DYNA
http://dyna.medellin.unal.edu.co/

UNIVERSIDAD NACIONAL DE COLOMBIA

Data mining algorithms for predicting the behavior of environmental indicators

Liz Pérez-Martínez [a], Manuel Alejandro Naranjo-Rey [a], Orlando Santos-Pérez [b], Juan Alfredo Cabrera-Hernández [a] & Dianelys Nogueira-Rivera [a]

[a] *Facultad de Ciencias Técnicas, Universidad de Matanzas, Matanzas, Cuba.*
lizy.perez@umcc.cu, manuel.naranjo@umcc.cu, alfredo.cabrera@umcc.cu, dianelys.nogueira@umcc.cu
[b] *Empresa de Proyectos de Arquitectura e Ingeniería de Matanzas (EMPAI). Matanzas, Cuba. orlando-santos@empai.cu*

Received: April 12th, 2021. Received in revised form: November 2nd, 2021. Accepted: November 11th, 2021.

Abstract
The need to embrace adequate entrepreneurial focuses to achieve a better environmental performance constitutes an imminent task. Developing predictive models for environmental indicators constitutes the main objective of this paper. An information-technology tool that backs up the decision making will enable an efficient entrepreneurial environmental management, in such a way that they avoid errors that are commented at the present time. The application of techniques of data mining enabled capturing the last bosses and to reply to them, in addition to accomplish estimates with new data or out of sample, as well as inferring behaviors and future results, for the sake of anticipating possible situations of deterioration that compromise the environmental sustainability. The experiments designed to compare to the results in classification using the predictive models, prove that the percentage of error oscillates between the 4 % and the 5 %, what demonstrates extremely good degree of precision (height), according to the scales of verification.

Keywords: prediction; classification; ARIMA; temporal serie.

Algoritmos de minería de datos para la predicción del comportamiento de indicadores ambientales

Resumen
La necesidad de adoptar adecuados enfoques empresariales para lograr un mejor desempeño ambiental constituye una tarea inminente. Desarrollar modelos predictivos para indicadores ambientales constituye el objetivo principal de este artículo. Para ello se emplearon tecnologías que demostraron ser competentes para el logro del mismo. La aplicación de técnicas de minería de datos permitió capturar los patrones pasados y replicarlos, además de realizar estimaciones con datos nuevos o fuera de muestra, así como inferir comportamientos y resultados futuros, en aras de anticipar posibles situaciones de deterioro que comprometan la sostenibilidad ambiental. Los experimentos diseñados para comparar los resultados en la clasificación al emplear los modelos predictivos, demuestran que el porcentaje de error oscila entre el 4% y el 5%, lo que evidencia un grado muy bueno (alto), de acuerdo con las escalas de comprobación.

Palabras clave: predicción; clasificación; ARIMA; serie temporal.

1. Introducción

La gestión ambiental cubana ha estado caracterizada por cambios bruscos e inesperados en direcciones muchas veces contrapuestas. Esto ha conllevado a repensar el empleo de las técnicas normalmente utilizadas en el tratamiento de una realidad que, de tan cambiante, se ha convertido en incierta. En el camino hacia la búsqueda de soluciones y la prevención de la crisis en el futuro, se hace imprescindible el empleo de herramientas y procesos que ayuden al correcto desenvolvimiento de las entidades empresariales en lo que a su gestión ambiental se refiere [1]. De ahí la importancia de contar con herramientas que ayuden en el análisis de la información, tales como los sistemas de apoyo a la toma de

How to cite: Pérez-Martínez. L., Naranjo-Rey, M.A., Santos-Pérez, O., Cabrera-Hernández, J.A. and Nogueira-Rivera, D., Algoritmos de minería de datos para la predicción del comportamiento de indicadores ambientales.. DYNA, 88(219), pp. 228-236, October - December, 2021.

decisiones. Su propósito es ayudar a la administración en la definición de tendencias, señalamiento de problemas y toma de decisiones inteligentes. Su función básica es recolectar datos operacionales del negocio y reducirlos a una forma que pueda ser empleada para analizar el comportamiento del mismo.

La constante mutabilidad a la que se ven sometidos los fenómenos ambientales no permite, en la mayor parte de los casos, tomar en consideración datos del pasado para poder establecer inferencias del futuro. Es por ello que la preparación de una decisión, simple o compleja, se convierte en una actividad organizativa del pensamiento en la que inevitablemente se combinan intuición y lógica.

Adicionalmente, son muchas las ocasiones en que está latente la ocurrencia de problemas relacionados con la obtención de resultados estadísticamente significativos, derivados del uso inapropiado de técnicas estadísticas y econométricas, además de la recurrente ausencia de datos o en otros casos y la duplicidad de los mismos. Luego, se está en presencia de modelos econométricos muy útiles, desde el punto de vista de las relaciones de causalidad que describen, pero que pierden su fuerza al mostrar resultados pobres en sus parámetros estadísticos formales o al utilizarse para plazos largos, cuestión no concebida para algunos de estos modelos.

Es en este sentido, que los modelos predictivos de minería de datos son apropiados para tratar problemas que tienen evidentes relaciones no lineales a largo plazo y donde se requieren pronósticos con un elevado nivel de fiabilidad formal (bondad del ajuste) y confiabilidad (apropiada selección de variables a relacionar).

En [2] se plantea la valoración de indicadores de sostenibilidad a partir de la creación de un grupo de expertos y la aplicación del método Delphi orientado a dehesas específicamente. Sousa, S. N., Estruch-Guitart, V. & García, C. (2020) [3] plantean el uso de indicadores de Fuerzas Motrices-Presiones-Estado-Impactos-Respuestas (FPEIR) para evaluar la explotación sobre el agua en las islas Baleares. En [4] se comparan diferentes modelos para la predicción del nivel de material particulado (MP2.5) en Santiago de Chile y a partir de aquí inferir la calidad del aire. Mediante series de tiempo se intenta predecir el calentamiento global a partir de la temperatura media mundial en los últimos 165 años en la investigación [5]. Mateo Pérez, V., Mesa Fernández, J. M., Villanueva Balsera, J. & Terrados Cristos, M. (2021) [6] predicen los parámetros del agua de entrada en una planta de tratamiento, con el enfoque de mejorar la eficacia de la instalación. En [7] el objetivo de la investigación se encaminó a diseñar un sistema de indicadores medioambientales para medir y evaluar en forma sistemática los impactos medioambientales en la industria láctea de Sibanicú, Cuba. Establecer indicadores medioambientales en las empresas del sector químico del Área Metropolitana del Valle de Aburrá, de acuerdo con el Global Reporting Initiative Standards 2016, por medio de técnicas tales como encuestas, entrevistas y la realización de un *benchmarking* de revelación de indicadores medioambientales, donde se identifica el nivel de reporte respecto a cada indicador, es el tema tratado por Restrepo, J. M. H., Pérez, A. O., Gutiérrez, J. P. & García, J. A. C. (2021) [8]. En [9] se muestran los resultados del seguimiento realizado a la aplicación de Indicadores de Desempeño Ambiental en la industria manufacturera de la ciudad de Cali y Yumbo – Colombia desde el año 2005 hasta 2013. La calidad de agua hace referencia a los valores apropiados de los parámetros fisicoquímicos y/o biológicos del agua para un uso específico, se plantea en [10] y propone el monitoreo de este indicador para proporcionar información útil a fin de procesarla por herramientas de aprendizaje automático con fines predictivos. Este documento tiene como objetivo presentar una revisión de las técnicas de aprendizaje automático utilizadas en la estimación de la calidad de agua. El trabajo investigativo compara redes neuronales (RN), sistemas de inferencia neurodifusa (Anfis) y máquinas de vectores de soporte (MVS). En [11], la investigación realiza una integración de indicadores medioambientales e indicadores de desempeño operacional (KPIs) particularizado para gráneles sólidos sucios, de forma que una vez definida la lista de indicadores se realice un estudio de comportamiento e interrelación de unos parámetros con otros para conseguir describir tendencias o comportamientos, realizándose también un análisis Clúster.

Como se observa en la bibliografía consultada, se abordan temas relacionados con indicadores ambientales y sus parámetros desde diferentes ópticas. Algunos casos proporcionan soluciones a problemas muy específicos de determinada región o industria y en otros se realizan análisis puntuales del estado actual o posible tendencia de un indicador o parámetro en concreto. Sin embargo no se aprecia ninguna solución que brinde una herramienta informática que permita observar el estado actual de los indicadores medioambientales y proyecciones a futuro a partir de los parámetros estimados por modelos de predicción de alta fiabilidad, para el análisis en tiempo real por expertos en la materia.

La bibliografía consultada refleja el escaso empleo de herramientas informáticas para la gestión de indicadores ambientales, que asistan la toma de decisiones y viabilicen la actividad humana, lo que constituye un obstáculo en el objetivo estratégico de convertir la información en conocimiento; dicho conocimiento será trascendental para lograr un desarrollo sostenible. Por lo que, para la gestión ambiental cubana, contar con herramientas que apoyan la toma de decisiones es crucial para alcanzar el tan ansiado desarrollo sostenible.

A partir de lo expuesto la investigación desarrolla modelos predictivos para indicadores ambientales que apoyen la toma de decisiones para una eficiente gestión ambiental empresarial.

2. Materiales y métodos

2.1 Adquisición de los datos

La predicción de un indicador se hace posible si se logra encontrar un modelo capaz de pronosticar el comportamiento de los parámetros que lo sintetiza. Estos son predecibles si se cuenta con un amplio historial de datos y adaptamos un modelo capaz de simular su comportamiento. Para la creación y entrenamiento del modelo se empleó una base de

datos que almacena los valores de la temperatura, uno de los parámetros ambientales que más influyen en el comportamiento de los indicadores y que refleja una marcada estacionalidad, fenómeno muy recurrente en variables ambientales y que es conveniente si se desea construir un modelo extrapolable a otros parámetros con el fin de predecir el comportamiento del indicador.

Los datos que se utilizan en esta investigación son extraídos de la web Banco Mundial de Datos. Esta web ofrece datos de acceso abierto y gratuito sobre el desarrollo mundial. Fueron seleccionados para el entrenamiento de los modelos los datos Temperatura Mensual (˚C) – Cuba del Centro de Análisis de Información, División de Ciencias Ambientales del Laboratorio Nacional de Oak Ridge (Tennessee, Estados Unidos).

Esta base de datos posee la temperatura promedio en Cuba desde 1901 hasta 2016, reporta un total de 1392 observaciones puesto que tiene una frecuencia mensual.

El Banco Mundial de Datos brinda la posibilidad de descargas en diversos formatos como csv, xml y excel. Los datos fueron descargados en formato csv, e introducidos en una base de datos postgreeSQL.

La investigación se realiza en el software RStudio, un entorno de desarrollo integrado para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo. Este software es a menudo utilizado en este tipo de trabajos porque, a decir de [12], es un software libre muy potente para el análisis de datos, generador de gráficos y cuenta con paquetes especializados que sirven a diferentes ciencias.

RStudio está disponible para Windows, Mac y Linux o para navegadores conectados a RStudio Server o RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux). RStudio tiene proporciona el entorno informático estadístico R.

La elección de la herramienta se sustenta en que es software libre, maneja un entorno de trabajo agradable y cuenta con una vasta cantidad de librerías para el desarrollo de aplicaciones estadísticas. Ofrece opciones de representación a partir de gráficos y tablas y cuenta con la posibilidad de conectarse a innumerables bases de datos entre las que se encuentra postgreeSQL, necesaria para el desarrollo de la investigación.

La exploración de los datos se inició con una muestra como se aprecia en la Tabla 1. Se observa que existen 12 mediciones de la temperatura por año, una por cada mes. No existen celdas vacías o con errores por lo que no fue necesario eliminar ninguna fila, aunque el tratamiento a futuro en estos casos será la ponderación de las medidas anterior y posterior para determinar el valor faltante.

Estadísticamente se cuenta con una variable cuantitativa continua, que arroja un valor mínimo de 19.3687°C, máximo de 28.704°C y promedio de 25.2684°C aproximadamente. Las medidas de dispersión de aproximan a 1.6751 la desviación media, 3.8386 la varianza y 1.9585 la desviación típica. Se observa la distribución de frecuencias en la Tabla 2.

Tabla 1.
Muestra de los datos.

1901	Enero	21.3934
1901	Febrero	21.3323
1901	Marzo	22.6022
1901	Abril	22.7852
1901	Mayo	24.909
1901	Junio	26.673
1901	Julio	26.8807
1901	Agosto	26.8042
1901	Septiembre	26.8741
1901	Octubre	26.015
1901	Noviembre	21.8677
1901	Diciembre	21.5215

Fuente: Elaboración propia.

Tabla 2.
Tabla de frecuencias.

Categorías	Frecuencia	Porcentaje
[19-21)	26	1.867816092
[21-23)	184	13.2183908
[23-25)	386	27.72988506
[25-27)	469	33.69252874
[27-29)	327	23.49137931
Total:	1392	100

Fuente: Elaboración propia.

Para continuar el dataset se convierte en una serie temporal con el objetivo de convertir los datos en valores procesables por los modelos de análisis y predicción. En R la serie temporal es expresada con un mínimo de dos vectores, uno para la fecha y el resto para los valores de la variable estudiadas. El software utilizado ofrece facilidades para esta conversión, a partir de la especificación de los datos, la frecuencia de medición y la fecha de inicio podemos obtener la serie temporal correspondiente. En [13] se define como la serie de observaciones de una o más variables en el tiempo. Resalta la importancia del tiempo como dimensión en esta área, dado que los eventos pasados pueden influir en los futuros.

Según [14], el uso de series temporales es muy común en diferentes campos como la economía para la cotización de las existencias en los mercados o la cantidad de desempleados, las ciencias sociales para las inscripciones en las escuelas, la medicina para la cantidad de casos de influenza en un período dado o la evaluación de diferentes medicamentos en pacientes con hipertensión, la meteorología para el análisis de diferentes fenómenos naturales, entre otros.

Los datos se convierten en una serie temporal para realizar un análisis más profundo y la posterior aplicación de diferentes modelos de predicción. Para esto se consideró el inicio de las mediciones (1901) y la frecuencia en que se realizaron (12).

Una vez conformada la serie temporal en R, la forma más sencilla de comenzar fue mediante su representación en un gráfico de secuencia, representado en la Fig. 1.

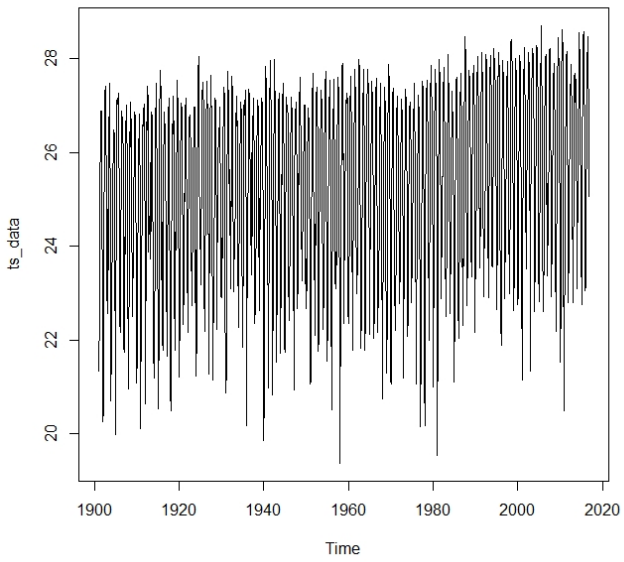

Figura 1. Gráfico de secuencia.
Fuente: Elaboración propia.

Una serie puede ser no estacionaria por una variación en la media, una variación en la varianza o por la presencia de estacionalidad. Esto significa que si existe alguno de estos casos es necesario aplicar transformaciones en la serie. Por lo observado se evidencia que la serie no es estacionaria en media porque esta varía en el tiempo. También presenta variación cada cierto periodo y es normal en este tipo de estudios porque las variables ambientales y principalmente la temperatura tienden a disminuir o aumentar en el tiempo (ejemplo en verano o invierno).

A partir del análisis anterior se decidió continuar mediante los métodos de descomposición y de suavizado exponencial para contrastar sus resultados y elegir el mejor modelo.

2.2 Método de descomposición

Estos métodos tratan de separar la serie en subseries de tendencia, estacionalidad y ruido por lo que son especialmente descriptivos. ARIMA es muy utilizado en trabajos de predicción. En algunas investigaciones como [15] y [16] se combina este modelo con redes neuronales para mejorar su precisión aunque si se logra una buena selección de los parámetros como en [17] y [18] se logran importantes resultados sin necesidad de mayores requerimientos de cómputo. [19] realiza una comparación del desenvolvimiento entre las redes neuronales artificiales y el modelo ARIMA para la predicción de la temperatura en Bangladesh y este último resultó ser el de mejores resultados.

Para aplicar un modelo ARIMA ajustado fue necesaria la transformación de la serie temporal en otra que fuera aproximadamente estacionaria. Para esto se emplearon técnicas como la diferenciación y logaritmos en dependencia de la no estacionariedad.

A pesar que R cuenta con una librería llamada *forecast* y que permite la o matemdescomposición del gráfico para su análisis visual, se aplicó el métodático que depende menos del error humano. Para este objetivo fue necesario realizar pruebas de presencia de raíz unitaria, porque en caso afirmativo esto implicaría la no estacionariedad.

Tabla 3.
Resultados de pruebas estadísticas en la serie original.

Prueba	Resultado (valor p)	Según la hipótesis nula
Dickey-Fuller Aumentada	< 0.01	estacionaria
Kwiatkowski-Phillips-Schmidt-Shin	< 0.01	no estacionaria
Phillips-Perron	< 0.01	no estacionaria

Fuente: Elaboración propia.

Las pruebas estadísticas utilizadas fueron Dickey-Fuller Aumentada, Kwiatkowski-Phillips-Schmidt-Shin, Phillips-Perron que demuestran la presencia de raíz unitaria, lo cual implica no estacionariedad. Los resultados son expuestos en la Tabla 3.

Por las pruebas estadísticas aplicadas la serie tiene características no estacionarias y se planteó la necesidad de su transformación. La no estacionariedad en media, según se observa en la Fig. 1, se puede eliminar al aplicar una diferenciación a la serie temporal.

Obtenida la serie transformada, se aplicaron nuevamente las pruebas estadísticas para conocer los valores de la nueva serie. En la Tabla 4 se muestran los resultados de los mismos.

La mayoría de los test aplicados demostraron estacionariedad excepto en el caso de Philips-Perron, cuando esto sucede se podría estar en presencia de una ruptura estructural, lo cual implica falsos reportes de no estacionariedad. Para conocer esto se aplicó la prueba de Zivot y Andrews que permitió además de conocer si una serie poseía raíz unitaria, saber si tenía ruptura estructural y en qué punto de esta existía. El resultado de esta última prueba arrojó la estacionariedad de la serie y la presencia de un cambio estructural en la misma.

Se graficó la serie transformada en la Fig. 2 y se observa cómo se eliminó la variación en media y se muestra una serie estacionaria.

Posteriormente se analizó la estacionalidad de la serie, si esta poseía una frecuencia superior a una medición por año. Si hay presencia de esta, significa que las estaciones influyen directamente en el valor de la variable. Para esta investigación, se conocía previamente la influencia, puesto que la temperatura aumenta en verano y disminuye en invierno.

Este tipo de modelos ARIMA en los que las variaciones en el tiempo influyen en la estacionalidad son conocidos como SARIMA. También pueden verse como ARIMA(p,d, q)(P,D,Q)h. En el modelo, la primera parte (p,d,q) expresa la parte no estacional del modelo (donde p es del orden autorregresivo AR, d es del orden de la integración y q es del orden de la media móvil MA). La otra parte (P,D,Q)h identifica el orden de la posible estacionalidad en la serie de tiempo (h = período). [20]

Para evaluar la estacionalidad en la serie se emplearon las siguientes pruebas: Prueba de raíz unitaria de Osborn, Chui, Smith y Birchenhall; Prueba de raíz unitaria de Hylleberg, Engle, Granger y Yoo y la Prueba de raíz unitaria de Canova y Hansen.

Tabla 4.
Resultados de pruebas estadísticas en la serie transformada.

Prueba	Resultado (valor p)	Según la hipótesis nula
Dickey-Fuller Aumentada	< 0.01	estacionaria
Kwiatkowski-Phillips-Schmidt-Shin	> 0.1	estacionaria
Phillips-Perron	< 0.01	no estacionaria

Fuente: Elaboración propia.

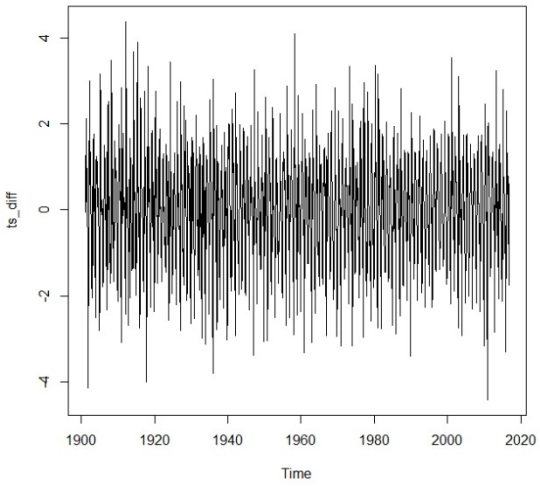

Figura 2. Gráfico de secuencia de la serie transformada.
Fuente: Elaboración propia.

El paquete de R, *forecast,* contiene una función llamada nsdiffs(). Esta función analiza las pruebas mencionadas anteriormente y devuelve un número igual a la cantidad de diferenciaciones en la parte estacional que serán necesarias para que la serie deje de ser estacional, en caso que la serie sea no estacional devuelve 0.

Cuando se aplicó esta función, la devolución del software fue 1, por tanto, fue necesaria una diferenciación en la parte estacional.

Una vez obtenida la serie aproximadamente estacionaria y no estacional similar a la original, se procedió a la construcción del modelo predictivo.

En estadística y econometría, en particular en series temporales, un modelo autorregresivo integrado de promedio móvil o ARIMA (acrónimo del inglés *autoregressive integrated moving average*) es un modelo estadístico de regresión lineal en otras palabras significa que utiliza sus propios retrasos con el fin de encontrar patrones para una predicción hacia el futuro. Las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes. El modelo ARIMA necesita identificar los coeficientes y número de regresiones que se utilizarán, se caracteriza por 3 términos p, d, q, donde p hace relación al orden del término AR, q es el orden del término MA y d es la cantidad de diferenciación requerida para que las series de tiempo sean estacionarias. Este modelo es muy sensible a la precisión con que se determinen sus coeficientes, cuando no se encuentran correlacionados y son independientes entre sí el modelo funciona de manera óptima.

En la investigación de [21] se define matemáticamente que le modelo de regresión automática (AR) es aquel donde Y(t) depende solo de sus propios retrasos.

$$Yt = \alpha + \beta 1\, Yt-1 + \beta 2\, Yt-2 + \ldots + \beta p\, Yt-p + \varepsilon 1 \quad (1)$$

Donde Yt-1 es el rezago 1 de la serie, β1 es el coeficiente del modelo y α es el término de intercepción, también estimado en el modelo. De la misma forma, el modelo de media móvil pura (MA) es aquel en el que Y(t) depende de los errores de pronóstico rezagados.

$$Yt = \alpha + \varepsilon t + \Phi 1\, \varepsilon t-1 + \Phi 2\, \varepsilon t-2 + \ldots + \Phi q\, \varepsilon t-q \quad (2)$$

Donde los términos de error son los errores de los modelos autorregresivos de los respectivos retrasos. Los errores εt y εt-1 son los errores de las ecuaciones 3 y 4:

$$Yt = \beta 1\, Yt-1 + \beta 2\, Yt-2 + \ldots + \beta 0\, Y0 + \varepsilon t \quad (3)$$

$$Yt-1 = \beta 1\, Yt-2 + \beta 2\, Yt-3 + \ldots + \beta 0\, Y0 + \varepsilon t-1 \quad (4)$$

A partir de estas ecuaciones se pudo llegar a definir el modelo ARIMA mediante las series de tiempo que se diferenciaron al menos una vez, para convertirlo en estacionario (ecuación 5):

$$Yt = \alpha + \beta 1\, Yt-1 + \beta 2\, Yt-2 + \cdots + \beta p\, Yt-p + \varepsilon t + \Phi 1\, \varepsilon t-1 + \Phi 2\, \varepsilon t-2 + \cdots + \Phi q\, \varepsilon t-q \quad (5)$$

Como la serie temporal presentó una tendencia, lo primero fue aplicar una diferenciación, de orden *d*. Una vez diferenciada la serie, se compararon los correlogramas de la función de autocorrelación (ACF) y la función de autocorrelación parcial (ACFP), proceso que arrojó orientación para la formulación del modelo orientativo.

Los procesos autorregresivos presentan función de autocorrelación parcial (ACFP) con un número finito de valores distinto de cero. Un proceso AR(p) tiene los primeros *p* términos de la función de autocorrelación parcial distintos de cero y los demás son nulos. En la práctica se considera que una muestra dada proviene de un proceso autorregresivo de orden *p* si los términos de la función de autocorrelación parcial son casi cero a partir del que ocupa el lugar *p*. Un valor se considera casi cero cuando su módulo es inferior a $\frac{2}{\sqrt{T}}$. Los programas de ordenador constituyen la franja ($-\frac{2}{\sqrt{T}}$; $\frac{2}{\sqrt{T}}$) y detectan los valores de la ACFP que caen fuera de ella.

Los procesos de medias móviles presentan función de autocorrelación con un número finita de valores distintos de cero. Un proceso MA(*q*) tiene los primeros *q* términos de la función de autocorrelación distintos de cero y los demás son nulos. Las dos propiedades descritas son muy importantes con vistas a la identificación de un proceso mediante el análisis de las funciones de autocorrelación y autocorrelación parcial, las Figs. 3 y 4 representan el gráfico de dichas funciones respectivamente

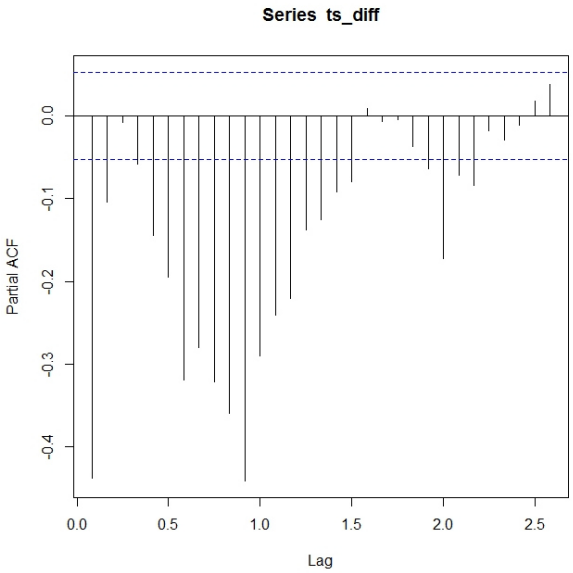

Figura 3. Gráfico de ACFP.
Fuente: Elaboración propia.

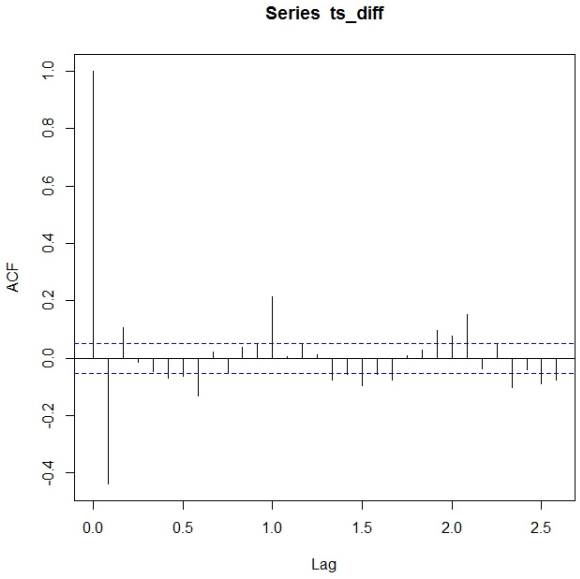

Figura 4. Gráfico de ACF.
Fuente: Elaboración propia.

Como el modelo ARIMA tuvo una parte estacional se analizaron los gráficos en los primeros retardos para la parte estacionaria y los retardos en cada periodo para el análisis de la parte estacional.

El grafico de autocorrelación parcial muestra que los dos primeros retardos son diferentes de 0 y q con el aumento de los periodos la función tiende a disminuir. A partir del aumento de los periodos pueden considerarse marcados el primer o los dos primeros rezagos.

El grafico de autocorrelación simple muestra que los dos primeros retardos son marcadamente diferentes de 0 y q a partir de esta comienzan a variar entre positivos y negativos, esto indica que se toman los dos primeros. Con el aumento de los periodos la función tiende a disminuir. A partir del aumento de los periodos pueden considerarse marcados el primer o los dos primeros rezagos.

A partir de aquí se derivaron dos variantes para el modelo ARIMA(p,d,q),(P,D,Q):

ARIMA(2,1,2)(1,1,1)
ARIMA(2,1,2)(1,1,2)
ARIMA(2,1,2)(2,1,1)
ARIMA(2,1,2)(2,1,2)

2.3 Método de suavizado exponencial

Las técnicas de suavizado exponencial son de tipo predictivo, por lo que puede utilizarse para predecir a corto plazo en las series temporales. Proporcionan previsiones razonables para horizontes de predicción inmediatos. Los resultados que se obtienen con ellas son satisfactorios, incluso cuando no se dispone de un gran número de datos históricos (aunque a mayor histórico de datos, más exactitud presentará el modelo en sus predicciones). A diferencia de los métodos de descomposición estacional, para aplicar los de suavizado no es necesario convertir la serie en una aproximadamente estacionaria, pues existen modelos para series no afectadas por tendencia ni estacionalidad, para series con tendencia y para series con tendencia y estacionalidad.

El modelo Holt-Winters reúne un conjunto de procedimientos que conforman el centro de la familia de series temporales de suavizado exponencial. A diferencia de muchas otras técnicas, este modelo fue seleccionado porque puede adaptarse fácilmente a cambios y tendencias, así como a patrones estacionales. En comparación con otras técnicas, como ARIMA, el tiempo necesario para calcular el pronóstico es considerablemente más rápido. Su aplicación en entornos de negocio es muy común, se utiliza habitualmente por muchas compañías para pronosticar la demanda a corto plazo cuando los datos de venta contienen tendencias y patrones estacionales de un modo subyacente.

Holt-Winters es un modelo estático de predicción aplicado a series de tiempo caracterizadas por estacionalidad y tendencia linear, basado en el método de medias móviles de peso exponencial (EWMA). Este modelo divide los datos analizados en tres partes, cada una es representada por una ecuación de tipo EWMA. [22]

A partir del modelo, R presenta la función HoltWinters() para la aplicación del mismo, esta cuenta con un parámetro que indica si el modelo se realizará con suavizado exponencial, y otro que indica si el modelo es estacional, por defecto son falsos y en correspondencia con las propiedades de los datos analizados, así permanecerán.

2.4 Validación de los modelos

La prueba de Ljung-Box es un tipo de prueba estadística de si un grupo cualquiera de autocorrelaciones de una serie de tiempo son diferentes de cero. En lugar de probar la aleatoriedad en cada retardo distinto, esta prueba la aleatoriedad en general basado en un número de retardos.

Tabla 5.
Resultados de la prueba estadística los modelos.

Modelos	Valor p
ARIMA(2,1,2)(1,1,1)	0.9674
ARIMA(2,1,2)(1,1,2)	0.9736
ARIMA(2,1,2)(2,1,1)	0.9864
ARIMA(2,1,2)(2,1,2)	0.9376
HoltWinters()	< 2.2e-16

Fuente: Elaboración propia.

Esta prueba también es conocida como la prueba Q de Ljung-Box, y está estrechamente relacionada con la prueba de Box-Pierce. Esta es una versión simplificada de la estadística de Ljung-Box para los cuales los estudios de simulación posteriores han demostrado un rendimiento deficiente.

H0: Los datos se distribuyen de forma independiente (es decir, las correlaciones en la población de la que se toma la muestra son 0, de modo que cualquier correlación observada en los datos es el resultado de la aleatoriedad del proceso de muestreo).

Ha: Los datos no se distribuyen de forma independiente.

Para que el modelo seleccionado sea validado tiene que tener los residuales estacionarios, normalizados e independientes. Para esto se realiza la prueba de ruido blanco o Ljung-Box. Un ruido blanco es una serie tal que su media es cero, la varianza es constante y no se puede correlacionar. Los resultados se exponen en la Tabla 5

Los resultados de este test aceptan en todos los modelos ARIMA la hipótesis nula. Esto significa que los residuales se distribuyen como un ruido blanco. Por tanto, estos presentan estacionariedad, normalidad e independencia, lo cual implica que se está en presencia de modelos adecuados para la predicción. Sucede diferente en el modelo HoltWinters por lo que desechamos el modelo.

Para seleccionar el modelo que mejor ajuste posee nos apoyaremos en AIC. El criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico en un conjunto dado de datos. El AIC proporciona un medio para la selección del modelo, valora un sacrificio entre la complejidad del modelo y la bondad de ajuste del modelo. Ofrece una estimación relativa de la información pérdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos. Por lo que el modelo que presente menor perdida de datos será el de más ajuste a los datos. [23]

En la Tabla 6 se comparan los resultados de los modelos.

Según los valores devueltos por el software por el criterio de selección aplicado el modelo ARIMA(2,1,2)(2,1,2) es el más ajustado a los datos estudiados.

Tabla 6.
Resultados de AIC en los modelos ARIMA.

Modelos	Valor AIC
ARIMA(2,1,2)(1,1,1)	3044.929
ARIMA(2,1,2)(1,1,2)	3043.881
ARIMA(2,1,2)(2,1,1)	3044.07
ARIMA(2,1,2)(2,1,2)	3042.96

Fuente: Elaboración propia.

Tabla 7.
Predicciones para el próximo año.

2017	Enero	23.71792
2017	Febrero	23.70863
2017	Marzo	24.71125
2017	Abril	25.67066
2017	Mayo	26.89284
2017	Junio	27.84484
2017	Julio	28.26223
2017	Agosto	28.33761
2017	Septiembre	27.92950
2017	Octubre	27.04149
2017	Noviembre	25.34470
2017	Diciembre	24.14931

Fuente: Elaboración propia.

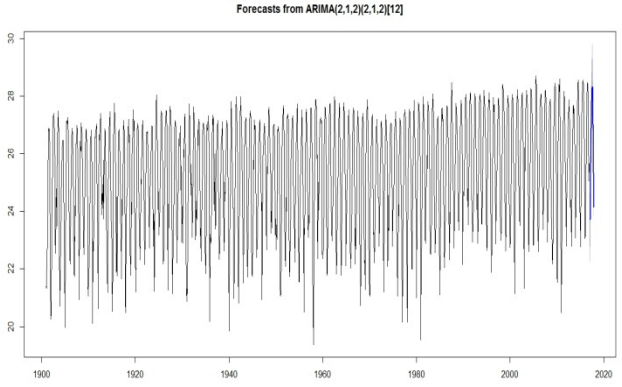

Figura 5. Predicciones para 2017.
Fuente: Elaboración propia.

3. Discusión de los resultados

Una vez obtenido el modelo más ajustado a los datos y que cumple con la prueba de Ljung-Box es posible realizar las estimaciones. La Tabla 7 muestra las estimaciones realizadas por el modelo para el próximo año.

Las predicciones obtenidas para el próximo año, que como tiene una frecuencia mensual fueron necesarias 12 predicciones, son representadas en la Fig. 5.

Para evaluar el modelo planteado para la solución será usado el indicador de Porcentaje de Error Medio Absoluto (MAPE) por su fácil interpretación [24], el cual se calcula con la ecuación 1.

$$MAPE = \frac{\sum_{t=1}^{n} \left| \frac{(y_t - \hat{y}_t)}{y_t} (100) \right|}{n} \quad (6)$$

Dónde:

$\boldsymbol{y_t}$: Es el valor observado (valor real del indicador)

$\boldsymbol{\hat{y}_t}$: Es el valor pronosticado (predicción del indicador)

$\boldsymbol{n}$: Es la cantidad de observaciones

Al aplicar el MAPE se obtuvo en el modelo aproximadamente 2.108366% de error absoluto que al comparar con el esquema de clasificación en la Tabla 8 se determina como alta precisión, por lo que es un modelo altamente confiable para la predicción de futuros indicadores.

Tabla 8.
Criterio de validación de los modelos de pronóstico.

% de Error MAPE	Clasificación del pronóstico
Menor de 10%	Alta precisión
10% -20%	Buena precisión
20% - 50%	Precisión razonable
Mayor del 50%	Poco fiable

Fuente: [24]

Tabla 9.
Rangos de error entre literaturas.

Response variable	Error metrics	Accuracy range	Unit	Paper
Temperature	MBE	(5.8,68)	%	Florita and Henze (2009)
Temperature	MAE	(0.1,5.82)	°C	Ferreira et al. (2012); Lazos et al. (2015)
Temperature	MAE	(2.26,7.69)	%	Raftery et al. (2005)
Temperature	MSE	(<0.1,0.89)	°C	Raftery et al. (2005)
Temperature	RMSE	(0.46,3.36)	°C	Chen and Athienitis (1996); Zhang and Hanby (2007)
Temperature	RMSE	(2.94,9.58)	%	Raftery et al. (2005)
Temperature	MAXAE	(2.4,14.25)	°C	Ren and Wright (2002)
Solar	MBE	(−4.7626,74)	%	Florita and Henze (2009); Akarslan and Hocaoglu (2016)
Solar	MBE	(−72,185)	W/m2	Perez et al. (2010); Verbois et al. (2018a)
Solar	MAE	(6.6,24.8)	%	Perez et al. (2010); Verbois et al. (2018a)
Solar	MAE	(8.8,207)	W/m2	Ferreira et al. (2012); Verbois et al. (2018a)
Solar	MSE	(9.81,47.46)	%	Lima et al. (2016)
Solar	RMSE	(11.7,142)	%	Aryaputera et al. (2015); Miller et al. (2018)
Solar	RMSE	(0.086,514)	W/m2	Aryaputera et al. (2015); Alzahrani et al. (2017);
Solar	RMSE	(473.38,554.23)	W/m2	Ren and Wright (2002)
Solar	MAPE	(6.36,55.47)	%	Chen et al. (2011)

Fuente: [25]

Dong (2021) realiza una revisión bibliográfica del estado actual de los modelos de predicción para la temperatura global y solar. En la Tabla 9 se muestran las métricas de los modelos consultados para una comparación con el modelo construido en esta investigación.

Se calcula el error medio absoluto (MAE) y el error cuadrático medio del modelo (RMSE) en ARIMA(2,1,2)(2,1,2) y se obtiene 0.5096998 y 0.7091234 respectivamente. Las métricas obtenidas en el modelo lo posicionan, entre los citados en la bibliografía, como el más ajustados y por tanto óptimo para la predicción de los parámetros.

4. Conclusiones

Para la implementación de algoritmos predictivos de alta precisión por los métodos utilizados es necesaria la adquisición de una extensa base de datos, ya que una estos modelos requieren un amplio histórico de datos. Los datos deben ser depurados, eliminar filas repetidas y valores nulos. Realizar un análisis minucioso para obtener las características de la serie temporal que permita estimar con mayor exactitud el modelo más asequible.

En la investigación se logró la implementación de modelos predictivos que permitirán procesar los datos y realizar inferencias futuras, donde el modelo ARIMA(2,1,2)(2,1,2) fue el de mayor exactitud en las predicciones.

Los resultados obtenidos comprobaron que los modelos implementados son fiables y que sus pronósticos tienen un ¨alto grado de precisión¨. Estos son derivados de la simulación y no de la subjetividad de los investigadores, lo cual provee de solidez y rigor en la toma de decisiones, y un mayor espectro para su uso a partir de sus propiedades estadísticas.

El hecho de que las predicciones del software sean muy cercanas a la realidad permite emitir criterios acertados para evaluar una situación en un espacio de tiempo determinado.

4.1 Trabajos futuros

Este trabajo queda en la creación de modelos de predicción a partir de una serie de datos obtenidos donde se respeta la periodicidad. En trabajos futuros se propone realizar un análisis del comportamiento de los modelos ante fallos en la frecuencia de medición de los datos para corroborar que este mantiene un alto grado de precisión en sus predicciones.

Referencias

[1] Selpa-Navarro, A.Y. y Espinosa-Chongo, D., La gestión del capital de trabajo como proceso de la gestión financiera operativa. Gestión Joven, Revista de la Agrupación Joven Iberoamericana de Contabilidad y Administración de Empresas. Asociación Española de Contabilidad y Administración de Empresas (AECA). (4), art. 2008-1029-53, 2009.

[2] Mesías, F.J., et al., Valoración de indicadores de sostenibilidad en dehesas por diferentes grupos de interés: aplicación de un estudio Delphi., XVII Jornadas sobre Producción Animal, 2017, pp. 15-17.

[3] Sousa, S.N., Estruch-Guitart, V. y García C., Uso de indicadores causa-efecto para el diagnóstico de la sostenibilidad hídrica en las

Islas Baleares (España). Boletín de la Asociación de Geógrafos Españoles, 85, art. 2833, 2020. DOI: 10.21138/bage.2833

[4] Rojas-Herrera, E.I., Comparación de un modelo híbrido obtenido de la mezcla de vectores autorregresivos y la metodología de redes neuronales artificiales ANN-VAR y un modelo econométrico de vectores autorregresivos (VAR) para la predicción del nivel de mp2.5 en Santiago de Chile, Tesis de grado, Departamento de Industrias, Universidad Tecnica Federico Santa Maria, Valparaíso, Chile, 2018, 82 P.

[5] Guerrero, J.B., Rangel, Y.U. y López, S.U. Predicción del calentamiento global mediante el desarrollo de un modelo de series de tiempo. Ambiente y Desarrollo, 21(40), pp. 125-139, 2017. DOI: 10.11144/Javeriana.ayd21-40.pcgm

[6] Mateo-Pérez, V., et al., Mejora del pretratamiento de una EDAR mediante la predicción de parámetros del agua de entrada. En: 25th International Congress on Project Management and Engineering, Alcoi, 6th–9th July, 2021.

[7] Gutiérrez, B.M.M. y Sánchez-Batistas, A., Evaluación de los impactos medioambientales en la gestión de Fábrica de Quesos Sibanicú. Revista Cubana de Finanzas y Precios, 3(4), pp. 72-88, 2019.

[8] Herrera-Restrepo, J.M., et al. Indicadores medioambientales y Objetivos de Desarrollo Sostenible (ODS) a revelar por parte de empresas del sector químico. Revista En-contexto, 9(14), pp. 151-184, 2021.

[9] Granada, F., Cooper, R. y Anholon, R., Evolución de indicadores de desempeño ambiental en Colombia: estudio de caso sector industrial Cali-Yumbo. In: 7th International Workshop. Advances in Cleaner Production – Academic Work, Barranquilla, Colombia 2018, 7 P.

[10] Aguilar-Aguilar, A.C. y Obando-Díaz. F.F., Aprendizaje automático para la predicción de calidad de agua potable. Ingeniare 28, pp. 47-62, 2020. DOI: 10.18041/1909-2458/ingeniare.28.6215

[11] Samanès, T., González-Cancelas, M.N. y Molina-Serrano, B., Integración de indicadores medioambientales y de desempreño operacional en terminales de graneles sólidos sucios del sistema portuario español. Rect@, 20(1), pp. 77-93, 2019. DOI: 10.24309/recta.2019.20.1.03

[12] Sánchez-Villena, A., Uso de programas estadísticos libres para el análisis de datos: Jamovi, Jasp y R. Revista Perspectiva, 20(1), pp. 112-114, 2019. DOI: 10.33198/rp.v20i1.00026.

[13] Liu, T., Liu, S. and Shi, L., Time series analysis using SAS enterprise guide, Springer, 2020, ISBN: 978-981-15-0320-7. DOI:10.1007/978-981-15-0321-4.

[14] Geoghegan, R., Time series analysis and its applications: with R examples, Springer, 2006. ISBN: 978-3-319-52452-8. DOI: 10.1007/978-3-319-52452-8.

[15] Xu, G., Cheng, Y., Liu, F., Ping, P., and Sun, J., A water level prediction model based on ARIMA-RNN, in: 2019 IEEE 5th International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 221-226, DOI: 10.1109/BigDataService.2019.00038.

[16] Hirata, T., Kuremoto, T., Obayashi, M., Mabu, S. and Kobayashi, K., Time series prediction using DBN and ARIMA, in: 2015 International Conference on Computer Application Technologies, 2015, pp. 24-29, Doi: 10.1109/CCATS.2015.15.

[17] Schmidt, F., Suri-Payer, F., Gulenko, A., Wallschläger, M., Acker, A. and Kao, O., Unsupervised anomaly event detection for cloud monitoring using online arima, in: 2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion), 2018, pp. 71-76, DOI: 10.1109/UCC-Companion.2018.00037.

[18] Wang, H., Huang, J., Zhou, H., Zhao, L. and Yuan, Y., An integrated variational mode decomposition and arima model to forecast air temperature, Sustainability, 11,(15), art. 4018. 2019. DOI: 10.3390/su11154018.

[19] Nury, A.H., Hasan, K. and Alam, M.J.B., Comparative study of wavelet-ARIMA and wavelet-ANN models for temperature time series data in northeastern Bangladesh, Journal of King Saud University-Science, 29(1), pp. 47-61, 2017. DOI: 10.1016/j.jksus.2015.12.002.

[20] Kumar, A.S. and Mazumdar, S., Forecasting HPC workload using ARMA models and SSA, in: 2016 International Conference on Information Technology (ICIT), pp. 294-297, 2016. DOI: 10.1109/ICIT.2016.065.

[21] Parra, J.A.P., Cruz, O.A.T. y Méndez, Y.L.A., Dispositivo basado en modelo arima para predicción de variables ambientales (temperatura, humedad, velocidad del aire) en el área agrícola del departamento del Meta, Revista GEON (Gestión, Organizaciones y Negocios), 7(2), pp. 1-12, 2020. DOI: 10.22579/23463910.193.

[22] Pena, E.H., de Assis, M.V. and Proença, M.L., Anomaly detection using forecasting methods arima and hwds, in: 2013 32nd International Conference of the Chilean Computer Science Society (SCCC), 2013, pp. 63-66, DOI: 10.1109/SCCC.2013.18.

[23] Chen, L. and Lai, X., Comparison between ARIMA and ANN models used in short-term wind speed forecasting, in: 2011 Asia-Pacific Power and Energy Engineering Conference, 2011, pp. 1-4. DOI: 10.1109/APPEEC.2011.5748446.

[24] Glen, S., Mean Absolute Percentage Error (MAPE). From StatisticsHowTo.com: Elementary Statistics for the rest of us!. [online]. s.a., Available at: https://www.statisticshowto.com/mean-absolute-percentage-error-mape/

[25] Dong, B., et al., Review of onsite temperature and solar forecasting models to enable better building design and operations, Building Simulation, 14(4), pp. 885-907, 2021. DOI: 10.1007/s12273-020-0759-2

L. Pérez, recibió sus títulos como Ing. Informático y MSc. en Administración de Empresas en la Universidad de Matanzas, Matanzas, Cuba en 2012 y 2015 respectivamente. En la actualidad es profesor auxiliar en el Departamento de Informática de la Universidad de Matanzas, Matanzas, Cuba. Sus principales intereses de investigación incluyen aplicaciones de la Inteligencia Artificial a la gestión de negocios y a la gestión ambiental.
ORCID: 0000-0001-6187-7875.

M.A. Naranjo, recibió su título como Ing. Informático en la Universidad de Matanzas, Matanzas, Cuba en 2020. En la actualidad es adiestrado en el Departamento de Informática de la Universidad de Matanzas, Matanzas, Cuba. Sus principales intereses de investigación incluyen aplicaciones de la Inteligencia Artificial a la gestión de negocios y a la gestión ambiental.
ORCID: 0000-0003-0380-8219.

O. Santos, recibió sus títulos como Ing. Civil, MSc. en Administración de Empresas, Dr. en Ciencias Técnicas Ingeniería Industrial y MSc. en Ciencias de la Educación en la Universidad de Matanzas, Matanzas, Cuba en 2016, 2018, 2020 y 2021 respectivamente. En la actualidad es profesor auxiliar en el Departamento de Ingeniería Industrial de la Universidad de Matanzas, Matanzas, Cuba y Esp. en Gestión de Ciencia, Tecnología e Innovación en la Empresa de Proyectos de Arquitectura e Ingeniería de Matanzas (EMPAI), Matanzas, Cuba. Sus principales intereses de investigación incluyen la gestión vial, gestión empresarial, planeamiento urbano y gestión postgraduada.
ORCID: 0000-0003-2420-5732.

J.A. Cabrera, recibió sus títulos como Lic. en Educación Especialidad Geografía y Dr. en Ciencias Geográficas en la Universidad Pedagógica de Matanzas, Matanzas, Cuba en 1979 y en la Universidad de La Habana, La Habana, Cuba en 1996 respectivamente. En la actualidad es profesor titular en el Departamento de Construcciones de la Universidad de Matanzas, Matanzas, Cuba. Sus principales intereses de investigación incluyen la gestión ambiental, gestión de playas, paisajes y manejo integrado de zonas costeras.
ORCID: 0000-0002-2723-3619.

D. Nogueira, recibió sus títulos como Ing. Industrial, MSc. en Administración de Empresas y Dr. en Ciencias Técnicas Ingeniería Industrial en la Universidad de Matanzas, Matanzas, Cuba en 1995, 1998 y 2002 respectivamente. En la actualidad es profesor titular en el Departamento de Ingeniería Industrial de la Universidad de Matanzas, Matanzas, Cuba. Sus principales intereses de investigación incluyen la gestión empresarial, gestión por procesos, control de gestión y gestión del conocimiento.
ORCID: 0000-0002-0198-852X.