MORFOMETRÍA DE LA RED DE CANALES DEL DELTA DEL RÍO PARANÁ

1er. Nicolás E. Ortiz, 2do. Martín Sabarots Gerbec, 3er. Santiago Guizzardi y 4to Rafael Grimson

[1,2 y 3] Subgerencia Laboratorio de Hidráulica, Instituto Nacional del Agua, Argentina
[4] Universidad Nacional de San Martín
nortiz@ina.gob.ar

RESUMEN:

En el siguiente trabajo se presenta una metodología basada en datos para predecir la profundidad media de cursos de agua en el Delta del río Paraná. Para ello se trabajó con cinco fuentes de datos topobatimétricos, con los cuales se calcularon una serie de atributos de interés. Este conjunto de datos se entrenó con un algoritmo de aprendizaje automático denominado Random Forest Regression, el cual permite estimar la profundidad media de cursos de agua a partir de variables explicativas fácilmente obtenidas de diversas fuentes de información. El desarrollo de esta metodología permite avanzar en la caracterización hidráulica de cursos de agua no relevados y que constituyen insumos de importancia para el desarrollo de un modelo hidrodinámico de la zona.

ABSTRACT:

The following paper presents a data-driven methodology to predict the mean depth of watercourses in the Paraná River Delta. For this purpose, five topo-bathymetric data sources were used to calculate a series of attributes of interest. This data set was trained with a machine learning algorithm called Random Forest Regression, which allows estimating the mean depth of watercourses from explanatory variables easily obtained from various sources of information. The development of this methodology allows us to advance in the hydraulic characterization of watercourses that have not been surveyed and that constitute important inputs for the development of a hydrodynamic model of the area.

PALABRAS CLAVES: Delta del río Paraná, Random Forest Regression, Aprendizaje Automático.

INTRODUCCIÓN

El Delta del río Paraná comprende un área de aproximadamente 1.500.000 ha desde su nacimiento en la localidad de Diamante en Entre Ríos, hasta su desembocadura en el estuario del Río de la Plata (Figura 1), presentando una relevante importancia comercial, productiva y ambiental.

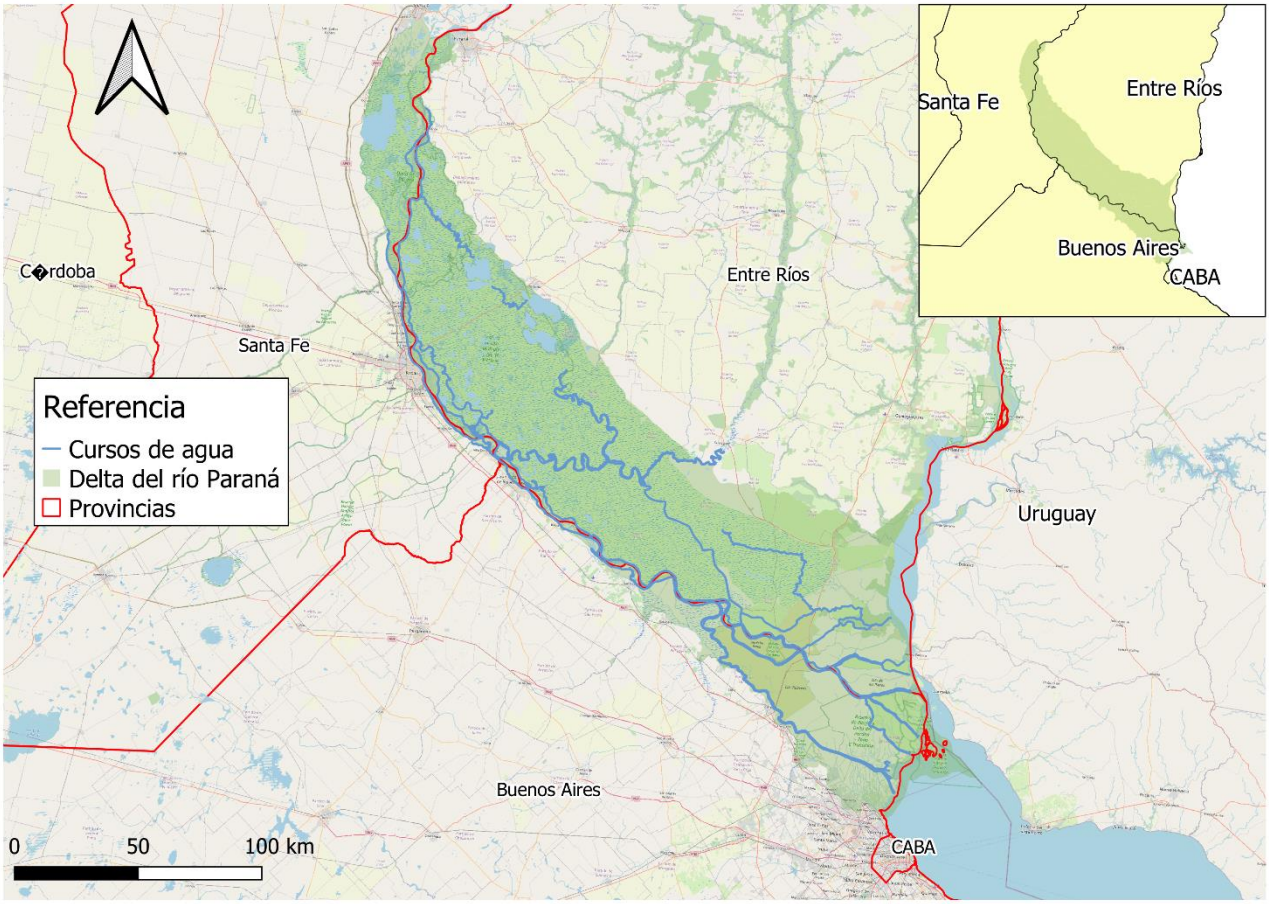


Figura 1.- Ubicación geográfica del Delta del Paraná.

Para entender el funcionamiento de este sistema complejo es necesario recurrir a la modelación numérica y al análisis de datos para dar respuesta a diferentes problemáticas que se presentan. En los últimos años se han realizado diversas campañas de medición (aforos líquidos, batimétricos, topográficos, entre otros) para la recopilación de datos (Morale et *al.*, 2018) que sirven de insumo para la puesta a punto de los modelos hidrodinámicos en desarrollo (Sabarots Gerbec, 2014; Guizzardi et *al.*, 2022).

Dada la compleja red hidrológica y geomorfológica presente, se configura una red de canales heterogénea y dinámica en el tiempo que es necesario abordar de una manera eficiente. Una particularidad importante es que no se dispone de mediciones hidráulicas detalladas para la mayoría de los cursos de agua de la región del Delta del río Paraná, en particular para las ramas secundarias, lo cual no es una tarea sencilla de resolver dada las limitaciones económicas, físicas y de infraestructura para relevar todos los cursos de agua de interés, contando su principal vía fluvial con una extensión de unos 448 km de extensión.

En el presente trabajo se describe una metodología para la determinación de la profundidad media de los cursos de agua del Delta del Paraná con un enfoque basado en datos.

METODOLOGÍA

Se partió de la hipótesis que las características morfométricas de los cursos de aguas principales y secundarios se ven influenciados principalmente por su historia geomorfológica y su régimen hidrológico.

Para ello se propone un abordaje de toma de decisiones basada en datos a partir del uso de fuentes provenientes de diferentes campañas realizadas a lo largo de los últimos años en la zona.

Se trabajó con una técnica de aprendizaje automático denominada Random Forest Regression, el cual es un algoritmo rápido, robusto y de fácil uso, que utiliza el método Bagging obteniendo resultados de diferentes predictores (árboles de decisión) los cuales son promediados. El objetivo es reducir la varianza del modelo y lograr un grado de generalización óptimo (James et *al.*, 2013). Se eligió este algoritmo a modo experimental, dada la cantidad de datos presentes y las cualidades previamente mencionadas.

Se trabajó con un conjunto de 1295 datos de secciones transversales (Figura 2) que se recopilaron a partir de cinco fuentes de datos diferentes.

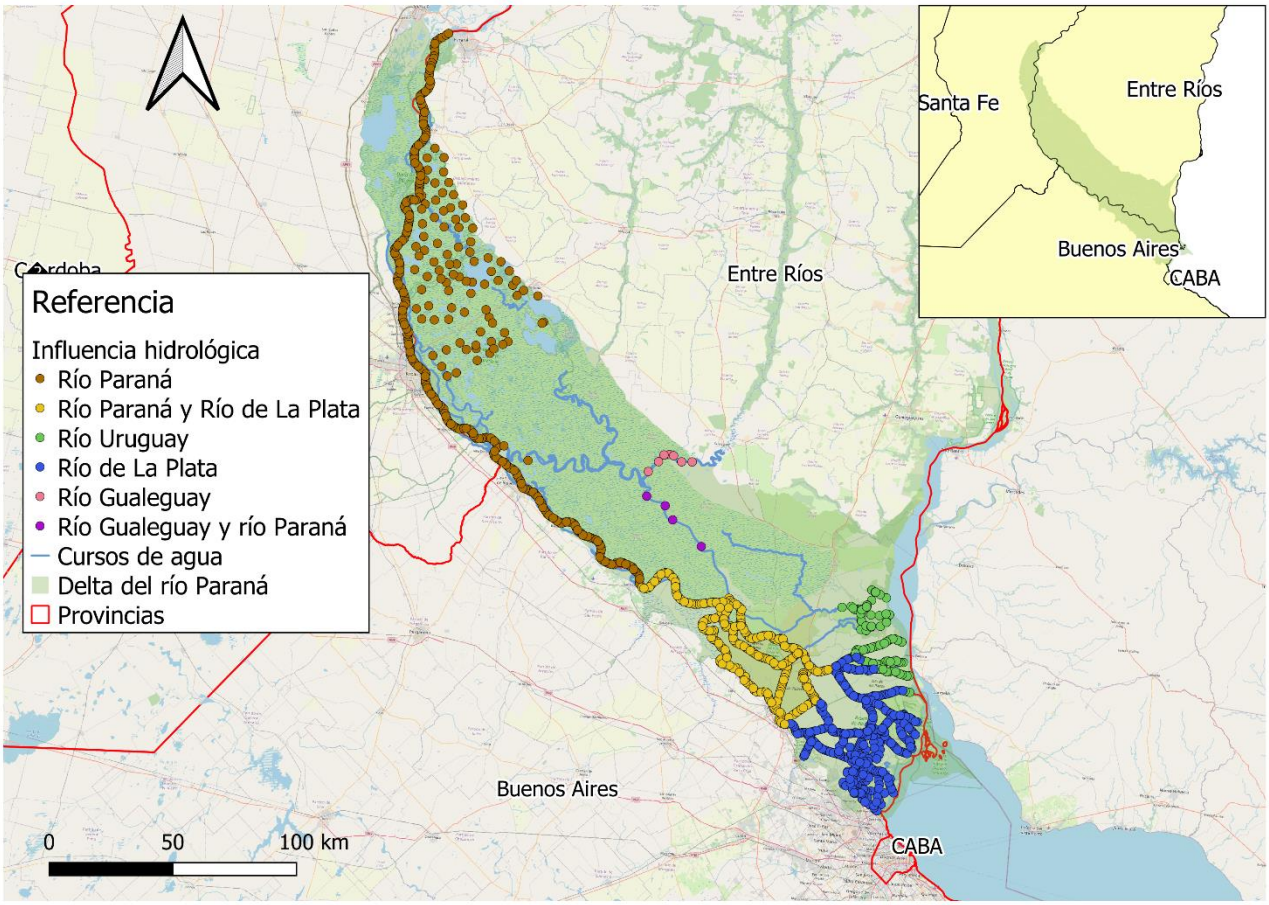


Figura 2.- Clasificación de secciones según influencia hidrológica.

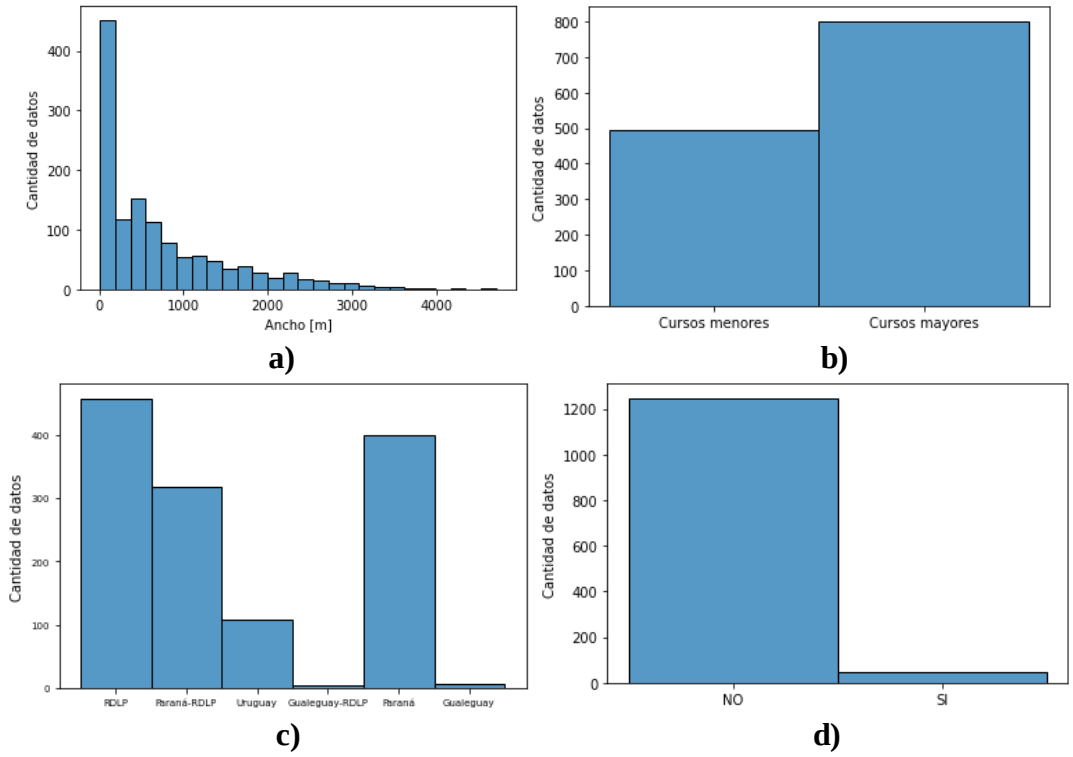
Para el entrenamiento del modelo, en primera instancia se dividió el conjunto de datos en un set de entrenamiento (80%) y testeo (20%) y posteriormente se normalizaron los datos.
Los hiperparámetros que se optimizaron fueron: *n estimators* (se seleccionaron 100), *max depth* (se optimizó con 22 en un rango de 1 a 25), *min samples leaf* (se optimizó con 2, con valores de 2, 5, 10 y 20) *y min samples split* (se optimizó con 2, con valores de 2, 3, 4 y 5). La métrica utilizada para evaluar la performance del modelo fue RMSE (Root Mean Squared Error).

ANÁLISIS EXPLORATORIO DE DATOS

Los datos con los cuales se trabajó fueron recopilados de diferentes campañas de medición realizadas en diferentes épocas y por diferentes organismos, por lo cual presentan incertidumbres propias de cada técnica de medición.

Las variables explicativas se seleccionaron de acuerdo a la disponibilidad de información y dada la hipótesis que presentan una relación física con la profundidad media de los cursos de agua, la cual es la variable a explicar (Figura 3). A su vez, se consideró que sean datos que se puedan determinar de manera sencilla. Entre los atributos asociados a los datos, se encuentran:

- Latitud
- Longitud
- Ancho del curso de agua
- Principal influencia fluvial bajo la que se encuentra el curso
- Unidades de paisaje
- Unidades geomorfológicas
- Cursos mayores o menores
- Cursos naturales o artificiales



a) **b)**

c) **d)**

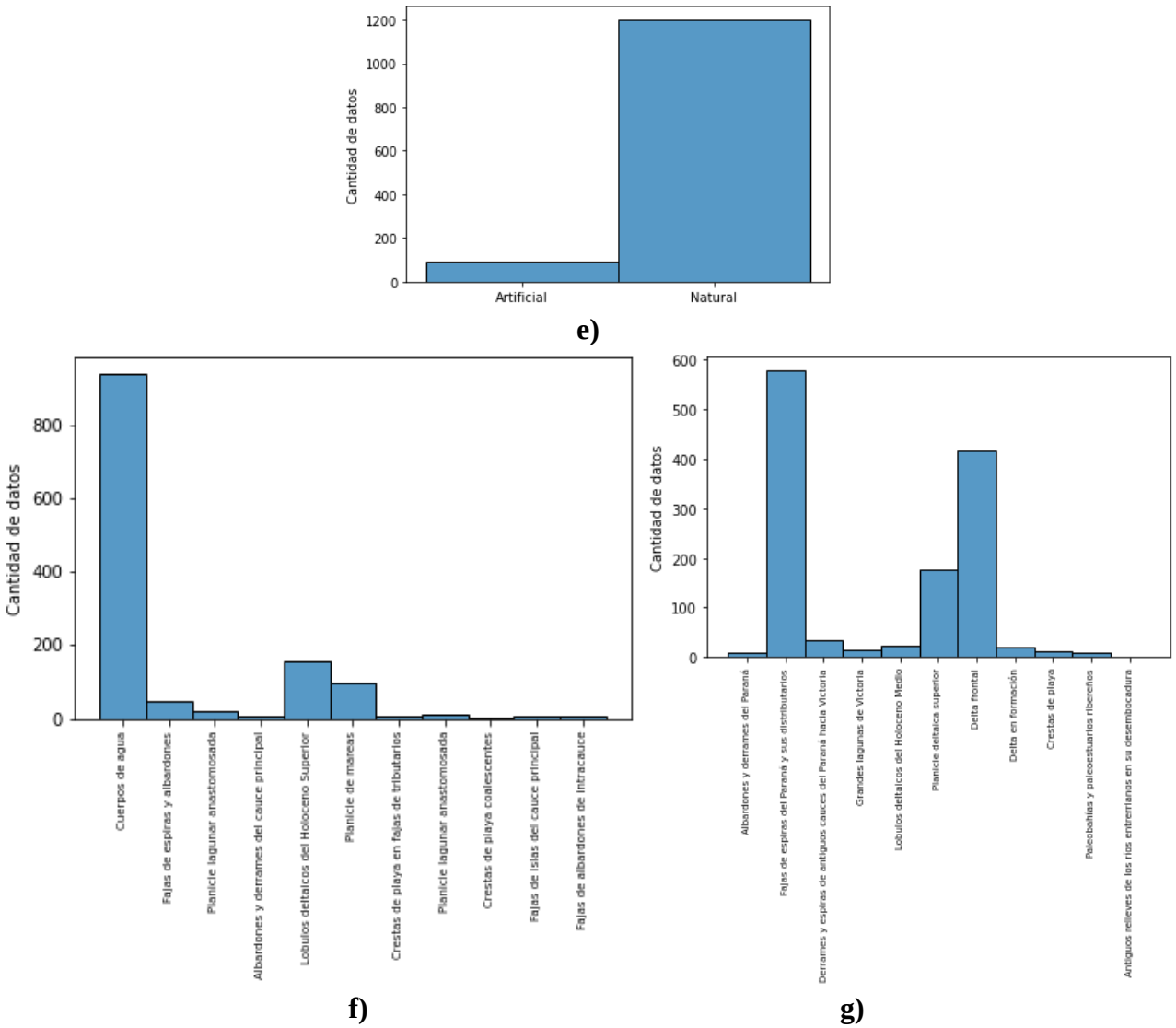


Figura 3.- Histogramas de variables explicativas. a).- Ancho de sección, b).- Dimensión, c).- Influencia hidrológica, d).- Influencia fluvial local, e).- Tipo, f).- Unidades de paisaje y g).- Unidades geomorfológicas.

RANDOM FOREST REGRESSION

Random Forest es un ensamble de árboles de decisión que dependen de una colección de variables aleatorias y que utilizan funciones de embolsado (Bagging) y aleatoriedad para crear un bosque aleatorio con árboles de decisión no correlacionados.

Cada conjunto de embolsado se compone de un árbol de entrenamiento de una muestra de datos extraída del conjunto con reemplazo, lo que significa que los datos se pueden utilizar más de una vez, llamada muestra de arranque. De esa muestra de entrenamiento, un tercio se reserva como datos de prueba, conocidos como muestra fuera de la bolsa (out of bag). Luego, se introduce otra instancia de aleatoriedad a través del empaquetamiento de atributos, lo que agrega más diversidad al conjunto de datos y reduce la correlación entre los árboles de decisión.

Estos modelos se entrenan de manera independiente, en el caso de un problema de regresión se promediarán las predicciones de los árboles de decisión individuales (Figura 4). Esta metodología es ampliamente utilizada para reducir la varianza en conjuntos de datos ruidosos.

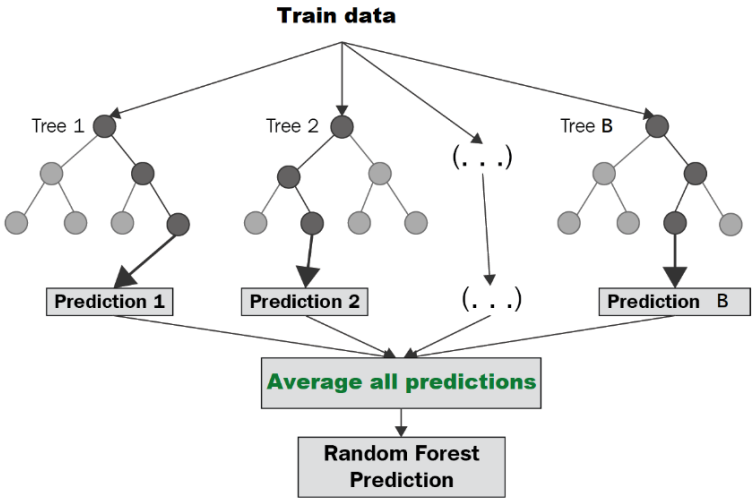


Figura 4.- Estructura de Random Forest.

Una característica importante del proceso de aplicación de algoritmos de aprendizaje automático es la selección y ajuste de los hiperparámetros del modelo. Este se realiza sobre el conjunto de datos de entrenamiento, en la cual se busca minimizar o maximizar una métrica de interés. La elección de los hiperparámetros y su rango de evaluación es arbitraria, pero es posible realizar múltiples pruebas para la elección de los más relevantes. Entre los hiperparámetros más relevantes en Random Forest encontramos: *n_estimators* (número de árboles que tendrá Random Forest); *max depth* (profundidad máxima del árbol de decisión); *min simples leaf* (número mínimo de muestras que debe haber en un nodo final u hoja) y *min samples split* (número mínimo de muestras necesarias antes de dividir el nodo).

Con este algoritmo se consigue mejorar la capacidad predictiva en comparación a los modelos basados en un único árbol de decisión, pero la interpretabilidad del modelo se reduce. Al tratarse de una combinación de múltiples árboles, no es posible obtener una representación gráfica sencilla del modelo y no es inmediato identificar de forma visual que predictores son más importantes.

RESULTADOS

Se presentan a continuación los resultados obtenidos de las predicciones de profundidad media realizadas en secciones pertenecientes a datos del conjunto de testeo (Figura 5).

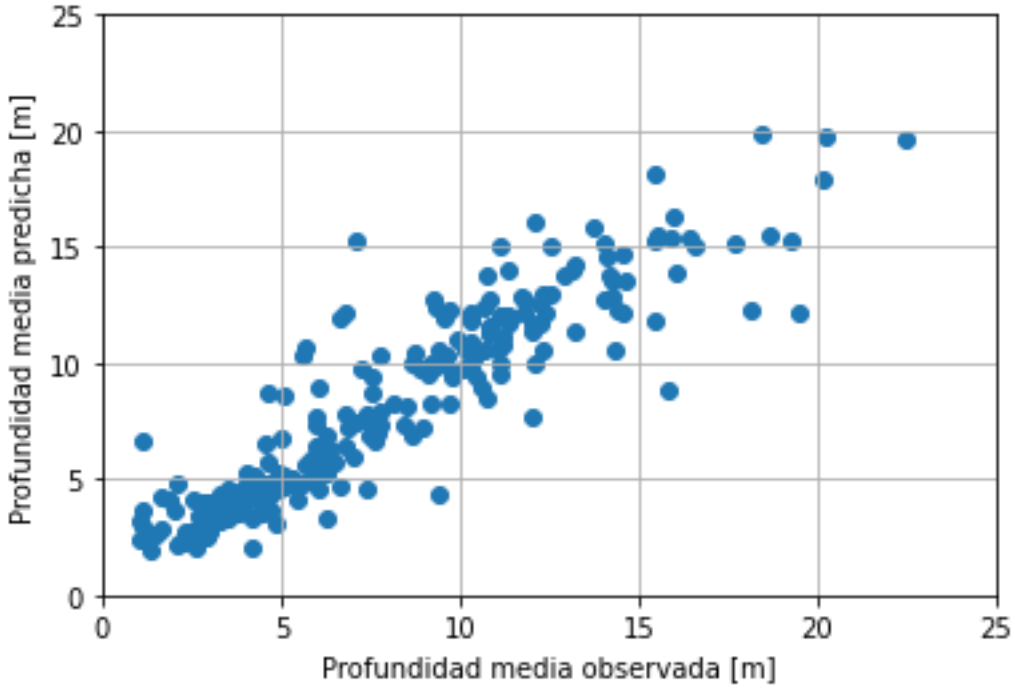


Figura 5.- Profundidad media observada versus profundidad media predicha.

Se obtuvo un RMSE de 1.79 m en el conjunto de datos de testeo. El atributo más explicativo de la profundidad media de los cursos de agua resultó ser el ancho de las secciones (Figuras 6 y 7).

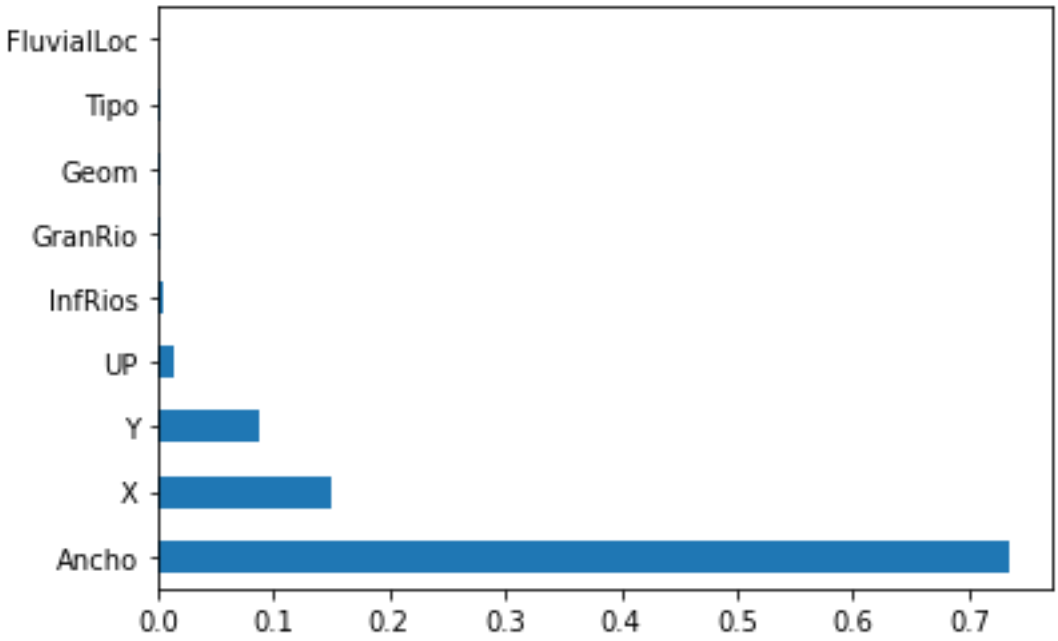


Figura 6.- Importancia de variables en la predicción.

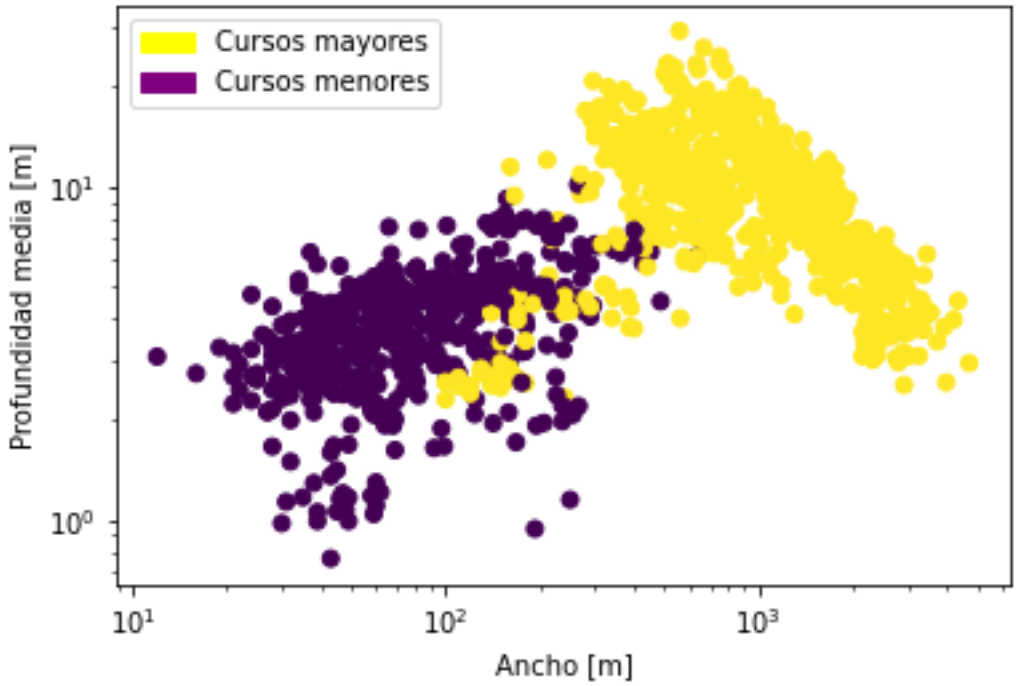


Figura 7.- Clasificación de cursos de agua en mayores y menores.

A partir de ello, se evaluó el error cuadrático medio en los cursos de agua que presentan un ancho mayor a 100 m y menor a 100 m, obteniéndose un RMSE de 1.94 m y 1.12 m respectivamente.

Por último, se evaluó el RMSE de cursos mayores y menores por zonas, las cuales fueron definidas por su influencia hidrológica principal (Tabla 1).

Tabla 1.- RMSE según influencia hidrológica en el conjunto de testeo.

Influencia hidrológica	Curso mayor	RMSE [m]
río Paraná	SI	1.37
río Paraná	NO	1.30
río Uruguay	SI	2.89
río Uruguay	NO	0.70
Río de La Plata y río Paraná	SI	2.54
Río de La Plata y río Paraná	NO	1.88
Río de La Plata	SI	1.63
Río de La Plata	NO	1.06
río Gueleguay	SI	2.35
río Gueleguay	NO	SIN DATOS
río Gueleguay y río Paraná	SI	SIN DATOS
río Gueleguay y río Paraná	NO	SIN DATOS

CONCLUSIONES

La aplicación de esta técnica permite el desarrollo de una línea de base para el establecimiento de una metodología que permita predecir una variable hidráulica de importancia y a su vez estandarizar los procedimientos de medición en campañas planteando un objetivo específico.

A partir del análisis de datos por clases se observa que las mayores diferencias se presentan en grandes ríos y en zonas en los cuales hay menor disponibilidad de datos.
Esta metodología presenta la desventaja de estar condicionada fuertemente a la calidad de los datos presentes, los cuales fueron conseguidos de fuentes diversas y presentan clases de datos desbalanceadas. Sin embargo, el modelo presenta la cualidad de ir aprendiendo a medida que se incorporan nuevos datos y puede ir mejorando su performance a lo largo del tiempo, lo cual permitiría ir ajustando las diferencias presentes.
El desarrollo de esta información permite avanzar en la caracterización de cursos de agua no medidos y en los cuales la posibilidad de avanzar en el conocimiento a partir de la medición in situ es poco viable. En particular estos datos son necesarios generarlos para la construcción de un MDE topobatimétrico que incluya gran parte de los cursos de agua del Delta del río Paraná.

REFERENCIAS

James, G., Witten, D., Hastie, T., & R., Tibshirani (2013). An introduction to statistical learning with applications in R. *Springer Science and Business Media, ISBN: 978-1-4614-7137-7.*
Guizzardi, S., Bianchi, J., Cortese, J.E., Uriburu Quirno, M., & Sabarots Gerbec, M. (2022). Forecast System Implementation in the Paraná Delta. *Proceedings of the 39th IAHR World Congress, Granada, Spain.*
Morale, M., Sabarots Gerbec, M. Re, M., Ortiz, N., & Bernal, J. (2018). Delta del Paraná: del territorio hacia la modelación hidrodinámica. *IFRH 2018.*
Sabarots Gerbec, M. (2014). Estudio de la dinámica superficial de la red de canales del Delta Medio del río Paraná. *II Encuentro de Investigadores en Formación en Recursos Hídricos, Ezeiza, Argentina*.