Protocolo para la Identificación y Tratamiento de Datos Inválidos Medidos en Sitio para Plantas Eólicas

Julio 11, 2023

Autores

María Alejandra Vargas Torres.

Camilo Andrés Sedano Quiroz, M.Sc.

Nelson Andrés Salazar Peña, M.Sc.

Oscar David Salamanca Gómez, M.Sc.

Andrés Leonardo González Mancera, Ph.D.

Revisión **Revisión**	**Fecha** **Fecha**	**Descripción** **Descripción**
0	2023-06-09	
1	2023-07-11	Se revisó de acuerdo a los comentarios públicos y acuerdos al interior del grupo de trabajo.

1. Objetivo

El presente protocolo define los criterios para identificar datos inválidos o atípicos de las series de medición en sitio, el número máximo admisible de datos ausentes de la serie, y el procedimiento para el llenado de datos ausentes para completar las series de mínimo un (1) año de datos de medición conforme con lo establecido en el Numeral 3 del Artículo 3 de la Resolución CREG 101 006 de 2023.

2. Ámbito de aplicación

Plantas de generación eólica, que van a participar en algún mecanismo de asignación de obligaciones del cargo por confiabilidad de que trata la Resolución CREG 071 de 2006 (o todas aquellas que la modifiquen, adicionen o sustituyan), que apliquen procedimientos relacionados con asignación de obligaciones del cargo por confiabilidad, y que tengan Obligaciones de Energía Firme (OEF) previamente asignadas a la expedición de la presente resolución.

3. Identificación de datos inválidos o atípicos

Las series de datos de mínimo un (1) año medidas en sitio en forma continua y con resolución diezminutal deben haber sido adquiridas conforme con lo dispuesto en el Acuerdo CNO 1715 de 2023, o aquel que lo modifique o sustituya, el Acuerdo CNO 1524 de 2022, o aquel que lo modifique o sustituya, según aplique, para las plantas en operación comercial.

La identificación de datos inválidos o atípicos en la serie de datos se realiza mediante la técnica de rango intercuartílico (IQR, por sus siglas en inglés). El IQR es una medida de variabilidad que permite considerar datos como atípicos a partir de la división de la serie de datos en cuartiles.

Al emplear la técnica de identificación de datos inválidos o atípicos se obtiene una serie de datos filtrada con resolución diezminutal, donde los datos de velocidad del viento (V), dirección del viento (D), temperatura ambiente (TA) y presión atmosférica (P) que no pasen el control de calidad deben ser considerados como ausentes en la serie de datos filtrada. La identificación de datos inválidos o atípicos para la temperatura ambiente solo aplica si dicha serie de datos fue medida en sitio.

Nota: Los datos ausentes son aquellos datos que no están almacenados o presentes para la serie de datos de la variable de interés.

Si cada serie de datos filtrada con resolución diezminutal presenta en total más del 10% de datos ausentes, de acuerdo a lo establecido por MEASNET en la Sección 7.2 y Sección 7.3 [5], dicha serie de datos no es aceptable.

3.1. Rango intercuartílico

El rango intercuartílico (IQR, por sus siglas en inglés) se estima mediante la Ecuación 1, donde Q_1 es el primer cuartil (i.e., correspondiente al percentil 25), Q_3 es el tercer cuartil (i.e., correspondiente al percentil 75) e IQR es el rango intercuartílico definido en la Ecuación 2. Aquellos datos de V, D, TA y P que se encuentren fuera del límite establecido son considerados como ausentes en la serie de datos filtrada.

$$Q_1 - 1.5 \cdot IQR < V, D, TA, P < Q_3 + 1.5 \cdot IQR \quad (1)$$

$$IQR = Q_3 - Q_1 \quad (2)$$

Los cuartiles Q_1 y Q_3 deben estimarse **para cada hora de cada mes** de la serie de datos de mínimo un (1) año para definir una tabla de búsqueda (ver Tabla 1), teniendo en cuenta para dicho cálculo la cantidad de datos diezminutales presentes en cada rango horario de cada mes.

Debe definirse una tabla de búsqueda para la serie de datos de velocidad del viento, dirección del viento, temperatura ambiente y presión atmosférica.

La verificación de la Ecuación 1 se realiza para cada dato de V, D, TA y P con los cuartiles e IQR correspondiente al mes y hora según la estampa temporal de la serie de datos.

Tabla 1. Tabla de búsqueda de primer (Q_1) y tercer (Q_3) cuartil para cada hora de cada mes. Debe definirse una tabla de búsqueda para la serie de datos de velocidad del viento, dirección del viento, temperatura ambiente y presión atmosférica.

Mes	Hora	Q_1	Q_3
1	0		
1	1		
1	2		
1	⋮		
1	23		
2	0		

2	1		
2	2		
2	⋮		
2	23		
⋮	⋮		
12	0		
12	1		
12	2		
12	⋮		
12	23		

4. Procedimiento para el llenado de datos ausentes

La metodología para el llenado de datos ausentes para completar mínimo un (1) año de datos de medición se desarrolla para la serie de datos filtrada con resolución diezminutal. Se asume que los datos de diferentes días, pero de una misma franja de tiempo, se comportan bajo una distribución normal.

Los parámetros de la distribución normal se calculan a partir de los datos correspondientes a la misma franja de tiempo de los datos ausentes en periodos alrededor del periodo faltante, donde:

1. El término *periodo* hace referencia a uno o más días. Así, por periodo faltante se entiende el día, o los días, para los cuales no se contó con datos en una o varias franjas de tiempo.
2. La *franja de tiempo* hace referencia a una franja diezminutal.

La metodología se basa en seleccionar un periodo anterior y un periodo posterior al periodo con datos ausentes (i.e., periodo faltante) con la misma duración temporal del periodo faltante. Es decir, se seleccionan los días (tantos días como el número de días del periodo faltante) inmediatamente anteriores y posteriores a los días con franjas de tiempo con datos ausentes.

La Figura 1 presenta un esquema de la metodología para el llenado de los datos ausentes. El recuadro blanco intermedio corresponde al periodo de datos ausentes (i.e., periodo faltante). Los recuadros azul y naranja corresponden al periodo anterior y posterior, respectivamente, con duración igual al periodo faltante. La región sombreada de la franja de tiempo indica los datos utilizados para el cálculo de los parámetros de la distribución normal. Los datos se deben tomar en la misma franja de tiempo (i.e., franja diezminutal) de los datos ausentes.

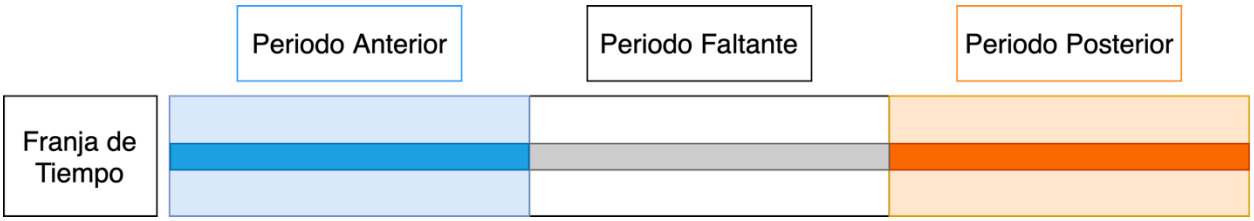

Figura 1. Esquemático para la selección de periodos y franjas diezminutales para el llenado de los datos ausentes.

El procedimiento para el llenado de los datos ausentes de las series de datos filtradas de V, D, TA y P consiste en los siguientes cuatro pasos [1]:

1. Identificar los periodos con datos ausentes.
2. Identificar los datos de los periodos anterior y posterior correspondientes a la misma franja de tiempo (i.e., franja diezminutal) de los datos ausentes. Los periodos anterior y posterior deben tener el mismo número de días del periodo faltante.
    a. Si en alguno de los periodos anterior o posterior no se alcanza a tener el mismo número de días del período faltante, se debe extender el periodo faltante (puede incluir datos válidos) hasta que los periodos anteriores y posteriores tengan el mismo número de días del periodo faltante. A partir de dichos periodos se calcularán los parámetros de la distribución normal (Paso 3.b.i), es de notar que únicamente se llenarán los datos ausentes (Paso 3.b.ii) mientras que los datos válidos se mantendrán inmutables.
3. Realizar el llenado de datos ausentes.
    a. En caso de haber datos ausentes en un único día, el llenado se realiza a partir del promedio simple entre el dato de la franja de tiempo del día anterior y el dato de la franja de tiempo del día siguiente.
    b. En caso de haber datos ausentes en varios días consecutivos para una misma franja de tiempo:
        i. Calcular los parámetros de la distribución normal, es decir, la media (μ) y la desviación estándar muestral (σ) a partir de los datos de los periodos anterior y posterior.
        ii. Seleccionar de manera aleatoria los datos con los que se completa los datos ausentes a partir de una distribución normal con media y desviación estándar muestral igual a la calculada en el paso anterior.

Si la distribución de los datos faltantes no permite la implementación de los métodos aquí propuestos, el agente puede presentar una propuesta técnica con un método alternativo para revisión y aprobación por parte del SURER.

Referencias

1. W. Menke and J. Menke, “Filling in missing data,” in Environmental Data Analysis with Matlab, Elsevier, 2016, pp. 223–237. doi: 10.1016/B978-0-12-804488-9.00010-0.
2. Y. Rochtus, “Filling gaps in time series in urban hydrology”, pp. 1-11, 2014.
3. N. V. D. Papale, “Filling the gaps in meteorological continuous data measured at FLUXNET sites with ERA Interim reanalysis”, pp. 157-171, 2015.
4. F. A. P. &. M. S.Kandasamy, “A comparison of methods for smoothing and gap filling time series of remote sensing observations- applications to MODIS LAI products”, pp. 4055-4071, 2013.
5. MEASNET, *Evaluation of Site-Specific Wind Conditions - V3,* 2022.

Anexo. Casos de ejemplo de llenado de datos ausentes

A continuación se presentan tres casos de ejemplo para ilustrar el procedimiento de llenado de datos ausentes. Aunque dichos casos de ejemplo se realizan para un número específico de días con la velocidad del viento (V), el procedimiento es independiente de la estampa temporal y, adicionalmente, es el mismo para la dirección del viento, temperatura ambiente y presión atmosférica.

Caso 1. Datos ausentes de un único día

La Tabla A.1 presenta la serie de datos en resolución diezminutal de velocidad del viento (V) para siete días. La columna color gris claro identifica el día con datos ausentes y las celdas color gris oscuro identifican las estampas de tiempo de dichos datos ausentes (i.e., el periodo faltante), las cuales abarcan desde las 10:10 hasta las 10:50h del día 3 (Paso 1 de Sección 4).

Las columnas color azul claro y naranja claro identifican los periodos anterior y posterior, respectivamente, con la misma cantidad de días del periodo faltante. Adicionalmente, las celdas color azul oscuro y naranja oscuro identifican la misma franja de tiempo del periodo faltante (Paso 2 de Sección 4).

Tabla A.1. Serie de datos en resolución diezminutal de velocidad del viento para siete días con datos ausentes en un único día (i.e., de 10:10 hasta 10:50h del día 3).

Hora	Minuto	Día						
		1	2	3	4	5	6	7
0	0	11.4	10.7	12.9	11.6	12.6	8.3	12.6
0	10	10.8	12.5	12.6	13.1	12.3	8.7	11.7
0	20	11.6	12.7	12.9	12.5	12.5	9.1	11.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	50	14.3	15.7	14.5	16.5	14.2	15.8	14.1
10	0	14.6	14.9	16.5	14.5	15.0	14.6	14.8
10	10	15.8	15.9		15.2	13.8	15.1	13.9
10	20	13.9	16.8		15.4	15.4	14.3	13.4
10	30	14.9	15.6		16.6	12.8	14.7	12.0
10	40	13.9	16.1		17.6	13.7	13.4	13.8
10	50	15.3	17.0		16.6	12.5	13.1	14.1
11	0	17.0	17.2	13.6	15.7	13.4	14.2	14.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	30	11.2	12.6	10.1	14.0	12.5	12.7	10.4
23	40	10.8	13.1	9.7	13.7	13.7	11.5	10.3

23	50	12.1	12.0	11.4	14.3	13.2	11.0	11.0

Debido a que los datos ausentes existen en un único día, el llenado se realiza a partir del promedio simple entre el dato de la franja de tiempo del día anterior y el dato de la franja de tiempo del día siguiente (Paso 3.a de Sección 4). Luego:

$$V_{10:10} = \frac{15.9 + 15.2}{2} = 15.6 \quad \text{(A.1)}$$

$$V_{10:20} = \frac{16.8 + 15.4}{2} = 16.1 \quad \text{(A.2)}$$

$$V_{10:30} = \frac{15.6 + 16.6}{2} = 16.1 \quad \text{(A.3)}$$

$$V_{10:40} = \frac{16.1 + 17.6}{2} = 16.8 \quad \text{(A.4)}$$

$$V_{10:50} = \frac{17.0 + 16.6}{2} = 16.8 \quad \text{(A.5)}$$

Caso 2. Datos ausentes en varios días consecutivos

La Tabla A.2 presenta la serie de datos en resolución diezminutal de velocidad del viento (V) para siete días. Las columnas color gris claro identifican los días consecutivos con datos ausentes y las celdas color gris oscuro identifican las estampas de tiempo de dichos datos ausentes (i.e., el periodo faltante), las cuales abarcan desde las 10:10 hasta las 10:50h de los días 3 y 4 (Paso 1 de Sección 4).

Las columnas color azul claro y naranja claro identifican los periodos anterior y posterior, respectivamente, con la misma cantidad de días del periodo faltante. Adicionalmente, las celdas color azul oscuro y naranja oscuro identifican la misma franja de tiempo del periodo faltante (Paso 2 de Sección 4).

Tabla A.2. Serie de datos en resolución diezminutal de velocidad del viento para siete días con datos ausentes en varios días consecutivos (i.e., de 10:10 hasta 10:50h de los días 3 y 4).

Hora	**Minuto**	**Día**						
		1	2	3	4	5	6	7
0	0	11.4	10.7	12.9	11.6	12.6	8.35	12.6
0	10	10.8	12.5	12.6	13.1	12.3	8.74	11.7

0	20	11.6	12.7	12.9	12.5	12.5	9.14	11.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
9	50	14.3	15.7	14.5	16.5	14.2	15.83	14.1
10	0	14.6	14.9	16.5	14.5	15.0	14.69	14.8
10	10	15.8	15.9			13.8	15.17	13.9
10	20	13.9	16.8			15.4	14.36	13.4
10	30	14.9	15.6			12.8	14.75	12.0
10	40	13.9	16.1			13.7	13.45	13.8
10	50	15.3	17.0			12.5	13.1	14.1
11	0	17.0	17.2	13.6	15.7	13.4	14.26	14.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
2	30	11.2	12.6	10.1	14.0	12.5	12.7	10.4
23	40	10.8	13.1	9.7	13.7	13.7	11.51	10.3
23	50	12.1	12.0	11.4	14.3	13.2	11.06	11.0

Debido a que los datos ausentes existen en varios días consecutivos, el llenado se realiza para cada franja de tiempo a partir de valores aleatorios obtenidos de una distribución normal con media (μ) y desviación estándar muestral (σ) estimada para para cada franja de tiempo a partir de los datos de los periodos anterior y posterior (Paso 3.b de Sección 4).

Por lo tanto, primero se estima μ y σ para cada franja de tiempo del periodo faltante:

$$\mu_{10:10} = \frac{15.8 + 15.9 + 13.8 + 15.1}{4} = 15.2 \qquad \sigma_{10:10} = \text{std}(15.8; 15.9; 13.8; 15.1) = 1.0 \qquad \text{(A.6)}$$

$$\mu_{10:20} = \frac{13.9 + 16.8 + 15.4 + 14.3}{4} = 15.1 \qquad \sigma_{10:20} = \text{std}(13.9; 16.8; 15.4; 14.3) = 1.3 \qquad \text{(A.7)}$$

$$\mu_{10:30} = \frac{14.9 + 15.6 + 12.8 + 14.7}{4} = 14.5 \qquad \sigma_{10:30} = \text{std}(14.9; 15.6; 12.8; 14.7) = 1.2 \qquad \text{(A.8)}$$

$$\mu_{10:40} = \frac{13.9 + 16.1 + 13.7 + 13.4}{4} = 14.3 \qquad \sigma_{10:40} = \text{std}(13.9; 16.1; 13.7; 13.4) = 1.2 \qquad \text{(A.9)}$$

$$\mu_{10:50} = \frac{15.3 + 17.0 + 12.5 + 13.1}{4} = 14.5 \qquad \sigma_{10:50} = \text{std}(15.3; 17.0; 12.5; 13.1) = 2.1 \qquad \text{(A.10)}$$

Luego, para cada día del periodo faltante se obtiene un valor aleatorio a partir de una distribución normal con los correspondientes μ y σ según la franja de tiempo:

$$V_{3,10:10}, V_{4,10:10} = N\sim(\mu_{10:10}, \sigma_{10:10}) \quad \text{(A.11)}$$

$$V_{3,10:20}, V_{4,10:20} = N\sim(\mu_{10:20}, \sigma_{10:20}) \quad \text{(A.12)}$$

$$V_{3,10:30}, V_{4,10:30} = N\sim(\mu_{10:30}, \sigma_{10:30}) \quad \text{(A.13)}$$

$$V_{3,10:40}, V_{4,10:40} = N\sim(\mu_{10:40}, \sigma_{10:40}) \quad \text{(A.14)}$$

$$V_{3,10:50}, V_{4,10:50} = N\sim(\mu_{10:50}, \sigma_{10:50}) \quad \text{(A.15)}$$

Caso 3. Periodo anterior o posterior sin misma cantidad de días

La Tabla A.3 presenta la serie de datos en resolución diezminutal de velocidad del viento (V) para doce días.

En orden cronológico, inicialmente se identifica que el periodo faltante abarca las estampas de tiempo desde las 10:10 hasta las 10:30 de los días 5 y 6. Consecuentemente, se identifica que el periodo anterior con la misma cantidad de días del periodo faltante abarca los días 3 y 4 en la misma franja de tiempo. No obstante, para el periodo posterior, el día 8 tiene datos ausentes en la misma franja de tiempo del periodo faltante. Por lo tanto, el periodo posterior no logra cumplir con la misma cantidad de días del periodo faltante.

Se sigue entonces el procedimiento indicado en el Paso 2.a: extender el periodo faltante hasta que los periodos anteriores y posteriores tengan el mismo número de días del periodo faltante. Así, el periodo faltante abarca la franja de tiempo desde las 10:10 hasta las 10:30 (celdas color gris oscuro) de los días 5, 6, 7 y 8 (celdas color gris claro). Las columnas color azul claro y naranja claro identifican los periodos anterior y posterior, respectivamente, con la misma cantidad de días del periodo faltante. Finalmente, las celdas color azul oscuro y naranja oscuro identifican la misma franja de tiempo del periodo faltante (Paso 2.a de Sección 4).

Tabla A.3. Serie de datos en resolución diezminutal de velocidad del viento para doce días mes con periodos anterior o posterior sin misma cantidad de días inicialmente.

Hora	Minuto	Día											
		1	2	3	4	5	6	7	8	9	10	11	12
0	0	11.4	10.7	12.9	11.6	12.6	8.35	12.6	8.4	9.5	9.4	10.4	9.8
0	10	10.8	12.5	12.6	13.1	12.3	8.74	11.7	7.6	9.6	10.2	10.7	10.7
0	20	11.6	12.7	12.9	12.5	12.5	9.14	11.8	7.6	9.3	10.0	10.4	10.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	50	14.3	15.7	14.5	16.5	14.2	15.83	14.1	10.4	11.9	10.0	12.8	16.8
10	0	14.6	14.9	16.5	14.5	15.0	14.69	14.8	10.8	12.1	10.5	14.0	17.5
10	10	15.8	15.9	15.3	15.2			13.9		13.3	11.2	15.6	16.9
10	20	13.9	16.8	15.0	15.4			13.4		13.1	10.0	13.9	16.7
10	30	14.9	15.6	15.0	16.6			12.0	11.2	13.1	11.1	13.6	15.7
10	40	13.9	16.1	15.8	17.6	13.7	13.45	13.8	11.6	12.9	10.7	13.0	16.8
10	50	15.3	17.0	15.6	16.6	12.5	13.1	14.1	12.7	12.3	11.2	13.8	16.7
11	0	17.0	17.2	13.6	15.7	13.4	14.26	14.2	11.2	12.0	11.8	13.6	15.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	30	11.2	12.6	10.1	14.0	12.5	12.7	10.4	9.9	9.6	10.0	10.4	11.7
23	40	10.8	13.1	9.7	13.7	13.7	11.51	10.3	10.0	10.2	10.9	11.6	11.5
23	50	12.1	12.0	11.4	14.3	13.2	11.06	11.0	9.3	10.5	10.9	10.7	12.6

Debido a que el periodo faltante extendido abarca más de un día, el llenado se realiza para cada franja de tiempo a partir de valores aleatorios obtenidos de una distribución normal con media (μ) y desviación estándar muestral (σ) estimada para para cada franja de tiempo a partir de los datos de los periodos anterior y posterior (Paso 3.b de Sección 4).

Por lo tanto, el procedimiento de estimación de los parámetros μ y σ es el mismo al indicado en la Ecuación A.6 a la Ecuación A.10 y, asimismo, para cada día del periodo faltante se obtiene un valor aleatorio a partir de una distribución normal con los correspondientes μ y σ según la franja de tiempo.

En este caso, el llenado de datos se realiza para las estampas de tiempo 10:10, 10:20 y 10:30 de los días 5 y 6, y para las estampas de tiempo 10:10 y 10:20 para el día 8. Los datos válidos del periodo faltante extendido se mantienen inmutables.