Data-driven modelling of the structural behavior of concrete dams using machine learning techniques – Extended abstract

João Francisco Henriques de Sena Cardoso[a], Juan Tomé Caires da Mata[b], João Carlos de Oliveira Fernandes de Almeida[a]

[a]*Instituto Superior Técnico, Portugal, joao.sena.cardoso@tecnico.ulisboa.pt, jalmeida@civil.ist.utl.pt*
[b]*Laboratório Nacional de Engenharia Civil, Portugal, jmata@lnec.pt*

Abstract The scope of this dissertation entails the structural health monitoring of concrete dams and different machine learning formulations, to support the analysis and interpretation of the observed behavior in concrete dams. In this dissertation, horizontal displacements, spanning 30 years, were analyzed, and interpreted at two points in the central block of Alto Lindoso dam. At each point, three formulations denominated Hydrostatic-Season-Time (HST), Hydrostatic-Temperature-Time (HTT) and Hydrostatic-Temperature-Foundation-Time (HTFT) were tested, the latter being an innovative proposal within the scope of this work. All formulations allow the consideration of the effects of the reservoir water level, temperature, and time. The differences between them lies in: i) HST considers the effect of air temperature by an annual period sinusoidal function; ii) HTT considers the effect of air temperature through the temperature measured in the dam body; and iii) HTFT considers the effect of air temperature through the temperature measured in the dam body and considers the effect of the deformation of the foundation. The following machine learning methods were used in the formulations: multiple linear regression, artificial neural networks, and random forest. The analysis considered 286 observation campaigns categorized according to learning period and prediction period. All the formulations showed a good estimation capacity, with some potential advantages in the use of the HTFT formulation. The machine learning methods proved to be suitable for this type of application, with the random forest method presenting more beneficial insights results when compared to the other methods.

***Keywords**:* Structural Health Monitoring of Dams; Analysis and Interpretation of Observed Behavior; Machine Learning; Artificial Neural Networks; Random Forest; Quantitative Interpretation.

1. Introduction

Structural health monitoring of dams consists on activities and procedures that provide information and data necessary for the assessment of the structural safety of dams. The structural safety control of a dam is a relevant subject since a possible failure or incident may cause significant damage, including human casualties, economical or environmental ones. In order to guarantee structural dam safety and to detect changes in the structural behavior, and ultimately prevent failures and incidents, the structural behavior is measured through a monitoring system. The structural behavior is then compared to predictive models that represent the normal dam behavior. These models can be either statistical, deterministic, hybrid or mixed [1]. Statistical models prevail for the life cycle of dams after construction due to their simplicity of formulation, easiness of correlation between governing and dependent variables and allowing for the separation of effects. The main disadvantage is the fact that these correlations do not have a physical meaning and therefore are not able to return a general rule for all structures alike.

Another factor that is changing the approach to the statistical models of dams is the application of machine learning techniques to compute the correlations between the governing and the dependent variables. Several of these techniques have been applied to the monitoring of dams [2] with promising results.

In dam related engineering activities, the traditional approach is based on the relationship between the dependent variable and the main loads through a linear expression which incorporates terms related to the hydrostatic load, the thermal load and the time effects, where the coefficients are obtained by the multiple linear regression method. The most common way to depict the thermal load is to use an annual periodical sinusoidal-type function which provides a model called the Hydrostatic-Season-Time (HST) [3].

This dissertation follows the work of Mata et al [4] [5] and provides a comparison between the traditional model (HST) and approach via a case study considering a Hydrostatic-Temperature-Time model (HTT) and a Hydrostatic-Temperature-Foundation-Time model (HTFT) which is an innovative new type of model. These two models consider the thermal load through the temperature measured in the dam body and, for the latter case the foundation deformation is also considered.

Several machine learning techniques to compute the models are also applied, specifically, the multiple linear regression method, the artificial neural networks method and the random forest method with the help of the R programming language and libraries [6].

The case study covers the application of these models and methods to two points in the central block of the Alto Lindoso dam to model the horizontal displacement along the radial direction.

2. Dam Classification

Dams have existed since ancient times and the first records of this kind of structure report to the Mesopotamia civilization [7]. Their main use was to control the water flow in rivers, to provide access to water for agriculture or consumption and/or to prevent flooding. In the Iberian Peninsula there are still some Roman examples of these structures [8]. Some of the current designs for dams are still based on these ancient examples. The Romans used several types of materials for their infrastructures and designs depending on site conditions and material availability.

Generally, dams can be categorized as per three major criteria: i) to the type of use of the dam such as, irrigation purposes, flood control, human consumption among others; ii) classification according to material and design, mainly used in engineering, where the two main types are embankment dams and concrete dams; and iii) in a risk assessment base which is used for legislative purposes.

2.1. Embankment Dams

Embankment dams provide structural stability mainly due to their self-weight and have a trapezoidal shape where the faces are relatively steep. These structures can be subdivided into the following types according to their construction and seepage resisting mechanisms: homogeneous - Figure 1 - a); zoned - Figure 1 - b), c), d); impermeable face lining - Figure 1 - e); and internal impermeable wall - Figure 1 - f). Different materials, such as rock, soil, sand and clay, may be used while the foundation must be impervious.

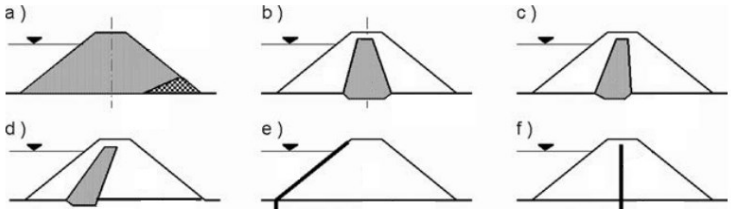

Figure 1 – Types of embankment dams (adapted from [9])

Homogeneous embankment dams are built through the compression of consecutive layers of the same type of soil which provide a low hydraulic conductivity barrier so that the seepage through the dam is rather low. Normally they have a filter zone either on the upstream side or on the downstream side.

If the mostly available material is not impervious enough to ensure a very low flow within the dam body, it is necessary to provide an impermeable zone or lining to prevent seepage. In the first case the materials are selected and placed to ensure that there is an impervious zone within the dam, while in the latter case, the majority of the material only provides the stability necessary whereas the lining made of an artificial material provides the flow control.

The most common causes of failure or incidents are: i) internal erosion, due to seepage; ii) overtopping, due to extreme events and/or lack of spillway capacity; iii) slope instability, due to poor quality control during construction and iv) operational and assessment errors.

2.2. Concrete Dams

Concrete dams can take different shapes, since this type of dam is able to resist the loads in two main ways. It can act like an embankment dam where the friction between the self-weight and the foundation opposes the loads and it is generally shaped as a triangle. These dams are classified as: gravity dams - Figure 2 -a), b); buttress dam - Figure 2- c) where a concrete wall is supported by a buttress along the spans; and arch dams - Figure 2 - d), where the loads are transmitted to the abutments and foundation by a concrete arch shaped wall. Since dams are designed mainly for horizontal loads, they normally display an effective structural behavior under seismic loads [10].

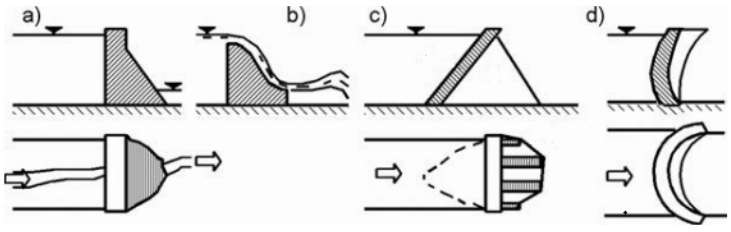

Figure 2 – Types of concrete dams (adapted from [9])

Gravity dams are solid structures of concrete that are able to deliver safe structural solutions due to their large mass. This type of dam displays good adaptability to different site and foundation conditions and resists well to the ageing phenomena of concrete [11].

Arch dams are rather slender structures compared to gravity dams and their shape provides a safe structural solution. Arch dams distribute the loads to the abutments and foundation by compression of the concrete wall, although tension may arise near the foundation on the upstream side [9].

In this type of dam, ageing phenomena of concrete and thermal loads are greater concerns than with gravity dams, since they have a much lower thermal inertia and the integrity of the concrete is fundamental to distribute the loads to the foundation.

3. Structural Health Monitoring of Dams

During the life cycle of a dam, ranging from construction, operational period to dam retirement, the dam is exposed to risks which may cause incidents or failure. These occurrences have to be controlled and monitored since they may have a detrimental effect. From recorded accidents, these effects may cause human casualties, and economical or environmental damages [12][13].

There were two major incidents in concrete dams: the failure of the Malpasset arch dam in France [14] and the overtopping of the Vajont arch dam in Italy [15], due to geotechnical problems and lack of monitoring. In Portugal, although there are no records of major incidents hitherto, there were cases in which was necessary intervention, namely the decommission of Alto Ceira dam, due to concrete ageing (alkali-aggregate reactions) [16]. Other interventions included the impermeabilization of the concrete surface with an impermeable PVC lining at the Pracana dam to reduce seepage through the dam body. The monitoring and analysis of these situations allowed for a supported decision.

The structural health monitoring of dams relies on the compromise between the dam monitoring through measurements and visual inspections, the analysis and interpretation of the structural behavior and the assessment of the dam condition in order to support the stakeholders decisions with and throughout time [17] (Figure 3).

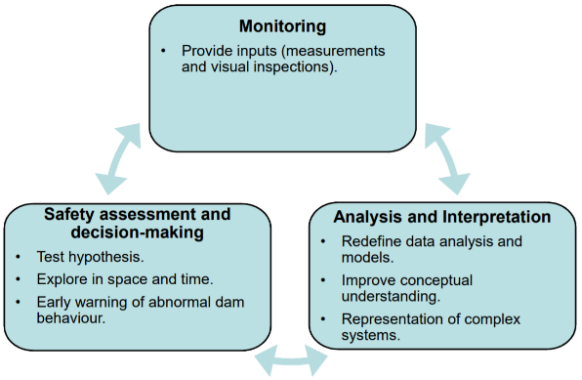

Figure 3 – Main activities of structural health monitoring of dams [16].

Generally, large concrete dams have instrumentation in several locations that record several physical quantities, such as: displacements, joint movements, concrete stresses, strains, temperature and cracking, air temperature, leakage, seepage, water level and temperature, precipitation, and dynamic loads (seismic load or swell) (Figure 4). The structural dam behavior is measured and recorded using devices that compose the monitoring system, being the records stored in a dam database [18]. With the data gathered, it is possible to create statistical models that try to estimate the structural behavior of the parameters in study [2].

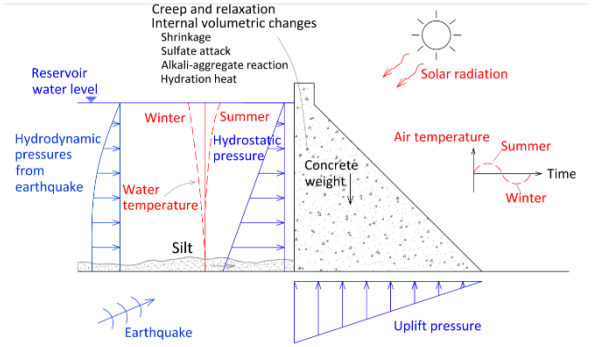

Figure 4 – Loads that affect the structural response of the dam [16].

The analysis of past behavior allows the interpretation of the local and global structural behavior of the dam. The new observed behavior is then compared with one of the models. These models are updated constantly as the devices gather more data on the current structural behavior of the concrete dam. If the observed behavior is different from the behavior predicted by the models, additional studies are performed in order to assess and to understand the main causes and to recalibrate the previous models.

3.1. The Hydrostatic-Season-Time Model

In the hydrostatic-season-time (HST) statistical model the effect on the structural parameter (U) is given in terms related to the hydrostatic effect of the water level in the reservoir, and it is generally defined by a 4th degree polynomial of the water level (h) as shown in equation (1). The thermal effect is usually given by an annual variation of the temperature which can be represented by sinusoidal functions with a one-year period, as shown in equation (2), where $\theta_1 = 2\pi t_d/365$ and t_d is the number of days since the 1st of January to the day of the observation. The time effect can take some different types of functions (some examples are in equation (3)), where t are the days since construction. There are two main variants: one that considers a horizontal asymptote and the second that does not. This model also considers a constant k [1].

$$U_h(h) = \beta_1 \times h + \beta_2 \times h^2 + \beta_3 \times h^3 + \beta_4 \times h^4 \quad (1)$$

$$U_{\theta_1}(\theta_1) = \beta_1 \times cos\,\theta_1 + \beta_2 \times sen\,\theta_1 + \beta_3 \times sen^2\,\theta_1 + \beta_4 \times cos\,\theta_1\,sen\,\theta_1 \quad (2)$$

$$U_t(t) = \begin{cases} \beta_1 \times e^{-ct}; \\ \beta_1 \times t + \beta_2 \times \left(ln\left(1 + \frac{t}{c_1}\right)\right)^{c_2} + \beta_3 \times \left(1 - e^{\frac{-t}{c_3}}\right); \\ \beta_1 \times log\,t + \beta_2 \times e^t; \\ \beta_1 \times t + \beta_2 \times t^2 + \beta_3 \times t^3. \end{cases} \quad (3)$$

3.2. The Hydrostatic-Temperature-Time Model

The hydrostatic-temperature-time (HTT) statistical model is similar to the HST model, except on terms related to the effect of the thermal load on the structural parameters. In this case, instead of a seasonal variation of the temperature, the observed measures of temperature in the dam body are considered (equation (4)), where T_j is the measured temperature in device j [5].

$$U_{T_j}(T_j) = \sum \beta_j \times T_j \quad (4)$$

3.3. The Hydrostatic-Temperature-Foundation-Time Model

The proposed hydrostatic-temperature-foundation-time (HTFT) statistical model uses the HTT model as a basis where a linear relation related to the foundation deformations is added to the model, as presented in equation (5), where EF_j is the measured deformation of the foundation in device j.

$$U_{EF_j}(EF_j) = \sum \beta_j \times EF_j \quad (5)$$

4. Machine Learning Methods

Machine learning is part of the artificial intelligence spectrum and studies the development and application of algorithms that can learn from data [19]. The goal regarding the use of machine learning methods in the statistical models is to find the best relationship between the effects observed in the structural parameter in analysis and the independent variables that represent the loads [20].

4.1. Multiple Linear Regression

The multiple linear regression (MLR) method is the traditional statistical method for a data-driven model approach for the structural interpretation of the behavior of the dam. This method has proven reliable across the years and is mathematically sound [20]. To find the coefficients β_j of each model, the equations are normally solved by the least squares method, where the model is fit to minimize the sum-of-squares of the differences between the observed and predicted values. In matrix notation, the coefficients are given by solving the equation, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{U}$, which provides a model in the form of $\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{\beta}}\boldsymbol{X}$, which then produces the residuals in vector form as $\hat{\boldsymbol{\varepsilon}} = \mathbf{U} - \widehat{\mathbf{U}}$ [5]. The matrix $\boldsymbol{X}^T\boldsymbol{X}$ must be invertible, which implies that $\boldsymbol{X}$ must have linearly independent columns, and the number of observations is greater than the number of the independent variables, in order for this method to provide a valid and computable solution.

4.2. Artificial Neural Networks

Artificial neural networks (NN) can be applied to different types of problems and can be applied to both supervised learning and unsupervised learning [22]. The unit basis for such a method is the artificial neuron (Figure 5), where the independent variables are the inputs, these inputs are weighted and summed at the core of the artificial neuron, then this weighted sum is transferred to the output via a transfer function, which is, normally, a step function or is sigmoidal shape.

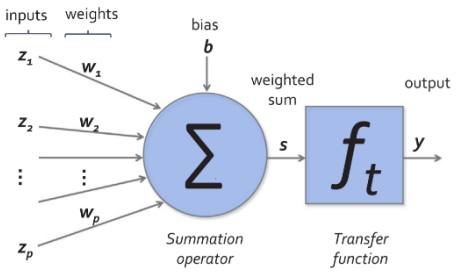

Figure 5 – Artificial neuron [23].

The multi-layer perceptron (MLP) is a technique within the artificial neural networks, that computes these kinds of neurons. The MLP architecture can have multiple hidden layers, multiple inputs, multiple neurons and multiple outputs in the system. A model created with this method begins with assigning random values to the weights according to the number of input values, number of neurons in the hidden layers and number of outputs [24]. After the first computation, the model calculates the value of the cost function, for example, $E = 0{,}5 \times \sum_{k=1}^{M}(\widehat{u_k} - u_k)^2$, where M is the number

of observations. The goal of the method is to minimize this function for all observations. To minimize this function the method used is the backpropagation of the error, where the weight is updated by calculating the partial derivative of the cost function regarding to the specific weight to be updated and a learning rate (a), in a similar way to the following equation: $* w_j = w_j - a(\partial E/\partial w_j)$. This process continues until a specific level of error and generalization are satisfied or a finite number of iterations is reached [25].

4.3. Random Forest

The Random Forest method (RF) is both used for regression or classification problems [26]. This method is one of the most used in machine learning since it is able to tackle different types of problems in a simple way with, generally, good results while also proving to be an accurate method for data control, making it an excellent starting point to analyze a set of data. However, it does not provide as good results as a specialized algorithm may provide [27]. The basis for this method is an unpruned regression tree. An unpruned regression tree is the successive division (branches) of a set of data (root node) into two (nodes), until the end nodes are reached. The end node is the final subset of data that must be homogeneous and independent [21]. The split at each node is determined by analyzing the variance at each node, given by, $V(t) = 1/M \times \sum_{k=1}^{M}(y_k - \overline{y_t})^2$, where the aim is to maximize the following expression, $V(t) - ((nt_e)/M \times V(t_e) + (nt_d)/M \times V(t_d))$, where t is the parent node, M is the total number of observations at the parent node, t_e is the left node, nt_e is the number of observations on the left node, t_d is the right node, nt_d number of observations on the right node, y_k is the observation and $\overline{y_t}$ is the average of the predictors at the node [28].

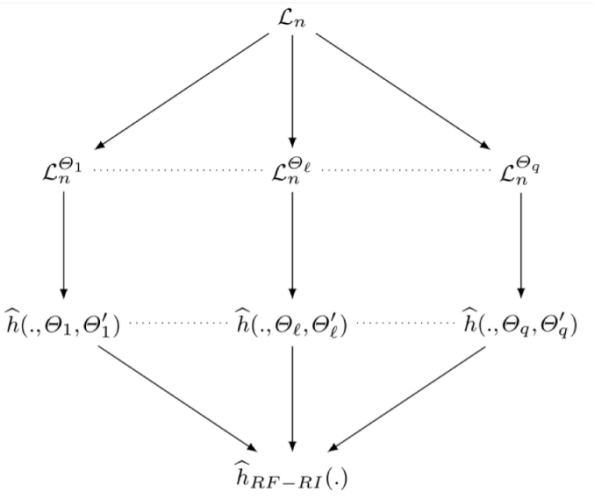

Figure 6 – Diagram of random forest with random inputs [29].

A model by the Random Forest method is a randomized set of multiple unpruned regression trees. The most used method of Random Forest is the Random Forest with random inputs (Figure 6) which gives another randomness effect that helps to obtain better results [29]. In this case the original set of data is subdivided into q random subsets where are applied to the unpruned regression tree method to split each node, the difference being that at each subset there is a n number of random independent variables to be analyzed. The biggest drawback of this method is the inability to go beyond the range of the learning data [27].

5. Case Study

The Alto Lindoso dam is a double curvature concrete arch dam on the Lima River in the north of Portugal and it was concluded in January 1992. It is 110 meters high, and the crest is at 339,0 meters height with a total length of 297 meters. At the central block of the dam (between joint 11 and joint 12) the thickness at the crest is 4 meters and at the foundation is 21 meters. There are 3 galleries inside the dam body and a drainage gallery along the foundation. The foundation consists of a heterogenous granite basin. The first reservoir filling started in January 1992 and reached peak capacity in April 1994 [16]. The dam has multiple instrumentation across its body and around it. This instrumentation has the ability to measure data regarding loads and aforementioned structural behavior.

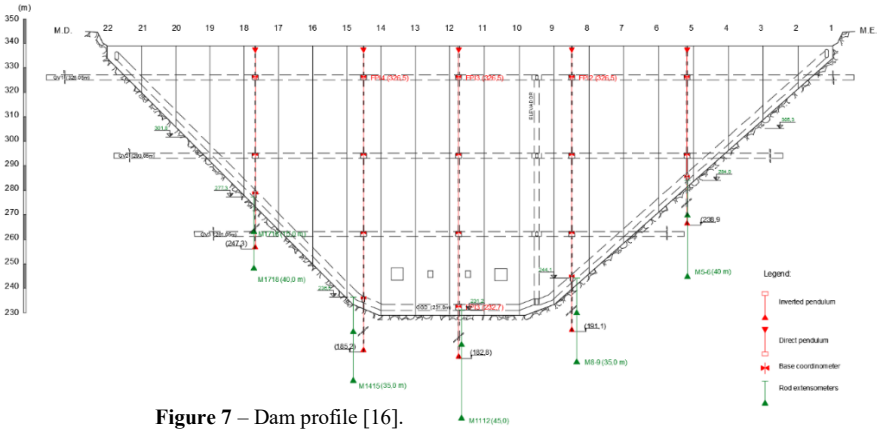

Figure 7 – Dam profile [16].

The dependent variables chosen to test the proposed models and methods were the horizontal displacements along the radial direction at two locations of the central block. These displacements are measured in the FP3 pendulum, at the heights of 326,5 meters and 232,7 meters (Figure 7 and Figure 8).

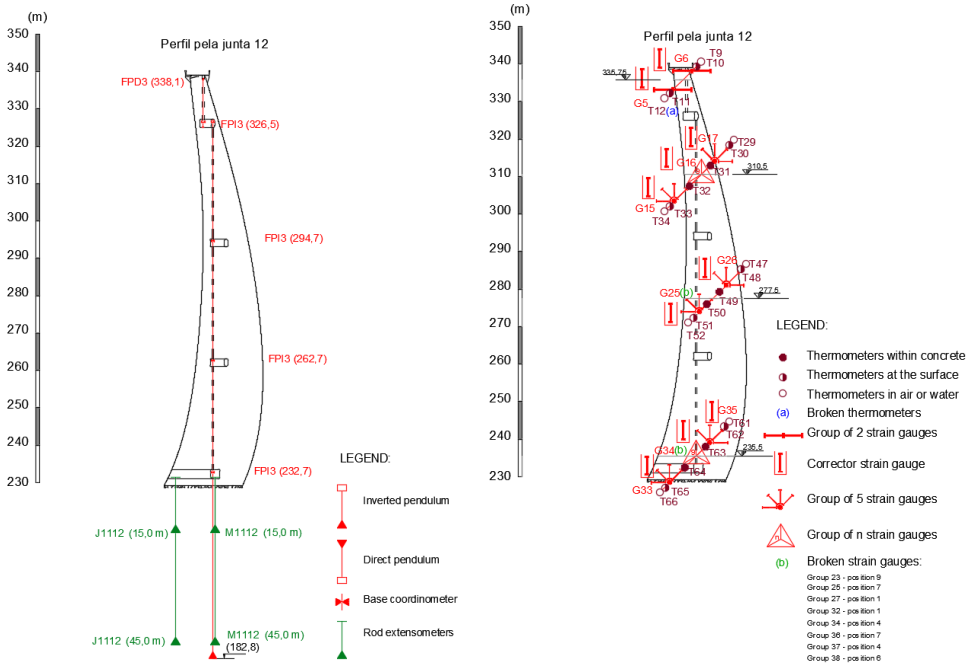

Figure 8 – Central block instrumentation [16].

5.1. Goals and Methodology

This case study follows the developments of Mata et al [16] among others [24] [2] [30] [31], and specifically, the use of artificial neural networks and random forests to obtain data-driven models for observed dependent variables in dams and comparing them with the traditional model. They also proposed and tested different formulations for these models.

It was proposed to research: i) whether the improvement of the performance of the models is possible by adding information related to the relative movements of the rock mass foundation measured via rod extensometers in the foundation (HTFT); ii) whether the application of the aforementioned machine learning techniques is reliable and comparable with the performance of the traditional method; iii) whether it is possible to reduce the number of independent variables in the formulations.

To achieve these goals, the following governing variables were used: the water level; the number of days since the first day of the year; the observed temperatures in the devices G06, G15, G34, T31 and T49; and the deformations observed in the rod extensometers M1112-45,0m (upstream) and J1112-45,0m (downstream). The dependent variables are the observed horizontal displacements along the radial direction at FP3-326,5m and FP3-232,7m.

To achieve a sound set of data it was necessary to create a cleaning data process. This cleaning process involved selecting the data such that each set of the measurements is taken at most +/- 3 days apart. This process resulted in 286 observation campaigns between January 1992 and January 2020. The data was divided into learning sets between January 1992 and April 2017 (280 observation campaigns) and prediction sets until January 2020.

The process to reduce the governing variables (temperatures in the dam body) was performed by selecting the most important variables among the measured temperatures. The most important input variables were selected as the ones with the lowest p-value in the traditional Multiple Linear Regression model, and also compared to the parameter, percentage of the mean square error (%incMSE), given by the Random Forest models.

5.2. Models

The proposed methodology resulted in the formulations depicted in Table 1 and Table 2 for each observed displacement and model. Models 6 and 7 are obtained from models 4 and 5 by doing the following linear transformations: $EF_T = (EF_{M1112\_45} + EF_{J1112\_45})/2$ that represent the vertical movement of the block foundation and $EF_R = EF_{M1112\_45} - EF_{J1112\_45}$ that represent the block rotation (upstream-downstream direction). Time effects were not considered in the time period studied. The relevance of the input related to the temperature effect was studied in order to minimize the number of inputs.

The three machine learning methods described previously: Multiple Linear Regression (IQ), artificial Neural Network (NN) and Random Forest (RF) were applied to these models. The models were computed using the R programming language [6] and respective libraries [32] [33].

Table 1 – Governing variables for each model for the displacement observed at FP3-326,5m

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Const.	k^*	k^*	k^*	k^*	k^*	k^*	k^*
Hydrostatic effect	h^4	h^4	h^4	h^4	h^4	h^4	h^4
Temperature effect	$sen\,(d)$	$G_{06\_33575\_T}$	$G_{06\_33575\_T}$	$G_{06\_33575\_T}$	$G_{06\_33575\_T}$	$G_{06\_33575\_T}$	$G_{06\_33575\_T}$
	$cos\,(d)$	$G_{15\_31050\_T}$	$G_{15\_31050\_T}$	$G_{15\_31050\_T}$	$G_{15\_31050\_T}$	$G_{15\_31050\_T}$	$G_{15\_31050\_T}$
		$G_{34\_23550\_T}$	$T_{31\_311\_T}$	$G_{34\_23550\_T}$	$T_{31\_311\_T}$	$G_{34\_23550\_T}$	$T_{31\_311\_T}$
		$T_{31\_311\_T}$		$T_{31\_311\_T}$		$T_{31\_311\_T}$	
		$T_{49\_278\_T}$		$T_{49\_278\_T}$		$T_{49\_278\_T}$	
Foundation deformation effect				EF_{M1112_45}	EF_{M1112_45}	EF_T	EF_T
				EF_{J1112_45}	EF_{J1112_45}	EF_R	EF_R

*There is no constant in the Random Forest models.

Table 2 – Governing variables for each model for the displacement observed at FP3-232,7m

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Const.	k^*	k^*	k^*	k^*	k^*	k^*	k^*
Hydrostatic effect	h^4	h^4	h^4	h^4	h^4	h^4	h^4
Temperature effect	$sen\,(d)$	$G_{06\_33575\_T}$	$G_{34\_23550\_T}$	$G_{06\_33575\_T}$	$G_{34\_23550\_T}$	$G_{06\_33575\_T}$	$G_{34\_23550\_T}$
	$cos\,(d)$	$G_{15\_31050\_T}$	$T_{31\_311\_T}$	$G_{15\_31050\_T}$	$T_{31\_311\_T}$	$G_{15\_31050\_T}$	$T_{31\_311\_T}$
		$G_{34\_23550\_T}$	$T_{49\_278\_T}$	$G_{34\_23550\_T}$	$T_{49\_278\_T}$	$G_{34\_23550\_T}$	$T_{49\_278\_T}$
		$T_{31\_311\_T}$		$T_{31\_311\_T}$		$T_{31\_311\_T}$	
		$T_{49\_278\_T}$		$T_{49\_278\_T}$		$T_{49\_278\_T}$	
Foundation deformation effect				EF_{M1112_45}	EF_{M1112_45}	EF_T	EF_T
				EF_{J1112_45}	EF_{J1112_45}	EF_R	EF_R

*There is no constant in the Random Forest models.

5.3. Main Results

From Table 3 and Table 4 it is possible to compare the performance of the selected models in both the learning and prediction period respectively for the upper displacement (FP3-326,5m), while from Table 5 and Table 6 the same applies for the lower displacement (FP3-232,7m). Being the residuals (ε) obtained from the difference between the measured value and the value obtained by the model, each of these tables present the following performance indicators: the minimum residual (ε_{min}), the maximum residual (ε_{max}), the standard deviation of the residuals (σ_ε), the coefficient of determination (R^2) and the sum of the squared error ($\sum \varepsilon^2$). In order to assess the quality of the generalization capacity and the performance of the selected models Figures 9 and 10 are presented.

The following figures and tables display the modelling fit, observed behavior of the radial displacement at FP3-326,5m and performance parameters, for both the learning period and prediction period, and for the three machine learning methods of four selected models.

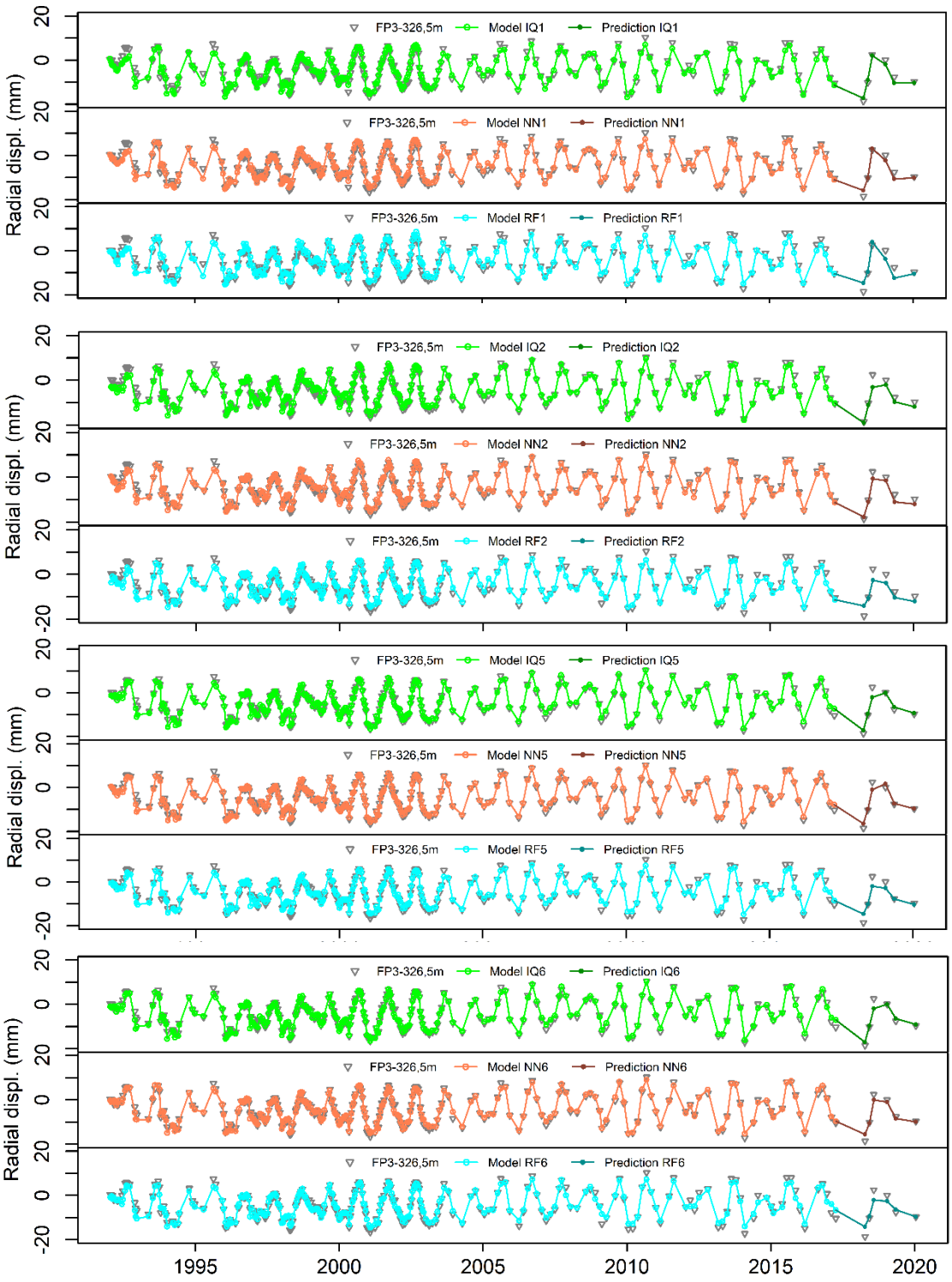

Figure 9 – Selected fitted models and prediction for the observed horizontal displacement at FP3-326,5m.

Table 3 – Model performance parameters for the learning period for the displacement observed at FP3-326,5m.

		Model IQ1	Model NN1	Model RF1	Model IQ2	Model NN2	Model RF2	Model IQ5	Model NN5	Model RF5	Model IQ6	Model NN6	Model RF6
ε_{min}	(mm)	-3,60	-3,06	-5,84	-3,21	-3,18	-6,20	-3,14	-2,82	-4,67	-3,31	-2,97	-4,27
ε_{max}	(mm)	6,33	5,01	6,32	6,78	5,46	4,98	5,20	5,32	4,55	5,22	4,03	4,34
σ_ε	(mm)	1,67	1,50	1,95	1,54	1,28	1,74	1,13	0,96	1,41	1,13	0,95	1,45
R^2	-	0,93	0,94	0,90	0,94	0,96	0,93	0,97	0,98	0,95	0,97	0,98	0,95
$\sum \varepsilon^2$	(mm^2)	776,90	631,09	1057,56	668,94	455,71	840,78	356,94	256,35	556,29	355,25	252,48	588,05

Table 4 – Model performance parameters for the prediction period for the displacement observed at FP3-326,5m.

		Model IQ1	Model NN1	Model RF1	Model IQ2	Model NN2	Model RF2	Model IQ5	Model NN5	Model RF5	Model IQ6	Model NN6	Model RF6
ε_{min}	(mm)	-2,74	-2,71	-4,84	-5,60	-0,87	-5,10	-4,53	-2,12	-4,23	-4,44	-3,06	-5,00
ε_{max}	(mm)	1,37	3,06	4,09	-0,39	3,40	4,56	1,49	3,43	3,94	1,51	2,55	4,48
σ_ε	(mm)	1,64	2,07	3,35	1.82	1,61	3,44	2,21	1,94	2,82	2,19	1,86	3,25
R^2	-	0,96	0,90	0,81	0,97	0,96	0,93	0,95	0,93	0,99	0,95	0,95	0,98
$\sum \varepsilon^2$	(mm^2)	15,04	22,64	58,26	45,68	30,36	75,21	24,50	19,09	42,15	24,03	18,04	53,64

The following figures and tables display the modelling fit, observed behavior of the radial displacement at FP3-232,7m and performance parameters, for both the learning period and prediction period, and for the three machine learning methods of four selected models.

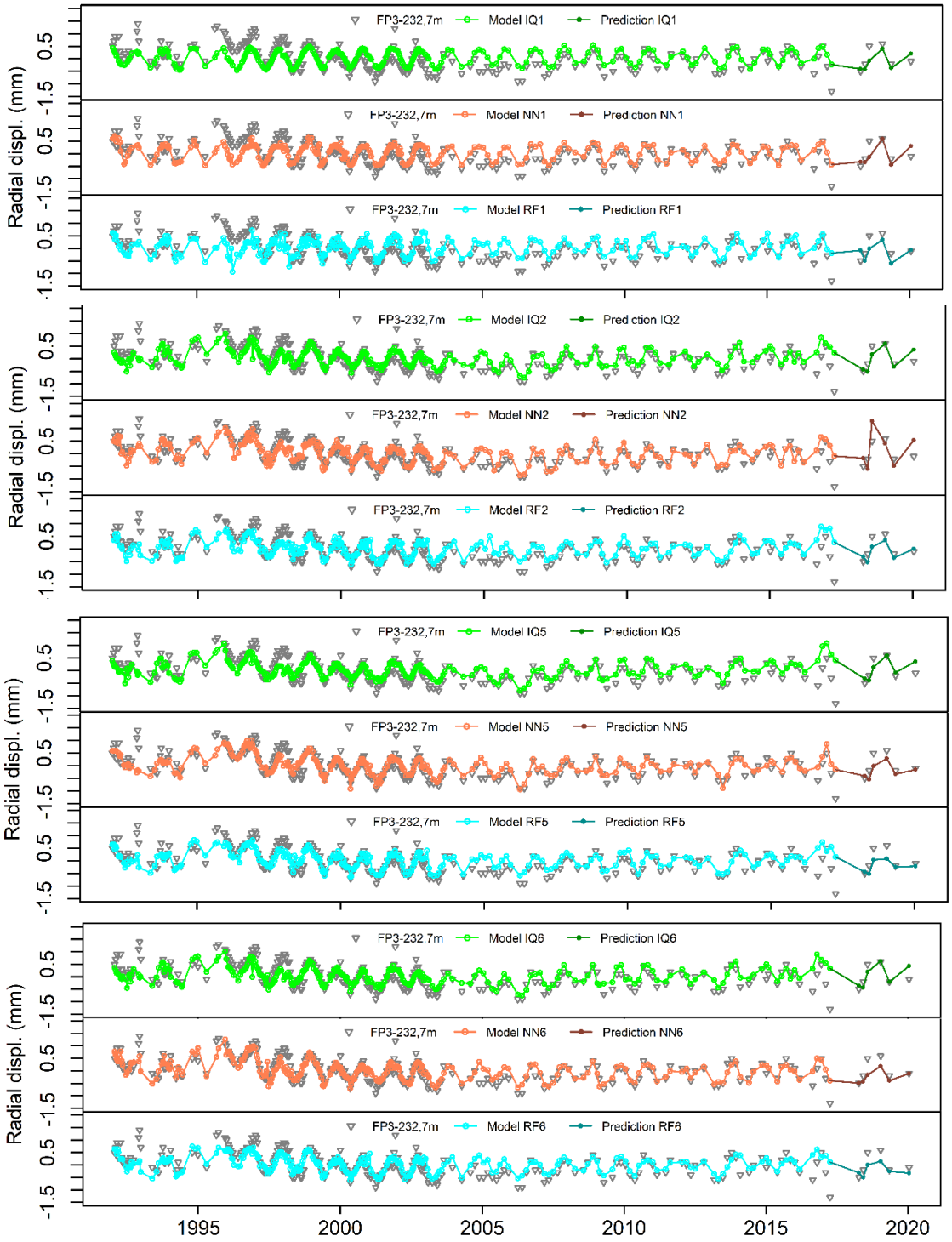

Figure 10 – Selected fitted models and prediction for the observed displacement at FP3-232,7m.

Table 5 – Model performance parameters for the learning period for the displacement observed at FP3-232,7m.

		Model IQ1	**Model NN1**	**Model RF1**	**Model IQ2**	**Model NN2**	**Model RF2**	**Model IQ5**	**Model NN5**	**Model RF5**	**Model IQ6**	**Model NN6**	**Model RF6**
ε_{min}	(mm)	-1,07	-0,90	-1,31	-1,55	-1,22	-1,16	-1,71	-1,15	-1,21	-1,63	-0,91	-1,32
ε_{max}	(mm)	1,20	1,15	1,14	1,48	1,01	1,59	1,57	1,58	1,50	1,44	1,08	1,48
σ_{ε}	(mm)	0,38	0,36	0,42	0,35	0,29	0,36	0,36	0,30	0,31	0,35	0,27	0,32
R^2	-	0,34	0,43	0,25	0,45	0,63	0,44	0,43	0,60	0,57	0,45	0,66	0,54
$\sum \varepsilon^2$	(mm2)	41,29	35,67	49,17	34,51	23,09	35,18	35,50	16,66	26,79	34,18	21,03	28,92

Table 6 – Model performance parameters for the prediction period for the displacement observed at FP3-232,7m.

		Model IQ1	**Model NN1**	**Model RF1**	**Model IQ2**	**Model NN2**	**Model RF2**	**Model IQ5**	**Model NN5**	**Model RF5**	**Model IQ6**	**Model NN6**	**Model RF6**
ε_{min}	(mm)	-0,57	-0,40	-0,51	-0,34	-0,80	-0,43	-0,36	-0,10	-0,51	-0,31	-0,01	-0,52
ε_{max}	(mm)	0,30	0,64	0,43	0,46	0,40	0,19	0,46	0,51	0,04	0,53	0,66	0,18
σ_{ε}	(mm)	0,30	0,36	0,35	0,29	0,51	0,24	0,29	0,22	0,23	0,30	0,26	0,26
R^2	-	0,53	0,42	0,40	0,62	0,51	0,69	0,57	0,78	0,83	0,59	0,72	0,76
$\sum \varepsilon^2$	(mm2)	0,54	0,68	0,76	0,44	1,43	0,44	0,43	0,51	0,59	0,45	0,69	0,67

Every proposed model was tested, validated and had the ability in providing good performance regarding the displacement at the upper point, while at the lower point the HST model has proven to experience less accuracy. The thermal effect showed better correlation with the displacements when considering the measured temperatures (HTT models) than with the seasonal formulation (HST model).

Although the MLR method and RF method diverged on which governing variables were more important at the lower point, it is also pointed out that reducing the governing variables did not significantly decrease the performance of the models. For both cases the HTFT models displayed better performance particularly based on the sum squared residuals parameter, although more data is necessary for validation purposes.

Both machine learning methods (NN and RF) showed good and comparable results with the traditional method (MLR). The NN method showed better results at the upper point and the RF method at the lower point, but both proved to be reliable in either situation, especially, in the learning period.

The linear transformation applied to the measurements regarding the foundation deformation did not have much effect on the performance of the models.

6. Final Remarks

The development of several models based on the use of several machine learning techniques (multiple linear regression (IQ), artificial neural networks (NN) and Random Forest (RF)) to the analysis and interpretation of the observed structural behavior of Alto Lindoso dam was presented.

The three formulations: Hydrostatic-Season-Time (HST), Hydrostatic-Temperature-Time (HTT) and Hydrostatic-Temperature-Foundation-Time (HTFT) were tested. The HTFT is an innovative proposal presented within the scope of this work. All the formulations showed to be suitable to model the pattern of the observed structural dam behavior based on the knowledge of recorded data.

As final remarks, it is possible to conclude that the use of the measurement of the temperatures in the dam body are relevant to represent the thermal effect and to depict the behavior of the horizontal displacement in concrete dams.

All the three machine learning methods referred before were used for the HST, HTT and HTFT formulations, and they have shown capable in providing reliable models with good generalization capacity to be used in the structural health monitoring of dams.

Acknowledgements

The author acknowledges his dissertation supervisors, PhD. Juan Mata who was a major force to reach the desired outcome and Professor João Almeida. Also, a word of appreciation is extended to the company EDP - Energias de Portugal S.A. that provided the data for the procedures addressed in this research project and to Laboratório Nacional de Engenharia Civil (LNEC) which provided some of the tools necessary to draw the conclusions in this dissertation.

References

[1] Swiss Committee on Dams (2003). Methods of analysis for the prediction and the verification of dam behaviour. 21st Congress of ICOLD, June 16-20, 2003, Montreal

[2] Salazar, F. (2017). A machine learning based methodology for anomaly detection in dam behaviour. PhD Thesis in Structural Analysis, Universitat Politecnica de Catalunya.

[3] Simon, A.; Royer, M.; Mauris, F.; Fabre, J. (2018). Analysis and Interpretation of Dam Measurements using Artificial Neural Networks. 9th ICOLD European Club Symposium, April 10-12, 2013, Venice, Italy.

[4] Mata, J. (2011). Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models. Engineering Structures, Volume 33, Issue 3, March 2011, p903-910.

[5] Mata, J.; Castro, A.; Costa, J. (2012). Constructing statistical models for arch dam deformation. Structural. Control Health Monitoring, Volume 22, Issue 3, March 2014, p.423-437.

[6] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[7] Helms, S. (1977). Jawa Excavations 1975: Third Preliminary Report. Levant, Volume 9, January 1977, p21-35.

[8] Quintela, A.; Cardoso, J.; Mascarenhas, J. M. (1986). *Aproveitamentos Hidráulicos Romanos a Sul do Tejo. Contribuição para a sua inventariação e caracterização*. EPAL – Empresa Portuguesa das Águas Livres, Lisbon

[9] Tanchev, L (2014). Dams and Appurtenant Hydraulic Structures (2nd edition). CRC Press/Balkema, Leiden, ISBN: 978-1-138-00006-3.

[10] Cheng, S. (2015). Hydraulic Structures. Springler-Verlag, Berlim, ISBN 978-3-662-47330-6.

[11] Jansen, R. (1988). Advanced dam engineering for design, construction, and rehabilitation. Van Nostrand Reinhold (1st edition), New York, ISBN:978-1-4612-8205-1.

[12] Ellingwood B.; Corotis, R.; Boland, J.; Jones, N. (1993). Assessing Cost of Dam Failure. Journal of Water Resources Planning and Management, Volume 119, Issue 1, January 1993, p64-82.

[13] Bharti, M.; Sharma, M.; Islam, N. (2020). Study on the Dam & Reservoir, and Analysis of Dam Failures: A Data Base Approach. International Research Journal of Engineering and Technology, Volume 7, Issue 5, May 2020, p1661-1669.

[14] Duffaut, P. (2013). The traps behind the failure of Malpasset arch dam, France, in 1959. Journal of Rock Mechanics and Geotechnical Engineering, Volume 5, Issue 5, October 2013, p.335–341.

[15] Genevois, R.; Ghirotti, M. (2005). The 1963 Vaiont Landslide. Gironale di Geologia Applicata, Volume 1, January 2005, p 41-52.

[16] Mata, J. (2013). Structural Safety Control of Concrete Dams Aided by Automated Monitoring Systems. PhD Thesis in Civil Engineering, Instituto Superior Técnico, Universidade de Lisboa.

[17] Karbhari, V.; Ansari, F. (2009). Structural health monitoring of civil infrastructure systems. CRC Press, New York. ISBN 978-1-4398-0131-4.

[18] Bukenya, P.; Moyo, P.; Beushausen, H.; Oosthuizen, C. (2014). Health monitoring of concrete dams: A literature review. Journal of Civil Structural Health Monitoring, Volume 4, Issue 4, November 2014, p235-244.

[19] Lee, I.; Shin, Y. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons, Volume 63, Issue 2, March-April 2020, p157-170.

[20] Salehi, H.; Burgueño, R. (2018). Emerging artificial intelligence methods in structural engineering. Engineering Structures, Volume 171, September 2018, p170-189.

[21] Sobral, M. (2014). *Regressão Linear e Árvores de Regressão: Previsão do desempenho na disciplina de Matemática.* MSc Dissertation in Statistics, Mathematics and Computation, Universidade Aberta.

[22] Worden, K.; Staszewski, W; Hensman, J. (2010). Natural computing for mechanical systems research: A tutorial overview. Mechanical Systems and Signal Processing, Volume 25, Issue 1, January 2011, p4-111.

[23] Granrut, M.; Simon, A.; Dias, D. (2019). Artificial neural networks for the interpretation of piezometric levels at the rock-concrete interface of arch dams. Engineering Structures, Volume 178, January 2019, p616-634.

[24] Kao, C.; Loh, C. (2013). Monitoring of long-term static deformation data of Fei-Tsui arch dam using artificial neural network-based approaches. Structural Control and Health Monitoring, Volume 20, Issue 11, November 2013, p282-303.

[25] Basheer, I.; Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods, Volume 43, Issue 1, December 2000, p3-31.

[26] Sullivan, W. (2017). Machine Learning for beginners: Algorithms, decision tree & random forest introduction (eBook Kindle). Healthy Pragmatic Solutions Inc, ASIN : B074YJYWDY.

[27] Hartshorn, S (2016). Machine Learning with Random Forests and Decision Trees: A Visual Guide for Beginners (eBook Kindle). Amazon, ASIN: B01JBL8YVK.

[28] Gates, M. (2017). Machine Learning: For Beginners – Your Definitive guide For Neural Networks, Algorithms, Random Forests and Decision Trees Made Simple (eBook Kindle). Amazon, ISBN: 978-1547039043.

[29] Genuer, R; Poggi, J (2020). Random forests with R. Springer, Cham. ISBN 978-3-030-56484-1.

[30] Amberg, F. (2009). Interpretative models for concrete dam displacements. 23rd ICOLD Congress, May 25-29, 2009, Brasilia, Brasil.

[31] Belmokre, A.; Mihoubi, M.; Santillan, D. (2019). Seepage and dam deformation analyses with statistical models: support vector regression machine and random forest. Procedia Structural Integrity, Volume 17, 2019, p698-703.

[32] Venables, W.N., Ripley, B. D. (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

[33] Liaw, A.; Wiener, M. (2002). Classification and Regression by Random Forest. R News, Volume 2/3, December 2002, p18-22.