RELLENO DE DATOS DE VELOCIDADES DE VIENTO MEDIANTE LA APLICACIÓN DE MÉTODO DE HOT DECK PARA LA ESTIMACIÓN DE PRODUCCIÓN DE ENERGÍA ELÉCTRICA EN BASE AL RECURSO EÓLICO

UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE QUITO

CARRERA:
INGENIERÍA ELÉCTRICA

Trabajo de titulación previo a la obtención del título de
INGENIERO ELÉCTRICO

TEMA:

RELLENO DE DATOS DE VELOCIDADES DE VIENTO MEDIANTE LA APLICACIÓN DE MÉTODO DE HOT DECK PARA LA ESTIMACIÓN DE PRODUCCIÓN DE ENERGÍA ELÉCTRICA EN BASE AL RECURSO EÓLICO

AUTOR:
JORGE EDUARDO QUISHPE FREIRE

TUTOR:
SILVANA FABIOLA VARELA CHAMORRO

Quito, febrero 2020

. Datos de Catalogación Bibliográfica .

Jorge Eduardo Quishpe Freire

RELLENO DE DATOS DE VELOCIDADES DE VIENTO MEDIANTE LA APLICACIÓN DE MÉTODO DE HOT DECK PARA LA ESTIMACIÓN DE PRODUCCIÓN DE ENERGÍA ELÉCTRICA EN BASE AL RECURSO EÓLICO.

Universidad Politécnica Salesiana, Quito – Ecuador 2020

Ingeniería Eléctrica

Breve reseña histórica e información de contacto.

Jorge Eduardo Quishpe Freire (Y'1994). Realizó sus estudios de nivel secundario en el Colegio "Policía Nacional" de la ciudad de Quito. Egresado de Ingeniería Eléctrica de la Universidad Politécnica Salesiana. Su trabajo se basa en la aplicación en el relleno de datos de series incompletas de viento para la producción de energía eléctrica.
jquishpef@est.ups.edu.ec

Dirigido por:

Silvana Fabiola Varela Chamorro (Y'1975). Se graduó en Ingeniería Eléctrica en la Escuela Politécnica Nacional en el año 2001 y de Máster en Ciencias en Ingeniería Eléctrica en el Instituto Tecnológico de Morelia. Actualmente se encuentra trabajando como docente en la Universidad Politécnica Salesiana. Áreas de interés: Transitorios Eléctricos, Sistemas de Distribución.
svarela@ups.edu.ec

DECLARATORIA DE COAUTORÍA DEL DOCENTE TUTOR

Yo, Silvana Fabiola Varela Chamorro declaro que bajo mi dirección y asesoría fue desarrollado el trabajo de titulación "*RELLENO DE DATOS DE VELOCIDADES DE VIENTO MEDIANTE LA APLICACIÓN DE MÉTODO DE HOT DECK PARA LA ESTIMACIÓN DE PRODUCCIÓN DE ENERGÍA ELÉCTRICA EN BASE AL RECURSO EÓLICO"* realizado por Jorge Eduardo Quishpe Freire, obteniendo un producto que cumple con todos los requisitos estipulados por la Universidad Politécnica Salesiana para ser considerados como trabajo final de titulación.

Quito, febrero 2020

……

Ing. Silvana Fabiola Varela Chamorro

C.C.: 1713565818

CESIÓN DE DERECHOS DE AUTOR

Yo, Jorge Eduardo Quishpe Freire, con documento de identificación N° 1725288409, manifiesto mi voluntad y cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que soy autor del trabajo de grado/titulación intitulado: "*RELLENO DE DATOS DE VELOCIDADES DE VIENTO MEDIANTE LA APLICACIÓN DE MÉTODO DE HOT DECK PARA LA ESTIMACIÓN DE PRODUCCIÓN DE ENERGÍA ELÉCTRICA EN BASE AL RECURSO EÓLICO*", mismo que ha sido desarrollado para optar por el título de: Ingeniero Eléctrico, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En aplicación a lo determinado en la Ley de Propiedad Intelectual, en mi condición de autor me reservo los derechos morales de la obra antes citada. En concordancia, suscribo este documento en el momento que hago entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

……

Nombre: Jorge Eduardo Quishpe Freire

Cédula: 1725288409

Fecha: Quito, febrero 2020

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RELLENO DE DATOS DE VELOCIDADES DE VIENTO MEDIANTE LA APLICACIÓN DE MÉTODO DE HOT DECK PARA LA ESTIMACIÓN DE PRODUCCIÓN DE ENERGÍA ELÉCTRICA EN BASE AL RECURSO EÓLICO.

Resumen

Durante las campañas de medición de viento en las estaciones meteorológicas, pueden sucederse condiciones atípicas que producen la pérdida de datos ya sea por falla del equipo, por falla en el suministro eléctrico de respaldo, por saturación de espacio de almacenamiento, entre otras. Por tanto, es necesario que las series de datos sean completadas, tratando de reducir la incertidumbre en el proceso. En el presente trabajo se trabaja con datos de velocidades de viento proporcionadas por la estación meteorológica instalada en la Universidad Politécnica Salesiana. Sede Quito – Campus Sur. Los datos registrados de forma horaria se encuentran completos y validados por el método de las Rachas. En base a la serie completa se obtienen 3 series de datos adicionales quitando de manera aleatoria el 10, 40 y 70% de datos. Aplicando el método de Hot-Deck se completan las series construidas y se realizan comparaciones con la serie de datos original completa. Para la estimación de producción de energía eléctrica se utiliza la Distribución de Weibull. Finalmente, se muestran los resultados en los que se analizan la efectividad del llenado de datos conforme a los escenarios propuestos. Para el desarrollo del trabajo se ha empleado las ayudas computacionales RStudio y Matlab.

Abstract

During the wind measurement campaigns in the weather stations, atypical conditions can occur that produce the loss of data either by equipment failure, by backup power supply failure, by storage space saturation, among others. Therefore, it is necessary that the data series be complete, try to reduce the uncertainty in the process.

In this work we work with wind speed data provided by the weather station installed at the Salesian Polytechnic University. Quito Headquarters - South Campus. The data recorded on an hourly basis is complete and validated by the Rachas method. Based on the complete series, 3 additional data series will be needed, randomly citing 10, 40 and 70% of data. Applying the Hot-Deck method, the constructed series are completed and comparisons are made with the complete original data series. Weibull Distribution is used for the modification of electric energy production. Finally, there are the results in which the effectiveness of data submission is analyzed according to the proposed scenarios. RStudio and Matlab computational aids have been used for the development of the work.

Palabras Clave: Generación eléctrica, energía eólica, Weibull, Hot Deck, homogenización.

Keywords: Electricity generation, wind energy, Weibull, Hot Deck, homogenization.

1. Introducción

La escasez de datos en mediciones es un tema que se ha venido tratando hace mucho tiempo, de modo que varios investigadores sugieren varios métodos de imputación, planteamientos y técnicas, desde la más sencilla hasta la más compleja, cada una con sus ventajas y desventajas que conllevan [1]–[3].

En las series de datos de viento es usual encontrarse con varios problemas como la escasez de datos y/o la carencia de homogeneidad, todo esto es debido a ciertos factores que influyen en las mediciones como errores en las codificaciones, en el entorno que se encuentra, en los dispositivos utilizados o en el modo de realizar las mediciones [4]–[6]. La serie de la velocidad del viento es no lineal y no estacionaria, y tiene una variación variable en el tiempo. Por lo tanto, la velocidad del viento a menudo se considera uno de los parámetros meteorológicos más difíciles de pronosticar [7], [8].

La homogenización de datos también es un tema muy importante ya que detectará si los datos con los que se trabajarán se encuentran con valores atípicos, es decir que los datos obtenidos no se encuentren dentro de los valores correctos, poniendo en riesgo la investigación ya que no se podría obtener resultados correctos [9]–[11].

Llenar el valor faltante es la tarea principal del procesamiento de datos, actualmente se prefiere la imputación de Hot Deck [12]–[14]. Para abordar este problema , han propuesto autores implementar modelos de distribución para imputación de datos faltantes [1], [14]–[16], otros autores presentan como modelo, la red Bayesiana Dinámica útil para completar los valores faltantes en los datos, además el algoritmo emplea una regresión vectorial de soporte para predecir los valores faltantes [17], [18].

El método que presentan [19], [20],corresponde a un método de k-vecino más cercano basado en el perfil para estimar los datos de índice espectral que faltan y proporcionan un llenado de huecos preciso bajo cambios graduales y abruptos.

Otro método empleado [21], son matrices transitorias de categoría mínima y máxima para llenar los datos a partir de matrices iniciales conocidas y son empleadas para modelos digitales de elevación, necesarias en hidrología para determinar las rutas del agua, la red de drenaje, la división de la cuenca, el sedimento y el movimiento.

También proponen [22], interpolación lineal a través del espacio. seleccionado una corrección dinámica adicional basada en las diferencias para datos faltantes para series de tiempo constantes e ininterrumpidas del nivel del agua como aplicaciones que incluyen el monitoreo de mareas de tormenta y el manejo general de emergencias, planificación costera, mapeo de costas, restauración de hábitat y actividades operativas como el apoyo de librado y navegación.

En [23], proponen aplicaciones de microarrays de bioinformática por medio de datos estadísticos y algoritmos de agrupación, usados en biogenética para llenado de datos de tipificación genética.

En [24], plantean relleno de datos por vectores espaciales, de acuerdo a las coordenadas de datos conocidos y cercanos de mapeo.

En [25], muestra el método de suavizado de optimización de puntos de control por medio de una red neuronal, modelo de red wavelet y el modelo de regresión automática. El método se basa en las dependencias entre los datos del propio conjunto de datos para completar algunos de los valores faltantes, de modo que el conjunto de datos lleno pueda reflejar con mayor precisión la integridad del conjunto de datos, la mayoría de estos estudios utilizan formularios web o archivos de texto como fuentes de datos externas,

rellenando los valores faltantes de datos con resultados de búsqueda de fuentes de datos externas big Data.

En [26], presenta un algoritmo de duplicación paralela, llamado FER-APARDA, mediante el uso de vinculación de registros probabilísticos, para detectar con éxito réplicas en conjuntos de datos sintéticos con más de 1 millón de registros aproximadamente, logrando una aceleración casi lineal para mejorar la calidad de la información y llenado de datos

Los métodos a menudo requieren supuestos de distribución y conocimiento previo sobre los datos, esto puede causar cierta dificultad para la investigación de ingeniería. Por lo tanto, el objetivo del presente trabajo consiste en la implementación del Método de Hot Deck Múltiple [27]–[30], para el relleno de datos faltantes en las series de velocidades de viento, siendo un tema de interés, ya que, da una solución a un problema muy frecuente que se encuentra al momento de realizar las mediciones de este recurso, no obstante queda aclarar que el relleno de los valores faltantes será con valores aproximados a los reales, debido a que estos métodos no pueden reproducir los valores verdaderos [31],[32], [33].

Los datos de las series de velocidades de viento son proporcionados por el Instituto Nacional de Meteorología e Hidrología del Ecuador (INAMHI) de la estación meteorológica que se encuentra ubicada en la Universidad Politécnica Salesiana, Sede Quito, Campus Sur, con código M1274, latitud -0.271900 longitud -78.55000 y altitud 2886.00 msnm [34]–[36].

A partir del análisis generaran valores razonables que se encuentren dentro del rango de la serie que será analizada, para poder estimar la producción de energía eléctrica mediante la Distribución de Weibull [37]–[42].

La distribución del presente artículo es de la siguiente manera: Sección 1: introducción y antecedentes a las series de viento del recurso eólico, Sección 2: marco teórico, Sección 3: modelado del método Hot Deck Múltiple, Sección 4: implementación del modelo matemático, Sección V: análisis de resultados de las series de viento imputadas junto con la distribución de Weibull, Sección 6: conclusiones y trabajos futuros y Sección 7: referencias y estado del arte.

2. Marco teórico

2.1. Series de Velocidades de Viento

El recurso eólico en los últimos años ha sido una fuente importante para la generación de energía eléctrica, por lo tanto, se dice que es una fuente de energía madura, confiable y renovable, esto se puede observar en la magnitud de potencia instalada, en costos y en desempeño, lo cual la hace llamativa dentro del mercado eléctrico [35].

Al viento se lo considera que es una fuente de energía, ya que tiene varias propiedades que intervienen en su disponibilidad presentando variaciones en diferentes escalas espaciales y temporales, tanto en la superficie como en la altura [31].

Los datos de las series de velocidades de viento a utilizar en el presente trabajo son obtenidas a través de un anemómetro instalado en la estación meteorológica, y hacen referencia al periodo que va desde el 21/10/2018 hasta el 21/10/2019, por lo tanto, este periodo contiene 365 días, para esto, las mediciones han sido registradas cada hora durante estos días, por esta razón, se trabajará con 8760

datos en total los cuales han sido encolumnados para poder trabajar de una manera más organizada.

Los datos de las series de viento obtenidos se representan en la Figura 1.

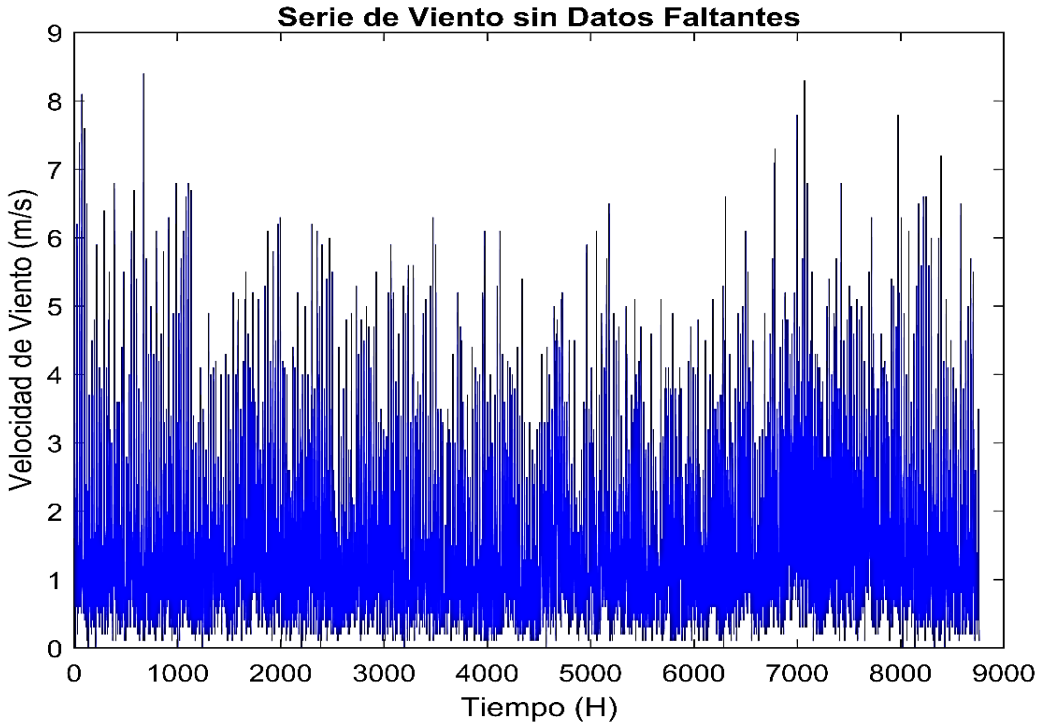

Figura 1. Series de viento obtenidas del INAMHI por hora.

2.2. Homogenización de datos

Antes de empezar a trabajar con los datos obtenidos de las series de viento es necesario hacer un control de homogenización, el cual nos permitirá saber si los datos son adecuados para su utilización, para esto se utilizará el método La Racha o La Ronda ya que posee un procedimiento de manejo muy fácil y también tiene un alto grado de confiabilidad estadística [9], este método sirve para poder detectar si en las data obtenida de las series de datos de velocidades de viento existen valores dudosos.

Este método consiste en el siguiente procedimiento [10]:

Algoritmo del Método de Rachas

Paso 1: Se calcula la mediana de las series de datos obtenidas, estas pueden ser mensuales (N) y se elabora una tabla con todos estos valores.

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i$$

Paso 2: Anotar las veces que las medianas calculadas superen la mediana general (+) o no sobrepasen la mediana general (-).

Paso 3: Se cuentan el número de rachas que hay (NS). Una racha es cada cambio de signo que exista, referenciándose al paso 2.

Paso 4: Se considera que NA es el número de valores por encima de la mediana general y NB es el número de valores por debajo de la mediana general, entonces le corresponden un número determinado de rachas, con una probabilidad entre 10% y 90% de que sean homogéneas.

Paso 5: Todos estos valores dependerán del número de elementos que posea la serie y deberán ser observados en la tabla que Castillo y Sentis proponen en [11].

Terminar

Una vez aplicado el método a todos los datos se comprobó que la data a

utilizar es homogénea, por lo tanto, se puede utilizar estos datos de manera confiable.

2.3. Datos faltantes

Estos datos faltantes pueden producirse por varias razones, por ejemplo: cuando una persona se olvida responder una pregunta, cuando el entrevistador se olvida hacer una pregunta, o cuando un dispositivo tiene patrones de salto [43].

Para el registro de series de viento comúnmente se plantea una matriz de datos $\boldsymbol{X}$, en donde estas bases de datos abarcan las mediciones de K años y N casos (por lo general son registros horarios o registros diarios) [12].

$$X=[x_{ij}] \quad (1)$$

Donde:
$i = 1,2, \dots, N$
$j = 1,2, \dots, K$

Al momento en el que se desea trabajar con datos faltantes, es necesario conocer si los datos faltantes generados por los dispositivos lo hacen de forma aleatoria o no. Para poder identificarlos se han clasificado en tres tipos [13], [14], [43] como lo muestra la Tabla 1.

Tabla 1. Tipos de Datos Faltantes

Tipo	Descripción
Completamente aleatoria (**MCAR**: "Missing Completely at Random")	Si la probabilidad de que el valor faltante no dependa de alguna otra variable.
Aleatoria (**MAR**: "Missing at Random")	Si la probabilidad de que el valor faltante dependa de otras variables. El valor que falta se puede estimar usando otras variables.
No aleatorio (**NMAR**: "Not Missing at Random")	Si la probabilidad del valor faltante depende de otros valores que también son faltantes, por ello, los valores faltantes no pueden ser estimados a partir de variables que son existentes.

2.4. Métodos para imputación de datos faltantes

La escasez de datos en mediciones es un tema que se ha venido tratando hace mucho tiempo, de modo que varios investigadores sugieren demasiados métodos de imputación, planteamientos y técnicas, desde la más sencilla hasta la más compleja, cada una con sus ventas y desventajas que conlleva cada una [1]. Entre estos están:

a) *Sustitución media:* Se considera como un método arcaico. Este método consiste en que todos los valores faltantes en la variable fueron reemplazados por la media del registro total. Por ejemplo, si la edad media de los participantes en un estudio es 68.2, entonces 68,2 se usará para reemplazar todos los valores faltantes. La sustitución es sencilla y veloz, pero el método se basa en la suposición de que todos los valores faltantes son de tipo MCAR, lo cual es muy poco común [14], [43].

b) *Imputación de regresión o imputación media condicional:* Este método usa la matriz de correlación para encontrar varios predictores de valores faltantes. Se elije al mejor predictor y se lo establece como una variable independiente en una ecuación de regresión. Las variables dependientes se crean desde variables que poseen datos incompletos. Los eventos con datos completos se emplean para crear la ecuación de regresión; esta ecuación se emplea para poder predecir los datos de valores faltantes para los eventos incompletos. Es un

proceso repetitivo, hasta que los exista poca diferencia entre los valores que se predecirán, es decir convergen, los valores predictores de la última repetición serán los valores que se utilizaran para reemplazar los valores faltantes. Este método puede ser un modelo sobreestimado y de menor significancia [14], [43].

c) *Eliminación Listwise:* Este método también es llamado análisis de caso completo [15]. Cuando este método es usado, automáticamente el programa de computador elimina cualquier caso que tenga datos perdidos para cualquier bivariado o multivariado análisis. Por consiguiente, una considerable parte de los datos no pueden ser usados ya que los casos se eliminan. Esta condición elevará el peligro de sesgo cuando el valor faltante esté presente [14], [15], [43].
d) *Imputación Múltiple:* Este método es propuesto por Rublin ya que abarco el tema de la incertidumbre en los datos imputados [27].

   En este método los valores faltantes son imputados n-veces para representar la incertidumbre de los posibles valores que serán imputados. Los valores de las n-veces se examinan para obtener una única estimación, pero combinada [44].

   Los resultados generados por este método son un poco cuestionados ya que solo es efectivo con una pequeña cantidad porcentual de valores faltantes de la muestra a analizar [1].
e) *Hot Deck Simple:* Este método consiste en que el valor faltante puede ser reemplazado por otros valores que son semejantes con características de correlación identificadas. Se emplea una matriz de correlación para poder determinar que variable se encuentra más correlacionada con los datos faltantes [14], [43].

   Ford [45] define este método como un procedimiento en el cual los elementos faltantes y/o perdidos son reemplazados por valores de uno o varios registros semejantes. La división de registros semejantes en grupos disjuntos y homogéneos se realiza de manera los registros "correctos" (donantes) siguen la misma distribución que los registros "incorrectos" (receptores). Por esta razón y a la propiedad de respuesta, todos los conjuntos de datos imputados contienen solo valores plausibles, que la mayoría de métodos no garantizan [46]. De este método se deriva otro el cual es conocido como Hot Deck Múltiple, este método es una variación del método Hot Deck Simple combinado con el método de imputación repetida y el método tradicional de la imputación múltiple paramétrica [29].

Todos estos métodos antes mencionados han sido aplicados en distintas áreas de estudio demostrando tanto su efectividad al igual que sus falencias

2.5. Función de Distribución de Weibull

En 1939 el físico sueco Waloddi Weibull, anuncio esta función en un estudio de resistencia de materiales, la

cual desde entonces se ha utilizado en distintas áreas para representar datos continuos. Su extenso uso se basa en la modelación de diversas formas: decreciente, exponencial, Rayleigh, normal y normal con asimetría negativa [37].

Entre las principales ventajas de esta función comparada con otras funciones de densidad de probabilidad son: permite estimar satisfactoriamente la asimetría de la distribución de densidad de probabilidad; si para la disposición de velocidad de viento la función tiene un factor de forma, en ese caso la distribución de la velocidad al cubo de igual manera sigue la función de Weibull [38].

La función de densidad de probabilidad de Weibull [38] [39] [40], se encuentra definida por:

$$f(v) = \left(\frac{k}{c}\right)\left(\frac{v}{c}\right)^{k-1} e^{-\left(\frac{v}{c}\right)^{k}} \qquad (2)$$

Donde:

v Es la velocidad del viento en (m/s).

c Es el parámetro de escala con unidades de (m/s).

k Es el parámetro de forma sin dimensiones.

Para determinar los parámetros de la función de Weibull se utilizan las ecuaciones a continuación [47]:

$$k = \left(\frac{\sum_{i=1}^{N} v_i^k \, ln(v_i)}{\sum_{i=1}^{N} v_i^k} - \frac{\sum_{i=1}^{N} ln(v_i)}{N}\right)^{-1} \qquad (3)$$

y

$$c = \left(\frac{1}{N}\sum_{i=1}^{N} v_i^k\right)^{\frac{1}{k}} \qquad (4)$$

Donde:

N Representa el número de observaciones

v_i Es la velocidad promedio del viento registrada en un intervalo de tiempo.

Una vez obtenido los resultados de los parámetros de las Ecuaciones 3 y 4, se podrá obtener el resultado de la ecuación 2, así podremos determinar la función de Weibull [42] como se observa en la Figura 2.

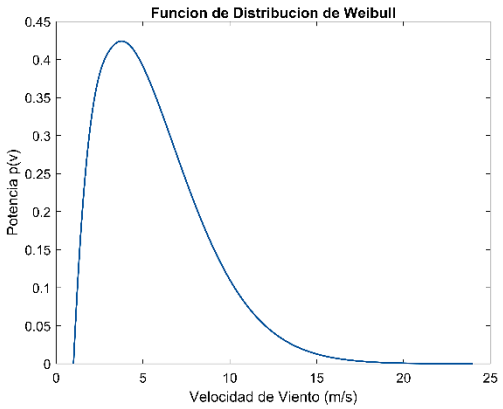

Figura 2. Función de Distribución de Weibull.

3. Método Hot Deck Múltiple para imputación de datos

Se ha elegido utilizar este método, porque los datos que utilizaremos de las series de viento fueron obtenidas a través de mediciones hechas por anemómetros, los cuales pueden presentar ausencia de valores de tipo MCAR, lo cual es un indicativo para poder aplicar perfectamente el método.

Cranmer y Gill nos indican en [29] que este método posee una serie de ventajas que indican que el método es sustancial para la imputación de datos, entre ellas están:

- Vence un problema significativo del método Hot Deck tradicional, ya que imputa diversos conjuntos de datos para luego efectuar varios análisis y la

combinación entre coeficientes y matrices de covarianza.

- Las cualidades de los datos discretos son conservadas sin redondeo y a su sesgo asociado y también realiza la reducción de los errores estándar, lo cual significa una gran mejora frente a otras técnicas de imputación múltiple.
- El método Hot Deck múltiple es en verdad un enfoque no paramétrico de la imputación. La imputación múltiple típica requiere suposiciones de normalidad y se fundamenta en modelos paramétricos de los valores faltantes en cambio el Hot Deck Múltiple evitan todos estas estas suposiciones.
- El Hot Deck Múltiple funciona con valores continuos, pero funciona mejor donde la imputación múltiple funciona peor que es con valores discretos.
- El algoritmo del Hot Deck Múltiple es fácil de entender ya que es muy intuitivo.

Al analizar cada una de estas ventajas podemos darnos cuenta de que al ser un método que abarca varias técnicas nos permite realizar una imputación de manera confiable, precisa y sencilla.

En este método en caso de existir carencia de valores para el i-enésimo caso los valores examinados del conjunto de datos (x,y) son usados para llenar los valores faltantes. Una observación muy importante es que los valores que han sido imputados son valores extraídos de los valores reales, en vez de ser extraídos de valores creados, de tal forma que, en el momento que las variables discretas se imputan utilizando un método hot deck, estas mantienen sus propiedades discretas [29].

El método Hot Deck Múltiple se diferencia de otros métodos Hot Deck, en que el Hot Deck Múltiple utiliza diferentes valores para un dato faltante [32].

3.1. Valoración de Afinidad

La elección de los valores donadores es de suma importancia para la validación de los valores a imputar. Para medir el grado en que un dato con un valor faltante es semejante a otro dato, se crea un conjunto de valoraciones de afinidades de 0 y 1, cuyos elementos se denotan como α_{ij} y evalúan el grado de semejanza que tiene el receptor *i* con cada donador *j*. La afinidad se especifica en términos del grado en que cada donador potencial encaja con los valores del receptor en todas las variables distintas de la imputada [29].

Para cada valor se tiene el vector (y_i,x_i) donde y_i muestra la variable de resultado y x_i es un vector de *k*-distancia de variables explicativas puramente discretas, cualquiera de estas puede poseer valores faltantes. Si el *i*-enésimo caso sujeto a consideración posee q_i valores faltantes en x_i, entonces un vector donante potencial, x_j, $j \neq i$, tendrá entre 0 y k-q_i coincidencias exactas con *i*.

Ahora se define z_{ij} como el número de variables para cada donador potencial *j* y el receptor *i* tiene diferentes valores. Por esta razón k-q_i - z_{ij} es el número de variables en las que *j* y *i* coinciden impecablemente. Este valor a gran escala por el mayor

número de posibles coincidencias (k-q_i) es el valor de afinidad:

$$\alpha_{ij} = \frac{k - q_i - z_{ij}}{k - \in q_i} \qquad (5)$$

Se observa que el valor de afinidad es definido en la ecuación 2 tiene las propiedades deseables para $\alpha_{ij}=1$ para $i \in \boldsymbol{D}_R$ (datos con respuestas) y $\alpha_{ij}=1$ para $i \in \boldsymbol{D}_{RN}$ (datos sin respuesta). Se escribe el vector de todos los α_{ij} como $\boldsymbol{\alpha}_i$. En la Tabla 2 se muestra como el valor de α_{ij} disminuye al igual que el número de coincidencias [29].

Tabla 2. Valores de Afinidad.

Coincidencia	**Sin coincidencia**	**α_{ij}**
k-q_i	0	$(k\text{-}q_i)/(k\text{-}q_i)$
k-q_i-1	1	$(k\text{-}q_i\text{-}2)/(k\text{-}q_i)$
k-q_i-2	2	$(k\text{-}q_i\text{-}3)/(k\text{-}q_i)$
⋮	⋮	⋮
2	k-q_i-2	$2/(k\text{-}q_i)$
1	k-q_i-1	$1/(k\text{-}q_i)$
0	k-q_i	$0/(k\text{-}q_i)$

3.2. Imputación usando valores de afinidad

Cranmer y Gill en [29] nos explica que se considera un conjunto de datos $\boldsymbol{D}$, con vectores de fila (y_i,x_i) y valores faltantes en alguna o en todas las variables, en el cual $\boldsymbol{D}_R$ continúa denotando respuestas analizadas y $\boldsymbol{D}_{RN}$ continúa denotando no respuestas. Para cada variable h, el valor de $x_{i[h]}$ se encuentra perdido.

Cada valor con una $xi_{[h]}$ faltante puede ser considerado miembro de una celda de imputación.

Las examinaciones que constituyen la mejor celda de imputación, $\boldsymbol{C}$, son aquellas para las cuales α_{ij}= max $\boldsymbol{\alpha}_i$ para todas las j en la celda $\boldsymbol{C}$. En algunas situaciones, el conjunto solo consistirá en coincidencias correctas y en otras situaciones solo consistirá en las mejores coincidencias.

Al subdividir apropiadamente los valores en las celdas de imputación, se puede tratar las resoluciones en la mejor celda de imputación $\boldsymbol{C}$ como la realización de variables aleatorias independientes e idénticamente distribuidas (iid) con media μ_c y varianza σ^2_c.

Entonces, todos los valores de $x_{i[h]}$ en la celda se tomarán de la misma distribución, no importa la distribución que sea.

Algo importante para el propósito final, es dividir a los valores en celdas independientes de los sistemas de muestreo y respuesta implica que la asignación en la ecuación 3 es efectiva para valores observados como no observados.

$$x_{i\,[h]} \,|\, (\boldsymbol{D}, \boldsymbol{D}_R, \boldsymbol{D}_{NR}) \,| \sim f\,(\mu_c, \sigma^2_c) \qquad (6)$$

Para todo i en $\boldsymbol{C}$.

Esto indica que la distribución de $x_{i[h]}$ en la celda de imputación es la misma indistintamente de si está condicionada a todos los datos $\boldsymbol{D}$, los datos guardados $\boldsymbol{D}_R$, o los datos faltantes $\boldsymbol{D}_{RN}$.

Puesto que los valores faltantes dentro de la mejor celda de imputación son de tipo MCAR y hay realizaciones de $x_{[h]}$ que son independientes del diseño, se extraerá al azar de los valores de $x_{[h]}$ para imputar el valor faltante $x_{i[h]}$. Aún se necesitan múltiples imputaciones para capturar la varianza de imputación. A continuación, se extrae con reemplazo $M \geq 2$ valores de

$x_{[h]}$, luego asigne la m-enésima extracción para el m-enésimo conjunto de valores duplicados como $\boldsymbol{D_m}$, m=1,…,M. El proceso se reitera con cada valor faltante en cada columna del conjunto de datos. El resultado será M conjunto de datos imputados cada uno ya sin valores faltantes [29].

3.3. Estimación después de la imputación Hot Deck Múltiple

Se ejecutará M copias de la especificación del modelo, utilizando la media de los coeficientes estimados y la media ponderada de los errores estándar. Dejando a θ_m, m=1,…,M ser coeficientes estimados calculados individualmente desde los conjuntos de datos imputados M, y $\sum_m$, m=1,…,M son las variaciones asociadas para θ_m. Una estimación individual de θ, es creada desde la media de los m-valores [29].

$$\theta_M = \frac{1}{M}\sum_{m=1}^{M}\theta_m \qquad (7)$$

Al calcular la variabilidad de estimación de θ es un poco más difícil que la media usada para generar θ_M ya que el total de la varianza está compuesto de variaciones de coeficientes estimados dentro de cada conjunto de datos imputados y la variación de coeficientes estimados entre los conjuntos de datos imputados. La varianza en la imputación es la media de las variaciones de coeficientes individuales entre los modelos:

$$W_M = \frac{1}{M}\sum_{m=1}^{M}\sum_m \qquad (8)$$

La varianza entre imputaciones es la varianza de las estimaciones de M coeficientes.

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\theta_m - \theta_M)^2 \qquad (9)$$

La varianza total de θ_M es una suma ponderada de tipo ANOVA:

$$T_M = W_M + \left(1 + \frac{1}{M}\right)B_M \qquad (10)$$

Donde [1+1/M] es el ajuste para M finitos y los nuevos grados de exención son:

$$df_{M1} = (M-1)\left[1 + \frac{1}{M+1}\frac{W_M}{B_M}\right] \qquad (11)$$

Estos valores se generan automáticamente para el usuario mediante un software estadístico que implementa la imputación múltiple [29].

4. Implementación del modelo matemático

La modelación matemática propuesta en este artículo de investigación es resuelta en el software RStudio [33]. Este modelo matemático puede ser aplicado a distintos tipos de series con datos faltantes, estas pueden ser de dimensiones extensas como también de una dimensión reducida.

Para imputar los datos faltantes de las series de viento aplicando la metodología descrita, se emplea el modelo en la data de series de viento.

Algoritmo de Imputación Múltiple Hot Deck [29]

Paso 1: Crea M ≥2 copias del conjunto de datos D, cada una con valores perdidos y observados. Estos se denotan $\boldsymbol{D_m}$ para m=1,…,M.

Paso 2: Busca secuencialmente en cada una de las columnas del conjunto de datos los valores que estén faltantes.

a) Cuando se encuentra un valor faltante, crea un vector de valoraciones de afinidad, el cual calcula la proximidad de las otras filas que poseen los datos faltantes.
b) Crea la mejor celda de imputación para el valor faltante de la cual se toman imputaciones aleatoriamente, para producir un vector de imputaciones.
c) Imputa uno de estos valores en la celda apropiada de cada conjunto de datos duplicados.

Paso 3: Repite el paso 2 hasta que se hayan logrado imputar todos los valores faltantes en los conjuntos de datos *M*.

Paso 4: Estima la estadística de interés para cada conjunto de datos.

Paso 5: Combina las estimaciones estadísticas en una única estimación.

Terminar

Para la estimación de energía, a las series imputadas con distintos porcentajes de datos, se les aplica la función de Distribución de Weibull, la metodología de este modelo matemático se encuentra descrita por el algoritmo que está a continuación.

Algoritmo de la Función de Distribución de Weibull

Paso 1: Se toma la data de viento disponible y se elabora una tabla, clasificándolos por intervalos.

Paso 2: Contar el número de frecuencias y calcular la frecuencia relativa y la frecuencia acumulada de cada intervalo.

Paso 3: Calcular el parámetro *k* de la función de Weibull.

Paso 4: Calcular el parámetro *c* de la función de Weibull.

Paso 5: Reemplazar los parámetros *c* y *k* en la ecuación general de la función de Distribución de Weibull.

Paso 6: Se procede a reemplazar uno por uno cada valor de los intervalos del paso 1 para poder obtener los valores de *p(v)*.

Paso 7: Se procede a crear una tabla con los valores de los intervalos junto con su valor de *p(v)*.

Paso 8: Graficar la tabla del paso 7 y observar la gráfica de la función de distribución de Weibull.

Terminar

5. Análisis de resultados

Como resultado de la aplicación del modelo matemático propuesto en las series de viento, se obtiene una nueva serie con valores imputados lo cual permite llenar los valores faltantes en las series. En el caso de estudio se comparará series de viento con distintos porcentajes de datos faltantes

5.1. Caso de estudio

La serie de viento proporcionada por el INAMHI [34] de la estación meteorológica de la Universidad Politécnica Salesiana después de ser homogenizada por el método de las Rachas se considerará como nuestra serie de viento con el 100% de datos disponibles, la cual contiene 8760 muestras las cuales fueron tomadas cada hora, durante el periodo del 21/10/2018 hasta el 21/10/2019.

A esta serie de datos de viento que se encuentra al 100% de datos disponibles, se eliminará aleatoriamente distintos datos, de esta manera se obtendrán distintas series de viento con distintos porcentajes de datos disponibles, las cuales luego serán comparadas con la serie que se encuentra al 100%, por lo tanto, se analiza 1 caso de estudio con tres escenarios como se describe a continuación:

- Escenario 1: serie de viento con 90% de datos disponibles.

- Escenario 2: serie de viento con 60% de datos disponibles.
- Escenario 3: serie de viento con 30% de datos disponibles.

En la Figura 3 se muestra cómo se encuentran distribuidas las series con los distintos porcentajes de datos disponibles.

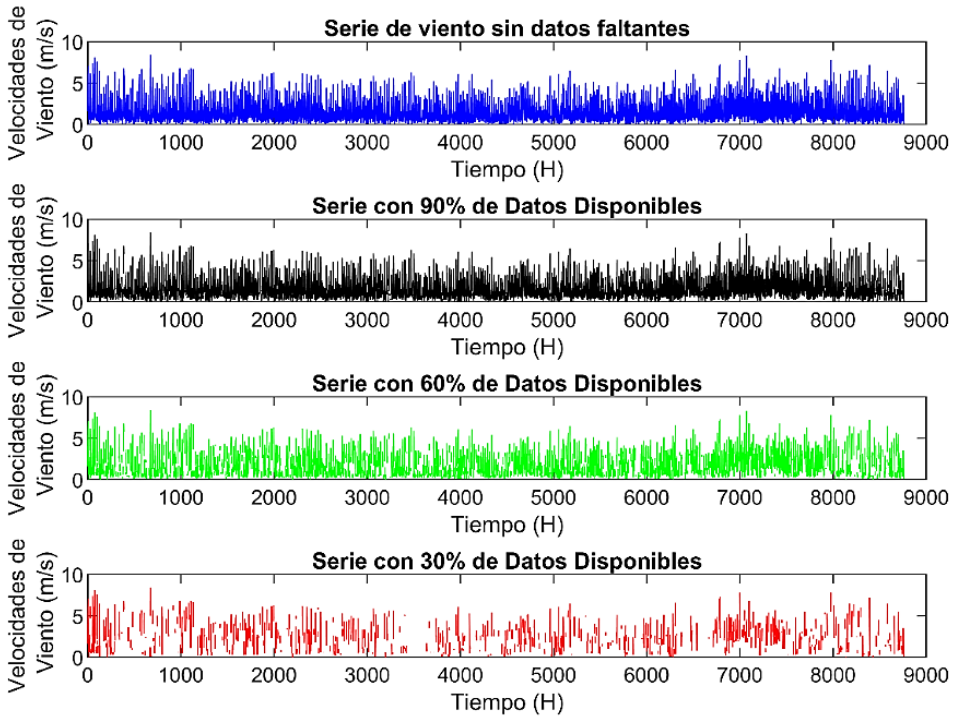

Figura 3. Series de viento con distinto porcentaje de valores disponibles

Para el caso de estudio se propone este tipo de metodología para poder observar como el Método de Hot Deck Múltiple funciona ante distintos porcentajes de valores faltantes en las series de datos de viento.

5.2. Resultados

La simulación es para cada uno de los escenarios. Cada escenario se conforma por diferentes porcentajes de datos disponibles de series de viento. En cada escenario podemos comparar el comportamiento de 90%, 60% y 30% de la serie de datos disponibles, esto a su vez poder generar valores para la estimación de producción de energía eléctrica en base al recurso eólico.

En la Figura 4 se muestra como las series de datos con valores faltantes han sido imputadas con el método Hot Deck.

En la Tabla 3 se observan los valores de la media de cada serie de datos con distintos porcentajes de datos disponibles.

En la Figura 5 se representan las velocidades medias por mes de cada serie aplicada el Método de Hot Deck con distintos porcentajes de valores disponibles. Como se puede observar la serie con el 30% de datos disponibles se encuentra muy dispar de las otras series, pero también se observa que las series con el 60%, el 90% y el 100% de datos disponibles no se encuentran muy disparejas entre ellas.

Los resultados de imputación de las series con distintos porcentajes de datos se muestran en la Tabla 4, donde se observan valores que son representativos al momento de comparar las series imputadas.

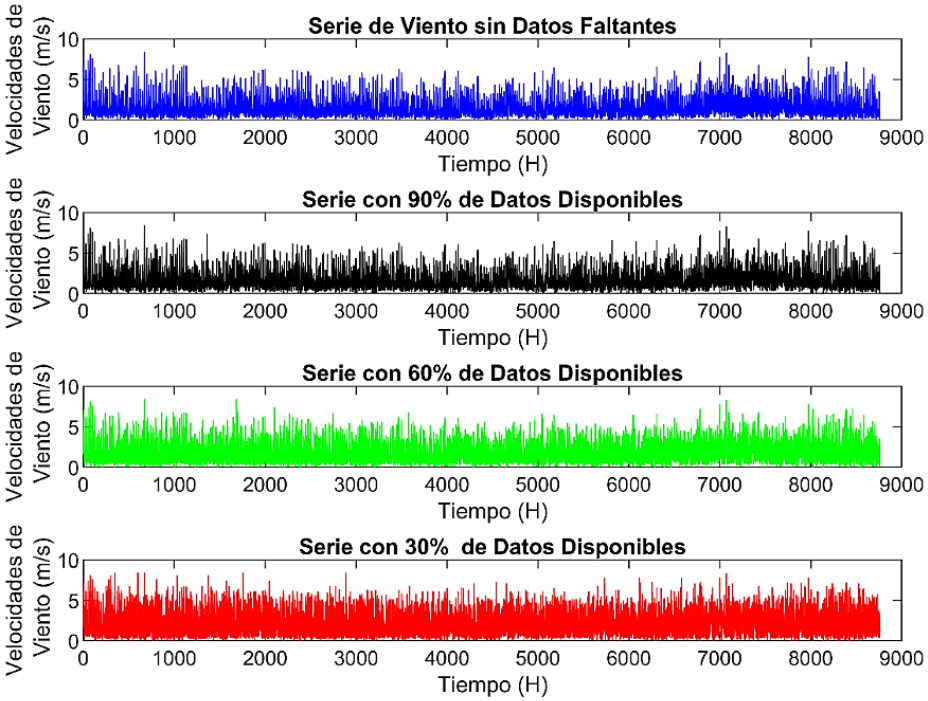

Figura 4. Series de viento aplicadas el método de Hot Deck

Tabla 3. Medias con distintos porcentajes de datos disponibles

		Medias con distintos porcentajes de datos disponibles			
Año	**Mes**	**30%**	**60%**	**90%**	**100%**
2018	Octubre	2,22	1,93	1,85	1,75
	Noviembre	2,24	1,72	1,63	1,51
	Diciembre	2,27	1,77	1,64	1,53
2019	Enero	2,19	1,78	1,73	1,63
	Febrero	2,07	1,55	1,47	1,38
	Marzo	1,96	1,53	1,44	1,32
	Abril	1,99	1,55	1,42	1,34
	Mayo	2,21	1,65	1,52	1,45
	Junio	2,17	1,67	1,55	1,42
	Julio	2,20	1,78	1,62	1,55
	Agosto	2,42	2,20	2,03	1,97
	Septiembre	2,34	1,91	1,83	1,76
	Octubre	2,28	1,80	1,60	1,49

En base a los resultados de la tabla 4 y la Figura 5, se observa que entre mayor porcentaje de datos faltantes exista en las series de datos, mayor será su variación respecto a la serie original, como lo demuestra la mediana y la media de las series imputadas, deduciendo que los valores rellenados variarán dependiendo a los datos con los que se trabaje con el método Hot Deck,

Tabla 4. Comparación de valores de las series imputadas

% de Datos completos	% de Datos Faltantes	Mediana	Media
100%	0%	1,2	1,5377
90%	10%	1,3	1,6325
60%	40%	1,3	1,7471
30%	70%	2	2,1959

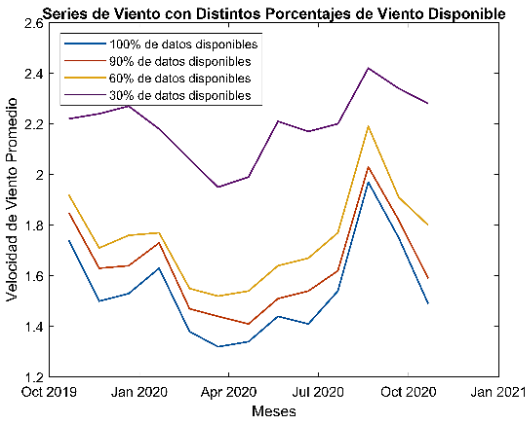

Figura 5. Series de viento promedio mensuales con distintos porcentajes de datos disponibles

Se ha realizado el cálculo del coeficiente de correlación entre las series de datos que han sido aplicadas el método de Hot deck estos resultados se muestran en la Tabla 5.

Tabla 5. Coeficientes de correlación entre series de viento imputadas con distintos porcentajes de datos faltantes

Coeficientes de Correlación entre Series con distintos porcentajes de datos disponibles	
	100%
100%	1
90%	0,9044
60%	0,7171
30%	0,3952

Es posible observar en la Tabla 5 que la falta de datos en las series es un factor importante en la imputación de datos ya que al aplicar el método de Hot Deck se observa que:

- Entre la serie con 0% de datos faltantes y la serie con el 10% de datos faltantes el coeficiente de correlación es muy alto, a lo cual se puede decir que la relación si es lineal entre estas dos series.
- Entre la serie con 0% de datos faltantes y la serie con el 40% de datos faltantes el coeficiente de correlación es alto a lo cual se puede decir que la relación si es lineal entre estas dos series.
- Entre la serie con 0% de datos faltantes y la serie con el 70% de datos faltantes el coeficiente de correlación es muy pequeño a lo cual se puede decir que la relación no es lineal entre estas dos series.

Tabla 6. Factores k y c para la Función de Distribución de Weibull con distintos porcentajes de datos en las series de viento

Factores	**Series completas en distintos porcentajes de datos disponibles**			
	100%	**90%**	**60%**	**30%**
k	1,57	1,57	1,30	1,29
c	1,78	1,78	1,87	2,40

En la Tabla 6 se muestran los valores obtenidos de los factores *k* y *c*, los cuales serán reemplazados en la ecuación 2, junto con los valores obtenidos de los intervalos de las series de viento, para poder obtener los valores de la Tabla 6.

En la Tabla 7 se tiene los valores de la distribución de Weibull para cada intervalo de viento, de las series de viento imputadas con distintos porcentajes de valores faltantes, se han dividido en 25 intervalos, estos serán de gran aporte para la gráfica de la función de Weibull.

Tabla 7. Valores de las Funciones de Distribución de Weibull con distintos porcentajes de datos en las series de viento

Intervalos de viento	**Series completas en distintos porcentajes de datos disponibles**			
	100%	**90%**	**60%**	**30%**
Velocidad (m/s)	**p100(v)**	**p90(v)**	**p60(v)**	**p30(v)**
0,336	0,3177	0,3177	0,3720	0,2825
0,682	0,4093	0,4093	0,3925	0,3070
1,028	0,4228	0,4228	0,3678	0,3006
1,374	0,3907	0,3907	0,3253	0,2804
1,72	0,3350	0,3350	0,2774	0,2538
2,066	0,2711	0,2711	0,2302	0,2249
2,412	0,2091	0,2091	0,1870	0,1960
2,758	0,1548	0,1548	0,1492	0,1685
3,104	0,1104	0,1104	0,1173	0,1432
3,45	0,0761	0,0761	0,0910	0,1205
3,796	0,0509	0,0509	0,0697	0,1005
4,142	0,0331	0,0331	0,0528	0,0832
4,488	0,0209	0,0209	0,0397	0,0684
4,834	0,0129	0,0129	0,0295	0,0558
5,18	0,0077	0,0077	0,0217	0,0453
5,526	0,0045	0,0045	0,0159	0,0365
5,872	0,0026	0,0026	0,0115	0,0293
6,218	0,0015	0,0015	0,0083	0,0234
6,564	0,0008	0,0008	0,0060	0,0186
6,91	0,0004	0,0004	0,0042	0,0147
7,256	0,0002	0,0002	0,0030	0,0116
7,602	0,0001	0,0001	0,0021	0,0091
7,948	0,0001	0,0001	0,0015	0,0071
8,294	0,0000	0,0000	0,0010	0,0055

Intervalos de viento	Series completas en distintos porcentajes de datos disponibles			
	100%	90%	60%	30%
Velocidad (m/s)	p100(v)	p90(v)	p60(v)	p30(v)
8,4	0,0000	0,0000	0,0009	0,0051

En la Tabla 8 se ha realizado el cálculo del coeficiente de correlación entre las funciones de Weibull creadas a partir de las series de viento con distintos porcentajes de valores faltantes, esto nos permitirá poder conocer de mejor manera si los datos obtenidos se encuentran relacionados entre sí.

Tabla 8. Coeficientes de correlación entre las Funciones de Distribución de Weibull con las series de viento imputadas con distintos porcentajes de datos

Coeficientes de correlación entre las Funciones de Distribución de Weibull con distintos porcentajes de datos disponibles	
	100%
100%	1
90%	1
60%	0,98806
30%	0,97701

Una observación importante en la Tabla 8 es que a pesar de que en la Tabla 5 los coeficientes de correlación son bajos entre las series imputadas, al momento de realizar el cálculo de la función de Weibull y comparar las correlaciones observamos que:

- Entre la serie con 0% de datos faltantes y la serie con el 10% de datos faltantes el coeficiente de correlación es igual a 1, por lo tanto, se puede decir que la relación si es lineal entre estas dos series.
- Entre la serie con 0% de datos faltantes y la serie con el 40% de datos faltantes el coeficiente de correlación es muy alto casi 1, a lo cual se puede decir que la relación si es lineal entre estas dos series.
- Entre la serie con 0% de datos faltantes y la serie con el 70% de datos faltantes el coeficiente de correlación es muy alto casi 1, a lo cual se puede decir que la relación si es lineal entre estas dos series.

Por lo tanto, nos damos cuenta de que no han afectado significativamente el que las series hayan tenido distintos porcentajes de valores faltantes ya que como se observa en la Tabla 8 los coeficientes de correlación son altos para las Funciones de Distribución de Weibull.

En la Figura 6 se muestra las funciones de Weibull para las series de viento con distintos porcentajes de valores faltantes.

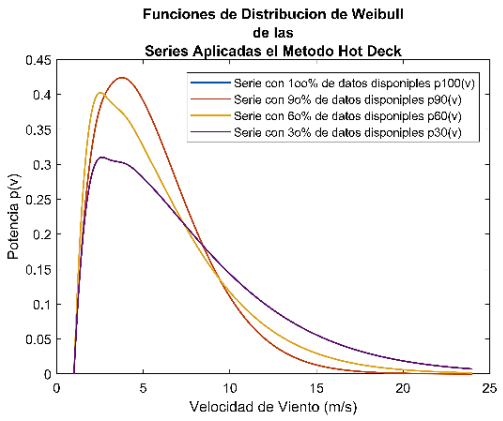

Figura 6. Funciones de Distribución de Weibull de las series aplicadas el Método Hot Deck.

Con los resultados de la Tabla 7 y la Figura 6, se puede estimar la producción de energía eléctrica, esto se puede realizar eligiendo un aerogenerador de entre varios catálogos, que se ajuste a los datos obtenidos para que pueda operar y generar electricidad a su máxima capacidad.

6. Conclusiones

Para garantizar la correcta imputación de datos faltantes de series incompletas, se debe trabajar con series que no posean muchos datos faltantes, ya que esto puede verse afectado de una manera significativa en el cálculo de valores medios ya que no se ajustan de manera real.

Se comprobó que el método Hot Deck Múltiple funciona perfectamente con series de datos que poseen distintos porcentajes de datos disponibles, ya que entre más donadores posea la serie más fácil será la imputación de los datos faltantes en las series de viento y por lo tanto se aproximará más a la serie real.

El método de relleno de datos es muy valedero, ya que aun para el caso de 30% de datos disponibles la función de Distribución de Weibull tan solo se pierde un 3% en potencia, es decir el método me permite ajustar los valores muy cercanos a los valores reales.

El nivel de incertidumbre entre las series de datos imputadas y las series de datos originales para la estimación de generación eléctrica mediante la función de distribución de Weibull es muy pequeña.

Como se pudo observar en la Tabla 8 las funciones de Weibull, no importó el porcentaje de datos faltantes en las series ya que los valores de correlación que existen entre cada serie no tuvieron una variación muy grande, haciendo que la estimación para generar energía eléctrica no varíe entre cada serie.

El método de Hot Deck junto con la función de Distribución de Weibull, son procedimientos matemáticos convenientes para trabajar con series de vientos y poder crear ecuaciones que permitan la estimación de producción de energía eléctrica como se puede observar en la Figura 6.

6.1. Trabajos futuros

Los resultados alcanzados en el presente trabajo puede ser el comienzo para llevar a cabo investigaciones relacionadas a las siguientes temáticas:

Calcular energía eléctrica a partir de la función de Weibull obtenida.

Comparar el método utilizado para la imputación de datos con otros métodos propuestos por otros autores y comprobar que método es el más destacado.

7. Referencias

[1] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, “A review of missing values handling methods on time-series data,” *2016 Int. Conf. Inf. Technol. Syst. Innov.*, no. October, 2016.

[2] Y. Tian, Q. Liu, Z. Hu, and Y. Liao, “Wind speed forecasting based on Time series - Adaptive Kalman filtering algorithm,” *FENDT 2014 - Proceedings, 2014 IEEE Far East Forum Nondestruct. Eval. New Technol. Appl. Increasingly Perfect NDT/E*, pp. 315–319, 2014.

[3] N. Rab, F. Leimgruber, and T. Esterl, “Synthetic wind speed time series with Markov and ARMA models: Comparison for different use cases,” *Int. Conf. Eur. Energy Mark. EEM*, vol. 2015-Augus, pp. 1–5, 2015.

[4] D. Jijón, J. Constante, M. Moya, and G. Guerrón, “Métodos para homogenizar y rellenar datos de viento de la torre meteorológica del Parque Eólico Villonaco en Loja-Ecuador,” *Av. en Ciencias e Ing.*, vol. 7, no. 2, pp. 44–52, 2016.

[5] Álvarez O, Maldonado J, Montaño T, and Tenechagua L, "Análisis Climático de la Velocidad del Viento en la Región Sur del Ecuador," *Rev. Politécnica*, vol. 35, no. 3, 2014.

[6] H. O. Alvarez, T. Montaño, E. Quentin, J. Maldonado, and J. Solano, "Homogeneización de series de velocidad del viento mensuales en las estaciones meteorológicas del INAMHI en Loja, Ecuador," *Rev. Climatol.*, vol. 13, pp. 35–44, 2013.

[7] J. Heckenbergerova, P. Musilek, and J. Marek, "Analysis of wind speed and power time series preceding wind ramp events," *Proc. 2014 15th Int. Sci. Conf. Electr. Power Eng. EPE 2014*, no. 1, pp. 279–283, 2014.

[8] I. Colak, S. Sagiroglu, M. Yesilbudak, E. Kabalci, and H. Ibrahim Bulbul, "Multi-time series and-time scale modeling for wind speed and wind power forecasting part II: Medium-term and long-term applications," *2015 Int. Conf. Renew. Energy Res. Appl. ICRERA 2015*, vol. 5, pp. 215–220, 2015.

[9] S. Cartaya, S. Zurita, and V. Montalvo, "Métodos de ajuste y homogenización de datos climáticos para determinar índice de humedad de Lang en la provincia de Manabí, Ecuador," *La Técnica Rev. las Agrociencias. ISSN 2477-8982*, no. 16, p. 94, 2016.

[10] M. de la paz Almeida Román, "Instructivos De Procesamiento De Información Hidrometeorológica," p. 299, 2010.

[11] F. E. Castillo and F. C. Sentis, "Agrometereología," p. 517, 2001.

[12] A. M. Ferreira, "Metodología de análisis e imputación de datos faltantes en series de velocidad del viento," *VI Congr. Galego Estatística e Investig. Operacións*, p. 8, 2005.

[13] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Inf. Sci. (Ny).*, vol. 233, pp. 25–35, 2013.

[14] Y. S. Afrianti, S. W. Indratno, and U. S. Pasaribu, "Imputation algorithm based on copula for missing value in timeseries data," *Proc. 2014 2nd Int. Conf. Technol. Informatics, Manag. Eng. Environ.*, pp. 252–257, 2014.

[15] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychol. Methods*, vol. 7, no. 2, pp. 147–177, 2002.

[16] I. Song, Y. Yang, J. Im, T. Tong, H. Ceylan, and I.-H. Cho, "Impacts of Fractional Hot-Deck Imputation on Learning and Prediction of Engineering Data," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2019.

[17] S. P. Susanti and F. N. Azizah, "Imputation of missing value using dynamic Bayesian network for multivariate time series data," *Proc. 2017 Int. Conf. Data Softw. Eng. ICoDSE 2017*, vol. 2018-Janua, pp. 1–5, 2018.

[18] J. Sessa and D. Syed, "Techniques to deal with missing data," *Int. Conf.*

Electron. Devices, Syst. Appl., pp. 1–4, 2017.

[19] M. Pattanodom, N. Iam-On, and T. Boongoen, "Clustering data with the presence of missing values by ensemble approach," *2016 2nd Asian Conf. Def. Technol. ACDT 2016*, pp. 151–156, 2016.

[20] L. Malambo and C. D. Heatwole, "A Multitemporal Profile-Based Interpolation Method for Gap Filling Nonstationary Data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 252–261, 2016.

[21] X. Jingwen, Z. Wanchang, and L. Chuansheng, "A novel method for filling the depressions in massive DEM data," *Int. Geosci. Remote Sens. Symp.*, pp. 4080–4083, 2007.

[22] P. E. Tissot, W. B. Zhu, S. Duff, M. Rink, J. Rizzo, and D. Martin, "Development, assessment and implemenation of an automated gap filling method for tide stations with dual water level sensors," *2014 Ocean. - St. John's, Ocean. 2014*, pp. 1–10, 2015.

[23] P. Valarmathie and K. Dinakaran, "An efficient technique for missing value imputation in microarray gene expression data," *Proc. ICCCS 2014 - IEEE Int. Conf. Comput. Commun. Syst.*, no. Icccs 114, pp. 73–80, 2014.

[24] J. Chen, F. Miao, H. Lu, and Y. Duan, "Study on automatic vectorization of thematic map based on Neighborhood Filling," *2011 Int. Conf. Remote Sensing, Environ. Transp. Eng. RSETE 2011 - Proc.*, pp. 6321–6324, 2011.

[25] L. H. Yang and T. Sen Zhan, "One treatment of missing data based on control-points optimization smoothing method," *World Autom. Congr. Proc.*, pp. 1–4, 2012.

[26] J. Song, Q. Yu, and Y. Guo, "The Data Integrity Error Repair Method for Filling Missing Values with External Data," 2019.

[27] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, John Wiley. New York, 1987.

[28] J. K. Kim and W. Fuller, "Fractional hot deck imputation," *Biometrika*, vol. 91, no. 3, pp. 559–578, 2004.

[29] S. J. Cranmer and J. Gill, "We have to be discrete about this: A non-parametric imputation technique for missing categorical data," *Br. J. Polit. Sci.*, vol. 43, no. 2, pp. 425–449, 2013.

[30] Z. H. Wang, "Numeric missing value's hot deck imputation based on cloud model and association rules," *2nd Int. Work. Educ. Technol. Comput. Sci. ETCS 2010*, vol. 1, pp. 238–241, 2010.

[31] J. A. La Cal Herrera, "La energía eólica en Sierra Mágina," *Sumuntán*, vol. 23, pp. 127–151, 2006.

[32] G. Kalton and L. Kish, "Some Efficient Random Imputation Methods," *Commun. Stat. - Theory Methods*, vol. 13, no. 16, pp. 1919–1939, 1984.

[33] A. S. Cranmer, J. Gill, N. Jackson, A. Murr, D.

Armstrong, and M. D. Armstrong, "Package 'hot.deck,'" pp. 1–10, 2016.

[34] INAMHI, "RED DE ESTACIONES AUTOMÁTICAS HIDROMETEOROLÓGICAS, " 2019. [Online]. Available: http://186.42.174.236/InamhiEmas/.

[35] W. Mogrovejo and F. Quezada, "Aprovechamiento del recurso eólico en el ecuador mediante la estimación de producción eléctrica y análisis de viabilidad económica," *ResearchGate*, p. 12, 2016.

[36] G. Shi, Z. Wang, M. Zhu, X. Cai, and L. Yao, "Variable speed control of series-connected DC wind turbines based on generalized dynamic model," *IET Conf. Publ.*, vol. 2013, no. 623 CP, pp. 3–8, 2013.

[37] O. S. Vallejos and D. M. Aedo, "Weibull-fit.xla: Programa para el ajuste óptimo de la función de densidad de probabilidad weibull de tres parámetros," *Inf. Tecnol.*, vol. 21, no. 1, pp. 91–99, 2010.

[38] J. Serrano Rico, "Comparación de métodos para determinar los parámetros de Weibull para la generación de energía eólica," *Sci. Tech.*, vol. 18, no. 2, pp. 315–320, 2013.

[39] K. Ulgen and A. Hepbasli, "Determination of Weibull parameters for wind energy analysis of Izmir, Turkey," *Int. J. Energy Res.*, vol. 26, no. 6, pp. 495–506, 2002.

[40] P. Ramírez and J. A. Carta, "Influence of the data sampling interval in the estimation of the parameters of the Weibull wind speed probability density distribution: A case study," *Energy Convers. Manag.*, vol. 46, no. 15–16, pp. 2419–2438, 2005.

[41] A. Genc, M. Erisoglu, A. Pekgor, G. Oturanc, A. Hepbasli, and K. Ulgen, "Estimation of wind power potential using weibull distribution," *Energy Sources*, vol. 27, no. 9, pp. 809–822, 2005.

[42] C. Gavriluta, S. Spataru, I. Mosincat, C. Citro, I. Candela, and P. Rodriguez, "Complete methodology on generating realistic wind speed profiles based on measurements," *Renew. Energy Power Qual. J.*, vol. 1, no. 10, pp. 1757–1762, 2012.

[43] J. A. Saunders, N. Morrow-Howell, E. Spitznagel, P. Doré, E. K. Proctor, and R. Pescarino, "Imputing missing data: A comparison of methods for social work researchers," *Soc. Work Res.*, vol. 30, no. 1, pp. 19–31, 2006.

[44] A. T. Sree Dhevi, "Imputing missing values using Inverse Distance Weighted Interpolation for time series data," *6th Int. Conf. Adv. Comput. ICoAC 2014*, pp. 255–259, 2015.

[45] K. Strike, K. El Emam, and N. Madhavji, "Software cost estimation with incomplete data," *IEEE Trans. Softw. Eng.*, vol. 27, no. 10, pp. 890–908, 2001.

[46] D. W. Joenssen and U. Bankhofer, "Hot deck methods for imputing missing data: The effects of limiting donor usage,"

Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7376 LNAI, no. September, pp. 63–75, 2012.

[47] C. R. Ranganathan, B. Centre, and T. Nadu, "Estimation of Wind Power Availability in Tamil Nadu," *Renew. Energy*, vol. 1, no. 3, pp. 429–434, 1991.

7.1. Matriz de Estado del Arte

Tabla 9. Matriz de estado del arte

RELLENO DE DATOS DE VELOCIDADES DE VIENTO MEDIANTE LA APLICACIÓN DE MÉTODO DE HOT DECK PARA LA ESTIMACIÓN DE PRODUCCIÓN DE ENERGÍA ELÉCTRICA EN BASE AL RECURSO EÓLICO.																						
	DATOS			TEMÁTICA					FORMULACIÓN DEL PROBLEMA			RESTRICCIONES DEL PROBLEMA			PROPUESTAS PARA EL PROBLEMA					SOLUCIÓN AL PROBLEMA		
ITEM	AÑO	TÍTULO DEL ARTÍCULO	CITAS	RELLENO DE DATOS FALTANTES	HOMOGENIZACIÓN DE DATOS	RECURSO EÓLICO	DATOS FALTANTES	FUNCIÓN DE DISTRIBUCIÓN DE WEIBULL	SERIES DE DATOS COMPLETADAS	ESTIMACIÓN DE GENERACIÓN DE ENERGÍA ELÉCTRICA	PERDIDA DE DATOS ALEATORIOS	TIPOS DE VARIABLES DE LOS DATOS PERDIDOS	PORCENTAJES DE DATOS PERDIDOS	DATA NO HOMOGENIZADA	MÉTODO REGRESION	MÉTODO INPUTACION MULTIPLE	MÉTODO LISTWISE	METODO HOT DECK	MÉTODO IMPUTACIÓN MEDIA	APLICACIÓN MÉTODO HOT DECK MULTIPLE	HOMOGENIZACIÓN DE DATOS	APLICACIÓN DE LA FUNCIÓN DE DISTRIBUCIÓN DE WEIBULL
1	2016	Aprovechamiento del recurso eólico en el ecuador mediante la estimación de producción eléctrica y análisis de viabilidad económica	0			✠		✠		✠												✠
2	2006	La energía eólica en Sierra Mágina	4			✠				✠												
3	2013	Homogeneización de series de velocidad del viento mensuales en las estaciones meteorológicas del INAMHI en Loja, Ecuador	4		✠	✠								✠	✠						✠	
4	2016	A review of missing values handling methods on time-series data	17	✠			✠				✠	✠			✠	✠	✠		✠			

5	2014	Evaluación del potencial solar y eólico del campus centra de la Universidad Industrial de Santander y la ciudad de Bucaramanga, Colombia	37	✠		✠				✠												
6	2016	Métodos de ajuste y homogenización de datos climáticos para determinar índice de humedad de Lang en la provincia de Manabí, Ecuador	6		✠	✠	✠						✠			✠			✠	✠	✠	
7	2010	Instructivos De Procesamiento De Información Hidrometeorológica	2			✠			✠													
8	2006	Imputing missing data: A comparison of methods for social work researchers	100	✠			✠		✠		✠	✠	✠		✠	✠	✠		✠			
9	2005	Metodología de análisis e imputación de datos faltantes en series de velocidad del viento	5	✠		✠	✠		✠		✠	✠	✠		✠				✠			
10	2013	A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm	61	✠			✠		✠		✠	✠										
11	2014	Imputation algorithm based on copula for missing value in timeseries data	4	✠			✠		✠		✠	✠					✠	✠	✠			
12	2002	Missing data: Our view of the state of the art	5239	✠			✠		✠		✠	✠			✠		✠		✠			
13	1987	Statistical Analysis With Missing Data	425	✠			✠		✠		✠	✠							✠	✠		
14	2004	Fractional hot deck imputation	56	✠			✠		✠		✠	✠	✠									
15	2013	We have to be discrete about this: A non-parametric imputation technique for missing categorical data	56	✠			✠		✠		✠	✠	✠		✠		✠	✠		✠		
16	2010	Weibull-fit.xla: Programa para el ajuste óptimo de la función de densidad de probabilidad Weibull de tres parámetros	1					✠		✠												✠
17	2013	Comparación de métodos para determinar los parámetros de Weibull para la generación de energía eólica	15			✠		✠		✠												✠
18	2002	Determination of Weibull parameters for wind energy analysis of Izmir, Turkey	58			✠		✠		✠												✠
19	2005	Influence of the data sampling interval in the estimation of the parameters of the Weibull wind speed probability density distribution: A case study	126			✠		✠		✠												✠
20	1991	Estimation of Wind Power Availability in Tamil Nadu	12			✠		✠		✠												✠
21	2012	Complete methodology on generating realistic wind speed profiles based on measurements	32			✠		✠		✠												✠

22	1984	Some Efficient Random Imputation Methods	95	✠			✠		✠	✠					✠							
23	2011	A Hot-Deck Multiple Imputation Procedure for Gaps in Longitudinal Recurrent Event Histories	16	✠		✠	✠			✠		✠	✠							✠		
24	2010	A Review of Hot Deck Imputation for Survey Non-response	251	✠		✠	✠		✠	✠	✠	✠			✠					✠		
25	1997	Analysis of Incomplete Multivariate Data	3248	✠		✠	✠		✠		✠	✠			✠					✠		
26	2015	Analysis of wind energy prospect for power generation by three Weibull distribution methods	42			✠		✠		✠		✠			✠							✠
27	2013	Comparison of missing value imputation methods in time series: the case of Turkish meteorological data	85			✠		✠					✠		✠							✠
28	1993	Data analysis using hot deck multiple imputation	68	✠			✠		✠		✠	✠	✠							✠		
29	2005	Hot Deck Imputation for the Response Model	83	✠			✠		✠		✠	✠	✠		✠					✠		
30	2019	Hot Deck Multiple Imputation for Handling Missing Accelerometer Data	1	✠			✠		✠		✠	✠	✠							✠		
31	2018	Multiple hot-deck imputation for network inference from RNA sequencing data	6	✠			✠		✠		✠	✠	✠		✠					✠		
32	2001	Software cost estimation with incomplete data	43	✠			✠		✠							✠		✠				
33	2012	Hot deck methods for imputing missing data: The effects of limiting donor usage	6	✠			✠		✠							✠		✠				
			CANTIDAD:	19	2	17	20	9	18	13	15	17	11	1	13	5	5	4	7	10	2	9

7.2. Resumen de indicadores

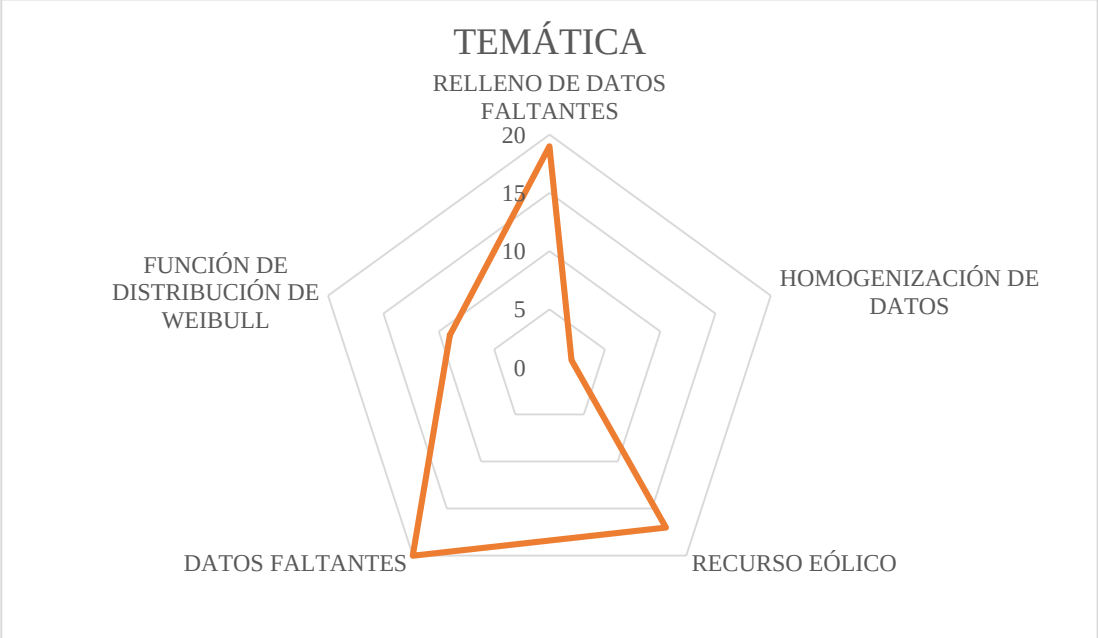

Figura 7. Resumen e indicador de la temática - Estado del arte

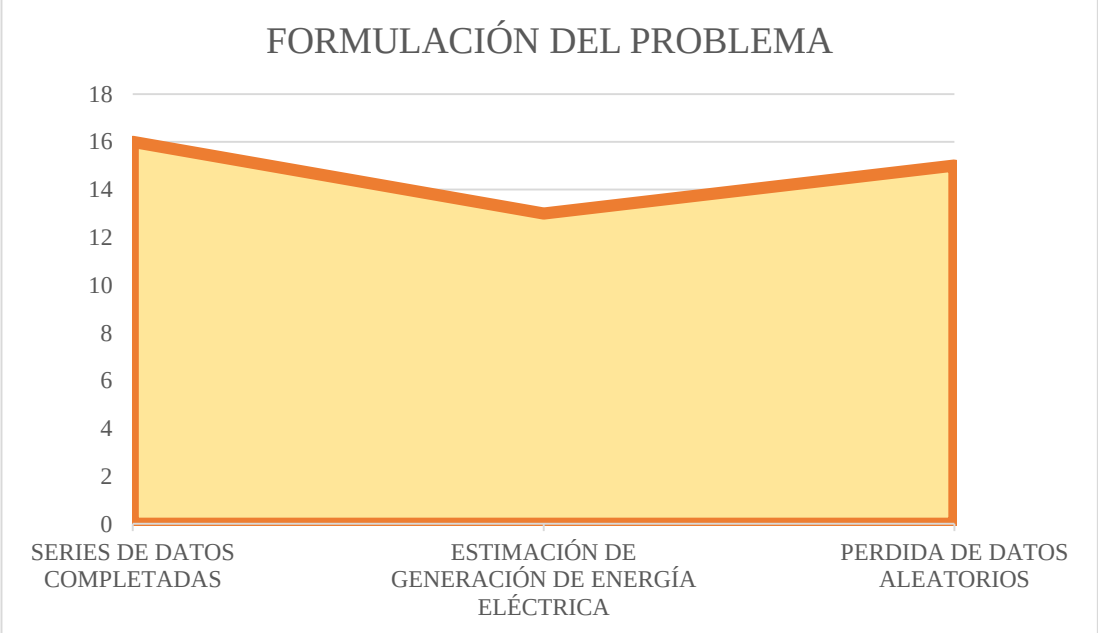

Figura 8. Indicador de formulación del problema - Estado del arte

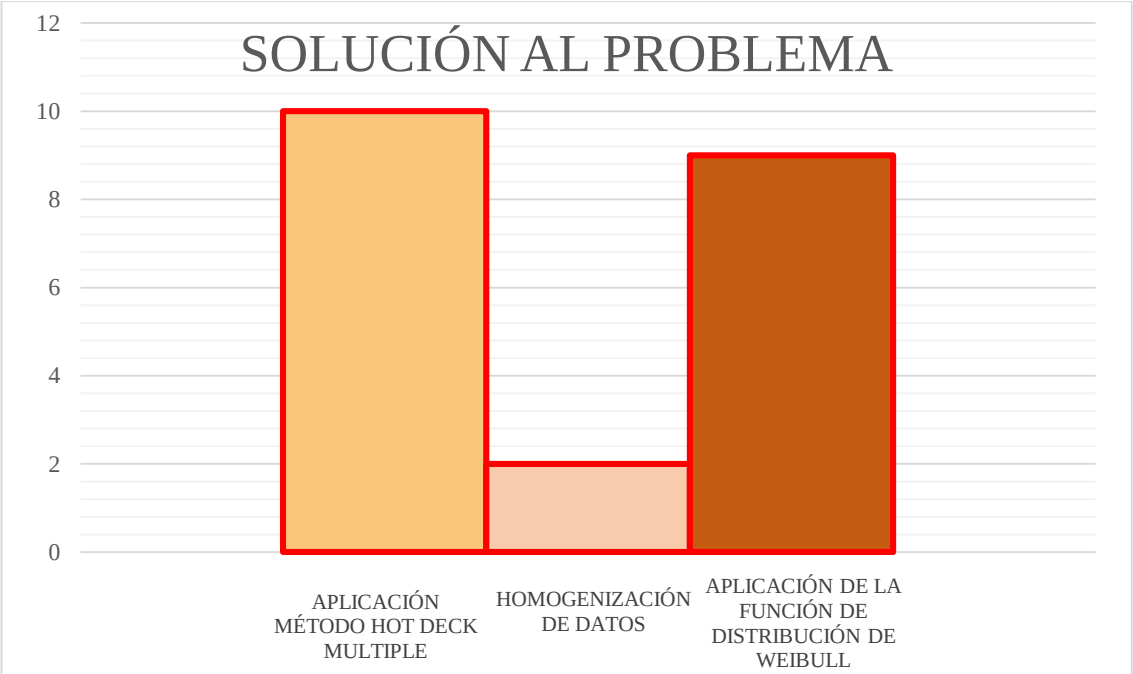

Figura 9. Indicador de solución - Estado del arte