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

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

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]

Citation

Contreras Vargas, C.; Quintero-Ibáñez, J.; Solanilla, A. (2022). 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. *Bol. Cient. CIOH*; 41(2): 71-83. Printed ISSN 0120-0542 and online ISSN 2215-9045. DOI: https://doi.org/10.26640/22159045.2022.604

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 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.

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

[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

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.*

Introduction

The Center for Oceanographic and Hydrographic Research (CIOH) of the General Maritime Directorate (Dimar), carries out a diverse number of researches in a variety of subjects such as oceanography, hydrography, marine biology, environmental protection, and coastal zone management among others. Through management of information and report elaboration, a few research groups have conducted academic research, applied studies, and policy advice for more than 45 years (CIOH, 2022). Additionally, Dimar has the Network for the Measurement of Oceanographic and Marine Meteorology Parameters (RedMpomm), which aims to obtain, store, standardize and make available oceanographic and marine meteorology data, measured on the Colombian Caribbean and Pacific coasts and insular areas, manage and provide access to national oceanographic data resources, in order to improve comprehensive maritime, river and port security, preserve human life at sea and make decisions related to the exercise of the National Maritime Authority (Dimar, 2022a).

The RedMpomm is made up of meteorological and tide gauge stations, directional wave buoys and metocean buoys, whose transmission is in real time via cellular and satellite. The data generated in this network is used by the CIOH to carry out the institutional and scientific projects. However, there are gaps in the data generated from some stations, attributed to technical problems of sensors, which due to their location are difficult to access for timely repair; transmission inconveniences or some extreme event to which the equipment or its infrastructure may be exposed, thus affecting the continuity of the measurements.

The study of the climate and its variations, requires the use of long historical series. These datasets are essential since they are the basis for assessing century-scale trends, for the validation of climate models, as well as detection and attribution of climate change at a regional scale (Acquaotta and Fratianni, 2014). Completing the time series data set will increase the robustness of the information, which make it a high-quality input for better analysis and an improvement of the decision-making process in a variety of fields such as those related to maritime security, navigation, and climate change. Climatologists and environmentalists are striking to extract meaningful information from huge amount of observational record and simulation data for the climate system (Zhang, Zhang and Khelifi, 2018). Colombia's climate change goals cover strategies for mitigations but focuses on adaptation, which requires improving the data available on the natural capital of the country, as a first step in tracking changes of strategic ecosystems such as mangroves, wetlands, coral reefs, glaciers, oceans and tropical forests recognizing their intrinsic value and the environmental services they provide for Colombia and the world (NDC de Colombia, 2020). In this way, all the efforts to improve the data series of climate variables will result in better inputs for the execution of the activities of the maritime authority.

Taking into account the importance of the data generated in the time series, as well as the existence of gaps in them, the CIOH identified this issue and posed it as a data science challenge to reconstruct time series of meteorological variables of a coastal station in the Colombian Caribbean. To carry out this challenge and as a first pilot exercise, a coastal meteorological station belonging to the RedMpomm (Islas del Rosario) was selected. The selected variables were air temperature and wind velocity, both series cover the period from 2013 to 2021 and present gaps associated with observations with missing values. This study presents a methodological approach for the reconstruction of these time series, in which, the techniques of data science and statistics modeling

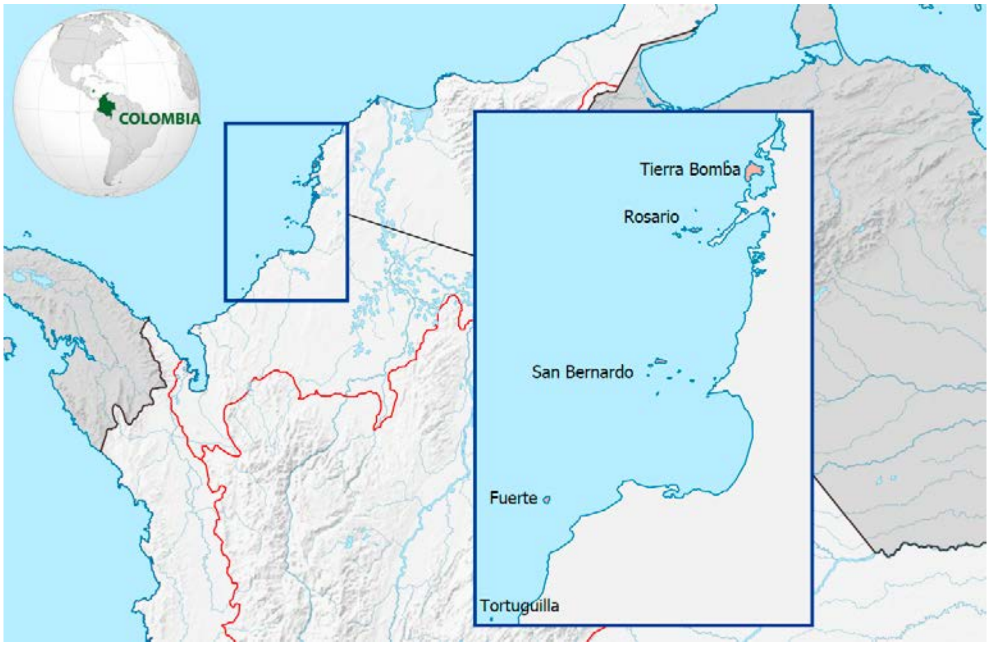

Figure 1. Map of the Rosario Islands Archipelago, Caribbean Sea.

were used to complete them with the aim that they can be used later by the researchers of the national maritime authority. The approach used constitutes a first exercise for the implementation of future models for completing times series facing gaps of information not only at Islas del Rosario but at different meteorological stations.

Study area

The study area is located 52 kilometers approximately from South-West of bay of Cartagena, between 75º41'47''W - 75º48'15''W and 10º14'33''N - N 10º41'45''N (Figure 1). The area includes 28 islands and cays built upon a set of ancient and successive coralline formations currently lying at least 3 m above mean sea level (Cendales, Zea and Díaz, 2002) and since 1977, the area has been legally protected as "Corales del Rosario and San Bernardo National Natural Park" (Parques Nacionales de Colombia, 2022).

The archipelago's marine dynamics is influenced significantly by the intensity and seasonality of the trade winds, that affects wave propagation in the shallow waters and rising sea levels (Restrepo *et al*., 2012). For the bay of Cartagena (the closest major city to the study area), the dry season refers to the increase in wind speeds in the area and the decrease in local rainfall. This begins in December and goes until March, registering winds from the northeast between 2.6 and 5.1 m/s and occasional maximums of 15.43 m/s. On the other hand, in the wet season (between August and November), the climatic conditions are characterized by low wind intensities, between 1.0 and 3.0 m/s, occasionally showing maximums of 5.0 m/s (Rueda, Otero and Pierini, 2013). In the Caribbean there is also an important synoptic event in meteorology, the "Veranillo de San Juan", a period without rain that normally occurs between July and August (Andrade and Barton, 2000). Puerres *et al*. (2018) shown that storms typically occur during the dry season associated with cold fronts and the major storms and other high-wind, extreme-wave events shape the beach ridges on Rosario Island, every 70 years.

Regarding the air temperature and based on an analysis of 59 years of information derived from reanalysis data (1950-2009), Gutiérrez *et al*. (2011) describe that the air temperature in the archipelago ranged between 26.2 °C and 27.36 °C. On the other hand, the data reported for the bay of Cartagena, indicate that the average temperature presents its highest values between May and June, with averages between 28.3 ºC and 28.4 ºC; in the same way, the minimum values of the average temperature are presented during the months of January, February and March, oscillating between 26.8 ºC and 27.1 ºC. The maximum temperature presents a multiannual average of 31.5 ºC, whit their highest values in June, July and August with averages between 31.9 ºC and 32.0 ºC (CIOH, 2022b).

Methodology

A Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology developed by Wirth and Hipp (2000) was implemented. This is described as a hierarchical model (from the general to the specific), which is divided into 6 phases in the life cycle of a project. This section describes the data understanding, data preparation, modeling, evaluation, and deployment phases (Figure 2).

The development of the methodology was implemented in Python, which is a high-level interpreted programming language. The community around the available tools and libraries, makes Python particularly attractive for workloads in data science, machine learning, and scientific computing (Raschka, Patterson, and Nolet, 2020). In addition, an interactive dashboard was designed to upload the files with the incomplete time series. This was implemented using Amazon Web Service EC2 as a virtual machine, a Docker container for run application, and this application was built using the Python Dash framework (Figure 3).

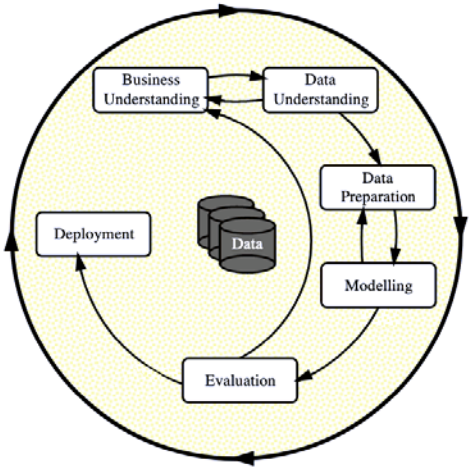

Figure 2. Phases of the CRISP-DM Process Model (Wirth and Hipp., 2000).

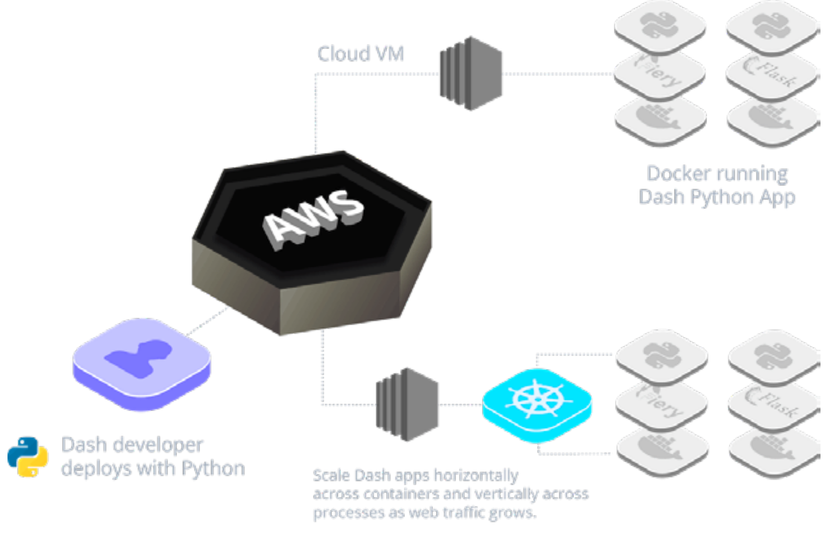

Figure 3. Interactive dashboard deployment process. Source https://plotly.com/

Data understanding

Data collection

The data was collected by an automatic weather station located in Isla Naval of the archipelago Islas del Rosario, which belongs to the RedMpomm (Figure 4). The wind speed and the air temperature time series were supplied by the Colombian Oceanographic Data Center (Cecoldo) since August 19, 2013, to December 31, 2021, in a CSV format, with a periodicity of every 10 minutes for the wind speed (Dimar 2022b) and every hour for the air temperature (Dimar 2022c).

Data exploration

In this process, the extreme values in the series were analyzed in the climatic context of the dates on which these extreme values were recorded.

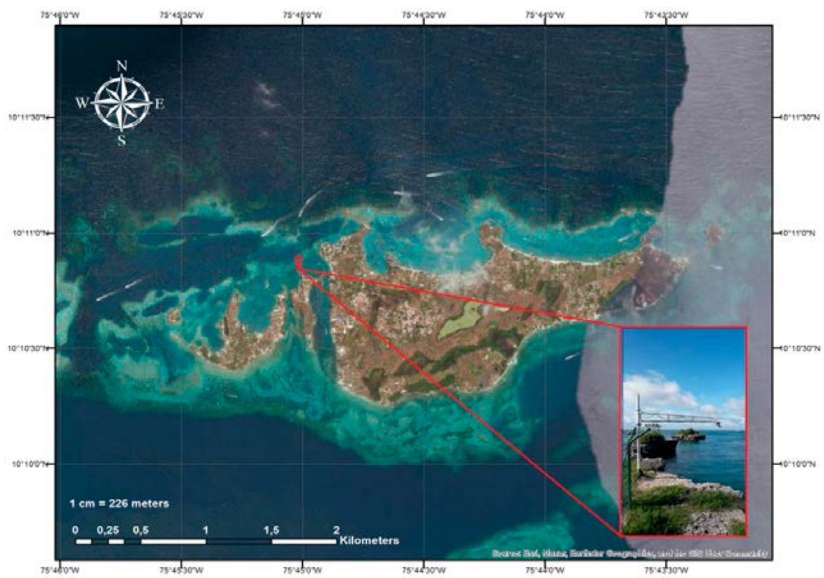

Figure 4. Location of the automatic weather station located in Isla Naval (Islas del Rosario, Colombian Caribbean) belonging to RedMpomm.

Data preparation

Data cleaning

According to the standard used by Cecoldo, the data not reported by the sensors are considered missing values and they are filled with the value: -99999. For data processing purposes, these values were converted to null. In addition, there are some values that do not correspond to the environmental conditions of the place where they are measured. According to the climatological context for the Colombian Caribbean coastal zones (minimum, average and maximum air temperatures), values outside the range between 10 °C and 55 °C are not expected (Instituto de Hidrología, Meteorología y Estudios Ambientales [IDEAM], 2022). While, smaller values are possible in the measurement of the sensors, it would be incongruous in the data; therefore, those values were becoming additional gaps. This procedure was also carried out with the wind magnitude, where the threshold was 75 m/s, lower limit for a category 5 hurricane on the Saffir Simpson scale (Taylor *et al*. 2010). Finally, the total percentage of gaps for each variable was obtained.

Data description

To characterize the variables, air temperature and wind magnitude, two box plots were made. These graphs allow visualizing the distribution of the variables, the interquartile range, and the atypical values. In addition, a stationarity test was performed for both time series. This is important for the development of models and imputation methods, for which the Dickey-Fuller test was implemented, which has the null hypothesis that a unit root is present in an autoregressive time series model. The alternative hypothesis is different depending on the version of the test used, but it is usually one of stationarity or trend-stationarity (Hatanaka, 1996). In addition, the variables of interest are grouped by time of day, that is, obtaining the average air temperature or wind speed for each hour of the day, in order to explore the possible cyclical behavior of these series.

Data reconstruction

An imputation method based on the median was implemented. The near point imputation method uses nearby values to order, then selects the median as a replacement for the missing value, with the advantage that the replacement belongs to the data (Majidpour *et al*., 2016). Even, imputation using the median, outperforms imputations based on k-nearest neighbor algorithms in some data science applications (Sessa and Syed, 2016). The approach used is summarized as follows (Figure 5):

1. For each year in the time series, the variable of interest is grouped by the hour of the day and day of the year.
2. The median is obtained for each day and specific hour of all the years.
3. Finally, with "Median" the time series is imputed in the years where there are gaps.

Figure 5. The process implemented for median-based imputation.

Data modeling

Only for the air temperature series, two models were built. For this, the total data was first divided non-randomly with the following values 80% - 20% in training and testing respectively (Gholamy, Kreinovich and Kosheleva, 2018). The first of these models was an autoregressive (AR) with a model parameter of 27 lags and the other a recurrent neural network (RNN) with GRU units, both implemented in Python. These have different advantages depending on the application, where those based on neural networks have an extra factor of complexity in terms of the correct search for hyperparameters and the computation necessary to train them, with the classic models being easier to use (Almqvist, 2019).

Model evaluation

The root mean square error metric has been used to evaluate the fit of both models. It is commonly used to measure model performance in meteorology, air quality, and climate research studies (Chai and Draxler, 2014).

Results and discussion

Data exploration

Figure 6 shows the time series of the wind magnitude for the evaluated period. From it, an interannual seasonality is observed, as well as extreme values, which were registered during February 2019 and 2020 (since April 27 to September 27). These values are probably calibration errors or damage to the sensor, since there is no record of events that have occurred in the area during that period that have generated sustained wind speeds above 70 m/s. Despite the hurricane Iota happened in 2020, the dates of the extreme values (April-September), do not agree with those of formation and passage of the hurricane through the Colombian Caribbean, which happened in November (National Hurricane Center, 2022). Therefore, these values were converted as additional null values.

Data cleaning

The time series after the cleaning process are shown in Figures 7 and 8. The total percentage of null values were 12.12% and 3.3% for the wind magnitude and air temperature, respectively. For the wind magnitude, the highest values were recorded in the dry season and the lowest, in the rainy season, agreeing with that described by other authors for the study area (Rueda, Otero and Pierini, 2013). This behavior contrasts with air temperature, variable for which the lowest values are obtained at the beginning of the year and the highest are recorded at the end.

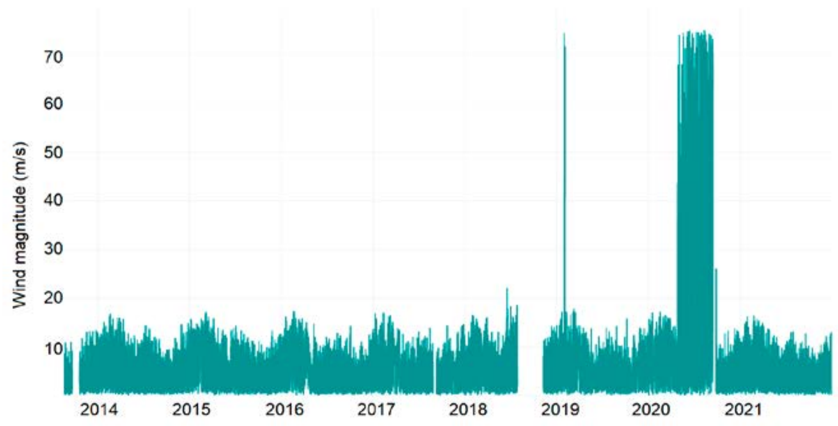

Figure 6. Original time series for wind magnitude.

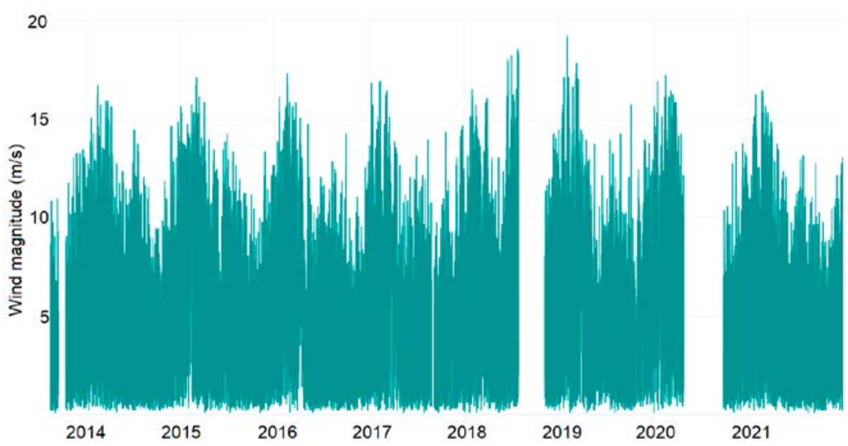

Figure 7. Wind magnitude (m/s) time series after the cleaning process.

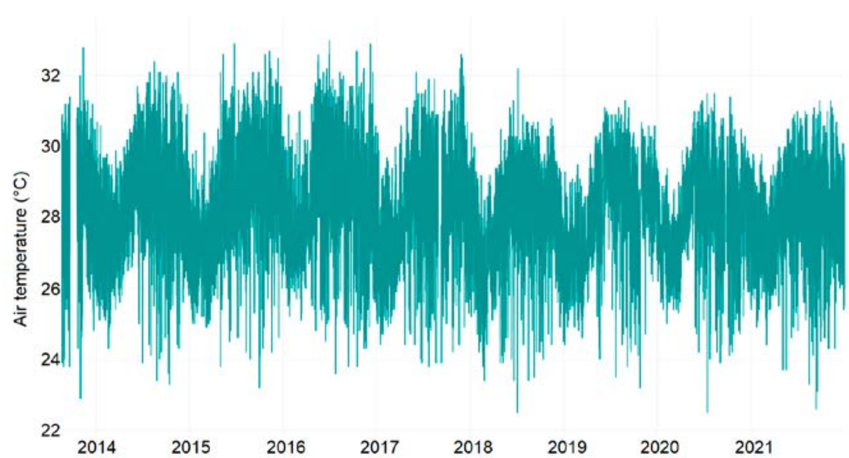

Figure 8. Air temperature (°C) time series after the cleaning process.

Data description

From the box plots, the atypical values were identified (red dots in Fig. 9). For the wind magnitude series, the median was 4.7 m/s and this presented a maximum outlier of 19.3 m/s (Fig. 9a); while for the air temperature series, the median was 28.2 °C with a maximum outlier of 33 °C (Fig. 9b).

These graphs allowed to identify that for the air temperature, outliers can be presented both in maximums and in minimums, while in the wind magnitude, only maximums are presented. On the other hand, the stationarity tests for each of the series, first eliminating the null values, confirmed the stationary behavior.

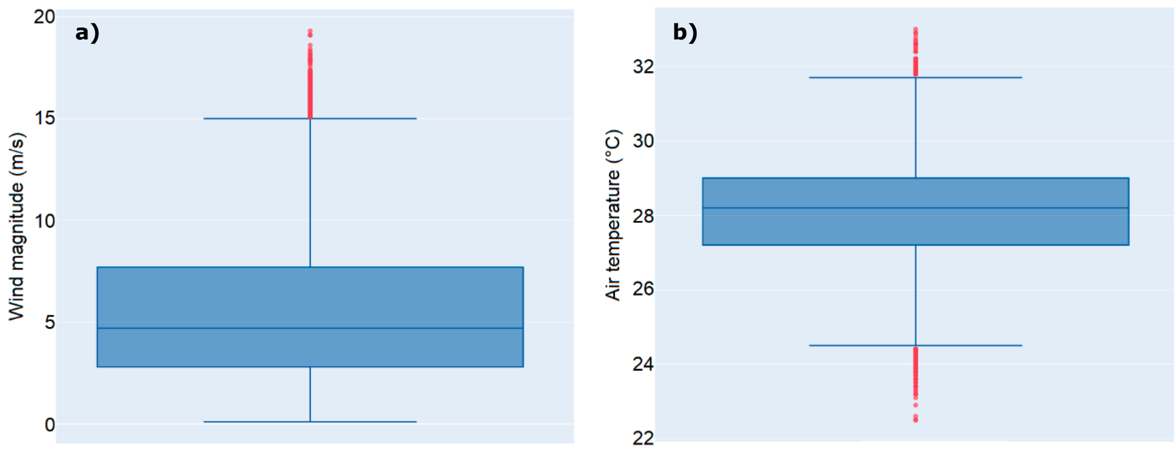

Figure 9. Boxplot for wind magnitude (a) and air temperature (b).

Additionally, the daily average behavior is shown in Figure 10. The average wind speed at 00:00 starts at 6.2 m/s, decreasing through the following hours until its lower value of 4.5 m/s at 6:00. From this hour, the average wind speed keeps the same value until 10:30, point in which starts to increase for the next 10 hours reaching its highest points of 7.4 m/s at 20:00 (Fig. 10a). On the other hand, the air temperature at 00:00 is 27.8 °C, decreasing through the following hours until its lower value of the average of 27.3 °C at 06:00. From this point, average temperature starts to increase through the next 6 hours, reaching its highest points of 28.9 °C at 12:00, the value is nearly the same from this hour until 14:00, starting to decrease until 18:00 reaching a value of 28.3 °C.

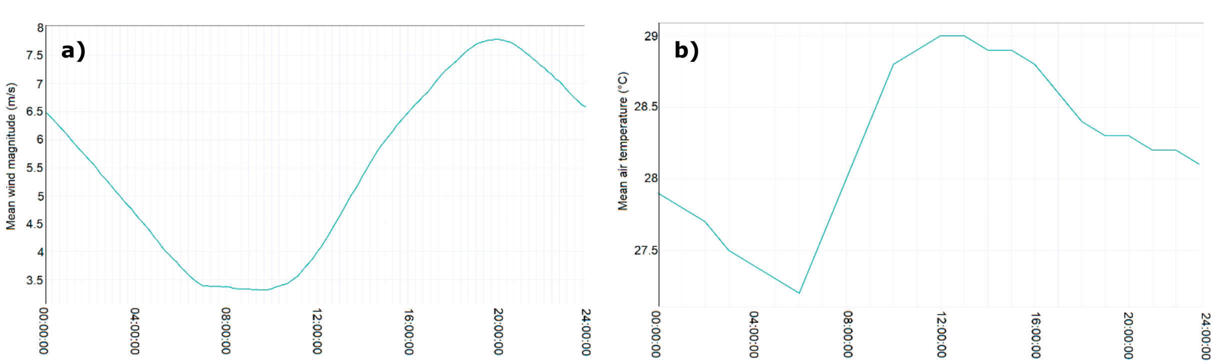

Figure 10. Daily average of wind magnitude (a) and air temperature (b).

Data reconstruction

As a visual representation of the reconstruction method, Figure 11 shows the air temperature (°C) from 2016 to 2021, where the median for all years of the series is called "Median". These time series reflect the behavior described by other authors, which describe the air temperature ranged between 26.2 °C and 27.36 °C (Gutiérrez *et al*., 2011) and the highest values between May and June and the minimum values during January, February and March (CIOH, 2022b).

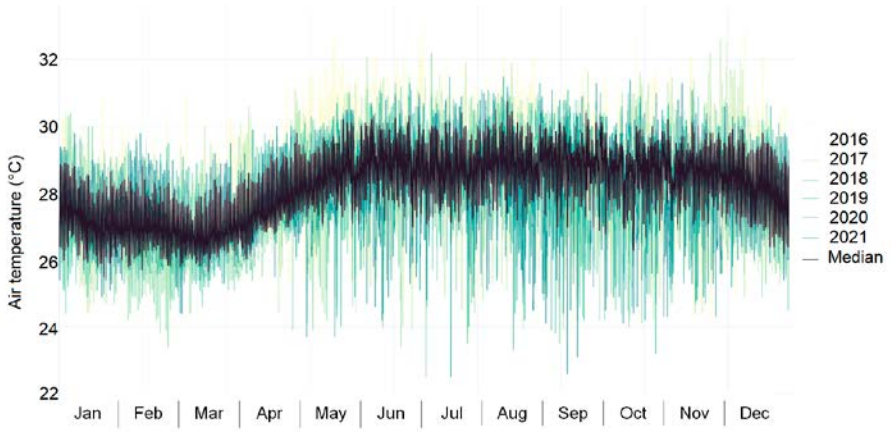

Figure 11. Air temperature series (°C) from 2016 to 2021 superimposed, including the median of all years.

This approximation to the gap reconstruction of both time series were satisfactory, that is; 52824 wind magnitude values and 2421 air temperature values were completed. Figures 12 show the result after completing the time series, air temperature and wind magnitude from the May to July of 2019. This statistical treatment, based on the median, allows not to alter the behavior of the time series, and is not biased towards extreme values, since the median is a robust statistic.

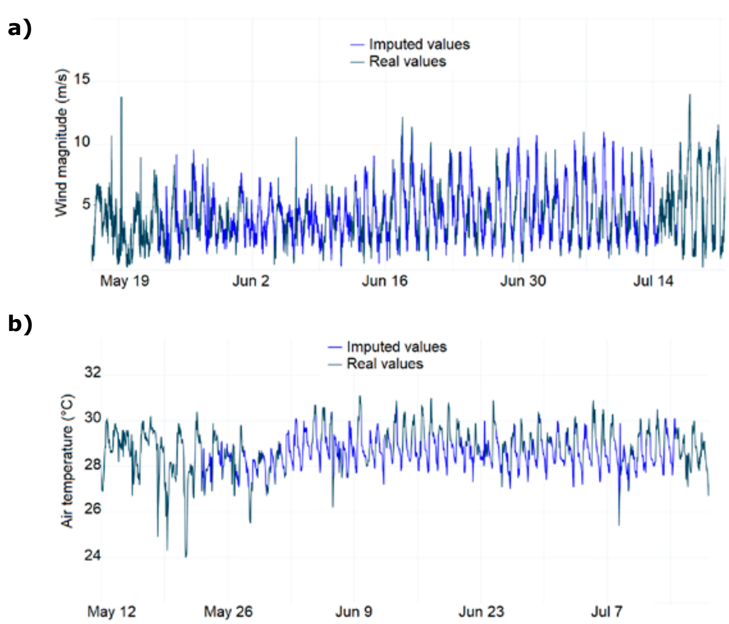

Figure 12. Imputed and real values in 2019 for wind magnitude (a) and air temperature (b).

On the other hand, it was evident that for prolonged consecutive gaps, the imputation method reduces variance of series. As an illustration of this, the reconstruction of a gap of 264 consecutive hours for the air temperature series is shown in Figure 13 and the reconstruction of a gap of 220320 consecutive minutes for the wind magnitude series is shown in Figure14. Even so, the gaps were filled regardless of their size.

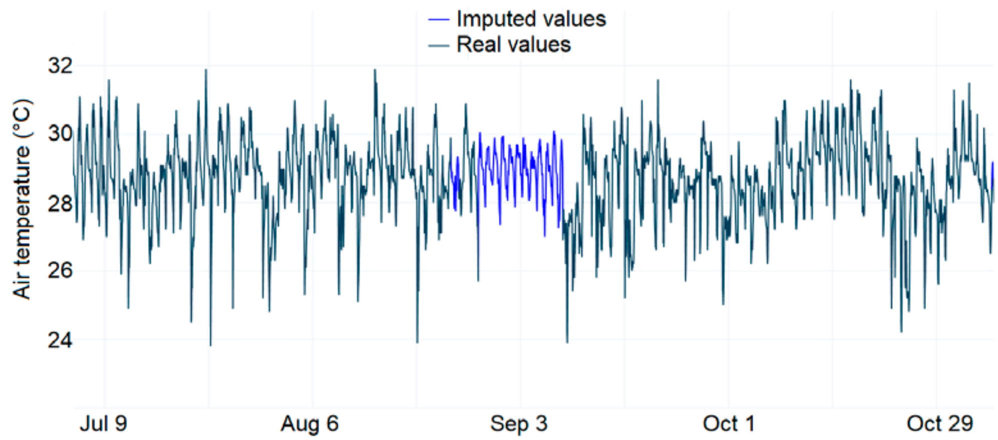

Figure 13. Reconstructed air temperature series in the period from August to September 2017.

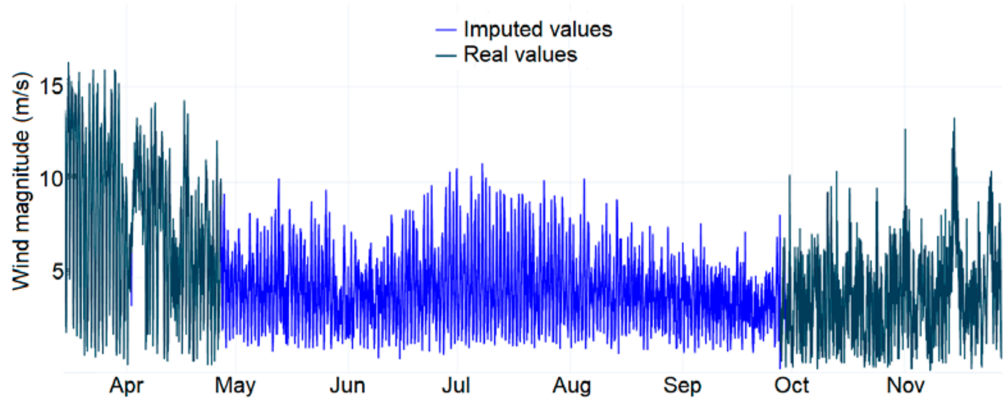

Figure 14. Reconstructed wind magnitude series in the period from April to September 2020.

Data evaluation

From the models made for validation of the reconstructed air temperature time series, one autoregressive model and another of recurrent neural networks were successfully trained and evaluated with the root-mean-square error evaluation metric, reporting a value of 0.47 °C and 0.31 °C for the AR and RNN models, respectively. This is a first approximation to the usefulness of reconstruction methodology carried out in this work. Finally, Figure 15 shows the fit with the RNN model for the last month of 2021 compared to the real values.

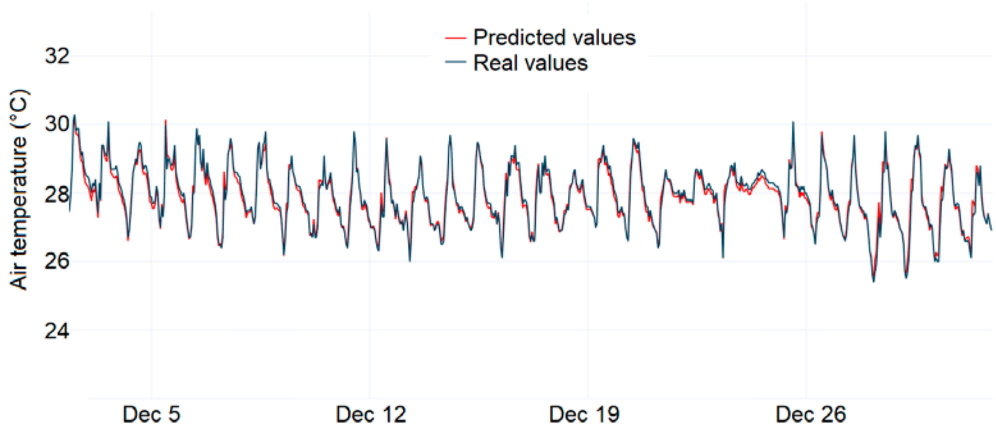

Figure 15. Data modeling in the last month of 2021 with the model of RNN for the air temperature series.

Conclussions

From the perspective of a common framework for applied data science, it was possible to develop a first approach for the reconstruction of meteorological data from a coastal station in the Colombian Caribbean. For this, a specific treatment of exploration, cleaning and extraction of the data characteristics was developed, which allowed to carry out a substantiated imputation, in addition to verifying the reconstructed series work as an input for the creation of data, thinking in future models.

In this study, the time series reconstruction fulfilled the main goal to reconstruct a whole time series of air temperature and wind speed. Regardless to all the possible uses that a time series can provide, the different purposes require a whole timeline. The generated model provides an approach for time series reconstruction in needs of sustain a continuity of weather data; nevertheless, the inclusion of other statistical methods and the creation of a report for the reconstructed data can be considered as a next phase of the prototype.

Finally, the technology used, allows the easy deployment of a web application for the data visualization, considering this first approach as a product with the potential to be scaled to other databases related to another weather stations throughout the national territory.

Aknowledgments

This study is part of the results of the initiative "Bank of challenges applying data science" sponsored by the Ministry of Information Technology and Communications of Colombia. The CIOH was selected with the challenge "Reconstruction of time series of meteorological variables" which was developed by the Team 122 within the training program of DS4A - Colombia and supported by the Innovation and Technology Section of the CIOH as well as by Cecoldo and RedMpomm belonging to Dimar.

Bibliography

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.