1. ANÁLISIS DE DATOS

La herramienta de "Análisis de datos" en WAYREapp se encuentra habilitada desde la versión 3.0.

1.1. Datos

Esta herramienta se compone de dos análisis que permiten evaluar el estado de los datos importados.

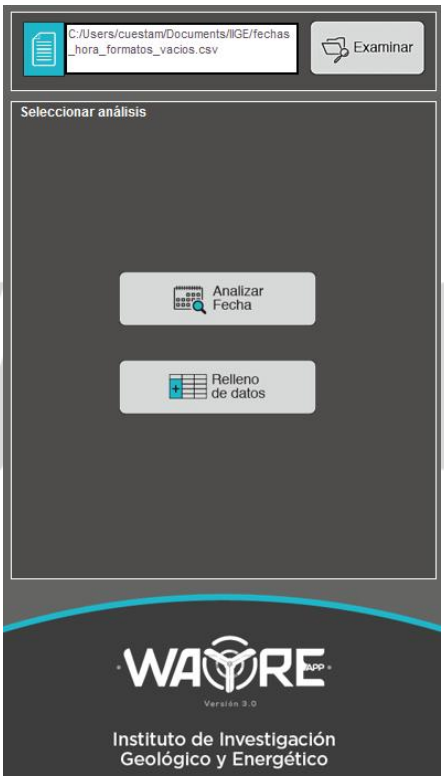

Figura 1. Opciones de la herramienta de Análisis de datos.

La herramienta "Analizar fecha" explora el estado de la fecha y la hora en la tabla y la convierte al formato de fecha usada en los análisis de WAYREapp (formato ISO "yyyy-mm-dd HH:MM:SS"). La herramienta "Relleno de datos" permite rellenar los datos de una columna que tenga datos en vacíos o en blanco.

1.2. Analizar fecha

Una fecha puede escribirse de diferentes formas según los formatos que existen para representar los elementos de la fecha, como el año, mes, día, hora, minuto y segundo. En la Tabla 1 se observa los formatos y codificación para describir los elementos de una fecha [1].

Tabla 1. Formatos y codificación para elementos de una fecha.

Código		Ejemplo	Descripción
FECHAS			
Año	%Y	2013	Año con siglo como número decimal.
	%y	13	Año sin siglo como número decimal rellenado con ceros (00 a 99).
Mes	%m	09	Mes como número decimal rellenado con ceros (01 a 12).
	%B	September, Septiembre	Nombre completo del mes según la configuración regional.
	%b	Sep, Sep	Nombre abreviado del mes según la configuración regional.
Día	%A	Sunday, Domingo	Nombre completo del día de la semana según la configuración regional.
	%a	Sun, Dom	Nombre abreviado del día de la semana según la configuración regional.
	%d	8	Día del mes como número decimal rellenado con ceros (01 a 31).
	%w	0	Día de la semana como número decimal (0=Sunday, domingo; 6=Saturday, sábado).
	%j	051	Día del año como número decimal rellenado con ceros (001 a 366).
Semana	%U	36	Número de semana del año (domingo como primer día) rellenado con ceros.
	%W	35	Número de semana del año (lunes como primer día) rellenado con ceros.

HORA			
Hora	%H	7	Hora (formato 24h) como número decimal rellenado con ceros (00 a 23).
	%I	7	Hora (formato 12h) como número decimal rellenado con ceros (01 a 12).
	%p	AM	Equivalente regional de AM o PM (en mayúsculas).
Minuto	%M	6	Minuto como número decimal rellenado con ceros (00 a 59).
Segundo	%S	5	Segundo como número decimal rellenado con ceros (00 a 59).
Microsegundo	%f	0	Microsegundos (000000–999999) como número decimal rellenado con 6 dígitos.

En la Tabla 2 se muestran los símbolos más usados para separar los elementos de la fecha

Tabla 2. Símbolos usados para separar elementos de fecha.

Separadores	Ejemplo
Slash o barra oblicua (/)	2025/01/11
Guion medio (-)	2025-01-11
Punto (.)	11.01.2025
Coma (,)	11 January, 2025
Espacio ( )	11 Jan 2025
Sin espacio ()	20250111

Al combinar los elementos de la fecha y los separadores se forman varios tipos de formatos de escritura para una fecha. En la Tabla 3 se muestran las combinaciones más comunes en las que se puede presentar la fecha, la primera parte muestra combinaciones de los elementos de la fecha usando como separador el símbolo de guion medio (-) y en la segunda parte se observa otros formatos usados que incluyen el día de la semana. Se puede reemplazar el separador usado en la primera parte de la Tabla 3 por cualquiera de las opciones antes mencionadas en la Tabla 2 y se reconocerá como un formato válido de fecha.

Tabla 3. Combinaciones de formatos de fecha.

Formato	**Formato (strftime/strptime)**	**Ejemplo real (2025-01-11)**
Separador usado (-)		
yyyy-mm-dd	%Y-%m-%d	2025-01-11
yyyy-dd-mm	%Y-%d-%m	2025-11-01
dd-mm-yyyy	%d-%m-%Y	11-01-2025
mm-dd-yyyy	%m-%d-%Y	01-11-2025
yyyy-bb-dd	%Y-%b-%d	2025-Jan-11
yyyy-dd-bb	%Y-%d-%b	2025-11-Jan
dd-bb-yyyy	%d-%b-%Y	11-Jan-2025
bb-dd-yyyy	%b-%d-%Y	Jan-11-2025
yyyy-BB-dd	%Y-%B-%d	2025-January-11
yyyy-dd-BB	%Y-%d-%B	2025-11-January
dd-BB-yyyy	%d-%B-%Y	11-January-2025
BB-dd-yyyy	%B-%d-%Y	January-11-2025
yy-mm-dd	%y-%m-%d	25-01-11
yy-dd-mm	%y-%d-%m	25-11-01
dd-mm-yy	%d-%m-%y	11-01-25
mm-dd-yy	%m-%d-%y	01-11-25
yy-bb-dd	%y-%b-%d	25-Jan-11
yy-dd-bb	%y-%d-%b	25-11-Jan
dd-bb-yy	%d-%b-%y	11-Jan-25
bb-dd-yy	%b-%d-%y	Jan-11-25
yy-BB-dd	%y-%B-%d	25-January-11
yy-dd-BB	%y-%d-%B	25-11-January
dd-BB-yy	%d-%B-%y	11-January-25
BB-dd-yy	%B-%d-%y	January-11-25
Otros formatos		

dd bb, yyyy	%d %b, %Y	11 Jan, 2025
dd BB, yyyy	%d %B, %Y	11 January, 2025
bb dd, yy	%b %d, '%y	Jan 11, 25
aa dd bb yyyy	%a %d %b %Y	Sat 11 Jan 2025
AA dd BB, yyyy	%A %d %B, %Y	Saturday 11 January, 2025
AA, dd BB yyyy	%A, %d %B %Y	Saturday, 11 January 2025
aa, dd bb yyyy	%a, %d %b %Y	Sat, 11 Jan 2025
AA dd de BB de yyyy	%A %d de %B de %Y	Saturday 11 de January de 2025

Los formatos de escritura más comunes de las horas se muestran en la Tabla 4.

Tabla 4. Formatos de escritura de la hora

Formato (strftime/strptime)	**Ejemplo real (14:30:00)**
%H:%M	14:30
%H:%M:%S	14:30:00
%I:%M %p	02:30 PM
%I:%M:%S %p	02:30:00 PM
%I:%M:%S%p	02:30:00PM
%I:%M:%S.%f %p	02:30:00.123456 PM
%H:%M:%S.%f	14:30:00.123456
%H.%M	14.30
%H.%M.%S	14.30.00
%I.%M %p	02.30 PM
%Hh%M	14h30
%Hh%Mmin	14h30min
%H horas, %M minutos	14 horas, 30 minutos
%H%M%S	143000

Una fecha completa incluye la fecha y la hora en cualquiera de los formatos de la Tabla 3 y la Tabla 4, por ejemplo: 2025-Jan-11 02:30:00 PM.

1.2.1. Fecha completa

En la Tabla 3 se detalla los formatos de las fechas y se puede observar que la posición de los elementos de la fecha puede variar. En este sentido, el software no puede reconocer

automáticamente si la fecha empieza con el día, mes o el año, por ejemplo, si se tiene la fecha "11-02-2025 02:30:00", no se puede saber si el número 11 corresponde a un mes o a un día. Por este motivo se debe seleccionar si las fechas en la columna empiezan con el año, mes o el día, tal como se muestra en la Figura 2.

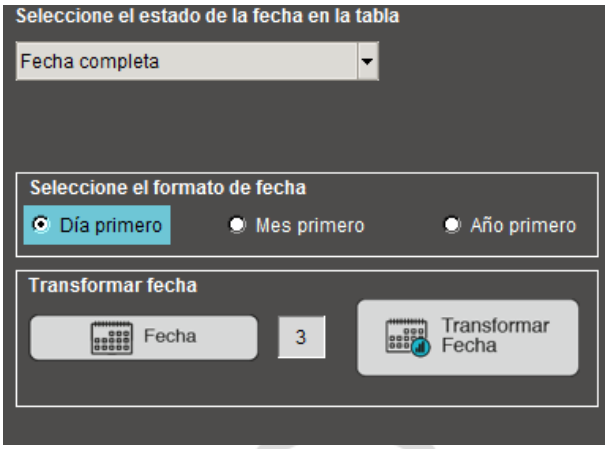

Figura 2. Parámetros para la transformación de la fecha en una sola columna.

La fecha completa debe estar en una sola columna y en cualquiera de los formatos mostrados en la Tabla 3 seguido de la hora en cualquiera de los formatos mostrados en la Tabla 4.

1.2.2. Fecha y hora separadas

Se debe seleccionar si el formato de la fecha tiene el día, mes o año primero. Después se debe seleccionar la columna de la fecha y la columna de la hora (Figura 3).

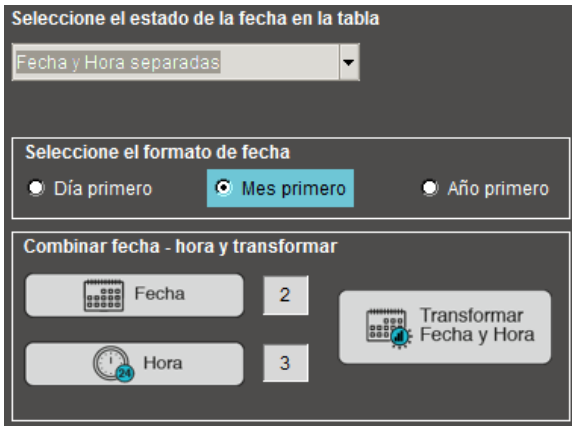

Figura 3. Parámetros para la transformación de la fecha y hora en columnas separadas.

La fecha debe estar en una columna en cualquiera de los formatos mostrados en la Tabla 3. La hora debe estar en otra columna en cualquiera de los formatos mostrados en la Tabla 4.

1.2.3. Columnas separadas

Al seleccionar la opción "Columnas separadas" se debe seleccionar las columnas donde se encuentran los elementos de la fecha (Figura 4).

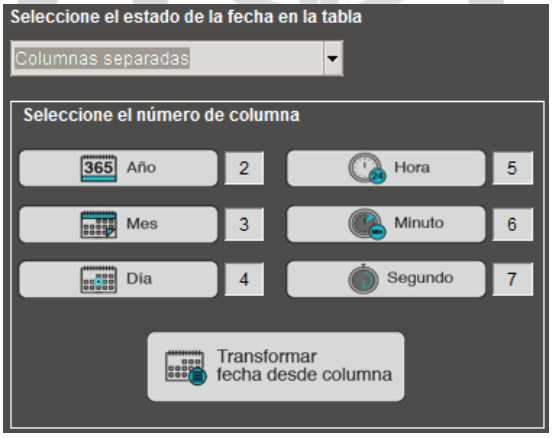

Figura 4. Parámetros para la transformación de la fecha con los elementos en columnas separadas

La fecha debe estar separada en año, mes y día, cada una en una columna diferente. De igual manera la hora debe estar separada en hora, minuto y segundo, cada una en una columna diferente.

1.3. Relleno de datos

Al ingresar a "Relleno de datos" se debe escoger el método con el que se hará el relleno de datos y se debe seleccionar la columna que contiene datos vacíos que se deseen rellenar (Figura 5). Las columnas que se seleccionen para rellenar datos deben estar en formato numérico.

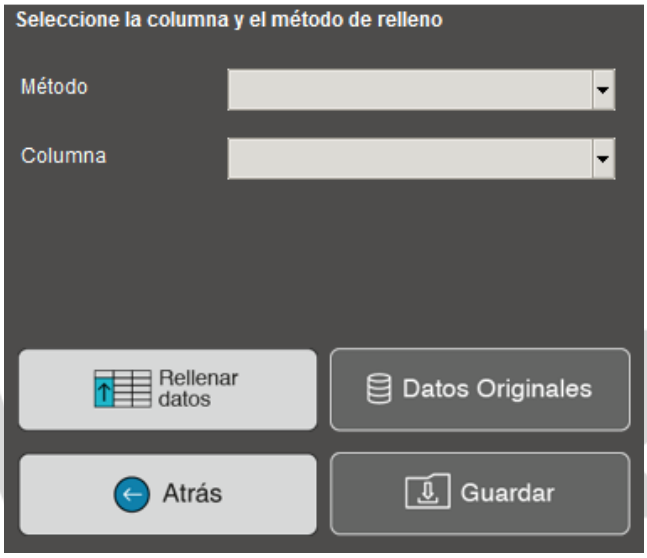

Figura 5. Parámetros de selección para el relleno de datos.

Los métodos de relleno de datos se detallan a continuación.

1.3.1. Media Aritmética

Los valores faltantes en la columna se rellenan con la media aritmética de todos los valores no vacíos de la columna.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

Donde:

- $\bar{x}$: Media aritmética.
- n : Numero de datos no vacíos de la columna.
- x_i : Cada dato en la columna.

1.3.2. Moda

La columna seleccionada de rellena con el valor más frecuente en el set de datos analizados.

1.3.3. Mediana

Es el valor que divide al grupo de datos en dos de igual tamaño. Primero los datos se ordenan de menor a mayor. Si se tienen dos valores que dividen al grupo de datos en dos iguales, se calcula su media aritmética para obtener la mediana. Los datos vacíos se rellenan con este valor.

$$Me = \begin{cases} X\left[\frac{n+1}{2}\right] & \text{si } n \text{ es par} \\ \frac{X\left[\frac{n}{2}\right] + X\left[\frac{n+1}{2}\right]}{2} & \text{si } n \text{ es } impar \end{cases} \quad (2)$$

Donde:

- Me : Mediana.
- $X[]$: El valor del set de datos en una posición específica.

1.3.4. Regresión lineal

Los valores vacíos de la columna se rellenan usando una interpolación lineal. Cada valor se calcula mediante la Ecuación (3).

$$V_{Rl} = y_1 + (y_2 - y_1) \times \frac{(x - x_1)}{(x_2 - x_1)} \quad (3)$$

Donde:

- V_{Rl} : Valor interpolado por regresión lineal.
- x_1, x_2 : Índices de los valores válidos antes y después del valor vacío.
- y_1, y_2 : Valores válidos antes y después del valor vacío.
- x : Índice del valor vacío.

1.3.5. Regresión polinomial

Los valores vacíos se rellenan ajustando los datos de la columna a una ecuación polinómica de segundo grado mostrada en la Ecuación (4), tal que pase por los puntos (x_i, y_i).

$$P(x) = a_0 + a_1 x + a_2 x^2 \tag{4}$$

$$y_i = P(x_i) \tag{5}$$

Donde:

- $P(x) = y$: Valor interpolado por regresión polinomial de segundo grado.
- a_0, a_1, a_2 : Índices de los valores válidos antes y después del valor vacío.
- x : Índice del valor vacío.

Los coeficientes a_0, a_1 y a_2 se encuentran resolviendo un sistema lineal para después poder evaluar la ecuación y encontrar los valores vacíos de la columna mediante los índices x_i con la Ecuación (5).

1.3.6. Regresión spline cubico

Teniendo un conjunto de puntos (x_n, y_n), los valores vacíos se rellenan ajustando los datos de la columna a una función por partes, construida con polinomios de tercer grado que tengan la forma de la Ecuación (6). Cada tramo o cada parte de la función se evalúa en el intervalo $[x_i, x_{i+1}]$.

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{6}$$

Donde:

- $S_i(x)$: Valor interpolado por la función spline de tercer grado.

- $a_i, b_i, c_i \; y \; d_i$: Coeficientes de la función spline cubica.
- x: Valor independiente en el intervalo $[x_i, x_{i+1}]$.

Para resolver la ecuación se debe tomar en cuenta las siguientes condiciones [2]:

- El polinomio debe pasar por los puntos y_i y y_{i+1}, tal como se muestra en la Ecuación (7) y Ecuación (8).

$$S_i(x_i) = y_i \tag{7}$$

$$S_i(x_{i+1}) = y_{i+1} \tag{8}$$

- La primera derivada (Ecuación (9)) y segunda derivada (Ecuación (10)) deben ser continuas.

$$S'_i(x_{i+1}) = S'_{i+1}(x_{i+1}) \tag{9}$$

$$S''_i(x_{i+1}) = S''_{i+1}(x_{i+1}) \tag{10}$$

- Las segundas derivadas en los extremos deben ser cero.

$$S''_0(x_0) = 0 \tag{11}$$

$$S''_{n-1}(x_n) = 0 \tag{12}$$

Con todas estas ecuaciones se puede resolver un sistema de ecuaciones lineales para encontrar los coeficientes $a_i, b_i, c_i \; y \; d_i$.

1.3.7. Valor anterior

Los valores vacíos de la columna se rellenan hacia atrás y se reemplazan con el valor no nulo más cercano por debajo.

1.3.8. Valor posterior

Los valores vacíos de la columna se rellenan hacia adelante y se reemplazan con el valor no nulo más cercano por arriba.

1.3.9. Ingresar valor

Esta opción permite ingresar un valor por teclado para que se rellene en todas las celdas vacías de la columna seleccionada (Figura 6).

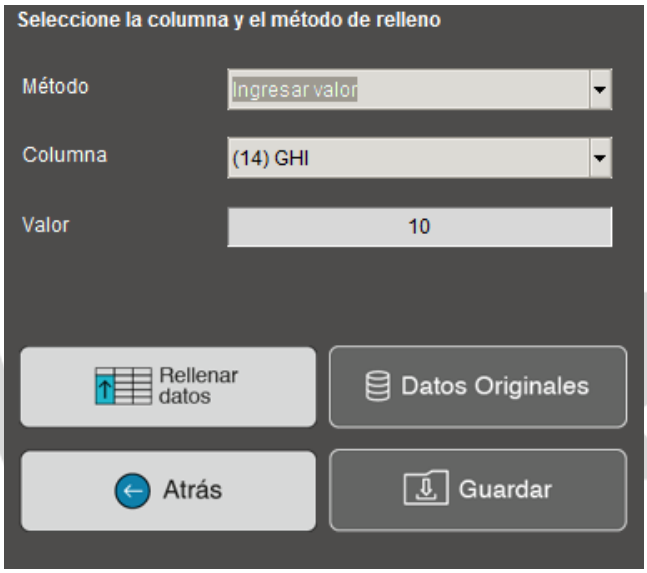

Figura 6. Ingreso del valor que se desea llenar en las celdas vacías.

1.3.10. Mejores vecinos

Al seleccionar el método de mejores vecinos se debe seleccionar la columna que se desea rellenar y la cantidad de vecinos (k) que serán tomados en cuenta para el análisis (Figura 7).

Los valores vacíos de la columna se rellenan con información de las columnas del resto de la tabla importada. Solo se tomará en cuenta las columnas que estén en formato numérico. Un valor faltante x_{ij} se calcula usando información de sus k vecinos más cercanos, donde esos vecinos sí tienen un valor observado en las celdas [3].

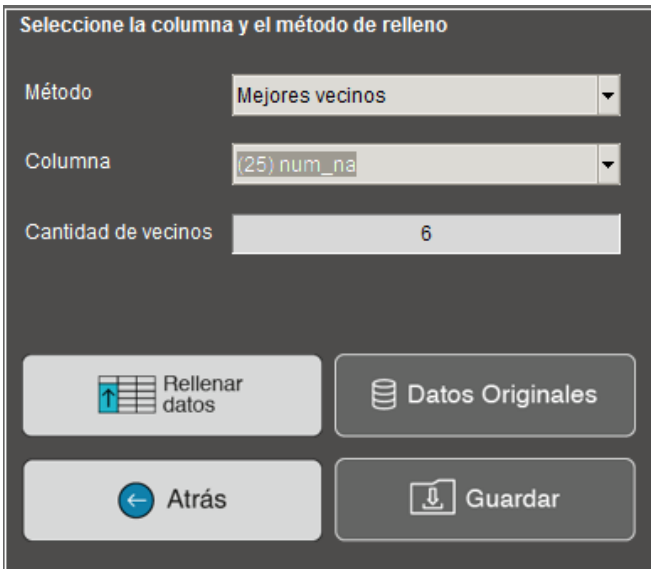

Figura 7. Ingreso de la cantidad de vecinos considerados para el análisis.

La Ecuación (13) representa a la celda en donde se encuentra el valor vacío en la fila y columna y la Ecuación (14) describe a los valores que se encuentran en la fila donde existe la celda vacía.

$$x_{ij} = NaN \tag{13}$$

$$x^{(i)} = (x_{i1},\ x_{i2}, x_{i3} \dots\ x_{id}) \tag{14}$$

Donde:

- i : Índice de la fila de la celda vacía.
- j : Índice de la fila de la columna vacía.
- d : Cantidad total de columnas.
- x_{ij} : Valor de la celda vacía en la fila i columna j.
- $x^{(i)}$: Valores de la fila i.

Para imputar un valor vacío se calcula la distancia euclídea de la fila con el valor faltante con respecto a las demás filas en la tabla, usando la Ecuación (15).

$$dist\,(i,m) = \sqrt{\sum_{l \in L_{ijm}} (x_{il} + x_{ml})^2} \quad (15)$$

Donde:

- m : Índice de otra fila (potencial vecino).
- $L_{ijm} \subseteq \{1, \ldots, d\}/\{j\}$: Conjunto de columnas sin NaN tanto en la fila i como en la fila m, y que no incluyen la columna j que queremos imputar.
- x_{il} : Valor de la celda en la fila i columna l.
- $dist\,(i,m)$: Distancia euclídea entre la fila i y la fila m.

Los k vecinos más cercanos serán las filas cuya distancia euclídea sean menores, es decir, si se usan 5 vecinos cercanos se usarán las 5 filas que tengan menor distancia euclídea [4]. Para imputar el valor faltante se usa el promedio de los valores en la columna j de las filas encontradas como mejores vecinos, utilizando la Ecuación (16).

$$\hat{x}_{ij} = \frac{1}{k} \sum_{r=1}^{k} x_{m_r,j} \quad (16)$$

Donde:

- $\hat{x}_{ij}$: Es el valor imputado que reemplaza al valor NaN en la posición (i,j).
- k : El número de mejores vecinos
- $x_{m_r,j}$: Valores en la columna j de las filas designadas como mejores vecinos.

2. BIBLIOGRAFIA

[1] P. S. Foundation, "datetime — Basic date and time types, Python 3.12.3 Documentation." Accessed: Jun. 01, 2025. [Online]. Available: https://docs.python.org/3/library/datetime.html#module-datetime

[2] J. Anderson, R. W. B. Ardill, K. J. M. Moriarty, and R. C. Beckwith, "A cubic spline interpolation of unequally spaced data points," *Comput. Phys. Commun.*, vol. 16, no. 2, pp. 199–206, 1979, doi: 10.1016/0010-4655(79)90088-2.

[3] Scikit-learn, "'3.1. Cross-validation: evaluating estimator performance,' Scikit-learn v1.4.1 Manual." Accessed: Jun. 01, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html#stratified-k-fold

[4] L. Li, J. Zhang, F. Yang, and B. Ran, "Robust and flexible strategy for missing data imputation in intelligent transportation system," *IET Intell. Transp. Syst.*, vol. 12, no. 2, pp. 151–157, 2018, doi: 10.1049/iet-its.2017.0273.

DOCUMENTO TÉCNICO: METODOLOGÍA DE "ANÁLISIS DE SERIES TEMPORALES"

Elaborado por:	Elaborado por:
Firmado electrónicamente por: JESSICA ESTEFANIA CONSTANTE ARGUELLO. Validar únicamente con FirmaEC	Firmado electrónicamente por: MANUEL ALEJANDRO CUESTA RUGEL. Validar únicamente con FirmaEC
Ing. Jessica Constante Analista Técnico de Repositorio Institucional 3	Ing. Alejandro Cuesta Analista Técnico de Repositorio Institucional 1

Elaborado por:
Firmado electrónicamente por: JUAN DIEGO JIJON VALDIVIESO. Validar únicamente con FirmaEC
Fís. Diego Jijón, MSc. Analista Técnico de Servicios Especializados 3

Revisado por:	Aprobado por:
Firmado electrónicamente por: MIGUEL ERNESTO YANEZ CEVALLOS. Validar únicamente con FirmaEC	Firmado electrónicamente por: GEOVANNA PATRICIA VILLACRESES RIVADENEIRA. Validar únicamente con FirmaEC
Ing. Ernesto Yánez Director de Gestión de la Información	Mgs. Geovanna Villacreses Subdirector Técnico