# Aplicación de ciencia de datos para la reconstrucción de series de tiempo de variables meteorológicas en Islas del Rosario (Caribe colombiano) entre los años 2013 – 2021

## *Application of Data Science for the reconstruction of time series of meteorological variables in the Islas del Rosario (Colombian Caribbean), between the years 2013-2021*

**DOI:** 10.26640/22159045.2022.604 **Fecha de recepción:** 2022-09-20 / Fecha de aceptación: 2022-10-19

**Camilo Contreras Vargas[1], Julián Quintero-Ibáñez[2], Ángela Solanilla[3]**


## Citar como:

**Contreras Vargas, C.; Quintero-Ibáñez, J.; Solanilla, A. (2022).** Aplicación de ciencia de datos para la reconstrucción de series de tiempo de variables meteorológicas en Islas del Rosario (Caribe colombiano) entre los años 2013 – 2021. *Bol. Cient. CIOH*; 41(2): 67-80. ISSN impreso 0120-0542 e ISSN en línea 2215-9045. DOI: https://doi.org/10.26640/22159045.2022.604



## Resumen

Este estudio revisa dos series de tiempo de variables meteorológicas medidas por una estación automática ubicada en Islas del Rosario (Caribe colombiano), perteneciente a la Red de Medición de Parámetros Oceanográficos y de Meteorología Marina (RedMpomm) de la Dirección General Marítima (Dimar). Las series de tiempo corresponden a datos de temperatura ambiente y magnitud del viento en el periodo 2013-2021, los cuales presentan algunos valores faltantes. El objetivo del estudio fue desarrollar un modelo que permitiera completar automáticamente los diferentes vacíos existentes en las series de tiempo, utilizando las ventajas de la ciencia de datos al completar información con valores estimados. La importancia de obtener series reconstruidas radica en lograr tener bases de datos más sólidas para ser utilizada en los trabajos de investigación y académicos que realiza la Dimar. La metodología desarrollada consistió en el uso de imputación de medianas a partir de datos existentes sobre fechas y horas asociadas a valores faltantes, todo esto mediante el uso de desfases de datos e información complementaria como relaciones de periodicidad sobre el conjunto de datos. Los resultados mostraron que se logró implementar una metodología capaz de estimar el valor más adecuado para completar las diferentes series temporales, la cual constituye una primera aproximación para la reconstrucción de datos meteorológicos.


**Palabras claves:** serie de tiempo, meteorología, valores faltantes, temperatura del aire, magnitud del viento.


## Abstract

*This study reviews two time series of meteorological variables measured by an automatic station located in Islas del Rosario (Colombian Caribbean), belonging to the Network for Measurement of Oceanographic Parameters and Marine Meteorology (RedMpomm) of the General Maritime Directorate (Dimar). The time series correspond to data of air temperature and wind magnitude in the period 2013-2021, which present some missing values. The objective of the study was to develop a model that would*



[1] Orcid: 0000-0003-0066-6294. DS4A Program Colombia. Correo electrónico: camilocontrerasvargas910@gmail.com

[2] Orcid: 0000-0003-3026-0391. Centro de Investigaciones Oceanográficas e Hidrográficas del Caribe. Correo electrónico: jquinteroi@dimar.mil.co

[3] Orcid: 0000-0003-1516-9100. University of California San Diego. Correo electrónico: asolanilla@ucsd.edu

*automatically reconstruct missing values in the time series, using the advantages of data science to complete information with estimated values. The importance of obtaining reconstructed series lies in having more solid databases to be used in the research and academic work carried out by Dimar. The methodology developed consisted of the use of imputation of medians from existing data on dates and times associated with missing values, all this through the use of data lags and complementary information such as periodicity relationships on the data set. The results showed that it was possible to implement a reliable methodology capable of estimating the most appropriate value to complete the different time series, which constitutes a first approximation for the reconstruction of meteorological data.*

***Keywords:*** *time series, meteorology, missing values, air temperature, wind magnitude.*

## Introducción

El Centro de Investigaciones Oceanográficas e Hidrográficas del Caribe (CIOH) de la Dirección General Marítima (Dimar), lleva a cabo un gran número de investigaciones en diversas áreas tales como la oceanografía, hidrografía, biología marina, protección ambiental y manejo de zonas costeras, entre otros. A través de la gestión de la información y la elaboración de informes, algunos grupos de investigación han realizado investigaciones académicas, llevado a cabo estudios y prestado asesorías en materia de políticas marinas durante más de 45 años (CIOH, 2022a). Adicionalmente, Dimar cuenta con la Red de Medición de Parámetros Oceanográficos y de Meteorología Marina (RedMpomm) que tiene como objetivo obtener, almacenar, estandarizar y poner a disposición, los datos oceanográficos y de meteorología marina obtenidos en las costas del Caribe y Pacífico colombiano y en áreas insulares, gestionar y brindar acceso a los recursos de datos oceanográficos nacionales, con el fin de mejorar la seguridad integral marítima, fluvial y portuaria, preservar la vida humana en el mar y tomar decisiones relacionadas con las operaciones de la autoridad mMarítima nacional (Dimar, 2022a).

La RedMpomm se conforma de estaciones meteorológicas y mareográficas, boyas direccionales de oleaje y boyas MetOcean, las cuales transmiten información en tiempo real vía celular y satelital. El CIOH utiliza los datos generados en esta red para llevar a cabo los proyectos institucionales y científicos. Sin embargo, existen vacíos en los datos generados en algunas de las estaciones, los cuales se atribuyeron a problemas técnicos de los sensores, cuya oportuna reparación es limitada debido al difícil acceso a los mismos, considerando su ubicación; los equipos o la infraestructura que se ven expuestos a problemas de transmisión o eventos extremos que terminan por afectar la continuidad de las mediciones.

El estudio del clima y sus variaciones implica el uso de series históricas de larga duración. Estos conjuntos de datos resultan esenciales, ya que constituyen la base para evaluar tendencias a grandes escalas temporales para la validación de modelos climáticos, así como para la detección del cambio climático a escala regional (Acquaotta y Fratianni, 2014). Complementar el conjunto de datos de las series de tiempo, aumentará la solidez de la información, transformándola en un insumo de alta calidad para llegar a un análisis más detallado y una mejora del proceso de toma de decisiones en una variedad de campos, como aquellos relacionados con la seguridad marítima, la navegación y el cambio climático. Actualmente, climatólogos y ambientalistas propenden por extraer información útil de una gran cantidad de registros de observación y datos de simulación para el sistema climático (Zhang, Zhang y Khelifi, 2018).

Las metas de cambio climático de Colombia incluyen estrategias de mitigación; sin embargo, están enfocadas en la adaptación y esto implica mejorar la información disponible sobre el capital natural del país, como un primer paso en el monitoreo de los cambios de ecosistemas estratégicos como manglares, humedales, arrecifes de coral, glaciares, océanos y bosques tropicales, reconociendo su valor intrínseco y los servicios ambientales que ofrecen a Colombia y al mundo (NDC de Colombia, 2020). De esta forma, todos los esfuerzos por mejorar las series de datos de las variables climáticas redundarán en mejores insumos para la implementación de las actividades de la autoridad marítima.


**Figura 1.** Mapa del Archipiélago de las Islas del Rosario, Mar Caribe.

Teniendo en cuenta la relevancia de la información consolidada en las series temporales y los vacíos que existen en ellas, el CIOH identificó este problema y planteó reconstruir series temporales de variables meteorológicas de una estación costera en el Caribe colombiano, como parte de un reto de ciencia de datos. Para llevar a cabo este reto y como un primer piloto, se seleccionó una estación meteorológica costera que hace parte de la RedMpomm (Islas del Rosario). Las variables seleccionadas fueron temperatura del aire y velocidad del viento; ambas series cubren el período comprendido entre 2013 y 2021, y presentan vacíos asociados a observaciones con valores faltantes.

Este estudio presenta un enfoque metodológico para la reconstrucción de estas series de tiempo, en el que se utilizan técnicas de análisis de datos y modelado estadístico para la complementación de las mismas, de manera que puedan ser utilizadas posteriormente por los investigadores de la autoridad marítima nacional. El enfoque utilizado constituye un primer ejercicio para la implementación de futuros modelos que permitan reconstruir series de tiempo, en las cuales existan vacíos de información, no solo en la estación de Islas del Rosario, sino en diferentes estaciones meteorológicas.

## Área de estudio

El área de estudio se ubica a 52 kilómetros aproximadamente al Sur-Oeste de la bahía de Cartagena, entre 75º41’47’’W – 75º48’15’’W y 10º14’33’’N – N 10º41’45’’N (Figura 1). El área incluye 28 islas y cayos sobre un conjunto de formaciones coralinas antiguas y sucesivas que actualmente se encuentran por lo menos a tres metros sobre el nivel medio del mar (Cendales, Zea y Díaz, 2002). Desde 1977, esta área ha sido protegida legalmente bajo el nombre de “Parque Nacional Natural Corales del Rosario y San Bernardo” (Parques Nacionales de Colombia, 2022).

La dinámica marina del archipiélago está influenciada significativamente por la intensidad y la estacionalidad de los vientos alisios, afectando la propagación del oleaje en las aguas poco profundas y el aumento del nivel del mar (Restrepo *et al*., 2012). Para la bahía de Cartagena (la ciudad principal más cercana al área de estudio), la época seca hace referencia al aumento de la velocidad del viento en el área y a la disminución de las precipitaciones locales. Comienza en diciembre y se extiende hasta marzo, registrándose vientos del noreste entre 2.6 y 5.1 m/s y máximas ocasionales de 15.43 m/s. Por otra parte, durante

la época de lluvias (entre agosto y noviembre), sobresalen condiciones climáticas caracterizadas por vientos de baja intensidad, entre 1.0 y 3.0 m/s, presentando esporádicamente máximos de 5.0 m/s (Rueda, Otero y Pierini, 2013).

En el Caribe también existe un evento sinóptico importante en cuestiones de meteorología, el "Veranillo de San Juan". Este es un período carente de lluvia que, por lo general, ocurre entre julio y agosto (Andrade y Barton, 2000). Puerres *et al*. (2018) demostraron que las tormentas generalmente ocurren durante la estación seca, asociadas a los frentes fríos, a las tormentas principales, y a otros eventos de vientos fuertes y olas extremas que dan forma a las crestas de las playas en las Islas del Rosario cada 70 años.

Respecto a la temperatura del aire y con base en un análisis de 59 años de información obtenida de datos de reanálisis (1950-2009), Gutiérrez *et al*. (2011) describieron que la temperatura del aire en el archipiélago oscila entre 26.2 °C y 27.36 °C. Por otra parte, los datos obtenidos de la bahía de Cartagena indican que la temperatura promedio presenta sus valores más altos entre mayo y junio, con promedios entre 28.3 ºC y 28.4 ºC. Así mismo, los valores mínimos de la temperatura media se presentan durante los meses de enero, febrero y marzo, oscilando entre 26.8 ºC y 27.1 ºC. La temperatura máxima presenta una media multianual de 31.5 ºC, con sus valores máximos en junio, julio y agosto, con medias entre 31.9 ºC y 32.0 ºC (CIOH, 2022b).

## Metodología

Se implementó una metodología de Proceso Estándar Intersectorial para Minería de Datos (CRISP-DM) desarrollada por Wirth y Hipp (2000). Este se describe como un modelo jerárquico (es decir, que va de lo general a lo específico), dividido en las seis fases del ciclo de vida de un proyecto. Esta sección describe las fases de entendimiento del fenómeno, preparación de datos, modelado, evaluación e implementación (Figura 2).

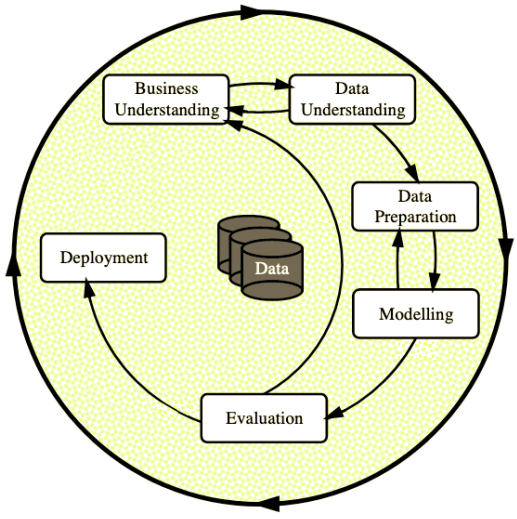


**Figura 2.** Fases del Modelo del Proceso CRISP-DM (Wirth y Hipp, 2000).

El desarrollo de la metodología se implementó en Python, un lenguaje de programación interpretado de alto nivel . La comunidad en torno a las bibliotecas disponibles de uso abierto, hace que Python sea particularmente atractivo para trabajo en ciencia de datos, aprendizaje automático y computación científica (Raschka, Patterson y Nolet, 2020). Adicionalmente, se diseñó un tablero interactivo para cargar los archivos con las series de tiempo que están incompletas. Esto se implementó mediante la utilización de Amazon Web Services EC2 como máquina virtual, un contenedor de Docker para ejecutar la aplicación, la cual se construyó a través del marco de trabajo de Python (Figura 3).

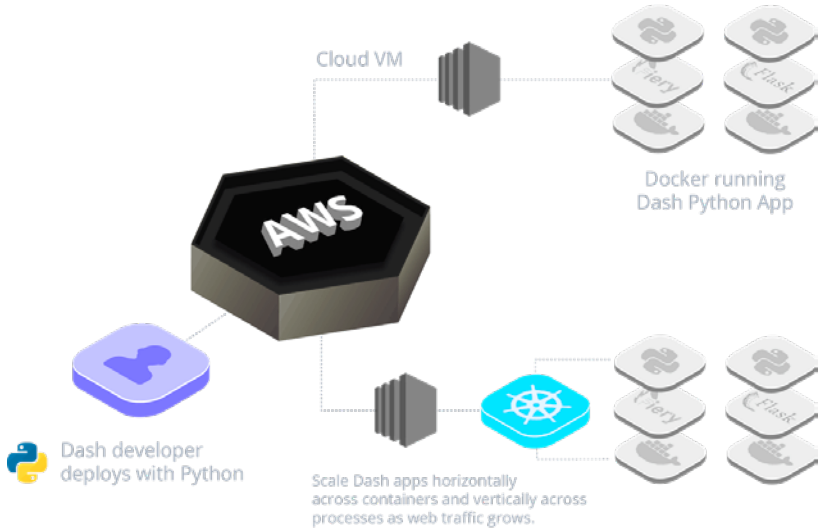


**Figura 3.** Proceso de implementación del tablero interactivo. Fuente: https://plotly.com/

## Entendimiento del fenómeno

### *Recolección de los datos*

Los datos fueron medidos por una estación meteorológica automática ubicada en Isla Naval del archipiélago Islas del Rosario, perteneciente a la RedMpomm (Figura 4). Las series de tiempo de magnitud del viento y temperatura del aire fueron suministradas por el Centro Colombiano de Datos Oceanográficos (Cecoldo) desde el 19 de agosto de 2013, al 31 de diciembre de 2021, en formato CSV, con una periodicidad de cada 10 minutos para la velocidad del viento (Dimar 2022b) y cada hora para la temperatura del aire (Dimar 2022c).

### *Exploración de los datos*

En este proceso se analizaron los valores atípicos de la serie en el contexto climático de las fechas en las que se registraron dichos valores.

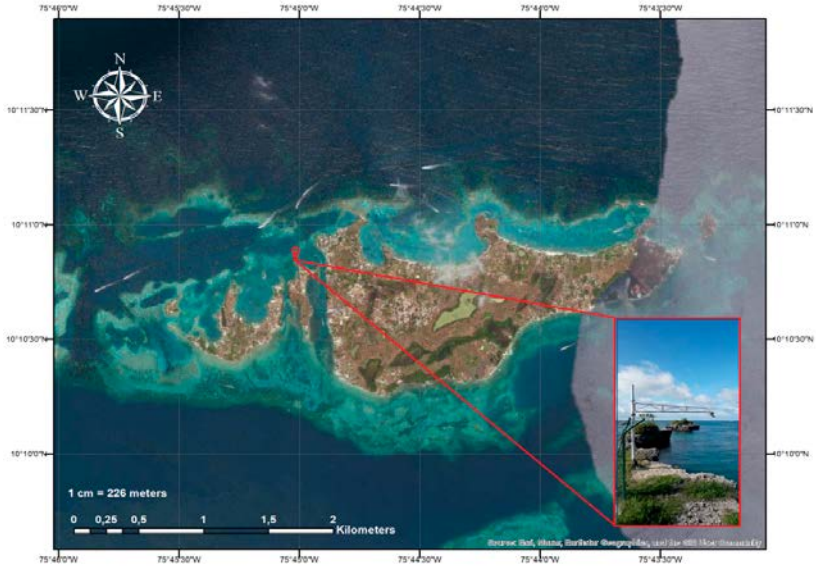


**Figura 4.** Ubicación de la estación meteorológica automática en Isla Naval (Islas del Rosario, Caribe colombiano) perteneciente a RedMpomm.

## Preparación de los datos

### *Limpieza de datos*

Según el estándar utilizado por Cecoldo, los datos no reportados por los sensores se consideran valores faltantes y se les asigna el valor: -99999. Con fines de procesamiento, estos valores fueron convertidos a nulos y constituyen datos a reconstruir. Adicionalmente, existen algunos valores que no corresponden con las condiciones ambientales del lugar en el que se midieron. De acuerdo con el contexto climatológico de las áreas costeras del Caribe colombiano (temperaturas mínimas, medias y máximas del aire), no se esperan valores fuera del rango entre 10 °C y 55 °C (Instituto de Hidrología, Meteorología y Estudios Ambientales [IDEAM], 2022). Aunque los sensores pueden registrar valores fuera del intervalo esperado para la zona debido a su amplio rango de medición, estos resultan incongruentes para el área de estudio; por lo tanto, esos valores fueron revisados y convertidos a nulos, lo que a su vez corresponden a vacíos adicionales. Este procedimiento también se realizó con la magnitud del viento, en donde el umbral fue de 75 m/s, límite inferior para un huracán de categoría cinco en la escala Saffir-Simpson (Taylor *et al*. 2010). Por último, se obtuvo el porcentaje total de vacíos para cada variable.

### *Descripción de los datos*

Se realizaron diagramas de cajas y bigotes para caracterizar las variables temperatura del aire y magnitud del viento. Estos gráficos permiten visualizar la distribución de las variables, el rango intercuartil, y los valores atípicos. Adicionalmente, se llevó a cabo una prueba de estacionariedad para ambas series de tiempo. Esto es importante para el desarrollo de modelos y métodos de imputación, para lo cual se implementó la prueba de Dickey-Fuller, cuya hipótesis nula afirma que existe una raíz unitaria en un modelo de serie de tiempo autorregresivo. La hipótesis alternativa es diferente según la versión de la prueba utilizada, pero suele ser de estacionariedad o con tendencia estacionaria (Hatanaka, 1996). Además, las variables de interés están agrupadas de acuerdo a la hora del día; es decir, mediante la obtención de la temperatura media del aire o la magnitud del viento para cada hora del día, con el fin de explorar el posible comportamiento cíclico de estas series.

### *Reconstrucción de los datos*

Se implementó un método de imputación basado en la mediana. El método de imputación de punto cercano usa valores cercanos para ordenar y después selecciona la mediana como reemplazo del valor faltante, con la ventaja de que el reemplazo pertenece a los datos (Majidpour *et al*., 2016). Incluso, la imputación que utiliza la mediana supera a las imputaciones basadas en algoritmos de los vecinos más cercanos en algunas aplicaciones de análisis de datos (Sessa y Syed, 2016). El enfoque utilizado se resume de la siguiente manera (Figura 5):

1. La variable de interés se agrupó por hora del día y día del año, para cada año de la serie de tiempo.
2. La mediana se obtuvo para cada día y hora específica de todos los años.
3. Por último, con dicha mediana se imputó la serie de tiempo en los años en los que habían vacíos.


**Figura 5.** Proceso implementado para la imputación basada en la mediana.

### *Modelado de datos*

Esta fase solo se implementó para la variable temperatura del aire. Se construyeron dos modelos para dicha serie; primero, se dividieron los datos totales de forma no aleatoria con los siguientes valores: 80 % - 20 % en entrenamiento y prueba, respectivamente (Gholamy, Kreinovich y Kosheleva, 2018). El primero de estos modelos fue un modelo autorregresivo (AR) con un parámetro de modelo de 27 rezagos, y el otro una red neuronal recurrente (RNN) con unidades GRU, ambos implementados en Python. Según la aplicación, estos presentan diferentes ventajas; aquellos basados en redes neuronales presentan una mayor complejidad en términos

de la búsqueda correcta de hiperparámetros y el cómputo necesario para entrenarlos, siendo los modelos clásicos más fáciles de usar (Almqvist, 2019).

## Evaluación del modelo

Para evaluar el ajuste de ambos modelos se utilizó la métrica del error cuadrático medio (RMSE). Comúnmente, se usa para medir el rendimiento del modelo en estudios de meteorología, calidad del aire e investigación climática (Chai y Draxler, 2014).

## Resultados y discusión

### *Exploración de datos*

La figura 6 muestra la serie de tiempo de la magnitud del viento para el período evaluado. A partir de la misma se observa una estacionalidad interanual, así como valores extremos, que se registraron durante febrero de 2019 y 2020 (del 27 de abril al 27 de septiembre). Estos valores probablemente sean errores de calibración o daños en el sensor, ya que no hay un registro de eventos que hayan ocurrido en la zona durante ese periodo, que hayan generado velocidades de viento sostenidas superiores a 70 m/s. A pesar de que el huracán Iota ocurrió en 2020, las fechas de los valores extremos (abril-septiembre), no concuerdan con las fechas de formación y paso del huracán por el Caribe colombiano, los cuáles ocurrieron en noviembre (Centro Nacional de Huracanes, 2022). Por lo tanto, estos valores se convirtieron en valores nulos adicionales.

### *Limpieza de datos*

En las Figuras 7 y 8 se muestra la serie de tiempo después del proceso de limpieza. El porcentaje total de valores nulos fue de 12.12 % para la magnitud del viento y 3.3 % para la temperatura del aire. En cuanto a la magnitud del viento, los valores más altos se registraron en la época seca y los más bajos, en la época lluviosa, lo cual coincide con lo descrito por otros autores para la zona de estudio (Rueda, Otero y Pierini, 2013). Este comportamiento contrasta con la temperatura del aire, variable para la cual se obtienen los valores más bajos al principio de año y los más altos al final del mismo.

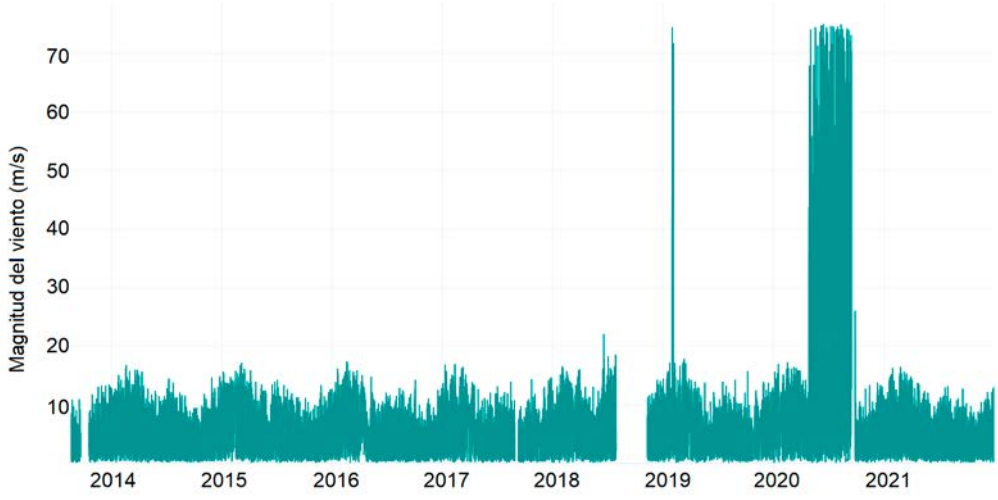


**Figura 6.** Serie de tiempo original para la magnitud del viento.

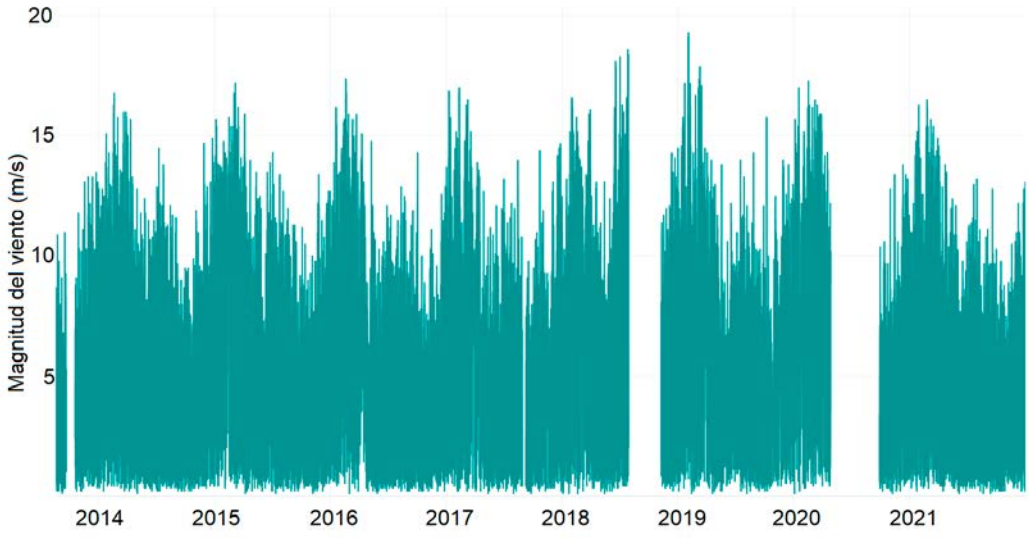


**Figura 7.** Serie de tiempo de la magnitud del viento (m/s) después del proceso de limpieza.

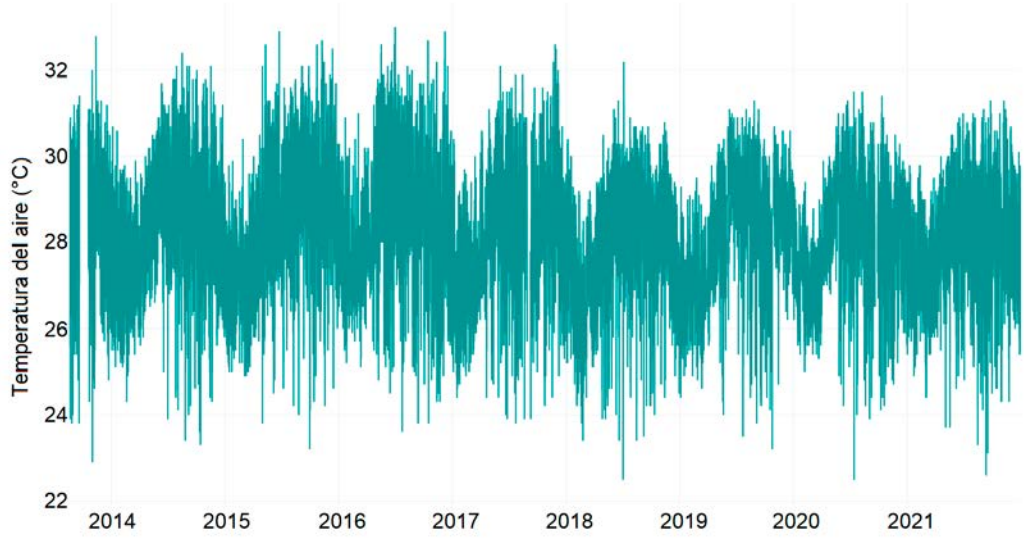


**Figura 8.** Serie de tiempo de la temperatura del aire (°C) después del proceso de limpieza.

### *Descripción de los datos*

Se identificaron los valores atípicos a partir de los diagramas de cajas y bigotes (puntos rojos en la Figura 9). La mediana fue de 4.7 m/s para la serie de la magnitud del viento, que así mismo presentó un máximo atípico de 19.3 m/s (Figura 9a). En cuanto a la serie de temperatura del aire, la mediana fue de 28.2 °C con un máximo atípico de 33 °C (Figura 9b).

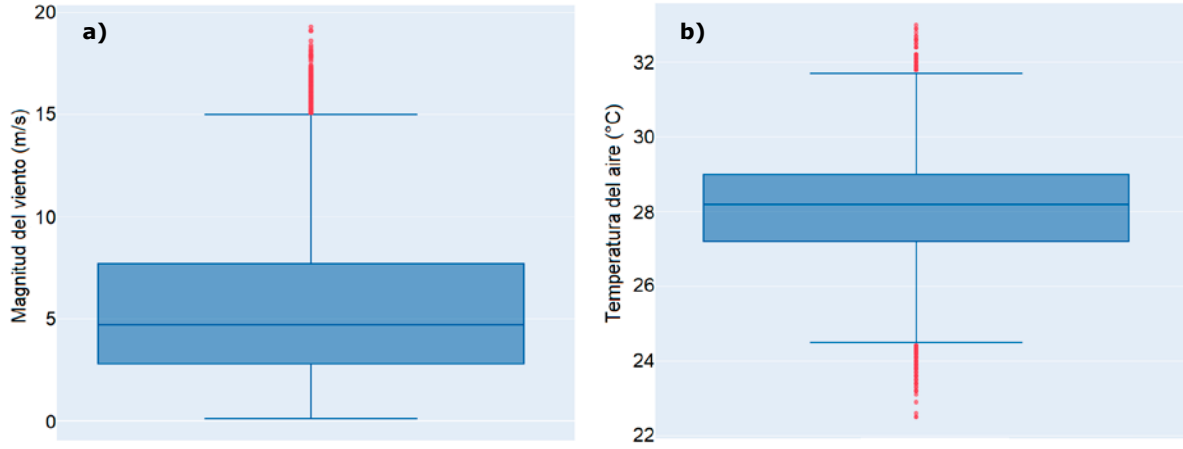


**Figura 9.** Diagrama de caja y bigotes para la magnitud del viento (a) y la temperatura del aire (b).

Estos gráficos permitieron identificar que, para la temperatura del aire, se pueden presentar valores atípicos, tanto en las máximas como en mínimas, mientras que, en la magnitud del viento, solo se presentan máximas. Por otro lado, las pruebas de estacionariedad para cada una de las series, previa eliminación de valores nulos, confirmaron el comportamiento estacionario.

Adicionalmente, la Figura 10 muestra el comportamiento promedio diario. La velocidad promedio del viento inicia en 6.2 m/s a las 00:00 y disminuye en las horas siguientes, hasta alcanzar su valor mínimo de 4.5 m/s a las 6:00. A partir de esta hora, la velocidad promedio del viento se mantiene en el mismo valor hasta las 10:30, momento en el que comienza a aumentar durante las siguientes 10 horas alcanzando su punto máximo de 7.4 m/s a las 20:00 (Figura 10a). Por otra parte, la temperatura del aire es de 27.8 °C a las 00:00 horas, disminuyendo progresivamente durante las siguientes horas hasta alcanzar su valor inferior al promedio de 27.3° a las 06:00 horas. A partir de este punto, la temperatura promedio comienza a aumentar durante las seis horas siguientes, alcanzando su punto más alto de 28.9° a las 12:00; el valor es casi el mismo desde esta hora hasta las 14:00, punto desde el cual comienza a disminuir hasta las 18:00, alcanzando un valor de 28.3 °C.

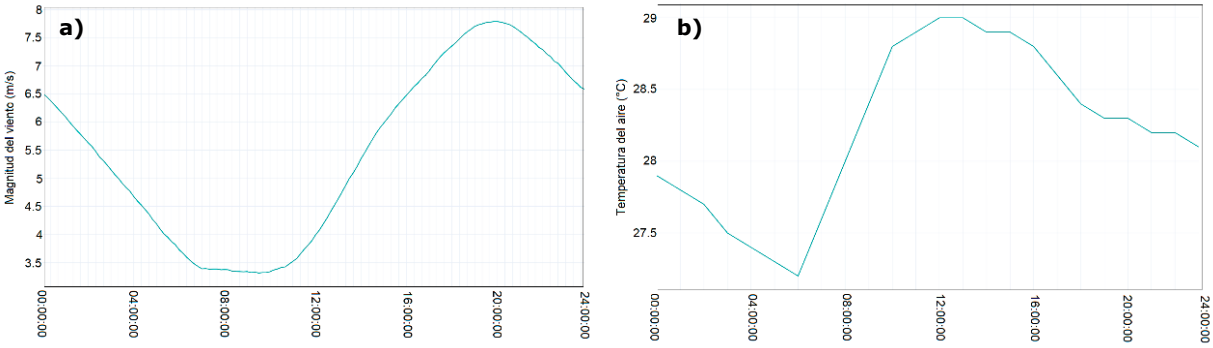


**Figura 10.** Promedio diario de la magnitud del viento (a) y la temperatura del aire (b).

### *Reconstrucción de los datos*

Como representación visual del método de reconstrucción (Figura 11), se muestra la temperatura del aire (°C) de 2016 a 2021, en donde la mediana para todos los años de la serie se denomina "Median". Estas series de tiempo reflejan el comportamiento que ha sido descrito por otros autores; los cuales señalan que la temperatura del aire oscila entre 26.2 °C y 27.36 °C (Gutiérrez *et al*., 2011), con los valores más altos registrados entre mayo y junio, y los valores mínimos durante enero, febrero y marzo (CIOH, 2022b).

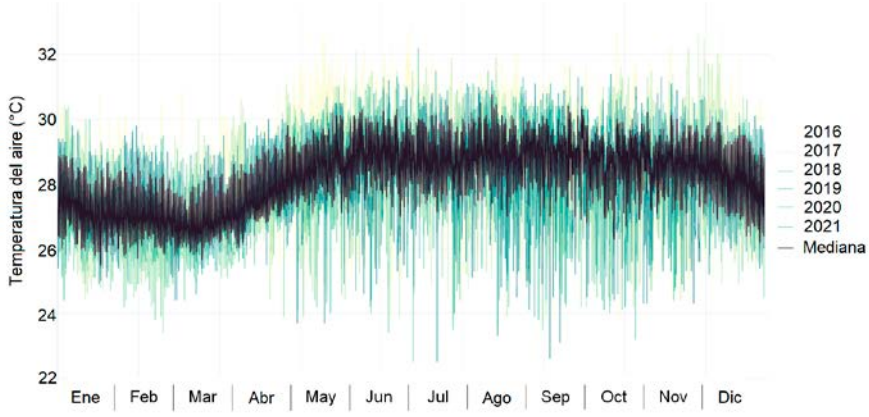


**Figura 11.** Serie superpuesta de la temperatura del aire (°C), de 2016 a 2021. Incluye la mediana de todos los años.

Esta primera aproximación para la reconstrucción de los vacíos de ambas series de tiempo resultó satisfactorio; ya que se completaron 52824 valores de magnitud del viento y 2421 valores de temperatura del aire. La Figura 12 muestra el resultado después de completar las series de tiempo de temperatura del aire y la magnitud del viento, entre los meses de mayo a julio de 2019. Gracias a este proceso estadístico, el cual se basa en la mediana, fue posible evitar alterar el comportamiento de la serie de tiempo. Adicionalmente, no está sesgado hacia valores extremos, ya que la mediana es una estadística consistente.

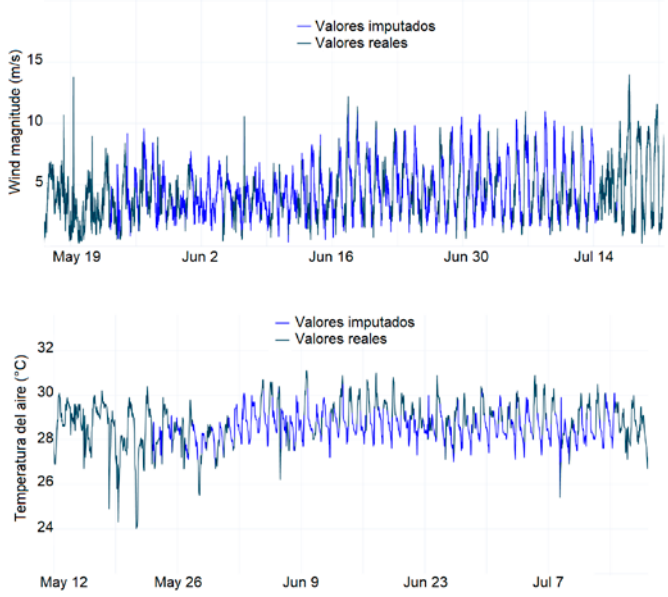


**Figura 12.** Valores imputados y reales en 2019 para la magnitud del viento (a) y la temperatura del aire (b).

Por otra parte, se evidenció que, para desfases consecutivos prolongados, el método de imputación reduce la varianza de la serie. Como ejemplo de esto, la Figura 13 muestra la reconstrucción de un intervalo de 264 horas consecutivas para la serie de temperatura del aire, y la Figura 14 la reconstrucción de un intervalo de 220320 minutos consecutivos para la serie de magnitud del viento. A pesar de lo anterior, se completaron los valores faltantes, independientemente de su tamaño.

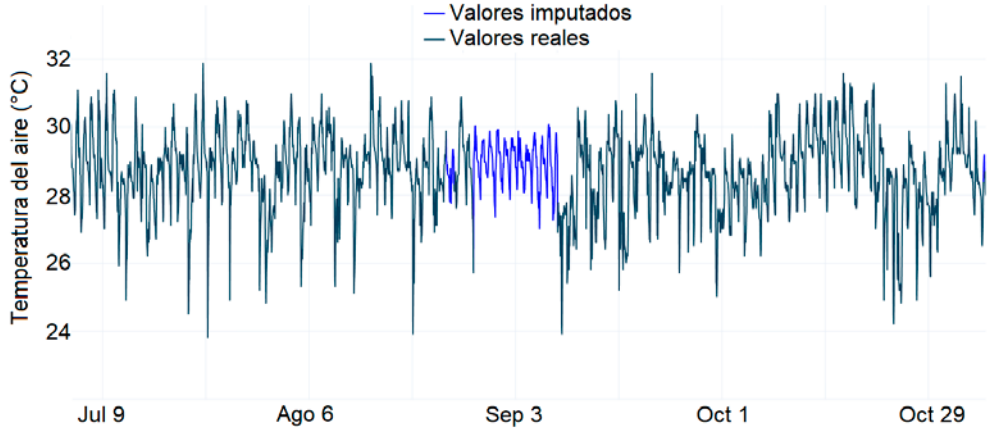


**Figura 13.** Serie reconstruida de temperatura del aire, desde agosto a septiembre de 2017.

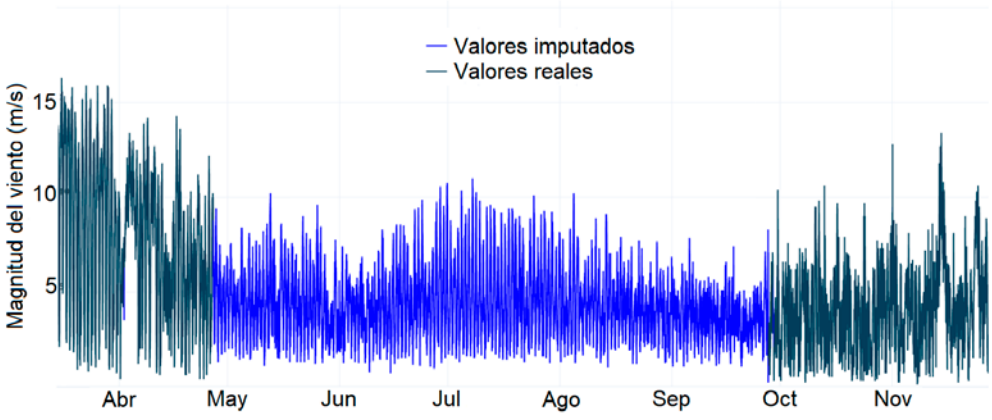


**Figura 14.** Serie reconstruida de la magnitud del viento, desde abril a septiembre de 2020.

### *Evaluación de los datos*

A partir de los modelos realizados para validar la serie de tiempo reconstruida de la temperatura del aire, se entrenaron y evaluaron exitosamente el modelo autorregresivo y el modelo de redes neuronales recurrentes con la métrica del error cuadrático medio, reportando un valor de 0.47 °C y 0.31 °C para los modelos AR y RNN, respectivamente. Esta es una primera aproximación sobre la utilidad de la metodología de reconstrucción implementada en este trabajo. Finalmente, la Figura 15 muestra el ajuste con el modelo RNN para el último mes de 2021, comparado con los valores reales.

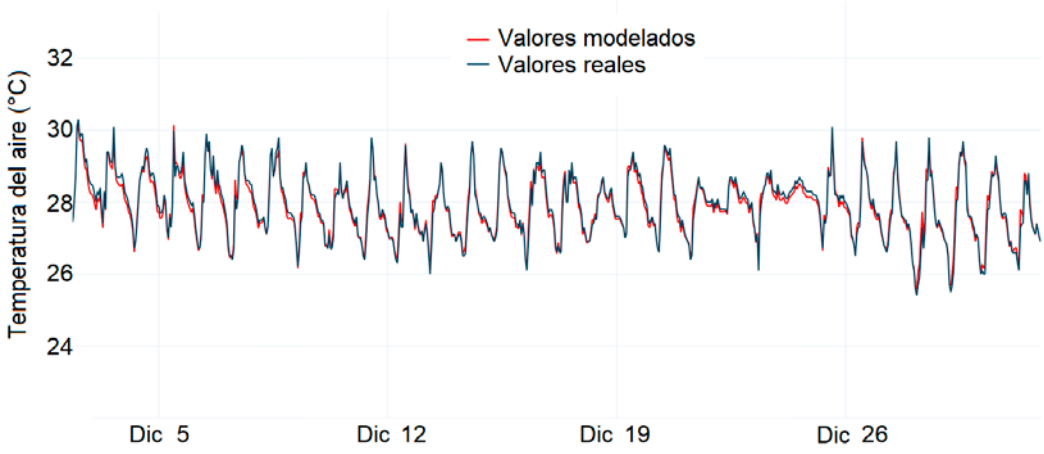


**Figura 15.** Modelado de datos en el último mes de 2021 con el modelo de RNN para la serie de temperatura del aire.

## Conclusiones

Desde la perspectiva de un marco común para la aplicación de la ciencia de datos, fue posible desarrollar una primera aproximación para la reconstrucción de datos meteorológicos de una estación costera en el Caribe colombiano. Para ello, se desarrolló un procedimiento específico de exploración, limpieza y extracción de las características de la información, que permitió realizar una imputación fundamentada, además de verificar el trabajo de la serie reconstruida como insumo para la creación de datos, pensando en modelos futuros.

En este estudio, la reconstrucción de las series de tiempo, permitió alcanzar el objetivo principal de reconstruir una serie temporal completa de temperatura del aire y magnitud del viento. Independientemente de todos los usos posibles que se puedan asignar a una serie de tiempo, los diferentes propósitos requieren de una línea de tiempo completa. El modelo generado proporciona un enfoque para la reconstrucción de series de tiempo bajo la necesidad de mantener la continuidad de los datos meteorológicos. Sin embargo, se puede considerar a la inclusión de otros métodos estadísticos y a la creación de un informe sobre los datos reconstruidos, como una próxima fase del prototipo.

Por último, la tecnología utilizada permite el fácil despliegue de una aplicación online para la visualización de datos, considerando este primer enfoque como un producto con el potencial para ser escalado a otras bases de datos relacionadas con otras estaciones meteorológicas en el ámbito nacional.

## Agradecimientos

Este estudio hace parte de los resultados del proyecto "Banco de Retos aplicando Ciencia de Datos", auspiciado por el Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia. El CIOH fue seleccionado con el reto "Reconstrucción de series temporales de variables meteorológicas", el cual fue desarrollado por el Equipo 122 en el marco del programa de capacitación de DS4A – Colombia, y apoyado por la Sección de Innovación y Tecnología del CIOH, así como por Cecoldo y la RedMpomm de la Dimar.

## Referencias bibliográficas

Acquaotta, F.; Fratianni, S. (2014). The importance of the quality and reliability of the historical time series for the study of climate change. Revista Brasileira de Climatologia 14, 20-38.

Almqvist, O. (2019). A comparative study between algorithms for time series forecasting on customer prediction: An investigation into the performance of ARIMA, RNN, LSTM, TCN and HM

Andrade, C.A.; Barton, E.D. (2000). Eddy development and motion in the Caribbean Sea. Journal of Geophysical Research: Oceans, 105(C11), 26191-26201.

Cendales, M. H.; Zea, S.; Díaz, J. M. (2022). Geomorfología y unidades ecológicas del complejo de arrecifes de las Islas del Rosario e Isla Barú (Mar Caribe, Colombia). Revista de La Academia Colombiana de Ciencias Exactas, Fisicas y Naturales, 26(101), 497–510.

Centro de Investigaciones Oceanográficas e Hidrográficas (28 de septiembre 2022 a). ¿Qué es CIOH? https://www.cioh.org.co/index.php/es/quienes-somos/mision-y-vision-quienes-somos

Centro de Investigaciones Oceanográficas e Hidrográficas (18 de septiembre 2022 b). https://www.cioh.org.co/meteorologia/Climatologia/ResumenCartagena4.php#:~:text=Temperatura%3A%20Las%20temperaturas%20m%C3%A1ximas%20en,31.0%C2%BAC%20y%2031.1%C2%BAC.

Chai, T.; Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions*, *7*(1), 1525-1534.

Dirección General Marítima (28 de septiembre de 2022 a). Red de Medición de Parámetros Oceanográficos y de Meteorología Marina (RedMpomm). https://dimar.maps.arcgis.com/apps/opsdashboard/index.html#/48d2c76148af428789abae6b3a8789de

Dirección General Marítima. (2022b). Serie de datos de velocidad y dirección del viento obtenidos con estación meteorológica automática ubicada en Islas del Rosario, Colombia. [2013-2021]. Centro Colombiano de Datos Oceanográficos (Cecoldo).

Dirección General Marítima. (2022c). Serie de datos de temperatura del aire obtenidos con estación meteorológica automática ubicada en Islas del Rosario, Colombia. [2013-2021]. Centro Colombiano de Datos Oceanográficos (Cecoldo).

Gholamy, A.; Kreinovich, V.; Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation.

Gutiérrez, C.; Marrugo, M.; Lozano-Rivera, P.; Sierra-Correa, P.; Andrade, C. (2011). Clima Marino. En: Zarza, E (Eds). El entorno ambiental del Parque Nacional Natural Corales del Rosario y de San Bernardo, 39-47.

Hatanaka, M. (1996). *Time-series-based econometrics: unit roots and co-integrations*. OUP Oxford.

Instituto de Hidrología, Meteorología y Estudios Ambientales (30 de septiembre 2022). *Meterología Aeronáutica*. http://bart.ideam.gov.co/web/inicio.htm

National Hurricane Center. (2 de octubre 2022). *Tropical cyclone report – Hurricane Iota*. https://www.nhc.noaa.gov/data/tcr/AL312020_Iota.pdf

Parques Nacionales de Colombia (30 de septiembre 2022). *Parque Nacional Natural Corales del Rosario y de San Bernardo*. https://www.parquesnacionales.gov.co/portal/es/ecoturismo/parques/region-caribe/parque-nacional-natural-corales-del-rosario-y-de-san-bernardo/

Majidpour, M.; Qiu, C.; Chu, P.; Pota, H.R.; Gadh, R. (2016). Forecasting the EV charging load based on customer profile or station measurement?. *Applied energy*, *163*, 134-141.

NDC de Colombia. (2020). Actualización de la Contribución Determinada a nivel nacional de Colombia (NDC). *Bogotá: MinAmbiente, DNP, CANCILLERÍA, AFD, Expertise France, WRI*.

Puerres, L.Y.; Bernal, G.; Brenner, M.; Restrepo-Moreno, S.A.; Kenney, W.F. (2018). Sedimentary records of extreme wave events in the southwestern Caribbean. Geomorphology, 319, 103-116.

Raschka, S.; Patterson, J.; Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193.Restrepo, J. C., Otero, L., Casas, A. C., Henao, A., & Gutiérrez, J. (2012). Shoreline changes between 1954 and 2007 in the marine protected area of the Rosario Island Archipelago (Caribbean of Colombia). Ocean & Coastal Management, 69, 133-142.

Restrepo, J.C.; Otero, L.; Casas, A.C.; Henao, A.; Gutiérrez, J. (2012). Shoreline changes between 1954 and 2007 in the marine protected area of the Rosario Island Archipelago (Caribbean of Colombia). Ocean & Coastal Management, 69, 133-142.

Rueda J.G., Otero, L.J.; Pierini, J.O. (2013). Caracterización hidrodinámica en un estuario tropical de Suramérica con régimen micro-mareal mixto (Bahía de Cartagena, Colombia). Boletín Científico CIOH, 31, 159-174.

Sessa, J., Syed, D. (2016). Techniques to deal with missing data. In *2016 5th international conference on electronic devices, systems and applications (ICEDSA)* (pp. 1-4). IEEE.

Taylor, H.T.; Ward, B.; Willis, M.; Zaleski, W. (2010). The Saffir-Simpson hurricane wind scale. Atmospheric Administration: Washington, DC, USA.

Wirth, R.; Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, pp. 29-39).

Zhang, Z.; Zhang, K.; Khelifi, A. (2018). Multivariate time series analysis in climate and environmental research. Cham: Springer International Publishing.