COMMISION INTERNATIONALE DES GRANDES BARRAGES VINGT-SEPTIÈME CONGRÈS DES GRANDES BARRAGES *Marseille, Juin 2021*	

A FREE SOFTWARE FOR DAM MONITORING DATA ANALYSIS: EXPLORATION, CURATION AND MACHINE LEARNING MODEL FITTING*

Fernando Salazar
Project Development Director, CIMNE, Spain
André Conde
Research Engineer, CIMNE, Spain
Carlos Barbero
Head of Dam Safety Department, Catalan Water Agency

1. INTRODUCCIÓN

Safety has always been a priority for dam engineers throughout history. Today, it is becoming even more important as the average age of existing structures worldwide increases: there is a growing number of dams with several decades of operation, which are approaching the end of their service life. Many of these structures, however, continue to provide a fundamental service without viable alternatives, which aroused interest in intensifying control and safety measures.

The monitoring system is a crucial element in the control of dam safety. Advances in the performance of measuring devices allow greater accuracy and reading frequency in the most recent dams, but at the same time require advanced tools for the analysis of the large amount of data recorded.

In other technical and engineering fields, and especially in other sectors such as social networks or communication, the volumes of data handled are much greater and complex than those of the best instrumented dams, which has led to the development of very powerful analysis tools and techniques. This technology has been introduced in dam engineering in recent years at research level, and is close to penetrating professional practice. In this regard, the Technical Committee on Dam Surveillance (TCDS) of the ICOLD is preparing a bulletin provisionally

* *Un logiciel gratuit pour l'analyse des données de surveillance des barrages: exploration, curation et ajustement du modèle de machine learning*

entitled "Acquisition and Interpretation of Dam Surveillance Data and Observations". This document compiles the experiences accumulated in various countries by different actors on methods for processing and representing monitoring data and determining behavior patterns.

The authors of this contribution are collaborating in the writing of the mentioned document, on the basis of the knowledge acquired in the last years through the development of several research projects related to the use of advanced techniques of data analysis for the improvement of dam safety [1,2,3]. One of the results of this work is a software tool that includes functionalities specifically developed to meet the needs of dam safety managers in terms of data acquisition, curation and exploration and its subsequent use for the creation of predictive models using machine learning techniques.

In this contribution, some of the functionalities of the mentioned software are briefly described. The tool is free to use. All features have been developed using the programming language R [4]: the operations on the data, the generation of models and graphical utilities [5]. Monitoring data from the La Baells dam in the period 2008-2019 have been used for illustrating the functionalities. La Baells is a double-curvature arch dam with 100 m height above foundation, located in Barcelona.

2. DESCIPTION OF THE SOFTWARE TOOL

2.1 DATA EXPLORATION

The traditional and simplest way of representing monitoring data involves showing the time evolution of one or a group of variables. When several variables of different nature are to be shown, a secondary vertical scale is added. However, interesting information can be obtained by exploring other types of graphs, such as a scatterplot of response variables of the dam as a function of the reservoir level or temperature, or the relationship among response variables (e. g. a pendulum against another). Thus, one of the functionalities of the software allows the user to plot three-dimensional and interactive graphs where up to 4 variables of any type can be represented, including the reading date.

Figure 1 shows an example: one of the radial displacements measured in pendulums is plotted on the vertical axis, while the reservoir level and the 90-day moving average of the air temperature are shown on the horizontal axes.

Additionally, the points are colored according to the date of reading. The usefulness of this representation can only be appreciated by interacting with the graph, which can be rotated and zoomed. The figures below show the same data from different points of view. It is observed (Fig. 1 bottom left) how in this case a view can be found that clearly reflects a high linear correlation between a combination of the variables represented on the horizontal axes and the radial displacement. This relationship can only be seen by rotating the graph (not seen in the upper and bottom right views).

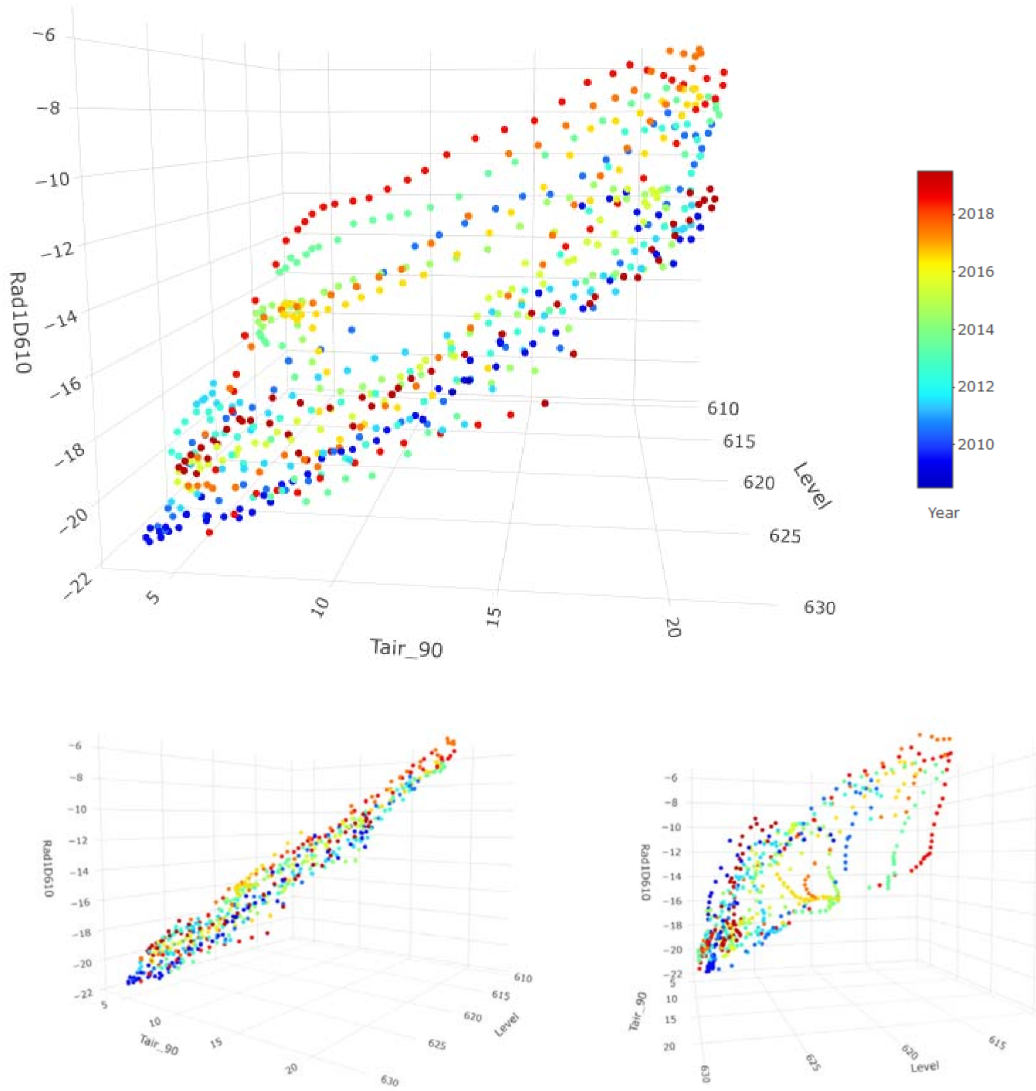

Fig.1. Data exploration. A radial displacement is plotted in the vertical axis, while reservoir level and air temperature (90-days moving average) are plotted in the 2 horizontal axes. The colors represent the date of reading.

2.2 DATA CURATION

Dam monitoring databases often include errors of different nature: missing data, different reading frequencies between devices, and reading errors, among others. Sometimes, changes in the systems cause the same series to present different reading frequencies in different periods. The formats in which the data is stored also vary between dam owners.

Different file formats can be used to load data into the application, including some of the most common ones such as csv or Excel. Exploratory charts allow the identification of those errors and periods without data, which can be corrected interactively: missing data can be interpolated, reading errors can be fixed. Further features, which include generating derived variables (e.g. moving averages), among others, are described in [6].

2.3 MACHINE LEARNING BASED MODEL FITTING

The development of machine learning techniques and the improvement in computing power allowed the use of sophisticated methods for the generation of prediction models based on monitoring data. Although the HST model [7] and similar linear regression approaches are still the most common in professional practice, there is a growing interest in the use of machine learning techniques. The list of published works in this line is very extensive, as well as the diversity of techniques used and cases analyzed. Some of them are already considered in the draft of the mentioned bulletin in progress by the TCDS. The authors are also collaborating in the preparation of a theme for the next benchmark workshop on numerical analysis of dams (to be held in Slovenia), focusing on the application of data-based prediction models.

The software includes an interface (Fig. 2) for the generation of predictive models based on a machine learning algorithm called Boosted Regression Trees (BRTs). It was selected after a comparative study among several of the most popular techniques [8] which showed that BRTs offered a greater prediction capability for different types of response variables, besides being robust and simple to apply: it depends on a small set of parameters and these have a minor effect on the final result.

Fig. 2. BRT-based predictive model generation.

The software allows to generate a prediction model with the mentioned algorithm. To do this, the variable to be predicted is selected, as well as the predictor variables to use. In the case of predicting radial displacements of arch dams, the thermal inertia of the dam can be considered by using as input the moving average of the ambient temperature in different periods, as shown in previous studies [9].

2.4 IDENTIFICATION OF BEHAVIOUR PATTERNS

One of the main obstacles for the extensive use of advanced models is the difficulty in their interpretation and therefore in drawing conclusions about the behavior of the dam. It is true that linear regression models allow to extract the effect of the input variables simply by analyzing the value of the coefficients. Machine learning based models require the use of more sophisticated methods to draw equivalent conclusions. However, these methods exist and are commonly used in other fields, and, more importantly, they offer more reliable results because they generally better represent the true behavior of the system. Numerous works have shown that these models allow consideration of non-linear effects, as well as collinearity between input variables.

BRT-based models can be analyzed by studying two metrics: the relative influence of the variables measures the decrease in prediction accuracy when each one of them is discarded: it is based on the assumption that if a variable not associated to the dam response is removed, the accuracy of the model should not change, and vice versa. Figure 3 shows the result in the example used, where it can be seen that the radial displacement considered is influenced by the reservoir level and the 90-day moving average temperature. It is interesting to see that the ambient temperature on the day of the measurement features very low influence, which reflects the thermal inertia of the dam.

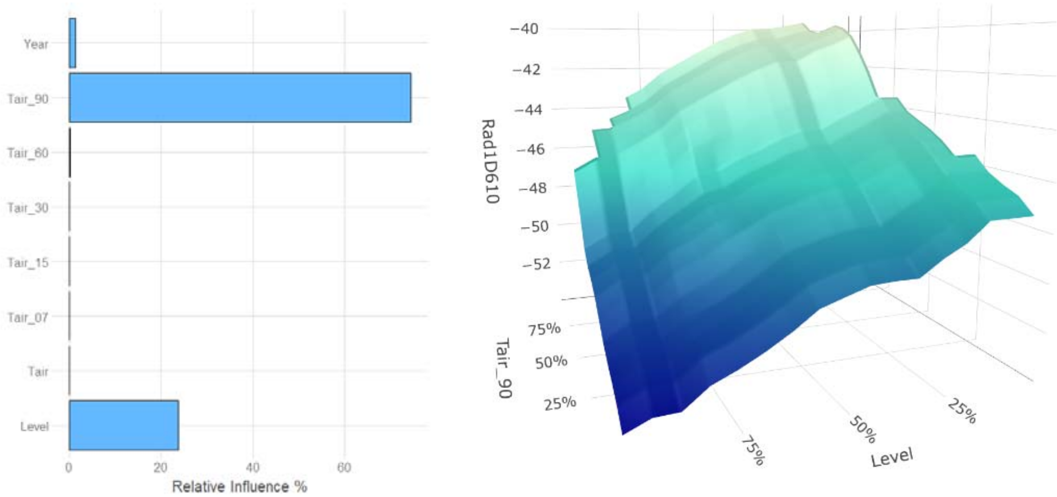

Fig. 3. Model interpretation. The relative influence (left) shows the thermal effect and its inertia, as well as the minor change in time (low influence of "Year"). The partial dependence plot (right) for the two most relevant inputs confirms that, as expected, the displacements towards downstream take place with low temperatures and high levels, and vice versa.

As for the partial dependence plot (Fig. 3 right), the result obtained in the example corresponds to engineering knowledge: the highest downstream displacement occurs under high hydrostatic load and low air temperature, and vice versa. What is relevant in the example is that the model has detected it automatically, without the need for preliminary variable selection. In particular, it identifies the greater relevance of the 90-day moving average of the temperature despite the introduction of other moving averages, which are highly correlated.

3. CONCLUSIONS

A software tool has been developed that allows loading dam monitoring data for exploration, error correction, completion of missing data and generation of derived variables, among other functionalities. With these data, prediction models can also be created based on BRTs, a robust, flexible and easy-to-apply machine learning algorithm. It offers in general a good accuracy and has procedures for the analysis of the model, also implemented in the tool. This allows to identify the actions that have greater influence on the response of the dam, and analyze that effect. The software is free to use, allowing it to be available to dam safety engineers worldwide, even those in less developed countries.

REFERENCES

[1] SALAZAR, F. A machine learning based methodology for anomaly detection in dam behaviour. PhD Thesis. Universitat Politècnica de Catalunya. 2017.

[2] SALAZAR, F., MORÁN, R., TOLEDO, M. Á., & OÑATE, E. (2017). Data-based models for the prediction of dam behaviour: a review and some methodological considerations. Archives of computational methods in engineering, 24(1), 1-21.

[3] SALAZAR, F., TOLEDO, M. Á., GONZÁLEZ, J. M., & OÑATE, E. (2017). Early detection of anomalies in dam performance: A methodology based on boosted regression trees. *Structural Control and Health Monitoring*, *24*(11), e2012.

[4] R DEVELOPMENT CORE TEAM, R. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (Vol. 1). Vienna.

[5] CHANG, W., CHENG, J., ALLAIRE, J., XIE, Y., MCPHERSON, J. 2018. Shiny: Web Application Framework for R. R package version 1.1.0 https://CRAN.R-project.org/package=shiny.

[6] SALAZAR, F., KOHLER, A., CONDE, A., LANDSTORFER, F. 2019b. Interpretation of dam monitoring data combining visualisation tools and machine learning. Eberlaste Dam Case Study. 15th International Benchmark Workshop on Numerical Analysis of Dams. Milano, 9-11 Sep. 2019

[7] WILLM AND BEAUJOINT: Les méthodes de surveillance des barrages au service de la production hydraulique d'Electricité de France, problèmes anciens et solutions nouvelles, in IXth International Congress on Large Dams. Istanbul, pp. 529–550. (In French) (1967)

[8] 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.

[9] SALAZAR, F., TOLEDO, M. A., OÑATE, E., & SUÁREZ, B. 2016. Interpretation of dam deformation and leakage with boosted regression trees. Engineering Structures, 119, 230-251.