Revista de Investigación
Hatun Yachay Wasi 2(2), 72 - 84 (2023)
ISSN: 2955 - 8255

ARTICULO ORIGINAL

Recibido: Mayo 2023
Aceptado: Agosto 2023

MODELOS DE REGRESIÓN LINEAL Y NO LINEAL EN EL AJUSTE DE CURVAS PARA DATOS DE HUMEDAD RELATIVA

LINEAR AN NO-LINEAR REGRESSION MODELS IN THE FITTING OF CURVES FOR RELATIVE HUMIDITY DATA

Jesús Espinola Gonzales[1]

[1]Universidad Tecnológica de los Andes-Abancay-Perú

Correspondencia:
Jesús Espinola Gonzales
jespinolag@utea.edu.pe

Como citar este articulo: Espinola, J. (2023). Modelos de regresión lineal y no lineal en el ajuste de curvas para datos de humedad relativa. *Hatun Yachay Wasi, 2* (2), 72–84. https://doi.org/10.57107/hyw.v2i2.48

RESUMEN

Esta investigación tuvo como objetivo obtener una curva que mejor se ajuste a un conjunto de datos; en este caso son mediciones de humedad relativa de una estación meteorológica. Se analizaron métodos de regresión lineal y no lineal, planteados como problemas de mínimos cuadrados. La curva de ajuste de tipo sinusoidal es la que mejor ajuste ofrece, aunque también se analizan situaciones en la que una función de tipo cúbica puede ser buena solución. Además, se analizó el ajuste de estos datos mediante la función logística, que da muy buenos resultados para ciertos periodos del día.

Palabras clave: ajuste de curvas, humedad relativa, modelos de regresión.

ABSTRACT

This research aimed to obtain a curve that best fits a data set; in this case they are relative humidity measurements from a weather station. Linear and non-linear regression methods were analyzed, posed as least squares problems. The sinusoidal-type fit curve is the one that offers the best fit, although situations in which a cubic-type function can be a good solution are also analyzed. In addition, the fit of these data was analyzed using the logistic function, which gives very good results for certain periods of the day.

Keywords: curve fitting, relative humidity, regression model.

INTRODUCCIÓN

El estudio del comportamiento de los diferentes parámetros meteorológicos como temperatura, humedad relativa, radiación solar, entre otros, es de vital importancia; especialmente hoy en día en que el cambio climático se está haciendo cada vez más notorio. La humedad relativa, como los otros parámetros meteorológicos, tiene efectos positivos y negativos según su porcentaje de presencia en el aire.

La humedad relativa puede tener efectos negativos en la conservación de alimentos (Carrillo & Reyes, 2013), conservación de viviendas (Zorrilla & Agulló, 2018), conservación de documentos de archivo y biblioteca (Sánchez, 1996), entre otras áreas.

En esta investigación se aborda el problema de ajuste de curvas a los datos de la humedad relativa, tomados en una estación meteorológica. Cabe mencionar que, actualmente, las medidas de la humedad relativa, también se pueden tener mediante dispositivos electrónicos de uso cotidiano.

Determinar una buena curva (función) que permita representar tales datos puede ser una buena herramienta, para estudiar el comportamiento de este parámetro meteorológico en el tiempo.

Así mismo, de forma relativamente sencilla se pueden obtener estas curvas de ajuste a datos de humedad relativa, con buenos resultados, haciendo uso de metodologías de regresión lineal y no lineal.

Se analizarán funciones de ajuste de tipo cúbica y sinusoidal; para condiciones particulares, la función logística. Por otra parte, es importante señalar que para los cálculos y elaboración de gráficas existe lo que se conoce como “software libre”, muy fácil de usar, tales como Octave, wxMaxima, entre otros.

Recolección de datos

Los datos del parámetro meteorológico humedad relativa fueron proporcionados por el Centro de Investigación Ambiental para el Desarrollo de la Universidad Nacional “Santiago Antúnez de Mayolo”. Estos datos se registraron en la estación meteorológica de Cañasbamba, Callejón de Huaylas en el departamento de Ancash; este lugar tiene un clima particular, debido a que se encuentra por encima de 2250 m s.n.m., y está ubicado entre la cordillera blanca y la cordillera negra; esta particularidad ha sido uno de los motivos, para tomar datos de esta estación meteorológica, para modelizarlos.

Exploración de los datos

Se han considerado datos de entre las 07:00 hasta las 20 h, que se podría considerar como periodo de tiempo con la mayor actividad económica de la zona. Se analizaron los datos de diferentes días, y se encontró similitud en el comportamiento de los datos registrados (Tabla 1 y Fig. 1, 2 y 3) de los días 1, 3 y 5, respectivamente.

TABLA 1

Datos Humedad relativa día 1, día 3 y día 5

Hora	X HR – día 1 (%)	X HR – día 3 (%)	X HR – día 5 (%)
0:50	89.00	86.00	89.00
1:50	88.00	86.00	90.00
2:50	88.00	88.00	90.00
3:50	89.00	90.00	89.00
4:50	91.00	90.00	86.00
5:50	91.00	90.00	85.00
6:50	92.00	87.00	85.00
7:50	92.00	86.00	84.00
8:50	88.00	85.00	68.00
9:50	78.00	70.00	56.00
10:50	66.00	51.00	44.00
11:50	49.00	42.00	37.00
12:50	41.00	31.00	34.00
13:50	37.00	26.00	32.00
14:50	32.00	25.00	32.00
15:50	33.00	31.00	34.00
16:50	36.00	32.00	39.00
17:50	39.00	39.00	46.00
18:50	45.00	53.00	51.00
19:50	57.00	66.00	59.00
20:50	65.00	72.00	67.00
21:50	70.00	81.00	75.00
22:50	75.00	78.00	78.00
23:50	83.00	77.00	80.00

FIGURA 1

Representación datos día 1 – humedad relativa (%)

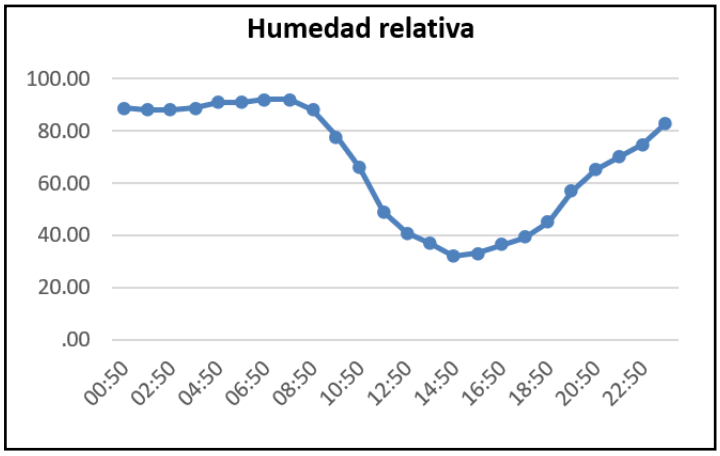


FIGURA 2

Representación datos día 3– humedad relativa (%)

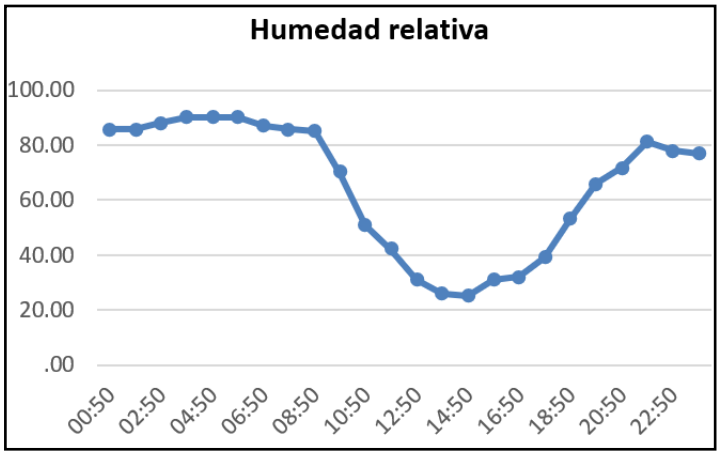
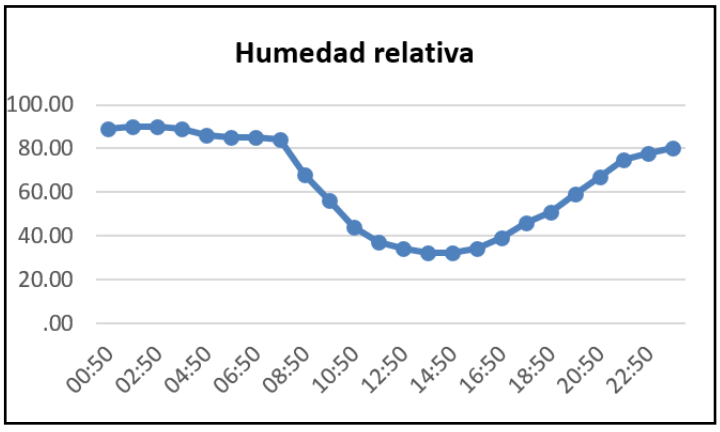


FIGURA 3

Representación datos día 5– humedad relativa (%)



Debido a la similitud se consideró trabajar directamente con los datos del día 1.

Modelados

Los resultados gráficos mostrados en las Figuras 1, 2 y 3, permiten suponer que las curvas de ajuste para estos datos pueden ser de tipo parábola cúbica o sinusoidal. Por otro lado, si se quisiera aproximar los datos entre las 07:00 h y el momento en que se alcanza su mínimo valor la curva puede ser del tipo logístico.

De manera clásica este tipo de problemática se puede abordar por técnicas de mínimos cuadrados (Aster et al., 2019). En lenguaje matemático este problema consiste en minimizar la expresión $\sum\left(y_i-\hat{y}_i\right)^2$, simbólicamente expresado como $Min\sum\left(y_i-\hat{y}_i\right)$, donde representa los valores del parámetro meteorológico humedad relativa referente a la variable independiente , que este caso representa la hora de medición; por otro lado, representa el valor estimado mediante una función de ajuste *f*: $\hat{y}=f\left(x\right)$, así $\widehat{y_i}=f\left(x_i\right)$.

La parábola cúbica como función de ajuste corresponde a un modelo de regresión lineal, mientras que las funciones de tipo sinusoidal y logística se corresponden a un modelo de regresión no lineal.

Modelo de regresión lineal

Analizamos la parábola cúbica, es decir, la determinación de los parámetros *a*, *b*, *c*, y *d* para la expresión $f\left(x\right)=a+bx+cx^2+dx^3$. El problema de mínimos cuadrados para esta función de ajuste se puede resolver mediante técnicas de álgebra lineal (Aster et al., 2019; Lay et al., 2016). La solución se calcula mediante la expresión $\bar{X}=\left(A^TA\right)^{-1}A^Tb$, donde A^T representa la matriz transpuesta de la matriz *A*. La matriz *A* está formada por filas de la forma $\left(1\ x_i\ x_i^2\ x_i^3\right)$ y el vector *b* está formado por las filas $\left(y_i\right)$.

Construcción de la función cúbica de ajuste de datos

Para este caso los x_i representan las horas en las que se ha medido la humedad relativa del día 1, y los y_i representan la humedad relativa correspondiente, por lo que las expresiones matriciales son

$$A=\begin{pmatrix}1&7&7^2&7^3\\1&8&8^2&8^3\\1&9&9^2&9^3\\1&10&10^2&10^3\\1&11&11^2&11^3\\1&12&12^2&12^3\\1&13&13^2&13^3\\1&14&14^2&14^3\\1&15&15^2&15^3\\1&16&16^2&16^3\\1&17&17^2&17^3\\1&18&18^2&18^3\\1&19&19^2&19^3\\1&20&20^2&20^3\end{pmatrix}\qquad \bar{X}=\begin{pmatrix}a\\b\\c\\d\end{pmatrix}\qquad b=\begin{pmatrix}92.0\\92.0\\88.0\\78.0\\66.0\\49.0\\41.0\\37.0\\32.0\\33.0\\36.0\\39.0\\45.0\\57.0\end{pmatrix}$$

Al hacer los cálculos correspondientes para $\bar{X} = \left(A^T A\right)^{-1} A^T \boldsymbol{b}$, se tiene como solución

$$\bar{X} = \begin{pmatrix} a = 34.454707 \\ b = \ \ 27.840814 \\ c = -3.453495 \\ d = 0.106232 \end{pmatrix}$$

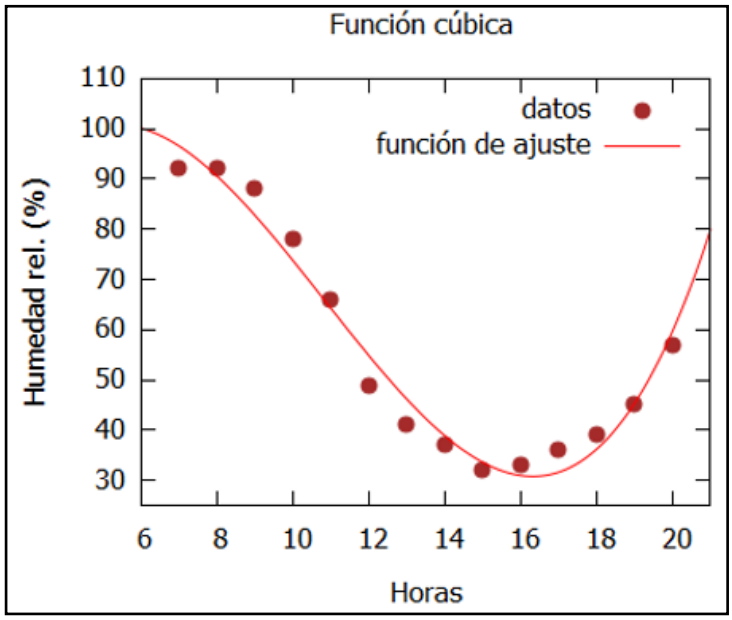
De esta manera se tiene que la función de ajuste parabólica de grado 3 que mejor ajusta los datos es

$$\hat{y} = f(x) = 34.454707 + 27.840814\,x - 3.453495\,x^2 + 0.106232\,x^3$$

En la Figura 4 se presenta, en color marrón, los datos (x_i, y_i) correspondiente a la hora de medición y el valor medido de la humedad relativa; así mismo, se presenta, en color rojo, la curva de la función cúbica de ajuste.

Figura 4

Datos de humedad relativa y función cúbica de ajuste



Modelo de regresión no lineal

Las otras funciones de ajuste consideradas en este trabajo son de tipo sinusoidal y de tipo logística; que como se verá hace que la determinación de los parámetros sean problemas de minimización de tipo no lineal (Aster et al., 2019), generalmente este tipo de problemas se pueden resolver mediante métodos de aproximación clásicos (Escudero, 1978) y/o técnicas metaheurísticas (Espinola et al., 2022; Talbi, 2009).

La función seno tiene como expresión general $\hat{y} = f(x) = a\ sin\left(\ b\left(x - dx\right)\right) + dy$, donde los parámetros a determinar son a, b, *dx*, y *dy*; los parámetros a y b son factores de escala, mientras que *dx* y *dy* representan las traslaciones en la dirección horizontal y vertical, respectivamente.

Por otro lado, la función logística en su expresión general queda expresada como

$$\hat{y} = f(x) = \mathrm{a}\frac{1 + \mathrm{m}\,\mathrm{e}^{\left(-\frac{\mathrm{x} - \mathrm{dx}}{\mathrm{T}}\right)}}{1 + \mathrm{n}\,\mathrm{e}^{\left(-\frac{\mathrm{x} - \mathrm{dx}}{\mathrm{T}}\right)}} + \mathrm{dy}$$

En este caso los parámetros a determinar son *a, m, n, T dx* y *dy*. Como en el caso anterior los parámetros *dx* y *dy* también representan las traslaciones en la dirección horizontal y vertical, respectivamente.

Es evidente que, en las dos curvas de ajuste, sinusoidal y logística, los parámetros a determinar están en expresiones no lineales.

Construcción de la función sinusoidal de ajuste de datos

se trata de un problema de minimización:

$$\text{Min}\sum_{i=1}^{14}\left(\,y_i - \hat{y}_i\,\right)^2$$

Cada y_i es la humedad relativa de cada hora x_i, por otro lado, $\hat{y}_i = a\,sin\left(b\left(x_i - dx\right)\right) + dy$. El objetivo es, con los datos que se tienen obtener los mejores parámetros *a, b, dx, dy* para minimizar la sumatoria. Para resolver este tipo de problemas existen diferentes métodos, generalmente, métodos da aproximación, partiendo de una solución inicial. En este caso se ha usado el método del gradiente reducido generalizado (Escudero, 1978). En la Tabla 2 se muestran los valores de los parámetros respectivos.

TABLA 2

Parámetros de la función sinusoidal de ajuste

Variable	Valores hallados	Valor mínimo de la función
a	32.152172	84.3436
b	0.340766	
dx	1.999861	
dy	62.589495	

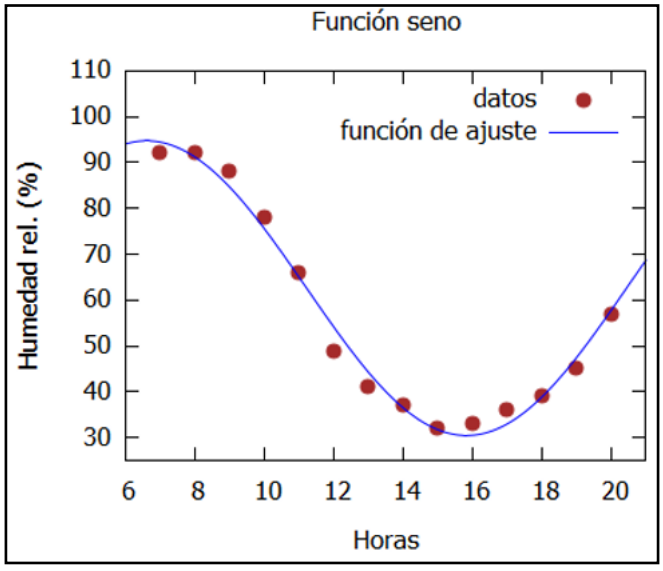
Así, la función sinusoidal de ajuste, para los datos de la humedad relativa, queda definida por:

$$\hat{y} = f\left(x\right) = 32.152172\ \ sen\left(0.340766\left(x - 1.999861\right)\right) + 62.589495$$

En la Figura 5 se muestran los datos (color marrón), mientras en que, en color azul, la curva de la función sinusoidal de ajuste.

Figura 5

Datos de humedad relativa y función cúbica de ajuste



Gráficamente se puede ver que, la función sinusoidal ofrece mejor ajuste de datos que la función cúbica, pero esto se analiza detalladamente en la sección de resultados y discusión.

Construcción de la función logística de ajuste de datos

Aquí, se obtuvo una curva que permite ajustar los datos entre las 07:00 hasta las 16:00 h, en la que, estadísticamente, la humedad relativa alcanza su mínimo valor. En algunos casos puede resultar de especial interés este periodo de tiempo.

Así mismo, es necesario resolver el problema de minimización

$$\text{Min}\sum_{i=1}^{10}\left(y_i - \hat{y}_i\right)^2$$

en el que se necesitan hallar los parámetros a, m, n, T, dx y dy de la siguiente expresión

$$\hat{y} = f(x) = a\frac{1 + m\, e^{\left(-\frac{x - dx}{T}\right)}}{1 + n\, e^{\left(-\frac{x - dx}{T}\right)}} + dy$$

Es evidente que los parámetros están en forma no lineal.

Este problema también se ha resuelto mediante el algoritmo del gradiente reducido generalizado (Escudero, 1978). En la Tabla 3 se muestra la solución del problema, es decir los parámetros que hacen que la expresión $\sum_{i=1}^{10}\left(y_i - \hat{y}_i\right)^2$ sea mínima.

TABLA 3

Parámetros de la función logística de ajuste

Variable	Valores hallados	Valor mínimo de la función
a	30.932566	10.055736
m	6.101863	
n	2.044765	
T	0.985852	
dx	10.420272	
dy	1.450884	

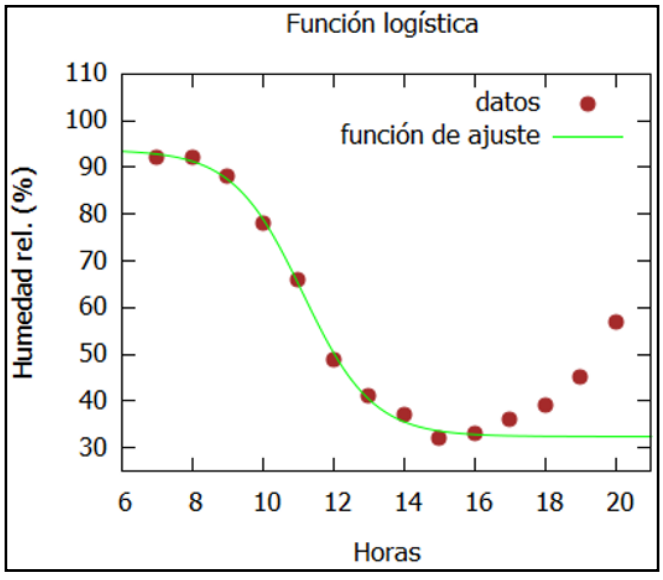
Así, la función logística que mejor ajusta los datos dados tratados es

$$\hat{y} = f(x) = 30.932566\frac{1 + 6.101863\, e^{\left(-\frac{x - 10.420272}{0.985852}\right)}}{1 + 2.044765\, e^{\left(-\frac{x - 10.420272}{0.985852}\right)}} + 1.450884$$

En la Figura 6 se observa la forma clara que la función logística, en color verde, permite ajustar muy apropiadamente los datos (color marrón), en el periodo de tiempo precisado.

Figura 6

Datos de humedad relativa y función logística de ajuste



RESULTADOS Y DISCUSIÓN

Periodo de 07:00 a 20:00 Hs.

En la Tabla 4, para cada función de ajuste, se muestra el error cuadrático medio, es decir $\frac{\Sigma(y_i-\hat{y}_i)^2}{n}$, donde "n" es el número de datos considerados, en este caso *n*=14. Con esto se puede comparar e identificar cuál es la función que se ajusta mejor a los datos (humedad relativa). El error cuadrático medio para la función sinusoidal es 6.0245, mientras que para la función cúbica es 12.6292.

Así pues, se tiene que la función sinusoidal es mejor función de ajuste para los datos de humedad relativa en el periodo analizado. Sin embargo, hay que tener especial cuidado si se trabaja con datos posteriores a las 20:00 h, pues la función seno es simétrica respecto de un eje vertical que pasa por donde la función alcanza su mínimo; entonces en ese caso será una buena función de ajuste si los valores a la derecha de esta línea vertical son simétricos respecto de los de la izquierda. Es decir, si el crecimiento de los valores de la humedad relativa es simétrico a su decrecimiento; de no ser así es recomendable considerar la función cúbica.

TABLA 4

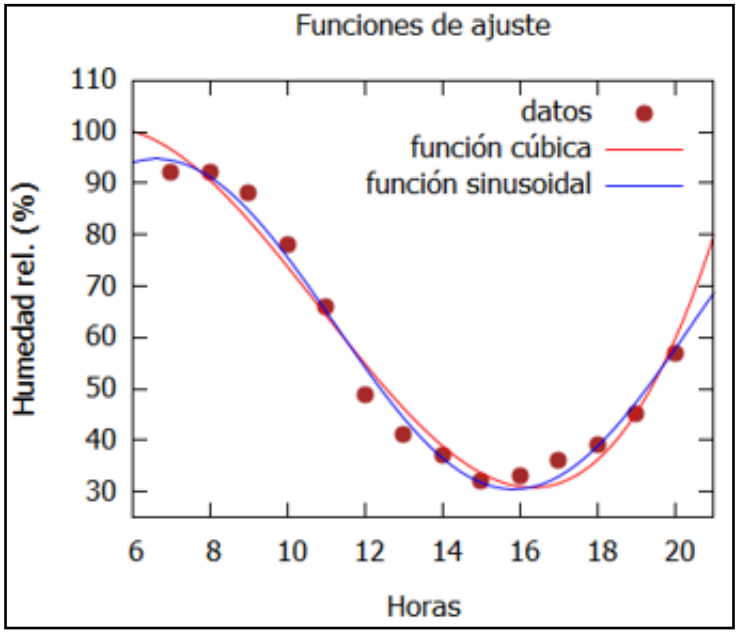
Error cuadrático medio de las funciones cúbica y sinusoidal

Datos: Hora / HR (%)		Función cúbica de ajuste		Función sinusoidal de ajuste	
x_i	**y_i**	**Aprox. ÿ_i**	**(y_i - ÿ_i)^2**	**Aprox. ÿ_i**	**(y_i - ÿ_i)^2**
7	92	96.5567	20.7638	94.4574	6.0387
8	92	90.5483	2.1074	91.1992	0.6413
9	88	82.7321	27.7511	84.6508	11.2172
10	78	73.7453	18.1021	75.5653	5.9277
11	66	64.2256	3.1486	64.9876	1.0249
12	49	54.8101	33.7572	54.1341	26.3591
13	41	46.1363	26.3820	44.2530	10.5820
14	37	38.8417	3.3918	36.4806	0.2698
15	32	33.5635	2.4447	31.7108	0.0836
16	33	30.9393	4.2466	30.4921	6.2896
17	36	31.6063	19.3045	32.9646	9.2134
18	39	36.2020	7.8288	38.8441	0.0243
19	45	45.3638	0.1323	47.4543	6.0237
20	57	59.7290	7.4474	57.8051	0.6482
		Error cuadrático medio:	**12.6292**	**Error cuadrático medio:**	**6.0245**

Lo expresado anteriormente, aquí se ve refrendado por la representación gráfica de las funciones de ajuste y los datos respectivos mostrados en la Figura 7. Los datos están representados como puntos (color marrón), la función de ajuste cúbica está en color rojo, mientras que la función de ajuste sinusoidal está en color azul.

FIGURA 7

Gráficas comparativas cúbica y sinusoidal



Periodo de 07:00 a 16:00 h

Otro de los resultados mostrado es el ajuste de los datos para el periodo de decrecimiento de los valores de la humedad relativa, esto es entre las 07:00 y las 16:00 h.

En la Tabla 5 se muestra el error cuadrático medio para cada una de las funciones de ajuste. Para la función logística el error cuadrático medio es 1.0056, para la función sinusoidal es 6.8434. mientras que para la función cúbica es 14.2095. Evidentemente, se tiene que en este caso la mejor función de ajuste es la función logística, lo que se corrobora en la Figura 8.

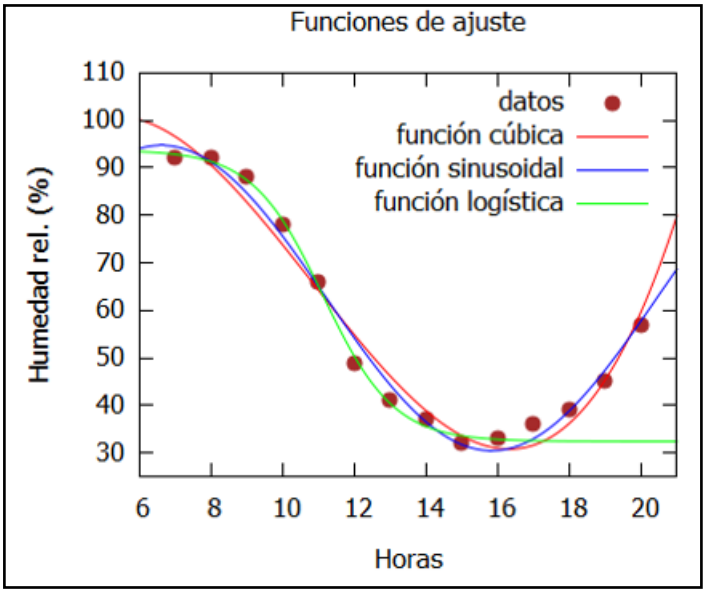
TABLA 5

Error cuadrático medio para cada una de las funciones de ajuste

Datos: Hora / HR (%)		Función cúbica		Función sinusoidal		Función logística	
x_i	y_i	Aprox. ÿ_i	(y_i - ÿ_i)^2	Aprox. ÿ_i	(y_i - ÿ_i)^2	Aprox. ÿ_i	(y_i - ÿ_i)^2
7	92	96.5567	20.7638	94.4574	6.0387	92.8374	0.7013
8	92	90.5483	2.1074	91.1992	0.6413	91.2846	0.5117
9	88	82.7321	27.7511	84.6508	11.2172	87.3887	0.3737
10	78	73.7453	18.1021	75.5653	5.9277	78.9036	0.8164
11	66	64.2256	3.1486	64.9876	1.0249	65.0203	0.9597
12	49	54.8101	33.7572	54.1341	26.3591	50.2867	1.6557
13	41	46.1363	26.3820	44.2530	10.5820	40.3586	0.4114
14	37	38.8417	3.3918	36.4806	0.2698	35.5367	2.1412
15	32	33.5635	2.4447	31.7108	0.0836	33.5656	2.4513
16	33	30.9393	4.2466	30.4921	6.2896	32.8175	0.0333
		Error cuadrático medio:	**14.2095**	**Error cuadrático medio:**	**6.8434**	**Error cuadrático medio:**	**1.0056**

FIGURA 8

Gráficas comparativas cúbica, sinusoidal y logística



CONCLUSIONES

La función sinusoidal es una buena función de ajuste para los datos de humedad relativa en el periodo del tiempo analizado.

El comportamiento simétrico, de la función seno, respecto de la recta vertical pasa por el punto donde alcanza su mínimo; pues en ciclos completos del día podría no ser la mejor opción, por lo que no hay que descartar la función cúbica como buena función de ajuste para ciertos casos.

La función logística permite ajustar, con muy buenos resultados, los datos de la humedad relativa en el periodo decreciente hasta cuando alcanza su valor mínimo.

En general, las técnicas de regresión no lineal permiten mejores resultados, los valores de los parámetros respectivos se pueden calcular con facilidad, teniendo en cuenta que hay diferentes herramientas informáticas de cálculo para ello.

AGRADECIMIENTO

Al Centro de Investigación Ambiental para el Desarrollo (CIAD) de la UNASAM, en Huaraz-Perú por facilitar la información meteorológica necesaria para el desarrollo de esta investigación.

REFERENCIAS BIBLIOGRÁFICAS

Aster, R., Borchers, B., & Thurber, C. (2019). *Parameter Estimation and Inverse Problems.* Elsevier. https://www.sciencedirect.com/book/9780128046517/parameter-estimation-and-inverse-problems

Bautista, J. (2020). *Metaheurísticas en ingeniería.* Madrid: Dextra Editorial S.L. https://www.researchgate.net/publication/346656195_Metaheuristicas_en_Ingenieria

Carrillo, M., & Reyes, A. (2013). Vida útil de los alimentos. *Revista Iberoamericana de las Ciencias Biológicas y Agropecuarias, 2*(3), 32-56. https://www.ciba.org.mx/index.php/CIBA/article/view/20

Escudero, L. (1978). Programación general no-lineal con restricciones: algoritmos aplicables (y II). Qüestiió *(Quaderns d'estadística i investigació operativa), 2*(4), 275-288. http://hdl.handle.net/2099/4598

Espinola, J., Cobo, Á., & Rocha, R. (2022). Metaheurísticas con Python. Casos Prácticos. *Hatun Yachay Wasi*, 1(2), 43-57. https://doi.org/https://doi.org/10.57107/hyw.v1i2.23

Lay, D., Lay, S., & McDonald, J. (2016). *Álgebra Lineal y sus aplicaciones.* Pearson. https://books.google.com.pe/books/about/Algebra_Lineal_Y_Sus_Aplicaciones.html?hl=es&id=lTIVrKT9CMIC&redir_esc=y

Poole, D. (2011). *Linear Algebra. A Modern Introduction* Brooks/Cole Cengage Learnig. https://edisciplinas.usp.br/pluginfile.php/5572235/mod_resource/content/1/David%20Poole%20-%20Linear%20Algebra_%20A%20Modern%20Introduction-Brooks%20Cole%20%282011%

Sánchez, A. (1996). Variables del deterioro ambiental: Humedad relativa y calor. *Beletín ANADAB*(2), 97-111. https://dialnet.unirioja.es/servlet/articulo?codigo=51009

Talbi, E. (2009). *Metaheuristics*. John Wiley & Sons Inc. http://dx.doi.org/10.1002/9780470496916

Zorrilla, V., & Agulló, M. (2018). Un estudio de caso de la presencia de humedades en viviendas de mujeres propietarias. *Técnica Industrial*, 321, 34-41. https://e-archivo.uc3m.es/bitstream/handle/10016/35129/humedades_TI_2018.pdf?sequence=1&isAllowed=y