Available online at www.sciencedirect.com

ScienceDirect

Procedia Structural Integrity 47 (2023) 325–330

Structural Integrity Procedia

www.elsevier.com/locate/procedia

27th International Conference on Fracture and Structural Integrity (IGF27)

Machine Learning tools applied to the prediction and interpretation of the structural behavior of existing dams

Caterina Nogara, Gabriella Bolzon*

Department of Civil and Environmental Engineering, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy

Abstract

The safety of existing dams is mainly ensured by the correct interpretation of monitoring data recorded during the whole lifetime of these structures. In this context, an increasing number of devices are being installed to provide more and more frequent measurements. Several Machine Learning tools have emerged as possible alternatives to traditional prediction approaches in recent years. Neural Networks have shown the ability to adapt to complex interactions and, therefore, to reach greater accuracy than conventional methods. However, this technique is susceptible to parameter tuning and difficult to generalize. Other recent studies have focused on Boosted Regression Trees. Less frequently used in dam engineering, they have proved to be equally accurate compared to Neural Networks, simpler to implement, and not sensitive to noisy and low relevant predictors. However, applications are limited to a few specific cases. The present contribution aims to evaluate the performances of this novel approach on dam data with a different specificity from previous research. The case study corresponds to a double-curvature arch dam introduced as a benchmark test by the International Commission on Large Dams. The input data include raw environmental variables, some derived variables, and time-related variables. Predictions of displacements under varying environmental conditions are performed, and relative influence indices are identified to determine the strength of each input-output relationship.

Keywords: Safety assessment; existing dams; structural monitoring; Macnine Learning tools.

* Caterina Nogara.
E-mail address: caterina.nogara@polimi.it
Gabriella Bolzon. Tel.: +39-02 2399 4219; fax: +39-02 2399 4330.
E-mail address: gabriella.bolzon@polimi.it

10.1016/j.prostr.2023.07.093

1. Introduction

Assessing the safety of dams is a fundamental and complex task due to the uniqueness of each structure and the uncertainty of local material properties and boundary conditions. The integrity evaluations are generally based on visual inspections and monitoring of the dam body and its foundation (ICOLD, 2000; ICOLD, 2012).

The properties to be measured for control purposes can be divided into causes and effects. The former group corresponds to state variables such as the water level and air and water temperatures. The latter identifies the structural response, represented by measurable quantities such as displacements, rotations, leakages, and piezometric pressures.

A predictive model, calibrated over a certain period of observations, uses the environmental variables, also named predictors, as inputs to return the value of a corresponding effect, known as the output (or prediction) of the model. The difference between each measurement and the relevant prediction is evaluated, verifying whether any discrepancy is contained within a tolerance limit, representative of the uncertainties of the problem. Otherwise, structural behavior anomalies are identified (Lombardi, 2005).

Technological improvements in recording and handling large amounts of data are associated with the development of reference models based on Artificial Intelligence (AI) and Machine Learning (ML) approaches. Several ML algorithms, such as Neural Networks (NNs) (Mata, 2011), Support Vector Machine (SVM) (Ranković et al., 2014), and Gaussian Process Regression (GPR) (Lin et al., 2019) have proven to be effective prediction tools when properly optimized and validated. Some studies have recently identified Boosted Regression Trees (BRT) as an algorithm particularly suitable for dam monitoring (Salazar et al., 2015; Salazar et al., 2016). However, applications are limited to a few specific cases. Therefore, the present contribution aims to evaluate the performances of this novel approach on dam data with a different specificity from previous research.

The considered case study is introduced in Section 2. Section 3 describes the BRT algorithm and its implementation for the given application. Finally, Section 4 presents some selected results.

2. Case study

The present case study corresponds to a double-curvature arch dam constructed between 1957 and 1960 and proposed in the 16th Benchmark Workshop by the International Commission on Large Dams (ICOLD) (Malm et al., 2022). The reference dam is equipped with several instruments, including pendulums, crack opening sensors, piezometers, and seepage measuring devices. Monitoring data have been regularly acquired since the first impound. Water level and air temperature are data available from 1995 to 2017, with a daily frequency.

The radial displacements of two points located at the crest and in the foundation (CB2 and CB3, respectively) of the central block of the dam are considered in this analysis. The time series of these displacements are provided from 2000 to 2012, with the frequency of one measurement every 1.5 weeks. This information is displayed in Fig. 1.

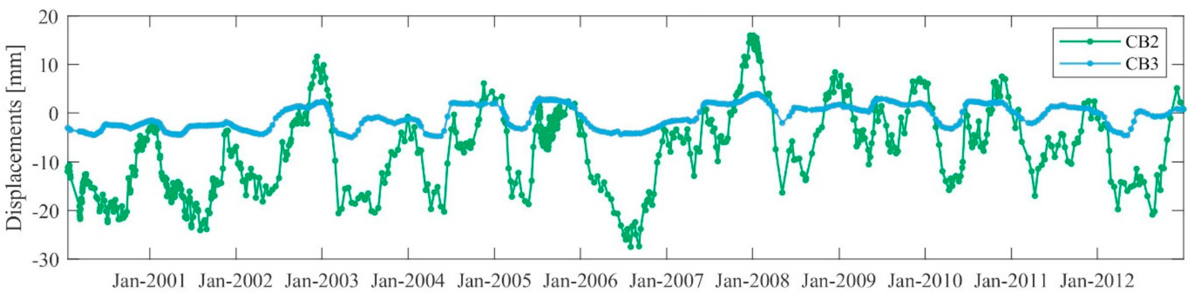

Fig. 1. Measured radial displacements.

3. Machine Learning model

3.1. Boosted Regression Trees

Boosted Regression Trees (BRTs) combine two algorithms: regression trees, which belong to the group of decision tree methods, and boosting, which builds and merges a set of models. Modern decision trees were described in detail by Breiman et al. (1984) and then by Hastie et al. (2001). As with any regression algorithm, the training data consists of p inputs and one response for each of the N observations (x_i, y_i), with $i = 1,2..N$ and $x_i = (x_{i1}, x_{i2} \dots x_{ip})$. In the present application field, the inputs correspond to the water level in the basin (WL), the air temperature (T), the number of days since the first recording, etc. The p data, recorded every 1.5 weeks, represent the entries x_i, while y_i is the corresponding radial displacement of either point CB2 or CB3.

The algorithm aims to subdivide the observations into a certain number of regions R_m according to the values of the input variables, assuming that the system response y_i can be represented by a constant within each sub-domain R_m, as schematized in Fig. 2(a) and Fig. 2(b). During the training phase, the algorithm employs a greedy heuristic approach. It selects the best option available at the moment to identify the splitting variables X and the splitting point x_k that define the regions, as shown, for instance, in Fig. 2(a) with the pairs T and t_k $(k = 1)$, as well as WL and wl_k $(k = 1, 2)$. For each input variable X, each of its x_k values is used as a threshold that divides the output into two partitions. The errors between the actual output values and the mean values associated with each of the two regions are then evaluated by an index. The pair variable-value (X, x) that minimizes the assumed error-index defines the node of the regression tree, as shown in Fig. 2(c). The averages of the output values become the predictions of the two branches derived from that node. The recursive binary partitioning process is repeatedly applied to each new region until some stopping criterion is reached.

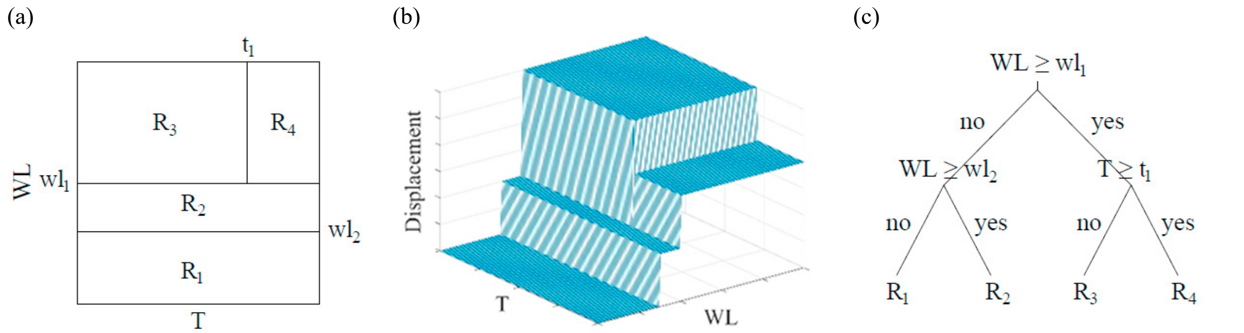

Fig. 2. (a) Partition of a two-dimensional input space by recursive binary splitting; (b) perspective plot of the prediction surface; (c) tree corresponding to the partition.

Boosting, on the other hand, is a method for improving the accuracy of a single algorithm by sequentially combining the results of several models (Friedman, 2001). For BRTs, the first regression tree is the one that minimizes the assumed loss function (for example, the Mean Squared Error, or MSE) for the selected tree size. For each subsequent step, the focus is on the residuals, i.e., on the variation in the response that the actual model does not explain.

Each step of the iterative procedure, which starts from $m = 1$ and ends with a defined number of trees ($m = nt$), consists of the following computations:

- the prediction error on the training set is computed as:

$$\tilde{y}_i = y_i - F_{m-1}(x_i) \tag{1}$$

where y_i are the actual values of the response and $F_{m-1}(x_i)$ are the predictions of the model at the generic m-th step of boosting method;

- a random subsample S_m of the training set is introduced and used to fit a new regression tree to the residual of the previous ensemble $\tilde{y}_i$:

$$y_i \approx f_m(X),\ i \in S_m \tag{2}$$

- the ensemble is updated, adding the contribution of the new regression tree modulated by a regularization parameter called learning rate ν ($0 \leq \nu \leq 1$):

$$F_m(X) \leftarrow F_{m-1}(X) + \nu f_m(X) \tag{3}$$

The fitted outputs in the final model are calculated as the sum of all trees multiplied by the learning rate. These results are much more stable and accurate than those obtained from a single regression tree model.

Therefore, for the implementation of a BRT model, it is necessary to define the values of three hyper-parameters:

1. the learning rate ν, which determines the contribution of each tree to the growing model;
2. the maximum number of splits, *ns*, which determines the complexity of each tree;
3. the number of trees to be considered, *nt*, which strongly depends on the two previous values.

3.2. Application of a BRT model to the dam monitoring data

The BRT algorithm is now applied to the data available for the considered dam, as introduced in Section 2. The aim is to create the two prediction models for the radial displacements at points CB2 and CB3, shown in Fig. 1. The considered dataset, made of input-target samples, is divided into two parts: 80% of the information is used for training, and 20% for testing. The predictors consist of 20 input variables: the reservoir level and some moving averages over different periods; the air temperature and its moving averages; time-related variables; the rate of change of the water level. The targets are the displacements measured at points CB2 and CB3, respectively.

As a common requirement for all prediction problems, overfitting of the training data should be avoided as this feature would reduce generality. Therefore, regularization methods are introduced to balance the model fit and predictive performances (Hastie et al., 2001). These incremental processes aim at the joint optimization of the number of trees *nt*, the learning rate *ν*, and the tree complexity *ns*.

The performance of several BRT models characterized by an increasing number of trees (*nt* varying from 1 to 5000), and different values of *ν* (0.002; 0.005; 0.01) and *ns* (1; 2; 4) is evaluated on the training set for the point CB2 through a 5-fold cross-validation technique (Elith et al., 2008). In this methodology, the training set is randomly divided into a number (in this case, 5) of subsets named folds. The model is then trained five times on different combinations of four folds, calculating the error-index on the excluded subset. The final validation error is then defined as the average of the five computations. It is generally observed that the cross-validation error decreases exponentially as the number of trees increases, but the computation time also expands dramatically. A reasonable compromise suggests limiting *nt* while the number of splits *ns* and the learning rate *ν* increase. However, $n_t > 1000$ is generally recommended not to jeopardize the accuracy of the model.

Based on the output of this analysis, the BRT model of CB2 displacements assumes *ν* equal to 0.01, *ns* equal to 4, and *nt* equal to 2000. The same cross-validation method is applied for optimizing the parameters of the BRT related to CB3. Both models are then trained again using all the data of the training set and tested on the remaining independent information.

4. Predictive performance

The accuracy of the predictions obtained with the tuned BRT model can be checked from the graph in Fig. 3. It displays the prediction (black points) of CB2 displacement compared to the measured data (blue points). The training set is represented on the left of the vertical dashed line, while test data are on the right.

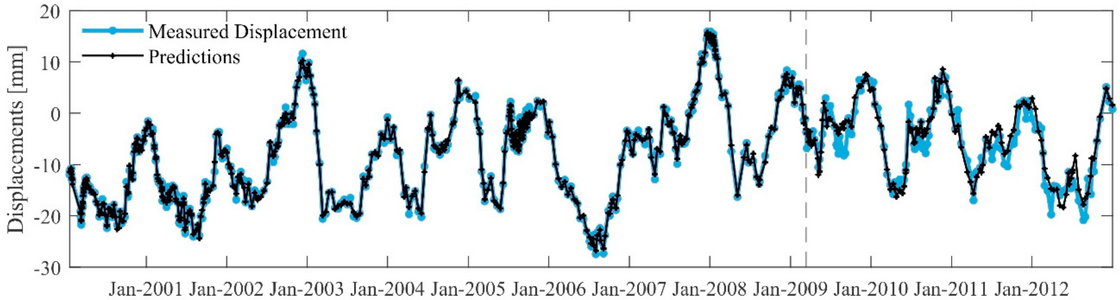

Fig. 3. Measured data and predictions of the BRT model for the training and test period of the CB2 radial displacement.

Table 1. MSE, MAE and NRMSE of the two BRT models for CB2 and CB3, on the training data and testing data.

Target	MSE training [mm^2]	MSE testing [mm^2]	MAE training [mm]	MAE testing [mm]	NRMSE training [-]	NRMSE testing [-]
CB2	0.262	4.899	0.394	1.762	0.012	0.051
CB3	0.005	0.313	0.051	0.451	0.008	0.063

Table 1 lists the Mean Squared Error (MSE), the Mean Absolute Error (MAE), and the Normalized Root Mean Squared Error (NRMSE) for each target (CB2 and CB3) in both the training and testing phase. The error indexes on the latter give an estimate of the predictive performance of the model. The values in both locations are comparable.

In addition to forecasting, BRT models can also be used to analyze the relationship between the causes (water level, temperature) and their effects (displacements). The formulae developed by Friedman (2001) estimate the relative influence of predictors by considering the number of times a variable is selected for splitting, weighted by the squared improvement in the model as a result of each split, and averaged over all trees (Friedman and Meulman, 2003). The relative influence of each variable is scaled so that the sum equals 100.

The relative importance of the predictors for CB2 and CB3 estimates is displayed in Fig. 4. It can be seen that the contribution of the hydrostatic load on radial displacements prevails on that due to temperature. The latter, however, is more significant for CB2. It can also be observed that the structural response at CB2 is more strongly correlated with long-term averages of water level and short-term averages of temperature. At the same time, CB3 displacement is more directly correlated with short-medium-term averages of water level and long-term averages of temperature. A meaningful influence of time, expressed in days, on both displacements should be further investigated.

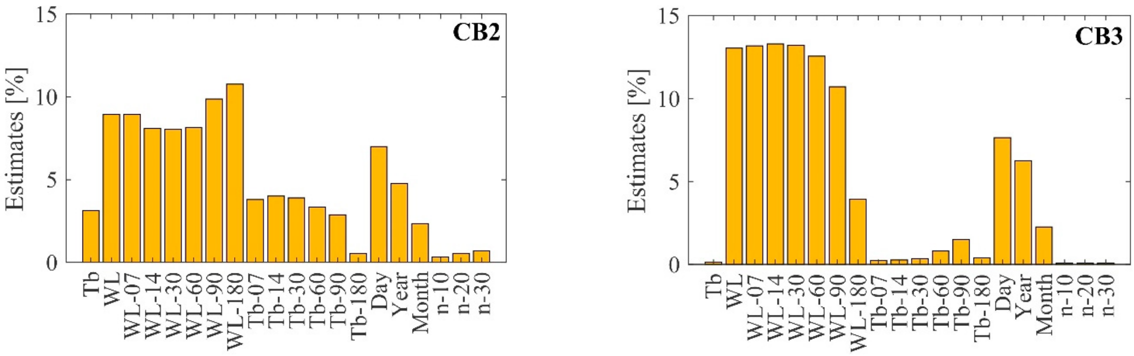

Fig. 4. Relative influence of inputs on the radial displacements, by considering: the water level (WL) and its moving averages over different time windows (WL – number of days); the air temperature (Tb) and its moving averages over different time windows (Tb – number of days); the number of days since the first recording (Day); the month and year corresponding to the recordings (Month, Year); the rate of variation of the water level over different periods (n – number of days).

5. Conclusion

Optimized BRT models allow for predicting the displacements occurring in a dam as a consequence of different external actions in order to support safety evaluations. The models also permit calculating the relative influence of each input variable on the structural response. Other ML methods are likely suitable tools for developing data-driven models, but BRT methodology is especially promising in terms of accuracy and interpretability. Furthermore, it is easy to implement, requiring a few hyper-parameters to be tuned. The next goal in structural safety assessment, based on continuous monitoring, is to develop a reliable detection system of possible anomalies with the identification of their origin.

References

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., 1984. Classification and regression trees. Wadsworth & Brooks. Cole Statistics/Probability Series.

Elith, J., Leathwick, J. R., Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77(4), 802–813.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232.

Friedman, J. H., Meulman, J. J., 2003. Multiple additive regression trees with application in epidemiology. Statistics in Medicine 22(9), 1365–1381.

Hastie, T., Tibshirani, R., Friedman, J., 2001. Data mining, inference, and prediction. The Elements of Statistical Learning; Springer: New York, NY, USA.

ICOLD, 2000. Automated dam monitoring systems: guidelines and case histories. International Commission on Large Dams, Technical Report B-118.

ICOLD, 2012. Dam surveillance guide. International Commission on Large Dams, Technical Report B-158.

Lin, C., Li, T., Chen, S., Liu, X., Lin, C., Liang, S., 2019. Gaussian process regression-based forecasting model of dam deformation. Neural Computing and Applications 31(12), 8503–8518.

Lombardi, G., 2005. Structural Safety Assessment of Dams: Advanced data interpretation for diagnosis of concrete dams.

Malm, R., Hellgren, R., Klun, M., Simon, A., Salazar, F., 2022. Theme A: Behaviour prediction of a concrete arch dam. 16th International Benchmark Workshop on Numerical Analysis of Dams. Ljubljana, Slovenia.

Mata, J., 2011. Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models. Engineering Structures 33, 3, 903–910.

Ranković, V., Grujović, N., Divac, D., Milivojević, N., 2014. Development of support vector regression identification model for prediction of dam structural behaviour. Structural Safety 48, 33–39.

Salazar, F., Toledo, M. A., Oñate, E., Morán, R., 2015. An empirical comparison of machine learning techniques for dam behaviour modelling. Structural Safety 56, 9–17.

Salazar, F., Toledo, M. T., Oñate, E., Suárez, B., 2016. Interpretation of dam deformation and leakage with boosted regression trees. Engineering Structures 119, 230–251.