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

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 ausentes de las series de medición en sitio, el número máximo admisible de datos ausentes de la serie, y la metodología para el llenado de datos ausentes para completar mínimo un (1) año de datos de medición requerido conforme con lo establecido en el Numeral 3 del Artículo 3 de la Resolución CREG 101 007 de 2023.

Las series de datos de mínimo un (1) año medidas en sitio en forma continua y con resolución horaria deben haber sido adquiridas conforme con lo dispuesto en el Acuerdo CNO 1725 de 2023 o aquel que lo modifique o sustituya.

2. Ámbito de aplicación

Plantas de generación solar fotovoltaica 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 aquellas que la modifiquen, adicionen o sustituyan), a plantas solares fotovoltaicas que apliquen procedimientos relacionados con asignación de obligaciones del cargo por confiabilidad, y a plantas solares fotovoltaicas que tengan Obligaciones de Energía Firme (OEF) previamente asignadas a la expedición de la presente resolución.

3. Identificación de datos inválidos o atípicos

La identificación de datos inválidos o atípicos en la serie de datos se realiza mediante las siguientes técnicas:

1. Procedimiento de control de calidad de redes de referencia de radiación de superficie (BSRN, por sus siglas en inglés) propuesta por Long y Dutton [1].
2. 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.

Para la serie de datos de irradiancia global horizontal (GHI, por sus siglas en inglés) se debe realizar una doble verificación a partir de las técnicas anteriormente mencionadas, mientras que para la serie de datos de temperatura ambiente (TA) la verificación se realiza únicamente con la técnica del rango intercuartílico.
Al emplear las dos técnicas de identificación de datos ausentes se obtiene una serie de datos filtrada con resolución horaria, donde:

- Los datos de GHI que **no pasen los dos controles de calidad** deben ser considerados como ausentes en la serie de datos filtrada.
- Los datos de GHI que **pasen solo uno de los dos controles** de calidad deben ser considerados como atípicos válidos en la serie de datos filtrada.
- Los datos de TA que **no pasen el control de calidad** deben ser considerados como ausentes en la serie de datos filtrada.

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 horaria presenta en total más del 10% de datos ausentes, dicha serie de datos no es aceptable.

3.1. Corrección según el ángulo cenital del sol

El ángulo cenital del sol (Z) corresponde al ángulo (en unidades de grados) entre la dirección del sol y el horizonte ideal (ver Figura 1). El valor de Z puede estimarse mediante el modelo SPA propuesto por Reda y Andreas (2004) disponible en [10], o a través de las herramientas NOAA Solar Geometry Calculator, PVLIB Solar Position o NOAA Solar Position Calculator.

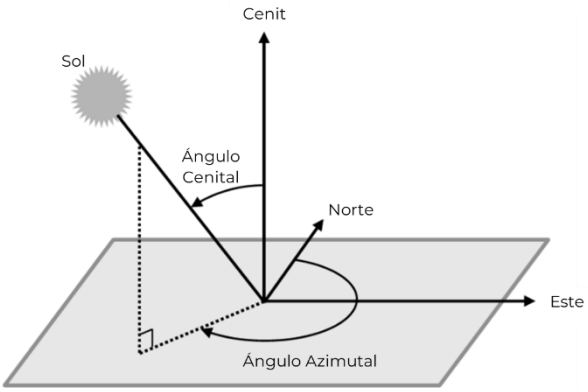

Figura 1. Ángulos que describen la posición del sol. Adaptado de [11].

Para aquellos valores donde $Z > 90°$:

- Los datos nocturnos (i.e., aquellos valores donde $Z > 90°$) pueden contener información valiosa para el control de calidad, e.g., desviaciones (*offset*) de instrumentación. Sin embargo, los datos procesados de irradiancia, energía generada por la planta fotovoltaica y otras cantidades que se espera que sean cero por la noche, deben establecerse en cero durante la noche (i.e., aquellos

valores donde $Z > 90°$) después de realizar los controles de calidad, para evitar valores extraños. Dichos datos nocturnos no deben ser procesados a nivel de programación del datalogger.

- Si existen datos ausentes de GHI, estos deben ser llenados con un valor de cero. Estos no deben ser tenidos en consideración para el límite establecido de 10% de datos ausentes.

Esta corrección según el ángulo cenital del sol debe realizarse sobre las series de datos de medición en sitio y las series de datos filtradas; es decir, antes y después de emplear las dos técnicas de identificación de datos ausentes.

3.2. Procedimiento de control de calidad de BSRN

El procedimiento consta de un nivel de control de calidad para asegurar que los datos están dentro de los límites establecidos y así garantizar la fiabilidad de los mimos. Dicho nivel de control de calidad verifica la congruencia con límites físicos mediante la Ecuación 1, donde I_{ext} es la irradiancia extraterrestre en unidades de W/m^2 y Z es el ángulo cenital del sol en unidades de grados. Aquellos datos que se encuentren fuera del límite establecido no pasan el control de calidad.

$$-4 \leq GHI \leq I_{ext} \cdot 1.5 \cdot (\cos(Z))^{1.2} + 100 \quad (1)$$

Para aquellos valores donde $Z > 90°$ se debe establecer que $\cos(Z) = 0$.

La irradiancia extraterrestre (I_{ext}) es la cantidad teórica de irradiancia solar (en unidades de W/m^2) que estaría disponible en la superficie de la tierra perpendicular al sol y fuera de la atmósfera. El valor de I_{ext} puede estimarse a través de la herramienta PVLIB Extraterrestrial Radiation o mediante la Ecuación 2, donde DOY es el número del día en el año.

$$I_{ext} = 1361 \cdot \left(1 + 0.033 \cdot \cos\left(\frac{2\pi \cdot DOY}{365}\right)\right) \quad (2)$$

3.3. Rango intercuartílico

El rango intercuartílico (IQR, por sus siglas en inglés) se estima mediante la Ecuación 3, 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 4. Aquellos datos que se encuentren fuera del límite establecido no pasan el control de calidad.

$$Q_1 - 1.5 \cdot IQR < GHI, TA < Q_3 + 1.5 \cdot IQR \quad (3)$$

$$IQR = Q_3 - Q_1 \quad (4)$$

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). **Debe definirse una tabla de búsqueda tanto para la serie de datos de GHI como para la serie de datos de TA.** De esta manera se establece el comportamiento histórico y la variabilidad del recurso para un horario mensual específico.

La verificación de la Ecuación 3 se realiza para cada dato de GHI y TA 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 tanto para la serie de datos de irradiancia global horizontal como para la serie de datos de temperatura ambiente.

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 horaria. Se asume que los datos de diferentes días, pero de una misma franja horaria, se comportan bajo una distribución normal.

Los parámetros de la distribución normal se calculan a partir de los datos correspondientes a la misma franja horaria de los datos ausentes en periodos de tiempo 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 horaria.

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 horarias con datos ausentes.

La Figura 2 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 horaria indica los datos utilizados para el cálculo de los parámetros de la distribución normal. Los datos se deben tomar en la misma franja horaria de los datos ausentes.

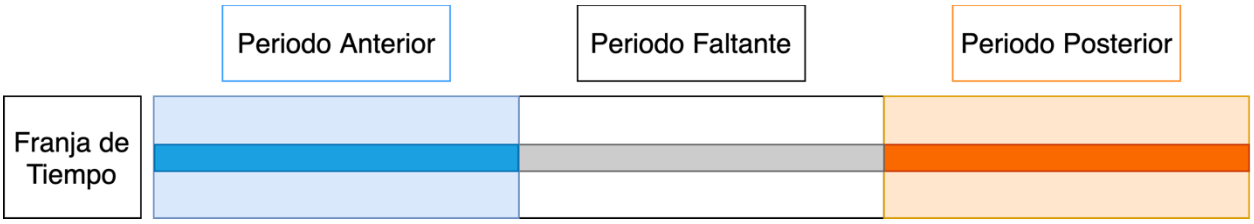

Figura 2. Esquemático para la selección de periodos y franjas horarias para el llenado de datos ausentes.

El procedimiento para el llenado de los datos ausentes de las series de datos filtradas de GHI y TA consiste en los siguientes cuatro pasos [7, 12, 13]:

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 horaria) de los datos ausentes. Los periodos anterior y posterior deben tener el mismo número de días del periodo ausente.
    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), mas únicamente se llenará 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 horaria:
        i. Calcular los parámetros de la distribución normal, es decir, la media (μ) y la desviación estándar muestral (σ) 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. Long, C.N.; Dutton, E.G. BSRN Global Network Recommended QC Tests. V2.0. 2002. Available online: https:// bsrn.awi.de/fileadmin/user_upload/bsrn.awi.de/Publications/BSRN_recommended_QC_tests_V2.pdf
2. C. Toledo, A. M. Gracia Amillo, G. Bardizza, J. Abad, and A. Urbina, “Evaluation of Solar Radiation Transposition Models for Passive Energy Management and Building Integrated Photovoltaics,” Energies, vol. 13, no. 3, p. 702, Feb. 2020, doi: 10.3390/en13030702.
3. C. A. Gueymard, “A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar

projects," Renewable and Sustainable Energy Reviews, vol. 39, pp. 1024–1034, Nov. 2014, doi: 10.1016/j.rser.2014.07.117.
4. F. E. Vignola, A. C. McMahan, and C. N. Grover, "Bankable Solar-Radiation Datasets," in Solar Energy Forecasting and Resource Assessment, Elsevier, 2013, pp. 97–131. doi: 10.1016/B978-0-12-397177-7.00005-X.
5. M. Sengupta, A. Habte, S. Wilbert, C. Gueymard, and J. Remund, "Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Third Edition," NREL/TP-5D00-77635, 1778700, MainId:29561, Apr. 2021. doi: 10.2172/1778700.
6. F. Vignola, C. Grover, N. Lemon, and A. McMahan, "Building a bankable solar radiation dataset," Solar Energy, vol. 86, no. 8, pp. 2218–2229, Aug. 2012, doi: 10.1016/j.solener.2012.05.013.
7. 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.
8. M. Ernst and J. Gooday, "Methodology for generating high time resolution typical meteorological year data for accurate photovoltaic energy yield modelling," Solar Energy, vol. 189, pp. 299–306, Sep. 2019, doi: 10.1016/j.solener.2019.07.069.
9. Photovoltaic system performance. Part 1, Monitoring, Edition 1.0. Geneva, Switzerland: International Electrotechnical Commission, 2017.
10. I. Reda and A. Andreas, Solar position algorithm for solar radiation applications. Solar Energy, vol. 76, no. 5, pp. 577-589, 2004.
11. Nou, Julien & Chauvin, Rémi & Thil, Stéphane & Grieu, Stéphane. (2016). A new approach to the real-time assessment of the clear-sky direct normal irradiance. Applied Mathematical Modelling. 40. 10.1016/j.apm.2016.03.022.
12. IEC (2021), The International Electrotechnical Commission, Photovoltaic system performance Part 1: Monitoring, IEC 61724-1, IEC, Geneve, 2021.
13. IEC (2021), The International Electrotechnical Commission, Photovoltaic system performance Part 2: Power evaluation method, IEC 61724-2, IEC, Geneve, 2021.

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 irradiancia global horizontal (GHI), el procedimiento es independiente de la estampa temporal y, adicionalmente, es el mismo para la temperatura ambiente.

Caso 1. Datos ausentes de un único día

La Tabla A.1 presenta la serie de datos en resolución horaria de irradiancia global horizontal (GHI) 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 11 hasta las 13h 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 horaria de irradiancia global horizontal para siete días con datos ausentes en un único día (i.e., de 11 hasta 13h del día 3).

Hora	Día						
	1	2	3	4	5	6	7
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	326.7	498.3	328.9	699.9	506.7	261.1	481.0
10	825.8	491.8	313.1	760.5	976.0	256.6	663.1
11	612.3	425.1		664.1	917.6	829.4	880.5
12	378.6	292.7		551.8	840.6	1039.4	734.5
13	481.6	203.6		656.2	661.1	483.2	628.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0

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:

$$GHI_{11} = \frac{425.1 + 664.1}{2} = 544.6 \quad \text{(A.1)}$$

$$GHI_{12} = \frac{292.7 + 551.8}{2} = 422.25 \quad \text{(A.2)}$$

$$GHI_{13} = \frac{203.6 + 656.2}{2} = 429.9 \quad \text{(A.3)}$$

Caso 2. Datos ausentes en varios días consecutivos

La Tabla A.2 presenta la serie de datos en resolución horaria de irradiancia global horizontal (GHI) 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 11 hasta las 13h 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 horaria de irradiancia global horizontal para siete días con datos ausentes en varios días consecutivos (i.e., de 11 hasta 13h de los días 3 y 4).

Hora	Día						
	1	2	3	4	5	6	7
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	326.7	498.3	328.9	699.9	506.7	261.1	481.0
10	825.8	491.8	313.1	760.5	976.0	256.6	663.1
11	612.3	425.1			917.6	829.4	880.5
12	378.6	292.7			840.6	1039.4	734.5
13	481.6	203.6			661.1	483.2	628.4

⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0

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:

$$\mu_{11} = \frac{612.3 + 425.1 + 917.6 + 829.4}{4} = 696.1 \qquad \sigma_{11} = \text{std}(612.3; 425.1; 917.6; 829.4) = 191.9 \quad \text{(A.4)}$$

$$\mu_{12} = \frac{378.6 + 292.7 + 840.6 + 1039.4}{4} = 637.8 \qquad \sigma_{12} = \text{std}(378.6; 292.7; 840.6; 1039.4) = 311.7 \quad \text{(A.5)}$$

$$\mu_{13} = \frac{481.6 + 203.6 + 661.1 + 483.2}{4} = 457.4 \qquad \sigma_{13} = \text{std}(481.6; 203.6; 661.1; 483.2) = 163.7 \quad \text{(A.6)}$$

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:

$$V_{3,11}, V_{4,11} = N\sim(\mu_{11}, \sigma_{11}) \quad \text{(A.7)}$$

$$V_{3,12}, V_{4,12} = N\sim(\mu_{12}, \sigma_{12}) \quad \text{(A.8)}$$

$$V_{3,13}, V_{4,13} = N\sim(\mu_{13}, \sigma_{13}) \quad \text{(A.9)}$$

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

La Tabla A.3 presenta la serie de datos en resolución horaria de irradiancia global horizontal (GHI) para doce días.

En orden cronológico, inicialmente se identifica que el periodo faltante abarca las estampas de tiempo desde las 11 hasta las 13h 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 11 hasta las 13h (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 horaria de irradiancia global horizontal para doce días mes con periodos anterior o posterior sin misma cantidad de días inicialmente.

Hora	Día											
	1	2	3	4	5	6	7	8	9	10	11	12
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	326.7	498.3	328.9	699.9	506.7	261.1	481.0	292.0	524.2	415.3	618.7	250.5
10	825.8	491.8	313.1	760.5	976.0	256.6	663.1	634.3	870.1	674.9	694.1	385.7
11	612.3	425.1	448.1	664.1			880.5		721.5	599.8	726.4	1047.1
12	378.6	292.7	501.7	551.8			734.5		847.2	723.7	404.3	1064.2
13	481.6	203.6	348.6	656.2			628.4	785.2	491.2	332.7	625.4	551.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

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.4 a la Ecuación A.6 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 11, 12 y 13 de los días 5 y 6, y para las estampas de tiempo 11 y 12 para el día 8. Los datos válidos del periodo faltante extendido se mantienen inmutables.