aci avances en ciencias e ingenierías

Artículo/Article
Sección B/Section B
Vol. 16, nro. 1
e3274

Evaluación de métodos estadísticos y matemáticos para estimar datos pluviométricos faltantes en la microcuenca del río Pita, Pichincha, Ecuador

Santiago Bonilla-Cáceres[1]*, Teresa Alejandra Palacios Cabrera[1]

[1]Universidad Central del Ecuador, Facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental, Carrera de Ingeniería Ambiental, Quito, Ecuador.

*Corresponding author's name, email: osbonilla@uce.edu.ec

Evaluation of statistical and mathematical methods to estimate missing pluviometric data in the microbasin of the Pita river, Pichincha, Ecuador

Abstract

The absence of time series data on meteorological variables is a drawback in environmental sciences, especially with regard to precipitation, which is a key variable in several fields. Now, the present study aimed to compare several statistical and mathematical methods to generate missing pluviometric data in the microbasin of the Pita River, such as the Paulhus and Kohler method, multiple linear regression (MLR), Wavelet transform and artificial neural networks, using information from the hydrometeorological network of the Fund for Water Protection (FONAG) of Quito. The artificial neural networks were highly effective in generating pluviometric data in the study area, with coefficients of determination (R^2) higher than 0.64; and root mean squared error (RMSE) lower than 3.4. In addition, multiple linear regression showed good correlations between real data and generated data; however, the insufficient linearity between independent variables makes it lose statistical reliability. In contrast, the Paulhus and Kohler method, together with the Wavelet transform, proved to be less effective, showing poor correlation and high errors in the simulated data. These findings underscore the importance of carefully choosing methods for estimating rainfall data in paramo areas to ensure the accuracy and reliability of results in water resources management.

Keywords: precipitation, meteorology, mathematical modeling, statistics, water resource

Editado por / Edited by:
Eva O.L. Lantsoght

Recibido / Received:
27/02/2024

Aceptado / Accepted:
11/04/2024

Publicado en línea / Published online:
09/05/2024

Resumen

La ausencia de datos en series temporales de variables meteorológicas es un inconveniente en las ciencias ambientales, especialmente en lo que respecta a la precipitación, que es una variable clave en varios campos. Ahora bien, el presente estudio se propuso comparar varios métodos estadísticos y matemáticos para generar datos pluviométricos faltantes en la microcuenca del río Pita, tales como el método de Paulhus y Kohler, la regresión lineal múltiple (RLM), la transformada de *Wavelet* y las redes neuronales artificiales; utilizando información de la red hidrometeorológica del Fondo para la Protección del Agua (FONAG) de Quito. Las redes neuronales artificiales fueron altamente efectivas para generar datos pluviométricos en la zona de estudio,

 DOI: https://doi.org/10.18272/aci.v16i1.3274

con coeficientes de determinación (R^2) superiores a 0.64; y raíces del error cuadrático medio menores (RMSE) a 3.4. Además, la regresión lineal múltiple presentó buenas correlaciones entre los datos reales y los datos generados. Sin embargo, la insuficiente linealidad entre variables independientes hace que se pierda confiabilidad estadística. En contraste, el método de Paulhus y Kohler, junto con la transformada de *Wavelet*, demostraron ser menos eficaces, mostrando una correlación deficiente y altos errores en los datos simulados. Estos hallazgos subrayan la importancia de elegir cuidadosamente los métodos de estimación de datos pluviométricos en zonas de páramo para garantizar la precisión y la fiabilidad de los resultados en la gestión de recursos hídricos.

Palabras clave: precipitación, meteorología, modelo matemático, estadística, recurso hídrico

INTRODUCCIÓN

Los fenómenos climáticos influyen en la producción y suministro de recursos para la población, por lo tanto, cada país tiene la responsabilidad de supervisar las condiciones climáticas y sus cambios para realizar predicciones a corto, mediano y largo plazo [1]. Por tal razón, la ausencia de datos en series temporales de distintas variables meteorológicas (temperatura, precipitación, humedad relativa, velocidad del viento, etc.) es un inconveniente en las ciencias ambientales [2]. Cabe mencionar que algunos procedimientos de análisis pueden adaptarse a esta situación, pero otros requieren series completas [3]. Problemas comunes asociados con la falta de datos incluyen: la operación de estaciones meteorológicas de forma manual, la recopilación de información en momentos inoportunos, el mal funcionamiento de sensores automáticos y situaciones externas como interrupciones en el suministro eléctrico [4]. Además, la presencia de valores atípicos (*outliers*) puede considerarse como carencia de información en algunos estudios ambientales, dado que se descartan al ser tomados como errores instrumentales de los equipos de medición; caso contrario, tienen el potencial de afectar negativamente los resultados de un modelo numérico [5]. No obstante, es importante mencionar que, aunque estos tienen características diferentes al resto de información, numerosos estudios los incluyen dado que su eliminación podría resultar en la pérdida de información valiosa del fenómeno investigado [6].

Dentro de este marco, la precipitación es crucial, tanto en investigaciones hidrogeológicas, considerando que constituye el insumo principal para calcular balances hídricos y emitir alertas tempranas sobre posibles riesgos de sequía [7], como para la agricultura, en estudios de disponibilidad de lluvia para el diseño de mecanismos de recolección para lugares donde el acceso al agua es limitado [8]. Asimismo, es fundamental en el análisis de eventos extremos dentro del contexto del cambio climático [9]. Esta variable desencadena el ciclo hidrológico en la etapa terrestre, presentándose de manera aleatoria en relación al tiempo y espacio [10]. Por consiguiente, el análisis de eventos meteorológicos para la elaboración de modelos hidrológicos o la planificación de proyectos hidráulicos requieren principalmente de datos pluviométricos de alta calidad [11].

Ahora bien, cerca del 85 % del abastecimiento de agua para el Distrito Metropolitano de Quito (DMQ) tiene su origen en los páramos [12] y, una de las zonas más importantes,

DOI: https://doi.org/10.18272/aci.v16i1.3274

es la microcuenca del río Pita, reconocida como una de las fuentes hídricas clave para la ciudad; su río se integra a la cuenca alta del río Guayllabamba, que a su vez está incluida en la cuenca del río Esmeraldas [13]. El río Pita es responsable del 38 % de agua potable para el sur y centro de Quito, aportando un caudal de 1.6 m^3/s a través del Sistema Pita-Puengasí. Es el segundo sistema más relevante en la distribución de agua potable para el DMQ después del sistema Papallacta [14]. En consecuencia, resulta necesario establecer estrategias que aborden la carencia de información en sistemas hidrológicos, con el fin de estudiar la disponibilidad del recurso hídrico.

Se emplean distintos enfoques para estimar datos faltantes en series temporales. Entre los más comunes de estas metodologías, se encuentran: la regresión lineal, la razón normal, la regresión múltiple y los modelos geoestadísticos; sin embargo, según Melo et al. [15], estos últimos requieren trabajar sobre semivariogramas, lo que puede llegar a aumentar de manera significativa su complejidad. También se utiliza la aplicación de redes neuronales para analizar los datos meteorológicos [16], y la transformada de *Wavelet*, la cual se ha popularizado en las últimas décadas como una herramienta de análisis espectral para bases de datos ambientales [17]. En la actualidad, la estimación mediante técnicas estadísticas y matemáticas se lleva a cabo utilizando sistemas informáticos, lo que facilita el manejo eficiente de grandes conjuntos de datos en un tiempo reducido y con una menor carga de trabajo humano [18].

El objetivo del presente estudio es comparar varios métodos estadísticos y matemáticos para generar datos pluviométricos faltantes en la microcuenca del río Pita, tales como el método de Paulhus y Kohler, la regresión lineal múltiple (RLM), la transformada de *Wavelet* y las redes neuronales artificiales, que son los más utilizados en las ciencias ambientales y de la tierra, mediante el empleo del *software* estadístico RStudio. Para su ejecución, se utilizó información de las estaciones de la red hidrometeorológica del Fondo para la Protección del Agua (FONAG), con el propósito de identificar la metodología más adecuada para posteriores investigaciones en regiones de páramo con características climáticas similares. En la Tabla 1 se presentan algunos trabajos previos que abordan metodologías similares a las propuestas en este estudio. Esta pequeña recopilación destaca la relevancia y el interés de las metodologías planteadas, subrayando la necesidad de un análisis y el estudio de su aplicabilidad en contextos nacionales.

DOI: https://doi.org/10.18272/aci.v16i1.3274

Tabla 1. Trabajos previos relacionados con las metodologías aplicadas en la presente investigación

Metodología aplicada	Referencia
Paulhus y Kohler	[19]
	[20]
	[21]
	[22]
	[23]
Regresión Lineal Múltiple (RLM)	[24]
	[25]
	[26]
	[27]
	[28]
Transformada de *Wavelet*	[29]
	[30]
	[31]
	[32]
	[33]
Redes Neuronales Artificiales	[34]
	[35]
	[36]
	[37]
	[38]

METODOLOGÍA

Área de estudio

La microcuenca del río Pita está políticamente en tres cantones: el DMQ, Mejía y Rumiñahui, abarcando mayoritariamente las parroquias de Píntag y Machachi [39]. La distribución parroquial del territorio dentro de la microcuenca se detalla de la siguiente manera: Píntag abarca el 55.8 %, Machachi un 38.2 %, Rumipamba un 3.2 % y Sangolquí el 1.7 % [40]. Los páramos de la vertiente occidental del volcán Sincholagua complementados por una fracción de los deshielos del volcán Cotopaxi, constituyen las principales fuentes de flujo para el río Pita [41]. La Figura 1 presenta el estado de los páramos en la zona alta de la microcuenca, en donde se encuentran los humedales de páramo que actúan como las principales fuentes de agua de calidad para la capital.

DOI: https://doi.org/10.18272/aci.v16i1.3274

Figura 1. Páramos de la zona alta de la microcuenca del río Pita

La Tabla 2 ofrece un desglose de las coordenadas de las estaciones utilizadas en la investigación, incluyendo tanto las estaciones meteorológicas como las pluviométricas. Asimismo, la Figura 2 ilustra la ubicación específica de estas estaciones en la microcuenca de estudio.

Tabla 2. Estaciones Red Hidrometeorológica FONAG (Sistema de Referencia EPGS: 4326 – WGS 84)

Código	Tipo	Nombre	Longitud	Latitud	Altura (msnm)
M5028	Meteorológica	Hcda. Prado Miranda	-78.39071414	-0.48330906	3 526
M5029	Meteorológica	El Carmen	-78.33336768	-0.50165975	4 100
M5026	Meteorológica	Cotopaxi Control Norte	-78.44334571	-0.56382380	3 670
M5076	Pluviométrica	Potrerillos	-78.40224661	-0.61684014	3 866
M5030	Pluviométrica	Hcda. Gordillo	-78.35721535	-0.41833358	3 248

DOI: https://doi.org/10.18272/aci.v16i1.3274

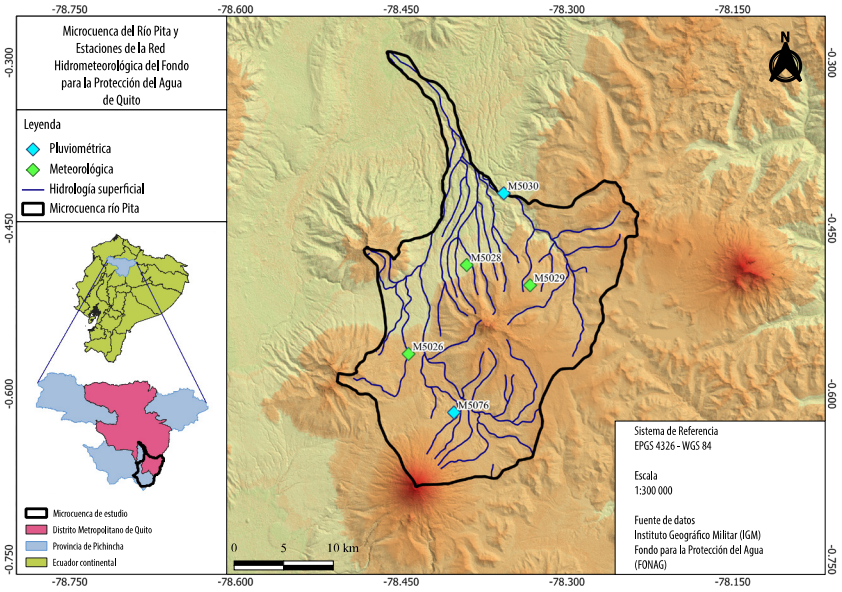

Figura 2. Mapa de microcuenca del río Pita y estaciones de la red hidrometeorológica de FONAG

Fuente y tratamiento de datos

Los registros pluviométricos fueron adquiridos de la red hidrometeorológica del Fondo para la Protección del Agua (FONAG) de Quito, que son de acceso libre en su página web: www.sedc.fonag.org.ec. Estos datos comprenden las precipitaciones mensuales acumuladas de cinco estaciones situadas dentro de la microcuenca. Se empleó como criterio de selección de estaciones a aquellas que proporcionaran la información más completa posible. Por ello, se determinó analizar el periodo entre 2014 y 2023 (10 años), durante el cual la ausencia de datos de cada estación no excedió el 5 % (ver Tabla 3).

Tabla 3. Porcentaje de datos ausentes en información pluviométrica

Código	Tipo	Nombre	Datos disponibles	% Datos faltantes
M5028	Meteorológica	Hcda. Prado Miranda	116	3.33
M5029	Meteorológica	El Carmen	114	5.00
M5026	Meteorológica	Cotopaxi Control Norte	120	0.00
M5076	Pluviométrica	Potrerillos	119	0.83
M5030	Pluviométrica	Hcda. Gordillo	119	0.83

DOI: https://doi.org/10.18272/aci.v16i1.3274

Se optó por completar los pocos datos faltantes utilizando la mediana de cada serie temporal, debido a que los diferentes conjuntos de datos presentaban una alta dispersión en términos de desviación estándar y varianza; esto se evidencia en los diagramas de caja y bigotes (Figura 3). Según Das e Imon [42], para conjuntos de datos con una alta dispersión, la mediana es menos susceptible a verse afectada por *outliers* o por alta variabilidad. Luego, para simular la generación de datos pluviométricos, se procedió a eliminar aleatoriamente el 20 % de los datos. En el estudio de Maharana et al. [43] se menciona que, al trabajar con bases de datos que sobrepasen el 20 % de información ausente, los modelos elaborados pierden robustez. De este modo, se permitió llevar a cabo la posterior comparación de los métodos establecidos sin comprometer la confiabilidad de los resultados.

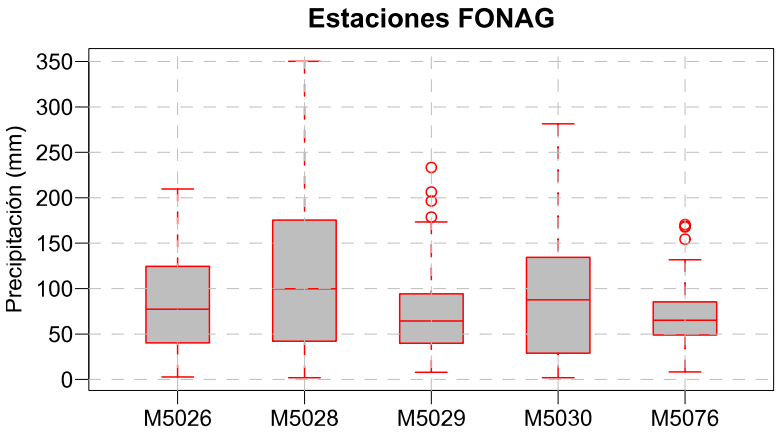

Figura 3. *Boxplots* de estaciones con información pluviométrica

En la Figura 4 se presenta un diagrama de flujo que esquematiza de manera general la metodología empleada en esta investigación. Este diagrama proporciona una visión panorámica de los pasos seguidos durante el desarrollo del estudio, desde la recolección y tratamiento de datos, hasta la ejecución, evaluación y análisis de los modelos estadísticos y matemáticos.

DOI: https://doi.org/10.18272/aci.v16i1.3274

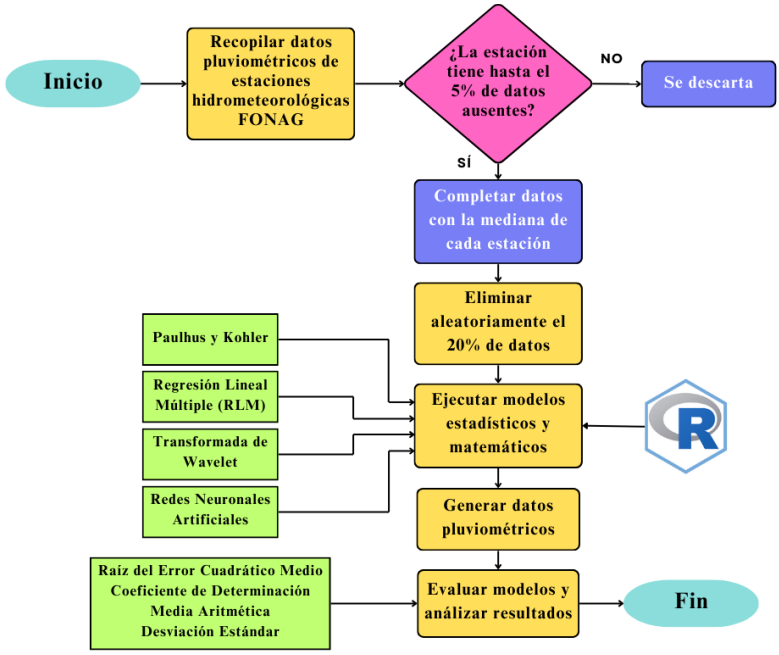

Figura 4. Proceso global de investigación

Método de Paulhus y Kohler (1952)

También llamado método de razón normal, implica estimar el valor incompleto: $x(t)$ de una serie, utilizando los datos de estaciones cercanas y simultáneas que muestren una fuerte correlación con la serie a completar [44]. Esto se realiza mediante la Ecuación 1.

$$x(t) = \frac{1}{3}\left[\frac{\bar{x}}{\bar{x}_1}x_1(t) + \frac{\bar{x}}{\bar{x}_2}x_2(t) + \frac{\bar{x}}{\bar{x}_3}x_3(t)\right] \quad (1)$$

Donde:

$\bar{x}$: media aritmética de datos pluviométricos
$\bar{x}_1, \bar{x}_2, \bar{x}_3$: media aritmética de estaciones vecinas
$x_1(t), x_2(t), x_3(t)$: datos pluviométricos de series vecinas

Para su ejecución se utilizó el paquete *climatol* del *software* estadístico RStudio donde, además de la opción de normalizar los datos dividiéndolos por sus valores medios, *climatol* también brinda la posibilidad de realizar esta normalización restando las medias o llevando a cabo una estandarización completa. Por lo tanto, tras denominar m_x y s_x a la media y desviación estándar de una serie X, a continuación se muestran las alternativas disponibles para la normalización de datos pluviométricos [45]:

DOI: https://doi.org/10.18272/aci.v16i1.3274

- Restar la media: $x = X - m_X$
- Dividir por la media: $x = X/m_X$
- Estandarizar: $x = (X - m_X)/s_X$

El principal desafío de este método radica en el desconocimiento de los valores de medias aritméticas y de desviaciones estándar de las series durante el periodo de estudio, lo que es común en las bases de datos reales. Por lo tanto, *climatol* aborda este problema al calcular inicialmente estos parámetros con los datos disponibles en cada serie. Luego, rellena los datos faltantes utilizando estas medias y desviaciones estándar provisionales, y vuelve a calcularlos con las series rellenadas. Posteriormente, se recalculan los datos inicialmente faltantes utilizando los nuevos parámetros, lo que resulta en nuevas medias y desviaciones estándar. Este proceso se repite hasta que ninguna media cambie al redondearla con la precisión inicial de los datos [45]. Una vez que las medias han sido estabilizadas, se lleva a cabo la normalización de todos los datos, seguida de la estimación de los mismos, tanto en las series existentes como en las que no están completas, utilizando la Ecuación 2.

$$\hat{y} = \frac{\sum_{j=1}^{j=n} w_j x_j}{\sum_{j=1}^{j=n} w_j} \tag{2}$$

Donde $\hat{y}$ representa el valor estimado utilizando los n datos x_j más cercanos disponibles en cada intervalo de tiempo, y w_j es el peso asignado a cada uno de ellos.

Regresión Lineal Múltiple (RLM)

Hay una variedad de técnicas de regresión que varían dependiendo del tipo de variables y de la relación funcional supuesta entre ellas. Las técnicas más básicas, aunque muy efectivas en términos de la cantidad de información que pueden proporcionar, son las regresiones lineales [46]. La regresión lineal múltiple se construye a partir de una regresión lineal simple, la cual se utiliza cuando se tiene más de una variable independiente [47]. En este estudio, el modelo de regresión se aplica para datos pluviométricos y se adapta a las condiciones y necesidades del análisis, tal como se ilustra en la Ecuación 3.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \epsilon \tag{3}$$

Donde:

y: valor de precipitación que se quiere estimar
$x_1, x_2, \ldots, x_k$: datos pluviométricos de estaciones hidrometeorológicas vecinas
b_0: intercepto o valor de precipitación cuando todas las estaciones tienen valores de 0
$b_1, b_2, \ldots, b_k$: coeficientes de regresión
ϵ: errores aleatorios

Cuando se tienen n observaciones o filas en el conjunto de datos pluviométricos, se obtiene el siguiente modelo:

DOI: https://doi.org/10.18272/aci.v16i1.3274

$$y_1 = b_0 + b_1x_{11} + b_2x_{12} + \cdots + b_kx_{1k} + \epsilon_1$$
$$y_2 = b_0 + b_1x_{21} + b_2x_{22} + \cdots + b_kx_{2k} + \epsilon_2$$
$$y_3 = b_0 + b_1x_{31} + b_2x_{32} + \cdots + b_kx_{3k} + \epsilon_3$$
$$\cdots$$
$$y_n = b_0 + b_1x_{n1} + b_2x_{n2} + \cdots + b_kx_{nk} + \epsilon_n$$

Utilizando matrices, se puede representar el sistema de n ecuaciones mediante la Ecuación 4.

$$y = Xb + \epsilon \tag{4}$$

Donde:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_n \end{bmatrix} \qquad \text{y} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix}$$

En general, y es un vector $(n \times 1)$ de datos pluviométricos, X es una matriz $(n \times p)$ de los niveles de las variables independientes (información pluviométrica de estaciones hidrometeorológicas vecinas), b es un vector $(p \times 1)$ de los coeficientes de regresión y ϵ es un vector $(n \times 1)$ de los errores aleatorios. De esta manera, los estimadores de mínimos cuadrados se calculan mediante la Ecuación 5.

$$L = \sum \epsilon_i^2 = \epsilon'\epsilon = (y - Xb)'(y - Xb) \tag{5}$$

El estimador de mínimos cuadrados $\hat{b}$ es la solución para el vector b (Ecuación 6).

$$\frac{\partial L}{\partial b} = 0 \tag{6}$$

Finalmente, al resolver la ecuación diferencial anterior se obtiene la Ecuación 7, donde se determinan los coeficientes de regresión para el modelo.

$$\hat{b} = (X'X)^{-1}X'y \tag{7}$$

DOI: https://doi.org/10.18272/aci.v16i1.3274

En la regresión lineal múltiple, se utilizan múltiples variables explicativas, lo que posibilita el aprovechamiento de una mayor cantidad de información en la construcción del modelo y, por consiguiente, la obtención de estimaciones más precisas para completar las series pluviométricas.

Transformada de *Wavelet*

El tercer método utilizado corresponde a las transformadas de *Wavelet*, que son herramientas matemáticas que permiten analizar señales de manera similar a la transformada de Fourier de tiempo corto, proporcionando información tanto en el dominio del tiempo como en el de la frecuencia [48]. Las transformadas de *Wavelet* permiten estudiar características en la serie espacial con un detalle ajustado a su escala, es decir, rasgos amplios a gran escala y rasgos finos a pequeña escala. Esta característica es útil para las variaciones espaciales que son significativamente no estacionarias y tienen componentes transitorios de corta duración [49]. De esta manera, el análisis *Wavelet* tiene distintas aplicaciones, desde la dinámica de fluidos [50], la geofísica [51] y la hidrología [52], como en esta investigación. Las *wavelets*, fundamentales en la transformada *wavelet* madre, representan una señal mediante versiones desplazadas y escaladas de una onda finita que pueden generarse a partir de un conjunto de datos experimentales. Esta transformada no solo es local en el dominio del tiempo, sino también en el dominio de la frecuencia [53]. Una vez que se tiene una *wavelet* madre, se pueden generar *wavelets* mediante las operaciones de dilatación y traslación [54]. Para números enteros j, k se utiliza la Ecuación 8.

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) \tag{8}$$

Resulta que estas ondículas pueden formar un conjunto ortonormal (Ecuación 9).

$$< \psi_{j,k}, \psi_{j',k'} >= \int_{-\infty}^{\infty} \psi_{j,k}(x)\psi_{j',k'}(x)dx = \delta_{j,j'}\delta_{k,k'} \tag{9}$$

Donde $\delta_{m,n} = 1$ si $m = n$, y $\delta_{m,n} = 0$ si $m \neq n$. En este caso $<\cdot,\cdot>$ es el producto interior. Además, ese conjunto de ondículas puede formar bases para varios espacios de funciones. Por ejemplo, y más técnicamente, $\{\psi_{j,k}(x)\}_{j,k\,\epsilon\,\mathbb{Z}}$ puede ser una base ortonormal completa para $L^2(\mathbb{R})$. Así, dada la función $f(x)$, se procede a descomponerla en una serie de Fourier generalizada, como indica la Ecuación 10.

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k}\,\psi_{j,k}(x) \tag{10}$$

Donde, debido a la ortogonalidad de las ondículas, se obtiene la Ecuación 11.

$$d_{j,k} = \int_{-\infty}^{\infty} f(x)\psi_{j,k}(x)dx =< f, \psi_{j,k} > \tag{11}$$

DOI: https://doi.org/10.18272/aci.v16i1.3274

Para enteros j, k, los números $\{d_{j,k}\}_{j,k \in \mathbb{Z}}$ se denominan coeficientes de *wavelet* de f, generando una función que se asemeja al conjunto de datos ingresados, que en esta investigación son las observaciones pluviométricas de cada estación hidrometeorológica. Para varias situaciones, las *wavelets* resultan útiles, aunque hay numerosos casos donde otros métodos disponibles son igualmente eficientes o incluso superiores. El paquete *WaveletComp* de RStudio permite trabajar esta metodología de una manera más rápida, y a su vez, ofrece opciones de trazado que facilitan un ajuste óptimo del modelo.

Redes Neuronales Artificiales

Dado que las redes neuronales artificiales se diseñaron intencionalmente como modelos conceptuales de la actividad cerebral humana, resulta útil comprender primero cómo funcionan las neuronas biológicas. La Figura 5 ilustra como las señales entrantes son recibidas por las dendritas de la célula a través de un proceso bioquímico, y su vez, emitiendo una señal de salida por el axón [55].

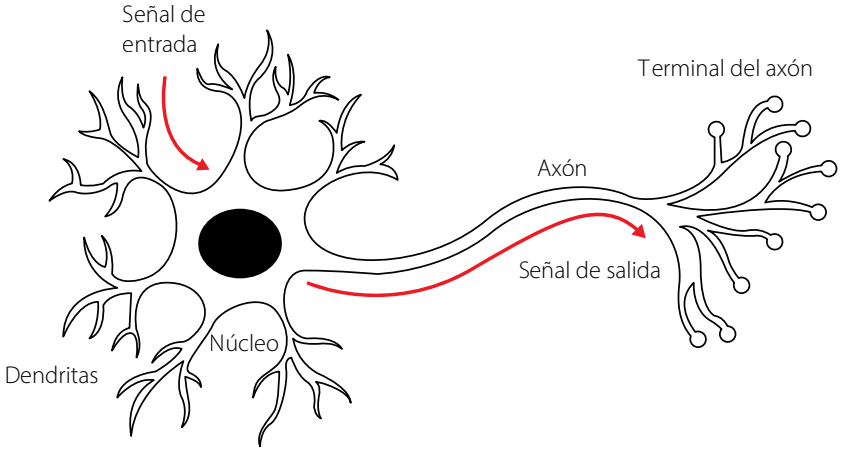

Figura 5. Representación artística de una neurona biológica. Imagen con base en [55]

Así, una red neuronal artificial es un modelo de regresión o clasificación en dos etapas, generalmente suele representarse mediante un diagrama de red, como lo muestra la Figura 6.

DOI: https://doi.org/10.18272/aci.v16i1.3274

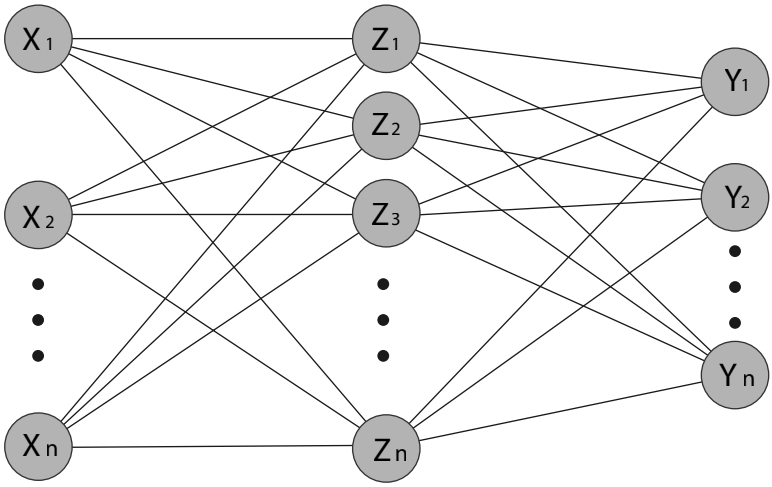

Figura 6. Esquema de una red neuronal con una capa oculta. Imagen con base en [56]

Para regresión, normalmente hay una sola unidad de salida Y_1 en la parte superior, como es el caso de esta investigación (Figura 7), en donde las estaciones hidrometeorológicas vecinas se transforman en las dendritas de entrada, y la estación con datos faltantes es la variable de salida donde se generarán los nuevos datos pluviométricos. Sin embargo, es importante mencionar que estas redes pueden manejar múltiples respuestas cuantitativas de forma fluida [56].

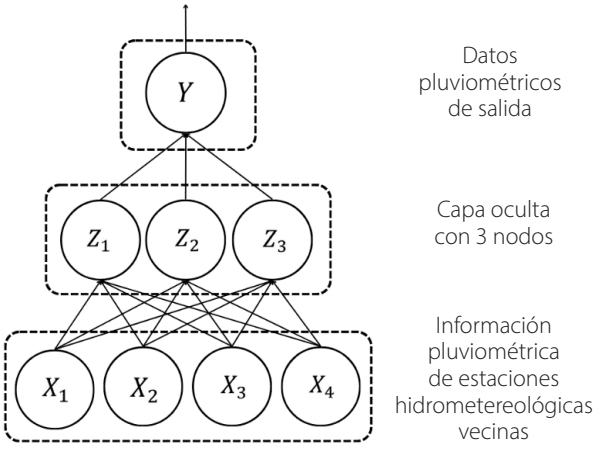

Figura 7. Representación de una estructura de red neuronal artificial con una capa oculta de 3 nodos para estimación de datos pluviométricos

DOI: https://doi.org/10.18272/aci.v16i1.3274

Una neurona artificial típica con n dendritas de entrada puede representarse mediante la Ecuación 12. Los pesos (w_i) permiten que cada una de las n entradas de x contribuya en mayor o menor medida a la suma de las señales de entrada. El valor acumulado se pasa a la función de activación, $f(x)$, y la señal resultante, $y(x)$, es el axón de salida [57].

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{12}$$

En el presente estudio se utilizó el algoritmo de retropropagación (*backpropagation*), que es el método de entrenamiento predominante en redes neuronales. Este método de aprendizaje supervisado emplea el descenso del gradiente, que se divide en dos fases: en primer lugar, se introduce un patrón de entrada que se propaga a través de las diferentes capas de la red neuronal hasta generar la señal de salida. Luego, esta salida se compara con la salida deseada para calcular el error en cada neurona y los errores se retropropagan desde la capa de salida hacia todas las neuronas de las capas intermedias [58]. Cada neurona recibe un error que refleja su influencia en el error global de la red. A partir de este error recibido, se realizan ajustes en los pesos sinápticos de cada neurona. El propósito consiste en reducir al mínimo el error entre la salida producida por la red y la salida deseada por el usuario cuando se presenta un conjunto de patrones p, conocido como conjunto de entrenamiento. Por consiguiente, el error se distribuye en sentido opuesto al flujo normal de información de la red. Así, el algoritmo identifica y corrige los errores durante el proceso de aprendizaje, comenzando desde las capas más profundas y retrocediendo hacia la entrada. Para simplificar este procedimiento, se empleó el paquete *neuralnet* de RStudio, el cual facilita la implementación de este método al especificar los parámetros de entrada, como el número de nodos en las capas ocultas, y definir variables dependientes e independientes.

La Figura 8 presenta un resumen del funcionamiento del algoritmo de *backpropagation* utilizado en la ejecución de redes neuronales artificiales.

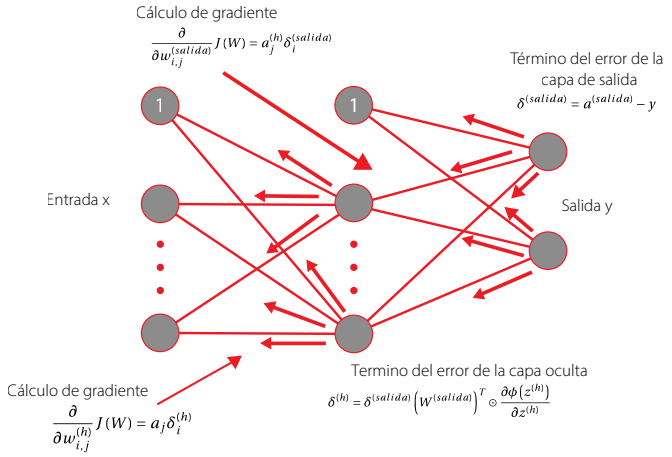

Figura 8. *Backpropagation* en redes neuronales artificiales. Imagen con base en [59]

DOI: https://doi.org/10.18272/aci.v16i1.3274

Métricas de evaluación

Con el fin de determinar el método óptimo de generación de datos pluviométricos, se establecieron dos métricas principales: Raíz del Error Cuadrático Medio ($RSME$) y Coeficiente de Determinación (R^2). La métrica RMSE es comúnmente empleada para evaluar la efectividad de un modelo de regresión. Su función es determinar la discrepancia entre dos conjuntos de datos, comparando las predicciones del modelo con los valores reales (Ecuación 13) [60].

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{13}$$

Donde:

y_j: serie pluviométrica original
$\hat{y}_j$: serie pluviométrica estimada

Mientras que el coeficiente de determinación proporciona información sobre el grado de relación entre las dos variables que explican la fluctuación de los datos (Ecuación 14) [61]. Para el caso de estudio, las dos variables serán los datos estimados y los datos reales.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{14}$$

Donde:

y_i: serie pluviométrica original
$\hat{y}_i$: serie pluviométrica estimada
$\bar{y}_i$: media de datos pluviométricos

No obstante, también se aplicó la media aritmética y la desviación estándar a los conjuntos de datos antes y después de la simulación.

RESULTADOS

Con *climatol* se pudo visualizar el conjunto de datos de las cinco estaciones, donde los espacios en blanco constituyen la información pluviométrica ausente que se eliminó de manera aleatoria para simular las metodologías presentadas anteriormente (Figura 9).

DOI: https://doi.org/10.18272/aci.v16i1.3274

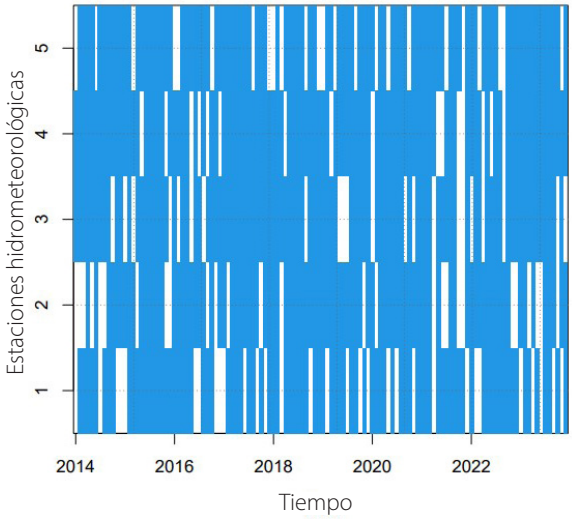

Figura 9. Disponibilidad de datos pluviométricos de estaciones hidrometeorológicas

La Figura 10 muestra la distribución espacial de las estaciones en la zona de estudio, en donde *climatol* realizó automáticamente un *clustering* jerárquico para identificar patrones. Se observaron dos clústeres distintos, representados en el gráfico por estaciones marcadas en verde y estaciones marcadas en rojo. Los clústeres identificados en el gráfico sugieren la presencia de dos áreas geográficas distintas, cada una con características climáticas únicas.

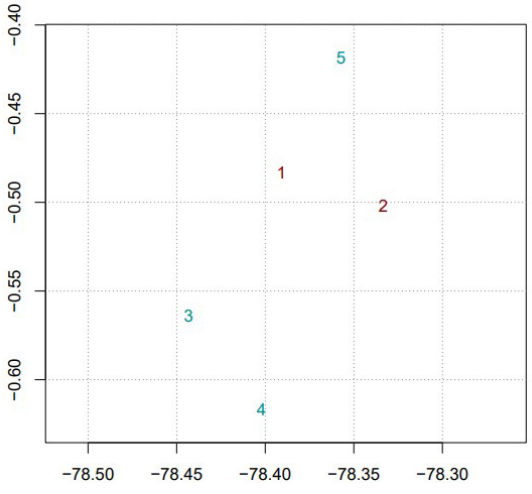

Figura 10. Agrupamiento de estaciones hidrometeorológicas mediante *climatol*

DOI: https://doi.org/10.18272/aci.v16i1.3274

Los diagramas de anomalías (Figura 11) incluyen dos líneas suplementarias en la sección inferior, las cuales indican la mínima separación entre los datos adyacentes (en verde) y la cantidad de datos de referencia empleados (en naranja), ambas utilizando la escala logarítmica del eje derecho. El análisis de los diagramas de anomalías de precipitación acumulada mensual es esencial para examinar las desviaciones en los patrones de lluvia a lo largo del tiempo. Estos diagramas muestran claramente los periodos donde se han observado cambios significativos en la precipitación acumulada, destacando tanto los excesos como los déficits de lluvia respecto a las condiciones climáticas durante 10 años de estudio (2014-2023). En las cinco estaciones también se señalan con una línea vertical discontinua las posibles fechas de cambio tras la evaluación de la homogeneidad de la serie.

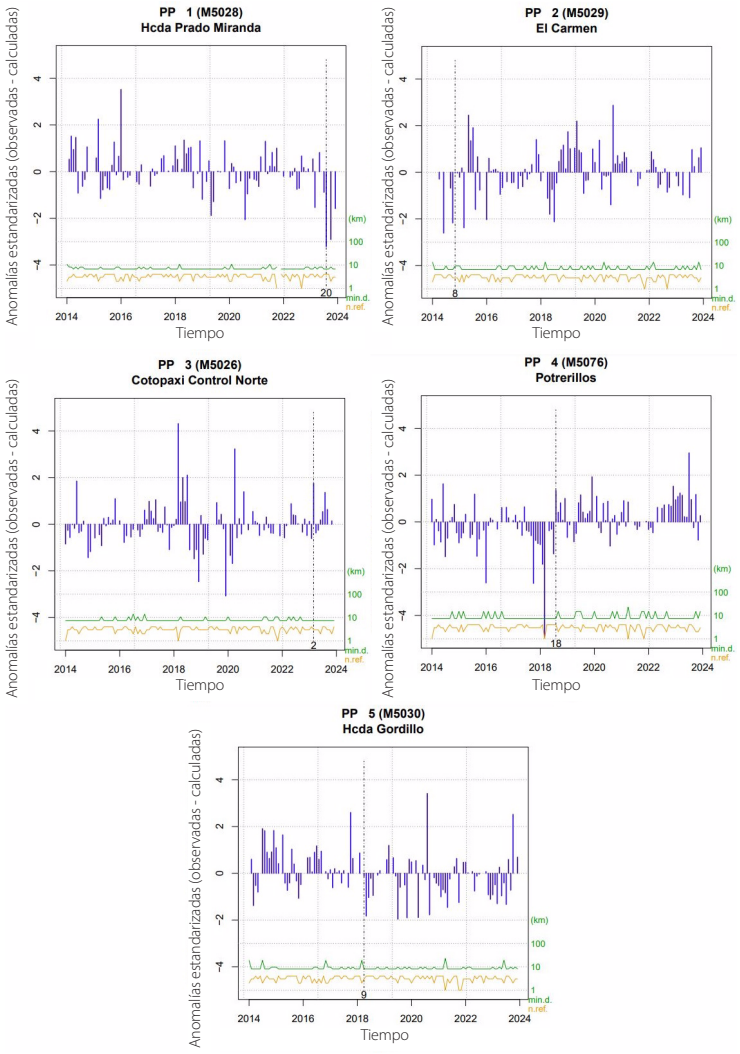

Figura 11. Anomalías climáticas en información pluviométrica

DOI: https://doi.org/10.18272/aci.v16i1.3274

Una vez completadas las series temporales, se observaron diferencias significativas entre los valores originales y los datos generados, tal como se refleja en las métricas de evaluación (Tabla 5). El RMSE revela la magnitud promedio del error de predicción, evidenciando discrepancias considerables entre los valores observados y los generados, con RMSEs de 120.96, 69.83, 81.77, 81.04 y 69.54 para las estaciones M5028, M5029, M5026, M5076 y M5030, respectivamente. Ahora bien, los valores de R^2 fueron extremadamente bajos, oscilando entre 0.000 y 0.067, lo que sugiere una variabilidad significativa no explicada por el modelo. Al comparar las medias aritméticas de los datos originales con los completados, se evidenció una variación diferencial entre estaciones, lo que sugiere una influencia heterogénea de la estimación en la tendencia central de las series temporales. Por otro lado, la comparación de las desviaciones estándar de los datos originales y completados mostró cambios más sutiles, lo que indica una relativa estabilidad en la dispersión de los datos después de la aplicación de los métodos de estimación.

Para la generación de datos a través de la regresión lineal múltiple (RLM), se generó un modelo específico para cada estación, empleando las estaciones restantes como variables independientes en el proceso. Ante todo, se analizó la relación estadística entre los datos pluviométricos de todos los conjuntos de datos. La consideración de esta información es fundamental para determinar los predictores óptimos del modelo, identificar variables con relaciones no lineales que no deben ser consideradas y detectar posibles problemas de multicolinealidad entre los predictores. Al mismo tiempo, se sugiere complementar este análisis representando la distribución de cada variable a través de histogramas (Figura 12), que demuestran una forma asimétrica, lo que sugiere que la variable de precipitación acumulada mensual no sigue una distribución normal.

Los histogramas muestran una distribución levemente sesgada hacia la derecha, indicando una mayor frecuencia de valores de precipitación menores que la media. Además, los datos recopilados de las diversas estaciones en la microcuenca del río Pita exhiben multicolinealidad, lo que indica que varias variables están linealmente relacionadas entre sí. Por último, los coeficientes de Pearson superan el 0.65 en todas las estaciones, excepto en M5030 con M5076, lo que impide la identificación clara del efecto individual de cada variable sobre la variable respuesta.

DOI: https://doi.org/10.18272/aci.v16i1.3274

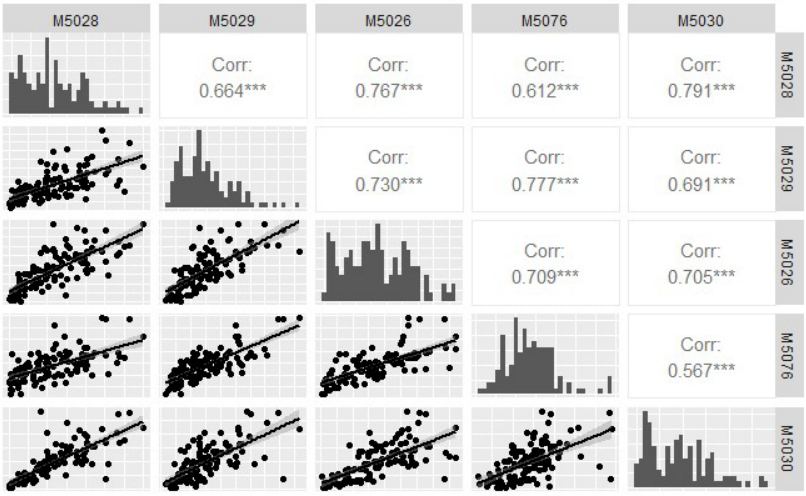

Figura 12. Matriz de correlación de datos pluviométricos

En la Tabla 4 se describen las ecuaciones generadas por regresión lineal múltiple para cada estación, cabe señalar que el valor *p-value* es estadísticamente significativo para cada modelo generado (2.2e-10), lo que sugiere que los modelos no son aleatorios y al menos uno de los coeficientes de regresión parciales es diferente de cero.

Tabla 4. Ecuaciones de regresión para cada estación en la microcuenca del río Pita

Estación	Ecuación
M5028	$M5028 = 2.4790 + 0.6598\ M5026 + 0.6124\ M5030$
M5029	$M5029 = -4.7318 + 0.1605\ M5026 + 0.6926\ M5076 + 0.1924\ M5030$
M5026	$M5026 = 2.5013 + 0.2916\ M5028 + 0.2682\ M5029 + 0.4076\ M5076$
M5076	$M5076 = 24.4884 + 0.1834\ M5026 + 0.3871\ M5029$
M5030	$M5030 = 4.4462 + 0.4821\ M5028 + 0.4418\ M5029$

La validación de los métodos se llevó a cabo mediante el análisis de la linealidad entre las variables independientes y los residuos del modelo, un aspecto clave para determinar la homocedasticidad. Esta condición se verifica cuando los residuos muestran una distribución aleatoria alrededor de cero. Los valores de RMSE fluctúan entre 19.71 y 53.56, y el coeficiente de determinación (R^2) varía entre 0.552 y 0.727 (Tabla 5). Estos resultados indican que los modelos explican más del 50 % de la variabilidad en los datos de precipitación. Además, se observa que la media aritmética se mantiene constante con la aplicación del modelo, mientras que la dispersión de datos disminuye, lo cual se refleja en una menor desviación estándar después de aplicar los modelos de regresión.

Por otro lado, la aplicación de la transformada de *Wavelet* posibilitó la reconstrucción de las diversas series de datos mediante el empleo de herramientas matemáticas de vanguardia. La

DOI: https://doi.org/10.18272/aci.v16i1.3274

Figura 13 muestra la descomposición de *wavelet* de la serie temporal de datos pluviométricos recopilados durante el período de estudio utilizando la función *wt.image* de la librería *WaveletComp* en RStudio. Esta imagen resultante presenta una representación visual de la distribución de energía en diferentes escalas temporales, destacando patrones y estructuras de variabilidad multiescalar en los datos pluviométricos. El eje horizontal representa el tiempo (10 años), mientras que el eje vertical representa la escala o frecuencia. Además, se pueden identificar visualmente áreas de alta o baja variabilidad, coincidiendo con las mismas áreas de la Figura 11 de anomalías pluviométricas. A su vez, los cambios temporales en la estructura de descomposición proporcionan información crucial sobre la dinámica temporal de los datos pluviométricos. En esta representación, las áreas de color rojo intenso indican niveles más altos de periodicidad, mientras que las áreas delimitadas por líneas de contorno blancas representan componentes periódicos significativos en la serie temporal. Se observan patrones mensuales que reflejan el comportamiento de la precipitación en diferentes periodos de tiempo, como las temporadas de invierno en la región de la sierra, que generalmente abarcan los primeros meses del año, desde enero hasta mayo.

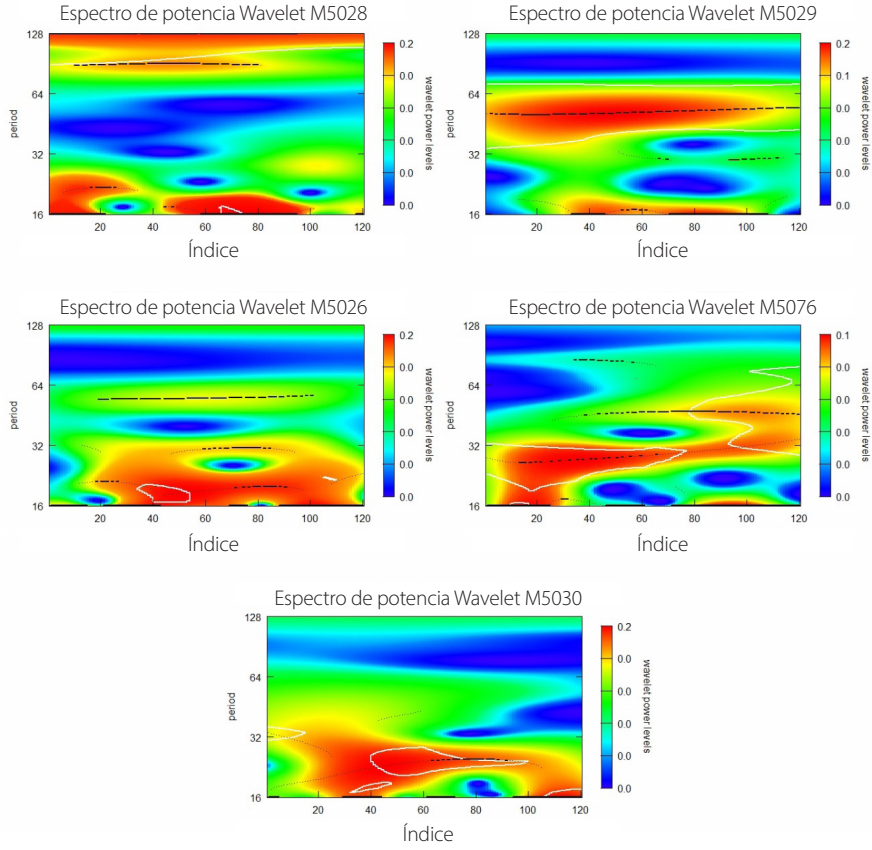

Figura 13. Espectrogramas *Wavelet*

DOI: https://doi.org/10.18272/aci.v16i1.3274

La Figura 14 muestra la reconstrucción de la serie temporal de datos pluviométricos mediante la aplicación de la función *reconstruct*. Esta visualización representa la serie temporal reconstruida a partir de la descomposición de la *wavelet* previamente realizada. Los nuevos datos pluviométricos estimados se ilustran en color rojo, mientras que de color negro está la serie de datos originales. La reconstrucción resalta las tendencias temporales, donde al final del periodo de tiempo se identifica una tendencia a la baja en la mayoría de estaciones, indicando un cambio gradual en los patrones de lluvia en la microcuenca. Igualmente, los ciclos estacionales revelan fluctuaciones notables en la cantidad de precipitación durante diferentes periodos anuales, mientras que las variaciones multiescalares presentes en los datos pluviométricos ofrecen la presencia de ciclos climáticos de diferente duración, como en el caso de eventos climáticos extremos, característicos en la región.

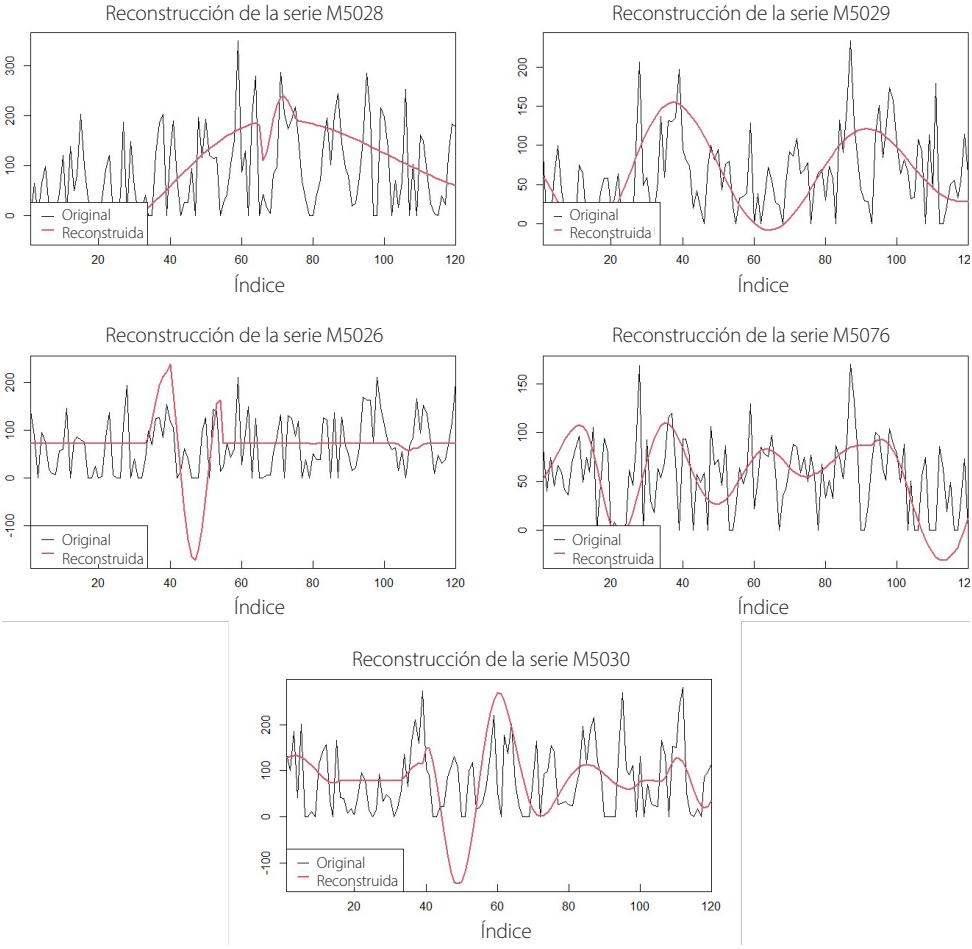

Figura 14. Reconstrucción de series temporales

DOI: https://doi.org/10.18272/aci.v16i1.3274

Los resultados del método (Tabla 5) muestran una amplia variabilidad en los valores de RMSE, que oscilan entre 56.24 y 118.96. Además, los valores de R^2 varían entre 0.012 y 0.253. Antes de la implementación de los métodos de rellenado, se evidencia una cierta inestabilidad en los datos originales, reflejada en una amplia gama de valores para la media aritmética y la desviación estándar, que van desde 67.30 hasta 113.05 y desde 44.22 hasta 81.04, respectivamente. Sin embargo, después de la aplicación de los modelos de transformada de *wavelet*, se observa una mejora en la estabilidad de los datos, con una reducción en la dispersión y una ligera ajuste en la media aritmética, que varía entre 61.48 y 101.93, y en la desviación estándar, que fluctúa entre 47.97 y 77.65.

Para la implementación del método de redes neuronales artificiales, se utilizó la librería *neuralnet*, la cual permitió configurar arquitecturas de red con dos capas ocultas para cada modelo asociado a las estaciones de estudio, compuestas por 5 y 3 nodos respectivamente. Después de probar varias configuraciones, se determinó que aumentar el número de nodos conlleva a mayores exigencias computacionales, mientras que reducir el número de nodos resulta en un incremento de errores. Para el entrenamiento de estos modelos, se asignó aleatoriamente el 70 % de los datos como conjunto de entrenamiento y el 30 % restante como conjunto de prueba. La visualización de los modelos generados por las redes neuronales para cada estación, junto con su correspondiente evaluación de errores y etapas de procesamiento, se presenta en la Figura 15.

DOI: https://doi.org/10.18272/aci.v16i1.3274

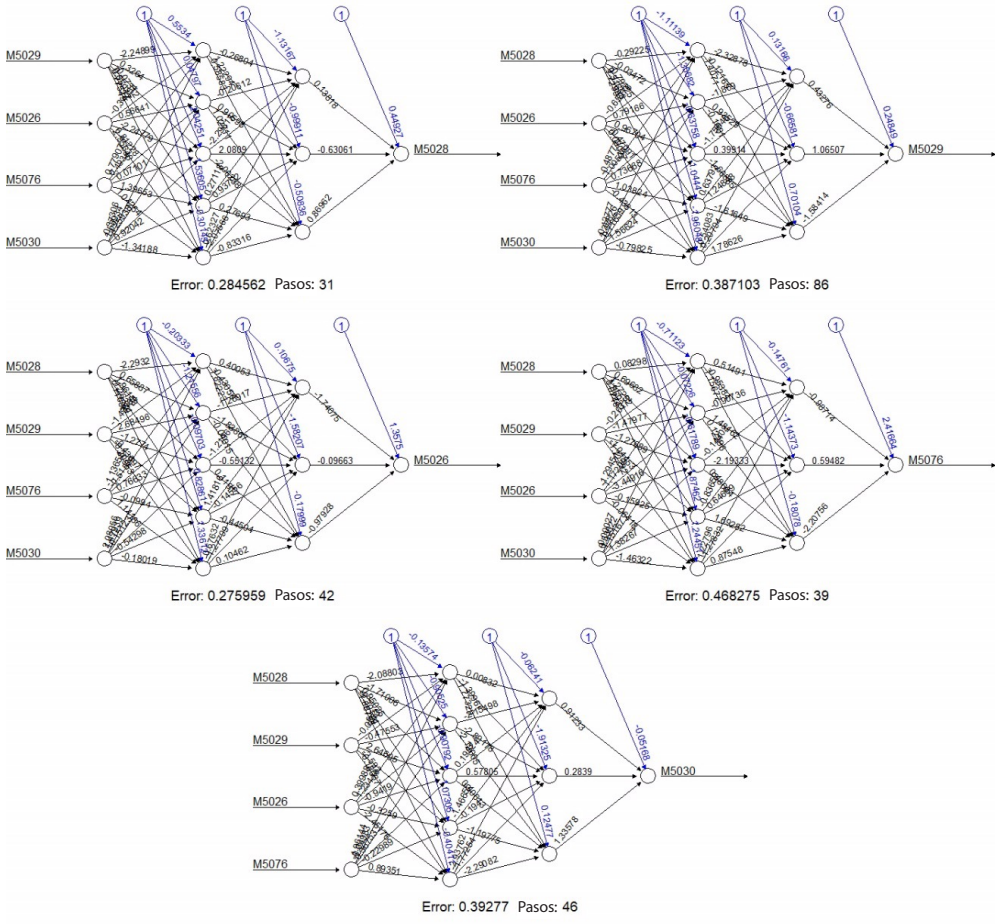

Figura 15. Redes neuronales artificiales por retropropagación

En los resultados del modelo de redes neuronales al rellenar datos pluviométricos, el error cuadrático medio (RMSE) muestra una variabilidad entre 1.56 y 3.41, lo que indica diferencias en la exactitud de las estimaciones en relación con los valores observados. Esta variación está asociada con factores geográficos y climáticos específicos de cada estación. Por otro lado, el R^2 revela la capacidad de los modelos para explicar la variabilidad en los datos de precipitación, con valores que oscilan entre 0.643 y 0.805. Esto sugiere diferencias en la capacidad predictiva de las redes neuronales en cada estación, posiblemente relacionadas con la complejidad de los patrones de precipitación locales.

DOI: https://doi.org/10.18272/aci.v16i1.3274

Además, al examinar los cambios en la media aritmética y la desviación estándar antes y después de aplicar el método, se observa una estabilización general o una ligera mejora en la precisión de los datos estimados, indicando una mayor consistencia y fiabilidad en las predicciones generadas por las redes neuronales artificiales.

Tabla 5. Evaluación de métodos de generación de datos pluviométricos

Paulhus y Kohler					
Métrica de evaluación	**M5028**	**M5029**	**M5026**	**M5076**	**M5030**
RMSE	120.96	69.83	81.77	81.04	69.54
$R2$	0.026	0.009	0.002	0.000	0.067
Media aritmética antes	113.05	73.25	82.23	68.03	91.31
Media aritmética después	102.95	80.51	86.16	75.60	96.61
Desviación estándar antes	81.04	44.22	51.12	30.84	65.82
Desviación estándar después	73.95	45.42	55.23	40.07	66.18
Regresión lineal múltiple					
Métrica de evaluación	**M5028**	**M5029**	**M5026**	**M5076**	**M5030**
RMSE	53.56	26.09	32.67	19.71	25.13
$R2$	0.672	0.600	0.552	0.727	0.719
Media aritmética antes	113.05	73.25	82.23	68.03	91.31
Media aritmética después	112.27	74.93	83.76	67.99	91.71
Desviación estándar antes	81.04	44.22	51.12	30.84	65.82
Desviación estándar después	76.94	43.44	50.03	28.66	65.32
Transformada de Wavelet					
Métrica de evaluación	**M5028**	**M5029**	**M5026**	**M5076**	**M5030**
RMSE	118.96	56.24	94.18	60.75	87.80
$R2$	0.024	0.074	0.207	0.012	0.253
Media aritmética antes	113.05	73.25	82.23	68.03	91.31
Media aritmética después	101.93	67.30	76.28	61.48	93.64
Desviación estándar antes	81.04	44.22	51.12	30.84	65.82
Desviación estándar después	77.65	47.97	55.63	34.83	66.09
Redes neuronales artificiales					
Métrica de evaluación	**M5028**	**M5029**	**M5026**	**M5076**	**M5030**
RMSE	2.14	2.78	1.89	3.41	1.56
$R2$	0.671	0.782	0.805	0.764	0.643
Media aritmética antes	113.05	73.25	82.23	68.03	91.31
Media aritmética después	114.15	75.20	81.79	67.32	90.58
Desviación estándar antes	81.04	44.22	51.12	30.84	65.82
Desviación estándar después	82.76	44.69	49.67	28.45	64.02

DOI: https://doi.org/10.18272/aci.v16i1.3274

DISCUSIÓN

La precipitación es un fenómeno ampliamente reconocido como un proceso complejo y no lineal [62]. Esta complejidad se refleja en los altos valores de la raíz del error cuadrático medio y bajos valores del coeficiente de correlación obtenidos mediante el método de Paulhus y Kohler, así como en el análisis realizado mediante la transformada de *Wavelet*. A pesar de las ambigüedades observadas en la implementación de estas metodologías, el uso de herramientas computacionales, como la herramienta *climatol*, ha demostrado ofrecer ventajas significativas. Esta herramienta no solo facilita el rellenado de datos faltantes, sino que también permite la homogeneización de las series temporales. De hecho, investigaciones previas, como la de Cartaya et al. [63], han empleado esta herramienta con el fin de homogenizar datos meteorológicos, obteniendo series temporales de mejor calidad. Esta práctica se justifica debido a las posibles discrepancias en la recopilación de datos mediante equipos meteorológicos, lo que puede afectar la fiabilidad estadística de los resultados obtenidos.

Asimismo, el *clustering* jerarquizado de la región de interés proporciona una explicación para ciertas discrepancias y errores en la generación de datos. Sin embargo, Poblete et al. [64] afirman que el enfoque jerárquico utilizado en la formación de los grupos presenta una limitación al generar combinaciones iniciales no deseables que pueden persistir durante el análisis, lo que podría resultar en interpretaciones incorrectas. Para garantizar una mayor confianza en los resultados del análisis de agrupamiento, se requiere realizar múltiples aplicaciones bajo diferentes condiciones, considerando estaciones atípicas como candidatas a revisión, y recalculando los grupos o utilizando diversas medidas de similitud y otros métodos de agrupación [65]. Entonces, se ha demostrado la nula eficacia del método de Paulhus y Kohler para precipitaciones acumuladas mensuales para la presente investigación, sin embargo, en el estudio de Pinthong et al. [19] este método demuestra tener una alta capacidad de generación de datos pluviométricos cuando la escala temporal es diaria, debido a que los errores generados en el recalculo de precipitaciones son más pequeños al tomar tiempos más cortos [66].

Además, la transformada de *Wavelet* mostró limitaciones en la estimación de datos faltantes, lo cual puede atribuirse a la longitud y la naturaleza aleatoria de la serie temporal. Este desafío se acentúa en estaciones recientes, porque la cantidad de datos disponibles es insuficiente. Idealmente, esta técnica se desempeña mejor en bases de datos con más de 30 años, donde la significancia estadística es más sólida [67]. A pesar de eso, la transformada *Wavelet* ofrece descomposiciones valiosas de las series de tiempo originales, lo que permite que los datos transformados en *wavelet* enriquezcan la capacidad de un modelo de pronóstico al capturar información relevante en varios niveles de resolución. Se ha observado que esta metodología parece ser más efectiva que la transformada de Fourier en el tratamiento de series de tiempo no estacionarias, según estudio previo de Salazar [68]. También, Sifuzzaman et al. [69] mencionan que una de las principales ventajas de la transformada de *Wavelet* es su robustez, dado que excluye cualquier sospecha errónea o procedimiento de prueba paramétrica.

Añadido a eso, la regresión lineal múltiple (RLM) presentó una óptima correlación de datos generados. Cabe destacar que, al estar en una misma microcuenca, las estaciones actúan de manera similar, por lo que su papel como variables independientes es muy

DOI: https://doi.org/10.18272/aci.v16i1.3274

utilizada. Sin embargo, es importante considerar la distribución normal de los datos modelados, requisito que no suele cumplirse al trabajar con precipitaciones acumuladas mensuales. Sin este requisito la validación y confiabilidad del modelo son nulas, al ser un modelo multivariado. Alfaro y Pacheco [70] observaron que este método muestra mejoras notables en comparación con los enfoques que dependen únicamente de la información de una estación. A pesar de ello, según Toro et al. [1], se desaconseja la aplicación de métodos de regresión cuando los coeficientes de determinación son inferiores a 0.8. Dado que en este estudio dichos coeficientes están por debajo, se sugiere optar por otro método.

Finalmente, las redes neuronales artificiales representan una estrategia efectiva para estimar la precipitación pluviométrica con mayor precisión, esto es corroborado por la alta correlación entre valores reales y valores simulados, y sus bajos valores de error.

Según Tealab et al. [71], en los últimos años ha habido un crecimiento notable en el interés y la investigación en torno al uso de redes neuronales, lo que ha generado un cuerpo creciente de literatura científica sobre este tema. Este aumento en la atención académica ha resultado en una diversidad de opiniones entre los investigadores, con algunos respaldando entusiastamente el potencial de las redes neuronales y otros señalando sus limitaciones [72]. En el estudio de Baño y Gutiérrez [73] mencionan que la incorporación de un predictor adicional podría potencialmente mejorar los resultados obtenidos. Sin embargo, su aplicación podría desaconsejarse en la elaboración de proyecciones climáticas a largo plazo, dado que algunas variables son altamente parametrizables. Mientras que una limitación podría surgir de la estructura multicapa seleccionada para las redes neuronales artificiales, la cual podría resultar insuficiente en términos del número de capas ocultas. Incrementar su complejidad podría facilitar que el método capture un aprendizaje más profundo de las relaciones físicas entre las variables predictoras y la variable objetivo.

CONCLUSIONES

Tras evaluar diversos métodos estadísticos y matemáticos para rellenar datos pluviométricos en la microcuenca del río Pita, se observó que las redes neuronales artificiales sobresalieron como el método más efectivo. Estas redes demostraron una alta capacidad de generación de datos pluviométricos, con coeficientes de correlación superiores a 0.6 y una proximidad cercana entre los datos observados y los datos simulados, lo que indica un ajuste adecuado sin caer en sobreajuste. Además, se evidenció que el método de Paulhus y Kohler y la transformada de *Wavelet* presentaron desempeños menos satisfactorios. Se destaca que la longitud de los datos de entrada, que abarcan 10 años de registros mensuales (120 datos en total), tuvo un impacto significativo en la calidad de la transformada de *Wavelet*, lo que sugiere que la cantidad de datos puede influir en la captura y representación de variaciones temporales en los datos. Adicionalmente, la resolución temporal de la información debe considerarse, ya que se ha observado que los datos con resolución mensual pueden limitar la eficiencia. Por otro lado, la regresión lineal múltiple también mostró estimaciones prometedoras, especialmente al considerar la naturaleza multivariada del modelo.

DOI: https://doi.org/10.18272/aci.v16i1.3274

El estudio de generación de datos pluviométricos resalta una brecha en la investigación, particularmente en áreas propensas a sequías e inundaciones recurrentes, subrayando la necesidad de abordar esta problemática a nivel de cuenca o microcuenca hidrográfica para mantener información precisa y actualizada, esencial para la gestión del agua y la mitigación de desastres naturales. Se propone también la generalización de las metodologías desarrolladas en este estudio para su aplicación en otras áreas geográficas, especialmente en zonas de páramo, verificando valores de pluviosidad altos, como es el caso de la microcuenca del río Pita. Esto implica adaptar los modelos y técnicas a las particularidades de cada microcuenca, abordando así la escasez de datos pluviométricos y fortaleciendo la capacidad de respuesta ante eventos climáticos extremos.

AGRADECIMIENTOS

Este trabajo se ha desarrollado a partir de las experiencias adquiridas en el Fondo para la Protección del Agua (FONAG), particularmente en el Programa de Educación Ambiental (PEA). El conocimiento y las experiencias compartidas por todo el equipo han sido fundamentales para la elaboración de un trabajo científico sobre recursos hídricos, destinado a contribuir a futuras investigaciones en el Distrito Metropolitano de Quito y otras ciudades de Ecuador.

CONFLICTOS DE INTERESES

Se declara no tener conflicto de intereses en relación a la presente investigación.

CONTRIBUCIONES DE LOS AUTORES

Los autores contribuyeron en todas las etapas de elaboración del presente artículo.

DOI: https://doi.org/10.18272/aci.v16i1.3274

REFERENCIAS

[1] Toro Trujillo, A. M., Arteaga Ramírez, R., Vázquez Peña, M. A. y Ibáñez Castillo, L. A. (2017). Relleno de series diarias de precipitación, temperatura mínima, máxima de la región norte del Urabá Antioqueño. *Revista Mexicana de Ciencias Agrícolas*, *6*(3). doi: https://doi.org/10.29312/remexca.v6i3.640

[2] Benítez-Gilabert, M. y Álvarez-Cobelas, M. (2008). Reconstrucción de series temporales en ciencias ambientales. *Revista Latinoamericana de Recursos Naturales, 4*(3). http://hdl.handle.net/10261/22205

[3] Alfaro, E. J. y Soley, F. J. (2009). Descripción de dos métodos de rellenado de datos ausentes en series de tiempo meteorológicas. *Revista de Matemática: Teoría y Aplicaciones*, *16*(1). doi: https://doi.org/10.15517/rmta.v16i1.1419

[4] Altamirano, C. y Carrillo, P. (2023). *Comparación de técnicas de relleno de datos faltantes de variables meteorológicas en la provincia de Chimborazo*. Escuela Superior Politécnica del Chimborazo. http://dspace.espoch.edu.ec/handle/123456789/19932

[5] Muñoz Herrera, W., Bedoya, O. F. y Rincón, M. E. (2020). Aplicación de redes neuronales para la reconstrucción de series de tiempo de precipitación y temperatura utilizando información satelital. *Revista EIA*, *17*(34). doi: https://doi.org/10.24050/reia.v17i34.1292

[6] Pérez Pelea, L. (2019). Valores atípicos en los datos, ¿cómo identificarlos y manejarlos? *Revista Del Jardín Botánico Nacional*, *40*. https://revistas.uh.cu/rjbn/article/view/6537

[7] Herrera Oliva, C. S., Campos Gaytán, J. R. y Carrillo González, F. M. (2017). Estimación de datos faltantes de precipitación por el método de regresión lineal: Caso de estudio Cuenca Guadalupe, Baja California, México. *Investigación y Ciencia de La Universidad Autónoma de Aguascalientes*, *71*. doi: https://doi.org/10.33064/iycuaa201771598

[8] Sayl, K., Adham, A. y Ritsema, C. J. (2020). A GIS-based multicriteria analysis in modeling optimum sites for rainwater harvesting. *Hydrology*, *7*(3). doi: https://doi.org/10.3390/HYDROLOGY7030051

[9] Cardoso Pereira, S., Marta-Almeida, M., Carvalho, A. C. y Rocha, A. (2020). Extreme precipitation events under climate change in the Iberian Peninsula. *International Journal of Climatology*, *40*(2). doi: https://doi.org/10.1002/joc.6269

[10] Carranza, J. M. G., Cortes, M. A. R., Hernandez, L. A. A., Vega, F. C., Vargas, F. L. R., Belmán, J. U. G. y Rangel, J. C. G. (2021). Relleno de datos faltantes en series de datos de precipitación para la ciudad de Guanajuato. *Jóvenes En La Ciencia: XXVI Verano de La Ciencia,10*.

[11] Gómez Guerrero, J. S. y Aguayo Arias, M. I. (2019). Evaluación de desempeño de métodos de relleno de datos pluviométricos en dos zonas morfoestructurales del Centro Sur de Chile. *Investigaciones Geográficas*, *99*. doi: https://doi.org/10.14350/rig.59837

[12] Palma, K. (2020). *Evaluación del estado del humedal Pugllohuma, perteneciente al Área de Conservación Hídrica Antisana (ACHA), mediante análisis de índices espectrales de imágenes capturadas desde una aeronave no tripulada (UAV)* [Tesis Ingeniería, Escuela Politécnica Nacional]. Repositorio Digital Escuela Politécnica Nacional. http://bibdigital.epn.edu.ec/handle/15000/21131

[13] FONAG. (2014). *Caracterización biofísica y socioeconómica de le cuenca alta del río Guayllabamba, con énfasis en las subcuencas de los ríos Pita y San Pedro y las microcuencas de los ríos orientales Papallacta y Antisana*. FONAG.

[14] Tufiño, S. (2019). *Comportamiento hidrológico de la cuenca del río Pita: Perspectiva con el modelo de planificación hídrica* [Tesis Ingeniería, Universidad de las Américas]. Repositorio Digital Universidad de las Américas. https://dspace.udla.edu.ec/handle/33000/11325

[15] Melo Martínez, C. E., Malagón Márquez, D. A. y Ramírez Forero, D. D. (2019). *Interpoladores determinísticos espacio-temporales, series de tiempo y análisis de datos funcionales para el estudio y predicción de la precipitación en Cundinamarca y Bogotá D.C* [Tesis de Grado, Universidad Distrital Francisco José de Caldas]. Repositorio Institucional Universidad Distrital Francisco José de Caldas http://hdl.handle.net/11349/14699

[16] OMM. (2011). *Guía de prácticas climatológicas N° 100*. Organizacion Mundial Meteorologica.

[17] Rhif, M., Abbes, A. Ben, Farah, I. R., Martínez, B. y Sang, Y. (2019). Wavelet transform application for/in non-stationary time-series analysis: A review. In *Applied Sciences (Switzerland), 9*(7). doi: https://doi.org/10.3390/app9071345

[18] Montgomery, K. (2013). Big Data Now. *Journal of Chemical Information and Modeling, 53*(9). https://pubs.acs.org/toc/jcisd8/53/9

DOI: https://doi.org/10.18272/aci.v16i1.3274

[19] Pinthong, S., Ditthakit, P., Salaeh, N., Hasan, M. A., Son, C. T., Linh, N. T. T., Islam, S. y Yadav, K. K. (2022). Imputation of missing monthly rainfall data using machine learning and spatial interpolation approaches in Thale Sap Songkhla River Basin, Thailand. *Environmental Science and Pollution Research*. doi: https://doi.org/10.1007/s11356-022-23022-8

[20] Portuguez-maurtua, M., Arumi, J. L., Lagos, O., Stehr, A. y Arquiñigo, N. M. (2022). Filling Gaps in Daily Precipitation Series Using Regression and Machine Learning in Inter-Andean Watersheds. *Water (Switzerland)*, *14*(11). doi: https://doi.org/10.3390/w14111799

[21] A. De Asis, C. (2021). Comparison of Normal Ratio Method and Distance Power Method for Estimating Missing Rainfall Data with Three Neighboring Stations. *Journal of Engineering Research and Reports*. doi: https://doi.org/10.9734/jerr/2021/v21i617469

[22] Polpinij, J. y Namee, K. (2021). Comparison of Methods to Estimate Missing Values in Monthly Rainfall Data. *25th International Computer Science and Engineering Conference*. doi: https://doi.org/10.1109/ICSEC53205.2021.9684588

[23] Curci, G., Guijarro, J. A., Di Antonio, L., Di Bacco, M., Di Lena, B. y Scorzini, A. R. (2021). Building a local climate reference dataset: Application to the Abruzzo region (Central Italy), 1930–2019. *International Journal of Climatology*, *41*(8). doi: https://doi.org/10.1002/joc.7081

[24] Papailiou, I., Spyropoulos, F., Trichakis, I. y Karatzas, G. P. (2022). Artificial Neural Networks and Multiple Linear Regression for Filling in Missing Daily Rainfall Data. *Water (Switzerland)*, *14*(18). doi: https://doi.org/10.3390/w14182892

[25] Gunawardena, N., Durand, P., Hedde, T., Dupuy, F. y Pardyjak, E. (2022). Data Filling of Micrometeorological Variables in Complex Terrain for High-Resolution Nowcasting. *Atmosphere*, *13*(3). doi: https://doi.org/10.3390/atmos13030408

[26] Liyew, C. M. y Melese, H. A. (2021). Machine learning techniques to predict daily rainfall amount. *Journal of Big Data*, *8*(1). doi: https://doi.org/10.1186/s40537-02100545-4

[27] Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M. y Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications*, *27*(1). doi: https://doi.org/10.1002/met.1873

[28] Llamas, R. M., Guevara, M., Rorabaugh, D., Taufer, M. y Vargas, R. (2020). Spatial gap-filling of ESA CCI satellite-derived soil moisture based on geostatistical techniques and multiple regression. *Remote Sensing*, *12*(4). doi: https://doi.org/10.3390/rs12040665

[29] Sentop, M. S., Yucel, M. y Ustundag, B. B. (2023). Spatio-Temporal Missing Data Reconstruction by Using Deep Neural Networks in Agricultural Monitoring Systems. *11th International Conference on Agro-Geoinformatics: Agro-Geoinformatics*. doi: https://doi.org/10.1109/Agro-Geoinformatics59224.2023.10233578

[30] Achite, M., Katipoglu, O. M., Şenocak, S., Elshaboury, N., Bazrafshan, O. y Dalkılıç, H. Y. (2023). Modeling of meteorological, agricultural, and hydrological droughts in semi-arid environments with various machine learning and discrete wavelet transform. *Theoretical and Applied Climatology*, *154*(1–2). doi: https://doi.org/10.1007/s00704-023-04564-4

[31] Narimani, R., Jun, C., De Michele, C., Gan, T. Y., Nezhad, S. M. y Byun, J. (2023). Multilayer perceptron-based predictive model using wavelet transform for the reconstruction of missing rainfall data. *Stochastic Environmental Research and Risk Assessment*, *37*(7). doi: https://doi.org/10.1007/s00477-023-02471-8

[32] Vivas, E., de Guenni, L. B., Allende-Cid, H. y Salas, R. (2023). Deep Lagged-Wavelet for monthly rainfall forecasting in a tropical region. *Stochastic Environmental Research and Risk Assessment*, *37*(3). doi: https://doi.org/10.1007/s00477-022-02323-x

[33] Ghamariadyan, M. y Imteaz, M. A. (2021). A wavelet artificial neural network method for medium-term rainfall prediction in Queensland (Australia) and the comparisons with conventional methods. *International Journal of Climatology*, *41*(S1). doi: https://doi.org/10.1002/joc.6775

[34] Park, J., Müller, J., Arora, B., Faybishenko, B., Pastorello, G., Varadharajan, C., Sahu, R. y Agarwal, D. (2023). Long-term missing value imputation for time series data using deep neural networks. *Neural Computing and Applications*, *35*(12). doi: https://doi.org/10.1007/s00521-022-08165-6

[35] Gholami, V. y Sahour, H. (2022). Simulation of rainfall-runoff process using an artificial neural network (ANN) and field plots data. *Theoretical and Applied Climatology*, *147*(1–2). doi: https://doi.org/10.1007/s00704-021-03817-4

DOI: https://doi.org/10.18272/aci.v16i1.3274

[36] Katipoğlu, O. M. (2022). Prediction of missing temperature data using different machine learning methods. *Arabian Journal of Geosciences*, *15*(1). doi: https://doi.org/10.1007/s12517-021-09290-7

[37] Ilaboya, I.R y E, I. O. (2019). Performance of Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) for the Prediction of Monthly Maximum Rainfall in Benin City, Nigeria. *International Journal of Engineering Science and Application*, *3*(1).

[38] Canchala-Nastar, T., Carvajal-Escobar, Y., Alfonso-Morales, W., Loaiza Cerón, W. y Caicedo, E. (2019). Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural networks. *Data in Brief*, *26*. doi: https://doi.org/10.1016/j.dib.2019.104517

[39] FONAG. (2012). Análisis Gobernanza de la Microcuenca de Río Pita. *Fondo Para La Protección Del Agua*. FONAG.

[40] Andrade, A. y Yépez, H. (2014). *Almacenamiento de agua y cuantificación de carbono en el ecosistema páramo dentro de un esquema Global Environment Oulook (GEO), caso de estudio: Páramo de Píntag-Cuenca Alta del Río Pita* [Tesis Ingeniería, Escuela Politécnica Nacional]. Repositorio Digital Escuela Politécnica Nacional. https://bibdigital.epn.edu.ec/handle/15000/7386

[41] Simbaña, K., Romero, D., Yánez, G., Benavides, D., & Navarrete, H. (2019). Evaluación de la calidad del agua del río Pita (Ecuador), implicación para la conservación de la vida acuática y silvestre. *InfoANALÍTICA, 7*(2). doi: https://doi.org/10.26807/ia.v7i2.104

[42] Das, K. R. y Imon, A. H. M. R. (2014). Geometric median and its application in the identification of multiple outliers. *Journal of Applied Statistics*, *41*(4). doi: https://doi.org/10.1080/02664763.2013.856385

[43] Maharana, K., Mondal, S. y Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1). doi: https://doi.org/10.1016/j.gltp.2022.04.020

[44] Barrera-Escoda, A. (2004). *Técnicas de completado de series mensuales y aplicación al estudio de la influencia de la NAO en la distribución de la precipitación en España* [Tesis Diploma de Estudios Avanzados, Universidad de Barcelona]. Grupo de Análisis de situaciones Meteorológicas Adversas. https://zucaina.net/Publicaciones/barrera-dea.pdf

[45] Guijarro, J. (2023). *Guía de uso del paquete de R climatol (versión 4)*. Climatol. https://www.climatol.eu/climatol4.1-es.pdf

[46] Montero, R. (2016). Modelos de regresión lineal múltiple. *Documentos de Trabajo En Economía Aplicada*, *3*(12). https://www.ugr.es/~montero/matematicas/regresion_lineal.pdf

[47] Hui, E. G. M. (2019). *Learn R for Applied Statistics*. APRESS. doi: https://doi.org/10.1007/978-1-4842-4200-1

[48] Gamboa, R. (2015). *Evaluación de modelos empíricos matemáticos y redes neuronales para estimar datos faltantes en estaciones meteorológicas en México* [Tesis, Institución de Enseñanza e Investigación en Ciencias Agrícolas]. Institución de Enseñanza e Investigación en Ciencias Agrícolas. http://colposdigital.colpos.mx:8080/jspui/bitstream/handle/10521/2621/Gamboa_Chel_RO_MC_Hidrociencias_2015.pdf?sequence=1

[49] Quiroz, R., Yarlequé, C., Posadas, A., Mares, V. y Immerzeel, W. W. (2011). Improving daily rainfall estimation from NDVI using a wavelet transform. *Environmental Modelling and Software*, *26*(2). doi: https://doi.org/10.1016/j.envsoft.2010.07.006

[50] Ilyas, H., Raja, M. A. Z., Ahmad, I. y Shoaib, M. (2021). A novel design of Gaussian Wavelet Neural Networks for nonlinear Falkner-Skan systems in fluid dynamics. *Chinese Journal of Physics*, *72*. doi: https://doi.org/10.1016/j.cjph.2021.05.012

[51] Katsavrias, C., Papadimitriou, C., Hillaris, A. y Balasis, G. (2022). Application of Wavelet Methods in the Investigation of Geospace Disturbances: A Review and an Evaluation of the Approach for Quantifying Wavelet Power. *Atmosphere, 13*(3). doi: https://doi.org/10.3390/atmos13030499

[52] Zamrane, Z., Mahé, G. y Laftouhi, N. E. (2021). Wavelet analysis of rainfall and runoff multidecadal time series on large river basins in western north africa. *Water (Switzerland)*, *13*(22). doi: https://doi.org/10.3390/w13223243

[53] Santamaría, F., Cortés, C. A. y Román, Y. F. J. (2012). Uso de la transformada de ondeletas (wavelet transform) en la reducción de ruidos en las señales de campo eléctrico producidas por rayos. *Informacion Tecnologica*, *23*(1). doi: https://doi.org/10.4067/S0718-07642012000100008

[54] Paparoditis, E. (2010). Wavelet Methods in Statistics with R. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *173*(1). doi: https://doi.org/10.1111/j.1467-985x.2009.00624_7.x

DOI: https://doi.org/10.18272/aci.v16i1.3274

[55] Lantz, B. (2015). *Machine Learning with R: Second Edition*. Packt Publishing. https://www.oreilly.com/library/view/machine-learning-with/9781784393908/

[56] Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. In *Mathematical Intelligencer, 27*(2). doi: https://doi.org/10.1007/BF02985802

[57] Dangeti, P. (2017). *Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R*. Packt Publishing. https://dl.acm.org/doi/book/10.5555/3164859

[58] Toral, J. (2012). *Redes Neuronales*. Universidad de Guadalajara.

[59] Raschka, S. y Mirjalili, V. (2019). *Python Machine Learning. Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing. https://books.google.com.ec/books/about/Python_Machine_Learning.html?id=sKXIDwAAQBAJ&redir_esc=y

[60] Sánchez, N. (2020). *Estudio comparativo de modelos de predicción estocásticos y heurísticos aplicados a la estimación de la calidad del aire*. Universitat Oberta de Catalunya.

[61] Ortega, R. M. M., Pendás, L. C. T., Ortega, M. M., Abreu, A. P. y Cánovas, A. M. (2009). El coeficiente de correlacion de los rangos de spearman caracterizacion. *Revista Habanera de Ciencias Medicas*, *8*(2). https://www.redalyc.org/pdf/1804/180414044017.pdf

[62] Ali, S. y Shahbaz, M. (2020). Streamflow forecasting by modeling the rainfall– streamflow relationship using artificial neural networks. *Modeling Earth Systems and Environment*, *6*(3). doi: https://doi.org/10.1007/s40808-020-00780-3

[63] Cartaya, S., Zurita, S. y Montalvo, V. (2016). Métodos de ajuste y homogenización de datos climáticos para determinar índice de humedad de Lang en la provincia de Manabí, Ecuador. *La Técnica: Revista de Las Agrociencias. ISSN 2477-8982*, *16*. doi: https://doi.org/10.33936/la_tecnica.v0i16.540

[64] Poblete, D., Arevalo, J., Nicolis, O. y Figueroa, F. (2020). Optimization of hydrologic response units (Hrus) using gridded meteorological data and spatially varying parameters. *Water (Switzerland)*, *12*(12). doi: https://doi.org/10.3390/w12123558

[65] Krleža, D., Vrdoljak, B. y Brčić, M. (2021). Statistical hierarchical clustering algorithm for outlier detection in evolving data streams. *Machine Learning*, *110*(1). doi: https://doi.org/10.1007/s10994-020-05905-4

[66] Xu, L., Chen, N., Moradkhani, H., Zhang, X. y Hu, C. (2020). Improving Global Monthly and Daily Precipitation Estimation by Fusing Gauge Observations, Remote Sensing, and Reanalysis Data Sets. *Water Resources Research*, *56*(3). doi: https://doi.org/10.1029/2019WR026444

[67] Kuriqi, A., Ali, R., Pham, Q. B., Montenegro Gambini, J., Gupta, V., Malik, A., Linh, N. T. T., Joshi, Y., Anh, D. T., Nam, V. T. y Dong, X. (2020). Seasonality shift and streamflow flow variability trends in central India. *Acta Geophysica*, *68*(5). doi: https://doi.org/10.1007/s11600-020-00475-4

[68] Salazar, A. (2019). *Estimación de niveles medios diarios en estaciones específicas en el río Magdalena a partir de esta variable en estaciones de aguas arriba y de afluentes mediante relaciones empíricas. Casos de estudio: Purificación, Puerto Berrío y Calamar* [Tesis de Maestría, Universidad de los Andes]. Repositorio Institucional Séneca Universidad de los Andes. http://hdl.handle.net/1992/43914

[69] Sifuzzaman, M., Islam, M. R. y Ali, M. Z. (2009). Application of Wavelet Transform and its Advantages Compared to Fourier Transform. *Journal of Physical Sciences*, *13*. https://www.researchgate.net/publication/242602743_Application_of_Wavelet_Transform_and_its_Advantages_Compared_to_Fourier_Transform

[70] Alfaro, R. y Pacheco, R. (2000). Aplicación de algunos métodos de relleno a series anuales de lluvia de diferentes regiones de Costa Rica. *Tópicos Meteorológicos y Oceanográficos*, *7*. https://www.researchgate.net/publication/237217878_Aplicacion_de_algunos_metodos_de_relleno_a_series_anuales_de_lluvia_de_diferentes_regiones_de_Costa_Rica

[71] Tealab, A., Hefny, H. y Badr, A. (2017). Forecasting of nonlinear time series using ANN. *Future Computing and Informatics Journal*, *2*(1). doi: https://doi.org/10.1016/j.fcij.2017.05.001

[72] García Valero, J. A. (2021). *Redes Neuronales Artificiales. Aplicación a la regionalización de la precipitación y temperaturas diarias.* Agencia Estatal de Meteorología AEMET. doi: https://doi.org/10.31978/666-20-028-5

[73] Baño-Medina, J. y Gutiérrez, J. M. (2018). *Deep Convolutional Neural Networks For Feature Selection in Statistical Downscaling*. 8th International Workshop in Climate Informatics.

DOI: https://doi.org/10.18272/aci.v16i1.3274