Validación de datos hidrológicos con redes neuronales artificiales. Aplicación a niveles en ríos

A. Madueño [1], M. López[2], J. Estévez[3], J.V. Giráldez[4]

[1] Dpto. Ingeniería Aeroespacial y Mecánica de Fluidos, Universidad de Sevilla, Ctra. De Utrera, Km.1, 41013 Sevilla, amadueno@us.es

[2] Dpto. Ingeniería del Diseño, Universidad de Sevilla, mlopezlineros@us.es

[3] Dpto. Ingeniería Rural, Universidad de Córdoba, ma2esguj@uco.es

[4] Dpto. de Agronomía, Universidad de Córdoba, ag1gicej@uco.es

Resumen

En el estudio hidrometeorológico la obtención de bases de datos de calidad es una herramienta esencial tanto para científicos como para responsables en la toma de decisiones en el ámbito operacional. La validación de los datos meteorológicos, garantiza la calidad de la información, identificación de valores incorrectos y detección de problemas que requieren la atención inmediata de los equipos de mantenimiento. El objetivo principal de este trabajo ha sido el desarrollo de un nuevo método de validación de datos de nivel en ríos basados en redes neuronales autorregresivas no lineales (NARNN). Para evaluar la eficiencia de este nuevo método se ha comparado los resultados con los métodos tradicionales, para ello se han introducidos errores artificialmente en las series de datos, resultando claramente más eficiente en la detección de datos el nuevo método con NARNN, detectando más del 90% de los errores introducidos, frente al 13% de los errores detectados por los métodos tradicionales.

Palabras clave: Analisis de datos, control de calidad, red neuronal autorregresiva.

Hydrological data validation using neural network. Applied in river stage data

Abstract

In the study of hydrological information, having a good quality in database is essential for research scientist as well as for decision- makers. The validation data ensure the quality of the base information, detecting incorrect records and problems that needs the immediate action by maintenance staffs. The main purpose of this work has been the develop of a new quality control method based on non-linear autoregressive neural networks (NARNN) for validating hydrological information of stage data level, for automatic detection of incorrect records. To assess the effectiveness of this new approach, a comparison with adapted conventional validation tests extensively used for hydro-meteorological data was carried out. A set of errors of different magnitudes was artificially introduced into the dataset to evaluate detection efficiency. The NARNN method detected more than 90%, while conventional tests detected only around 13%.

Keywords: Analysis of data, Neural network, Quality control, Non-linear autoregressive neural networks

Introducción y antecedentes.

El control de calidad es un requisito imprescindible y previo para el uso de la información hidrometeorológica. La obtención de bases de datos de calidad son herramientas esenciales tanto para científicos e ingenieros, como para responsables en la toma de decisiones. La validación de los datos meteorológicos, garantiza la calidad de la información, identificación de valores incorrectos y detección de problemas que requieren la atención inmediata de los equipos de mantenimiento. La aplicación de métodos de control de calidad

es especialmente necesario para datos hidrometeorológicos en tiempo real o casi en tiempo real para diferentes propósitos (WMO, 2008).
La información hidrometeorológica procedente de sensores suele contener datos erróneos, que dificultan su uso posterior, por ello se requiere personal especializado que proceda previamente a su revisión. Cuando se necesita este control en tiempo real, es deseable contar con un procedimiento previo automático que marque los datos potencialmente incorrectos para así reducir el análisis manual.
Han sido muchos los procedimientos desarrollados para para asegurar la calidad de datos procedentes de variables hidrometeorológicas, tales como la precipitación o variables climáticas de entrada (temperatura, humedad relativa, radiación solar, velocidad del viento) para la ecuación de la evapotranspiración de referencia. Sin embargo, la literatura relacionada con el control de calidad de datos en nivel de río es escasa.
Las series de datos hidrológicos normalmente son incompletas y consisten en una mezcla de señal y ruido. A esto se suma que el proceso hidrológico subyacente suele ser estocástico, lo cual hace más difícil la identificación de la señal. El objetivo del modelo es revelar el proceso subyacente de los datos. Este modelo es estimado a través de métodos estadísticos para encontrar regularidades y dependencias existentes en los datos.
El éxito de los modelos de redes neuronales (ANN), radica en su capacidad para aproximarse a cualquier función medible de Borel, con el grado de precisión que se desee, tal como indica Hornik et al. (1989). Las ANN se hicieron muy útiles en predicciones de series temporales debido a su capacidad de aprendizaje dentro de un gran volumen de datos potencialmente ruidosos. Destacando que en la mayoría de los casos, los modelos de predicción de las redes neuronales dan mejores resultados que otros modelos similares.
El primer paso para el diseño de una herramienta de control de calidad es el análisis de las principales fuentes de error. Este paso requiere el establecimiento de algunas normas para evaluar la calidad de los datos, incluyendo las sucesivas manipulaciones realizadas en los datos brutos adquiridos por los sensores. En el siguiente paso, los niveles de control se deben establecer en una revisión periódica de los datos. Estévez et al. (2011), propusieron una aplicación progresiva de las pruebas convencionales para el control de la información hidro-meteorológica. Hubbard et al. (2005), presentaron una serie de procedimientos de control de calidad basados en decisiones estadísticas. El sistema realiza cuatro análisis distintos: (i) cálculo de los umbrales según los distintos periodos del año; (ii) velocidad de cambio; (iii) persistencia estacional; y (iv) consistencia espacial. Esta última prueba se basa en un análisis de regresión lineal para la estimación de intervalos de confianza para cada una de las estaciones en cuestión respecto a la estación objeto. Los datos pueden estar clasificados en diferentes categorías según el método de control de calidad elegido.

Material y Métodos

Para la detección de errores dentro de una serie completa de datos, se ha utilizado datos procedentes de la cuenca del río Duero, el punto de control Villoldo, este es un punto localizado en la cuenca del río Carrión, un afluente del Duero en el Noroeste de la Península Ibérica. Esta estación pertenece en la actualidad al Sistema Automático de Información Hidrológica (SAIH) de la Confederación Hidrográfica del Duero, habiendo pertenecido, previa la existencia de este SAIH, a la Red Oficial Española de Aforos (ROEA). Con el fin de evaluar y comparar el porcentaje de registros erróneos detectados por los métodos tradicionales existentes hasta la fecha y los desarrollados con esta investigación, se han introducido errores conocidos en la serie inicial de datos diezminutales (comprendida entre

el 02/02/1999 al 20/07/2010 con un total de 602928 datos). Estas alteraciones introducidas en la serie original han sido errores aleatorios.

Métodos tradicionales para la validación de datos:

Han sido aplicados tres métodos tradicionales de control de calidad para su aplicación en la serie de datos con los errores aleatorios introducidos. Estos procedimientos son ampliamente utilizados en muchos trabajos, especialmente en el control de calidad de datos meteorológicos

Los principios básicos de las pruebas que se aplican a los datos registrados provienen de las tres reglas introducidas por Meek y Hatfield (1994), que están basadas en O'Brien y Keefer (1985). Estas reglas utilizan límites fijos o dinámicos para cada variable (lo que se conoce habitualmente como prueba de intervalo), límites fijos o dinámicos para los cambios entre observaciones sucesivas ("step test" o prueba de consistencia temporal) y, por último, límites para detectar medidas consecutivas iguales o de baja variabilidad (prueba de persistencia). Este último también se podría considerar como un intervalo de consistencia temporal.

Red neuronal no lineal auto-regresiva o NARNN por sus siglas en inglés:

Una estructura basada en una NARNN, es la base de esta investigación para el desarrollo de un nuevo método de validación. Este tipo de red incluye una serie de entradas (d términos de la propia salida) con retardo y(t-1), y(t-2), …y(t-d), una capa oculta de n neuronas con una función de activación sigmoidal y una capa de salida de una sola neurona con función de activación lineal (1). La NARNN se puede expresar matemáticamente como:

$$y(t) = \sum_{i=1}^{n} \beta_i \cdot \phi\left(\omega_{i0} + \sum_{j=1}^{d} \omega_{ij} . y(t-j)\right) + \beta_0 \qquad (1)$$

La función sigmoidal, φ, se emplea como función de activación. Los parámetros que incluyen son βi, ωij (pesos), y β0, ωi0 (sesgo).
Las condiciones iniciales son (y(0), y(1),…, y(d)), siendo el estado del sistema en el instante t (y(t),y(t+1),…, y(t+d)).
Los pesos se han ajustado mediante el algoritmo de Levenberg-Marquardt, por combinar el poco tiempo de cálculo y una alta eficiencia en la localización del mínimo de la función de error medio cuadrático (MSE).

Resultados y Discusión

Se ha comparado la capacidad de detección de errores entre los métodos tradicionales y la NARNN, empleando para ello los registros que van del dato 1001 al dato 6000 de la serie de datos estudiada en este trabajo. Los resultados obtenidos aparecen reflejados en la Tabla 2.1.
Por lo general, las pruebas tradicionales sólo detectan un máximo del 13% de los registros que han sido alterados en el caso del cociente de error mayor, e=1, y menos de un 6% para cocientes menores, e <1.
Por su parte el modelo propuesto basado en NARNN fue capaz de detectar todo tipo de errores independientemente de la magnitud de los mismos. Hay una ligera tendencia a la

baja con los valores de e, desde 453 errores detectados si e=1 a 387 errores detectados si e=0.2 con n=5 neuronas en la capa oculta.

Tabla 2.1.Resultados en la detección de errores: NARNN *versus* métodos tradicionales

Cociente de error	Errores introducidos	Errores detectados					
		Prueba de intervalo	Step Test	Prueba de Persistencia	Prueba tradicional	NN n=5, d=10, δ=0.08	NN n=1, d=1, δ=0.08
					Valor absoluto (%)		
1.0	*501*	*27*	*38*	*0*	*65 (12.97)*	*453 (90.4)*	*466 (93.0)*
0.8	*500*	*13*	*15*	*0*	*28 (5.60)*	*449 (89.8)*	*460 (92.0)*
0.6	*495*	*4*	*4*	*0*	*8 (1.62)*	*444 (89.7)*	*417 (84.2)*
0.4	*493*	*2*	*1*	*0*	*3 (0.61)*	*440 (89.3)*	*385 (78.1)*
0.2	*485*	*2*	*1*	*0*	*3 (0.62)*	*387 (79.8)*	*370 (76.3)*

NN: modelo NARNN; n= neuronas en la capa oculta; 1000= datos para el entrenamiento; 5000= datos para la validación; d= desfase de la retroalimentación; δ= umbral de selección.

Ha de destacarse la marcada diferencia en los cocientes 0.4 y 0.6.El método basado en la NARNN ha detectado entre un 89.6% y un 89.7% de las alteraciones introducidas en la serie, mientras que los métodos tradicionales no llegan al 2%.
De esta diferencia se deduce, la baja eficiencia de los métodos tradicionales en general y, especialmente en el caso de errores pequeños y medianos, lo que contrasta con el método NARNN que demuestra ser muy eficiente. Cuando el cociente de error es grande, el nuevo método puede detectar más del 90% de los registros alterados, mientras que las pruebas tradicionales sólo detectan alrededor del 13%.

Bibliografía

Estévez, J., Gavilán, P., Giráldez, J.V. 2011. Guidelines on validation procedures for meteorological data from automatic weather stations. Journal of. Hydrology. 402, 144-154.
Hubbard, K.G., Goddard, S., Sorensen, W.D., Wells, N., Osugi, T.T., 2005. Performance of quality assurance procedures for an applied climate information system. J.
Atmos. Ocean. Tech. 22, 105–112.
Hornik, K., Stinchcombe, M., White, H. 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359-366.
Meek, D.W., Hatfield, J.L., 1994. Data quality checking for single station meteorological databases. Agricultural and Forest Meteorology 36, 85–109.
O'Brien, K.J., Keefer, T.N. 1985. Real-time data verification: Computer applications in water resources. In Torno, H.C. (Ed). Proc. Specialty Conference ASCE, Buffalo, NY USA 764–770.
World Meteorological Organization, 2008. Guide to meteorological instruments and methods of observations. WMO-No.8, Ginebra.