Tratamiento y análisis inteligente de datos del comportamiento de estructuras hidráulicas

André CONDE[a], Fernando SALAZAR[b] y David J. VICENTE[c]

[a]International Center for Numerical Methods in Engineering, *aconde@cimne.upc.edu*, [b]International Center for Numerical Methods in Engineering, *fsalazar@cimne.upc.edu*, y [c]International Center for Numerical Methods in Engineering, *djvicente@cimne.upc.edu*.

Línea temática | (v) Tema Monográfico “Ingeniería del agua 4.0”

RESUMEN

El rendimiento de los dispositivos de monitorización ha experimentado mejoras permitiendo disponer de más información sobre el comportamiento de las estructuras. Sin embargo, las inversiones realizadas en modernización de sistemas de medida no se recuperan a menos que se complementen con aplicaciones capaces de manejar una información tan amplia y diversa.
En esta contribución, se presenta una herramienta de software para importar, explorar, limpiar y analizar datos de monitorización. Además, permite la generación de modelos de predicción basados en *machine learning*, así como la interpretación de la respuesta del sistema a las acciones o cargas en funcionamiento.
La metodología y la estructura general se dividen en dos fases: i) carga, depuración, completado y análisis datos, y ii) generación e interpretación de modelos predictivos basados en *machine learning*. El mismo modelo puede utilizarse para la detección de anomalías comparando las predicciones con el comportamiento registrado.

Palabras clave | exploración interactiva de datos; limpieza de datos; preprocesamiento de datos; interpretación de datos; *machine learning*.

INTRODUCCIÓN

El análisis de datos es la base del análisis de muchos sistemas hidroambientales en áreas como la hidrología, la gestión de embalses o la seguridad de presas. Este análisis de conjuntos de datos complejos se convierte a menudo en el factor limitante en estudios de ingeniería (Kotsiantis y Kanellopoulos, 2006). Independientemente del volumen de datos disponible, el análisis en tiempo real es necesario en todas las escalas de resolución, desde el conjunto de datos completo hasta valores singulares (Thorvaldsdóttir et al., 2013). La evaluación de la seguridad se basa fundamentalmente en el análisis experto de gráficos que muestran la evolución de las variables más relevantes, así como su relación con las principales cargas actuantes. Aunque gran parte del análisis puede realizarse automáticamente, la interpretación y la experiencia de los técnicos, basada en una visualización de datos rápida y flexible, son esenciales para distinguir entre relaciones complejas entre variables y errores o ruido aleatorio (Guyon y Elisseeff, 2003). Tradicionalmente, la representación de los datos de monitorización de estructuras hidráulicas se ha limitado a hojas de cálculo convencionales, lo que era suficiente cuando se manejaban datos manuales con baja frecuencia de lectura.

La tendencia hacia la instalación de sistemas automáticos de adquisición de datos (ADAS) ha dado lugar a un aumento de la cantidad de datos disponibles. Sin embargo, este aumento del volumen de datos disponibles también plantea problemas en lo que respecta a las herramientas de análisis y procesamiento de datos, así como a los métodos para generar modelos predictivos. Desafortunadamente, el tamaño y la diversidad de los conjuntos de datos producidos por los sistemas actuales a

 ISBN: XXX-XX-XXXXX-XX-X

menudo presentan grandes desafíos para las aplicaciones de visualización. Además, la exploración de datos se limita frecuentemente a observar la evolución de las series temporales de cada variable registrada. Los sistemas de monitorización avanzados requieren aplicaciones capaces de manejar una información tan amplia y diversa. Los gráficos de series temporales convencionales y los métodos estadísticos simples no son apropiados para maximizar la información extraída de estos datos. Aunque se han desarrollado algunas herramientas específicas que incorporan ciertas características (por ejemplo, Mora et al., 2008), las herramientas utilizadas en la ingeniería para la presentación y análisis de estos datos se limitan a menudo a los programas ofimáticos convencionales o a laboriosos y poco intuitivos scripts de programación.

En distintos campos de la ciencia y la ingeniería en los que se dispone de grandes bases de datos se están desarrollando herramientas para extraer información de los mismos. Estas herramientas incluyen entornos de visualización altamente personalizables e interactivos, que permiten presentar los datos en diferentes formatos, de forma que se puedan identificar determinados patrones y datos erróneos. Esto ha motivado a investigadores y profesionales a utilizar modelos predictivos basados en *machine learning* (ML) para la evaluación de los datos, como lo demuestra la creciente cantidad de publicaciones científicas en este campo (Salazar et al., 2017a).

Por otra parte, las mediciones obtenidas de las redes de monitorización, ya sean manuales o automáticas, suelen presentar errores que pueden deberse a la transcripción o a la comunicación, así como períodos de falta de datos debido a interrupciones del sistema de origen diverso. Además, las antiguas series registradas manualmente y los datos digitalizados más recientes obtenidos de los sistemas automáticos coexisten frecuentemente en las mismas bases de datos. La frecuencia de lectura y la calidad de estos datos son, en general, diferentes, lo que da lugar a series heterogéneas que requieren acciones de preprocesamiento para ser estandarizadas. Por todo esto, el pretratamiento es esencial para extraer la máxima información de los datos disponibles.

En esta comunicación se presentan dos aplicaciones informáticas desarrolladas en el lenguaje de programación R (Team and R Development Core Team, 2018) que hacen uso de la interactividad del paquete Shiny (Chang et al., 2018) y pueden ser ejecutadas localmente o alojadas en la nube con un acceso seguro adecuado. El objetivo principal de la primera aplicación es el tratamiento de datos de monitorización e incluye funcionalidades específicas de preproceso enfocadas a la generación de una base de datos adecuada para ajustar posteriormente modelos predictivos. La primera versión fue desarrollada específicamente para analizar el comportamiento de presas, pero también podría aplicarse, con pequeños cambios, para estudiar otros procesos hidroambientales como la detección de fugas en las redes de distribución de agua o la estimación del flujo de descarga en los aliviaderos sobre la base de resultados experimentales.

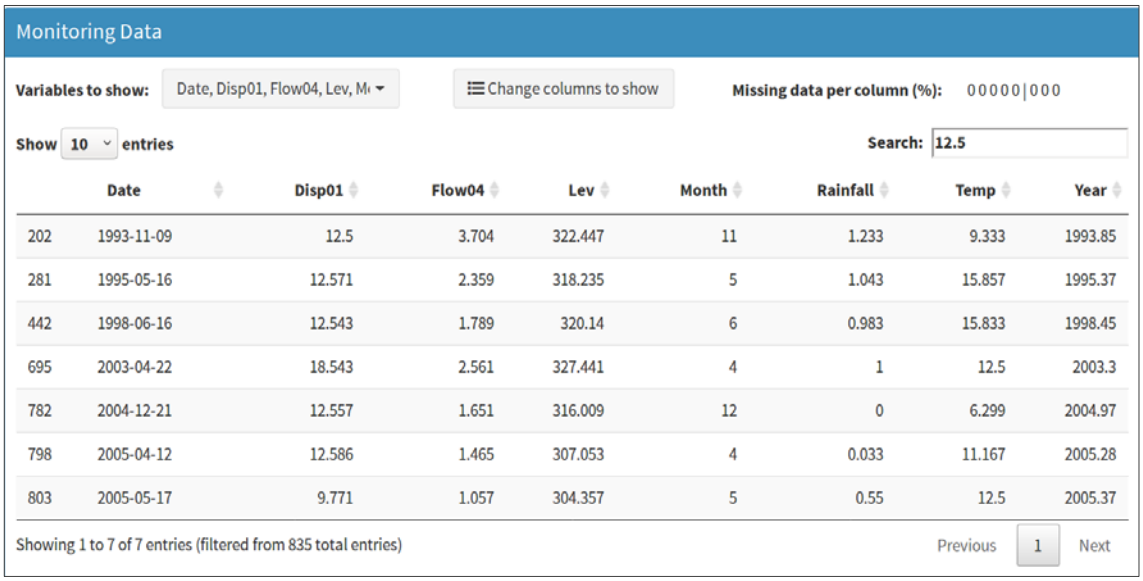
La segunda herramienta permite crear modelos predictivos del comportamiento del sistema. Las técnicas de ML se utilizan cada vez más con este fin en todo el mundo: se construyen modelos para estimar la respuesta de la estructura frente a una determinada combinación de cargas, y sus resultados se comparan con las mediciones reales para apoyar la toma de decisiones en las evaluaciones de seguridad. Además, la interpretación de estos modelos puede ser útil para el diseño del sistema de monitorización: el algoritmo calcula automáticamente la influencia relativa de las entradas (en este caso, los dispositivos de monitorización) en cuanto a su utilidad para estimar la respuesta, de forma que los más relevantes pueden ser seleccionados para ser automatizados. La flexibilidad del algoritmo permite analizar diferentes tipos de variables sin necesidad de determinar a priori cuáles son las cargas más influyentes o cómo afectan al valor objetivo.

TRATAMIENTO DE DATOS

Exploración de datos

La exploración de datos es un paso preliminar fundamental antes de cualquier análisis. Por un lado, permite conocer las características de los datos a analizar, como volumen disponible de datos, rangos de variación o la relación entre variables. Por otra parte, es útil para identificar errores tales como datos anómalos derivados de errores de medición o períodos de datos incompletos. Además, un experto con las herramientas adecuadas puede tener una primera idea de si el comportamiento responde a lo que se esperaba, observando cambios claros en las tendencias de las series, dispersión de los resultados, etc. La

herramienta desarrollada en este trabajo permite la visualización de tablas de datos interactivas (Figura 1) donde el usuario puede ver los valores ordenados por tiempo, pudiendo elegir las variables y el número de filas a mostrar, así como buscar un valor específico.



	Date	Disp01	Flow04	Lev	Month	Rainfall	Temp	Year
202	1993-11-09	12.5	3.704	322.447	11	1.233	9.333	1993.85
281	1995-05-16	12.571	2.359	318.235	5	1.043	15.857	1995.37
442	1998-06-16	12.543	1.789	320.14	6	0.983	15.833	1998.45
695	2003-04-22	18.543	2.561	327.441	4	1	12.5	2003.3
782	2004-12-21	12.557	1.651	316.009	12	0	6.299	2004.97
798	2005-04-12	12.586	1.465	307.053	4	0.033	11.167	2005.28
803	2005-05-17	9.771	1.057	304.357	5	0.55	12.5	2005.37

Figura 1 | Exploración de datos en tabla. Se muestra el resultado de la búsqueda del valor “12.5”.

La segunda opción de visualización es un gráfico de series de tiempo multivariable (Figura 2). A partir de la exploración de las series temporales, es posible identificar la existencia de errores de lectura o de períodos en los que faltan datos. En este gráfico, el usuario puede seleccionar las variables a representar mediante dos ejes independientes. Esto permite su comparación en una sola pantalla cuando su rango de variación es muy diferente. El gráfico incluye interactividad para explorar con detalle determinados periodos y mostrar los valores de cada variable en instantes específicos.

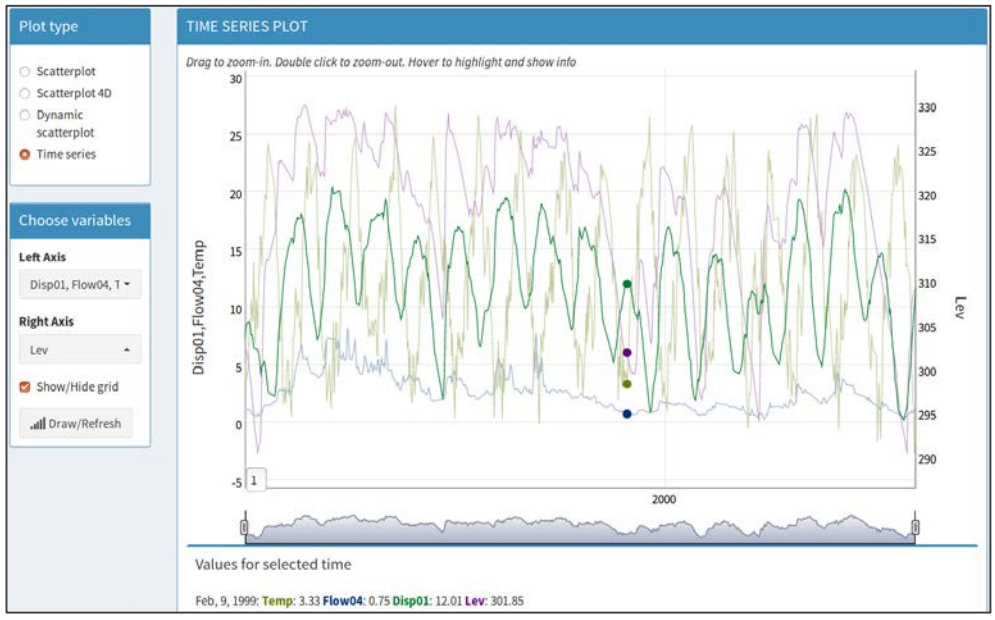


Figura 2 | Gráfico de series temporales con tres variables en el eje izquierdo y dos en el derecho.

La relación entre variables, así como su evolución en el tiempo, puede observarse mejor en una gráfica de dispersión cuasi-3D, en la que el usuario selecciona qué variables mostrar en cada eje, así como una tercera variable que permite ajustar el

color de los puntos en función de algunas de las variables del conjunto de datos. También es posible elegir los límites de la tercera variable; esta funcionalidad es muy útil para analizar la relación de dos variables durante períodos de tiempo específicos. Aquí, usamos *Year* para los colores (Figura 3 inferior); debe tenerse en cuenta que la variable *Year* se genera automáticamente como real (con dos posiciones decimales), no como un número entero. Este gráfico tiene funcionalidades para hacer zoom en los dos ejes, para un mejor análisis. Cuando se selecciona un punto específico en el gráfico, se muestra una tabla con los valores de la variable y todas las otras variables seleccionadas.

También se añade la posibilidad de gráficos de dispersión 3D, que permiten seleccionar las variables en cada uno de los tres ejes, con posibilidad de rotar el sistema de coordenadas y mostrar lo valores de cada variable en cada punto de forma interactiva. Un ejemplo de esto se puede ver en la Figura 3 superior.

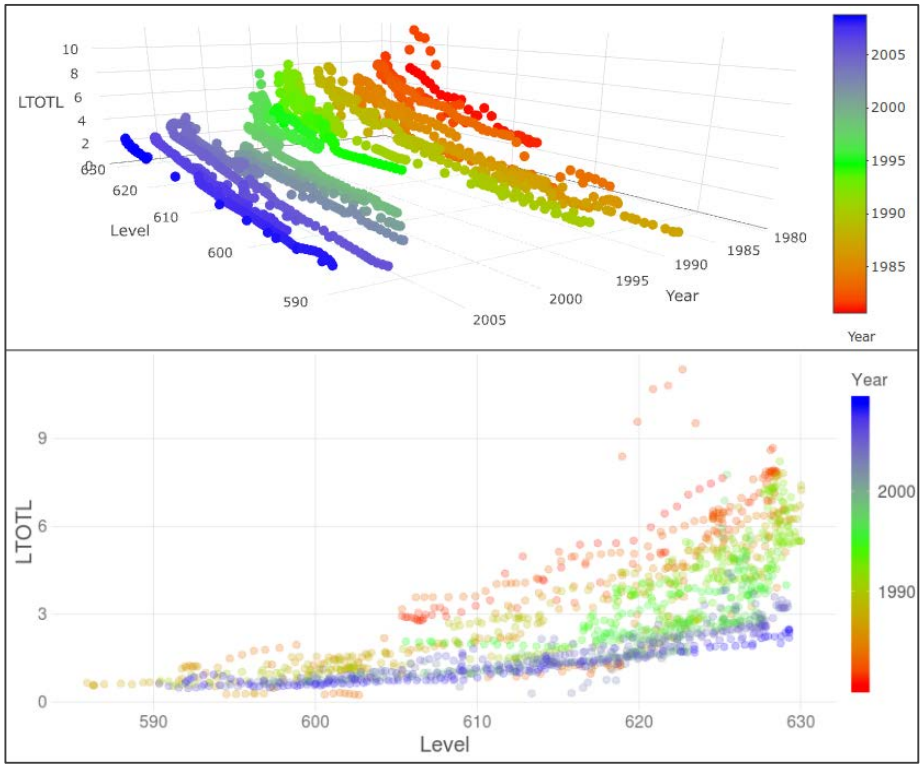


Figura 3 | Gráficos de dispersión para tres variables. En este ejemplo, los colores dependen de la variable *Year*, por lo que se puede observar la evolución temporal de la relación entre las variables en ambos ejes.

Otra opción de visualización es el gráfico de dispersión dinámico, donde se observa la evolución en el tiempo de varios sensores de una misma familia, lo que permite detectar cambios relativos entre las variables elegidas. Esta evolución se puede mostrar con respecto al tiempo y una variable seleccionada (eje horizontal) o el tiempo y dos variables seleccionadas (eje horizontal y tamaños).

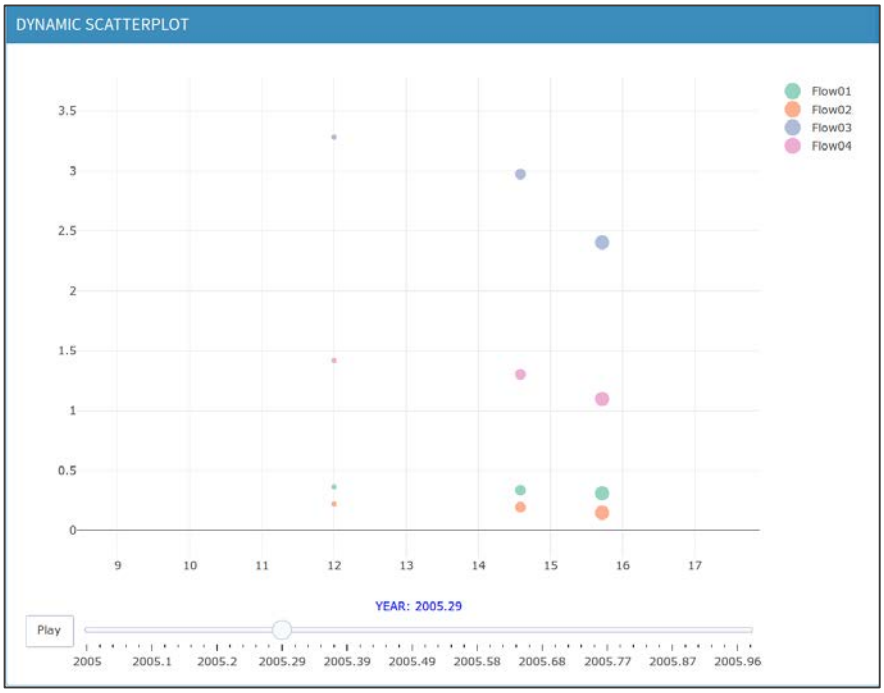


Figura 4 | Gráfico de dispersión dinámico. Evolución temporal de cuatro medidas de flujo con temperatura en el eje horizontal.

Por supuesto, cada variable puede en principio depender de muchas otras variables del sistema. Si ese es el caso, esta interpretación visual de las gráficas podría ser errónea. Para investigar estas relaciones puede hacerse uso de técnicas *machine learning*.

Reparación de datos: Datos incompletos

Es frecuente la existencia de periodos con datos ausentes en las series de medidas de auscultación, lo que limita la aplicación de determinados modelos de predicción. Varios métodos pueden ser útiles para compensar la falta de datos en circunstancias específicas, aunque la imputación de una alta proporción de datos puede conducir a un problema relevante, ya que ponen en duda la credibilidad de las conclusiones extraídas (Little et al., 2012). Cuando el análisis de variables con datos ausentes se considera inapropiado, es necesario considerar otras alternativas, como las siguientes:

A. Exclusión de las variables con datos ausentes del análisis.
B. Métodos de sustitución simples: cada valor que falta se rellena con un valor específico de la misma serie, como la observación anterior o la siguiente;
C. Métodos de estimación cuando la simple sustitución puede conducir a efectos sesgados (Molnar et al., 2009):
   a) Basado en la misma serie de datos (por ejemplo, el valor medio de un período determinado);
   b) Basado en series de datos de la misma naturaleza (es decir, medición de otros dispositivos de vigilancia del mismo fenómeno);
   c) Basado en series de datos de variables de diferente naturaleza, pero con una alta correlación.

La idoneidad de cada método depende de las causas que llevaron a la falta de datos y de la naturaleza de la variable considerada. Los datos que faltan pueden ser completamente aleatorios, es decir, no relacionados con las variables del estudio, o pueden mantener algún tipo de relación con el resto de datos de esa variable o alguna otra. Por ejemplo, la ausencia de datos de lluvia durante el invierno no debe ser reemplazada por los anteriores (otoño) ni ser ignorada por el cálculo de la media del resto del año, ya que implicaría ignorar la estación de lluvias. El estudio de las razones que subyacen a los datos que faltan es útil para seleccionar un método de imputación apropiado. Por ejemplo, los datos que faltan pueden estimarse a partir de alguna

variable relacionada. Esta técnica resulta en un sesgo más bajo que los métodos de análisis tales como la imputación múltiple o las ecuaciones de estimación (Little et al., 2012).

Tras el estudio de casos reales, se ha desarrollado una funcionalidad específica para corregir interactivamente estas imperfecciones seleccionando datos erróneos o períodos sin datos y sustituyéndolos por valores adecuados. Para ello se han implementado en el software algunos de los procedimientos más convencionales de imputación de datos:

A. Interpolación lineal entre los valores anterior y posterior;
B. Interpolación de una parábola basada en el valor correcto anterior más cercano y los dos posteriores;
C. Sustitución de cada uno de los puntos seleccionados por la media de los valores registrados el mismo día natural (u hora) del año (o día) anterior y posterior. Esto puede ser admisible en el caso de datos con estacionalidad anual (o diaria);
D. Sustitución de todos los puntos seleccionados por un valor fijo elegido por el usuario.

En la Figura 5 se muestra un ejemplo de introducción de valores ausentes mediante interpolación lineal.

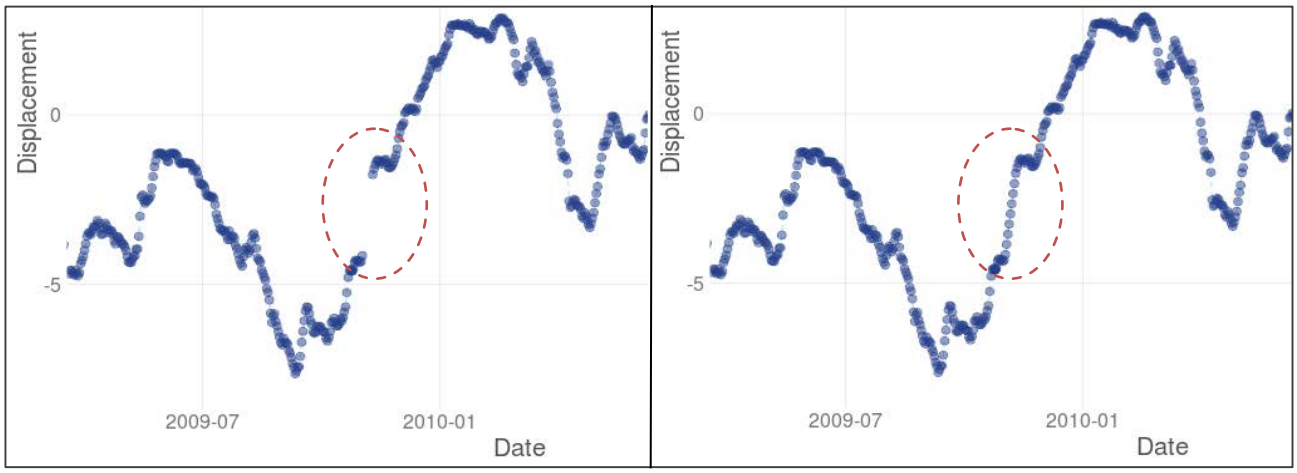


Figura 5 | Datos originales (izquierda). Introducción de datos mediante interpolación lineal (derecha).

Reparación de datos: Limpieza de errores

El proceso de limpieza de datos consiste en identificar datos corruptos, incorrectos o irrelevantes de una base de datos para luego reemplazarlos o eliminarlos (Wu, 2013). Estas inconsistencias detectadas o eliminadas pueden haber sido causadas originalmente por la entrada incorrecta de datos, por fallos en los sensores de medición, errores de transmisión o almacenamiento, o por diferentes definiciones de la misma variable. La identificación de los valores atípicos puede ser estricta (como rechazar cualquier valor que se encuentre fuera de un determinado rango) o difusa (como corregir registros que se encuentren dentro del rango global pero que supongan una variación local por encima de un determinado valor). El análisis de gráficas de dispersión es útil para observar cualquier valor atípico significativo en un conjunto de datos.

Esta funcionalidad también se ha desarrollado tras haber observado muchas imperfecciones en las bases de datos analizadas, que proceden de los sistemas de monitorización de estructuras hidráulicas. En la versión actual, además de las opciones ya mencionadas para los datos ausentes, el usuario puede sumar un valor fijo al período seleccionado o borrar determinados valores. La figura 6 muestra un ejemplo en el que parece claro que el error se corrige añadiendo un valor fijo.

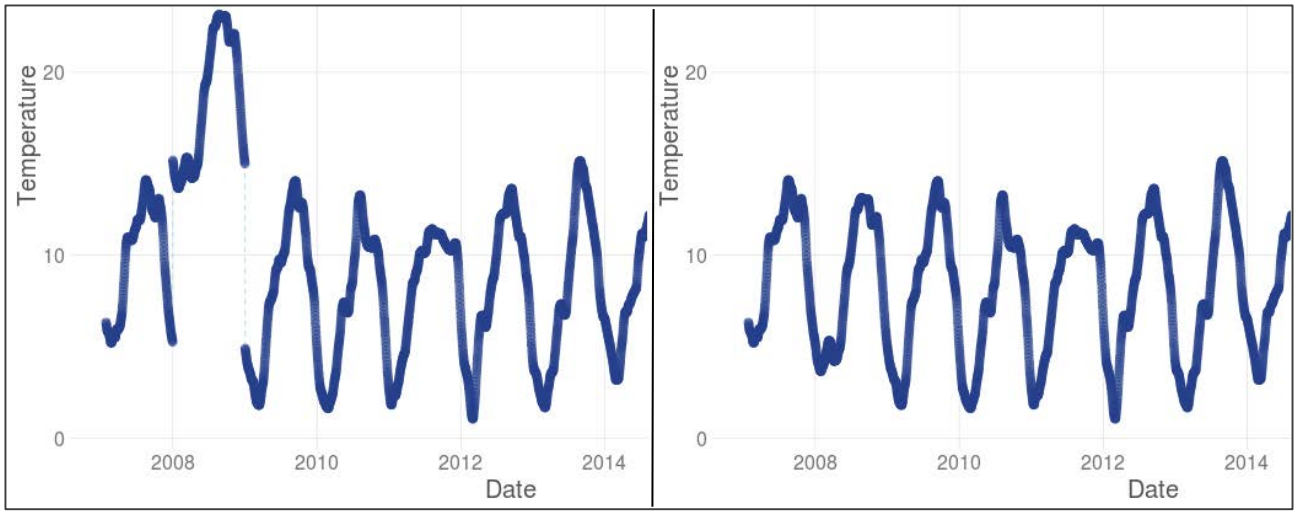


Figura 6 | Datos originales (izquierda). Corrección de errores mediante la adición de un valor fijo (derecha).

Generación de variables derivadas

Una vez realizada la exploración de los datos y corregidos los valores inexactos, y antes de ajustar los modelos predictivos, se pueden generar variables derivadas de los registros brutos para enriquecer el conjunto de entradas. La decisión de qué variables añadir corresponde al usuario experto, y depende del caso de estudio y del algoritmo a utilizar. Si el usuario desea analizar la evolución temporal del comportamiento del sistema, es necesario considerar explícitamente la variable tiempo. La aplicación automáticamente genera dos nuevas variables a partir de los registros de fecha: una categórica con los meses, para permitir analizar la estacionalidad de los datos, y una numérica con los años (añadiendo la parte decimal) para facilitar el análisis visual de la evolución de las variables con el tiempo.

El objetivo principal al añadir nuevas variables es obtener automáticamente nuevas características basadas en las existentes. Por ejemplo, las medias móviles son fáciles de realizar y ayudan a mejorar la relación señal/ruido de la serie de datos al sustituir cada valor de la serie por la media de un número determinado de valores anteriores (Dilawari, 2018).

Así, se incluye la posibilidad de añadir medias móviles, sumas acumuladas o combinaciones lineales de las variables existentes. Las dos primeras operaciones permiten crear variables derivadas que reflejan la principal evolución de la variable base a lo largo del tiempo, ocultando variaciones de ruido que pueden no ser relevantes.

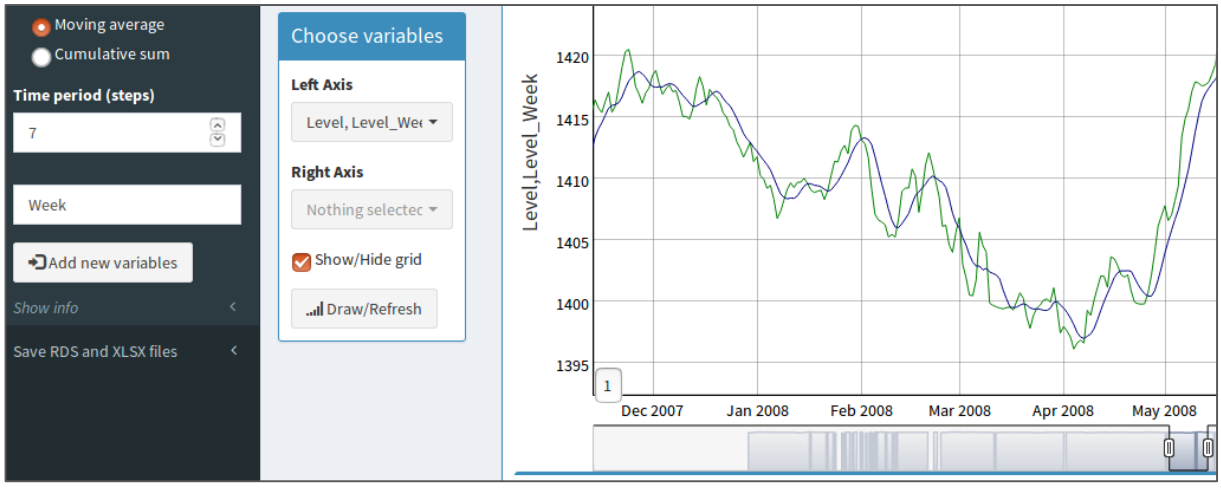


Figura 7 | Series temporales de una variable y su media móvil semanal.

Reducción de la base de datos

Por otra parte, el uso de bases de datos con una alta proporción de información o ruido irrelevante y redundante es menos eficiente para construir modelos predictivos. La preparación y el filtrado de los datos pueden ahorrar tiempo para su posterior procesamiento y ofrecer mejores resultados. Por ejemplo, además de eliminar variables o períodos de tiempo, la aplicación permite que las series registradas con una frecuencia de adquisición variable puedan reducirse a un único valor. Esta reducción se puede hacer en base a distintos períodos de tiempo (día, semana, quincena o mes) con los siguientes procedimientos:

a) Tomando como valor diario el registrado a una hora determinada del día (solo disponible en reducción a días);
b) Tomando el valor diario máximo o mínimo para cada período;
c) Agregando valores horarios, bien calculando el promedio o la suma de las mediciones totales.

ANÁLISIS DEL COMPORTAMIENTO DEL SISTEMA BASADO EN MACHINE LEARNING

Metodología

Es común utilizar modelos estadísticos para generar predicciones de ciertas variables de respuesta de estructuras hidráulicas y analizar la contribución de cada una de las cargas actuantes. Para el estudio del comportamiento de presas, uno de los ejemplos más utilizados es el modelo HST, desarrollado para analizar desplazamientos en presas de hormigón (Willm, 1967). Este método se ha utilizado a menudo tanto en la práctica profesional como en la investigación, aunque también se han identificado sus limitaciones (Salazar et al 2017a). Se han propuesto un buen número de alternativas para superar estas limitaciones del HST, que van desde modelos estadísticos avanzados, como HST-Grad (Tatin et al., 2015) o híbrido (Perner and Obernuber, 2009), hasta modelos ML, que se construyen exclusivamente a partir de los datos: por ejemplo, *Neural Networks* (De Granrut et al., 2019; Mata, 2011).

Los autores de esta comunicación han utilizado previamente algoritmos de *machine learning* para generar modelos predictivos del comportamiento de estructuras hidráulicas. En trabajos anteriores se compararon diferentes algoritmos en términos de precisión y facilidad de implementación (Salazar et al., 2015); se analizaron las posibilidades de interpretación de los modelos para extraer conclusiones sobre el comportamiento del sistema (Salazar et al., 2016) y se propuso una metodología para la aplicación a la detección de anomalías (Salazar et al., 2017b). Otras aplicaciones incluyen la estimación de la capacidad de descarga de aliviaderos con compuertas (Salazar et al., 2013) y en laberintos (Salazar y Crookston, 2019). Este conocimiento previo se ha aplicado al desarrollo de una segunda aplicación que permite utilizar los datos procesados para ajustar modelos de predicción, así como interpretar la respuesta del sistema a acciones o cargas en operación.

Habitualmente los modelos predictivos basados en datos se ajustan a parte de los datos disponibles (el conjunto de entrenamiento) y luego el modelo se aplica para predecir la respuesta para el período restante (el conjunto de pruebas o validación). La precisión de la predicción se mide comparando las predicciones obtenidas con el modelo con las lecturas reales. El objetivo principal de estos enfoques es la detección precoz de anomalías, para lo cual normalmente se establece algún umbral, de modo que si la desviación de la lectura real de la predicción del modelo es mayor que el umbral, se emite algún aviso.

Aunque los algoritmos ML se consideran a menudo como modelos de "caja negra", existen algunas herramientas disponibles para su análisis, que han demostrado ser útiles para comprender el comportamiento de la presa (Mata, 2011; De Granrut et al., 2019; Tinoco et al., 2018; Salazar et al., 2016). Esta interpretación debe basarse en el conocimiento del comportamiento histórico de las variables en diferentes situaciones.

Se ha creado una segunda aplicación que realiza un análisis de datos mediante los siguientes pasos:

a) Generación de un modelo predictivo basado en ML.
b) Interpretación del modelo en términos de los parámetros de entrada más influyentes y la naturaleza de su efecto sobre la variable considerada.

Generación del modelo

En este trabajo, utilizamos *Boosted Regression Trees* (BRT) (Elith et al., 2008), un algoritmo que demostró ser el más ventajoso en un estudio comparativo sobre detección temprana de anomalías (Salazar et al., 2015) y que ya se utilizó para el análisis del comportamiento de una presa arco (Salazar et al., 2017b). Las principales características de este algoritmo que lo hacen apropiado para este problema son las siguientes:

A. Permite considerar variables de diferente naturaleza y rango de variación sin necesidad de transformaciones adicionales.
B. Selecciona automáticamente las variables más relevantes en la respuesta de la estructura y descarta aquellas con poca influencia. Por lo tanto, no es necesario hacer una selección de variables previa.
C. Es robusto con respecto a parámetros de entrenamiento, a diferencia de otros modelos ML que requieren un conocimiento profundo del algoritmo para seleccionar cuidadosamente las opciones de ajuste del modelo.

La interfaz desarrolladla permite al usuario seleccionar las variables predictoras y la variable objetivo, así como los valores de los parámetros de entrenamiento (*Shinrkage, Number of trees, Interaction depth, Bag fraction*), aunque los valores por defecto suelen dar buenos resultados. También es posible calcular y comparar varios modelos con las mismas condiciones para verificar la influencia del componente aleatorio del algoritmo de entrenamiento. La principal decisión que se debe tomar en este paso es la selección del período de entrenamiento, es decir, los datos que se utilizarán para el ajuste de modelo; el período complementario se reserva para validación. Esto es importante para controlar el sobreajuste de los resultados (Lever et al., 2016).

Con este enfoque, la fiabilidad de las conclusiones obtenidas está relacionada con la precisión de las predicciones del modelo (Breiman, 2001). Por lo tanto, se calcula y muestra la discrepancia entre las predicciones y las observaciones, tanto para el periodo de entrenamiento como para el de validación. La aplicación permite analizar los resultados en cuanto a la precisión del modelo, mostrando los valores de error medio absoluto (MAE) y del coeficiente de determinación (R^2) y su evolución en el tiempo (Figura 8).

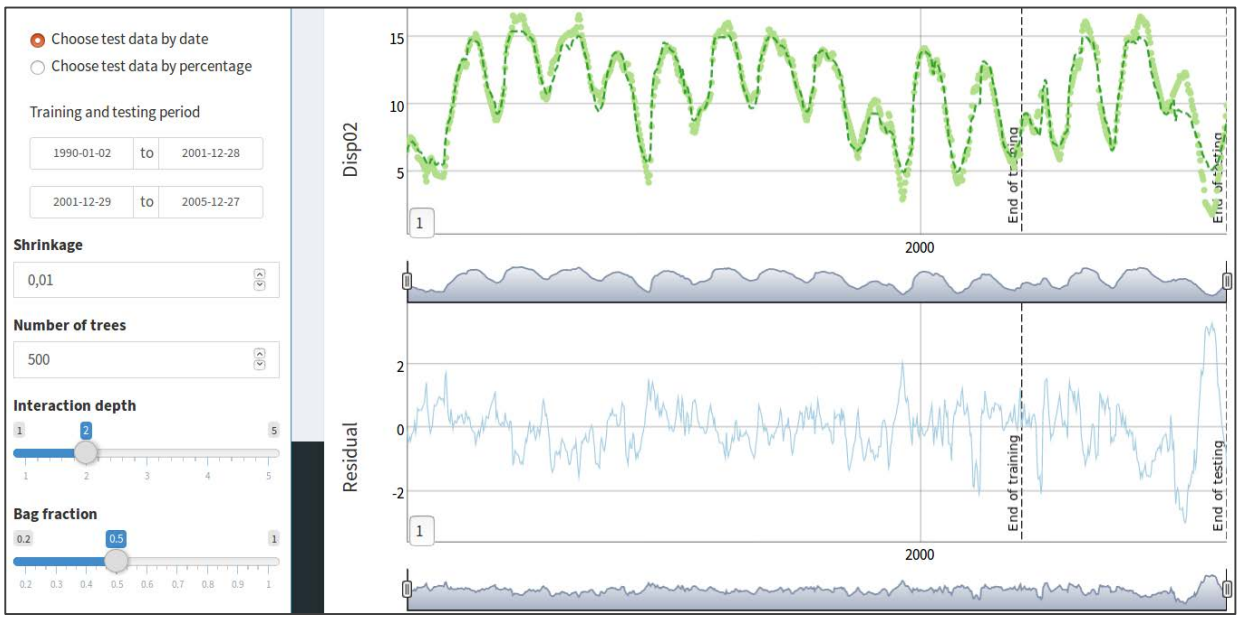


Figura 8 | Interfaz para ajustar y evaluar modelos predictivos.

Por defecto, se toma un periodo previo para entrenamiento del modelo, y el más reciente para validación. Este enfoque representa una aplicación realista, en la que datos existentes pueden utilizarse para construir un modelo, que puede aplicarse posteriormente a la evaluación de la seguridad en tiempo real. En la práctica, el modelo podría actualizarse con cierta frecuencia para ampliar el conjunto de entrenamiento y, por lo tanto, mejorar la precisión. El efecto del tamaño del conjunto de

entrenamiento sobre la capacidad de predicción se estudió en un trabajo previo, incluyendo criterios para actualizar el modelo (Salazar et al., 2017b).

Interpretación de resultados

Las Figuras 9 y 10 muestran un ejemplo de aplicación, en este caso para predecir el desplazamiento en una presa de hormigón. El modelo parte de las variables externas, que en este caso incluyen el nivel de embalse, la temperatura (valores diarios y medias móviles) y la precipitación acumulada en varios periodos. El primer resultado de la interpretación del modelo es la identificación de las variables de entrada más influyentes en el sistema, es decir, aquellas que están más relacionadas con la respuesta analizada. Dado que es habitual tener varias variables de una misma familia (por ejemplo medias móviles de diversos periodos de una misma variable), puede estudiarse también la importancia por familias. En el ejemplo mostrado se observa que la temperatura es más influyente, y en particular la media móvil de 60 días.

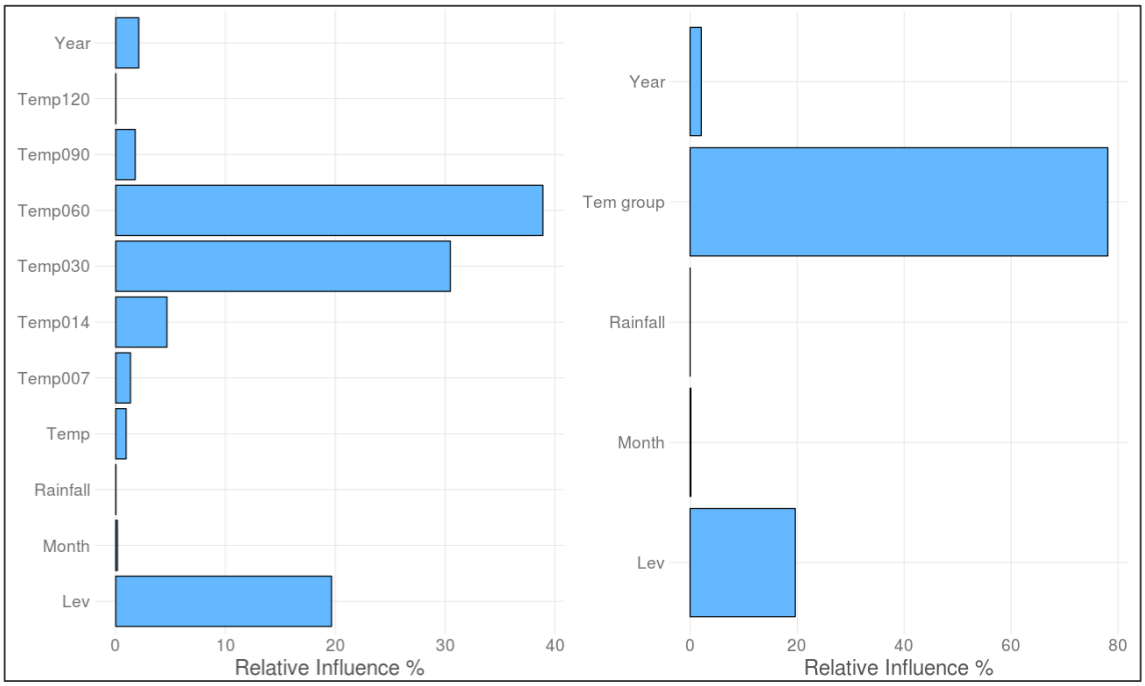


Figura 9 | Influencia de las variables predictoras en la respuesta.

Del mismo modo se puede estudiar la influencia parcial de cada una de las variables de entrada con respecto a la respuesta, individualmente o emparejadas. En la Figura 10 se incluye otro ejemplo, en este caso para estimar el caudal de filtración. Se observa que aumenta con el nivel de embalse, además de la evolución en el tiempo.

La influencia relativa calculada explica la interacción entre las variables: una pequeña diferencia en alguna de ellas puede dar como resultado una gran diferencia en la influencia calculada por el modelo. Esto muestra la capacidad del algoritmo para manejar variables altamente correlacionadas, como es el caso de las distintas medias móviles de temperatura. La interpretación del modelo permite identificar efectos que son difíciles de detectar mediante la exploración de datos, incluso si se han utilizado herramientas de visualización avanzadas.

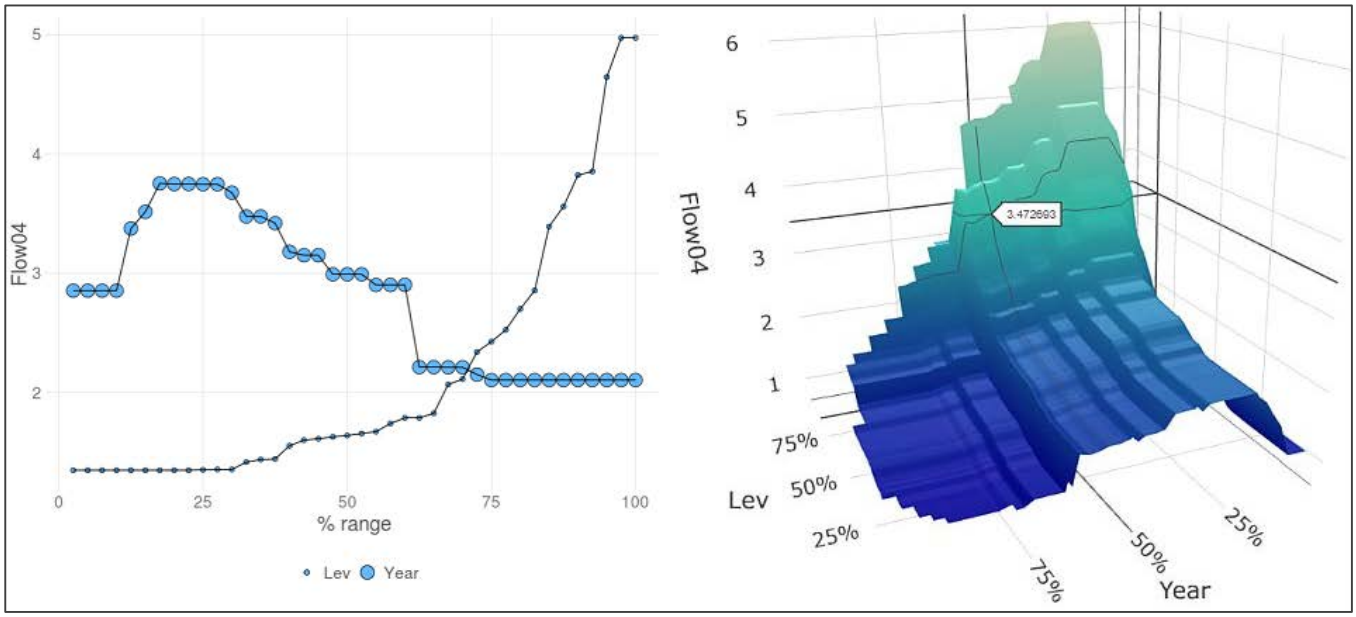


Figura 10 | Gráficos de dependencia parcial para dos predictores con respecto a la variable objetivo: líneas (izquierda) y superficie 3D (derecha)..

El modelo tiene una limitación fundamental que debe tenerse siempre en cuenta: en general, las predicciones fuera del rango de datos de entrenamiento no son fiables. Esto implica, por ejemplo, que la predicción del modelo será pobre para un nivel de embalse superior al máximo histórico registrado. Para controlar este efecto de extrapolación, se verifica si los datos de la prueba están dentro del rango de entrenamiento, y se emite una advertencia en caso de extrapolación.

CONCLUSIONES

Con la implementación de nuevos sistemas avanzados de monitorización y adquisición de datos, se generan bases de datos de tamaño creciente, lo que puede ser de gran utilidad tanto en el análisis de la respuesta como en la gestión de la seguridad. Estos avances en los instrumentos de medición, así como en las técnicas de transmisión y almacenamiento de información, han conducido a mejoras significativas en términos de precisión y fiabilidad, lo que se ha traducido en una mayor disponibilidad de información sobre el comportamiento del sistema en cuestión. La exploración de estos datos puede proporcionar información interesante, especialmente si se dispone de las herramientas adecuadas y la tarea es llevada a cabo por técnicos de ingeniería con un alto conocimiento del caso en estudio. Sin embargo, el comportamiento del problema es a menudo complejo, y las cargas actuantes están correlacionadas, lo que hace casi imposible identificar ciertos efectos simplemente observando los datos.

Se han presentado dos herramientas para el análisis e interpretación de datos de monitorización de estructuras hidráulicas. Una primera permite la exploración visual interactiva de los datos de monitorización en diferentes formatos con un alto grado de interactividad. Se pueden generar y analizar diferentes tipos de gráficos de dispersión y de series temporales. Además, los datos pueden corregirse y completarse en el mismo entorno y también de forma interactiva.

La segunda aplicación permite generar modelos de predicción basados en ML, mediante una implementación de modelos BRT que se adapta a variables de respuesta de diferente naturaleza. Estos modelos tienen la ventaja de ser muy flexibles en cuanto a la naturaleza de la respuesta del problema a analizar, así como a la cantidad de datos de entrada. Las variables de entrada de baja relevancia se excluyen automáticamente sin necesidad de una selección específica de variables.

Los modelos de predicción basados en BRT pueden ser útiles para interpretar el comportamiento de diferentes sistemas mediante el análisis de los datos de monitorización. Los estudios previos han mostrado buena capacidad de predicción del algoritmo, así como su robustez en relación con los parámetros de entrenamiento, por lo que no es necesario tener un conocimiento profundo del método para su uso práctico.

Como resultado, los modelos BRT pueden ser utilizados tanto para identificar cambios en el comportamiento del sistema, cuando se analiza el comportamiento pasado (análisis de datos retrospectivo), como para detectar desviaciones del funcionamiento normal, cuando se aplican a datos en tiempo real.

Además, estos modelos pueden ser analizados para explorar el grado de asociación entre las variables de entrada consideradas y la variable objetivo analizada, así como la forma de cada dependencia parcial.

Los resultados muestran que el método puede ser útil como apoyo para el análisis de la seguridad de estructuras hidráulicas, permitiendo la identificación de desviaciones del comportamiento normal. La herramienta puede ser utilizada para diferentes tipos de problemas y variables de salida, siempre que se disponga de datos de monitorización adecuados y los resultados sean interpretados por usuarios expertos en el sistema analizado.

AGRADECIMIENTOS

Los autores agradecen el apoyo financiero de CIMNE a través del Programa CERCA/Generalitat de Catalunya. También agradecen el apoyo financiero del Ministerio de Ciencia, Innovación y Universidades a través del proyecto TRISTAN (RTI2018-094785-B-I00), de la convocatoria 2018 de proyectos I+D+i «Retos Investigación» del programa estatal de I+D+i orientada a los Retos de la Sociedad.

REFERENCIAS

Breiman, L. 2001. Statistical Modeling: The Two Cultures. Statistical Science, 16(3), 199–231. #

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.

De 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, 178, 616-634.

Dilawari, M. 2018. Forecasting models for the displacements and the piezometer levels in a concrete arch dam.

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

Guyon, I., Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, 3(3), 1157–1182. Special Issue on Variable and Feature Selection

Kotsiantis, S., Kanellopoulos, D., P. P. 2006. Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, 1(2), 111–117.

Lever, J., Krzywinski, M. & Altman, N. Points of significance: Model selection and overfitting. *Nat. Methods* **13**, 703–704 (2016).

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., … Stern, H. 2012. The Prevention and Treatment of Missing Data in Clinical Trials. New England Journal of Medicine, 367(14), 1355–1360.

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

Molnar, F. J., Man-Son-Hing, M., Hutton, B., & Fergusson, D. A. 2009. Have last-observation-carried-forward analyses caused us to favour more toxic dementia therapies over less toxic alternatives? A systematic review. Open Medicine, 3(2), 1–20.

Mora, J., López, E., de Cea, J. C. 2008. Estandarización en la gestión de datos de auscultación e informe anual en presas de titularidad estatal. VIII Jornadas Españolas de Presas.

Perner, F. & Obernuber, P. 2009. Analysis of arch dam deformations. 2nd International conference for long term behaviour of dams. Graz

Salazar, F., Crookston, B. M. 2019. A performance comparison of machine learning algorithms for arced labyrinth spillways. Water. Special Issue "Machine Learning Applied to Hydraulic and Hydrological Modelling."

Salazar, F., Morán, R., Rossi, R., & Oñate, E. 2013. Analysis of the discharge capacity of radial-gated spillways using CFD and ANN - Oliana Dam case study. Journal of Hydraulic Research, 51(3), 244–252.

Salazar, F., Morán, R., Toledo, M. Á., & Oñate, E. 2017a. 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.

Salazar, F., Toledo, M. Á., González, J. M., & Oñate, E. 2017b. Early detection of anomalies in dam performance: A methodology based on boosted regression trees. *Structural Control and Health Monitoring*, *24*(11), e2012.

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.

Tatin, M., Briffaut, M., Dufour, F., Simon, A., & Fabre, J. P. 2015. Thermal displacements of concrete dams: Accounting for water temperature in statistical models. Engineering Structures, 91, 26-39.

Team, R. D. C., R Development Core Team, R. 2018. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (Vol. 1)*. Vienna.

Thorvaldsdóttir, H., Robinson, J. T., Mesirov, J. P. 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), 178–192.

Tinoco, J. A. B., Granrut, M. D., Dias, D., Miranda, T. F., &Simon, A. G. 2018. Using soft computing tools for piezometric level prediction. In Third International Dam World Conference (pp. 1-10).

Wu, S. 2013. A review on coarse warranty data and analysis. *Reliability Engineering and System Safety*, 114(1), 1–11.

Willm, Beaujoint. 1967. Les méthodes de surveillance des barrages au service de la production hydraulique d'Electricité de France, problèmes anciens et solutions nouvelles. IXth Int. Congr. Large Dams, Istanbul; p. 529–50.