Regresión lineal multivariable *versus* regresión simbólica a partir de programación genética. Aplicación a la caracterización espectroscópica de aguas residuales urbanas

Multivariate linear regression versus symbolic regression from genetic programming. Application to the spectroscopic characterisation of urban wastewater

Daniel Carreres-Prieto [a1*], Juan T. García [a2], Luis G. Castillo [a3], José M. Carrillo [a4] y Antonio Vigueras-Rodriguez [a5]

[a] Escuela Técnica Superior de Ingeniería de Caminos, Canales y Puertos y de Ingeniería de Minas, Universidad Politécnica de Cartagena, 30203 Cartagena, España.

E-mail: [a1] daniel.carreres@upct.es, [a2] juan.gbermejo@upct.es, [a3] luis.castillo@upct.es, [a4] jose.carrillo@upct.es, [a5] avigueras.rodriguez@upct.es

**Autor para correspondencia*

Recibido: 12/07/2022 Aceptado: 05/10/2022 Publicado: 31/10/2022

Citar como: Carreres-Prieto, D., García, J.T., Castillo, L.G., Carrillo, J.M., Vigueras-Rodriguez, A. 2022. Multivariate linear regression versus symbolic regression from genetic programming. Application to the spectroscopic characterisation of urban wastewater. *Ingeniería del agua*, *26*(4), 261-277. https://doi.org/10.4995/Ia.2022.18073

RESUMEN

Caracterizar en tiempo real las aguas residuales urbanas es clave para poder garantizar una correcta gestión de los recursos hídricos y la protección del medioambiente. A partir de mediciones indirectas, como la espectroscopía molecular que proporciona información sobre las propiedades físico-químicas del agua, es posible determinar la carga contaminante de las aguas residuales empleando modelos matemáticos de correlación. El presente trabajo compara la regresión lineal multivariable y los modelos de regresión simbólica basados en programación genética, para establecer una correlación con la carga contaminante de las aguas residuales. El estudio se ha centrado en la comparativa de modelos para la caracterización de nitrógeno total, fósforo total y nitrógeno en forma de nitrato, considerando 90 muestras de aguas residuales urbanas. Se observa que la regresión simbólica basada en programación genética proporciona una mejora en el ajuste (R^2) de entre el 72.76% y 146.39% respecto a la regresión lineal multivariable.

Palabras claves: Caracterización de aguas residuales, regresión simbólica, modelación heurística, espectrofotometría LED.

ABSTRACT

Characterising urban wastewater in real time is key to ensure the proper management of water resources and environmental protection. From indirect measurements, such as the molecular spectroscopy which provides information on the physicochemical properties of the water, it is possible to determine the pollutant load of wastewater from mathematical correlation models. The research compares multivariate linear regression models and symbolic regression models based on genetic programming to establish a correlation with the pollutant load of the wastewater. The study has focused on the comparison of models for the characterisation of total nitrogen, total phosphorus and nitrogen in the form of nitrate of 90 urban wastewater samples. It is observed that the symbolic regression based on genetic programming provides an improvement in goodness of fit (R2) of between 72.76% and 146.39% with respect to multivariate linear regression.

Keys word: *Wastewater characterisation, symbolic regression, genetic programming, multivariate linear regression, LED spectrophotometry.*

e ISSN: 1886-4996 ISSN: 1134-2196

INTRODUCCIÓN

Conocer la evolución de la carga contaminante de las aguas residuales urbanas en las redes de saneamiento a lo largo del tiempo, especialmente durante episodios de lluvia intensa, es crucial para una mejor optimización de las plantas de tratamiento y las redes de saneamiento, contribuyendo a un mejor cumplimiento de las políticas medioambientales de la Unión Europea (UE) y a alcanzar los objetivos del Pacto Verde Europeo.

El desarrollo de sensores para la monitorización continua de la carga contaminante de las aguas residuales representa un reto desde un punto de vista técnico, debido a la variabilidad de parámetros a caracterizar, a los diferentes procedimientos para llevarlos a cabo, y a las condiciones ambientales existentes, entre otras variables.

La espectrofotometría de longitud de onda variable es una técnica simple y rápida, que proporciona información valiosa sobre las propiedades físico-químicas de las aguas residuales. Diversos trabajos de investigación han generado desarrollos tecnológicos basados en el análisis multiespectral de bajo coste basado en iluminación de diodo emisor de luz (*light emitting diode*, LED), que permiten caracterizar las muestras de agua residual en cuestión de minutos y sin necesidad de utilizar reactivos químicos o someter las muestras a ningún tipo de pretratamiento (Carreres-Prieto *et al.,* 2019; 2020; 2022; Carreres-Prieto, 2021).

La respuesta espectral presenta una correlación con la carga contaminante de las aguas, aunque debido a la complejidad de la matriz de agua, donde se combinan y mezclan múltiples sustancias, es necesario disponer de una extensa campaña experimental que permita obtener modelos específicos para cada parámetro analítico que se desee estimar. A lo largo de la última década se han presentado numerosos trabajos de investigación que buscan determinar diversos parámetros contaminantes propios de las aguas residuales urbanas a partir de la correlación indirecta con la respuesta espectral de las muestras en un cierto rango de longitudes de onda (Lepot *et al.*, 2016; Mesquita *et al.*, 2017). Entre las técnicas más usadas están la regresión lineal multivariable (RLM) que requiere una distribución normal del parámetro que se desee estimar, y que la relación entre la variable respuesta y los regresores, sea de tipo lineal. En caso de ser una variable no lineal, en ocasiones es posible llevar a cabo transformaciones en la variable respuesta, pero no siempre. En caso contrario, puede que los resultados del estudio puedan no ser extrapolables a otros conjuntos de datos.

Dentro del campo de la inteligencia artificial, las técnicas de análisis basadas en programación genética (PG) permiten obtener modelos de estimación más eficaces que los proporcionados por la RLM. La PG es una metodología de modelización basada en modelos evolutivos de aprendizaje, que mediante un proceso de crecimiento, cruce y mutación análoga al observado en el desarrollo biológico de cualquier especie. Mediante la regresión simbólica, que es una técnica dentro de la PG, se busca la sintetización de los modelos en ecuaciones matemáticas fácilmente interpretables.

En este trabajo se analiza la bondad de la programación genética como herramienta para la obtención de modelos de la carga contaminante de las aguas a partir de mediciones indirectas, frente a los resultados obtenidos con los RLM.

El resto del artículo se organiza del siguiente modo:

- En la sección 2 se describe la naturaleza de los datos de partida para el desarrollo de este estudio, así como las características y principio de funcionamiento de la programación genética.
- En la sección 3 se compara el ajuste de los modelos de estimación basados en regresión lineal multivariable y programación genética (regresión simbólica) para los siguientes parámetros: Nitrógeno total (NT), Fósforo total (PT), y Nitrógeno en forma de nitrato ($N\text{-}NO_3^-$).
- En la sección 4 se muestra una comparativa entre el rendimiento de la PG y RLM como técnicas de modelización.
- En la sección 5, se recogen las principales conclusiones alcanzadas en el presente trabajo de investigación en base a los resultados de las comparativas entre la PG y RLM.

MATERIAL Y MÉTODOS

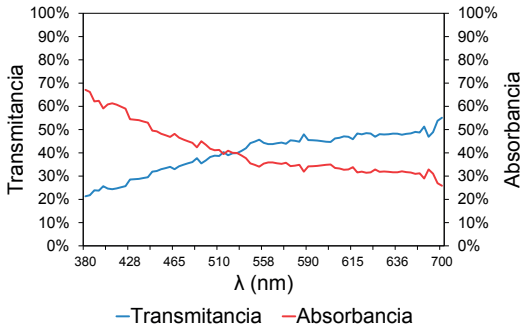
Campaña experimental

Para llevar a cabo el presente trabajo de investigación se han analizado y caracterizado 90 muestras de agua residual obtenidas en la línea de entrada (agua bruta) de la Estación Depuradora de Aguas Residuales (EDAR) de Cabezo Beaza, situada en la ciudad de Cartagena, entre los meses de junio de 2019 y abril de 2020.

Las muestras han sido caracterizadas en el laboratorio de la EDAR, así como por un equipo de espectrofotometría LED desarrollado por los autores (Carreres-Prieto *et al.*, 2020), que permite obtener la respuesta espectral de las muestras de agua dentro del rango de 380-700 nm sin someterla a pretratamientos o reactivos químicos.

Las muestras de agua bruta analizadas proceden de una muestra integrada en un volumen de 5 litros a lo largo de 24 horas, siendo tomadas de la etapa de pretratamiento de la planta. Las pruebas realizadas en el laboratorio se realizaron siguiendo las recomendaciones de los Métodos Estándar (SM) (APHA–AWWA–WPCF, 1998) y a la Organización Mundial de normalización (ISO), empleándose la técnica SM 4500-NC para el nitrógeno total, SM 4500-P B para el fósforo total, e ISO 7890-1 para el $N\text{-}NO_3^-$.

En la Figura 1 se muestra un ejemplo de respuesta espectral obtenido con una muestra de agua residual bruta. La figura también muestra su caracterización analítica para una mayor claridad.



Parámetros	Valor
Demanda Química de Oxígeno (DQO)	929 mg/L
Demanda Bioquímica de Oxígeno (DBO_5)	600 mg/L
Sólidos Suspendidos Totales (SST)	352 mg/L
Fósforo Total (FT)	11.3 mg/L
Nitrógeno Total (NT)	71 mg/L
Nitrógeno en forma de nitrato ($N\text{-}NO_3^-$)	0.5 mg/L
Potencial de Hidrógeno (PH)	7.43
Conductividad (C)	2870 µS/cm

Figura 1 | Ejemplo de respuesta espectral para una muestra de agua bruta.

Regresión lineal multivariable

Se trata de una técnica muy extendida por su sencillez de implementación, la cual trata de ajustar modelos de carácter lineal a partir de unos regresores de entrada que guardan alguna correlación de tipo lineal con la variable respuesta a estimar, a partir de una serie de predictores ponderados por unos coeficientes. La modelización mediante esta técnica se ve muy influenciada por la presencia de valores atípicos o extremos, que pueden afectar en la bondad del ajuste, sobrevalorándola. Por otro lado, dado el criterio de selección de regresores, este tipo de técnicas se ve influenciada por el tamaño de la población, dado que si este no es suficientemente alto, los predictores que no son realmente influyentes en el modelo, podrían parecer serlo, lo que afecta de forma negativa al proceso de modelización. Existen distintas técnicas para llevar a cabo este proceso de selección de predictores y su ponderación. Entre estas destaca la denominada "por Pasos", que en base a la significación o importancia (P-Valor) de los regresores, introduce aquellos que, dentro del modelo, presentan una mayor significación y extrae aquellos menos representativos, por medio de un proceso iterativo que permite una mejor optimización de los modelos.

Programación genética. Regresión simbólica

Existen diversas técnicas dentro del campo de la inteligencia artificial, como las redes neuronales que permiten lograr ajustes no lineales a partir de un proceso de aprendizaje similar al llevado a cabo por la programación genética, pero estos modelos

no se materializan en una función matemática. Sin embargo, las redes neuronales resultan de gran complejidad poder entender el funcionamiento de dicha red.

Dentro del campo de la programación genética, la rama de regresión simbólica permite sintetizar los resultados del aprendizaje en expresiones matemáticas sencillas que puedan ser implementadas en sistemas de bajo coste y con capacidades de computación, con la ventaja añadida de ser fáciles de comprender, representar y configurar (Zelinka *et al.*, 2005). Este tipo de técnica de análisis basado en programación genética, buscan la generación de árboles compuestos de constantes, variables y operaciones matemáticas, los cuales van generando/optimizando durante una fase de entrenamiento que permitan modelizar una cierta variable respuesta en base a unos regresores dados.

Esta técnica parten de una población inicial de individuos (expresiones matemáticas) denominadas "padres", donde su genoma está conformado por las variables (longitudes de onda en este caso) así como la disponibilidad de elección entre una serie operaciones matemáticas o relación entre los elementos. En la Figura 2 se muestra una vista simplificada del proceso de generación, cruce y mutación llevado a cabo durante la programación genética.

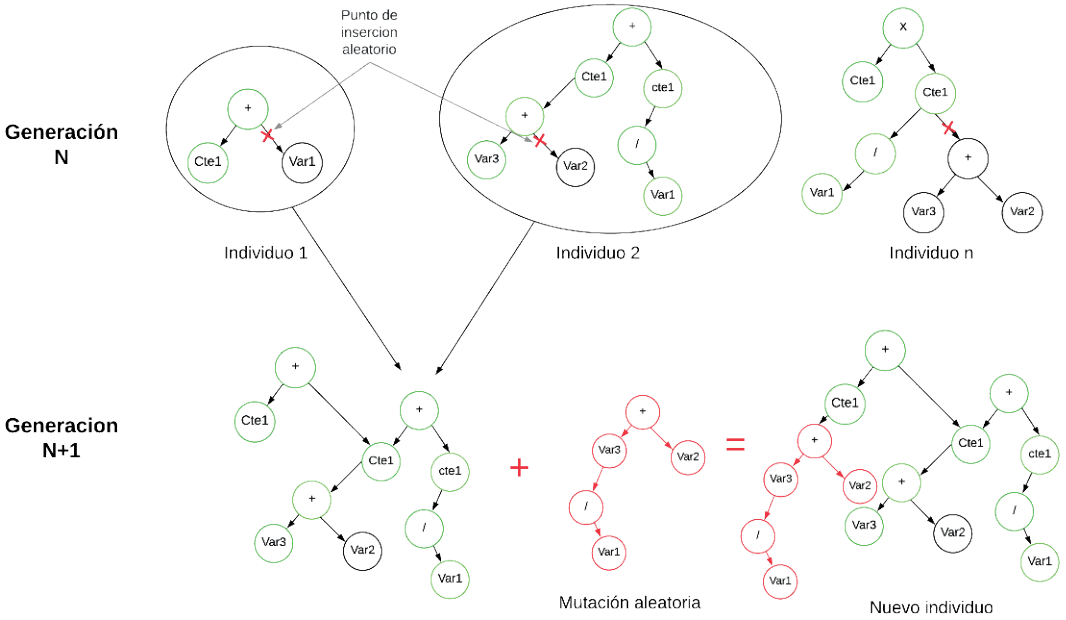


Figura 2 | Esquema de generación de individuos mediante programación genética.

De forma análoga a cualquier proceso biológico de evolución, los progenitores se cruzan entre sí para dar lugar a nuevos individuos llamados "hijos", que heredan parte de las características de sus padres. Sin embargo, si los progenitores poseen malos genes (lo cual es habitual en las primeras generaciones), los hijos serán al menos tan malos como los padres, lo que conduciría a una decadencia genética que se acentuaría generación tras generación, dificultando la obtención de modelos con un buen ajuste. Este problema se acentúa si el número de individuos de partida es bajo, dado que la falta de variedad genética conduce, al igual que sucede en la naturaleza, a una degradación genética muy rápida.

Con el fin de paliar esta situación, se considera que algunos de los nuevos individuos presenten una "mutación", definida como una función matemática generada al azar, la cual es insertada en algún punto aleatorio de su genoma (que por comodidad se suele representar en forma de árbol como en la Figura 2). De esta forma, los nuevos individuos consistirán en expresiones matemáticas que harán uso de parte (o toda) de la expresión matemática de sus padres (combinada de alguna forma) a la que se introducirán otras expresiones generadas de forma aleatoria.

En cada generación, nacen nuevos individuos con características diferentes. De estos, sólo aquellos que son capaces de desenvolverse mejor al entorno, es decir, de predecir con mayor precisión la variable respuesta, sobreviven para engendrar la siguiente generación, y así se repite el proceso anterior un cierto número de generaciones definidas por el usuario (ver Figura 2). En la última generación, el resultado óptimo será aquel individuo que presente el mejor ajuste.

Esta técnica, presenta la peculiaridad de que permite obtener en cada ejecución uno de los infinitos modelos posibles de estimación, presentando todos ellos un nivel de ajuste muy similar entre sí. Esto es debido a que el número de generaciones máximas permitidas en superior a las necesarias para la generación de modelos con altos niveles de estimación. Por ello, aunque exista un componente aleatorio en cuanto al cruce y mutación de los individuos en cada ejecución del entrenamiento, el alto número de generaciones (20), hacer que puedan converger en modelos con ajustes similares.

Como podemos observar en la Figura 2, tanto la mutación (función matemática generada de forma aleatoria), como el punto de cruce o inserción de genes, es aleatorio. Esto redunda en una riqueza genética que hace más factible que, al cabo de un cierto número de generaciones, y con el cruce y generación de un suficiente número de individuos en cada una de ellas, se logre alcanzar un modelo (individuo) con un ajuste superior al que podría lograrse con otras técnicas.

Sin embargo, el principal problema que presenta este tipo de técnicas es que precisan disponer de cierta cantidad de datos para poder llevar a cabo el estudio. Aunque no existe un número mínimo de datos necesario, se debe tener en cuenta que una parte de los datos debe emplearse única y exclusivamente para entrenar el modelo, es decir, para que aprenda a extraer patrones de los datos de entrada, mientras que el resto de datos se emplean para evaluar cómo se comporta el modelo con datos con los que nunca ha trabajado previamente. El porcentaje depende del número de datos que disponga el problema. Como criterio general, si el número de datos es menor 100, se recomienda una relación 80-20% con el fin de disponer de suficientes datos para la extracción de patrones, mientras que para muestras mayores se recomienda una relación en torno al 70-30%. Dado que el número total de muestras a analizar es de 90, para todos los modelos de regresión simbólica presentados en el presente trabajo de investigación, ha empleado el criterio 80-20%.

La versatilidad de esta técnica permite configurar tanto la longitud de las expresiones matemáticas resultantes, como las funciones que se incluyen en ella (funciones trigonométricas, exponenciales, logarítmicas, etc.), así como el número de variables a introducir en el modelo, entre otras cuestiones, facilitando lograr modelos con mejores ajustes.

Implementación

Como paso previo al cálculo de los modelos basados en regresión lineal multivariable, así como programación genética, se ha llevado a cabo un preprocesado de los datos con el fin de segregarlos en dos conjuntos con distribuciones similares (training y test) con el fin de poder lograr una modelización más fidedigna. Para llevar a cabo la división de los datos, se ha implementado un programa en Python, que divide el dataset en dos conjunto (training y test), de forma aleatoria y que analiza la función de distribución de cada uno de los conjuntos resultantes, realizando nuevas reordenaciones aleatorias hasta lograr una distribución similar en ambos conjuntos.

La regresión simbólica es una herramienta de estimación estadística que puede ser aplicada a una gran variedad de campos de estudio. Es posible implementar la regresión simbólica de una forma sencilla a partir de diferentes métodos, como la regresión dispersa (Quade *et al.*, 2016; Brunton *et al.*, 2018), el método AI-Feynman desarrollado por Udrescu *et al.*, 2020; así como mediante programación genética con ayuda de librerías como gpLearn orientadas al lenguaje de programación Python.

En la actualidad existen paquetes de software y librerías de acceso libre que permiten implementar este tipo de técnicas de una forma sencilla. Entre los paquetes de software disponible, destacan herramientas como HeuristicLab (Wagner *et al.*, 2014), TuringBot (TuringBot, 2020) o el módulo de Matlab llamado GPTIPS (Searson *et al.*, 2010). En este trabajo de investigación, se ha hecho uso del software gratuito HeuristicLab, el cual proporciona diversos algoritmos de programación genética, que permiten la generación de modelos con elevados ajustes, en un reducido número de generaciones mediante una interfaz de usuario sencilla.

Por otro lado, se ha hecho uso del software SPSS para el cálculo de los modelos de regresión lineal multivariable mediante la técnica de "por pasos", que permite agregar y eliminar regresores del modelo en función de su nivel de significación, logran modelos más precisos (Figura A2).

En el Apéndice, se recoge de forma más extensa este proceso de implementación, incluyendo indicaciones sobre el manejo de ambos paquetes de software.

Dada la popularidad creciente del uso de la programación genética, en el Apéndice, se describe el manejo de una librería para Python llamada gpLearn, que permite la implementación de modelos de programación genética de una forma simple, junto con una mayor flexibilidad en cuanto a la implementación de nuevos algoritmos de generación.

RESULTADOS

En esta sección se muestra una comparativa entre los niveles de ajuste obtenidos para los modelos generados mediante RLM y RS-AG para diferentes parámetros contaminantes, con el fin de dejar patente las características que cada técnica posee. Se ha designado como T_x y A_x, a los valores de transmitancia y absorbancia obtenidos a una longitud de onda x.

Fruto del proceso de aprendizaje evolutivo, los modelos de regresión simbólica determinan qué variables (valores de transmitancia y absorbancia a distintas longitudes de onda) tienen un mayor impacto en el proceso de caracterización de cada uno de los parámetros, con el fin de lograr estimaciones más próximas a los valores de referencia medidos en laboratorio.

Con el fin de obtener las mejores estimaciones, se ha empleado el software HeuristicLab para llevar a cabo el análisis de regresión simbólica y el software SPSS para el cálculo del modelo de regresión lineal multivariable. En el estudio no se han eliminado los atípicos estadísticos con el fin de lograr modelos más robustos y mostrar las ventajas de la regresión simbólica.

Nitrógeno Total

Regresión lineal multivariable

En la Ecuación (1) se muestra el modelo de estimación de la concentración total de nitrógeno (NT) a partir de la respuesta espectral, generado mediante regresión lineal multivariable, con un coeficiente de correlación ajustado $\overline{R^2}$= 30.5%.

$$NT_{(mg/L)}=184.431 - 224.268 \times T_{420} - 88.047 \times A_{380} \quad (1)$$

Programación genética

En la Ecuación (2) se muestra el modelo de estimación del NT a partir del análisis espectrofotométrico, generado mediante programación genética. Este modelo presenta una bondad en el ajuste del 75.15%.

$$NT_{(mg/L)}= (c_0 \times T_{583} - c_1 \times T_{480}) \times (c_2 \times A_{605} + c_3 \times A_{495}) + \frac{c_4 \times A_{600}}{(c_5 \times A_{650} - c_6 \times T_{650})} \times c_7 + c_8 \quad (2)$$

siendo: $c_0 = 13.818$, $c_1 = 12.365$, $c_2 = 45.425$, $c_3 = –26.948$, $c_4 = 0.895$, $c_5 = –11.916$, $c_6 = –14.660$, $c_7 = 9.486$, y $c_8 = 9.853$.

Fósforo Total

Regresión lineal multivariable

En la Ecuación (3) se muestra el modelo de estimación de la concentración total de fósforo (PT) a partir de la respuesta espectral, calculado mediante regresión lineal multivariable. Este modelo ha presentado un ajuste muy bajo, con un $\overline{R^2}$= 20.2%.

$$PT_{(mg/L)} = 16.482 - 19.869 \times T_{385} - 89.752 \times A_{550} + 85.294 \times A_{565} \quad (3)$$

Programación genética

En la Ecuación (4), se muestra el modelo de estimación de fósforo total calculado a partir de programación genética, el cual, pese a presentar una estructura más compleja que el anterior, permite lograr una bondad del ajuste del 48.29%.

$$PT_{(mg/L)} = \left(\frac{(c_1 \times T_{425} - c_1)}{\log(c_2 \times A_{580})} - \left((c_3 \times \mathrm{T}_{505} - c_4 \times T_{455}) - c_5 \times T_{435}\right)\right) \times c_6 + c_7 \quad (4)$$

donde: $c_0 = -1.150$, $c_1 = -0.183$, $c_2 = 2.164$, $c_3 = -7.987$, $c_4 = -6.320$, $c_5 = -3.411$, $c_6 = 9.207$, y $c_7 = 4.288$.

Nitrógeno en forma de nitrato

Regresión lineal multivariable

En la Ecuación (5) se muestra la expresión para la estimación de la concentración de nitrógeno en forma de nitrato ($N\text{-}NO_3^-$) en muestras de agua residual urbana mediante regresión lineal multivariable, logrando una estructura sencilla con regresores perteneciente a las regiones violeta (380 - 427nm) y azul (427 - 476nm) del espectro. Sin embargo, este modelo presenta un $\overline{R^2} = 47.1\%$.

$$\mathrm{N\text{–}NO_{3(mg/L)}^-} = -12.622 + 22.229 \times T_{385} + 16.158 \times A_{425} - 72.614 \times A_{428} + 71.913 \times A_{445} \quad (5)$$

Programación genética

El modelo mostrado en la Ecuación (6) calculado mediante programación genética, presenta $\overline{R^2}$ un mucho más alto, de 81.37%, a costa de emplear un mayor número de longitudes de onda que el modelo anterior, pertenecientes a las regiones violeta (380-427 nm), amarilla (570-571 nm), naranja (581-618 nm) y roja (618-780 nm) del espectro.

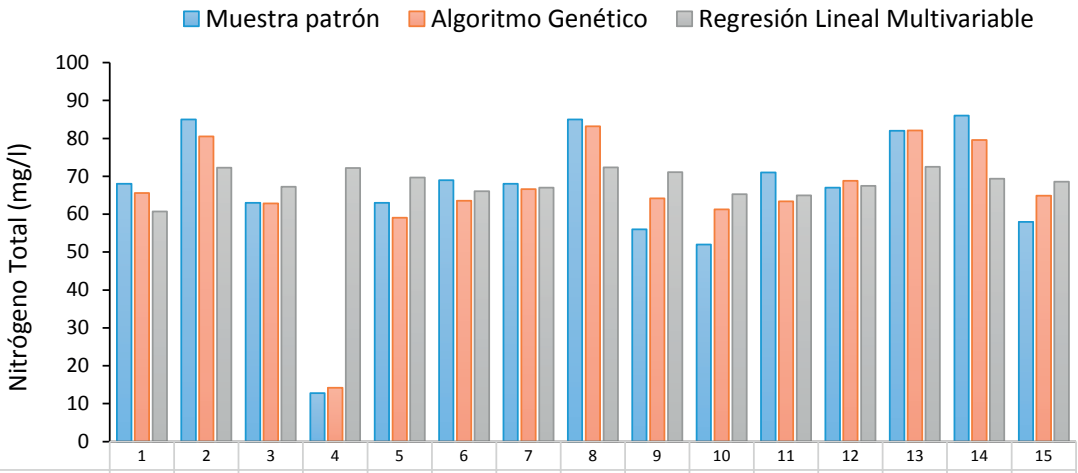
$$\mathrm{N\text{–}NO_{3(mg/L)}^-} = \frac{\left(\left(\left(\frac{c_0 \times T_{590}}{c_1 \times A_{609}^2 \times c_2} + c_3 \times c_4 \times c_5 \times A_{574}\right) + c_6 \times A_{380}\right) + c_7\right)}{c_8 \times A_{700}} \times c_9 + c_{10} \quad (6)$$

siendo: $c_0 = 1.288$, $c_1 = 1.787$, $c_2 = 1.823$, $c_3 = -5.336$, $c_4 = 7.704$, $c_5 = 2.330$, $c_6 = 0.872$, $c_7 = -9.879$ y, $c_8 = 0.990$, $c_9 = 0.016$, y $c_{10} = 3.155$.

DISCUSIÓN

Nitrógeno total

En lo relativo a los modelos de determinación de la concentración total de nitrógeno presente en las aguas residuales urbanas, los niveles de correlación obtenidos en base al modelo de programación genética recogido por la Ecuación (2) son más elevados que el obtenido mediante RLM (Ecuación (1)), presentando una estructura simple aunque con dos regresores más que el modelo de RLM. La mejor correlación se observa en la Figura 3, donde se ha realizado una comparativa entre 15 muestras tomadas al azar, observando que, para la mayoría de casos, los valores estimados de nitrógeno total mediante RS-AG (naranja), se aproximan más a los valores de referencia medidos en laboratorio (azul) que los proporcionados por la RLM (gris).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Muestra patrón	68,000	85,000	63,000	12,800	63,000	69,000	68,000	85,000	56,000	52,000	71,000	67,000	82,000	86,000	58,000
Algoritmo Genético	65,565	80,486	62,863	14,176	59,061	63,542	66,606	83,200	64,132	61,271	63,424	68,788	82,101	79,579	64,915
Regresión Lineal Multivariable	60,734	72,225	67,252	72,184	69,702	66,090	66,964	72,324	71,053	65,236	64,964	67,446	72,487	69,382	68,566

Figura 3 | Comparativa para 15 muestras tomadas al azar entre los niveles de estimación de NT obtenidos con el modelo basado en RLM (Ecuación (1), gris), el basado en RS-AG (Ecuación (2), naranja), y los valores de referencia medidos en laboratorio (azul).

Aunque hacen uso de un diferente número de longitudes de ondas, ambos modelos mantienen una distribución de pesos homogénea, es decir, que todos los regresores (longitudes de ondas) presentes en cada modelo, tiene una significación similar entre sí. Esto es consistente con trabajos previos llevados a cabo sobre la caracterización del nitrógeno a partir de la respuesta espectrofotométrica dentro del espectro visible (Carreres-Prieto *et al.*, 2022).

La estructura no lineal de la Ecuación (2) permite alcanzar niveles de correlación más elevados (75.15%). Esto se observa con claridad en la Figura 4, donde se muestra un diagrama de dispersión entre los valores de referencia (muestra patrón, laboratorio) y aquellos estimados mediante la RS-AG (Figura 4a) y mediante RLM (Figura 4b), apareciendo una mayor convergencia de los puntos entorno a la diagonal en el primer caso respecto al segundo.

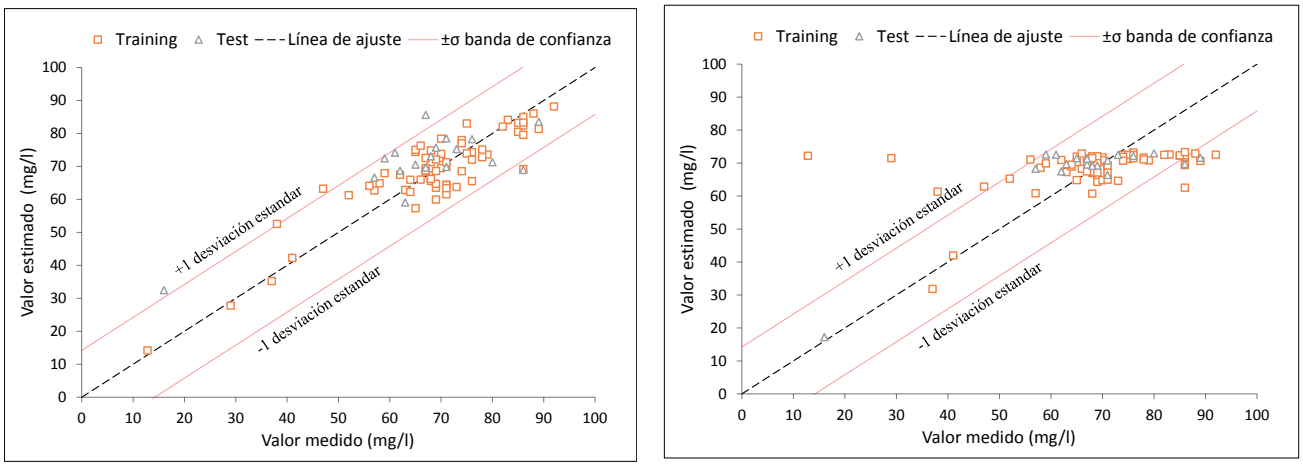
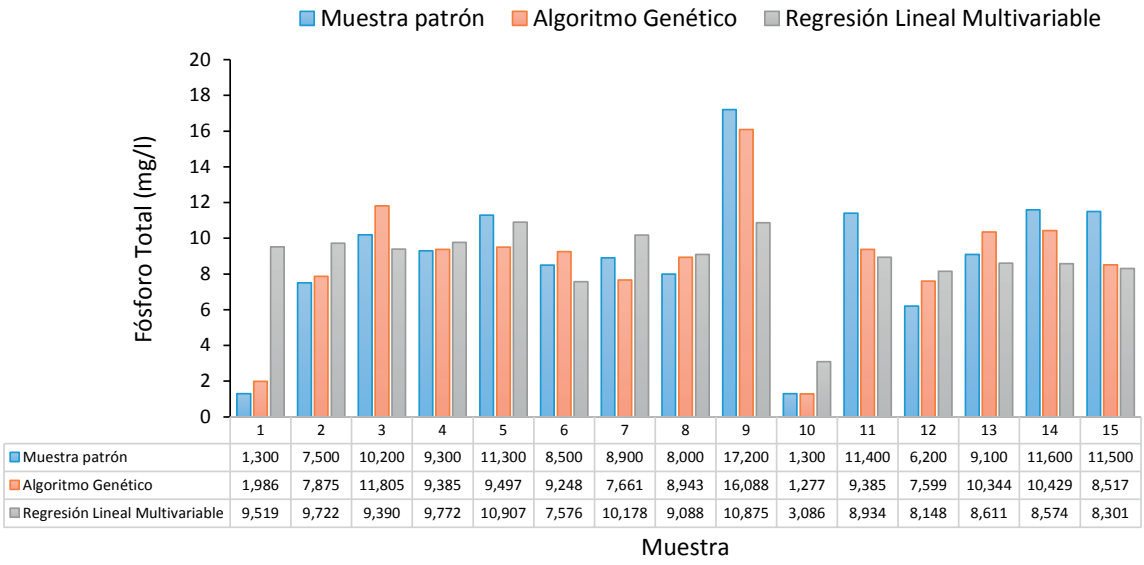


Figura 4 | Diagrama de dispersión modelos de nitrógeno total calculados mediante: a) programación genética; b) regresión lineal multivariable.

Fósforo total

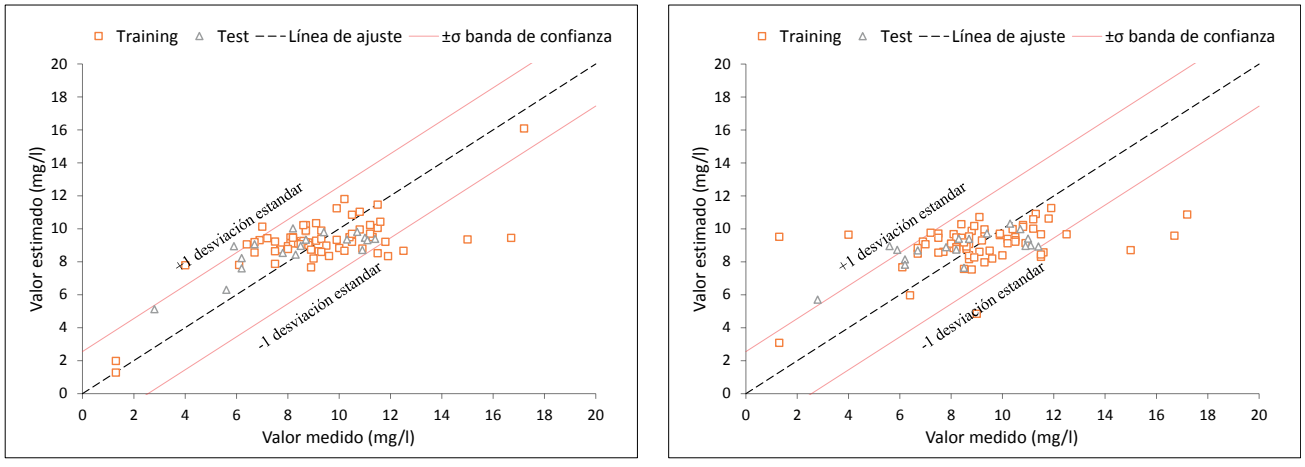
Resultados similares a los obtenidos el nitrógeno total, se muestran en los modelos de determinación de fósforo total (Figura 5). En la mayoría de las muestras seleccionadas al azar se aprecia que las estimaciones realizadas mediante el modelo RLM (gris) presentan deviaciones mayores respecto a las obtenidas mediante RS-AG (naranja). Destacan los valores obtenidos en la muestra número 1, donde para un valor de referencia de 1.3 mg/L, el modelo basado en programación genética estima un valor muy próximo de 1.99 mg/L, frente a los 9.52 mg/L de la regresión lineal.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Muestra patrón	1,300	7,500	10,200	9,300	11,300	8,500	8,900	8,000	17,200	1,300	11,400	6,200	9,100	11,600	11,500
Algoritmo Genético	1,986	7,875	11,805	9,385	9,497	9,248	7,661	8,943	16,088	1,277	9,385	7,599	10,344	10,429	8,517
Regresión Lineal Multivariable	9,519	9,722	9,390	9,772	10,907	7,576	10,178	9,088	10,875	3,086	8,934	8,148	8,611	8,574	8,301

Figura 5 | Comparativa para 15 muestras tomadas al azar entre los niveles de estimación de TP obtenidos con el modelo basado en RLM (Ecuación (3), gris), el basado en RS-AG (Ecuación (4), naranja), y los valores de referencia medidos en laboratorio (azul).

La introducción de operaciones matemáticas más complejas (logaritmo) en la Ecuación (4) permite alcanzar una mayor precisión, que se refleja en los diagramas de dispersión de la Figura 6, con mayor concentración de los puntos sobre la recta de ajuste en el caso de la RS-AG (Figura 6a).

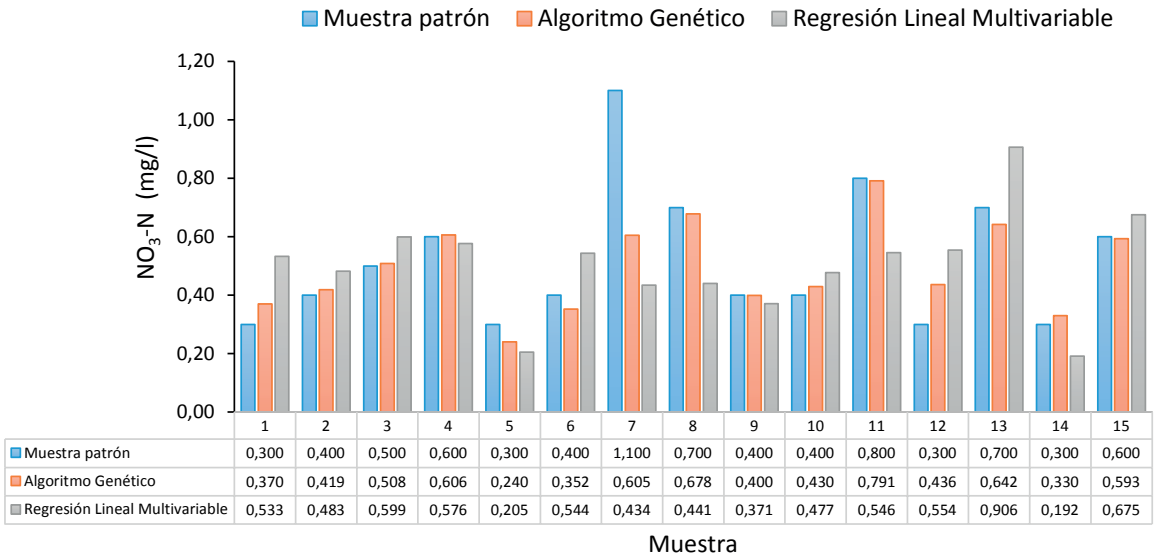


a) b)

Figura 6 | Diagrama de dispersión fósforo total calculado mediante: a) programación genética; b) regresión lineal multivariable.

Nitrógeno en forma de nitrato

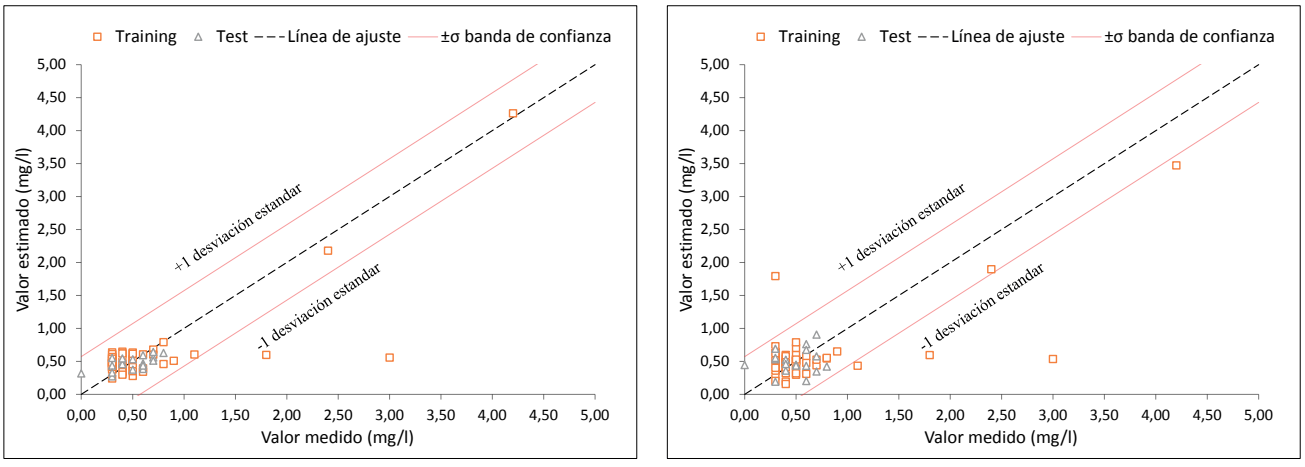
En la Figura 7 se observa que los valores estimados de nitrógeno en forma de nitrato a partir de la respuesta espectrofotométrica calculados mediante regresión lineal multivariable tienden a presentar mayores desviaciones a las estimaciones proporcionadas mediante programación genética, con datos muy próximos a los valores de referencia, especialmente cuanto mayor es el valor de referencia. Por ejemplo, en la muestra número 13, para un valor de 4.2 mg/L medido en laboratorio, el modelo basado en programación genética estima un valor de 4.26 mg/L, frente a los 3.47 mg/L de la regresión lineal multivariable. Sin embargo, en la muestra 7 ambos métodos tienden a subestimar el valor medido en laboratorio.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Muestra patrón	0,300	0,400	0,500	0,600	0,300	0,400	1,100	0,700	0,400	0,400	0,800	0,300	0,700	0,300	0,600
Algoritmo Genético	0,370	0,419	0,508	0,606	0,240	0,352	0,605	0,678	0,400	0,430	0,791	0,436	0,642	0,330	0,593
Regresión Lineal Multivariable	0,533	0,483	0,599	0,576	0,205	0,544	0,434	0,441	0,371	0,477	0,546	0,554	0,906	0,192	0,675

Figura 7 | Comparativa para 15 muestras tomadas al azar de los niveles de estimación de $N\text{-}NO_3^-$ obtenidos con el modelo basado en RLM (Ecuación (5), gris), el basado en RS-AG (Ecuación (6), naranja), y los valores de referencia medidos en laboratorio (azul).

El mejor ajuste puede observarse en una mayor concentración de puntos en el diagrama de dispersión de la Figura 8a (RS-AG) respecto al método basado en RLM (Figura 8b).



a) b)

Figura 8 | Diagrama de dispersión NO_3^--N calculado mediante: a) programación genética; b) regresión lineal multivariable.

Como se puede comprobar, la programación genética permite obtener mejores ajustes que otras técnicas de análisis tradicional como la RLM.

Las ventajas del uso de modelos de programación genética con respecto a técnicas tradicionales de análisis, queda constatado en los numerosos trabajos de investigación presentes en la bibliografía, como los trabajos de Leardi *et al.* (1992), que resalta la conveniencia de las técnicas de programación genética como método de selección de características (regresores), logrando modelos con mejores niveles de predicción, pero con un menor número de variables que otras técnicas, lo que los hace más robustos. Así mismo, los trabajos de Niazi *et al.* (2012), que aborda el uso de este tipo de técnicas en el campo de la quimiometría, destacando su utilidad cuando la variable respuesta no es continua, debiendo destacar el trabajo de Otto (2016), que indica que este tipo de técnicas es de gran utilidad en problemas complejos y con alta dimensionalidad, resaltando específicamente su adecuación en aplicaciones de selección de longitudes de onda en análisis espectrofotométrico multicomponente.

Además, dado a que este tipo de técnicas se basan en un proceso de aprendizaje basado en la extracción de patrones de los datos de entrada, se logra una mayor generalización de los modelos (siempre y cuando exista variabilidad en los datos de entrada, y estos a su vez, sean representativos), lo que permite superar en mejor medida, la limitación los modelos de RLM.

Por lo tanto, gracias a este proceso, los modelos calculados mediante PG son más generalizables y extrapolables que los de RLM, gracias a ese proceso de aprendizaje y extracción de patrones de los datos de entrada.

CONCLUSIONES

En este trabajo se ha llevado a cabo un estudio comparativo entre la aplicación de técnicas de regresión lineal multivariable (RLM) y de regresión simbólica a partir de programación genética (RS-PG), considerando 90 muestras de agua de entrada de una EDAR de la ciudad de Cartagena, sin eliminar atípicos estadísticos. Los resultados obtenidos dejan patente las ventajas que proporciona el uso de RS-AG a la hora de lograr modelos más precisos para caracterizar la carga contaminante de las aguas residuales brutas, concretamente, la determinación de nitrógeno total, fósforo total y nitrógeno en forma de nitrato.

Los modelos de programación genética aprenden a extraer patrones de los datos de entrada, lo que permite obtener tanto modelos lineales como no lineales más precisos. Además, no están sujetos a la necesidad de que la variable respuesta siga una distribución normal para poder extrapolar los resultados, limitación que sí presenta la regresión lineal multivariable.

Por su carácter evolutivo, la programación genética permite alcanzar en cada ejecución, una de las infinitas soluciones igualmente válidas, frente a los modelos de regresión lineal multivariable, basados en el *p*-valor que toman los regresores en relación a los ya presentes en el modelo tras cada iteración.

La programación genética, ha proporcionado niveles de ajustes muy elevados respecto a los proporcionados mediante la regresión lineal multivariable. El Nitrógeno Total, ha experimentado una mejora en su R^2 del 146.39%, similar a la mejora del Fósforo Total, con un 139.06%, y seguido por la mejora del Nitrógeno en forma de Nitrato, con un 72.76% de mejora de la bondad del ajuste.

Aunque el uso de técnicas de regresión lineal multivariable puede alcanzar modelos de correlación con niveles de ajustes similares a los proporcionados con técnicas de programación genética, dicha obtención está supeditada a la naturaleza de los datos y por supuesto, a la variable respuesta. En el caso particular de la caracterización de la carga contaminante de las aguas residuales, si bien es posible lograr ajustes similares entre ambas técnicas, el uso de programación genética permite obtener mejores ajustes, gracias al empleo de modelos con diferentes estructuras (operaciones matemáticas) con niveles de ajuste similares.

AGRADECIMIENTOS

El primer autor agradece la financiación recibida de la Fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia (España), a través del programa de capacitación para nuevos investigadores en áreas específicas de interés para la industria

y alta capacidad de transferencia de los resultados de investigación generados, titulado: "Subprograma Regional de Contratos de Formación de Personal Investigador en Universidades y OPIs" (Mod. B, Ref. 20320/FPI/17)".

El presente trabajo de investigación ha sido financiado mediante el proyecto MONITOCOES: *New intelligent monitoring system for microorganisms and emerging contaminants in sewage networks*. Referencia: RTC2019-007115-5, otorgado por el Ministerio de Ciencia e Innovación – Agencia Estatal de Investigación, dentro de la convocatoria RETOS COLABORACIÓN 2019.

El equipo desarrollado también ha recibido financiación para su industrialización a través del programa "Proof of Concept" de la Fundación Séneca, en el marco del proyecto " Equipo de MONITORIZACIÓN en Tiempo REAl de Contaminantes en Aguas Residuales (MONITOREA)." (21662/PDC/21.).

REFERENCIAS

APHA–AWWA–WPCF, Standard Methods for the Examination of Water and Wastewater, twentieth edition, Washington, DC, 1998

Brunton, S.L., Proctor, J.L., Kutz, J.N. 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, *113*(15), 3932-3937. https://doi.org/10.1073/pnas.1517384113

Carreres-Prieto, D., García, J.T., Cerdán-Cartagena, F., Suardiaz-Muro, J. 2019. Spectroscopy transmittance by LED calibration. *Sensors*, *19*(13), 2951. https://doi.org/10.3390/s19132951

Carreres-Prieto, D., García, J.T., Cerdán-Cartagena, F., Suardiaz-Muro, J. 2020. Wastewater quality estimation through spectrophotometry-based statistical models. *Sensors*, *20*(19), 5631. https://doi.org/10.3390/s20195631

Carreres-Prieto, D. 2021. Contribución al campo del IOT mediante el desarrollo de sensores inteligentes basados en espectrofotometría de longitud de onda variable. Aplicación a la monitorización en continuo de la carga contaminante en aguas residuales urbanas *Tesis Doctoral*. Universidad Politécnica de Cartagena.

Carreres-Prieto, D., García, J.T., Cerdán-Cartagena, F., Suardiaz-Muro, J., Lardín, C. 2022. Implementing Early Warning Systems in WWTP. An investigation with cost-effective LED-VIS spectroscopy-based genetic algorithms. *Chemosphere*, *293*, 133610. https://doi.org/10.1016/j.chemosphere.2022.133610

Leardi, R., Boggia, R., Terrile, M. 1992. Genetic algorithms as a strategy for feature selection. *Journal of chemometrics, 6*(5), 267-281. https://doi.org/10.1002/cem.1180060506

Lepot, M., Torres, A., Hofer, T., Caradot, N., Gruber, G., Aubin, J.B., Bertrand-Krajewski, J.L. 2016 Calibration of UV/Vis spectrophotometers: a review and comparison of different methods to estimate TSS and total and dissolved COD concentrations in sewers, WWTPs and rivers. *Water Research, 101*, 519-534. https://doi.org/10.1016/j.watres.2016.05.070

Mesquita, D.P., Quintelas, C., Amaral, A.L., Ferreira, E.C. 2017. Monitoring biological wastewater treatment processes: recent advances in spectroscopy applications. *Reviews in Environmental Science and Bio/Technology*, 16(3), 395-424. https://doi.org/10.1007/s11157-017-9439-9

Niazi, A., Leardi, R. 2012. Genetic algorithms in chemometrics. *Journal of Chemometrics, 26*(6), 345-351. https://doi.org/10.1002/cem.2426

Otto, M. 2016. *Chemometrics: statistics and computer application in analytical chemistry*. John Wiley & Sons. https://doi.org/10.1002/9783527699377

Quade, M., Abel, M., Nathan Kutz, J., Brunton, S.L. 2018. Sparse identification of nonlinear dynamics for rapid model recovery. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(6), 063116. https://doi.org/10.1063/1.5027470

Searson, D.P., Leahy, D.E., Willis, M.J. 2010. GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. *Proceedings of the International multiconference of engineers and computer scientists*, *1*, 77-80. Citeseer.

TuringBot, S. 2020. *Symbolic Regression Software*. URL: https://turingbotsoftware.com.

Udrescu, S.M., Tegmark, M. 2020. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances, 6*(16), eaay2631. https://doi.org/10.1126/sciadv.aay2631

Wagner, S., Kronberger, G., Beham, A., Kommenda, M., Scheibenpflug, A., Pitzer, E., Affenzeller, M. 2014. Architecture and design of the HeuristicLab optimization environment. *Advanced methods and applications in computational intelligence*, 197-261. Springer, Heidelberg. https://doi.org/10.1007/978-3-319-01436-4_10

Zelinka, I., Oplatkova, Z., Nolle, L. (2005). Analytic programming–Symbolic regression by means of arbitrary evolutionary algorithms. *International Journal of Simulation: Systems, Science and Technology*, *6*(9), 44-56.

APÉNDICE

Flujo de trabajo para la generación de modelos de RLM y PG

En la Figura A1, se muestra el diagrama de proceso seguido para el preprocesamiento de los datos y la generación de los modelos de regresión lineal multivariable y regresión simbólica.

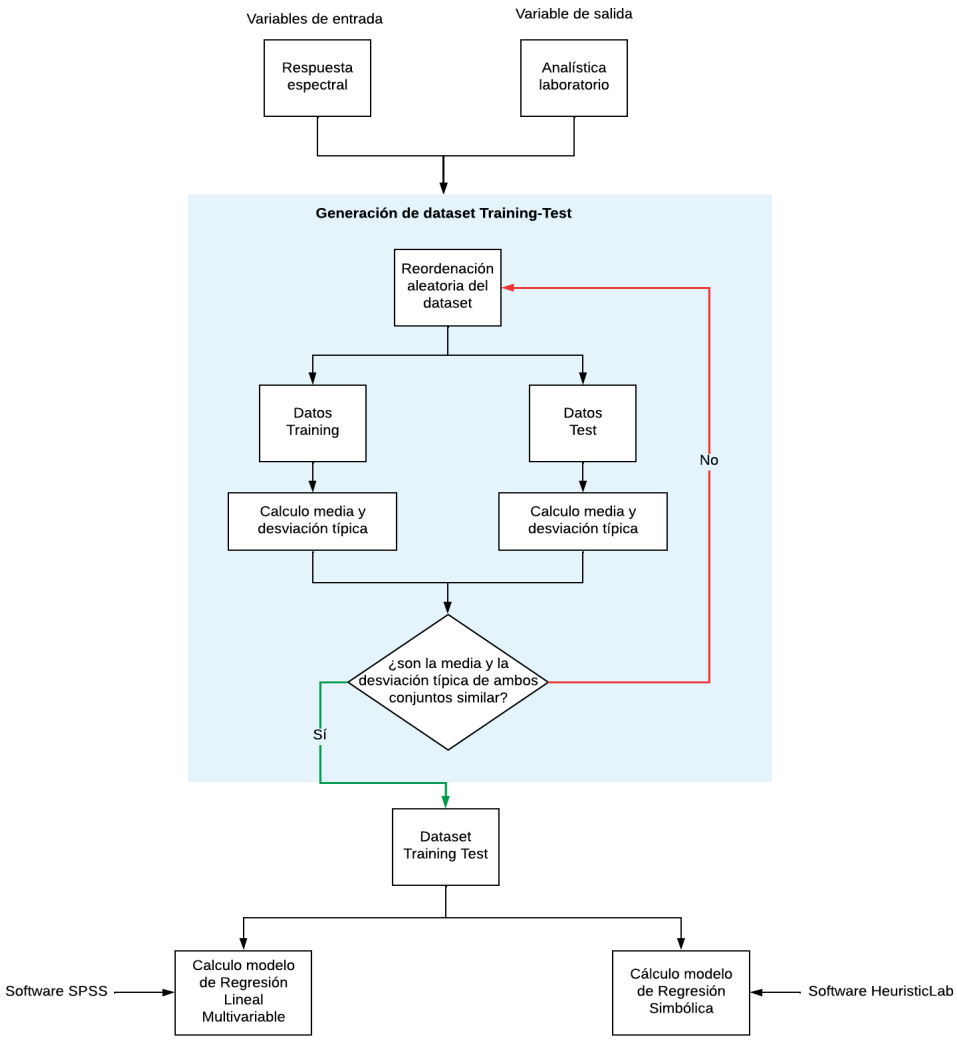


Figura A1 | Flujograma de proceso de preprocesamiento e implementación de modelos de regresión lineal multivariable y de regresión simbólica.

Para la generación de modelos de correlación, que permitan estimar una variable respuesta a partir de una serie de variables de entrada, partimos de dos conjuntos de datos: La respuesta espectrofotométrica de las muestras analizadas por el equipo de espectrofotometría desarrollado por los autores, que representa a las variables de entrada de los modelos, y por otro lado, los resultados analíticos de laboratorio, que es la variable objetivo.

Como paso previo a la generación de modelos de correlación, los datos deben segregarse en dos conjunto diferenciados: Training y Test, con una proporción del 80-20%. Para que los modelos sean más robustos y logren mejores estimaciones, debemos garantizar que exista heterogeneidad en cada dataset y que además, los datos de training y test sean lo más homogéneos posible entre sí, con el fin de lograr que los niveles de ajuste sean similares en ambos.

Para ello, mediante un script en Python y con ayuda de la librería Numpy, que implementa una gran variedad de funciones de procesado de datos, se ordenan de forma aleatoria y posteriormente se dividen en dos conjunto manteniendo la proporción 80-20%. Tras ello, se calcula la media y la desviación típica de cada conjunto y se comprueba si está es similar, asumiendo un margen de error del ±10%.

Si ambos conjuntos presentan discrepancias significativas entre sí, se vuelve a reordenar los datos de forma aleatoria y se repite el proceso hasta lograr que ambos sean homogéneos entre sí. Una vez generado el dataset de training y test, se procede a la generación de los modelos de regresión lineal multivariable y regresión simbólica.

Regresión lineal multivariable

Una vez importados los datos de training en SPSS, se procede a generar los modelos de regresión lineal multivariable mediante el método de cálculo de "por pasos", tal y como se muestra en la Figura A2, especificando los regresores del modelo, así como la variable de salida a predecir.

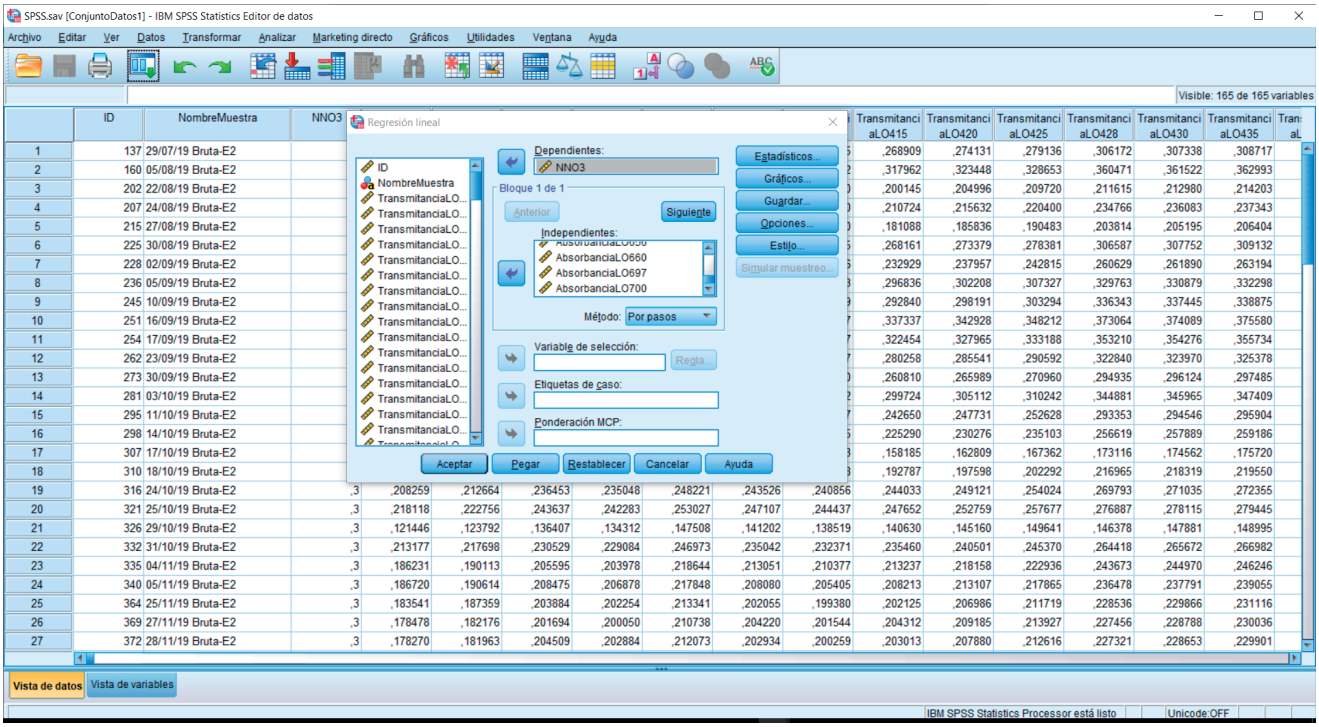


Figura A2 | Captura del proceso de configuración de los modelos de regresión lineal multivariable en SPSS.

Regresión simbólica

HeuristicLab proporciona una interfaz simple para configurar el proceso de aprendizaje del modelo. Existen una multitud de parámetros que pueden ser controlados para lograr mejores ajustes. Sin embargo, los principales parámetros a definir son:

- Cuál es la variable respuesta
- Qué regresores emplear
- La extensión del modelo de regresión simbólica resultante
- Número máximo de generaciones
- Población de cada generación (número de individuos que se crean en cada generación)
- Tasa de mutación o probabilidad de que un individuo en una generación sufra una mutación.

Tras el proceso de aprendizaje, el programa devuelve el modelo calculado, proporcionando una comparativa entre los valores de referencia (laboratorio) y aquellos estimados mediante el modelo calculado, tal y como se muestra en la Figura A3.

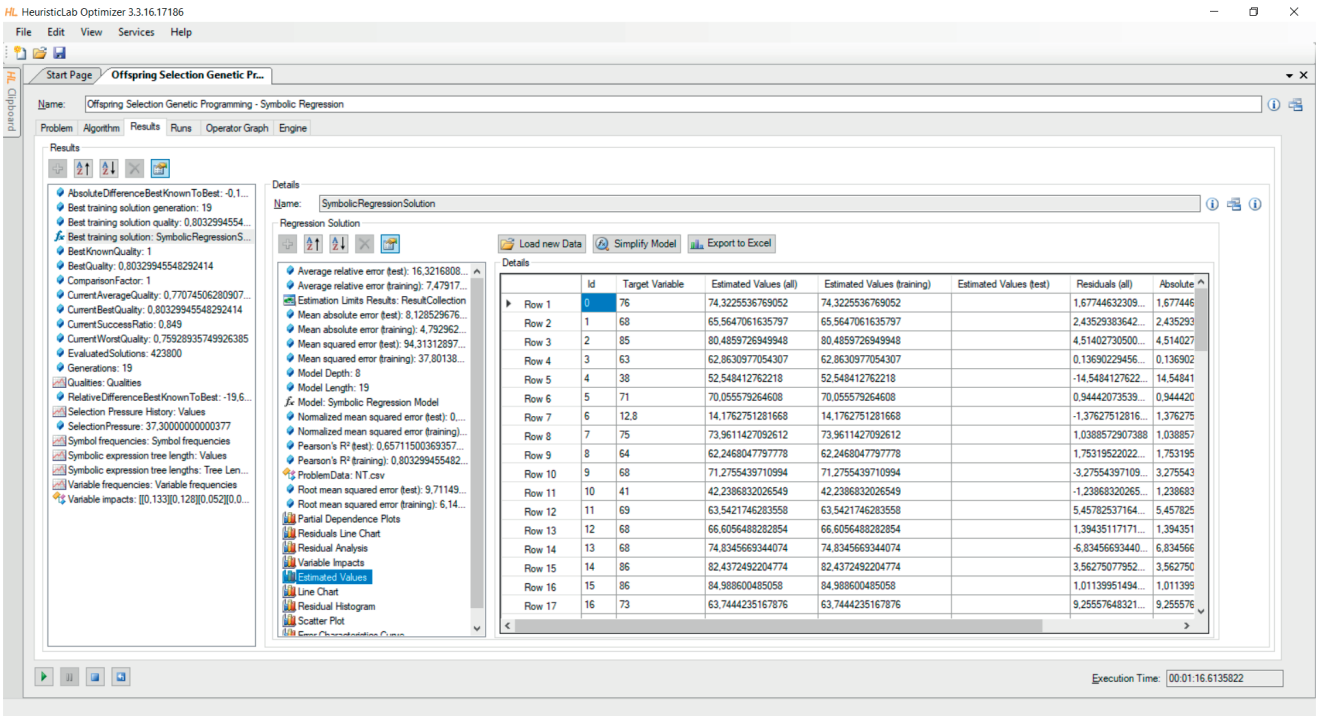


Figura A3 | Captura del software HeuristicLab tras el proceso de entrenamiento, donde se muestra una comparativa entre los valores de referencia y los estimados.

Ejemplo de implementación de Programación Genética mediante la librería gpLearn

En esta sección, se describe el proceso básico de manejo de la librería gpLearn, para la generación de modelos de regresión simbólica mediante Python.

Para el correcto desempeño de la librería, se precisa del uso de otras para llevar a cabo tareas auxiliares, como el preprocesamiento de los datos, entre las que destaca Scikit-learn (Sklearn). Esta se trata de una de las librerías más importantes dentro del campo de la inteligencia artificial, necesaria para poder implementar el script para llevar a cabo la regresión simbólica.

Esta librería contiene diferentes funciones de análisis, incluyendo la formación de *clusters*, análisis predictivos de datos, normalización de datos, transformaciones no lineales, entre otras. La ventaja de la programación genética radica en que es altamente configurable. No sólo es posible obtener modelos de estimación que se ajusten de forma adecuada a la variable respuesta, sino que también es posible configurar cómo se va a llevar a cabo el entrenamiento, así como el número de generaciones, probabilidad de que se produzca una mutación entre los individuos, número de variables que intervienen en el modelo, etc. En Código 1, se muestra un ejemplo de configuración de diversas variables del para la confección del modelo de programación genética, así como la ejecución del entrenamiento.

```
function_set = ['add', 'sub', 'mul', 'div','neg','inv']
est_gp = SymbolicRegressor(
population_size=3000,
function_set=function_set,
generations=20,
const_range =(-2000,2000),
init_method = 'half and half',
p_subtree_mutation=0.15,
p_hoist_mutation=0.05,
p_point_mutation=0.1,
p_crossover=0.15,
max_samples=0.9,
feature_names=LO_Soportadas_str,

verbose=1)


#Ejecución del entrenamiento del modelo

est_gp.fit(Regresores_training, Referencia_training)
```

Código 1 | Configuración del modelo de programación genética y ejecución del entrenamiento.

En función de la aplicación a realizar, puede interesar que el modelo final haga uso de funciones matemáticas específicas (uso de condicionales, aritméticas, exponenciales, logarítmicas, etc.) que pueden ser seleccionadas mediante el comando "function_set" entre funciones por defecto ya implementadas en gpLearn. Igualmente, existe la posibilidad de introducir funciones nuevas definidas por el usuario.

Un modelo de programación genética, no es más que un modelo evolutivo, que aprende a extraer patrones con el paso de las generaciones, definidas mediante el comando "generations". En el presente trabajo se ha considerado un número de 20. En cada generación se crea un cierto número de individuos, definidos por el comando "population_size". El tamaño de esta población intergeneracional depende del número de datos de entrenamiento. Si el número no es demasiado elevado, como en el caso que nos ocupa, se debe disponer de una elevada cantidad de individuos con el fin de tener suficiente riqueza genética para que sea más sencillo lograr un mejor ajuste. En este estudio se han considerado 3000 individuos.

La primera generación es sin duda la más importante, dado que sobre la misma se construirán todas las demás. Por ello, gpLearn incluye diferentes métodos de inicialización ("init_method"): "Full", "Grow " y "Half and Half", que permiten controlar cómo de grande serán los individuos, es decir, cuantos elementos compondrán su genoma. La elección depende de la naturaleza de los datos de partida.

La tasa de mutación, así como la forma en la que dicha mutación se integrará en el genoma de cada individuo, se define mediante los comandos: "p_subtree_mutation", "p_hoist_mutation", "p_point_mutation" y "p_crossover". En este estudio se han considerado tasas de mutación del 15%.

Mediante esta implementación de código, se puede entrenar modelos de programación genética de una forma simple y rápida, así como introducir modificaciones en el proceso de aprendizaje que permita lograr mejores resultados.