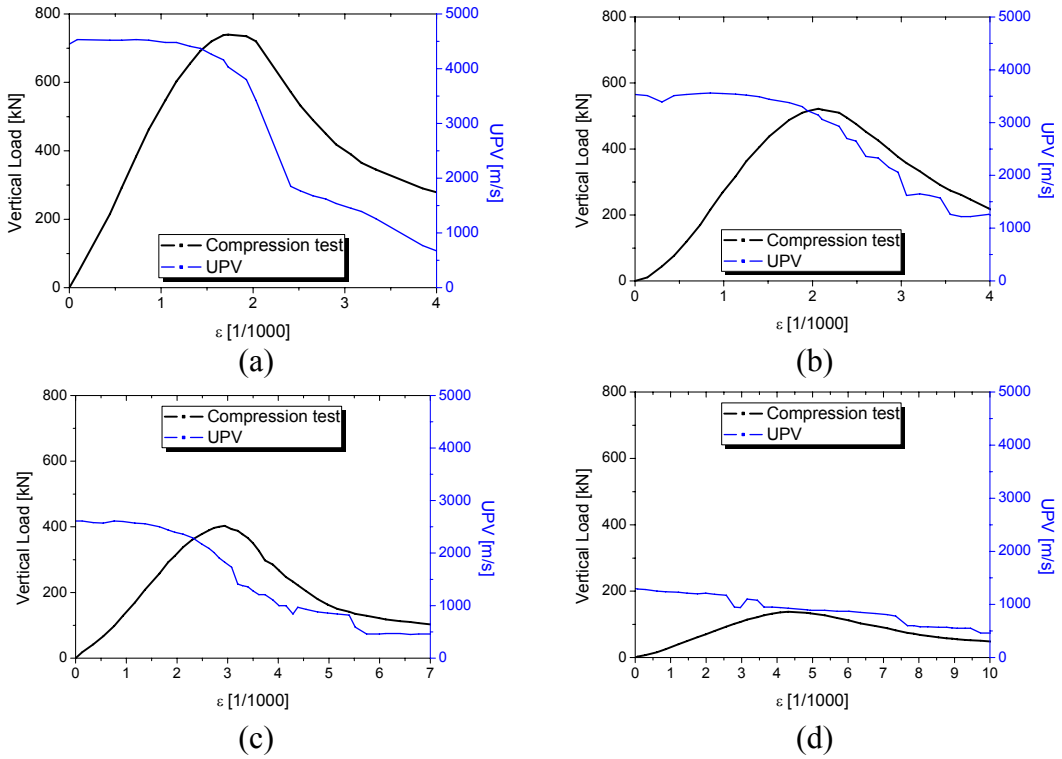


Figure 4: Vertical load and UPV versus axial strain (reference reinforcement arrangement): (a) reference; (b) 250 ºC; (c) 500 ºC; (d) 750 ºC.

The treated database, with 665 records, is composed of the following variables; UPV; level of temperature exposure (T); type of steel reinforcement arrangement (Reinforcement) and the relative axial strain, ε_r, of concrete column specimens. In Figure 5 a graphical representation of the database is shown, from which the complexity of the problem is suggested. Figure 5 shows a graphic matrix that represents, through points, associations of values for the quartet of variables UPV-T-Reinforcement-ε_r. The variable ε_r represents the ratio between the axial strain and the strain at peak compressive strength, so it assumes the unit value at the maximum compressive strength (see Figure 6).

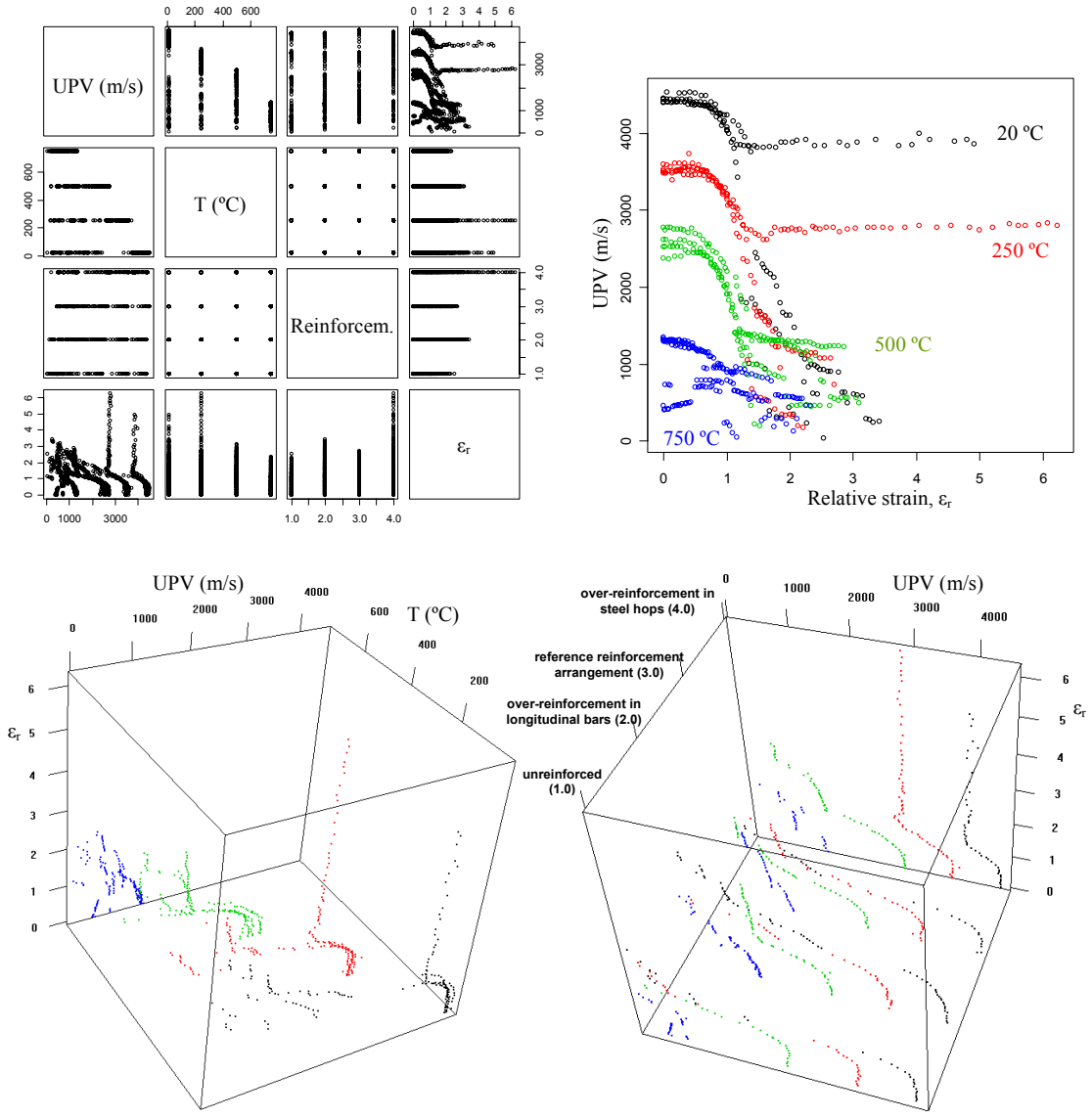
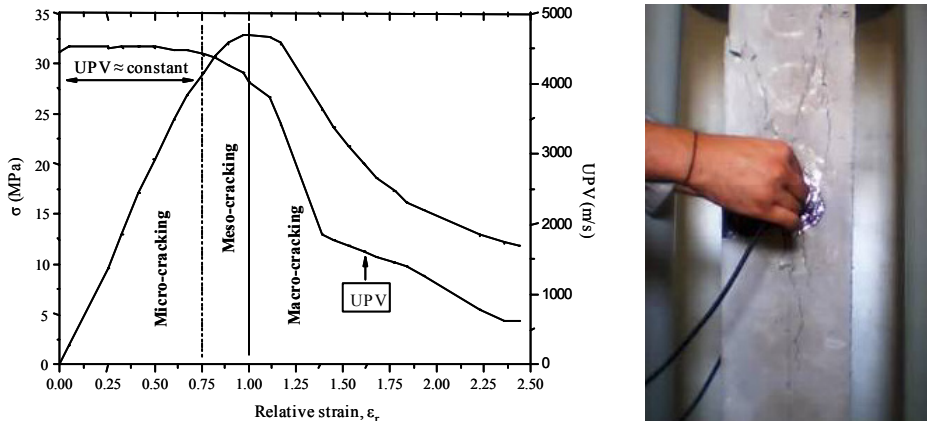


Figure 5: 2D and 3D representations of the database.

From a previous data analysis and physical interpretation of the problem, the goal for this investigation was the assessment of concrete elements after exposure to high temperatures, based on the adjustment of a model by DM techniques to predict the relative axial strain (ε_r) as function of UPV, T and Reinforcement. In this scenario, the ε_r predicted value can be considered as an assessment tool for the evaluation of concrete columns (Figure 6b). According to results shown in Figure 6a, if $\varepsilon_r \leq 0{,}75$ the damage level is composed by the formation of micro-cracks. For $0{,}75 < \varepsilon_r \leq 1$ the damage intensity is constituted by the presence of meso-cracks (of crack width in-between micro- and macro-cracks); if $\varepsilon_r > 1$ the concrete is in its softening phase, with the presence of macro-cracks that are the result of the coalescence of the meso- into cracks of large crack with.



(a) (b)

Figure 6: (a) Generic representation of compressive stress (σ) and UPV versus ε_r; (b) Apparatus to predict the pseudo-residual strength state of concrete elements.

2.2.2 Computation

The DM process was carried out in the environment of the statistical tool *R* [15]. The *RMiner* library, created by Cortez [13], was used because it contains a coherent set of functions that facilitates the application of different DM techniques, using packages available in the *R* program.

In this work, the regression model hypothesis used in the DM process was:

$$\varepsilon_r \sim UPV \text{ “+” } T \text{ “+” } Reinforcement \quad (3)$$

where the ε_r value (i.e., the dependent variable) appear as a function (~) of the independent variables (i.e., the right side of the formula).

The predictive capacities of several DM algorithms were compared according to the above regression model, through distinct measures of performance. In this work, the evaluation scheme is based on 10 executions (runs) of a 5-fold cross-validation [16], where the data is divided into 5 partitions of equal size. Sequentially, each different subset is tested and the remaining data is used for fitting the model. The overall performance is given by the average of error metrics on all 10 executions, and their confidence intervals under a *t-student* test with a 95% confidence level.

The first measure of performance is based on the concept of the *REC* curve, *Regression Error Characteristic* [17], which is the cartesian representation of the tolerance error (horizontal axis), measured in terms of absolute deviation, versus the percentage of points predicted within that tolerance (vertical axis). The ideal regressor has a REC area of 1.

The comparison of the REC curves in Figure 7a shows that the technique based on k-nearest neighbors (k-NN) demonstrates the best performance in the prediction of the relative axial strain, followed by the non linear techniques based on support vector machines (SVM) and neural networks (ANN).

To better evaluate the performance of the different DM techniques, scatter plots of the values predicted by the model versus the experimental values are also shown. Figures 5b-d show that the k-NN technique has the lowest scatter, which is very narrow in the range [0, 3] of relative strains.

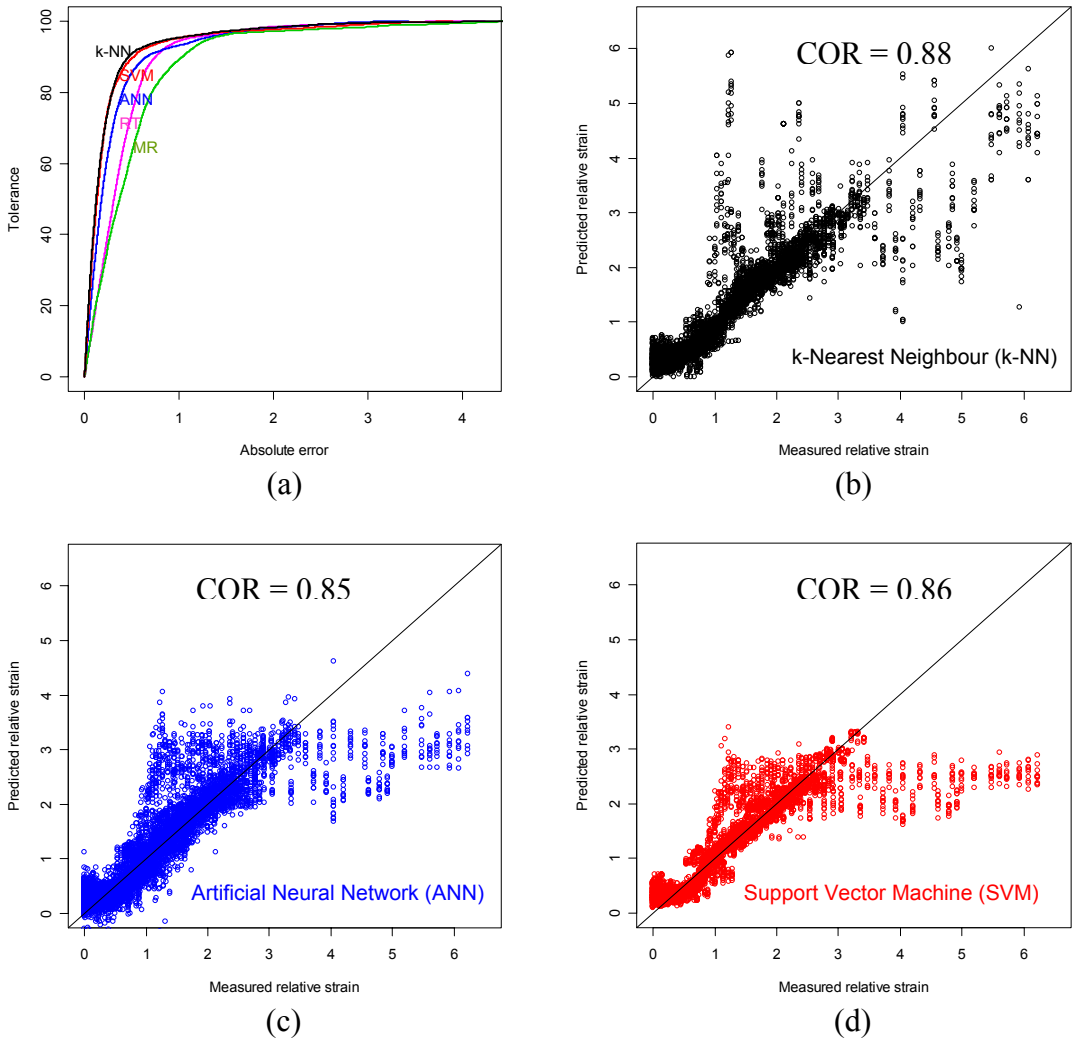


Figure 7: Predictive performance of different DM techniques to adjust ε_r model: (a) REC curves; (b–d) Predicted ε_r values versus experimental ε_r values.

The following metrics were computed to assess the performance of each technique: *Mean Absolute Deviation (MAD)*, *Relative Absolute Error (RAE)*, *Root Mean Squared (RMSE)*, *Relative Root Mean Squared (RRMSE),* and Pearson correlation coefficient (*COR*). These metrics, based on the error, are calculated according to the following equations [13]:

$$
\begin{aligned}
&MAD = 1/N \times \sum_{i=1}^{N} |y_i - \hat{y}_i| \\
&RAE = MAD / \sum_{i=1}^{N} |y_i - \bar{y}_i| / N \times 100\,(\%) \\
&RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N} \\
&RRMSE = RMSE / \sqrt{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2 / N} \times 100\,(\%) \\
&COR = \frac{\sum_{i=1}^{N} (y_i - \bar{y}_i)(\hat{y}_i - \bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2 \sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}}_i)^2}}
\end{aligned}
\tag{4}
$$

where N denotes the number of examples, y_i the recorded experimental value , $\hat{y}_i$ the predicted value, $\bar{y}_i$ the average of the recorded experimental values and $\bar{\hat{y}}_i$ the average of the predicted values.

Table 1 shows that the technique of k-nearest neighbors provided the lowest values for any of the error metrics in the prediction of ε_r, as well as a high COR.

	k-NN	ANN	SVM
MAD	0.244 ± 0.009	0.315 ± 0.006	0.262 ± 0.003
RAE (%)	31.65 ± 1.17	40.81 ± 0.79	33.96 ± 0.37
RMSE	0.494 ± 0.010	0.543 ± 0.007	0.538 ± 0.006
RRMSE (%)	48.26 ± 0.97	52.99 ± 0.67	52.51 ± 0.56
COR	0.878 ± 0.005	0.849 ± 0.004	0.856 ± 0.004

Table 1: Performance measures of different DM techniques in the modelling of ε_r.

2.2.3 Sensitivity analysis

The *Sensitivity Analysis* is a simple method that can be used to measure input relevance. It measures the variance (V_a) produced in the output $(\hat{y})$ when the input attribute (a) varies through its full range [18]:

$$
\begin{aligned}
&V_a = \sum_{i=1}^{L} (\hat{y}_i - \bar{\hat{y}}) / (L-1) \\
&R_a = V_a / \sum_{j=1}^{I} V_j \times 100\,(\%)
\end{aligned}
\tag{5}
$$

where I denotes the number of input attributes and R_a the relative importance of the a attribute. In the computation of V_a, the output $\hat{y}_i$ is obtained by holding all input variables at their average values, with exception of x_a, which varies through its entire range with L levels. In this work, it was assumed $L = 6$ for the continuous inputs.

Table 2 shows that for any of the DM models to predict the relative axial strain, the UPV, as expected, has the predominant effect on the value of this parameter.

The low values of the Reinforcement variable importance (see Table 2), particularly for the k-NN model, recommend to adjust ε_r only as function of UPV and T. However, if this procedure is followed, a lower reliability is obtained for the k-NN model, with a correlation coefficient of 0.836 ± 0.006. If only the input variable UPV is considered the correlation coefficient is 0.619 ± 0.012.

	Variable importance (%)		
	UPV	T	Reinforcement
k-NN	69.21	29.75	1.04
ANN	73.20	20.66	6.14
SVM	66.70	26.00	7.30

Table 2: Importance of variables in the ε_r prediction with different DM techniques.

3 Conclusions

The possibilities of using Data Mining (DM) techniques for the assessment of the compression behaviour of concrete columns after having been subject to temperature exposure were explored. For this purpose, the concept of relative strain, ε_r, (strain divided by the strain at peak stress) was selected as the key variable, and several DM techniques were used taking the ultrasonic pulse velocity (UPV), the exposed temperature (T) and the type of reinforcement arrangement (Reinforcement) as the known variables of the collected database.

The top-down hypothesis of predicting the ε_r as a function of the UPV, T and Reinforcement allowed a good reliability, particularly by using the k-nearest neighbour technique.

The highest capability of the k-nearest neighbour technique can be achieved from its capacity of clustering data, based on the T and Reinforcement variables. On the other hand, the most effective parameter influencing the accuracy of estimated relative strain of fire-damaged concrete elements is identified as the UPV in concrete.

Acknowledgements

The second author wishes to acknowledge the support provided by "CIVITEST – Pesquisa de Novos Materiais para a Engenharia Civil, Lda." and "António L. Rodrigues, Lda." Companies.

References

[1] G.A. Khoury, "Effect of fire on concrete and concrete structures", Prog. Struct. Engng Mater., 2, 429-447, 2000.

[2] B. Hobbs, M.T. Kebir, “Non-destructive testing techniques for the forensic engineering investigation of reinforced concrete buildings”, Forensic Science International, 167, 167-172, 2007.
[3] S. Lee, “Prediction of concrete strength using artificial neural networks”, Engineering Structures, 25, 849-857, 2003.
[4] J. Hoła, K. Schabowicz, “New technique of nondestructive assessment of concrete strength using artificial intelligence”, NDT&E International, 38, 251-259, 2005.
[5] M.A. Kewalramani, R. Gupta, “Concrete compressive strength prediction using ultrasonic pulse velocity through artificial neural networks”, Automation in Construction, 15, 374-379, 2006.
[6] G. Trtnik, F. Kavčič, G. Turk, “Prediction of concrete strength using ultrasonic pulse velocity and artificial neural networks”, Ultrasonics, 49, 53-60, 2009.
[7] B. Chen, T. Chang, J. Shih, J. Wang, “Estimation of exposed temperature for fire-damaged concrete using support vector machine”, Computational Materials Science, 44, 913-920, 2009.
[8] I.H. Witten, E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques”, 2nd Edition, Morgan Kaufmann, San Francisco, USA, 2005.
[9] J. Quinlan, “Induction of Decision Trees”, Machine Learning, 1, 81-106, 1986.
[10] K. Hechenbichler, K. Schliep, “Weighted k-Nearest-Neighbor Techniques and Ordinal Classification”. SFB 386, Paper 399, Ludwig-Maximilians University Munich, 2004.
[11] S. Haykin, “Neural Networks: A Comprehensive Foundation”, 2nd Edition. Prentice-Hall, New Jersey, 1999.
[12] C. Cortes, V. Vapnik, “Support Vector Networks”, Machine Learning, 20(3), 273-297, 1995.
[13] P. Cortez, “RMiner: Data Mining with Neural Networks and Support Vector Machines using R”, in “Introduction to Advanced Scientific Softwares and Toolboxes”, R. Rajesh, (Editor), IAEng, in press, 2009.
[14] L. Lourenço, J. Barros, P. Marques, R. Marques, “Ultrasonic pulse velocity in concrete specimens exposed to high temperatures”, Portuguese National Meeting “Betão Estrutural 2008”, Guimarães, CD-ROM, 2008 (in Portuguese).
[15] R Development Core Team, “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL: http://www.R-project.org, 2009.
[16] B. Efron, R. Tibshirani, “An Introduction to the Bootstrap”, Chapman & Hall, New York, USA, 1993.
[17] J. Bi, K. Bennett, “Regression Error Characteristic curves”, in “Proceedings 20th Int. Conf. on Machine Learning”, Washington, USA, 2003.
[18] R. Kewley, M. Embrechts, C. Breneman, “Data strip mining for the virtual design of pharmaceuticals with neural networks”, IEEE Transactions on Neural Networks, 11(3), 668-679, 2000.