III Jornadas de Ingeniería del Agua
La protección contra los riesgos hídricos

Posibilidades de la inteligencia artificial en el análisis de auscultación de presas

F. Salazar, E. Oñate
Centre Internacional de Mètodes Numèrics en Enginyeria (CIMNE)
Gran Capitán s/n. 08034 Barcelona.

M. Á. Toledo
Departamento de Ingeniería Civil. Hidráulica y Energética.
Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos
Universidad Politécnica de Madrid (UPM). Profesor Aranguren s/n. 28040 Madrid.

1. Introducción

El comportamiento estructural de las presas de embalse es difícil de predecir con precisión. Los modelos numéricos para el cálculo estructural resuelven bien las ecuaciones de la mecánica de medios continuos, pero están sujetos a una gran incertidumbre en cuanto a la caracterización de los materiales, especialmente en lo que respecta a la cimentación. Así, es difícil discernir si un estado que se aleja en cierta medida de la normalidad supone o no una situación de riesgo estructural.

Por el contrario, muchas de las presas en operación cuentan con un gran número de aparatos de auscultación, que registran la evolución de diversos indicadores como los movimientos, el caudal de filtración, o la presión intersticial, entre otros. Aunque hoy en día hay muchas presas con pocos datos observados, hay una tendencia clara hacia la instalación de un mayor número de aparatos que registran el comportamiento con mayor frecuencia (Restelli 2008)0. Como consecuencia, se tiende a disponer de un volumen creciente de datos que reflejan el comportamiento de la presa. En la actualidad, estos datos suelen tratarse con métodos estadísticos para extraer información acerca de la relación entre variables, detectar anomalías y establecer umbrales de emergencia.

El modelo general más común es el denominado HST (Hydrostatic-Season-Time), que calcula la predicción de una variable determinada a partir de una serie de funciones que tienen en cuenta los factores que teóricamente más influyen en la respuesta: la carga del embalse, el efecto térmico (en función de la época del año) y un término irreversible. El método fue desarrollado en 1967 (Simon et *al*. 2013) por ingenieros de Électricité de France (EDF) y se ha aplicado especialmente a la predicción de desplazamientos (Swiss Comittee on Dams, 2003).

Aunque se ha utilizado con éxito durante mucho tiempo, se han detectado algunas limitaciones del método, como el hecho de que asume que los tres efectos son independientes, y que las funciones deben definirse a priori. Esto permite un margen de mejora en caso de que no representen el efecto real en un caso concreto (Simon et *al*., 2013). Además, no permite reproducir la inercia de la presa en su respuesta frente a ciertas solicitaciones, como es el caso de la relación entre el nivel de embalse y la presión intersticial en presas de materiales sueltos (Bonelli y Radzicki 2008).

Puntualmente se han aplicado modelos más complejos para solventar las limitaciones mencionadas. En algunos casos se han introducido otras variables como la velocidad de variación del nivel de embalse (Sánchez Caro 2003), y en otros se han utilizado expresiones que se adaptan mejor al efecto de inercia, como la función impulso-respuesta (Bonelli y Radzicki 2008).

En otros campos de la ciencia, como la medicina o las telecomunicaciones, el volumen de datos es mucho mayor, lo que ha motivado el desarrollo de numerosas herramientas para su tratamiento y para la generación de modelos de predicción. Algunas de ellas, como las redes neuronales, ya han sido aplicadas al caso de la auscultación de presas (Santillán et *al*. 2010, Mata 2011, Simon et *al*. 2013) con resultados prometedores.

La aplicación de estas técnicas puede ayudar a mejorar la precisión de los modelos de predicción, y a entender mejor el comportamiento de la presa. Con esta idea se ha puesto en marcha el proyecto de investigación iComplex, en el que participan la empresa DACARTEC, la Universidad Politécnica de Madrid (UPM) y el Centro Internacional de Métodos Numéricos en Ingeniería (CIMNE).

Una de las tareas comprendidas en la primera fase del proyecto es la revisión de diversas herramientas de inteligencia artificial de cara a su aplicación a la predicción del comportamiento de presas: movimientos, tensiones, filtraciones, etc. Una de las herramientas que se está considerando como potencialmente útil está basada en los llamados bosques aleatorios. A continuación se describen someramente las bases de esta tecnología y se muestran resultados preliminares de su aplicación a un caso piloto.

2. Sobre los bosques aleatorios

Los bosques aleatorios (Breiman 2001) son modelos que permiten predecir el valor de una determinada variable (*variable objetivo*) a partir de una serie de *variables predictoras*, cuyo valor es conocido. Como los métodos estadísticos, requiere de unos datos de entrenamiento, a partir de los cuales se ajusta el modelo al caso de estudio. Un bosque aleatorio está formado por un conjunto de árboles de decisión. La predicción del bosque es el promedio de las predicciones de los árboles que lo forman. Por tanto, se trata de un *modelo de conjunto* (Martínez, 2006).

Los árboles de decisión se basan en la división sucesiva del conjunto de datos observados en grupos de casos “similares”. Suelen denominarse *árboles de regresión* aquellos cuya variable

objetivo es continua, y *árboles de clasificación* cuando es discreta o categórica. En adelante se utilizará por tanto el término árbol de regresión, ya que en auscultación de presas se trata de predecir variables continuas. La predicción un árbol de regresión es en general un valor constante para cada grupo, igual a la media de los valores observados. Para explicar el proceso de generación de un árbol de decisión, se utiliza un ejemplo sencillo relacionado con la auscultación de presas: supongamos que se desea ajustar un modelo para predecir el caudal de filtración en un determinado aforador a partir únicamente del nivel de embalse. La Figura 1 muestra la relación entre las variables de entrada (nivel de embalse) y objetivo (caudal de filtración).

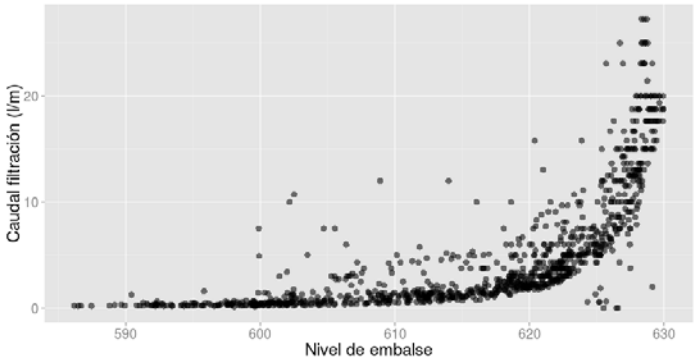

Figura 1. *Caudal de filtración en función del nivel de embalse en el caso de ejemplo*

En primer lugar, se dividen los datos en dos regiones según el nivel de embalse (en el ejemplo, según sea mayor o menor de 625,6; Figura 2). A continuación, una de las regiones creadas se subdivide a su vez en dos, y el proceso continúa hasta que se alcanza algún criterio de parada. En el ejemplo, el resultado final es la división de los casos en 7 grupos. La predicción del modelo es la media de los valores observados en cada grupo, y por tanto el resultado es una sucesión de escalones (Figura 3).

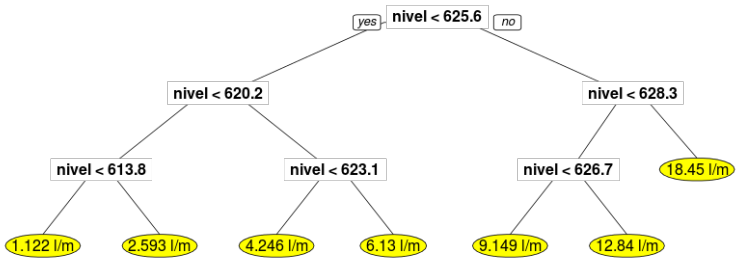

Figura 2. *Árbol de regresión para la predicción del caudal de filtración en función del nivel de embalse en el caso de ejemplo.*

El algoritmo de generación del árbol calcula la mejor división posible en cada paso como la que minimiza el error de predicción. Si hubiera más de una variable predictora, se calcula para cada una de ellas la división óptima y a continuación se selecciona la variable que produce una división más precisa (Hastie et *al.* 2009).

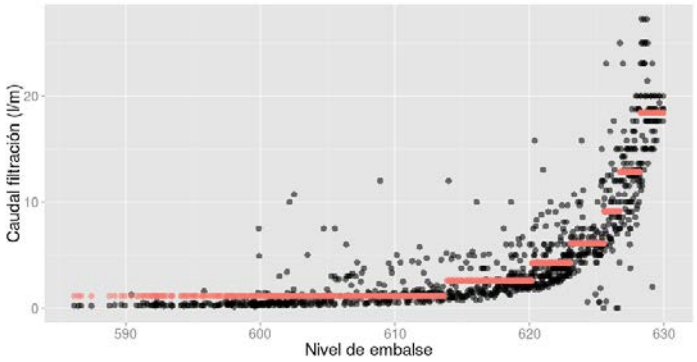

Figura 3. *Predicción del árbol de regresión (en rojo), en comparación con los datos registrados (en negro)*

Algunas de las propiedades más representativas de los árboles de decisión son las siguientes: a) su coste computacional es moderado; b) se adaptan bien a problemas no lineales; c) manejan sin problemas mezclas de variables continuas y discretas; además, las continuas pueden tener rangos muy diferentes lo que evita tener que transformarlas previamente, como ocurre con otros métodos; d) permiten considerar valores no medidos; e) los valores atípicos no modifican sustancialmente el resultado; f) no es necesario adoptar ninguna hipótesis a priori sobre la relación entre variables, ni sobre cuáles son más relevantes; g) son inestables, en cuanto a que una pequeña variación de los datos de entrenamiento puede producir una variación importante del resultado. Este problema puede convertirse en una ventaja si se utilizan métodos avanzados basados en árboles, como es el caso de los bosques aleatorios.

Los bosques aleatorios pertenecen a los denominados *modelos de conjuntos*, basados en la generación de un número (generalmente elevado) de modelos sobre una misma base de datos (o subconjuntos de ella). La predicción se calcula como la media de las predicciones de los modelos individuales. Los bosques aleatorios son un conjunto de árboles de decisión generados a partir de perturbaciones de los datos observados. El aspecto clave que caracteriza el método es que busca que los árboles sean independientes entre sí. Para ello, la diferencia principal con un árbol de decisión convencional es que en cada división, en lugar de considerar todas las variables predictoras disponibles para seleccionar la que minimiza el error, se analiza únicamente un **subconjunto aleatorio** de las mismas (Breiman 2001). De este modo, se aprovecha la propiedad de los árboles de decisión por la cual pequeñas perturbaciones en las primeras divisiones del espacio provocan resultados apreciablemente diferentes. Al introducir aleatoriedad en la construcción de cada árbol, se

consigue un conjunto de ellos sensiblemente independientes, de modo que se captura una proporción mayor de los patrones presentes en los datos de entrenamiento, y el resultado final mejora. Además, cada árbol se genera a partir de un conjunto de datos diferente, obtenido de los originales tomando una **muestra aleatoria con repetición**. Es decir, cada nuevo árbol se construye a partir de una muestra en la que aproximadamente un tercio de los datos originales aparece una vez, otro tercio aparece repetido, y el tercio restante no aparece.

Los bosques aleatorios han cobrado gran popularidad como método para generar modelos predictivos por su sencillez de programación y sus buenos resultados en diversas aplicaciones (Ganuer et *al*. 2008). Si bien suelen considerarse un modelo *de caja negra*, en cuanto a que no proporciona parámetros con interpretación física, se han desarrollado herramientas que permiten cuantificar cómo afecta cada variable al resultado final. En concreto, se define el índice de importancia de una variable como la variación del error de predicción que se produce al modificar aleatoriamente su valor, manteniendo el resto. Las variables más importantes provocarán un mayor aumento del error al ser permutadas.

3. Ejemplo de aplicación

Se ha elaborado un modelo de predicción en un caso de prueba basado en bosques aleatorios. Las variables que se pretende predecir son los caudales de filtración de una presa tomada como caso piloto. El periodo de datos disponible comprende desde la puesta en carga hasta el año 2008. Se han considerado los aforadores con un mayor número de datos registrados en ese periodo, que son los de la Tabla 1.

Además de los datos de aforo, del nivel de embalse y de otras magnitudes que no son objeto de este trabajo, se miden en la presa variables ambientales como precipitación y temperatura del aire.

Para esta primera prueba, las variables utilizadas para predecir el valor del caudal de filtración son: a) el número de día del registro, contado a partir de la puesta en carga de la presa; b) el año; c) el nivel de embalse medio el día de la lectura; d) la temperatura media ambiental; e) la precipitación acumulada en los 30 días anteriores a la lectura; f) la velocidad media de variación del nivel de embalse en los 10 días anteriores a la lectura; g) la media móvil de 60 días del nivel de embalse.

Aforador	Margen	Nº datos disponible
md50pr	Derecha	1023
md90pr	Derecha	748
totmd	Derecha	1071
mi50p	Izquierda	1066
mi90pr	Izquierda	1001
totmi	Izquierda	1021

Tabla 1. *Número de datos disponibles en cada aforador*

Se han dividido los datos disponibles en dos grupos. El primero se utiliza para ajustar los parámetros del modelo (datos de entrenamiento), y el segundo para comprobar la bondad del ajuste (datos de validación). La división se ha realizado de dos formas: a) el 60% de los datos más antiguos para entrenamiento, y el 40% más reciente para validación. Este es el criterio utilizado en el análisis de seguridad de la presa, así como en el trabajo de Santillán *et al.,* que utiliza redes neuronales (Santillán 2010); b) división aleatoria en todo el periodo registrado, con un 70% para entrenamiento y un 30% para validación.

En la 0 se muestra el error resultante en cada caso.

RMSE (l/min)				
División de los datos	60%-40% temporal		70%-30% aleatoria	
Aforador	Entrenamiento	Validación	Entrenamiento	Validación
md50pr	2,15	2,64	2,12	1,77
md90pr	0,40	1,58	0,55	0,47
totmd	2,56	4,42	2,41	2,64
mi50p	0,48	0,41	0,45	0,41
mi90pr	0,16	0,24	0,17	0,10
totmi	0,67	1,05	0,62	0,60

Tabla 2. *Errores de predicción del modelo (raíz del error cuadrático medio)*

Como ejemplo, en la Figura 4 muestra el resultado para el aforador "md50pr", que es al que corresponde también la Figura 3. Puede apreciarse la mejora de la predicción del bosque aleatorio respecto del árbol de regresión individual.

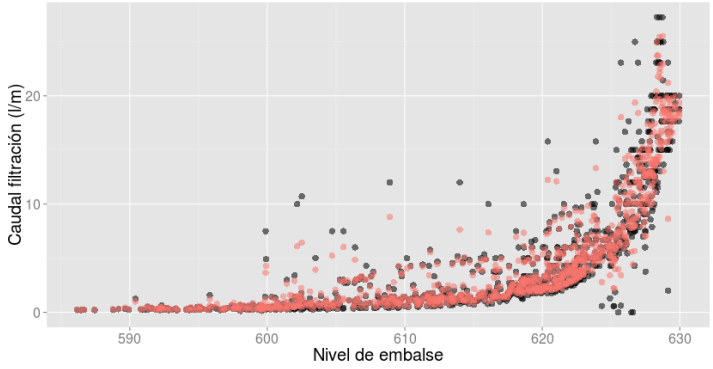

Figura 4. *Predicción del modelo basado en bosques aleatorios (en rojo), en comparación con los datos registrados (en negro). Aforador "md50pr".*

Se ha calculado también el índice de importancia de las variables, y llama la atención el hecho de que las relativas al periodo de medición (número de día desde la puesta en carga y

año de lectura) son siempre más relevantes que otras como la temperatura o la precipitación. En algunos casos, llegan a serlo más incluso que el nivel de embalse (Figura 5 izquierda). Para verificar estos resultados, se ha dibujando la relación entre el nivel de embalse y el caudal de filtración separando los registros por intervalos temporales. La Figura 6 muestra dos gráficos de este tipo, donde se observa claramente cómo en algunos casos (izquierda) el caudal de filtración depende en gran medida del periodo de la vida de la presa, lo cual no sucede en otros (derecha). Otro índice de la variación de la respuesta de la presa con el tiempo lo representa el hecho de que el error de entrenamiento es similar independientemente de cómo se dividan los datos, mientras que el de validación es sensiblemente inferior si se toman aleatoriamente. Si como se observa en la Figura 6 izquierda la respuesta de la presa varía con el tiempo, es lógico que un modelo entrenado con los datos de un periodo determinado se ajuste peor al aplicarlo a un periodo diferente.

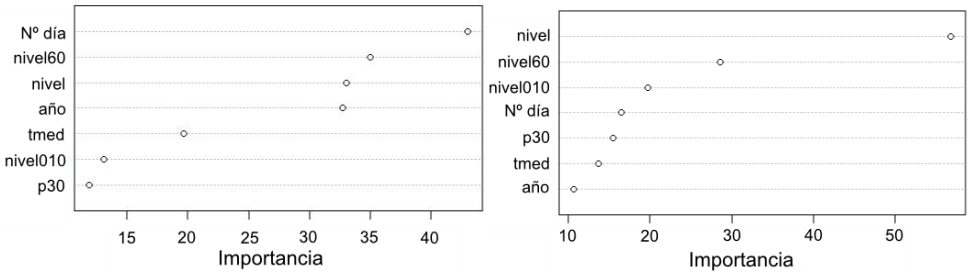

Figura 5. *Importancia de las variables predictoras en dos de los aforadores estudiados. Derecha: "totmi", donde el día del registro es la variable más importante, lo que denota una evolución temporal en el comportamiento. Izquierda: "md50pr", donde el nivel de embalse es claramente lo que más influye en la filtración.*

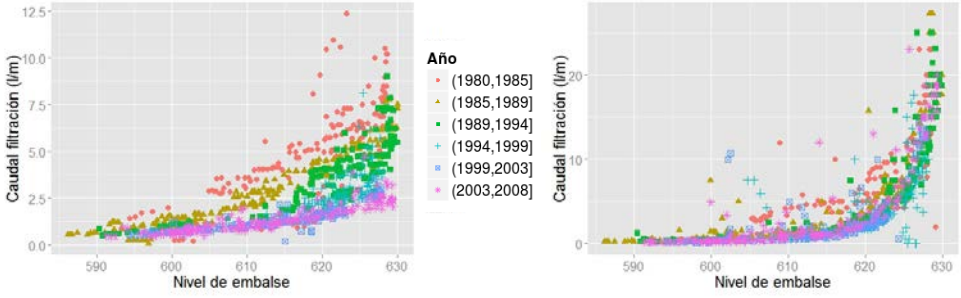

Figura 6. *Caudal de filtración "totmi" (izquierda) y "md50pr" (derecha) en función del nivel de embalse, separados por periodos de tiempo. Se observa que en el primer caso el comportamiento depende de manera importante del año de lectura, mientras que en el segundo depende más del nivel de embalse, lo cual es coherente con los índices de importancia calculados.*

Los resultados obtenidos hasta el momento con los bosques aleatorios sugieren que pueden ser una herramienta útil no solo como modelo de predicción, sino también para obtener información acerca del comportamiento de la presa, y del efecto de las variables de entorno.

En la actualidad se está trabajando para desarrollar criterios de selección de variables, así como para determinar cómo evoluciona la precisión del modelo en función del número de datos disponible.

Agradecimientos

Los autores quieren agradecer a la Agencia Catalana de l'Aigua la autorización para el uso de los datos de una de sus presas con fines de investigación, y a Ofiteco las gestiones realizadas para ello. También al Ministerio de Economía los medios facilitados para esta investigación dentro del proyecto "Desarrollo del software iComplex para el control y evaluación de la seguridad de infraestructuras críticas", del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008-11; IPT-2012-0813-390000.

Referencias

Bonelli, S., & Radzicki, K. 2008. Impulse response function analysis of pore pressures in earthdams. *European Journal of Environmental and Civil Engineering*, *12*(3), 243-262.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1), 5-32, 2001.

Ganuer, R., Poggi, J.M., Tuleau, C. 2008. Random Forests: some methodological insights. arXiv: 0811.3619.

Hastie, T., Tibshirani, R. y Friedman, J. 2009. The elements of statistical learning - Data mining, Inference and Prediction. Springer, 2ª edición.

Martínez, G. 2006. Clasificación mediante conjuntos. Tesis Doctoral. Universidad Autónoma de Madrid.

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

Restelli, F. 2008. Utilización de modelación conexionista para la determinación de patrones en respuestas pulsantes de sistemas automáticos. *V Congreso Argentino de Presas*, Tucumán, Argentina. 2008.

Sánchez Caro, F. 2007. Seguridad de presas: aportación al análisis y control de deformaciones como elemento de prevención de patologías de origen geotécnico. Tesis Doctoral. UPM.

Santillán, D., Morán, R., Fraile, J.J., Toledo, M.Á. 2010. Forecasting of dam flow-meter measurements using artificial neural networks. En Romeo García et *al.* (eds) Dam Maintenance and Rehabilitation II, CRC Press, Londres, 2010, pp 183-189.

Simon, A., Royer, M., Mauris, F. y Fabre, J.P. 2013. Analysis and Interpretation of Dam Measurements using Artificial Neural Networks. En *Proceedings of the 9th ICOLD European Club Symposium*, Venecia, 2013.

Swiss Committee on Dams. Methods of analysis for the prediction and the verification of dam behavior. En *21º Congress of the International Commission on Large Dams*, Montreal, 2003.