UNIVERSIDAD DE OVIEDO

Departamento de Ingeniería Eléctrica, Electrónica,
de Computadores y Sistemas

Programa de Doctorado: Ingeniería Eléctrica y Electrónica

TESIS DOCTORAL

IMPUTACIÓN DE DATOS FALTANTES EN REDES DE DISTRIBUCIÓN DE BAJA TENSIÓN: APLICACIÓN A EDIFICIOS DE PÚBLICA CONCURRENCIA

Mª. Concepción Crespo Turrado

Directores:

D. Francisco Javier de Cos Juez
D. Manuel García Melero

Oviedo, 2018

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: Imputación de datos faltantes en redes de distribución de baja tensión: aplicación a edificios de pública concurrencia	Inglés: Imputation of missing data in the field of energy management in buildings of public attendance

2.- Autor	
Nombre: María Concepción Crespo Turrado	DNI/Pasaporte/NIE:
Programa de Doctorado: Ingeniería Eléctrica y Electrónica (interuniversitario)	
Órgano responsable: Dpto. Ingeniería Eléctrica, Electrónica, de Computadores y Sistemas	

RESUMEN (en español)

Actualmente, la toma de datos es un proceso clave en el estudio de las instalaciones de energía eléctrica; no solo desde un punto de vista de calidad de la distribución, sino también desde un punto de vista económico, así pues, se hace necesario conocer cómo y dónde se consume energía en una instalación. En este contexto, la falta de datos de cualquiera de las principales variables eléctricas objeto de medida (tensión fase-neutro, tensión fase-fase, corriente en cada fase o factor de potencia) en cualquier momento, afecta negativamente al estudio realizado. Cuando esto ocurre, debe realizarse un proceso de imputación de datos para sustituir los datos que faltan por valores estimados. Para el desarrollo de la presente tesis doctoral, se inicia el trabajo demostrando la viabilidad de las técnicas de imputación como herramientas adecuadas para la gestión, monitorización y control de instalaciones de energía eléctrica.

Para llevar a cabo dicha demostración se emplean los datos recogidos en una instalación eléctrica tipo (Radiación solar recibida en varias plantas fotovoltaicas de Galicia bajo distintas condiciones atmosféricas) y una serie de algoritmos de imputación de alta eficiencia ya consolidados dentro de otros sectores de actividad como son: Inverse Distance Weighting (IDW), Multiple Linear Regressions Models (MLR Models) y Multiple Imputation by Chained Equations (MICE). En esta primera fase, no solo se demuestra la viabilidad del uso de las citadas técnicas como herramientas capaces de mejorar los sistemas de gestión y monitorización de una instalación, sino que también se evalúa la eficiencia de los citados algoritmos y se selecciona el más eficiente de ellos, MICE, para su uso como elemento de comparación con los futuros algoritmos a desarrollar en la presente tesis doctoral.

Posteriormente se presentan los resultados obtenidos al aplicar un nuevo método de imputación en la instalación objeto de estudio (Edificio Severo Ochoa de la Universidad de Oviedo). Este nuevo método, desarrollado, y al que se ha denominado AAA (Adaptive Assignation Algorithm), está basado en el algoritmo inteligente conocido como Multivariate Adaptive Regression Splines (MARS). Tal como ya se adelantó los resultados obtenidos con el nuevo método son comparados con los obtenidos tras la

aplicación del algoritmo de referencia MICE demostrándose las ventajas de AAA en términos de precisión y fiabilidad en la tarea propuesta. El estudio pormenorizado del rendimiento de la nueva metodología desarrollada (AAA) demostró que la fiabilidad del nuevo algoritmo decrecía sustancialmente cuando el número de variables faltantes en un mismo registro temporal era muy alto, lo que da pie a buscar nuevas metodologías de imputación, así como la hibridación de las mismas en la fase final del trabajo con objeto de superar la citada limitación.
Finalmente, con objeto de obtener imputaciones fiables en situaciones en las que el número de faltantes es muy elevado se desarrolló una nueva metodología basada en redes neuronales auto organizadas SOM (Self-Organized Maps) que se demostró más eficiente en ese tipo de situaciones muy adversas siendo, sin embargo, su rendimiento peor cuando el número de faltantes no es tan elevado.
Como la cantidad de faltantes por registro depende en gran medida de problemas en la red y en los equipos de medida y que estos no son a priori predecibles, es posible encontrarse en un corto espacio de tiempo con situaciones donde el número de datos faltantes varía considerablemente de un registro a otro. Por ello, la propuesta final de esta tesis se basa en la hibridación inteligente (o uso selectivo) de los distintos algoritmos desarrollados. Dicha hibridación se ha evaluado haciendo un uso combinado tanto de AAA con MICE como de AAA con SOM, habiéndose mostrado esta combinación como más adecuada independientemente del número de faltantes acontecido.

RESUMEN (en Inglés)

Nowadays, data measurement and collection are key processes in the study of the electric power system. Not only for power quality purposes, but also from an economic point of view, it is necessary to know how and where energy is consumed in a facility. In this context, missing data of any of the main electrical variables under measurement (phase-to-neutral voltage, phase-to-phase voltage, phase current or power factor) may negatively affect the underway study. When this takes place, a data imputation process must be conducted in order to replace the missing data with estimated values.
In the development of the present doctoral thesis, the effort is initially directed to prove the feasibility of imputation techniques as appropriate tools for the management, monitoring and control of electric power systems. To carry out this research, a data logger has been used in an electrical installation (global solar radiation received in many photovoltaic plants in Galicia under different atmospheric conditions) and a series of high efficiency imputation algorithms already consolidated within other activity sectors. Among this methods, it is important to highlight the Inverse Distance Weighting (IDW), Multiple Linear Regressions Models (MLR Models) and Multiple Imputation by Chained Equations (MICE). In this first stage of the thesis, the viability of using the aforementioned techniques as tools capable of enhancing the management and monitoring systems of an electrical installation has been demonstrated. Moreover, the efficiency of the said algorithms is evaluated and the one providing the best performance, the MICE, is used as a benchmark for comparison with the new algorithms developed in the present doctoral thesis.
Subsequently, the results of applying a new imputation method to the installation under study, the Severo Ochoa Building at the University of Oviedo, are presented. This new method, developed in the present work, and which has been called AAA (Adaptive

Assignation Algorithm), is based on an intelligent algorithm known as Multivariate Adaptive Regression Splines (MARS). As already mentioned, the results obtained with the new method are compared with those reached by applying the MICE benchmark algorithm. The advantages of AAA in terms of accuracy and reliability in the proposed task are clearly demonstrated. A detailed study of the performance of the new methodology (AAA) showed that the reliability of the new algorithm decreased substantially when the number of missing variables in the same time register is very high. This fact led to the development of new imputation technologies and their hybridization in the final stage of the work, in order to overcome the aforementioned limitation. The new methodology was developed based on self-organized neural networks SOM (Self-Organized Maps). It proved to be more efficient in very adverse situations being, however, its performance worse when the number of missing variables is not so high.

Given that the number of missing data per register depends largely on problems in the network and measurement equipment which are non-predictable or controllable, it is possible to find a short term with situations where the number of missing data changes considerably from a register to another. Therefore, the final proposal of this thesis is based on intelligent hybridization (or selective use) of different algorithms. This hybridization has been evaluated by making a combined use of both AAA with MICE and AAA with SOM. The latter combination results to be the most appropriate regardless of the number of missing events.

FORMULARIO RESUMEN DE TESIS POR COMPENDIO

1.- Datos personales solicitante	
Apellidos: Crespo Turrado	Nombre: María Concepción

Curso de inicio de los estudios de doctorado	2014/15

	SI	NO
Acompaña acreditación por el Director de la Tesis de la aportación significativa del doctorando	X	
Acompaña memoria que incluye		
Introducción justificativa de la unidad temática y objetivos	X	
Copia completa de los trabajos *	X	
Resultados/discusión y conclusiones	X	
Informe con el factor de impacto de la publicaciones	X	
Se acompaña aceptación de todos y cada uno de los coautores a presentar el trabajo como tesis por compendio	X	
Se acompaña renuncia de todos y cada uno de los coautores a presentar el trabajo como parte de otra tesis de compendio	X	

* Ha de constar el nombre y adscripción del autor y de todos los coautores asi como la referencia completa de la revista o editorial en la que los trabajos hayan sido publicados o aceptados en cuyo caso se aportará justificante de la aceptación por parte de la revista o editorial

FOR-MAT-VOA-033

Artículos, Capítulos, Trabajos

Trabajo, Artículo 1

Titulo (o título abreviado)	**Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions**
Fecha de publicación	29 de octubre de 2014
Fecha de aceptación	21 de octubre de 2014
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	Journal of Citation Report
Factor de impacto	2,245

Coautor2	X Doctor	□ No doctor .	María del Carmen Meizoso López
Coautor3	X Doctor	□ No doctor .	Fernando Sánchez Lasheras
Coautor4	X Doctor	□ No doctor .	Benigno Antonio Rodríguez Gómez
Coautor5	X Doctor	□ No doctor .	José Luis Calvo Rollé
Coautor6	X Doctor	□ No doctor .	Francisco Javier de Cos Juez

Trabajo, Artículo 2

Titulo (o título abreviado)	**A New Missing Data Imputation Algorithm Applied to Electrical Data Loggers**
Fecha de publicación	10 de diciembre de 2015
Fecha de aceptación	7 de diciembre de 2015
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	Journal of Citation Report
Factor de impacto	2,033

Coautor2	**X** Doctor	□ No doctor .	Indique nombre y apellidos	Fernando Sánchez Lasheras
Coautor3	**X** Doctor	□ No doctor .	Indique nombre y apellidos	José Luis Calvo-Rollé
Coautor4	□ Doctor	**X** No doctor .	Indique nombre y apellidos	Andrés José Piñón-Pazos
Coautor5	**X** Doctor	□ No doctor .	Indique nombre y apellidos	Francisco Javier de Cos Juez

Trabajo, Artículo 3

Titulo (o título abreviado)	**A Hybrid Algorithm for Missing Data Imputation and Its Application to Electrical Data Loggers**
Fecha de publicación	10 septiembre 2016
Fecha de aceptación	7 septiembre 2016
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	Journal of Citation Report
Factor de impacto	2,677

Coautor2	**X** Doctor	□ No doctor .	Indique nombre y apellidos	Fernando Sánchez Lasheras
Coautor3	**X** Doctor	□ No doctor .	Indique nombre y apellidos	José Luis Calvo-Rollé
Coautor4	□ Doctor	**X** No doctor .	Indique nombre y apellidos	Andrés-José Piñón-Pazos
Coautor5	**X** Doctor	□ No doctor .	Indique nombre y apellidos	Manuel García Melero
Coautor6	**X** Doctor	□ No doctor .	Indique nombre y apellidos	Francisco Javier de Cos Juez

UNIVERSIDAD DE OVIEDO

Departamento de Ingeniería Eléctrica, Electrónica,
de Computadores y Sistemas

Programa de Doctorado: Ingeniería Eléctrica y Electrónica

TESIS DOCTORAL

IMPUTACIÓN DE DATOS FALTANTES EN REDES DE DISTRIBUCIÓN DE BAJA TENSIÓN: APLICACIÓN A EDIFICIOS DE PÚBLICA CONCURRENCIA

Mª. Concepción Crespo Turrado

Directores:

D. Francisco Javier de Cos Juez
D. Manuel García Melero

Oviedo, 2018

UNIVERSIDAD DE OVIEDO

Departamento de Ingeniería Eléctrica, Electrónica,
de Computadores y Sistemas

TESIS DOCTORAL

IMPUTACIÓN DE DATOS FALTANTES EN REDES DE DISTRIBUCIÓN DE BAJA TENSIÓN: APLICACIÓN A EDIFICIOS DE PÚBLICA CONCURRENCIA

Mª. Concepción Crespo Turrado

Directores:
D. F. Javier de Cos Juez
D. Manuel García Melero

Oviedo, 2018

AGRADECIMIENTOS

A mis directores de tesis, Manuel García Melero, del Departamento de Ingeniería Eléctrica, Electrónica, de Computadores y Sistemas y F. Javier de Cos, del Departamento de Explotación y Prospección de Minas de la Universidad de Oviedo. Por todo lo que me han enseñado, por su apoyo incondicional y sobre todo por su paciencia y dedicación.

A los coautores de los artículos que han servido para desarrollar esta tesis por compendio de publicaciones. Sin su trabajo no hubiera sido posible.

A mis compañeros del Servicio de Mantenimiento e Instalaciones de la Universidad de Oviedo. Por su inestimable ayuda y colaboración en la elaboración de la parte experimental.

Por último, a mi familia y amigos. Por apoyarme en los momentos duros y alentarme a seguir adelante.

ACKNOWLEDGMENTS

To my thesis directors, Manuel García Melero, from the Electrical Engineering Department and F. Javier de Cos, from the Department of Mining at the University of Oviedo. For all they have taught me, for their unconditional support and above all for their patience and dedication.

To the co-authors of the articles that have served to develop this thesis by compendium of publications. Without their work would not have been possible.

To my colleagues in the Maintenance and Installations Service of the University of Oviedo. For their invaluable help and collaboration in the preparation of the experimental part.

Finally, to my family and friends. For supporting me in hard times and encouraging me to keep going.

RESUMEN

Hoy en día, la toma de datos es un proceso clave en el estudio de las instalaciones de energía eléctrica; no solo desde un punto de vista de calidad de la distribución, en la búsqueda de armónicos y en la falta de equilibrio entre fases, sino también desde un punto de vista económico, se hace necesario conocer cómo y dónde se consume energía en una instalación. En este contexto, la falta de datos de cualquiera de las principales variables eléctricas objeto de medida (tensión fase-neutro, tensión fase-fase, corriente en cada fase o factor de potencia) en cualquier momento, afecta negativamente al estudio realizado. Cuando esto ocurre, debe realizarse un proceso de imputación de datos para sustituir aquellos que faltan por valores estimados.

Para el desarrollo de la presente Tesis Doctoral, se inicia el trabajo demostrando la viabilidad de las técnicas de imputación como herramientas adecuadas para la gestión, monitorización y control de instalaciones de energía eléctrica.

Para llevar a cabo dicha demostración se emplean los datos recogidos en una instalación eléctrica tipo (radiación solar recibida en varias plantas fotovoltaicas de Galicia bajo distintas condiciones atmosféricas) y una serie de algoritmos de imputación de alta eficiencia, ya consolidados dentro de otros sectores de actividad. Entre estos algoritmos destacan los denominados como Inverse Distance Weighting (IDW), Multiple Linear Regressions Models (MLR Models) y Multiple Imputation by Chained Equations (MICE). En esta primera fase del trabajo, no solo se demuestra la viabilidad del uso de las citadas técnicas como herramientas capaces de mejorar los sistemas de gestión y monitorización de una instalación eléctrica, sino que también se evalúa la eficiencia de los citados algoritmos y se selecciona el más eficiente de ellos, MICE, para su uso como elemento de comparación con los futuros algoritmos a desarrollar en la presente tesis doctoral. [Crespo 1].

Posteriormente, se presentan los resultados de aplicar un nuevo método de imputación en la instalación objeto de estudio (Edificio Severo Ochoa de la Universidad de Oviedo). Este nuevo método, desarrollado en el presente trabajo, y al que se ha denominado AAA (Adaptive Assignation Algorithm), está basado en el algoritmo inteligente conocido como Multivariate Adaptive Regression Splines (MARS). Como ya se adelantó, los resultados obtenidos con el nuevo método son comparados con los obtenidos tras la aplicación del algoritmo de referencia MICE, demostrándose las ventajas de AAA en términos de precisión y fiabilidad en la tarea propuesta. El estudio pormenorizado del rendimiento de la nueva metodología desarrollada (AAA) demostró que la fiabilidad del nuevo algoritmo decrecía sustancialmente cuando el número de variables faltantes en un mismo registro temporal era muy alto. Esto dio pie al desarrollo de nuevas metodologías de imputación y a la hibridación de las mismas, lo cual, permitió superar la citada limitación en etapas posteriores [Crespo 2].

Finalmente, con objeto de obtener imputaciones fiables en situaciones en las que el número de faltantes era muy elevado, se desarrolló una nueva metodología basada en redes

neuronales auto-organizadas SOM (Self-Organized Maps). Esta técnica se demostró más eficiente en tales situaciones adversas; sin embargo, su rendimiento se reveló peor cuando el número de faltantes no era tan elevado.

Dado que la cantidad de faltantes por registro depende en gran medida de problemas en la red y en los equipos de medida, y que dichos problemas no son a priori predecibles o controlables, es posible encontrarse en un corto espacio de tiempo con situaciones donde el número de datos faltantes varía considerablemente de un registro a otro. Por ello, la propuesta final de esta tesis se basa en la hibridación inteligente (o uso selectivo) de los distintos algoritmos desarrollados. Dicha hibridación se ha evaluado haciendo un uso combinado tanto de AAA con MICE [Crespo 3] como de AAA con SOM [Crespo 4], resultando esta última combinación la más adecuada independientemente del número de faltantes acontecido.

ABSTRACT

Nowadays, data measurement and collection are key processes in the study of the electric power system. Not only for power quality purposes, but also from an economic point of view, it is necessary to know how and where energy is consumed in a facility. In this context, missing data of any of the main electrical variables under measurement (phase-to-neutral voltage, phase-to-phase voltage, phase current or power factor) may negatively affect the underway study. When this takes place, a data imputation process must be conducted in order to replace the missing data with estimated values.

In the development of the present doctoral thesis, the effort is initially directed to prove the feasibility of imputation techniques as appropriate tools for the management, monitoring and control of electric power systems. To carry out this research, a data logger has been used in an electrical installation (global solar radiation received in many photovoltaic plants in Galicia under different atmospheric conditions) and a series of high efficiency imputation algorithms already consolidated within other activity sectors. Among this methods, it is important to highlight the Inverse Distance Weighting (IDW), Multiple Linear Regressions Models (MLR Models) and Multiple Imputation by Chained Equations (MICE). In this first stage of the thesis, the viability of using the aforementioned techniques as tools capable of enhancing the management and monitoring systems of an electrical installation has been demonstrated. Moreover, the efficiency of the said algorithms is evaluated and the one providing the best performance, the MICE, is used as a benchmark for comparison with the new algorithms developed in the present doctoral thesis.

Subsequently, the results of applying a new imputation method to the installation under study, the Severo Ochoa Building at the University of Oviedo, are presented. This new method, developed in the present work, and which has been called AAA (Adaptive Assignation Algorithm), is based on an intelligent algorithm known as Multivariate Adaptive Regression Splines (MARS). As already mentioned, the results obtained with the new method are compared with those reached by applying the MICE benchmark algorithm. The advantages of AAA in terms of accuracy and reliability in the proposed task are clearly demonstrated. A detailed study of the performance of the new methodology (AAA) showed that the reliability of the new algorithm decreased substantially when the number of missing variables in the same time register is very high. This fact led to the development of new imputation technologies and their hybridization in the final stage of the work, in order to overcome the aforementioned limitation. The new methodology was developed based on self-organized neural networks SOM (Self-Organized Maps). It proved to be more efficient in very adverse situations being, however, its performance worse when the number of missing variables is not so high.

Given that the number of missing data per register depends largely on problems in the network and measurement equipment which are non-predictable or controllable, it is possible to find a short term with situations where the number of missing data changes

considerably from a register to another. Therefore, the final proposal of this thesis is based on intelligent hybridization (or selective use) of different algorithms. This hybridization has been evaluated by making a combined use of both AAA with MICE and AAA with SOM. The latter combination results to be the most appropriate regardless of the number of missing events.

CONTENIDO

1. PRESENTACIÓN DE LA TESIS DOCTORAL

Esta Tesis Doctoral se presenta como compendio de publicaciones que la doctoranda ha desarrollado en los últimos años, siguiendo una línea de investigación común y continua.

Se incluyen un total de tres artículos que han sido publicados en revistas indexadas por el Science Citation Index y un cuarto presentado en el 11th International Conference on Hybrid Artificial Intelligence Systems redactados desde el comienzo de los estudios de doctorado, hasta la fecha de entrega del documento de la Tesis.

Los directores de la Tesis Doctoral acreditan la participación de la doctoranda en todos los artículos, siendo su aportación muy importante para el desarrollo y publicación de los mismos.

Los cuatro artículos incluidos en esta Tesis y sus correspondientes referencias se enumeran a continuación:

1. Concepción Crespo Turrado, María del Carmen Meizoso López, Fernando Sánchez Lasheras, Benigno Antonio Rodríguez Gómez, José Luis Calvo Rollé, Francisco Javier de Cos Juez. **Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions.** *Sensors* **2014**, *14*(11), 20382-20399.doi:10.3390/s141120382.

2. Concepción Crespo Turrado, Fernando Sánchez Lasheras, José Luis Calvo-Rollé, Andrés José Piñón-Pazos, Francisco Javier de Cos Juez. **A New Missing Data Imputation Algorithm Applied to Electrical Data Loggers.** *Sensors* **2015**, *15*(12), 31069-31082.doi:10.3390/s151229842.

3. Concepción Crespo Turrado, Fernando Sánchez Lasheras, José Luis Calvo-Rollé, Andrés-José Piñón-Pazos, Manuel G. Melero, Francisco Javier de Cos Juez. **A Hybrid Algorithm for Missing Data Imputation and Its Application to Electrical Data Loggers.** *Sensors* **2016**, *16*(9), 1467. doi:10.3390/s16091467.

4. Concepción Crespo-Turrado, José Luis Casteleiro-Roca, Fernando Sánchez-Lasheras, José Antonio López-Vázquez, Francisco Javier de Cos Juez, José Luis Calvo-Rolle, Emilio Corchado**. Student Performance Prediction Applying Missing Data Imputation in Electrical Engineering Studies Degree. Chapter:** Hybrid Artificial Intelligent Systems. Volume 9648 of the series Lecture Notes in Computer Science pp 126-135

2. INTRODUCCIÓN

La gestión de la energía en edificios públicos es fundamental para garantizar el confort de los usuarios de los mismos al menor coste posible. Por un lado, se persigue modelizar los edificios de manera que sus instalaciones trabajen con un rendimiento óptimo y, por otro lado, conocer la distribución de consumos a fin de obtener las condiciones más ventajosas posibles en la contratación de suministros.

Para llevar a cabo la gestión energética, se hace necesario medir en tiempo real los distintos parámetros que van a influir en el coste final de la energía, medidas cuarto horarias de potencia, intensidad de corriente, etc. Durante el proceso de recopilación de datos, es habitual que ocurran incidencias como cortes de suministro, averías en instalaciones y equipos, fallos en la red de comunicación, etc. que originan la contaminación de ciertas medidas e incluso la pérdida total de las mismas.

Esta Tesis Doctoral surge como consecuencia del trabajo profesional desarrollado por la autora en el ámbito de la gestión de la energía eléctrica en edificios públicos, donde a través de la Oficina de Control de Instalaciones se verifican las redes de distribución de baja tensión y se establecen índices de calidad del suministro eléctrico para los distintos puntos de consumo pertenecientes a la Universidad de Oviedo.

Esta experiencia dirige el estudio del consumo de electricidad en la edificación hacía modelos numéricos para encontrar soluciones prácticas a la falta de datos en los registradores. La autora a través de este trabajo, busca mostrar cómo abordar la falta de registros. Es en este momento, cuando se hace necesario recurrir a técnicas de imputación de datos faltantes a fin de obtener un archivo de datos completo.

2.1. Antecedentes y Motivación de la Tesis Doctoral

Las redes de distribución son elementos cruciales en el sistema eléctrico: su extensión supone el 99% de la longitud total de la red y un altísimo porcentaje de toda la demanda eléctrica está directamente conectada a ella [IEA 1]. En las próximas décadas, se prevén importantes cambios en todas las infraestructuras energéticas [IEA 2] y, concretamente, en

la configuración del sistema eléctrico (Figura 1). Por un lado, la incorporación de los sistemas de generación distribuida está cambiando el concepto tradicional de red, según el cual, la energía se produce en una central generadora y llega al consumidor final a través de los sistemas de transporte y distribución. Las tecnologías de generación distribuida permiten al consumidor producir energía, fundamentalmente a partir de fuentes renovables, dando lugar a múltiples fuentes generadoras ubicadas en numerosos puntos de la red de distribución. Por otro lado, el incremento de vehículos que utilizan tracción eléctrica determina que los perfiles de carga se modifiquen y que, entre otras cosas, las infraestructuras hayan de ser diseñadas para soportar cargas adicionales y dispongan de la capacidad para tarificar correctamente la electricidad vendida a los usuarios. En consecuencia, la complejidad e importancia de las redes de distribución está yendo en aumento.

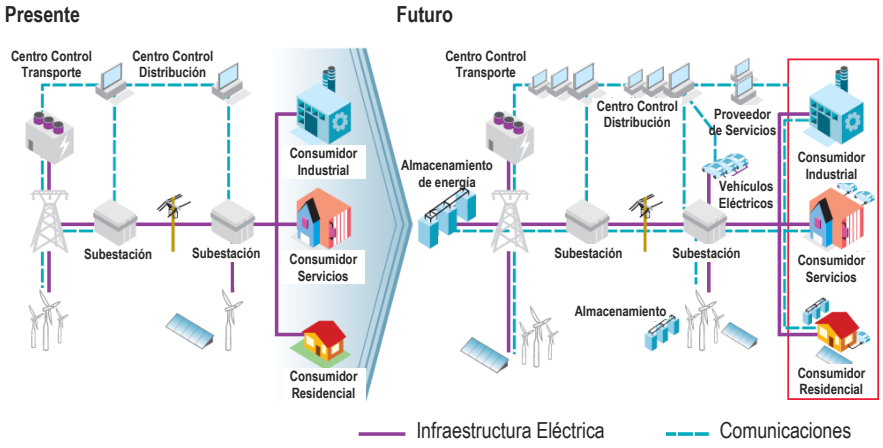


Figura 1: Presente y futuro de los sistemas eléctricos de potencia [IEA 1]

Además de la creciente integración en los sistemas eléctricos de la generación con fuentes renovables y la demanda creada por los vehículos eléctricos, actualmente, estos sistemas afrontan una serie de retos que incluyen el envejecimiento de las infraestructuras, el continuado aumento de la demanda, la necesidad de incrementar la seguridad del suministro y el desafío de disminuir las emisiones de CO_2. De esta forma, la previsión de la Agencia Internacional de la Energía es que las inversiones en la red de distribución supongan entre el 65% y más del 80% de todas las inversiones en la red hasta el año 2050 [IEA 1].

Las redes inteligentes (smart grids) y las tecnologías asociadas a ellas ofrecen alternativas para satisfacer estos retos y suministrar una energía más limpia y de forma más eficiente y sostenible. La Agencia Internacional de la Energía [IEA 3] define las redes inteligentes como redes de energía eléctrica que utilizan tecnologías digitales para monitorizar y gestionar el transporte de energía eléctrica desde los puntos de generación hasta los consumidores finales. Estas redes son capaces de coordinar las necesidades y capacidades de generadores, operadores de red, usuarios finales y resto de participantes del mercado eléctrico, de tal forma que pueden optimizar los activos de uso y operación de todo el sistema, minimizar costes e impacto medioambiental y mantener la fiabilidad, resiliencia y estabilidad del sistema. A partir del año 2009, el despliegue de proyectos piloto de redes

de este tipo ha experimentado un considerable desarrollo en países de todo el mundo, debido, sustancialmente, a los crecientes estímulos por parte de entes públicos y a incentivos gubernamentales. Esto ha motivado que las áreas tecnológicas que involucran (Figura 2) hayan experimentado grandes avances y que, en algunos casos, hayan alcanzado elevados niveles de madurez en su aplicación.

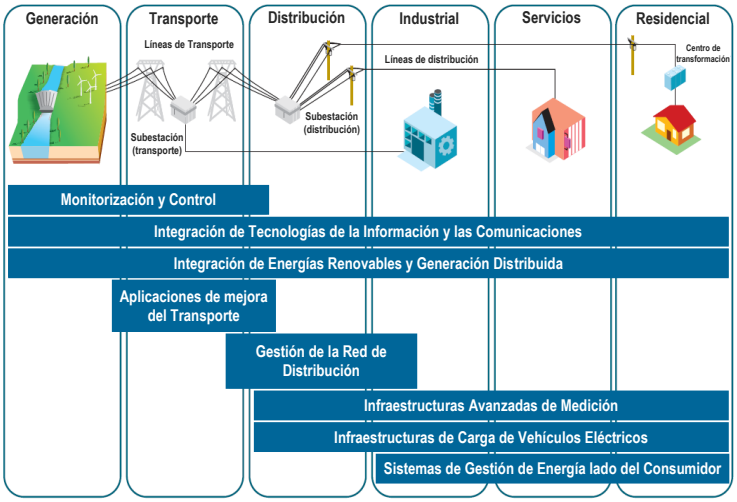


Figura 2: Áreas tecnológicas implicadas en una red inteligente [IEA 3]

Una de las áreas tecnológicas que más interés despierta en los proyectos de redes inteligentes es la relacionada con las infraestructuras avanzadas de medición (Advanced Metering Infrastructures - AMI). Ello se debe a dos razones: por un lado, los elementos relacionados con estas infraestructuras son los que, hasta la fecha, tienen una tendencia de implantación más rápida y un grado de madurez tecnológico más elevado. Por otro lado, las tecnologías relacionadas con las AMI son las que más se interrelacionan con otras áreas tecnológicas de las redes inteligentes. Es el caso, por ejemplo, de los sistemas de gestión de energía del lado del consumidor (Customer-Side Systems). La gestión de la energía en el hogar o en el negocio se lleva a cabo mediante sistemas de gestión que recopilan la información y proporcionan al usuario el apoyo suficiente para ajustar sus necesidades de energía eléctrica. Las aplicaciones inteligentes que se desarrollen en el futuro necesitarán de información del lado del usuario que en gran medida podrán ser proporcionadas por las AMI. Además, con la implantación de los contadores o medidores inteligentes, el número de sensores disponibles en la red pasará -está pasando- de unos pocos miles a millones. La necesidad de disponer de sensores adicionales en subestaciones y para la automatización de la etapa de distribución, permite prever que el número de sensores disponibles llegue a 1.5 por cada consumidor conectado a la red. Esto supone una modificación sustancial en la cantidad de información que han de manejar los sistemas de gestión, lo cual está dando lugar a una nueva generación de software en los centros de control que afecta al transporte y a la distribución. La evolución de estos sistemas hacia otros capaces de sugerir al operador las decisiones a tomar, o de tomarlas de forma automática, será la única forma de poder operar en el futuro una red en la que los puntos de generación distribuida serán innumerables y la proliferación de sensores y recolección de datos implicará que las técnicas como las que se presentan en esta tesis resultaran especialmente adecuadas para asegurar la fiabilidad del sistema ante posibles pérdidas de información.

Las AMI incluyen medidores inteligentes y un sistema de comunicación en campo que permite al medidor comunicarse con una unidad central. A partir de la información recogida, mediante sistemas basados en las tecnologías de la información y las comunicaciones se obtienen datos operativos que, entre otras utilidades, permiten la realización de tareas administrativas y de facturación. A día de hoy, no todas las redes tienen estas infraestructuras plenamente operativas y su utilidad está bastante limitada, quedando reducida, en muchos casos, a la facturación del consumo. La previsión es que, en el futuro, su utilidad se extienda a otras funciones que, además de suponer mejoras en el servicio, permitan la detección de fallos en el mismo, identificar fraudes, etc.

Los medidores (contadores) inteligentes son dispositivos electrónicos de medida usados para comunicar la información de facturación de clientes y operar sus sistemas eléctricos. Inicialmente, el uso de esta tecnología estaba reservado para clientes industriales y comerciales, ya que éstos requerían la aplicación de tarifas específicas en las que era necesario un conocimiento muy desagregado de los datos de facturación. De esta forma, el uso de los medidores electrónicos se implantó entre los consumidores más grandes. Posteriormente, conforme los requerimientos en los datos de consumo del resto de consumidores han ido aumentando y los costes de esta tecnología disminuyendo, su uso se ha ido extendiendo de forma gradual a todos los consumidores. A la combinación de los medidores electrónicos con las tecnologías bidireccionales de comunicación, utilizadas para el intercambio de información, monitorización y control, es lo que comúnmente se denomina AMI. Los sistemas previos, que utilizaban una comunicación unidireccional para recopilar los datos de medida, reciben el nombre de sistemas de lectura automatizada (Automated Meter Reading – AMR). En la Tabla 1, se pueden ver recopiladas las funciones que se han ido añadiendo a los medidores en su evolución desde los AMR a los AMI [EEI 1].

AMR	AMR Plus	AMI
Lecturas periódicas automatizadas.	Lecturas discrecionales.	Función de corte de servicio.
Detección de salidas de servicio.	Intervalos de datos más frecuentes.	Tarifas más desagregadas.
Detección de alteraciones/daños.	Notificación de salidas de servicio.	Programación remota.
Perfiles de carga.	Lectura de otras magnitudes.	Información calidad de suministro.

Tabla 1: Evolución en las funciones de los medidores

Como ya se ha venido comentando, las redes inteligentes están suponiendo un gran cambio de los aspectos de la red eléctrica relacionados con el transporte y la distribución. Funcionalmente, estas redes monitorizan y controlan las actividades que en ellas se llevan a cabo, asegurando un eficiente y fiable flujo bidireccional de energía e información de plantas generadoras, consumidores y puntos intermedios. Una red inteligente monitoriza el consumo

de energía mediante contadores inteligentes, los cuales transmiten la información sobre el uso de la energía a través de redes de comunicación. Los sistemas de medida inteligentes son, por tanto, un componente fundamental de las infraestructuras de las redes inteligentes en las tareas de recopilación de datos y comunicaciones. Asimismo, permiten al usuario realizar un seguimiento del uso de su propia energía y pueden, a través del envío de comandos, realizar maniobras sobre elementos de aparamenta de la red que contribuyen a obtener un sistema de distribución de la energía eléctrica más fiable.

A pesar de que existen varias tecnologías y diseños de sistemas de medida inteligentes, el esquema general en el que se basan es el mismo (Figura 3). Los medidores inteligentes recopilan de forma local la información y la transmiten, con una periodicidad variable, mediante una red de área local (LAN), a un punto de recolección. El elemento colector recupera estos datos y puede llevar a cabo algún tipo de procesamiento con ellos, tras lo cual la información se transmite mediante una red de área extendida (WAN) a un punto de recolección centralizado. En este punto, se realiza el procesamiento de la información y se gestiona su uso según las necesidades. Sobre este esquema general, las principales variantes de los sistemas de medida inteligentes vienen dadas por las tecnologías aplicadas en la transmisión de la información en la primera etapa (LAN) y que, fundamentalmente, se basan en el uso de la radiofrecuencia o de la propia línea de distribución de energía eléctrica. Adicionalmente a los elementos especificados, existe un componente de software necesario para la operatividad de los sistemas de medida inteligentes. Este software constituye el sistema de gestión de los datos de medida (Meter Data Management System – MDMS) y es un componente fundamental en el despliegue de los medidores inteligentes, puesto que almacena y verifica toda la información que se recibe de éstos y, a su vez, proporciona los datos necesarios a otras aplicaciones de gestión.

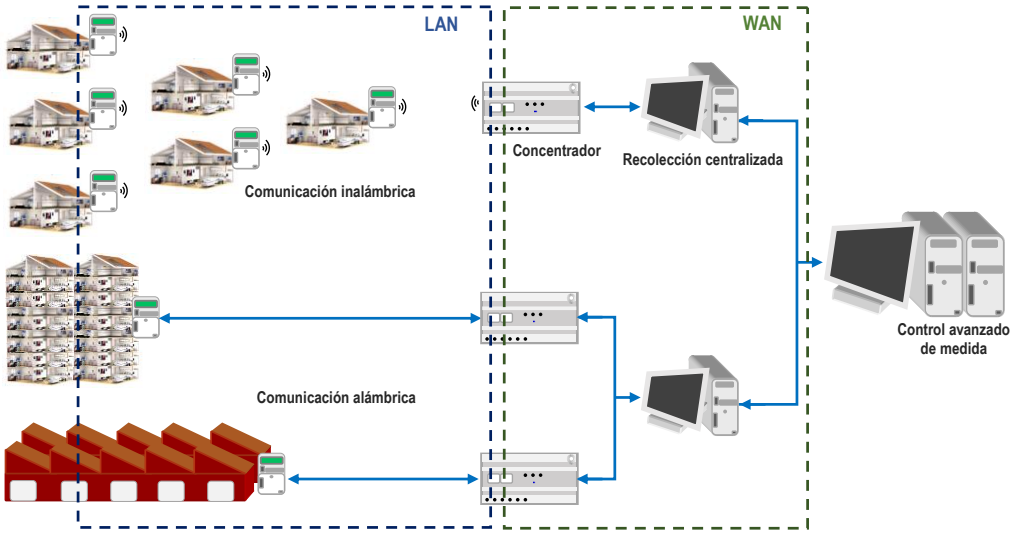


Figura 3: Esquema general de una infraestructura avanzada de medición

El uso de los medidores inteligentes incrementa de forma notable el número de medidas disponibles de la red de distribución. Paralelamente, el elevado número de dispositivos de medida instalados conlleva la aparición de nuevos retos relacionados con la

fiabilidad de las comunicaciones y la exactitud de las medidas [EEI 1], [IEA 3]. La exactitud de los medidores inteligentes suele ser alta, pudiendo oscilar entre ±0,5% y ±0,2%; sin embargo, múltiples factores pueden afectar a la exactitud de los datos obtenidos y a la fiabilidad en su adquisición y, por consiguiente, a las prestaciones de todo el sistema de medida y del MDMS. La selección inadecuada de los transformadores de medida, la falta de sincronización entre los medidores, la existencia de valores falsos o faltantes, los errores en la configuración de los medidores o la falta de homogeneidad en las unidades de medida son algunos de los factores identificados como causantes de la falta de exactitud en los datos obtenidos [Peppanen 1] y una de las motivaciones principales de la presente tesis.

No siendo los únicos, todos estos factores contribuyen notablemente a que la fiabilidad de los datos proporcionados por los sistemas de medida inteligente sea, en general, baja [Peppanen 2]. Según se recoge en [IEA 3], típicamente, un sistema de medida a gran escala, basado en AMI, puede perder hasta el 4% de los datos de medidas de potencia a la hora. En [Peppanen 2], concretamente, se estudia una instalación en la que los datos perdidos al cabo de un año representan el 2,72% del total. Esta pérdida de información puede deberse a que se haya producido, temporal o permanentemente, la interrupción de las comunicaciones entre el medidor y el sistema MDMS por múltiples causas: trabajos de mantenimiento, reparación o sustitución de equipos, averías en la instalación, etc. En cualquier caso, esta falta de datos y, en ocasiones, su baja fiabilidad, deben ser tenidas en cuenta y tratadas adecuadamente.

En este marco, en el que los sistemas eléctricos están evolucionando rápidamente hacia redes inteligentes, es imprescindible el manejo de ingentes cantidades de datos para su correcto funcionamiento. Adquiere especial importancia, entonces, la necesidad de discernir si estos datos, obtenidos con los sistema de medida inteligente, reflejan de forma correcta lo ocurrido en los intervalos de interés. Para ello, se hace necesario el desarrollo de métodos y procedimientos que permitan el pre-procesado de los datos, su discriminación y el tratamiento de los datos faltantes (Missing Data). En este contexto, se enmarca la motivación para el trabajo realizado en la presente Tesis. Con ella, precisamente, se pretende profundizar en el estudio de la imputación de datos faltantes en los sistemas de medida y monitorización de las magnitudes eléctricas en las redes de baja tensión.

Para el desarrollo de las técnicas de imputación de datos propuestas en la Tesis, se utilizó como laboratorio la propia red eléctrica de la Universidad de Oviedo; especialmente, el Edificio Severo Ochoa, cuyas instalaciones permitieron experimentar y validar las técnicas de imputación de datos desarrolladas.

La elección de este edificio, se debe a que el Severo Ochoa alberga al Servicio de Informática así como los Servicios Científico Técnicos de la Universidad de Oviedo. Esta dualidad en el uso, le confiere unas características especiales desde el punto de vista eléctrico, marcadas siempre por la necesidad de garantizar el suministro al centro de proceso de datos (CPD) del Servicio de informática 24 horas al día, 365 días al año con unos buenos estándares de calidad.

Entre los parámetros que caracterizan la calidad del suministro eléctrico en una instalación de baja tensión se encuentran:

- La amplitud y frecuencia de la tensión de suministro.
- El desequilibrio existente entre fases.
- La distorsión armónica de la tensión de suministro.

Así pues, la medición continuada de estos parámetros en una instalación eléctrica nos permitirá actuar sobre ella a fin de garantizar unos estándares de calidad adecuados a los requerimientos actuales de uso.

La presencia de armónicos en una instalación eléctrica se asocia con muchos problemas en su funcionamiento. Los principales problemas son el sobrecalentamiento en los conductores, especialmente en los neutros, que da lugar al disparo de los interruptores automáticos y a problemas en la continuidad del suministro. La distorsión de la forma de onda también puede provocar el mal funcionamiento de algunos dispositivos.

Otro problema que aparece con frecuencia en una instalación eléctrica es el desequilibrio entre fases. Aunque es bien sabido que el equilibrio se logra trabajando en los niveles más altos de la capacidad instalada, con el fin de aprovechar al máximo la instalación, a veces esto no es posible. Un desequilibrio lo provoca generalmente una mala distribución de cargas entre fases, lo que causa un alto retorno de corriente por el neutro. Estos problemas se incrementarán si estas cargas también están introduciendo distorsión armónica

En este contexto, la calidad de la electricidad es un problema que debe contemplarse desde todos los parámetros (voltaje, corriente, anomalías de la frecuencia, etc.) que puedan causar fallos o incapacidad de los dispositivos eléctricos o electrónicos para su funcionamiento.

Por todo lo expuesto, se hace imprescindible medir las variables de la instalación eléctrica, ya que así se podrá verificar que los parámetros de funcionamiento de la misma son los adecuados para satisfacer los requerimientos del edificio. Asimismo, las mediciones permitirán evaluar el rendimiento de los dispositivos instalados. Todo ello hace necesaria la implementación de una herramienta, como la desarrollada en la presente Tesis, que incremente el nivel de fiabilidad de los sistemas de medida. Esto se conseguirá dotándolos de un procedimiento de supervisión capaz de imputar eficientemente los potenciales datos faltantes que, como ya se ha mencionado, pueden alcanzar un porcentaje, nada despreciable, del 4%.

3. OBJETIVOS DE LA TESIS DOCTORAL

La aparición de huecos en las series de datos de medida en la red eléctrica es un problema habitual con el que se ha venido enfrentando el estudio estadístico de la calidad del suministro eléctrico desde hace tiempo. Disponer de un archivo de datos completos es ideal, pero aplicar métodos de imputación inapropiados para lograrlo, puede generar más problemas de los que resuelve. Así, durante las últimas décadas se han desarrollado procedimientos matemáticos a tal fin, que tienen mejores propiedades estadísticas que las opciones usadas tradicionalmente como la eliminación de datos (listwise) o el pareo de observaciones (pairwise).

Los algoritmos más eficientes hasta la fecha para buscar un valor matemático a datos faltantes en las series de medida son los de imputación multivariada (IM) y se pueden aplicar utilizando paquetes comerciales y de acceso gratuito.

Sin embargo, imputar información no debe entenderse como un fin en sí mismo y sus implicaciones en el análisis secundario de datos deben evaluarse con cautela partiendo de la base de que no existe un método de imputación ideal.

Cada situación es diferente, y la tasa de faltantes y su distribución espacial cambian en el tiempo, por lo que no es conveniente adoptar, a priori, el mismo procedimiento de imputación para todas las variables independientemente de su origen y contexto.

Se hace necesario, pues, desarrollar algoritmos de imputación que se adapten de manera específica a cada situación con objeto de maximizar su eficiencia y es por eso, que en este trabajo se han probado los algoritmos, buscados a priori para el estudio de la energía eléctrica en edificación en otras series de datos diferentes a objeto de comprobar su bonanza.

Así pues, el objetivo general de este trabajo es analizar el rendimiento de los algoritmos de imputación existentes cuando se aplican a las bases de datos de medida y control de las instalaciones consumidoras de energía eléctrica de la Universidad de Oviedo, así como desarrollar un novedoso algoritmo de imputación hibrido, basado en inteligencia artificial, que combine técnicas predictivas y estadísticas.

Partiendo de este objetivo general, pueden determinarse los siguientes objetivos específicos:

1. Seleccionar un método eficiente de imputación, dentro del estado del arte actual, que sirva como referencia de comparación para el futuro desarrollo de métodos específicos.
2. Desarrollar un método de imputación de alto rendimiento, basado en técnicas de inteligencia artificial predictivas, que permita imputar de manera inteligente en función de la evolución de la red.
3. Desarrollar un método de imputación de alto rendimiento, basado en técnicas de inteligencia artificial clasificativas, que permita realizar imputaciones eficientes en entornos con elevado número de datos faltantes por registro.
4. Hibridar los dos algoritmos inteligentes desarrollados de tal manera que se optimicen las ventajas de cada uno en función del número de faltantes por registro.

4. ESTADO DEL ARTE

4.1 Introducción.

La pérdida de datos supone un problema cuando se pretende realizar cualquier estudio que implique un tratamiento estadístico de la información. Los métodos estadísticos habituales asumen que se dispone de las medidas de todas las variables de un modelo especificado en todos los casos; sin embargo, esto no es así. Como se ha venido comentando, la pérdida de información es un fenómeno inherente a los procesos de medida de magnitudes eléctricas, por lo que, en un conjunto de datos, es habitual que esa pérdida afecte a algunas de las variables en algunos casos. Así pues, el inicio de los trabajos de esta tesis ha estado principalmente motivado por la necesidad de resolver el problema de falta de información de variables en los registros de energía llevados a cabo, utilizando los procedimientos que la literatura considera como el estado del arte en este ámbito.

Desde el punto de vista de la estadística, está asumido que se trabaja con series de datos completas y para ello, se incorporan opciones —que deberán ser adecuadas— para imputar observaciones faltantes. Está ampliamente documentado que la aplicación de procedimientos inapropiados de sustitución de información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio. Se hace por tanto imprescindible utilizar un buen procedimiento de imputación.

Los procedimientos de imputación que se utilizan con mayor frecuencia limitan o sobredimensionan el poder explicativo de los modelos y generan estimadores sesgados que distorsionan las relaciones de causalidad entre las variables, generan subestimación en la varianza y alteran el valor de los coeficientes de correlación [Acock].

Durante las últimas décadas se han propuesto distintas metodologías para sustituir datos faltantes; sin embargo, es frecuente que estos procedimientos se apliquen sin tener en cuenta sus fundamentos teóricos y sus limitaciones prácticas

Rubin [Rubin] sustenta que los procedimientos de imputación múltiple (IM) deben aplicarse en forma intensiva, pero no aclara que es altamente probable que en la práctica no se satisfagan los supuestos en que se fundamenta su metodología, ya que es común que el patrón de datos omitidos esté asociado a las características de la población de referencia, lo cual invalida el supuesto de aleatoriedad en el que se sustenta la técnica IM, y que asume que la falta de información tiene una distribución aleatoria en la población de referencia.

En relación a la imputación múltiple, cabe mencionar que Imputar significa sustituir observaciones, ya sea porque se carece de información (missing values) o porque se detecta que algunos de los valores recolectados no se corresponden con el comportamiento esperado (outliers). En esta situación, es común que se desee reponer las observaciones y se decida aplicar algún método de imputación de datos.

No obstante, utilizar algún procedimiento inapropiado puede generar más problemas de los que resuelve, introduciendo sesgos en el valor de los estimadores y en su error estándar, al tiempo que podría distorsionar la potencia de las pruebas de hipótesis [Little], lo que sugiere reflexionar acerca de la mejor manera de obtener estimadores que generen inferencia válida a partir de datos imputados. En [Rubin], se hace esta reflexión y se propone como solución el método de imputación múltiple (IM).

IM utiliza métodos de simulación de Monte Carlo y sustituye los datos faltantes a partir de un número ($m>1$) de simulaciones que, de acuerdo al autor, se ubica entre 3 y 10. La metodología consta de varias etapas, y en cada simulación se analizan la matriz de datos completos a partir de métodos estadísticos convencionales y posteriormente se combinan los resultados para generar estimadores robustos, su error estándar e intervalos de confianza.

Los métodos más usuales para el tratamiento de los valores ausentes, analizan ficheros de datos rectangulares en los que las filas de la matriz de datos representan las unidades medias y las columnas las variables que se miden para cada unidad. Las técnicas que vamos a enumerar aquí se ocupan de la falta parcial de datos en las filas, generalmente analizando la información del resto de las variables de la misma unidad para estimar un valor del dato ausente.

Rey del Castillo [Rey], especifica que independientemente de cual sea el método desde el que se pretende analizar el problema de la falta de datos, deben establecerse previamente una serie de conceptos válidos para todos ellos. De este modo:

Si se considera $Y=(y_{ij})$ como un fichero de datos rectangular de tamaño n x k, donde:

n: es el número de unidades de observación

k: es el número de variables

$y_i=(y_{i1}\ldots,y_{ik})$ la i-enesima fila

y_{ij}: el valor de la variable Y_j para la unidad i

Cuando en el fichero anterior existe falta de información, debe definirse la matriz indicadora de falta de datos "M" como: $M= (m_{ij})$, donde:

$m_{ij} = 1$, cuando y_{ij} está ausente

$m_{ij} = 0$, cuando y_{ij} no está ausente

Así pues, dado el siguiente fichero de datos Y:

$$Y = \begin{pmatrix} y11 & ? & y13 & \dots & ? & y1j & y1j+1 & \dots & y1k \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ yi1 & yi2 & ? & \dots & yij-1 & yij & ? & \dots & yik \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ ? & yn2 & yn3 & \dots & ? & ynj & ynj+1 & \dots & ynk \end{pmatrix} \quad (1)$$

Su correspondiente matriz "M" indicadora de la falta de datos sería:

$$M = \begin{pmatrix} 0 & 1 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \end{pmatrix} \quad (2)$$

Lo primero que ha de plantearse es el motivo de la falta de datos, es decir, el mecanismo que genera la matriz M y, en especial, conocer si el hecho de que falte un dato tiene relación con los valores de las variables. Según explican [Little], a este fin se trata la matriz M como una variable aleatoria y asignándole una distribución de probabilidad.

El mecanismo generador de los datos ausentes estará caracterizado por la distribución condicional de M dado Y, es decir, f (M/Y, Φ), donde Φ representa un conjunto de parámetros desconocidos. Si Y_{obs} e Y_{mis} son los datos registrados y los ausentes respectivamente. De Y, podemos decir:

1. **Datos ausentes en forma completamente aleatoria** (*MCAR: Missing Completely at Random*) si dicha ausencia no depende de Y. Esto es así, si: f(M/Y, Φ)=f(M │ Φ), para todo Y, Φ.
2. **Datos ausentes en forma aleatoria** (*MAR: Missing at Random*) Menos restrictivo que el anterior, la distribución de M depende de los componentes observados, Y_{obs}, y no de los ausentes, Y_{mis}. Así: f (M │ Y, Φ)= f(M │ Y_{obs}, Φ), para todo Y_{mis}, Φ.
3. **Ausencia no Aleatoria** (*MNAR: Missing Not at Random*), cuando la distribución de M depende de los valores ausentes de la matriz Y: Y_{mis}.

Little y Rubin [Little] señalaron, además, que se considera que la falta de datos es ignorable cuando el mecanismo es de tipo MAR o MCAR, mientras que es no ignorable cuando es no aleatoria, es decir, cuando el mecanismo es MNAR.

Los procedimientos a estudiar en el presente trabajo se construyen bajo el supuesto MCAR.

En los procedimientos de inteligencia computacional, según [Marwala], muchas mediciones pueden requerir de una solución inmediata a la falta de datos, es decir, en tiempo real. En estas situaciones los analistas necesitan soluciones independientes de las causas que producen las ausencias ya que no disponen de tiempo para estudiarlas. Por este motivo, el autor introduce un nuevo mecanismo de falta de datos: la ausencia por diseño natural (*MBND: Missing by Natural Design*), donde la falta de datos se produce como consecuencia de no poder llevar a cabo mediciones físicamente. La solución que se adopta más frecuentemente es la modelización matemática de los valores ausentes, siendo esto es posible cuando pueden describirse sus características físicas.

En lo que a los procedimientos empleados para el tratamiento de datos faltantes se refiere, según Rey del Castillo [Rey], y en base a su forma de actuar, podemos distinguir los siguientes niveles:

1. **Técnicas Basadas en registros completos**, conocidas también como "Técnicas de Ignorancia" porque, o bien ignoran los datos ausentes estudiando solo los casos completos o disponibles, o bien ignoran su ausencia considerando dicha falta como una categoría más.
2. **Técnicas de Imputación,** Cualquier estrategia que estime individualmente los valores ausentes de un conjunto de datos, de modo que se puedan aplicar después los métodos de análisis al conjunto de datos completado. Estas técnicas se consideran determinísticas, porque son las que históricamente aparecieron antes y constituyen las técnicas propias del aprendizaje automático.
3. **Técnicas de Tolerancia,** Estrategias de tratamiento para la ausencia de datos que trabajan directamente con conjuntos que incluyen datos faltantes. Dentro de estas técnicas se consideran, las técnicas indirectas de estimación (utilizadas en inferencia estadística para estimar parámetros) y las técnicas que actúan sobre algoritmos específicos de aprendizaje automático.

Los distintos procedimientos que se han venido utilizando a lo largo del tiempo para el tratamiento de datos, así como las técnicas más reconocidas de ellas se recogen a continuación [Rey]:

1. **Técnicas Basadas en Registros Completos:** también llamadas *técnicas de ignorancia* se fundamentan en ignorar la existencia de datos ausentes.
    a. Técnicas de casos completos (Listwise deletion): consiste en eliminar todos los registros que tienen alguna variable con datos ausentes.
    b. Técnicas de casos disponibles (Pairwise deletion): consiste en eliminar solamente los casos sin información para las variables que intervienen.
    c. Creación de nueva categoría (Dato ausente): consiste en incorporar una nueva categoría, la de *dato ausente* y tratarla en el análisis como otra cualquiera.

2. **Técnicas de Imputación:** consiste en sustituir el valor ausente por un valor concreto o de imputación.
    a. Técnicas de imputación determinística: son muy utilizadas en aprendizaje automático y ampliamente utilizadas como técnicas de clasificación o

agrupamiento, algunos ejemplos son el Cold-deck, hot-deck, k-vecino más próximo, GBK11, CM1, media, moda, regresión, MC1, POP.

b. Técnicas basadas en modelos probabilísticos: surgen para solucionar problemas generados por la utilización de técnicas deterministas en las imputaciones y pretenden medir la incertidumbre generada por la sustitución de valores ausentes por los correspondientes datos imputados. Como el algoritmo EM, y dentro de estas, los siguientes:
    i. Técnicas de remuestreo: Del estilo de Bootstrap, Jackknife, bootstrap, bayesiano y bootstrap bayesiano aproximado.
    ii. Imputación múltiple
    iii. Enfoque bayesiano.
    iv. Aumento de datos
    v. Imputación fraccional: imputación fraccional, imputación hot-deck fraccional, imputación fraccional del k-vecino más próximo.
c. Técnicas de Clasificación y Predicción para la Imputación: se corresponden con los métodos de aprendizaje automático que pueden ser utilizadas para estimar los valores individuales ausentes.
    i. Técnicas de clasificación no supervisada: Autoclass, FRCAR, MVC, SLLS, RRP, redes bayesianas: bayes ingenuo.
    ii. Técnicas de clasificación supervisada: BosqMue aleatorio, CL1P4, imputación por mínimos cuadrados, C4.5, KRIMP, redes neuronales: autoasociativas, redes min-max.

3. **Técnicas de tolerancia:** se refiere a estrategias de tratamiento de la falta de datos que son internas en los procedimientos de análisis, es decir, no precisan imputar los valores ausentes sino que trabajan directamente con conjuntos que incluyen datos faltantes, sin eliminarlos ni sustituirlos.
    a. Técnicas indirectas: técnicas de reponderación, calibración.
    b. Técnicas sobre algoritmos de aprendizaje específicos.
        i. Arboles de decisión: CART, C4.5, modelos de características reducidas.
        ii. Redes bayesianas
        iii. Redes neuronales: redes de extensión.
        iv. Otros: análisis paralelo, sistemas híbridos neuro-borrosos.

A la vista de la cantidad de métodos existentes para la determinación de datos faltantes, es importante definir cuál es el que mejor se adapta a cada circunstancia pudiendo aplicarse diferentes criterios para obtener un valor faltante o un conjunto de valores faltantes en una serie de datos (G). Así pues, las propiedades autoregresivas de la señal, por ejemplo, nos permiten tener en cuenta valores separados para autocompletar una serie de datos. Otras técnicas, como los modelos de ARIMA, modelo autoregresivo integrado de media móvil, (AutoRegresive Integrated Moving Average), popularizados en los años 70 por George Box y Gwilyn Jenkins, se han venido usando para pronosticar series temporales. Por lo tanto, estos modelos ARIMA también se pueden usar para imputar datos en series de tiempo [Reikard].

Varios autores, [Allison], [Enders], [Newman], [Schafer], [Buuren 1], han abogado por el uso de modernas técnicas de imputación de datos, como la imputación múltiple (MI), en lugar de enfoques tradicionales como la supresión por pares o por listas. Un requisito fundamental de MI es que el modelo de imputación debe ser al menos tan general como el modelo de interés para preservar las relaciones entre las variables [Enders]. En el caso de datos multinivel incompletos, es importante que el modelo de imputación tenga en cuenta la estructura multinivel para garantizar inferencias estadísticas válidas en los análisis multinivel subsiguientes [Black], [Graham], [Buuren 2].

El trabajo desarrollado en esta tesis tiene como objetivo evaluar la imputación multivariada de datos faltantes a escala de 15 minutos por el método de ecuaciones encadenadas (MICE), evaluar los resultados y posteriormente compararlos con los obtenidos al introducir variaciones en el método con técnicas de aprendizaje automático.

Así pues se ha desarrollado en las siguientes fases:

1. Analizado del rendimiento de los algoritmos de imputación existentes cuando se aplican a las bases de datos de medida y control de las instalaciones consumidoras de energía de la Universidad de Oviedo
2. Desarrollo de un novedoso algoritmo de imputación hibrido, basado en inteligencia artificial, que combine técnicas predictivas y estadísticas
3. Integración del algoritmo de imputación dentro de los sistemas de control de instalaciones de la Universidad de Oviedo.

Una vez identificado el problema de la falta de registro de datos en la instalación a la hora de llevar a cabo la modelización de la instalación, se plantea la necesidad de estimar los valores de esos datos faltantes de manera artificial.

Es así cómo, después de valorar distintas técnicas de imputación se decide comenzar el estudio con métodos conocidos, eliminando de manera artificial datos previamente registrados y calcular a posteriori el error cometido entre el dato medido y el dato calculado.

A continuación, se intenta mejorar el método aplicado con anterioridad, para buscar la mayor aproximación posible al dato real mediante el uso de diferentes algoritmos matemáticos que introducen técnicas de aprendizaje supervisado. Siempre buscando el mínimo error.

4.2 Aplicación al Sector Eléctrico.

En este apartado, se realiza una recopilación de los trabajos encontrados en el ámbito de las redes eléctricas en los que aparece reflejado este problema junto con la solución o el tratamiento que se ha propuesto para el mismo. Con objeto de acotar la revisión bibliográfica, sólo se recogen aquellos estudios en los que la falta de datos se ha encontrado explícitamente referenciado. Aunque, dada la transversalidad que presentan las situaciones en las que se producen los datos faltantes, los estudios que se referencian en esta sección

abarcan distintas aplicaciones prácticas que tiene la imputación de datos dentro de la rama de la Ingeniería Eléctrica, por ejemplo:

- Estimación de estado.
- Predicción de la demanda.
- Calidad del suministro eléctrico.
- Infraestructuras avanzadas de medición o depuración de perfiles de demanda.

Los anteriores, sólo son algunos de los temas en los que aparece reflejado el problema de los datos faltantes y donde la aplicación de los algoritmos objeto de estudio en el presente trabajo resolverían situaciones indeseadas.

A fin de fijar las aplicaciones prácticas, es necesario previamente referenciar cómo diferentes autores han venido estudiando y solventando el problema de la falta de datos, aunque ante la necesidad de acotar la revisión bibliográfica, se recogen únicamente aquellos estudios en los que se ha encontrado explícitamente referido este problema a la hora de estudiar los parámetros eléctricos.

El tratamiento de los datos faltantes en el ámbito de los sistemas eléctricos ha venido siendo objeto de estudio desde hace más de cuatro décadas [Merrill]. Podría decirse que se inicia con la irrupción del término bad data en la Ingeniería Eléctrica como consecuencia del apagón que sufrió, en 1965, una amplia zona del nordeste de Estados Unidos y parte del territorio de Ontario, en Canadá. Aquel incidente puso de manifiesto lo necesario que era el desarrollo de nuevas técnicas de operación de los sistemas eléctricos que permitieran elevar el nivel de seguridad en el servicio [Zarco] y, para ello, se realizó un considerable esfuerzo por conseguir más medidas e información de la red. Se desarrollaron así los sistemas de supervisión del control y adquisición de datos (Supervisory Control and Data Acquisition – SCADA), que supusieron un primer paso para analizar las condiciones de operación de la red y detectar las situaciones anómalas de funcionamiento. Pero no fue hasta 1970, con los trabajos de Schweppe y su estimación de estado [Schweppe 1], [Schweppe 2], [Schweppe 3], cuando se lograron grandes avances en el seguimiento y control del sistema y se sentaron las bases de lo que en la actualidad constituyen los sistemas de gestión de energía (Energy Management Systems – EMS).

4.2.1. La estimación de estado

Se trata de "un algoritmo de procesamiento de datos que convierte las medidas redundantes y otra información disponible en un estimado del estado del sistema eléctrico" [Zarco]. El estimador, a partir de toda la información disponible del sistema, genera una base de datos para llevar a cabo las funciones de control y de reparto. Los resultados de la estimación son los valores estimados de las variables del sistema eléctrico, que pueden incluir la generación, las cargas activas y reactivas, los flujos de potencia activa y reactiva de las líneas de transmisión y las magnitudes y fases de las tensiones de nudo. Los estimadores de estado, con su capacidad para resolver el problema de la inconsistencia en las medidas y de realizar el reparto de cargas en tiempo real, se han convertido en una parte

esencial de los EMS, al permitir obtener modelos de la red y monitorizar su estado, realizar tareas de planificación y de control y optimizar los flujos de carga, entre otras utilidades [Monticelli], [Huang 1].

Como consecuencia directa de los mencionados trabajos de Schweppe, comenzaron a aparecer las primeras publicaciones sobre datos incorrectos (bad data) y su influencia en la estimación de estado [Merrill], [Handschin]; ya que los estimadores de estado estáticos se basaban en funciones de mínimos cuadrados ponderados que podían ser muy sensibles a los datos erróneos y corromper, de esta forma, la información almacenada en la base de datos resultante [Mili]. A partir de entonces, el término bad data se empezó a utilizar profusamente en la literatura relacionada con los sistemas eléctricos y, desde entonces, los datos incorrectos han venido siendo objeto de estudio para su análisis, detección, identificación y procesamiento. En este ámbito, bad data se puede encontrar aludiendo a "datos mucho más inexactos de los que pueden ser asumidos por los modelos matemáticos, dando lugar a estimaciones muy deficientes" [Merrill] o, también, como "medidas con errores inusualmente grandes" [Doraiswami]. A pesar de que el término también aparece referido en [Kotiuga] como "medidas que se desvían del valor verdadero en, al menos, dos o tres veces la varianza asociada con dichas medidas" (es decir, errores que pueden ser mucho menos grandes), lo habitual es que este tipo de datos erróneos siempre vayan asociados a un mal funcionamiento o rotura de los equipos de medida o a fallos en los canales de comunicación con los mismos [Merrill], [Doraiswami], [Quintana], [Lo 1], [Zarco]. Una posible clasificación de los errores en las medidas que llegan al centro de control del sistema es la que se propone en [Lo 2]. En este trabajo, se establece una división en tres grupos:

1. Los errores pequeños, que serían los debidos a la inexactitud propia de los aparatos de medida.
2. Los errores groseros, que serían los asociados al mal funcionamiento o rotura de los equipos de medida.
3. Los datos faltantes, que serían principalmente el resultado de fallos en los canales de comunicación.

Una clasificación alternativa es la que se propone en [Zarco]. Según ésta, los errores de las medidas (diferencia entre el valor medido y el valor verdadero) se pueden clasificar de acuerdo al siguiente criterio:

1. Ruido normal, que sería aquel para el que el valor absoluto del error es inferior a 5 veces la desviación típica.
2. Error grosero; aquel para el que el valor absoluto del error está comprendido entre 5 y 20 veces la desviación típica.
3. Error extremo; aquel para el que el valor absoluto del error es superior a 20 veces la desviación típica.

Una buena parte de los errores de los tipos B y 2 son problemáticos, dado que pueden dar lugar a estimaciones del estado del sistema desviadas y poco fiables. El análisis de este tipo de errores exige, en primer lugar, la detección e identificación de los datos erróneos para, posteriormente, proceder a su eliminación o sustitución. Aunque el estudio realizado en esta Tesis no se centra en este tipo de datos, cabe mencionar que, entre las décadas de los

70 y los 90, se publicaron decenas de trabajos sobre la detección e identificación de bad data en la estimación de estado y también se propuso una formulación como alternativa para el cálculo de los datos a sustituir [García].

En el contexto de esta Tesis, son especialmente interesantes los errores de los tipos C y 3, así como algunos de los tipos B y 2. Antes de ser introducidos en el estimador de estado, los datos procedentes de las medidas se filtran para detectar los faltantes o aquellos que contienen errores gruesos evidentes. La alternativa de prescindir o eliminar estos datos no es deseable, ya que puede provocar problemas como hacer que parte de la red sea no observable o que se complique el proceso de cómputo del estimador [Lo 2]. En estos casos, una posible solución que se plantea para reemplazar los datos es la de recurrir a las pseudomedidas, que son valores que se obtienen a partir de datos históricos. K.L. Lo et al. [Lo 2] presentan dos algoritmos de sustitución de datos de estas características basados en técnicas de predicción de la demanda. En el primero de ellos, el modelo utilizado se plantea de utilidad a muy corto plazo (inferior a 10 minutos) y asume que no hay cambios significativos en la demanda. El segundo modelo se plantea a corto plazo (hasta una hora) y en él sí que se tienen en cuenta variaciones incrementales en la demanda.

La estimación de estado a la que se viene haciendo referencia es la que se conoce como estática; es decir, que, cuando se procesa un nuevo conjunto de medidas, los estados estimados previamente no se consideran para obtener la estimación del estado actual del sistema. Los estimadores capaces de proporcionar información válida a partir de una sucesión de estados estáticos que se van produciendo en el tiempo reciben el nombre de estimadores de estado dinámicos, para los cuales se ha acuñado el término Forecasting-Aided State Estimation (FASE) [Coutto 1]. Precisamente, una de las ventajas de estos estimadores es que su carácter predictivo permite sortear el problema de los datos faltantes, al ser éstos sustituidos por los valores proporcionados de forma predictiva por el estimador [Huang 1]. En la literatura, se pueden encontrar algoritmos para FASE basados en filtros de Kalman [Leite], [Valverde], redes neuronales y reconocimiento de patrones [Alves 1], [Alves 2], [Alves 3] o lógica difusa [Lin 1], [Huang 2].

4.2.2. Las unidades de medida fasorial (phasor measuremente units –PMU).

Son dispositivos que, merced a su sincronización mediante GPS, permiten obtener el valor eficaz y el ángulo de desfase de las tensiones y corrientes del sistema eléctrico. La frecuencia con la que son capaces de proporcionar esta información puede estar en torno a dos órdenes de magnitud por encima de la que permiten los sistemas SCADA tradicionales, por lo que se consigue obtener mucha más información y con mayor exactitud. La aplicación de la tecnología PMU (también referida en ocasiones como tecnología synchrophasor [Huang 1], [Jones] en los sistemas eléctricos es cada vez más importante para la monitorización del sistema y en aplicaciones de control y de protección [Shi]. Asimismo, la irrupción de esta tecnología también ha provocado un notable impacto en la estimación de estado [Huang 1], dado que se simplifica el proceso de cómputo, al no ser necesarias iteraciones, y se mejora el procesamiento de bad data [Chen 1], [Chen 2], [Zhu], [Coutto 2], [Shi]. Aunque la previsión es que la evolución hacia una red más inteligente hará que el despliegue de estos dispositivos en el futuro sistema eléctrico sea masivo, mientras tanto, se

han desarrollado alternativas para estimadores de estado que permitan la convivencia de las medidas de PMUs con las tradicionales del SCADA [Shi]. En cualquier caso, un estimador de estado que utilice las medidas procedentes de PMUs depende, de igual forma que los demás, de que los datos sean consistentes y fiables y se da la circunstancia de que estos dispositivos, al trabajar con una alta frecuencia de aporte de datos, son más vulnerables a congestiones de la red, fallos en los equipos, configuraciones defectuosas, etc. En [Ghiocel], se subraya la calidad que deben tener los datos asociados a PMUs y en [Jones] se propone un algoritmo de acondicionamiento de estos datos basado en un filtro de Kalman. Previamente a esta etapa de acondicionamiento, se plantean dos validaciones de los datos de entrada: una primera basada en la simple verosimilitud de los datos y una segunda basada en evaluar la relación señal-ruido (signal-to-noise ratio – SNR) de los mismos, sobre la consideración de que tanto el valor eficaz de la señal como su fase se pueden interpretar como valores de señales de tensión continua.

4.2.3. La calidad de onda.

En los estudios sobre este tema, es importante la etapa de pre-procesamiento que se da a la información disponible y que suele incluir la detección de outliers y el tratamiento de los datos faltantes [Terzija], [Lin 2]. Esta etapa de pre-procesamiento puede, también, añadir pasos adicionales interesantes como el borrado de datos que se presenta en [Yang]. En este trabajo, se toma como referencia para el borrado la clasificación de las variaciones de tensión (variaciones lentas, huecos e interrupciones de corta duración) adoptada en la normativa sobre la calidad del suministro eléctrico [IEEE Std. 1159]. De acuerdo a esta norma, se definen como eventos transitorios las variaciones de tensión con una duración comprendida entre medio ciclo de la forma de onda y un minuto, de modo que se borran los datos correspondientes a los eventos que excedan ese rango. Blair et al. [Blair] realizan un estudio sobre la detección y corrección de errores para la correcta monitorización de variables en la calidad del suministro eléctrico. Teniendo en cuenta las tendencias (diaria, semanal e, incluso, anual) y dependiendo de la extensión de los datos faltantes, se propone que los datos individuales sean obtenidos mediante la interpolación lineal entre datos adyacentes. Se advierte de la conveniencia de este método para valores intermedios, pero no para valores máximos o mínimos. En estos casos, se recomienda recurrir a un proceso iterativo que parta de la localización previa de todos los datos sospechosos. En [Lin 2], previamente a la aplicación de técnicas de minería de datos para el análisis de la calidad del suministro eléctrico, se lleva a cabo un proceso de imputación de datos faltantes mediante un procedimiento recursivo basado en la similitud. Partiendo de un periodo de muestreo de cinco minutos, se opera de la siguiente forma:

1. Para imputar los datos faltantes dentro de un intervalo de hasta una hora, se revisan los datos de las dos horas anteriores.
2. Si los datos faltantes son de entre una hora y un día se distingue entre si se trata de días laborables o de fin de semana. En el primer caso se promedian los datos de los dos días anteriores en los mismos instantes. En el segundo caso, se toman los datos de los fines de semana previos.

3. Por último, si los datos faltantes son de más de un día, se toma el mismo día del mes anterior.

4.2.4. Desarrollo de Medidores Inteligentes.

El despliegue de los medidores inteligentes en el lado del consumidor está creando redes de sensores que proporcionan muchos nuevos datos y en mucha mayor cantidad. Hay muchas causas por las que los datos faltantes pueden aparecer en estas redes: que los sensores se queden sin alimentación, que haya interferencias con otros dispositivos electrónicos o que falle la sincronización en el sistema de transmisión de datos. En estos casos, la consulta del usuario no puede ser respondida o, eventualmente, se responde con retraso; de cualquier modo, no suele ser conveniente que los sensores reenvíen los datos, puesto que la red pierde ancho de banda y se prolonga el tiempo de respuesta. En consecuencia, y dado que estas redes de sensores son partes fundamentales de las AMIs y de las redes inteligentes, el problema de los datos faltantes se ha trasladado, también, a estas infraestructuras y a otros campos como el de la previsión de la demanda. En este contexto, la importancia de los datos faltantes para poder realizar un correcto análisis de la información ha sido resaltada en diversos estudios [Lu], [Lankutis], [Jha]. Halatchev y Gruenwald [Halatchev] plantean la estimación de los datos faltantes procedentes de un sensor concreto en presencia de otros grupos de datos que puedan estar relacionados con el grupo de los faltantes. La técnica propuesta (Window Association Rule Mining) usa reglas de asociación para identificar los sensores que proporcionan los mismos datos (sensores relacionados), un determinado número de veces, en una ventana que se va deslizando. Los datos suministrados por estos sensores relacionados son los que se utilizan para la estimación de los datos faltantes. Las principales limitaciones de este método es que sólo es capaz de detectar las relaciones entre dos sensores e ignora los casos en los que los datos faltantes están relacionados con múltiples sensores. Sólo encuentra estas relaciones cuando los dos sensores proporcionan el mismo valor, ignorando los casos en los que los datos faltantes puedan ser imputados a partir de las relaciones entre sensores que puedan dar valores distintos [Jiang 1]. La gran cantidad de datos proporcionados por los medidores inteligentes en el lado de la carga permite profundizar en aspectos como la mayor desagregación en la previsión de la demanda o el conocimiento más detallado de los perfiles de carga o consumo. En consecuencia, esto ha dado lugar, también, a que los datos faltantes se tengan en cuenta en estos campos de investigación. Es el caso, por ejemplo, de Harvey et al. [Harvey], que proponen un método de clasificación de los perfiles de carga residenciales a partir de los datos de AMIs. En dicho método, no se utiliza ningún procedimiento de imputación de los datos faltantes, pero sí que se estudia su existencia, en distintos porcentajes, de cara a valorar la inmunidad del método ante la pérdida de datos.

4.2.5. Previsión de la Demanda de Instalaciones Eléctricas.

La publicación del Instituto Edison [EEI 2], en Estados Unidos, es una guía que recopila una serie de buenas prácticas industriales, cuyo objetivo es proporcionar una uniformización de las mismas entre todos los agentes implicados en los, cada vez más desagregados, servicios de medida de electricidad. Entre las prácticas que se proponen, están las que se refieren a la imputación de datos faltantes en las curvas de previsión de demanda.

En esta guía, aparecen de forma muy detallada los procedimientos a seguir en el caso de que los intervalos sean más cortos o más largos de dos horas. En el primer caso, se recurre a la interpolación lineal particularizando los casos en los que los datos faltantes puedan estar al principio, en el centro o al final del intervalo objeto de estudio. En el caso de intervalos superiores a dos horas, la aproximación que se realiza para imputar los datos se basa en usar el valor medio de días de referencia seleccionados. El documento especifica claramente el procedimiento y las reglas a seguir para la definición de los días de referencia, los cuales se clasifican de acuerdo a si son del mismo día de la semana (si son días laborables) y a si son del mismo tipo de día (laborable, fin de semana o festivo) y se escogen aquellos que resulten cronológicamente más cercanos al día en el que se ha de realizar la imputación. A partir de los días de referencia, el informe establece la forma en la que se ha de construir el perfil de demanda diario y con él la cumplimentación de los datos faltantes. En la misma línea de predicción de la demanda, Quilumba et al. [Quilumba 1], [Quilumba 2], proporcionan una visión general del pre-procesamiento que se debe dar a los datos de AMIs para obtener previsiones de mayor calidad. Aunque no se proporcionan detalles concretos acerca del tratamiento de los datos (se alude, sin concretar, al reemplazamiento por el valor medio, la mediana o por el valor más probable), sí se plantea una diferenciación entre el tratamiento de los datos perdidos como consecuencia de la interrupción de la comunicación con el medidor (intervalo de datos) y los debidos a la pérdida de información de los datos de sólo alguno de sus canales. La previsión de la demanda a muy corto plazo, así como el modelado y la estimación de estado del sistema de distribución se combinan en los trabajos publicados por Peppanen et al. En ellos, se encuentra presente el tratamiento de los datos faltantes, con distinta profundidad, obtenidos en la infraestructura de medida del sistema de distribución del campus del Georgia Institute of Technology (Atlanta –EEUU), el cual abarca más de 200 edificios. En [Peppanen 1], los datos procedentes de AMIs se someten a una depuración previa a su uso para validar el modelo de red de distribución propuesto, si bien no se entra en detalles acerca del procedimiento de imputación de datos. En [Peppanen 2], se propone un método de estimación de parámetros del sistema de distribución cuya validación se realiza a partir de los datos de los medidores inteligentes. Previamente, el reemplazamiento de los datos faltantes se realiza mediante la generación de pseudomedidas con una combinación ponderada de datos históricos y de datos interpolados o extrapolados a partir de medidas anteriores o posteriores. Este procedimiento se detalla en una publicación posterior [Peppanen 3], en la que aparece denominado como método de imputación de datos por promedio óptimamente ponderado. Entre las ventajas de dicho método, están el que no se requiera específica información adicional acerca de las medidas ni otras variables adicionales. Para ello se aprovechan dos características que poseen típicamente las cargas con las que se trabaja y que son las siguientes: por un lado, los datos no presentan tendencias en las que se observen cambios abruptos en los intervalos de tiempo cortos, de modo que esta apariencia "bastante continua" permite suponer que las muestras faltantes tienen similares características a las de los datos adyacentes disponibles. Por otro lado, dado que la evolución de las cargas está fuertemente condicionada por patrones de consumo humanos, los datos tienden a presentar características similares a lo largo de periodos de tiempo con similar actividad humana. Por ejemplo, los datos de los días de semana no festivos son diferentes de los de fin de semana, o los datos matutinos son distintos de los vespertinos.

Partiendo de estas premisas, el método de imputación propuesto obtiene el dato faltante mediante la suma del dato obtenido mediante interpolación lineal (más adecuada para los intervalos cortos) con el dato obtenido del promedio histórico (más adecuado en los intervalos largos). Ambos sumandos están afectados por sendos coeficientes de ponderación que dan mayor peso a uno u otro término. Los coeficientes de ponderación obedecen a expresiones exponenciales negativas cuyos exponentes dependerán de las características y longitud del intervalo de datos faltantes. De esta forma, se dará mayor peso al término de la interpolación lineal en los intervalos cortos y al término del promedio histórico en los largos.

El estudio de la previsión de demanda se ha particularizado para el caso de determinadas instalaciones, en las que el uso de medidores inteligentes puede permitir la construcción de perfiles de demanda específicos para ciertos tipos de consumidores. Majidpour et al. [Majidpour] realizan un estudio comparativo de cinco métodos de imputación aplicados a datos faltantes procedentes de instalaciones de carga de vehículos eléctricos. El objetivo final es suministrar los datos a un algoritmo que permita predecir la carga, en las siguientes 24 horas, de cada punto de suministro de los vehículos eléctricos. Los métodos estudiados utilizan imputación constante, media, mediana, de máxima probabilidad e imputación múltiple; sin embargo, los resultados que se presentan no se pueden considerar concluyentes al no disponerse de suficiente número de casos. Otro trabajo en el que se tiene en cuenta la imputación de datos en el estudio de estaciones de carga de vehículos eléctricos es el abordado por Soltani y Giannakis [Soltani]. En este trabajo, se trata el problema del aprendizaje en presencia de datos faltantes y, concretamente, se aplica al seguimiento de los hábitos de los usuarios de vehículos eléctricos cuando llevan a cabo las operaciones de recarga. En el estudio, se desarrolla un modelo que permite obtener la probabilidad con la que cada usuario de vehículo eléctrico realiza la recarga asumiendo que, en cada instante, las decisiones que tome el usuario para esas recargas pueden ser datos que se pierden en distinto número y ubicación.

La multiplicidad de medidores y la complejidad de las redes de transmisión de datos en AMIs han dado lugar a que las técnicas usadas para la limpieza, purificación o depuración de datos (data cleaning o data cleansing) hayan ido cobrando creciente interés para poder asegurar la calidad y el eficiente procesamiento de la información. A este respecto, Jiang y Chen [Jiang 2] proponen dividir en dos categorías los aspectos a considerar en los sistemas de procesamiento de señales cuando se trata de redes de sensores: por un lado, los aspectos referentes a data cleansing cuando se trata de bases de datos tradicionales. Por otro, los aspectos a tener en cuenta, de forma específica, cuando se trata del procesamiento de señales en redes de sensores que, eventualmente, pueden ser inalámbricas. En el caso de las bases de datos tradicionales, para el tratamiento de la información se tienen en cuenta los datos faltantes, la información duplicada, los datos incompletos o corruptos y la detección de ruido y outliers. Aspectos adicionales a tener en cuenta en las redes de sensores son el procesamiento on-line (a diferencia de las bases de datos tradicionales donde habitualmente es off-line), la desagregación de los datos en tiempo y en espacio y la contextualización de la información. Precisamente, Jiang y Chen [Jiang 2] proponen un proceso de imputación basado en minería de datos que trabaja con medidas previamente enriquecidas con información del dominio de cada sensor, de modo que la información suministrada por cada

uno de ellos quede contextualizada en espacio y tiempo y se facilite el reconocimiento de patrones.

Las técnicas de data cleansing se han aplicado en algunos trabajos a los datos de las curvas o perfiles de carga para poder realizar una predicción de la demanda de mayor calidad [Mateos 1], [Mateos 2]. En estos casos, el proceso completo de data cleansing incluye la validación de datos, la eliminación de ruido (datos corruptos en general, que incluyen los datos faltantes) y la preparación para el posterior análisis (que incluye la sustitución en la base de datos de aquellos corruptos). Posiblemente, el estudio más referenciado en este campo sea el de Chen et al. [Chen 3]. La esencia del método propuesto por estos autores para la detección de datos corruptos está en el modelado de los patrones de la curva de carga. El modelo obtenido puede ser usado para determinar la presencia en los datos de desviaciones anormales de los patrones y, de esta forma, identificar los datos corruptos. La clave, entonces, es encontrar la curva estimada que mejor se ajuste a la muestra de datos disponibles y establecer un intervalo de confianza respecto a ella, de modo que los puntos que caigan fuera de dicho intervalo serán calificados como corruptos y serán reemplazados por el valor estimado a partir de la curva de ajuste. Para la obtención de esta curva (proceso conocido como smoothing), los autores proponen el uso de la regresión no paramétrica, ya que el hecho de que las curvas de carga no obedezcan a una relación simple desaconseja el uso de la regresión paramétrica. Concretamente, las técnicas que utilizan para el smoothing se basan en el uso de funciones Spline y Kernel. Ambas técnicas son aplicadas con éxito a datos reales de curvas de carga de la British Columbia Transmission Corporation, tanto en casos de datos localmente como globalmente corruptos. Hoverstad et al. [Hoverstad] también utilizan el método anterior, con funciones Spline, con objeto de depurar las curvas de demanda antes de utilizarlas para hacer previsiones de la misma a corto plazo (24 horas). La principal aportación de estos autores, en la etapa de pre-procesamiento de los datos, es la de darles cierto tratamiento previo antes de obtener la curva de ajuste. De esta forma, consiguen mejorar la depuración de los datos en la medida que evitan la identificación como outliers de algunas puntas de carga. El tratamiento previo que proponen consta de dos etapas: una primera que consiste en eliminar de cada ciclo de carga diario, del periodo a estudiar, su valor medio. Se obtiene, de este modo, una curva de carga de valor medio nulo para cada día de estudio y con ello se consigue eliminar variaciones estacionales, haciendo que las diferencias en la señal entre días de verano y de invierno sean menores. La segunda etapa consiste en obtener lo que los autores denominan día o semana prototipo. El día prototipo se obtiene promediando, para cada hora del día, los valores de consumo que presentan las curvas obtenidas en la primera etapa (si en lugar de trabajar con las 24 horas del día, se trabajara con las 168 horas de la semana, se obtendría la semana prototipo). Restando de la curva de la primera etapa la de la segunda, se obtiene una señal sobre la que se aplica la regresión y la detección de los valores anómalos. La curva de carga depurada se obtendría revirtiendo todo el proceso anterior. En la misma línea de trabajo Tang et al. [Tang] proponen un método de detección de outliers basado en la identificación de tres posibles casos: a) datos con distribución normal; b) datos con distribución gamma y c) datos en lo que los autores llaman “small-size portrait”, los cuales se diferencian de los dos tipos anteriores por constituir una muestra de mucho menor tamaño. En el primer caso, el criterio de detección se basa en los parámetros media y desviación típica de la distribución normal. En el segundo

caso, los parámetros de referencia son β y γ de la distribución gamma. Por último, en el tercer caso, el criterio de detección se establece a partir de los valores que toman ciertas relaciones entre los percentiles 25 y 75. Una vez localizados los datos corruptos, los autores no determinan un criterio concreto para la imputación o el reemplazo, dejando abiertas las siguientes opciones:

1. En los casos primero y tercero, reemplazar los datos faltantes por el valor medio de todos los datos del intervalo.
2. En el segundo caso, reemplazar por el valor β/γ.

5. INSTALACIÓN DEL EDIFICIO SEVERO OCHOA

Como se ha venido refiriendo, gran parte de la investigación realizada utiliza el edificio Severo Ochoa como instalación de referencia. Dicho edificio, está ubicado en el campus universitario del Cristo, en Oviedo, y alberga a los Servicios Científico Técnicos. Estos Servicios son unidades de apoyo a la investigación dotadas de un equipamiento científico altamente sofisticado y tienen por objeto dar soporte a grupos de investigación de la Universidad de Oviedo o de otras instituciones públicas y empresas privadas. Asimismo, este edificio da cabida al Servicio de Informática de la Universidad de Oviedo. En su diseño eléctrico, se ha tenido en cuenta poder garantizar el suministro ininterrumpidamente a la instalación, especialmente al Servicio de Informática ubicado en la planta cuarta. Por esta razón, se ha dotado a la instalación de un sistema de alimentación ininterrumpida (SAI) y de un grupo electrógeno.

5.1. Elementos de la Instalación.

La instalación eléctrica del edificio Severo Ochoa está formada por los siguientes elementos:

Centro de transformación.

El edificio cuenta con un centro de transformación con acceso directo para la compañía suministradora, según normativa vigente, y compuesto por un conjunto de celdas prefabricadas bajo envolvente metálica, tipo monobloque, para alta tensión. Sus principales características son:

- Potencia: 1250 kVA.
- Trasformador seco, clase F, modelo Trihal de Merlin Gerin.

- Tensión nominal primaria: 22 kV
- Aislamiento: 24 kV(125 kV-50kV)
- Tensión nominal secundaria: 400/230V
- Frecuencia: 50 Hz

Figura 4: Celdas de corte y medida del centro de transformación

Figura 5: Transformador seco 1250 kVA

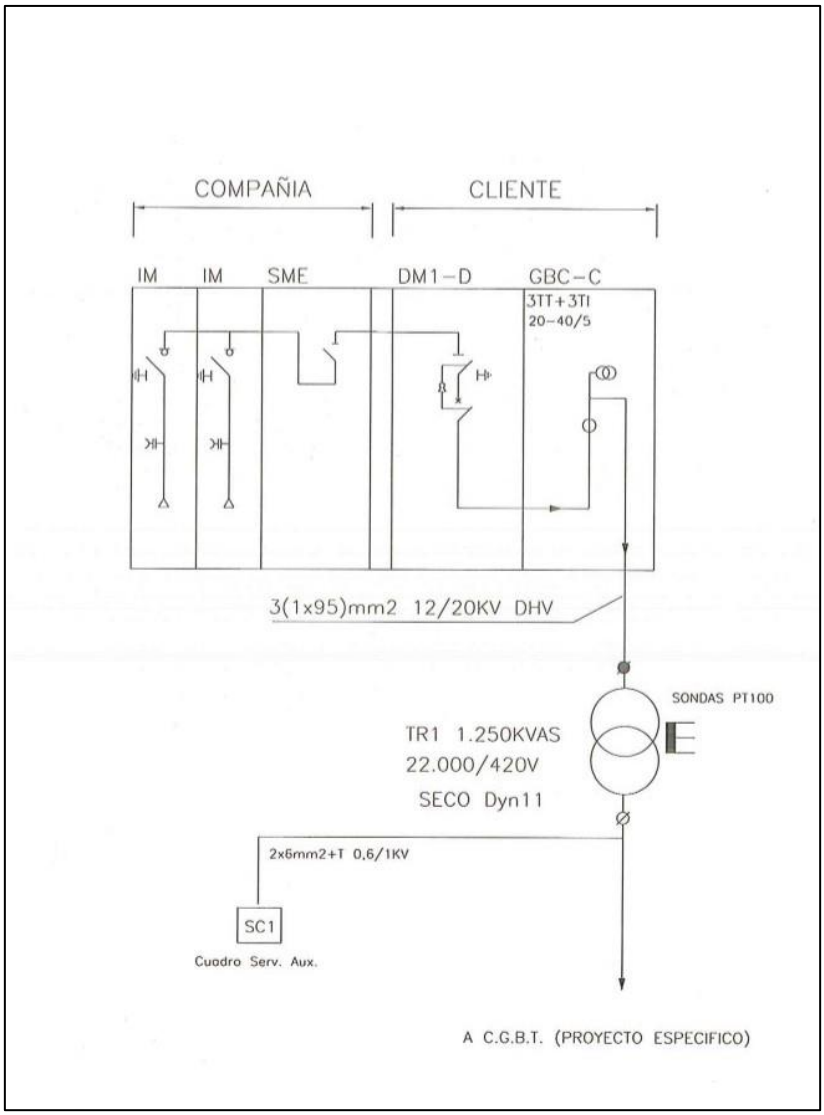


Figura 6: Esquema unifilar del centro de transformación (según proyecto de A.T. suscrito por D. Víctor Montes y visado con Nº 035605 por el C.O.I.T.I del Principado de Asturias, 1 de agosto de 2003)

Cuadro general de baja tensión (CGBT)

El edificio fue inaugurado en el año 2003. Esto significa que su diseño y ejecución fue llevado a cabo de acuerdo al Reglamento Electrotécnico para Baja Tensión (REBT) del año 73, ya que los trabajos de construcción del edificio fueron previos a la entrada en vigor del actual REBT del año 2002.

El cuadro general de baja tensión (CGBT) es el origen de todas las líneas generales de alimentación a los cuadros secundarios y a las acometidas independientes. Contiene, bajo la misma envolvente, tres subembarrados diferentes: red, grupo y SAI, así como la conmutación automática red-grupo. El CGBT está alimentado desde el centro de transformación y desde el cuadro general del grupo, realizándose en él la alimentación a los SAI desde la salida del grupo e incluyendo en él las líneas de salida del servicio de alimentación ininterrumpida. Según el proyecto específico de baja tensión, la potencia simultánea que permite alimentar es de 750 kW. La línea de alimentación desde el

transformador al cuadro general consiste en cable RZ1-k de dimensiones: 3(4x240) + (2x240) mm^2 y cuenta con un interruptor general tetrapolar de 1600 A. La línea del grupo es de dimensiones: 3x240 + 1x85 mm^2, y cuenta con un interruptor tetrapolar de 630 A.

Los dispositivos de protección del CGBT se pueden observar en la figura 7. Los interruptores automáticos son magnetotérmicos de caja moldeada y regulables en intensidad. Las protecciones en cascada cumplen criterios de selectividad amperimétrica y cronométrica y las actuaciones de los transformadores toroidales diferenciales sobre los interruptores automáticos son por bobina de emisión.

Todos los cuadros de planta, así como los cuadros secundarios de laboratorios y locales de características especiales están dotados de interruptores automáticos magnetotérmicos en cabecera y de interruptores magnetotérmicos en los circuitos de salida. También cuentan con interruptores diferenciales de 30mA de sensibilidad, que agrupan a cada grupo de salida. En el caso de alimentación a motores y algún equipamiento eléctrico especifico que así lo ha requerido, son de 300mA de sensibilidad. La implantación de estos cuadros se ha realizado teniendo en cuenta el uso de cada dependencia. En general, existe un cuadro por cada grupo de dependencias similares, y uno por local cuando este tiene suficiente identidad. Este es el caso de la sala de servidores del servicio de informática o el laboratorio de espectrometría de masas etc.

Figura 7: Cuadro general de baja tensión del edificio Severo Ochoa

Las líneas generales de distribución parten del CGBT, ubicado en el sótano del edificio y están realizadas en cable de cobre de polietileno reticulado (RZ1-0.6/1kV).

Grupo electrógeno.

El edificio, en el local anexo al centro de transformación, cuenta con un grupo electrógeno (Figura 8) que, además de suministrar el servicio complementario para circuitos

de alumbrado, vías de evacuación, pasillos, etc., y cumplir con la normativa vigente de evacuación de locales de pública concurrencia, permite suministrar energía a los SAI del servicio de informática. De esta manera se garantiza el suministro a este servicio en caso de corte general.

La conmutación integrará las funciones automáticas de mando del grupo electrógeno y de desconexión y reconexión de los circuitos no prioritarios, pudiendo regularse las temporizaciones de permutación.

Este grupo es de una potencia nominal de 350 kVA, y de una tensión estándar de 400 V/230 V.

Figura 8: Imagen del Grupo Electrógeno

SAI

Para dar alimentación al equipamiento informático se cuenta en la instalación con un SAI general (Figura 9) que dispone de un by-pass automático, así como de uno manual para realizar las tareas propias de mantenimiento. Las características del equipo son las siguientes:

- Potencia de salida: 120 kVA (cos φ = 0,80)
- Tensión de alimentación: 380 V ±15%
- Frecuencia de alimentación: 50 Hz ± 5%

- Tensión de salida: 400 V (ajustable entre 380V y 415V)
- Sobrecarga: 150% para 1 minuto, 125% para 10 minutos.

El conjunto de la instalación de SAI cuenta con una autonomía de 10 minutos.

Figura 9: Imagen del SAI general del edificio

El centro también cuenta con una sala de SAIs (Figura 10) destinados a suministrar alimentación ininterrumpida al sofisticado equipamiento científico ubicado en el edificio. Está instalación ha ido variando a lo largo del tiempo a medida que se han ido equipando los laboratorios, ajustándose a las necesidades de cada momento.

Figura 10: Sala de SAI de los Servicios Científico Técnicos.

Batería de Condensadores.

La instalación dispone de una batería de condensadores (Figura 11) para mejorar el factor de potencia. Se encuentra conectada al CGBT y cuenta con protección magnetotérmica en el cuadro. El regulador automático está alimentado desde la acometida general de red y los transformadores de intensidad están situados en esa misma acometida.

La batería instalada cuenta con una etapa fija de 60 kvar y otra automática de 300 kvar compuesta por 5 escalones de 60 kvar.

Figura 11: Batería de condensadores instalada en el edificio.

5.2. Variables Objeto de Estudio y Equipos de Medida

Los dispositivos instalados en el edificio Severo Ochoa y empleados en la medida y registro de datos, son equipos específicos para variables eléctricas. Todos ellos tienen características comunes en cuanto al mecanismo de medida; es decir, todos ellos aportan valores de: tensión fase-neutro, tensión entre fases, corriente de línea, entrada/salida de potencia, entrada/salida de energía, potencia reactiva, energía reactiva de entrada/salida, potencia aparente, energía aparente, factor de potencia y frecuencia. Todos estos valores se registran en tiempo real, devolviendo el equipo los valores instantáneos, máximos, mínimos y promedio según datos aportados en la Tabla 2.

La Tabla 3 muestra la precisión para cada uno de los dispositivos instalados durante las diferentes mediciones eléctricas. Los valores especificados en porcentaje aplican sobre el valor de la lectura. En todos los casos se trata de medidores de alto rendimiento,

catalogados como de clase 0.2 por los estándares de medida ANSI C 12.20 y el IEC 62053-22.

Variable Medida	Instantánea	Valor Promedio	Valor Máximo	Valor Mínimo
Tensión L-N	√		√	√
Tensión L-L	√		√	√
Corriente por fase	√	√	√	√
Corriente por neutro	√			
Potencia activa	√	√	√	√
Potencia reactiva	√	√	√	√
Potencia aparente	√	√	√	√
Factor de potencia	√	√	√	√
Energía activa máxima	√			
Energía activa mínima	√			
Energía activa acumulada	√			
Energía reactiva máxima	√			
Energía reactiva mínima	√			
Energía reactiva acumulada	√			
Energía aparente	√			
Frecuencia	√		√	√
THD (%)	√		√	√
Ángulos de voltaje y de corriente	√			
Nivel de carga (%)	√			
Forma de onda	√			

Tabla 2: Valores medidos por los equipos instalados en el edificio Severo Ochoa.

Los cuatro dispositivos instalados en el edificio Severo Ochoa y referenciados en el presente estudio pueden realizar todas las mediciones mencionadas y, además, cada uno de ellos tiene capacidades adicionales que se describen en los siguientes apartados.

Variable	Unidad	S100	S200	NEXUS 1252		MP200
				200 ms	1 s	
Tensión L/N	V, kV	0.1%	0.1%	0.1%	0.05%	0.3%
Tensión L/L	V, kV	0.1%	0.2%	0.1%	0.05%	0.5%
Corriente	A, kA	0.1%	0.1%	0.1%	0.025%	0.3%
Potencia activa	W	0.2%	0.2%	0.1%	0.06%	0.5%
Energía activa	Wh	0.2%	0.2%	N/A	0.04%	0.5%
Potencia reactiva	var	0.2%	0.2%	0.1%	0.08%	1.0%
Energía reactiva	varh	0.2%	0.2%	N/A	0.08%	1.0%
Potencia aparente	VA	0.2%	0.2%	0.1%	0.1%	1.0%
Energía aparente	VAh	0.2%	0.2%	N/A	0.08%	1.0%
Factor de Potencia	+/−0.5-1	0.2%	0.2%	0.1%	0.08%	1.0%
Frecuencia	Hz	1.10^{-2}	$+/-3.10^{-2}$	3.10^{-2}	1.10^{-2}	$+/-1.10^{-2}$

Tabla 3: Precisión de medida de cada variable en los distintos dispositivos.

5.2.1. Shark 100 (S-100)

El Shark 100 (Figura 12) es un medidor multifuncional diseñado para integrar en paneles de control o en cuadros eléctricos y está especialmente indicado para proporcionar información a los equipos de mantenimiento y operación de las instalaciones eléctricas. Una de las opciones incluidas para este equipo es el puerto óptico IrDA, que permite la programación del dispositivo en remoto utilizando un portátil o un asistente digital personal (PDA).

Figura 12: Dispositivo Shark 100

Además, el equipo incorpora tecnología V-Switch. Esta herramienta permite a los usuarios actualizar e incluir funciones nuevas con comandos de programación, incluso después de la instalación del dispositivo, sin necesidad de desmontarlos de su ubicación.

Asimismo, este dispositivo cuenta con un puerto RS485, lo que permite conectar uno más medidores Shark con un PC u otro dispositivo en un sitio alejado, haciendo la comunicación factible mediante protocolos Modbus o DNP 3.0. Además del RS485, el dispositivo también incorpora un pulso KYZ, protocolo homologado por las distribuidoras eléctricas para información instantánea sobre consumo de energía.

Figura 13: Sistema de comunicación de los dispositivos

5.2.2. Shark 200 (S-200)

El Shark 200 es un dispositivo compacto, similar al anterior físicamente, e igualmente utilizado para mediciones de potencia y energía. Proporciona además de las medidas de facturación y calidad de suministro un sistema de grabación avanzada de datos, que permite el registro de históricos. Este equipo también incluye la tecnología V-Switch del Shark 100 y está diseñado con arquitectura modular capaz de escalarse para funciones de comunicación altamente desarrolladas.

Como diferencia fundamental con el equipo anterior, cabe destacar que el Shark 200 incluye el registro de la forma de onda, lo que permite mejorar el análisis de la calidad de la energía.

Incorpora además la grabación de datos para las tendencias históricas, alertas de límite, desviaciones de entrada/salida y categorización de diferentes medidas asociadas a las variables objeto de estudio.

El equipo puede hacer un registro de curva CBEMA (Computer & Busines Equipment Manufacturer`s Associatión) describiendo las variaciones de tensión en la instalación que pueden ser toleradas sin interrupción del funcionamiento y fijando el tamaño

y duración de los posibles eventos que puedan ocurrir. También ofrece un análisis de armónicos desde el orden 40 hasta 255 para las entradas de corriente y tensión.

En cuanto a la comunicación, (Figura 13), este modelo incluye las siguientes características ya mendionadas en el apartado anterior:

- Un puerto RS485 permite la comunicación con protocolos Modbus o protocolo de red distribuida (DNP) v.3.0.
- Pulso KYZ - este dispositivo incorpora salidas de pulsos asignado al total de energía.
- Puerto óptico IrDA.

5.2.3. Shark MP200

El modelo MP200, (Figura 14) tiene las mismas prestaciones que el S-200 incorporando además información de uso de la energía de ocho circuitos trifásicos o de veinticuatro sistemas monofásicos. El sistema MP200 puede crear informes precisos del uso de la energía, analizar la demanda pico y proporcionar las señales de control para limitar este pico de demanda. Asimismo, permite el análisis de la facturación basada en el uso y demanda.

El MP200 ofrece las posibilidades de comunicación mencionadas para los modelos anteriores.

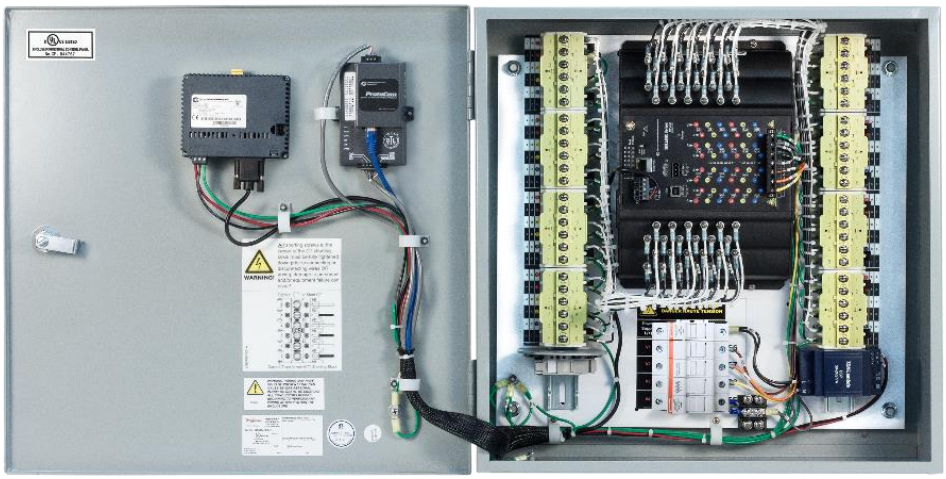


Figura 14: Shark MP200 instalado en cuadro.

5.2.4. Nexus 1252

En términos generales, este dispositivo (Figura 15), tiene características que ofrecen una visión global, tanto de la energía consumida, como de su uso dentro de la instalación. De esta forma se tiene una visualización de la calidad de la energía eléctrica dentro de una instalación receptora avanzada, estableciendo análisis de calidad de potencia estandarizado

con técnicas basadas en las normas EN50160 e IEC61000-4.30. Su uso está especialmente recomendado para la monitorización de transformadores.

El dispositivo es capaz de capturar un máximo de 512 muestras por ciclo, por lo que es capaz de llegar hasta el orden 255, en los armónicos de corriente y tensión. Si es necesario, puede medir los armónicos en tiempo real hasta el orden 128. El dispositivo proporciona el porcentaje THD y el Factor K de armónicos. Además, es posible monitorizar la distorsión de varios elementos de una instalación. Al igual que el anterior dispositivo, el Nexus 1252 también es capaz de hacer un registro CBEMA-ITIC, es decir, captura la magnitud y la duración de los picos y caídas de tensión en la instalación permitiendo realizar gráficamente la curva ITIC y CBEMA y relacionando la magnitud del evento con la duración del mismo.

En cuanto a la comunicación, el dispositivo cuenta con cuatro puertos y cada uno de ellos es capaz de comunicarse en varios protocolos.

Figura 15: Equipo Nexus 1252 instalado en el edificio Severo Ochoa

5.3. Descripción de los Datos obtenidos durante las Mediciones

A partir de las medidas realizadas por los equipos que se han descrito anteriormente, es posible llevar a cabo el análisis y estudio de las variables eléctricas y realizar la gestión energética del edificio, tanto en términos de demanda como de la calidad del suministro. La monitorización de los datos se lleva a cabo en tiempo real, así como la recogida y seguimiento de alarmas en la instalación. Esto permite hacer un seguimiento de todos los parámetros que intervienen en el suministro eléctrico, incluyendo los desequilibrios de corriente o la distorsión armónica. (Figura 16)

Figura 16: Registros gráficos de los datos de intensidades en la instalación

El conjunto de datos empleado para la presente investigación corresponde a mediciones de las siguientes variables: tensión fase-neutro (tres variables), tensión fase-fase (tres variables), la corriente en cada fase (tres variables) y el factor de potencia promedio (una variable) de un suministro eléctrico trifásico del ya referido edificio Severo Ochoa perteneciente a la Universidad de Oviedo, destinado a los Servicios Científico Técnicos y al Servicio de Informática. Los registros fueron tomados cada 15 minutos desde el 27 de noviembre de 2014 a las 18:45 a 31 de mayo de 2015 en 23:45 con el equipamiento de medida descrito con anterioridad. En la figura 17 se puede observar el perfil de carga de la instalación obtenida para la última quincena del año 2016.

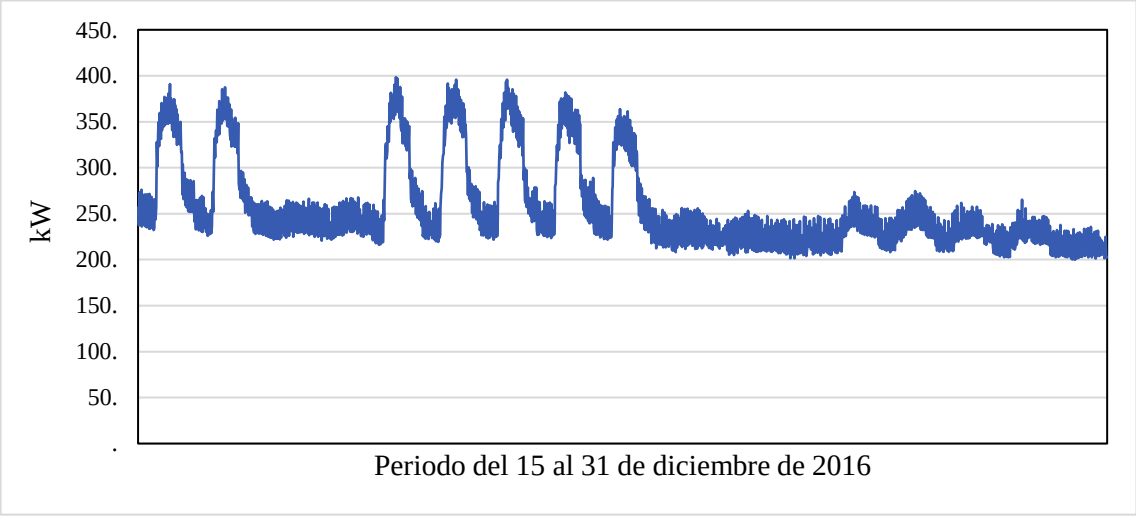


Figura 17: Curva cuarto horaria de la última quincena de diciembre de 2016 para el edificio Severo Ochoa

De la observación de figura 17, se puede afirmar que el edificio Severo Ochoa tiene instalado gran cantidad de equipamiento eléctrico que funciona de manera continua; ya que durante las noches, fines de semana y periodo vacacional la potencia demandada es superior a 200 kW a pesar de estar cerrado el edificio.

Al no estar destinado a actividades docentes, el consumo eléctrico en el edificio Severo Ochoa es bastante estable a lo largo del año, no siendo muy significativo el periodo vacacional de verano como ocurre en otros centros. En el último año, el consumo superó los 2.2 GWh. Esta condición se ha venido manteniendo desde la puesta en funcionamiento del edificio, tal y como puede apreciarse en la figura21 donde se recogen los consumos de los tres últimos años. El pequeño descenso de consumo en la segunda mitad del año 2017 ha estado únicamente motivado por el cese de actividad en dos laboratorios del centro.

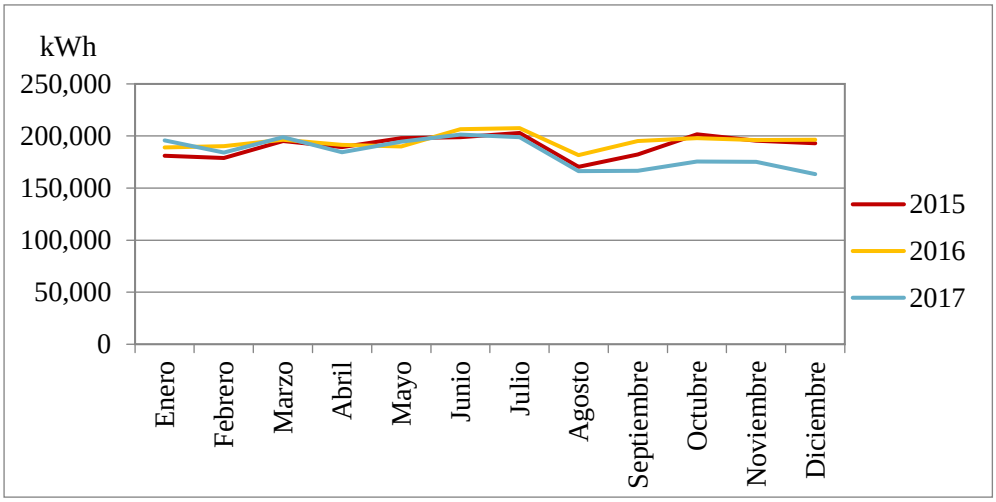


Figura 18: Consumos del Edificio Severo Ochoa en los últimos tres años.

Si se representa la curva cuarto horaria del edificio para el segundo semestre del año 2016, tendría una forma tal y como aparece en la figura 19 y facilitaría la siguiente información:

Demanda máxima:	497.31 kW
Demanda mínima:	205.51 kW
Demanda más frecuente:	248.27 kW
Demanda media:	264.77 kW

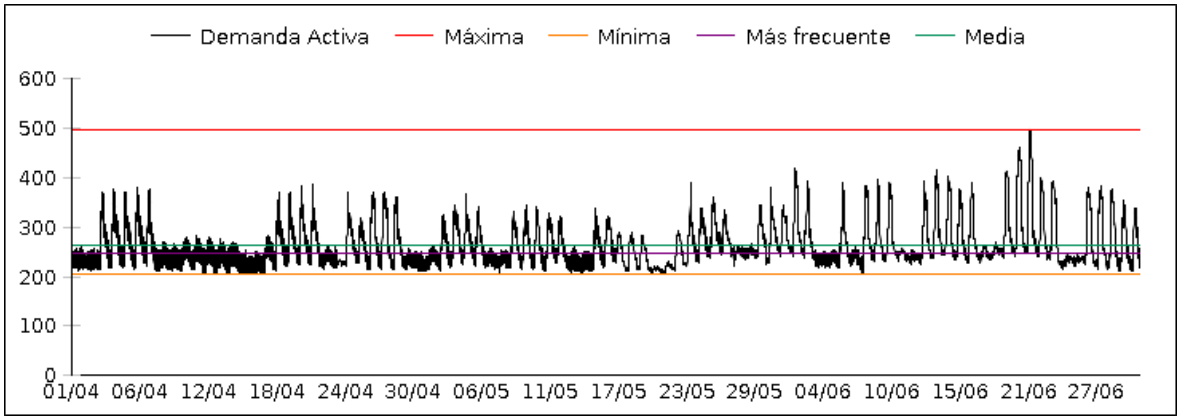


Figura 19: Curva cuarto horaria para el año 2016 del Edificio Severo Ochoa

6. MATERIAL Y MÉTODO

6.1. Filosofía del Trabajo.

Para la elección de los algoritmos de cálculo más adecuados en la imputación de datos faltantes aplicada el estudio de las instalaciones eléctricas de baja tensión de los edificios públicos se ha seguido la siguiente metodología de trabajo:

1º. Se ha elegido un conjunto de datos compuesto de un total de 17.763 muestras tal y como se enunciaba en la descripción de los datos y sobre el que se ha trabajado con los distintos algoritmos de cálculo.

2º. Se han eliminado todas aquellas filas donde había huecos.

3º. Una vez obtenido un conjunto homogéneo de datos, se ha sometido a este a un proceso de eliminación de datos de manera aleatoria. Este proceso presupone que la probabilidad de que haya un dato faltante no depende de las mediciones registradas o no registradas. Se llama falta completamente aleatoria (MCAR). El proceso de eliminación de datos aleatorios se repitió cinco veces para tres diferentes niveles de datos faltantes: 10%, 15% y 20% del total.

4º. Después de cada proceso de eliminación anterior, se probaron los distintos algoritmos objeto de estudio al subconjunto de datos resultante y se comparó entre si el rendimiento obtenido en los distintos métodos.

6.2. Descripción de los Algoritmos

6.2.1. Interpolación IDW

IDW es un método determinista de interpolación espacial. Se basa en la distancia entre la ubicación para la cual se debe interpolar un valor y las ubicaciones de las observaciones registradas. El punto es que la medida observada en un punto concreto presenta una alta correlación con los valores registrados en sitios cercanos. Por lo tanto, es

posible estimar un valor en cualquier punto a través de una combinación lineal de los valores medidos desde puntos próximos.

Este método fue utilizado durante los trabajos realizados al inicio de la investigación [Crespo 1] utilizando como variable de medida la irradiancia solar en las Estaciones Meteorológicas de Galicia, ante la falta de suficientes datos para aplicar en las variables eléctricas del Edificio Severo Ochoa.

Así pues, tal y como se explica en el citado trabajo, si G_E es la estimación de la irradiancia solar en un sitio sin medición, dicha G_E puede calcularse siguiendo las ecuaciones siguientes:

$$G_E = \frac{\sum_{i=1}^{n} W(r_{i,E}) G_i}{\sum_{i=1}^{n} W(r_{i,E})} \quad (3)$$

$$W(r_{i,E}) = \frac{1}{r_{i,E}^{p}} \quad (4)$$

donde:

$G_{i:}$ es el registro de irradiación solar medida en el sitio "i", con i = 1, 2, ..., n.

$W(r_{i,E})$: es la función de ponderación en el sitio i-ésimo.

$r_{i,E}$: es la distancia entre la i-ésima estación de medición de irradiación solar y el sitio de estimación.

p: es el parámetro de potencia utilizado en la interpolación. A menudo se elige p=2 para proporcionar aún más peso a los lugares más cercanos.

Consideramos tres casos: p =1, 1.5, 2. La altitud no se tuvo en cuenta para los cálculos de distancia porque todas las estaciones se encuentran prácticamente a la misma altura sobre el nivel del mar.

6.2.2. Regresión Lineal Múltiple MLR

Los modelos de MLR usan un conjunto de variables independientes que ayudan a explicar la variable independiente.

En este caso, al igual que en el apartado anterior, el método fue utilizado durante los trabajos realizados al inicio de la investigación [Crespo 1] por el mismo motivo. En el estudio, las medidas de las estaciones vecinas se eligieron como variables explicativas debido a las altas correlaciones entre ellas.

El coeficiente de correlación (CC) se expresa de acuerdo con la siguiente ecuación:

$$CC = \frac{\sum_{i=1}^{n}(Gx_i - \overline{Gx_i})(Gy_i - \overline{Gy_i})}{\sqrt{\sum_{i=1}^{n}(Gx_i - \overline{Gx_i})^2 \sum_{i=1}^{n}(Gy_i - \overline{Gy_i})^2}} \quad (5)$$

donde Gx_i y Gy_i son las medidas de diez minutos de la irradiación global en las estaciones x e y, y $\overline{Gx_i}$o $\overline{Gy_i}$ son la media de todas las medidas en dichas estaciones. La regresión lineal múltiple es un método estadístico que, en consecuencia, modela la relación entre una variable dependiente (y) y un conjunto de variables independientes (x1, x2, ..., xp). El modelo se puede representar de la siguiente manera:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (6)$$

donde α se denomina intersección, βi se denominan pendientes o coeficientes, ε es un error con media cero y varianza constante, y se acepta que cada variable independiente tiene una relación lineal con la variable dependiente.

La expresión anterior puede reescribirse en forma matricial de la siguiente forma:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (7)$$

En este estudio y_i son las observaciones de diez minutos de una estación, y x_{ij} son las observaciones de diez minutos de las ocho restantes. Por lo tanto, en este caso n=p=9 y βp es el coeficiente asociado a la ubicación "p".

Para obtener la intersección y los coeficientes tomamos el enfoque de mínimos cuadrados con un intervalo de confianza del 95%. Después de prestar atención a los valores de las estadísticas F y t, los coeficientes no significativamente diferentes de cero se establecen en cero para el modelo.

6.2.3. El algoritmo MICE

El algoritmo MICE (Imputación múltiple por ecuaciones encadenadas) es un nuevo sistema de referencia, desarrollado por van Buuren y Groothuis-Oudshoorn [Buuren 1]. Se trata de un método basado en cadenas de Markov Monte Carlo donde el espacio de estado es la colección de todos los valores imputados.

Al igual que cualquier otra cadena de Markov, para converger, el algoritmo MICE necesita satisfacer las siguientes tres propiedades:

Irreductible: la cadena debe ser capaz de llegar a todas las partes del espacio de estado.

Aperiódica: la cadena no debe oscilar entre diferentes estados.

Recurrencia: cualquier cadena de Markov se puede considerar como recurrente si la probabilidad de que la cadena de Markov que comienza desde i regrese a i sea igual a uno.

En la práctica, la convergencia del algoritmo MICE se logra después de un número relativamente bajo de iteraciones, generalmente entre 5 y 20. De acuerdo con la experiencia del creador del algoritmo, en general 5 iteraciones son suficientes, pero algunas circunstancias especiales requerirían un mayor número de iteraciones.

En el caso de la presente investigación, y debido a la realización de los resultados obtenidos en comparación con los otros métodos aplicados, 5 iteraciones se consideraron suficientes. Este número de iteraciones es mucho más bajo que en otras aplicaciones de los métodos de Markov Chain Monte Carlo, que a menudo requieren miles de operaciones. A pesar de esto, y desde el punto de vista de la experiencia en la investigación, también se debe señalar que en la aplicación más común cada iteración del algoritmo MICE llevaría varios minutos o incluso unas pocas horas. Además, el tiempo que dura cada iteración está vinculado principalmente al número de variables involucradas en el cálculo y no al número de casos. Debe tenerse en cuenta que los datos imputados pueden tener una cantidad considerable de ruido aleatorio, dependiendo de la fuerza de las relaciones entre las variables. Entonces, en aquellos casos en los que hay bajas correlaciones entre variables o son completamente independientes, la convergencia del algoritmo será más rápida. Finalmente, las altas tasas de datos faltantes (20% o más) ralentizarían el trabajo del proceso de convergencia.

El algoritmo MICE para la imputación de datos faltantes multivariados consiste en los siguientes pasos:

Especificado un modelo de imputación: $P(Y_j^{mis}|Y_j^{obs}, Y_{-j}, R)$ (8)

para la variable Y_j con $j=1,...,p$.

El algoritmo MICE obtiene la distribución posterior de *R* mediante el muestreo iterativo de la fórmula condicional representada anteriormente. Los parámetros *R* son específicos de las respectivas densidades condicionales y no son necesariamente el producto de una factorización de la verdadera distribución conjunta.

1. Para cada j, completa las imputaciones iniciales Y_j^0 por sorteos aleatorios de Y_j^{obs}
2. Repite para t = 1, ..., T (iteraciones).
3. Repite para j = 1, ..., p (variables).
4. Define $Y_j^t = (Y_1^t, \dots, Y_{j-1}^t, Y_{j+1}^{t-1}, \dots, Y_p^{t-1})$ como los datos actualmente completos excepto Y_j.
5. Dibuja $\emptyset_j^t \sim P\left(\emptyset_j^t \middle| Y_j^{obs}, Y_{-j}^t, R\right)$.
6. Saca imputaciones $Y_j^t \sim P\left(Y_j^{mis} \middle| Y_j^{obs}, Y_{-j}^t, R, \emptyset_j^t\right)$.

7. Termina repitiendo j.
8. Termina repitiendo t.

En el algoritmo al que se hace referencia:

Y representa una matriz $n \times p$ de datos de muestra parcialmente observados.

R es una matriz $n \times p$.

0-1 son los indicadores de respuesta de *Y*.

$\emptyset$ representa el espacio de los parámetros.

Téngase en cuenta que en la imputación MICE, se proporcionan conjeturas iniciales para todos los elementos faltantes para la matriz $n \times p$ de la muestra parcialmente observada. Para cada variable con elementos faltantes, los datos se dividen en dos subconjuntos, uno de los cuales contiene todos los datos faltantes. El subconjunto con todos los datos disponibles tiene una regresión en todas las demás variables. Luego, el subconjunto faltante se predice a partir de la regresión y los valores perdidos se reemplazan con los obtenidos de la regresión. Este procedimiento se repite para todas las variables con elementos faltantes. Después de esto, todos los elementos faltantes se imputan de acuerdo con el algoritmo explicado anteriormente, la regresión y las predicciones se repiten hasta que se alcanza el criterio de detención. En este caso, hasta que un cierto número de repeticiones consecutivas caigan dentro de la tolerancia especificada para cada uno de los valores imputados.

6.2.4. Modelos multivariantes de splines adaptativos regresivos (MARS).

Un problema común en muchas disciplinas es la adecuada aproximación de funciones de muchas variables, conocido únicamente el valor de dicha función para un grupo reducido de puntos del espacio de la variable independiente y, a menudo, perturbado por el ruido.

El objetivo es encontrar el modelo de dependencia entre la variable respuesta y las variables de entrada, $x_i,\ldots,x_n$, una vez que se ha realizado un muestreo

$$\{y_i, x_i, ..., x_n\}_1^N \tag{9}$$

El sistema que genera los datos se puede describir como:

$$y = f(x_1, ..., x_n) + \varepsilon \tag{10}$$

Sobre un dominio que contiene los datos, expresado como:

$$(x_1, K, x_n) \in D \subset R^n \tag{11}$$

La función *f* relaciona la variable de salida con las variables de entrada y ε es el ruido estocástico. El objetivo del análisis de regresión es encontrar una función:

$$\tilde{f}(x_1, L, x_n) \tag{12}$$

que sirva como aproximación de $f(x_1,\Lambda,x_n)$ sobre el dominio D.

Para ello se considera un tipo de funciones denominadas funciones básicas B_m de la forma:

$$B_m(x) = I\left[x \in R_m\right] \tag{13}$$

Donde:

I es una función que toma el valor 1 si el argumento es cierto y 0 en caso contrario.

$\{a_m\}_1^M$ Son los coeficientes de expansión cuyos valores son ajustados para obtener una buena adaptación a los datos.

$\{R_m\}_1^M$ Son las subregiones del espacio donde está definida la función. Si estas subregiones son disjuntas, sólo una función básica es distinta de 0 para cada x.

La principal limitación del método anterior es su falta de continuidad entre subregiones vecinas. Esta falta de continuidad limita la precisión de la adaptación. Para conseguir modelos continuos, con derivadas continuas, se desarrolló el método de splines regresivos adaptativos (Multivariable Adaptative Regressive Splines, MARS) [Friedman], Steinberg y Colla, 1995; Steinberg et al. 1999).

El único aspecto que introduce discontinuidades en el modelo es la función escalón. Si se remplaza esta función por otra que sea continua, el algoritmo 1 debería de producir modelos continuos. La función elegida para remplazar a la función escalón es un spline.

Las dos partes de la división de la función básica tienen la forma:

$$b_q^{\pm}(x - t^n) = \left[\pm(x - t^n)\right]_+^{q_S} \tag{14}$$

Donde:

t^n, es la localización del nodo, q_s es el orden del spline, y el subíndice indica la parte positiva del argumento. Para $q_s > 0$ la aproximación por splines es continua (Prenter, 1975), y con q_s -1 derivadas continuas. Las funciones escalón son un caso particular en que los splines son de grado cero, $q_s = 0$.

Este método produce unas funciones básicas que son el producto de splines univariantes, y tienen la forma:

$$B_m^{(q)}(x) = \prod_{K=1}^{K_m}\left[s_{km}.\left(x_{v(k,m)} - t_{km}\right)\right] \tag{15}$$

Es decir, remplazando las funciones escalón por splines de grado q_s, se consiguen modelos continuos, con q_s -1 derivadas continuas.

El modelo MARS se escribe de la siguiente forma:

$$\tilde{f}(x) = a_0 + \sum_{K_m=1} f_i(x_i) + \sum_{K_m=2} f_{ij}(x_i, x_j) + \sum_{K_m=3} f_{ijk}(x_i, x_j, x_k) + L \tag{16}$$

El primer sumatorio contiene todas aquellas funciones que dependen de una sola variable. El segundo contiene las funciones básicas que dependen de dos variables, y representa las interacciones entre dos variables. El tercer sumatorio representa la contribución de las interacciones entre tres variables, y así sucesivamente.

Consideremos la siguiente ecuación que representa al conjunto de variables asociada con la función básica *m*, $B_m(x)$:

$$V(m) = \{v(k,m)\}_1^{Km} \tag{17}$$

Cada función del primer sumatorio puede ser expresada como:

$$f_i(x_i) = \sum_{\substack{K_m=1 \\ i \in V(m)}} a_m B_m(x_i) \tag{18}$$

La ecuación anterior representa la suma de todas las funciones básicas que envuelven solamente la variable x_i y es el spline que representa la función univariante correspondiente.

Cada función bivariante del segundo sumatorio puede expresarse como:

$$f_i(x_i, x_j) = \sum_{\substack{K_m=2 \\ (i,j) \in V(m)}} a_m B_m(x_i, x_j) \tag{19}$$

La cual representa la suma de todas las funciones básicas que envuelven un determinado par de variables x_i y x_j. Sumándole la correspondiente contribución univariante para esas mismas variables se tendrá:

$$f_{ij}^*(x_i, x_j) = f_i(x_i) + f_j(x_j) + f_{ij}(x_i, x_j) \tag{20}$$

Que representa el conjunto de la contribución bivariante de x_i y x_j al modelo. Procediendo de la misma manera se obtienes las contribuciones de los términos correspondientes a grupos de tres o más variables.

El método MARS combina los métodos de proyección activa y proyección recursiva, utilizando regresión multivariante adaptativa con splines (Chevochot, 1999; Steinwart, 2001). El modelo utilizado por el MARS es el mismo que en proyección recursiva pero con distintas funciones base. Las funciones base utilizadas por el MARS son splines multivariantes, es decir, el producto tensorial de splines de una dimensión.

Con el fin de conocer la importancia de las variables que forman parte de un modelo MARS, el algoritmo muestra los resultados correspondientes a tres criterios. Estos son:

El criterio de nsubsets: este criterio cuenta el número de subconjuntos de modelos en el que se ha incluido cierta variable. Aquellas variables que se han incluido en un mayor número de subconjuntos de modelos son consideradas de mayor importancia. En este contexto, debe tenerse en cuenta que se denomina subconjunto a cada subconjunto de términos generado por el algoritmo en cada una de sus pasadas de poda. Es decir, existe un subconjunto para cada uno de los posibles tamaños del modelo, partiendo de una variable y llegando hasta el número final de variables que es elegido por el modelo. De modo que la estimación se realiza tomando en cuenta solo aquellos modelos de tamaño menor o igual que el del modelo final.

El criterio RSS: este criterio calcula el valor de la ecuación RSS (véase al respecto [Friedman]) para cada uno de los subconjuntos. Así, las variables cuya eliminación contribuyen de manera más significativa a la reducción del RSS son consideradas como las más importantes para el modelo final.

Finalmente, el criterio GCV es análogo al del RSS pero empleando esta vez la ecuación de GCV en vez de la del RSS (véase al respecto [Friedman]).

6.2.5. Mapas Autoorganizados – SOM.

T. Kohonen en el año 1982 presentó un modelo de red denominado mapas auto-organizados o SOM (Self-Organizing Maps), basado en ciertas evidencias descubiertas sobre el funcionamiento del cerebro.

Las redes con aprendizaje no supervisado (también conocido como aprendizaje autosupervisado) no requieren influencia externa para ajustar los pesos de las conexiones entre sus neuronas. La red no recibe ninguna información del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta porque no dispone de ninguna salida objetivo hacia la cual la red neuronal deba tender; por ello, suele decirse que estas redes son capaces de organizarse.

La red debe autoorganizarse en función de los datos procedentes del exterior, es decir, debe descubrir rasgos comunes, regularidades, correlaciones o categorías en los datos de entrada, e incorporarlos a su estructura interna de conexiones.

En el aprendizaje competitivo las neuronas existentes en la red compiten unas con otras de manera que cuando se presente a la red un patrón de entrada, sólo una de las neuronas de salida (o un grupo de vecinas) se active quedando finalmente una como neurona vencedora y anuladas el resto, que son forzadas a sus valores de respuesta mínimos.

El objetivo de este aprendizaje es categorizar los datos que se introducen en la red clasificando aquellos valores similares en la misma categoría, y éstas deben ser creadas por la propia red a través de los datos de entrada, puesto que se trata de un aprendizaje no supervisado.

En el caso concreto de las redes SOM, lo que se realiza es una identificación de características, obteniéndose en las neuronas de salida una disposición geométrica que representa un mapa topográfico de las características de los datos de entrada. De este modo si se presentan a la red informaciones similares, siempre serán afectadas las neuronas de salida próximas entre sí en la misma zona del mapa.

Un modelo SOM estará compuesto al menos por dos capas de neuronas. La capa de entrada, que constará de N neuronas, una por cada variable de entrada, y se encargará de recibir y transmitir a la capa de salida la información que recibe del exterior y la capa de salida que constará de M neuronas y es la encargada de procesar la información y formar el mapa de características.

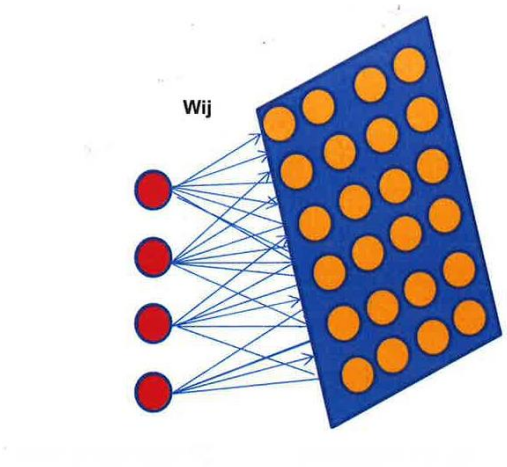


Figura 20: Esquema de arquitectura SOM

Las conexiones entre las dos capas que forman la red son siempre hacia delante, es decir la información se propaga desde la capa de entrada hacía la capa de salida. Cada neurona de entrada "*i*" está conectada con cada neurona de salida "*j*" mediante un peso *wij*, es decir, cada neurona de salida tiene asociado un vector de peso *wj*, llamado vector de referencia (codebook) debido a que constituye el vector promedio de la categoría representada por la neurona de salida j.

Entre las neuronas de la capa de salida, puede decirse que existen interacciones laterales de excitación e inhibición implícitas, pues aunque no estén conectadas, cada una de ellas va a influir sobre sus vecinas formando la topología o estructura del mapa. Este proceso se conoce como función de vecindad y durante la fase de entrenamiento, el SOM forma una red elástica que se pliega dentro de la nube de datos originales de manera que el algoritmo tiende a aproximar la densidad de datos, es decir, los vectores de referencia del codebook se acercan a las áreas donde la densidad de datos es elevada.

El funcionamiento de esta red es relativamente simple. Cuando se introduce una información $E_k=(e_1(k),...,e_N(k))$, cada una de las M neuronas de la capa de salida la recibe a través de las conexiones Feedforward con pesos w_{ji}. También estas neuronas reciben las correspondientes entradas debidas a las conexiones laterales con el resto de las neuronas de salida y cuya influencia dependerá de la distancia a la que se encuentren.

La formulación matemática del funcionamiento puede simplificarse mediante la siguiente expresión, que representa cuál de las M neuronas se activará al introducir la información E_k:

$$s_j = 1 \Leftrightarrow min\left\| E_k - w_j \right\| = min\left(\sqrt{\sum_{i=1}^{N} \left(e_i^{(k)} - w_{ij} \right)^2} \right)$$
$$s_j = 0 \; en \; otro \; caso \qquad (21)$$

Donde:

S_j es la salida generada por una neurona de salida j ante un vector de entrada Ek.

W_j es el vector de pesos de las conexiones entre cada una de las neuronas de salida j.

Lo que hace la red SOM es realizar una tarea de clasificación, ya que la neurona de salida activada ante una entrada representa la clase a la que pertenece dicha información de entrada.

6.2.6. Hibridación MARS-SOM

Un mapa autoorganizado tal y como se explicó en el apartado anterior, consta de componentes llamados nodos o neuronas, donde asociado a cada nodo existe un vector de ponderación de la misma dimensión que los vectores de datos de entrada y con una determinada posición en el espacio del mapa. La disposición habitual de los nodos es un espaciado regular en una cuadrícula hexagonal o rectangular.

El mapa de autoorganización describe un mapeo desde un espacio de entrada dimensional más alto a un espacio de mapa dimensional inferior. Para ubicar un dato de la muestra en un subconjunto determinado, los mapas autoorganizados utilizan aprendizaje competitivo. Cuando se alimenta un ejemplo de entrenamiento de la red, se calcula su distancia euclidiana a todos los vectores de ponderación existentes. La neurona con el vector de peso más similar a la entrada se llama la mejor unidad de coincidencia (BMU). Los pesos de la BMU y las neuronas cercanas a ella en la red SOM se ajustan hacia el vector de entrada. La magnitud del cambio disminuye con el tiempo y con la distancia desde el BMU. Este proceso se repite para cada vector de entrada una gran cantidad de ciclos (que se establece con anterioridad para cada estudio). La red termina asociando nodos de salida con grupos o patrones en el conjunto de datos de entrada.

Después del entrenamiento de la red, la información de entrada se agrupará en una serie de conjuntos igual a la dimensión de la red (o el número de neuronas utilizadas en el entrenamiento, que es lo mismo) y cada grupo vendrá inequívocamente identificado por el peso vector de la neurona. Este peso no es más que un registro eléctrico hipotético con valores típicos en las 10 variables que lo definen, siendo los valores de estas variables los que mejor representan el comportamiento del conjunto. El problema en este momento es identificar el número más adecuado de subconjuntos en los que se puede dividir las

condiciones óptimas de funcionamiento de la instalación eléctrica. Como se explicó anteriormente, los registros que forman la población serán sustituidos por los vectores directores de cada uno de los subconjuntos, creando una nueva población reducida de registros eléctricos óptimos. Este subconjunto reducido es representativo de la población original. Por lo tanto, el subconjunto debe ser lo suficientemente grande para representar la variabilidad de las múltiples situaciones de funcionamiento óptimo de la instalación eléctrica. Al mismo tiempo, el subconjunto debe reducirse lo suficiente para obtener una muestra equilibrada de casos representados por sus vectores directores versus casos anómalos de registro representado por cada fila de datos de la muestra original.

Estos tipos de mapas se utilizan para representar todas las observaciones disponibles (vectores de datos), con una precisión optimizada, mediante un conjunto reducido de modelos.

Sea *N* la dimensión de los "*n*" vectores directores $X(t) \in R^n$, $t = 1, 2, \ldots n$, donde cada vector de muestra se identifica con una etiqueta. La capa de salida bidimensional del mapa SOM contiene una malla rectangular de $k = 1, \ldots, x_{dim} \times y_{dim}$ nodos. Cada uno de estos nodos consta de un vector de referencia (codebook) *Wk*, de dimensión *N*.

El cálculo de los vectores de referencia se realiza utilizando el siguiente algoritmo para una cierta cantidad de iteraciones marcada con anterioridad

1. Se elegirá un vector de muestra $X(t)$ al azar.
2. Posteriormente se buscará el vector de peso más cercano:
   $W_c: \|X - W_c\| = min_j \|X - W_j\|$
3. Se ajustan los pesos W_i mediante la siguiente regla:
   $W_i(t+1) = W_i(t) + h_{ci}(t) \cdot [X(t) - W_i(t)]$

donde $h_{ci}(t)$ es la función vecina, que, en el caso de la presente investigación y que es muy común en la literatura [art. 3, 2016], es del tipo gaussiano:

$$h_{ci}(t) = \alpha(t) \cdot \exp\left(\frac{-\|W_c - W_i\|}{2 \cdot \sigma^2(t)}\right) \qquad (22)$$

El peso de las neuronas que se encuentran en el vecindario de la neurona ganadora $h_{ci}(t)$ se acerca más a $X(t)$. La tasa de aprendizaje $\alpha(t) \in [0,1]$ disminuye monótonamente a medida que aumenta el número de iteraciones, $\sigma(t)$ determina que el radio del vecindario también disminuye monótonamente. Después de muchas iteraciones y la lenta reducción de $\alpha(t)$ y $\sigma(t)$, el vecindario ocupa solo un nodo y se forma el mapa. Debe tenerse en cuenta que las neuronas, cuyos pesos están más cerca en el espacio del parámetro *W*, también están más cerca en la malla. Finalizado este proceso, los vectores directores obtenidos se desnormalizan. Así pues, tras la aplicación del algoritmo, el conjunto inicial de datos puede ser representado como un nuevo conjunto de datos de dimensión más reducida y formado por elementos representativos de la muestra original, denominados vectores directores y que podrían definirse como un conjunto de registros ficticios de tensiones, corrientes y factor de potencia que representa a la muestra original.

La cantidad de vectores directores elegidos para crear el mapa autoorganizado en el caso del algoritmo desarrollado está relacionada con el número de filas de la matriz obtenida empíricamente una vez eliminados los datos faltantes. En el caso de la muestra objeto de estudio del presente trabajo se probó con un número de vectores directores equivalente al 30%, 20%, 10% y 5% de la muestra original, siendo la cantidad finalmente elegida el 10%.

Una vez obtenida la muestra de vectores directores que, con una dimensión reducida, representa adecuadamente a la muestra original es preciso identificar de entre ellos cual es el más próximo al registro con datos faltantes que es preciso imputar. Para encontrar los vectores directores más cercanos se utilizará la distancia de Mahalanobis, que se trata de una medida de distancia bien conocida en estadística. Su utilidad radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales y se diferencia de la distancia euclídea en que tiene en cuenta la correlación entre las variables aleatorias [arti.3, 16]. Estas correlaciones permiten la identificación y el análisis de diferentes patrones y es una forma útil de determinar la similitud de un conjunto de muestra desconocido con uno conocido. Así pues, en la presente investigación, se usa para comparar cada una de las filas de la matriz de datos que tiene datos faltantes con todos los vectores directores calculados.

Se puede definir con la siguiente fórmula:

$$d_A(x_1, x_2) = \sqrt{(x_1 - x_2)^T \cdot A \cdot (x_1 - x_2)} \qquad (23)$$

donde x_1 y x_2 representan los conjuntos de variables de dos filas diferentes de la matriz de datos y $A \in R^{nxn}$ es una matriz totalmente semidefinida que representa el inverso de la matriz de covarianza de la clase {I}. Por medio de la descomposición del valor propio, A puede descomponerse en $A = W \cdot W^T$.

En el caso del presente algoritmo, se calcula la distancia de Mahalanobis de cada fila de vectores con dos o más puntos de datos faltantes para todos los vectores directores. Para hacer posible esta operación, téngase en cuenta que todas las variables con datos faltantes en la fila que provienen de la matriz de datos se eliminan en el vector director. Se selecciona el vector director con el valor de distancia Mahalanobis más bajo y las variables faltantes en esta fila de la matriz de datos se rellenan utilizando los valores presentes en la fila correspondiente del vector director. Finalmente, se reconstruye la matriz original y se imputa el valor de los datos faltantes de esas filas con solo uno o dos puntos de datos faltantes por medio del algoritmo AAA, explicado anteriormente y basado en una técnica multivariada no paramétrica denominada Splines de regresión adaptativa multivariante (MARS).

De forma esquemática puede representarse como se muestra en la Figura 21.

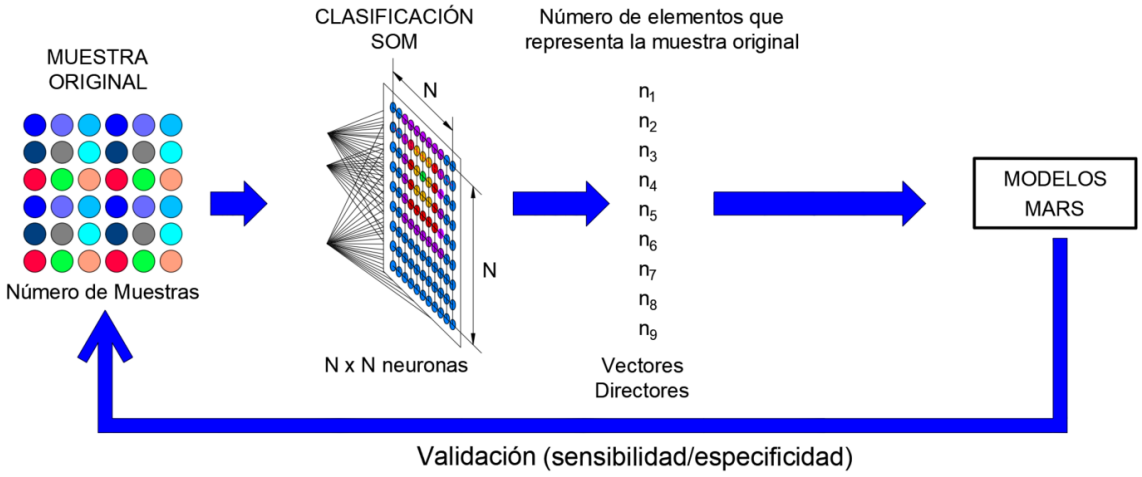


Figura 21: Representación gráfica de la hibridación MARS – SOM

7. RESULTADOS

La validación de los distintos métodos empleados en este trabajo para llevar a cabo la imputación, se ha realizado mediante técnicas de estadística comunes: Evaluación del error cuadrático medio (RMSE), y del error absoluto medio (MAE) expresados en las mismas unidades y como porcentaje de los valores medios medidos, es decir:

$$\text{RMSE (Unidad de medida)} = \sum_{i=1}^{n} \sqrt{\frac{1}{n}\left(\widehat{G_i}-G_i\right)^2} \quad (24)$$

$$\text{MAE (Unidad de medida)} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{G_i}-G_i\right| \quad (25)$$

$$\text{RMSE (\%)} = \frac{\text{RMSE}}{\frac{1}{n}\sum_{i=1}^{n} G_i} \text{x}100 \quad (26)$$

$$\text{MAE (\%)} = \frac{\text{MAE}}{\frac{1}{n}\sum_{i=1}^{n} G_i} \text{x}100 \quad (27)$$

Donde G_i y $\widehat{G_i}$ son respectivamente, las medidas y los valores estimados por modelo del parámetro a medir, y "n" es el número de puntos de datos de diez minutos del conjunto de validación.

RMSE es una regla de puntuación cuadrática que mide la magnitud promedio del error. En este caso, como los errores se cuadran antes de promediarlos, da un peso relativamente mayor a los errores grandes, es decir, pondera los errores de estimación grandes con más fuerza que los errores pequeños y se considera una métrica de validación del modelo muy importante.

MAE mide la magnitud promedio del error en un conjunto de pronósticos sin considerar su dirección, es una puntuación lineal que pondera todas las diferencias individuales por igual, siendo un buen complemento para el diagrama de dispersión de la medida modelada próxima a la línea 1-a-1.

Cuando los resultados se analizan utilizando ambas variables, se debe tener en cuenta que cuanto mayor sea la diferencia entre ellos, mayor será la varianza en los errores individuales en la muestra, y que cuanto menores sean sus valores, mejor será el modelo.

En los apartados siguientes, se analizan los resultados de los estudios llevados a cabo para los distintos tipos de algoritmos estudiados.

7.1. Validación de las técnicas de imputación como herramienta de mejora en aplicaciones eléctricas [Crespo 1]

En principio, con la presente investigación se pretende encontrar un algoritmo de imputación adecuado para el tipo de datos con que se va a trabajar y que esté incluido dentro del estado del arte actual como una técnica de alto rendimiento. El objeto final de este algoritmo será emplearlo como técnica de referencia.

En el inicio de la investigación, y ante la carencia en las instalaciones eléctricas de la Universidad de Oviedo de una serie de datos lo suficientemente extensa como para poder trabajar con ella, se recurrió a los valores de la radiación solar de las estaciones meteorológicas de la red de observación en tiempo real MeteoGalicia, capturados y almacenados cada diez minutos. Esta red es un servicio meteorológico regional con más de 100 ubicaciones en Galicia (España) que proporciona datos completos de radiación solar. La red integra estaciones con fines meteorológicos y agroclimáticos y el gobierno regional ofrece abiertamente observaciones de sus estaciones en Internet.

La diversidad de los instrumentos de medida de radiación solar (piranómetros) con diferentes niveles de precisión, junto con la necesidad de una calibración frecuente de los instrumentos, dificulta la obtención de una base de datos homogénea. Podemos encontrarnos con que no haya ninguna señal del sensor, que las señales sean inestables o que estén fuera de los límites físicos, que no haya registros almacenados, etc. Los motivos de las fallas son variados: un cable dañado o con corrosión, la pérdida de una conexión a tierra eléctrica adecuada; alteraciones en los programas de los sistemas de registro de datos, humedad dentro de un elemento del piranómetro, radiación reflejada desde nubes o cúmulos elevados etc.

Se pueden usar diferentes criterios para obtener un valor faltante o un conjunto de valores perdidos en una serie de datos de irradiación solar en las estaciones, pudiendo ser estos deterministas, aleatorios o mixtos. Para una primera aproximación en la imputación pueden emplearse métodos de estimación de la radiación basados en variables físicas (presencia de aerosoles, presión, nivel de ozono, etc.) que tienen la ventaja de su independencia espacial pero el inconveniente de que dependerán siempre de las condiciones meteorológicas. Una segunda aproximación pueden ser las relaciones de correlación, si dos estaciones están próximas, la radiación solar en ambas debe ser similar, aunque este argumento es válido para cielos nubosos o claros, pero no para aquellos parcialmente cubiertos en escalas de tiempo cortas.

Las propiedades autorregresivas de la señal también nos permiten completar las series, y de hecho han sido utilizadas técnicas ARIMA para imputar datos de radiación solar en series de tiempo. Lo que si parece evidente es que el rendimiento de todos los métodos está influenciado por la escala de tiempo.

Así, en la presente fase de la investigación se validan tres métodos: IDW (interpolación), MLR (Regresión Lineal múltiple) y MICE (Imputación Múltiple por Ecuaciones Encadenadas), utilizando la misma serie de datos y comparándolos posteriormente entre sí en términos de error cuadrático medio RMSE y error absoluto medio MAE.

En relación a los datos utilizados en la investigación, se tomó el conjunto de datos y se dividió en dos conjuntos; uno de entrenamiento con dos tercios de las muestras (14,553 muestras); y otro de prueba con la muestra restante (7,276 muestras). Así, el conjunto de entrenamiento es utilizado para aplicar los tres métodos referidos (IWD, MLR, MICE) y el conjunto de prueba se usa para validar los modelos con datos independientes, ya que este conjunto de prueba no proporcionó información alguna en la construcción de los modelos.

Supongamos que tenemos un conjunto de datos que está formado por *c* variables diferentes *v1*, *v2*, .., *vc* que son las columnas de una matriz de datos cuyo número total de filas es *r*. Estas filas se corresponden con las nueve estaciones fotovoltaicas donde se han recogido las mediciones de radiación solar.

Los algoritmos objeto de estudio se han aplicado a través de los siguientes pasos:

1. Creación de una nueva matriz con valores perdidos del conjunto de datos original.

Este paso del algoritmo no es necesario cuando se aplica a un conjunto de datos en el que se van a imputar datos faltantes, pero es obligatorio en la presente investigación validar el algoritmo mediante el uso de un conjunto completo de datos.

Sea A la matriz original (r x c) de *r* filas y *c* columnas. Como primer paso y para obtener una matriz con una cierta cantidad de datos faltantes, se elimina una proporción de *p* elementos en la matriz, esta proporción varía entre el 5 y el 20% de los datos.

Sea B la nueva matriz (r x c), con una proporción *p* de elementos faltantes. La eliminación se realiza completamente al azar; por lo tanto, el tipo de imputación que se probará para determinar el rendimiento del algoritmo es el que se conoce como ausente completamente al azar (MCAR).

2. Creación de la Matriz Reducida

Se crea una nueva matriz en la que se eliminan todas las filas con datos faltantes, esta nueva matriz se llama B^{red}. Aunque el número de filas s ($s \leq r$) de esta matriz cambiará dependiendo de la matriz que se vaya a imputar, en aquellos casos como los que aquí se estudian, donde la eliminación de datos se ha realizado completamente al azar y en una proporción *p*, el número de filas restante se representará por la siguiente fórmula:

$$u = r \cdot (1 - p)^c \tag{28}$$

Dónde:

p: Proporción de datos faltantes considerados.
r: Número de filas de la matriz original.
c: número de columnas de la matriz de datos.

Luego, la matriz B^{red} se normaliza.

Una vez que ha sido construida la matriz objeto de estudio, se calculan los valores faltantes aplicando los tres algoritmos objeto de estudio en el presente trabajo, IDW, MLR y MICE. Este procedimiento se lleva a cabo varias veces, de manera general, se realizan cinco iteraciones.

Finalizado este proceso, se calcula el RMSE y el MAE en porcentaje obtenido al compararlos con los datos originales y entre ellos. En los gráficos siguientes se representan los resultados obtenidos:

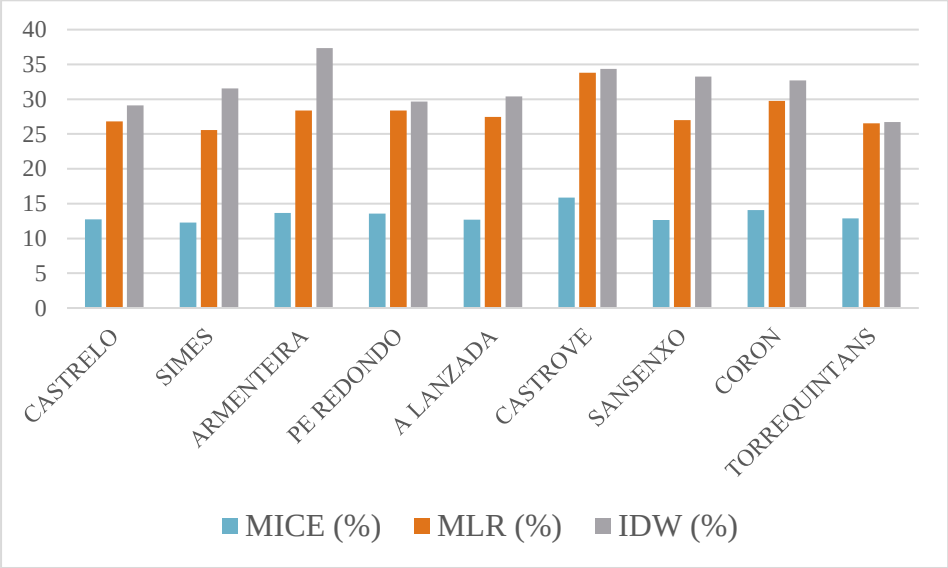


Figura 22: Comparativa para las nueve estaciones del error en término RMSE obtenido con los tres métodos

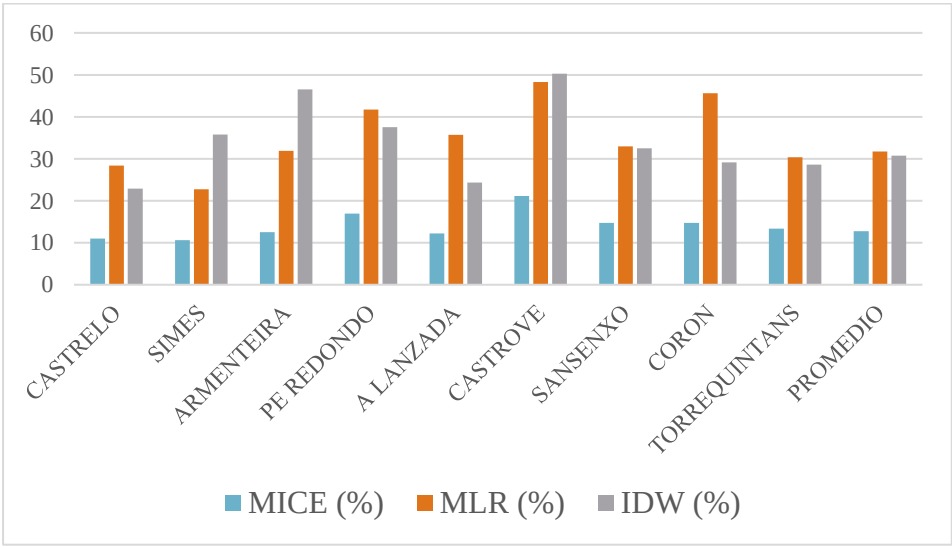


Figura 23: Comparativa para las nueve estaciones del error en término MAE obtenido con los tres métodos

Como se puede observar en las Figuras 22 y 23, los resultados, tanto en términos de RMSE como de MAE, para el algoritmo MICE, son mucho más bajos que los de los otros dos métodos para todas las estaciones objeto de estudio.

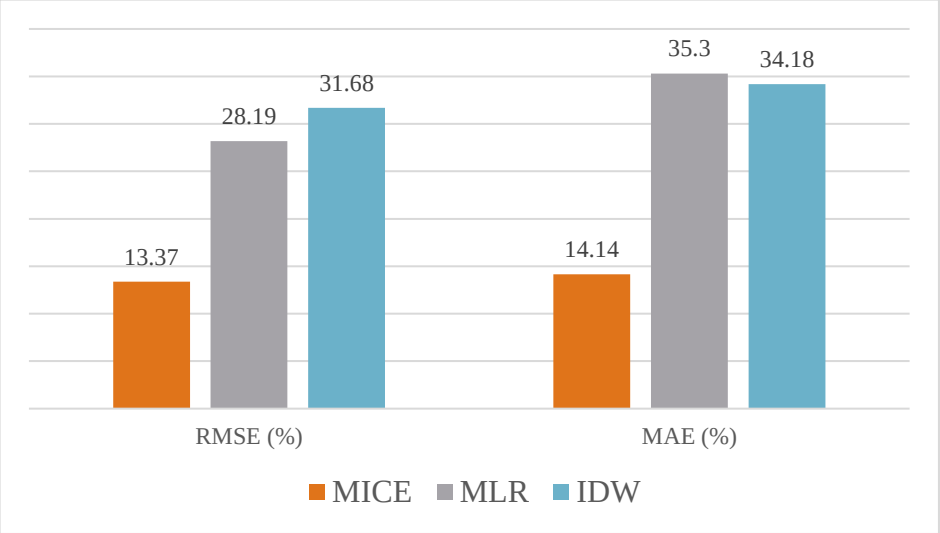


Figura 24: Valor Promedio de los errores en términos RMSE y MAE para los tres métodos

En la Figura 24, se representa el valor promedio en términos de RMSE y MAE para las nueve estaciones meteorológicas. En el caso del RMSE fue de 13.37% para el algoritmo MICE mientras que para MLR fue 28.19 % y 31.68% para el IDW. En el caso del MAE, fue del 14.14% para el MICE, mientras que los valores promedio para las mismas estaciones obtenidos con los métodos MLR e IDW fueron de 35.30% y 34.19% respectivamente.

En este punto de la investigación, podemos afirmar que MICE es una técnica adecuada y de alto rendimiento para la imputación de datos faltantes.

7.2. Desarrollo de una nueva técnica de imputación de aplicación en registradores de datos eléctricos [Crespo 2].

Una vez demostrada en el apartado anterior la viabilidad del algoritmo MICE como elemento de comparación para la imputación de datos faltantes, en este punto de la investigación se dispone de una muestra lo suficientemente extensa de datos de la instalación eléctrica del Edificio Severo Ochoa y se procede a desarrollar el primer método de imputación inteligente basado en MARS. Estos datos son medidas de las variables eléctricas; voltajes fase-neutro, voltajes de fase, intensidades y factor de potencia, que han proporcionado una serie de datos formada con diez columnas, una por cada variable con la que se ha trabajado y una fila para cada intervalo de 15 minutos desde el 27 de Noviembre de 2014 a las 18:45 al 31 de Mayo de 2015 a las 23:45.

Para probar los algoritmos objeto de estudio en este trabajo, se supone que tenemos un conjunto de datos formado por *n* variables diferentes *v1*, *v2*, ..., *vn*. Para calcular los

valores faltantes de la columna i-ésima, se emplean todas las filas sin valor faltante en dicha columna. Luego, se llevan a cabo una cierta cantidad de modelos mediante la eliminación al azar de un determinado porcentaje de datos según el método estudiado en el apartado anterior. Es posible encontrar filas con cantidades muy diferentes de datos faltantes desde 0 (sin datos faltantes) hasta *n* (faltan todos los valores). Aquellas columnas con todos los valores faltantes se eliminarán y no se usarán para el cálculo del modelo ni se imputarán. Por lo tanto, cualquier cantidad de datos faltantes de 0 a n-2 es factible (todas las variables menos una con valores perdidos).

En otras palabras, si el conjunto de datos está formado por las variables *v1*, *v2*, ..., *vn*. y se quieren estimar los valores perdidos en la columna *vi*, entonces el número máximo de diferentes modelos con el algoritmo objeto de estudio que se calcularían para esta variable (y en general para cada columna) es el siguiente: $\sum_{k=1}^{n-1}\binom{n-1}{k}$. Para el caso de los datos bajo estudio en esta investigación, con 10 variables diferentes, se entrenaría un máximo de 5,110 modelos distintos (511 para cada variable).

En la Figura 25, se muestra la representación gráfica de un conjunto de datos formado por las primeras 25 filas donde la *x* representa a la existencia de medida y *o* a la ausencia del dato:

Fila	*v1*	*v2*	*v3*	*v4*	*v5*	*v6*	*v7*	*v8*	*v9*	*v10*	M.1	M. 2	M. 3	M. 4	M. 5	M. 6	M.7	M. 8
1	X	X	X	X	X	X	X	X	X	X	Si	si	si	si	si	si	si	si
2	X	o	X	o	X	X	X	X	X	X	no	no	si	no	no	si	si	no
3	X	X	X	X	o	X	X	X	X	X	no	no	no	si	no	no	no	no
4	X	X	o	o	X	X	X	X	X	X	no	no	no	no	no	no	no	no
5	X	X	X	X	X	X	X	X	X	X	si	si	si	si	si	si	si	si
6	X	o	X	X	X	X	X	X	X	X	no	no	si	no	no	si	si	si
7	o	X	X	X	X	X	X	X	X	X	no	si	no	no	no	si	no	no
8	X	X	X	X	X	X	X	X	X	X	si	si	si	si	si	si	si	si
9	o	o	o	X	X	X	X	X	X	X	no	no	no	no	no	no	no	no
10	X	X	o	X	X	X	X	X	X	X	no	no	no	no	no	no	no	no
11	X	X	o	X	X	X	X	X	X	X	no	no	no	no	no	no	no	no
12	X	o	o	X	X	X	X	X	X	X	no	no	no	no	no	no	no	no
13	X	X	X	X	X	X	X	X	X	X	si	si	si	si	si	si	si	si
14	o	o	X	X	X	X	X	X	X	X	no	no	no	no	no	si	no	no
15	o	X	X	X	X	X	X	X	X	X	no	si	no	no	no	si	no	no

16	X	X	X	X	X	X	X	X	X	X	si	si	si	si	si	si	si	si
17	o	o	o	o	o	o	X	X	X	X	no	no	no	no	no	no	no	no
18	X	X	X	X	X	X	X	X	X	X	si	si	si	si	si	si	si	si
19	X	o	X	X	X	X	o	X	X	X	no	no	si	no	no	no	no	si
20	X	X	X	X	X	o	X	X	X	X	no	no	no	no	si	no	no	no
21	X	o	o	X	o	X	X	X	X	X	no	no	no	no	no	no	no	no
22	X	X	X	X	X	X	X	X	X	X	si	si	si	si	si	si	si	si
23	X	X	o	o	X	X	X	X	X	X	no	no	no	no	no	no	no	no
24	X	X	X	X	X	o	X	X	X	X	no	no	no	no	si	no	no	no
25	X	X	o	X	X	X	X	X	o	X	no	no	no	no	no	no	no	no

Figura 25: Representación gráfica de las 25 primeras filas de un conjunto de datos

Cuando el algoritmo se aplica por ejemplo a la tercera columna de este conjunto de datos (variable *v3*), todas las filas con datos faltantes (representados mediante el símbolo 'o') en la tercera columna no se emplean para el cálculo de los modelos (filas en rojo). Si se eliminaran esas filas, se entrenarían diferentes modelos para la predicción de *v3* utilizando diferentes subconjuntos de variables. Continuando con el ejemplo de la variable *v3* y teniendo en cuenta los datos que faltan en las primeras 25 filas, sería posible entrenar los siguientes modelos:

- Modelo 1: un modelo que utiliza como variable de salida *v3* y las otras 9 como variables de entrada (*v1, v2, v4, v5, v6, v7, v8, v9, v10*).
- Modelo 2: un modelo que utiliza como variable de salida *v3* y como variables de entrada *v2, v4, v5, v6, v7, v8, v9, v10*.
- Modelo 3: un modelo que utiliza como variable de salida *v3* y como variables de entrada *v1, v4, v5, v6, v7, v8, v9, v10*.
- Modelo 4: un modelo que utiliza como variable de salida *v3* y como variables de entrada *v1, v2, v4, v6, v7, v8, v9, v10*.
- Modelo 5: un modelo que utiliza como variable de salida *v3* y como variables de entrada *v1, v2, v4, v5, v7, v8, v9, v10*.
- Modelo 6: un modelo que utiliza como variable de salida *v3* y como variables de entrada *v4, v5, v6, v7, v8, v9, v10*.
- Modelo 7: un modelo que utiliza como variable de salida *v3* y como variables de entrada *v1, v5, v6, v7, v8, v9, v10*.
- Modelo 8: un modelo que utiliza como variable de salida *v3* y como variables de entrada *v1, v4, v5, v6, v8, v9, v10*.

Después del cálculo de todos los modelos disponibles, los datos faltantes de cada fila se calcularán utilizando aquellos modelos que emplean todas las variables disponibles no

perdidas de la fila. En aquellos casos en los que no se calculó ningún modelo, los datos faltantes se reemplazarán por la mediana de la columna. Debe tenerse en cuenta que los casos de conjuntos de datos grandes con un porcentaje no demasiado alto de datos faltantes, serán poco frecuentes. Como regla general para el uso de los algoritmos, se ha establecido que cuando se puede calcular un dato faltante utilizando más de un modelo, se debe de calcular utilizando aquellos modelos con el mayor número de variables de entrada; el valor sería estimado por cualquiera de ellos elegidos al azar. Finalmente, en aquellos casos excepcionales en los que no hay un modelo disponible para la estimación, se usará para la imputación el valor correspondiente a la mediana de la variable.

La metodología empleada es similar a la anterior: una vez eliminadas de las series de datos aquellas filas donde aparecían datos faltantes, para obtener un conjunto homogéneo, (matriz B^{red}) se procede a la eliminación aleatoria, mediante MCAR de un porcentaje determinado de datos (10, 15 y 20% del total). Sobre esta nueva serie, aplicaremos los algoritmos objeto de estudio que mediante técnicas de imputación nos facilitaran nuevos valores para datos que previamente hemos desestimado. Para garantizar la viabilidad del método utilizado, este procedimiento de inserción aleatoria se ha llevado a cabo cinco veces para cada uno de los porcentajes de datos faltantes estudiado. A cada una de esas veces que se realiza la prueba se la denomina iteración.

Posteriormente se compara los resultados obtenidos en cada una de las imputaciones con los valores reales y se calcula el error cometido en la aplicación de algoritmo MICE, y en la del nuevo, al que denominamos AAA (Algoritmo de Asignación Adaptativa), en cada una de las cinco iteraciones llevadas a cabo.

El rendimiento del algoritmo propuesto en comparación con MICE también se ha evaluado utilizando MAE y RMSE.

Los resultados de los errores obtenidos en términos de RMSE para cada iteración con un 15% de datos faltantes se muestran en las Tablas 6 y 7 del Anexo 1, siendo su representación gráfica la que se muestra en la Figura 26.
A la vista de los resultados mostrados en las gráficas anteriores, se puede afirmar que los resultados obtenidos para el algoritmo AAA son mejores que los obtenidos con MICE para un 15 % de datos faltantes en términos de RMSE.
Los resultados en términos de MAE se detallan en las Tablas 8 y 9 del Anexo 1 y su representación gráfica para el promedio de las cinco iteraciones se presenta en la Figura 27.

A la vista de lo mostrado, queda de manifiesto que los resultados obtenidos para el algoritmo AAA son mejores que los obtenidos con MICE para un 15 % de datos faltantes también en términos de MAE.

Para cada una de las diez variables involucradas se realizaron pruebas ANOVA bidireccionales. Las pruebas ANOVA son pruebas paramétrica también denominadas de análisis de la varianza, que requiere una serie de supuestos para poder ser aplicada correctamente. Esto va a servir, no solo para estudiar las dispersiones o varianzas de los grupos, sino para estudiar sus medias y la posibilidad de crear subconjuntos de grupos con medias iguales.

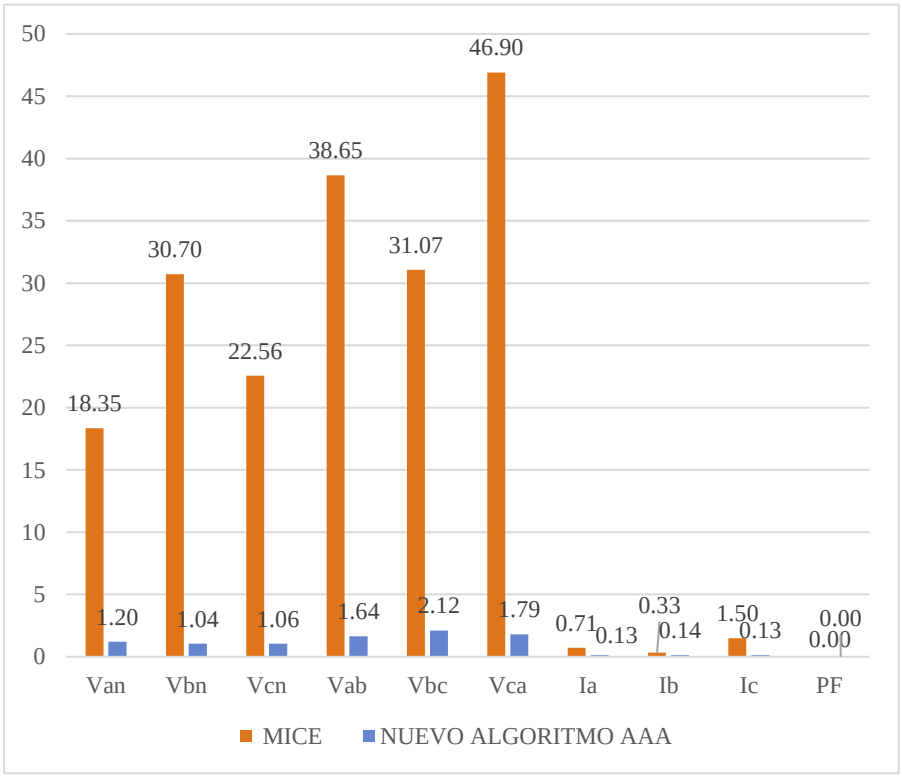


Figura 26: RMSE obtenido con 15% de datos faltantes aplicando MICE y el algoritmo AAA

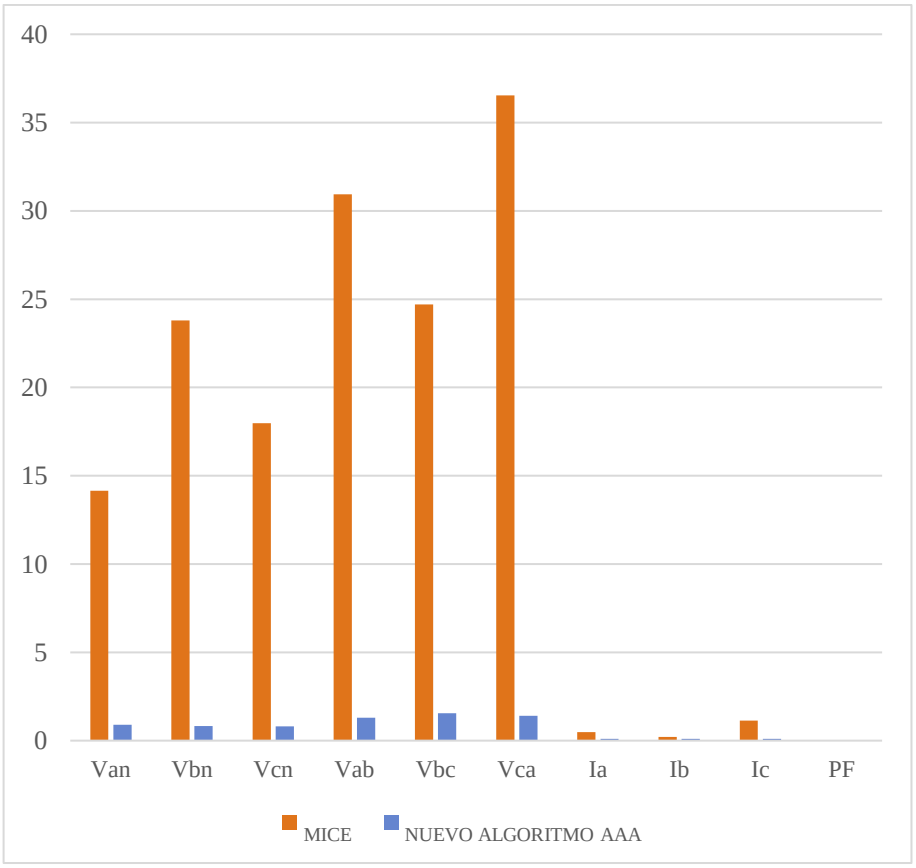


Figura 27: MAE obtenido con 15% de datos faltantes aplicando MICE y el algoritmo AAA

Se requiere que cada uno de los grupos a comparar tenga distribuciones normales, o lo que es más exacto, que lo sean sus residuales. Los residuales son las diferencias entre cada valor y la media de su grupo. Esto permite estudiar la dispersión o varianzas de los grupos, es decir su homogeneidad.

La realización de las pruebas ANOVA han permitido examinar los siguientes factores:

1. la influencia del tipo de algoritmo empleado para la imputación (MICE vs algoritmo AAA).
2. El nivel de datos faltantes (10%, 15% y 20%).
3. La interacción de los factores anteriores.

Estos estudios se llevaron a cabo para las métricas RMSE y MAE concluyéndose que existía influencia del modelo empleado en todas las variables, para ambas métricas. Ni el porcentaje de datos faltantes ni su interacción con el modelo empleado para la imputación fueron significativos en ninguna de las variables.

En cuanto a los valores de p, para el parámetro RMSE, fue prácticamente despreciable para el tipo de modelo (MICE vs AAA) en todas las variables. Al considerar el porcentaje variable de los datos faltantes, para el RMSE no hubo diferencias estadísticamente significativas entre los porcentajes (10, 15 y 20%). Los resultados numéricos obtenidos se muestran en la tabla 10 del anexo 1 donde el primer caso se representa por p y el segundo por p'.

En el caso del MAE métrico, el valor es menor a 0,001 para todas las variables para el tipo de modelo (MICE vs AAA). Para el caso del porcentaje variable de datos faltantes, en este caso, tampoco hay diferencias estadísticamente significativas entre lo porcentajes ensayados (10, 15 y 20%). Los resultados numéricos obtenidos pueden verse en la Tabla 11 del Anexo1, donde el primer caso se representa por *p* y el segundo por *p'*.

A la finalización de este trabajo, queda demostrado que, cuando aumentan considerablemente el número de datos faltantes en una misma fila, (es decir en un registro), el algoritmo pierde eficiencia rápidamente, poniéndose de manifiesto la necesidad de investigar en el desarrollo de una técnica alternativa o complementaria a esta.

7.3. Desarrollo de un nuevo algoritmo hibrido de imputación de aplicación en registradores de datos eléctricos [Crespo 4]

En este punto de la investigación, ha sido posible comprobar que los algoritmos anteriores pueden sustituir por valores estimados los datos de las variables eléctricas que faltan en una serie obtenida por un registrador de una instalación eléctrica cometiendo errores poco significativos siempre y cuando el número de datos faltantes por fila no supere a tres. Se ha pretendido avanzar en esta línea, presentando un nuevo algoritmo más robusto

ante variaciones altas de faltantes por fila. Para ello, se ha recurrido a un método de imputación de datos faltantes que combina mapas autoorganizados de redes neuronales (SOM) y distancias de Mahalanobis, para posteriormente, hibridarlo con el algoritmo estudiado en el apartado anterior. El nuevo algoritmo, basado en MARS, recibirá el nombre de Algoritmo de Asignación Adaptativa Hibrido, HAAA [art. 3, 16].

Los resultados obtenidos con el nuevo algoritmo, se han comparado con las técnicas ya aplicadas con anterioridad: MICE y AAA y se ha comprobado que el nuevo método propuesto supera a ambos algoritmos.

Los datos utilizados para la investigación son los mismos que los utilizados en el apartado anterior: los obtenidos durante aproximadamente 5 meses para las variables eléctricas de la instalación del Edificio Severo Ochoa, que en total suponen 17763 muestras.

Al igual que en las ocasiones anteriores, la prueba se realizó utilizando la metodología MCAR, eliminando el 10%, 15% y 20% de la información. Este proceso fue repetido cinco veces y el rendimiento de los tres algoritmos se comparó basándose en las métricas MAE y RMSE.

Para facilitar el tratamiento de la información, se presentan en este trabajo los resultados de todas las interacciones realizadas, así como los valores promedio de las cinco repeticiones para un porcentaje de datos faltantes del 20%. En las Tablas 12, 13 y 14 del Anexo 1, se recogen los valores numéricos obtenidos para la métrica RMSE con MICE, AAA y HAAA respectivamente. En las Tablas 15, 16 y 17, para la MAE con los mismos algoritmos. En las Figuras 28 y 29, se representan dichos resultados.

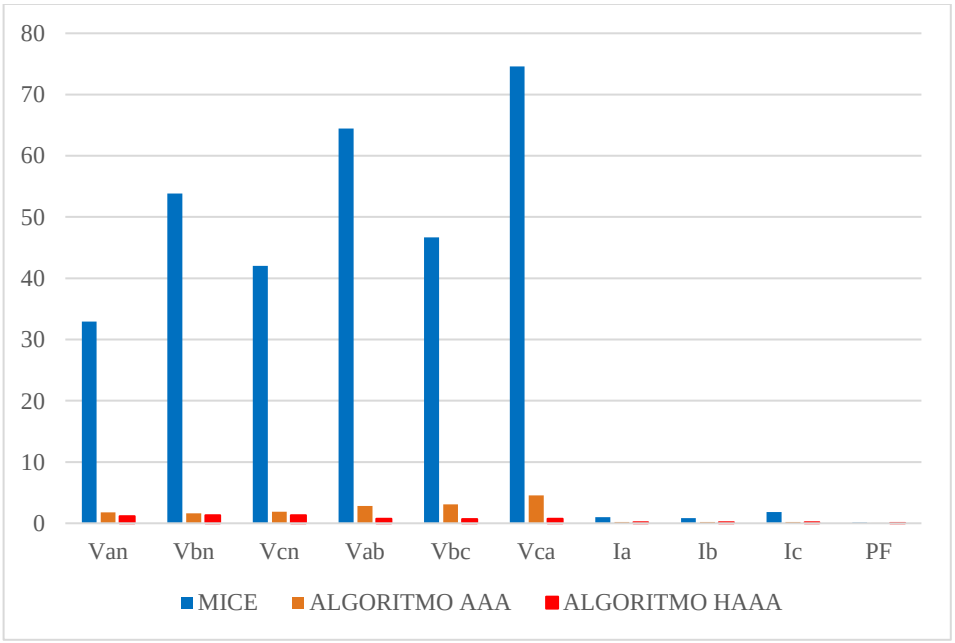


Figura 28 : Comparativa del MICE, AAA, Y HAAA en términos de RSME par un 20% de datos faltantes

A tenor de la evaluación utilizando datos de MCAR, el rendimiento general del nuevo algoritmo como puede apreciarse en las gráficas anteriores mejora a los anteriores. No

obstante, hay que destacar la existencia de un par de situaciones en las que la información no falta completamente al azar y son de gran interés para las mediciones eléctricas.

Estos son los siguientes:

1. *El caso en el que existe una correlación en la falta de datos.*

Una situación posible cuando se trabaja con datos eléctricos sería cuando toda la información faltante corresponde a la misma fase. Para simular este tipo de falla, se crearon cinco nuevos conjuntos de datos con un 20% de datos faltantes. Cada fase se representa por medio de cuatro variables diferentes: una variable de corriente de fase, dos variables de tensión de fase a fase y una variable de tensión de fase a neutro. Significa que cada fila con información incompleta faltante tiene cuatro variables faltantes o, en otras palabras, que solo un 5% del total de filas tendrá datos faltantes. En las filas referidas, seleccionadas al azar, se eliminó la información de las variables de una de las fases. Significa que, por ejemplo, cuando falta información para la variable Van, también falta la información de las variables, Vab, Vca e Ia. Los resultados obtenidos se presentan en las Tablas 18, 19 y 20 del anexo 1 para términos de RMSE y las Tablas 21,22 y 23 para términos de MAE. Como se puede observar, el rendimiento del algoritmo HAAA es peor que en el caso de MCAR, pero supera a MICE y AAA.

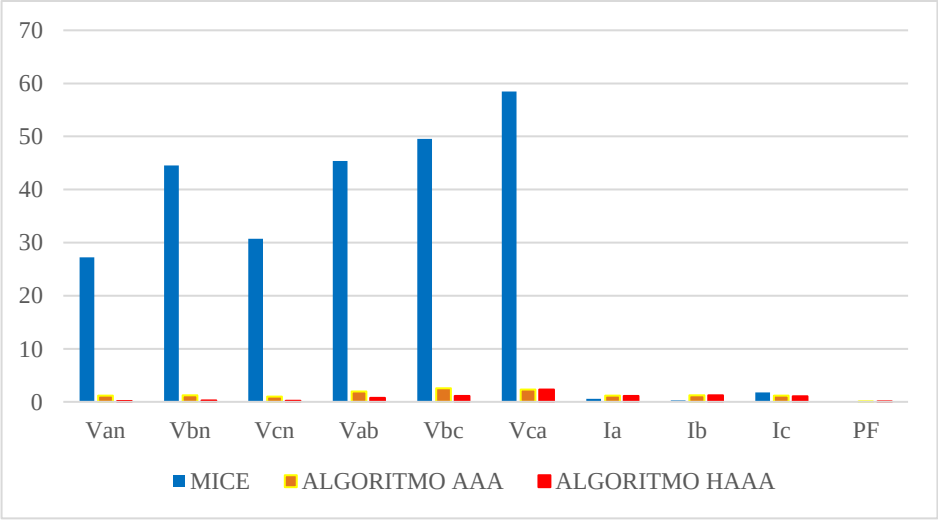


Figura 29: Comparativa del MICE, AAA, Y HAAA en términos de MAE par un 20% de datos faltantes

Los resultados obtenidos se representan en las figuras 30 y 31:

Queda de manifiesto que para las variables de voltaje, intensidad y factor de potencia empleadas en esta investigación, los valores de RMSE obtenidos por el nuevo algoritmo son considerablemente más bajos que los obtenidos usando los métodos AAA y MICE. Para el RMSE, la variable que se reduce a una cantidad menor recibe una reducción del 18%, mientras que la reducción promedio de todas las variables es del 48%. Para el MAE los datos

obtenidos son equivalentes a los anteriores tal y como queda de manifiesto en la figura anterior.

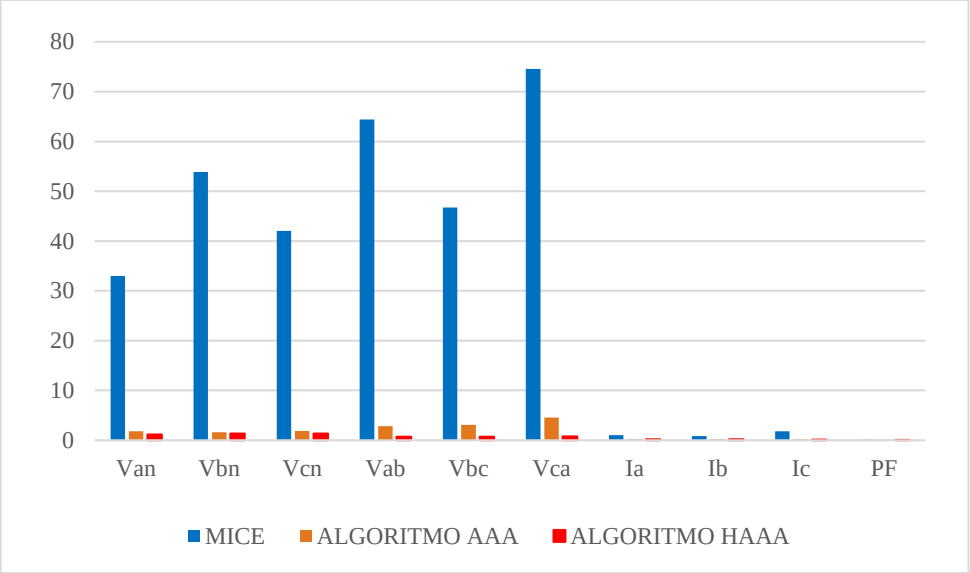


Figura 30: RMSE obtenido con un 20% de datos faltantes usando MICE, AAA, y HAAA cuando existe correlación en los datos faltantes.

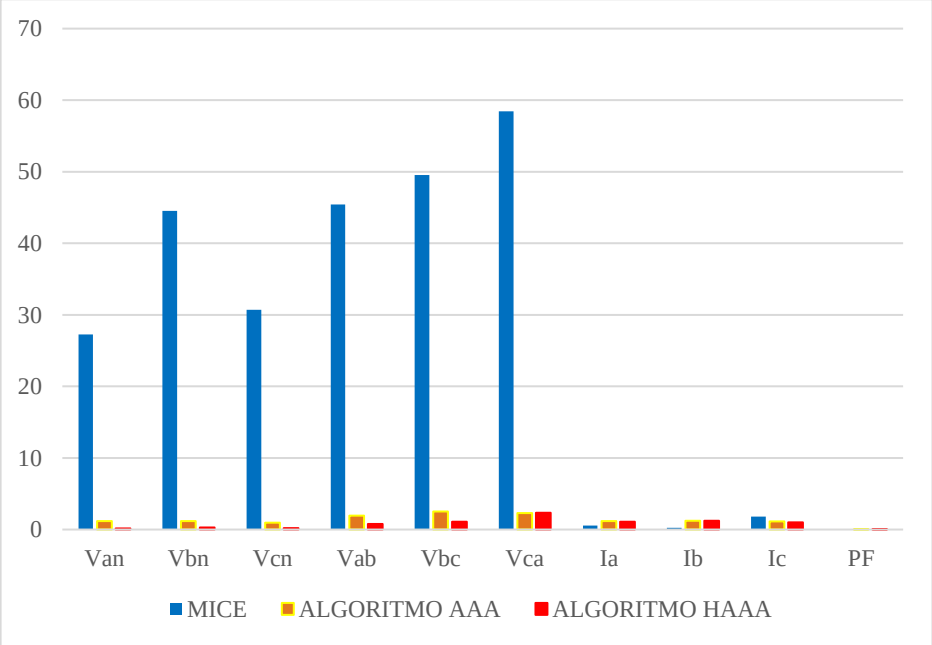


Figura 31: MAE obtenido con un 20% de datos faltantes usando MICE, AAA, y HAAA, cuando existe correlación en los datos faltantes

2. La mayoría de los datos faltantes corresponden a un cierto subconjunto de variables.

El segundo caso que hay que destacar es aquel en el que la mayoría de los datos faltantes corresponden a un cierto subconjunto de variables. Para simular este tipo de falla, se crearon cinco nuevos conjuntos de datos con un 90% de datos faltantes en una sola

variable. En cada conjunto de datos se eliminó una proporción del 90% de elementos en una sola columna, dejando el resto de las variables con sus valores originales.

Como se puede observar en la Figura 32, la precisión de imputación para todos los algoritmos disminuyó significativamente respecto a las pruebas con MCAR. Los resultados numéricos de las pruebas quedan recogidos en la tabla 24 del anexo 1.

Esto se esperaba en una situación tan desfavorable, sin embargo, es posible determinar que ambos algoritmos, HAAA y AAA, superan considerablemente al algoritmo de referencia MICE, siendo HAAA, el que obtuvo los mejores resultados.

A la vista de los resultados anteriores, puede afirmarse que el algoritmo hibrido es mejor que el AAA cuando el número de faltantes por fila es mayor de tres y cuando existe correlación entre las variables ausentes. Por el contrario, cuando el número de faltantes es menor de tres, los resultados obtenidos con el algoritmo AAA son mejores.

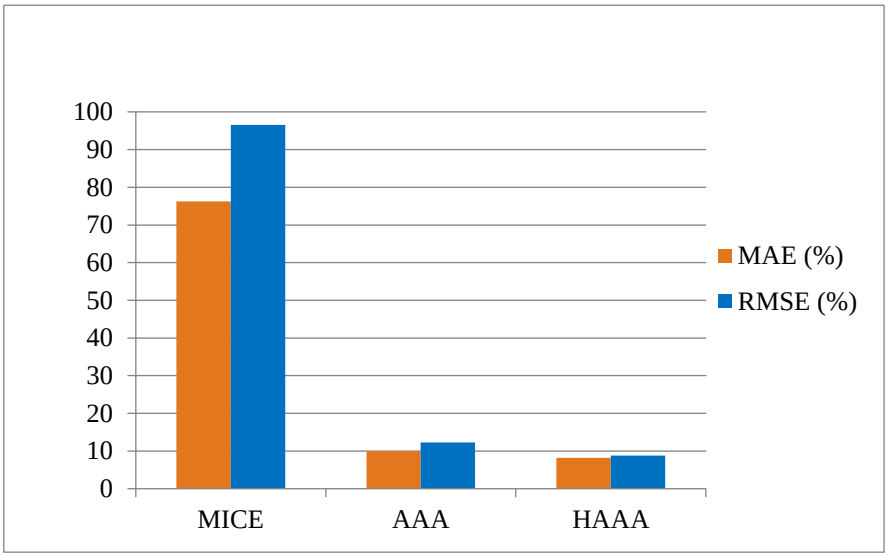


Figura 32: RMSE y MAE obtenido para un 90% de datos perdidos en una sola columna

8. CONCLUSIONES Y LÍNEAS FUTURAS

La existencia de armónicos y otras perturbaciones en las instalaciones eléctricas de baja tensión es un tema inevitable en la actualidad. Para garantizar una buena calidad del suministro eléctrico, es indispensable el uso de dispositivos de toma de datos en tiempo real que nos ofrezcan una visión clara y precisa de los parámetros de uso y permitan la toma de decisiones al servicio de mantenimiento de la instalación permitiendo de esta manera la prevención de fallos en el sistema. Durante el proceso del registro de datos, es posible que algunos falten y, en este contexto, es esencial el uso de técnicas de imputación que permitan cuantificar los datos de desaparecidos.

En relación a las distintas técnicas de imputación objeto de estudio se concluye:

- El algoritmo MICE es lo suficientemente robusto como para utilizarlo de referencia en las técnicas de imputación de datos aplicadas a la gestión de la energía en edificios públicos
- El algoritmo de asignación adaptativa AAA, es mejor que el MICE en la imputación de datos siempre y cuando el número de faltantes por fila no exceda de tres. Así pues, aunque podemos afirmar que el algoritmo propuesto en esta investigación (AAA) mejora en gran medida los resultados obtenidos por medio de una de las técnicas más reconocidas y comunes que se utilizan en la actualidad (MICE), debe tenerse en cuenta que el algoritmo propuesto, como muchos otros, tendría problemas de imputación en aquellos casos en que la mayoría de los datos faltantes pertenezcan a la misma columna o a un subconjunto reducido de columnas. Por este motivo, se hace necesario explorar el uso de máquinas de vectores de soporte (SVM) y métodos híbridos para encontrar un nuevo algoritmo con un rendimiento aún mayor, además de estudiar los sistemas no lineales que varían en el tiempo y otras características de calidad de energía, teniendo en cuenta propuestas como las aquí detalladas.
  El uso de técnicas de imputación de datos faltantes permite la creación de modelos de predicción utilizando conjuntos de datos no completos.

- El método de imputación clasificativa basado en el algoritmo SOM ofrece mejor rendimiento que el algoritmo AAA cuando el número de faltantes por fila excede de tres y supera al método MICE en todos los casos.
- La hibridación de AAA con el método basado en SOM mejora los resultados obtenidos con cualquiera de las otras técnicas cuando el número de faltantes por fila es aleatorio.

Así pues, aunque el algoritmo presentado supera a los demás, como los métodos anteriores con los que se compara, también tiene algunas limitaciones, tendría problemas de imputación en aquellos casos en los que la mayoría de los datos faltantes pertenecían a la misma variable o estaban concentrados en un cierto subconjunto de variables en lugar de distribuirse entre todas las variables del conjunto de datos.

Actualmente, se continúan desarrollando algoritmos híbridos que mejorarían los resultados de los existentes cuando tienen que abordar este tipo de problema. Dada la transversalidad de este tipo de técnicas, su aplicación en el ámbito de la Ingeniería Eléctrica puede tener otras variantes, además de la que ha sido objeto de estudio en esta Tesis. De una forma directa, cabe citar dos líneas de trabajo futuras que podrían suponer la aplicación inmediata de las técnicas de imputación expuestas. En primer lugar, estaría la monitorización y el diagnóstico de máquinas eléctricas. En la actualidad, las técnicas que se usan para el mantenimiento, ensayo y diagnóstico de estos dispositivos se basan en las tendencias que adoptan las variables de seguimiento que se obtienen mediante la realización de ensayos. En muchos casos, en estos ensayos se producen pérdidas de información o medidas erróneas que podrían dar lugar a un diagnóstico equivocado. Ante estas situaciones, los datos faltantes, derivados ocasionalmente de una mala práctica en la ejecución de los ensayos, podrían ser objeto de tratamiento conforme a los algoritmos desarrollados en la Tesis. De igual forma, la monitorización de la maquinaria eléctrica con fines de mantenimiento y diagnóstico implica el seguimiento continuo de múltiples variables: tensiones, corrientes y par figuran a menudo entre ellas. En estos casos, se pueden generar situaciones en las que se produzca la pérdida de datos o la existencia de información errónea y, en consecuencia, las técnicas expuestas en la Tesis pueden ser de aplicación.

El otro campo, dentro de la Ingeniería Eléctrica, en el que se podría profundizar mediante la imputación de datos faltantes, es el del estudio de las descargas parciales. Este fenómeno, habitual en el equipamiento de alta tensión, merece una mención aparte, ya que, además de ser estudiado en el ámbito del mantenimiento y diagnóstico de las máquinas eléctricas, también es objeto de estudio siempre que se manejan dispositivos e instalaciones de alta tensión. La caracterización del fenómeno de las descargas parciales se realiza de varias formas y a partir de múltiples variables (valor máximo, tasa de repetición, tensiones de aparición y extinción, resolución en fase, etc.). Dado que, en muchos casos, las medidas de estas variables se realizan en ambientes y condiciones propicios a que existan interferencias, es frecuente la pérdida de información o que ésta sea espúrea. Por ello, la aplicación de técnicas de imputación también tendría un perfecto encaje en este campo de investigación.

9. ARTÍCULOS PUBLICADOS

Sensors **2014**, *14*, 20382-20399; doi:10.3390/s141120382

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions

Concepción Crespo Turrado [1], María del Carmen Meizoso López [2], Fernando Sánchez Lasheras [3,*], Benigno Antonio Rodríguez Gómez [2], José Luis Calvo Rollé [2] and Francisco Javier de Cos Juez [4]

1 Maintenance Department, University of Oviedo, San Francisco 3, Oviedo 3307, Spain; E-Mail: ccrespo@uniovi.es

2 Departamento de Ingeniería Industrial, University of A Coruña, A Coruña 15405, Spain; E-Mails: mmeizoso@udc.es (M.C.M.L.); benigno@udc.es (B.A.R.G.); jlcalvo@udc.es (J.L.C.R.)

3 Department of Construction and Manufacturing Engineering, University of Oviedo, Gijón 33204, Spain

4 Project Management Area, Mining Department, University of Oviedo, Oviedo 33004, Spain; E-Mail: fjcos@uniovi.es

* Author to whom correspondence should be addressed; E-Mail: sanchezfernando@uniovi.es; Tel.: +34-984-833-135; Fax: +34-985-182-433.

External Editor: Vittorio M.N. Passaro

Received: 7 September 2014; in revised form: 8 October 2014 / Accepted: 21 October 2014 / Published: 29 October 2014

Abstract: Global solar broadband irradiance on a planar surface is measured at weather stations by pyranometers. In the case of the present research, solar radiation values from nine meteorological stations of the MeteoGalicia real-time observational network, captured and stored every ten minutes, are considered. In this kind of record, the lack of data and/or the presence of wrong values adversely affects any time series study. Consequently, when this occurs, a data imputation process must be performed in order to replace missing data with estimated values. This paper aims to evaluate the multivariate imputation of ten-minute scale data by means of the chained equations method (MICE). This method allows the network itself to impute the missing or wrong data of a solar radiation sensor, by using either all or just a group of the measurements of the remaining sensors. Very good results have been obtained with the MICE method in comparison with other methods employed in this field such as Inverse Distance Weighting (IDW) and Multiple Linear Regression

(MLR). The average RMSE value of the predictions for the MICE algorithm was 13.37% while that for the MLR it was 28.19%, and 31.68% for the IDW.

Keywords: missing data imputation; multivariate imputation by chained equations (MICE); multiple linear regression; solar radiation; pyranometer

1. Introduction

A meteorological or climate observation network is composed of a set of weather stations. They are usually placed at isolated points of a geographical zone, in order to determine the values of the meteorology and climatology variables of that area. Multiple variables such as air temperature, atmospheric pressure, wind speed and direction, relative humidity, rainfall, solar radiation, *etc.*, are measured and registered by each station and finally the data is sent to a central database of the network to be processed and stored [1].

Failures in the measurement process may occur for any variable, and consequently lack of and/or incorrect data can appear. These errors can often be detected, although only sometimes corrected, since the more variable the measurements are, the more erroneous or less precise the data imputation results.

Solar irradiation records in particular depend on the combined effects of both astronomical and meteorological events. Atmospheric conditions modify the extraterrestrial solar irradiation in such an ostensibly random manner that the global solar irradiation on the horizontal surface presents evolution randomness, with temporal and spatial variations due to weather conditions [2].

The amount of solar energy incident on the ground depends heavily on the state of the sky. Previous research has reported that among all the variables causing heterogeneity of solar radiation in the ground, atmospheric conditions such as cloud cover [3], aerosols and water vapor are the most important. In the case of cloud cover, factors such as cloud base height, cloud evaporation and formation and velocity have been reported as important [4]. In the case of aerosols a considerable reduction in the UV intensity has been observed during periods of high aerosol loading [5]. Previous research also demonstrated that the more water vapor solar radiation finds, the smaller the amount of solar energy present on the ground [6].

Global radiation includes radiation received directly from the solid angle of the sun's disc, as well as diffuse sky radiation that has been scattered in traversing the atmosphere. Global solar radiation (G) measurements at ground level are made primarily with pyranometers that use thermo-electric, photoelectric, pyro-electric or bimetallic elements as sensors [7].

The World Meteorological Organization (WMO) classifies pyranometers as secondary standard, first class and second class meters according to their measurement performance characteristics, such as spectral range, sensitivity, directional response, non-stability, temperature response, response time and non-linearity among others. Meteorological networks do not have the same type of equipment in all of their weather stations and, consequently, the radiation measurement quality varies from one to another. The diversity of the instruments installed with different accuracy levels, together with the need for the frequent instrument calibration, makes it very difficult to achieve a homogeneous data base [8].

The following problems have been reported [9] for solar global radiation: No signal from the sensor, an unstable signal, a lower or higher signal than physical limits, data not collected or stored and diurnal profiles systematically asymmetric with respect to the solar noon, among others. The following can be noted as likely reasons for such failures [10,11]: A damaged cable or with corrosion; the loss of proper electrical grounding; alterations in programs of data logger systems; moisture inside an element of the pyranometer; reflected radiation from properly-positioned towering cumulus clouds exceeding the solar constant for periods of less than ten minutes; half-melted frost or snow on the dome; communications failure, *etc.*

Some of these causes are temporary and may disappear spontaneously, but others require the intervention of a maintenance task force, and therefore errors persist for different periods of time. Lack of data or the presence of erroneous data adversely affects the study of any time series. For instance, solar energy applications need continuous radiation data time series to correctly assess the usefulness of the particular application and in its implementation, so additional procedures have been established to fill in missing values (where data is initially lacking or has been removed via quality checks) in the time series of solar radiation data [12]. Consequently, a process of data imputation may be followed for filling data series with estimated values.

Different criteria can be applied in order to obtain a missing value or a set of missing values in a series of solar irradiation data (G). Deterministic, random or mixed methods are available for this purpose. Some specific examples are discussed below.

Physical models such as MRM [13], ESRA [14], REST2 [15] or SIRAMix [16], which account for the estimated solar irradiation in terms of physical variables (aerosols, precipitable water, turbidity coefficients, total ozone in the vertical column, dry-bulb temperature, site pressure, *etc.*) are a good solution when the values of these auxiliary variables are known. The main advantage offered by these models is their spatial independence. In addition they do not require solar radiation data measured at the Earth's surface. However, the physical methods need complementary meteorological data to characterize the interactions of solar radiation with the atmosphere. SAs an example we can cite the ESRA method, which needs five inputs or the REST2 method, which needs a total of 10 inputs. Other methods that estimate solar radiation form satellite images have been used and tested by several authors [17,18].

Correlation relationships can be established as a second approximation. If weather stations are sufficiently close together, the difference in G records should be small, and then the space distance criteria can be applied. This argument is valid for clear or cloudy skies but performs worse for partly cloudy skies over short time scales. Also, when the missing data is only one, among known data, a simple interpolation can be the best solution [19].

Autoregressive properties of the signal allow us to take into account more separated values to auto-complete the series. ARIMA techniques were used to forecast solar radiation time series [20], so ARIMA models can be used to impute data in time series.

A similar study in terms of number of stations and time scales has been carried out [21], although in a different geographic area. This study proved that it was better to interpolate sequences of up to four missing values, and the first and the last missing values in longer sequences, using data from the same site. Otherwise it is better to use simultaneous data from the other sites of the network.

All the methods' performance is influenced by the time granularity. Weather stations provide data in several time scales: measuring equipment supplying data in the lowest scale and upper scales are built based on it. Normally, missing values in longer scales are estimated with less error than in shorter ones.

Several studies used Inverse Distance Weighting (IDW) to estimate solar irradiation [22–24], and also to interpolate other variables like temperature or precipitation [25–27].

Techniques for estimating data are essentially applications of statistics, but they should also rely on the physical properties of the system under consideration. IDW is a function based on one parameter, *i.e.*, distance, and it assumes that the region has a uniform characteristic [26]. A "cut-off" criterion is often used to limit either the distance to the locations considered or the number of observations considered [28].

In [22], working with hourly observations, it was concluded that the interpolations for distances beyond 34 km show an RMSE over 25% and it is suggested that in this case satellite measurements are more accurate. Consequently, the selected geographical area for this study is a small zone with a high density of measurement locations. The maximum distance between stations is less than 20 km.

In [23], working with monthly mean values, the distances between stations are over 95 km, and the relative error is under 29%. Finally in [24], the authors worked with 15-min observations but the RMSE was obtained by calculating daily aggregations and eliminating days identified as atypical. The values for RMSE are between 26% and 40%. Random errors tend to decrease when the data are averaged over a particular time period.

Kriging is a spatial interpolation approach that has been applied to estimate monthly irradiation [29]. However, the low number of stations and the high variability among the ten-minute scale data advise against the use of this method.

Regression models have been widely used for the estimation of global solar radiation. The most frequently selected variable is sunshine duration, where the Angstrom-Prescott-Page model is the main exponent for the monthly average daily global radiation [30]. Air temperature also appears in many models, being the Bristow-Campbell being one of the most widely-used models due to its simplicity and the availability of input data [31,32]. However, in most cases, the estimated radiation is a daily or monthly average daily value, and to our knowledge there are no sub-hourly regression models using temperature.

Cloud cover, relative humidity or wind velocity are other variables used to estimate solar radiation [33,34]. In [33], a method is developed to estimate hourly solar radiation by using relative humidity and ambient temperature in order to obtain a matrix of atmospheric transmittance coefficients that need to be adjusted to the particular area. The RMSE of this estimation method is 8.3%. The reason for this low value is that the method estimated hourly solar radiation. Time granularity is a key factor to compare the performance of the different methods. The longer the time scale, the fewer the errors. This reference is an example of a correlational method, but the error cannot be compared with the present research, as the authors used hourly data.

The aim of this paper is to evaluate a method which allows the network itself to fill the missing data of a sensor, using either all or just a group of the measurements of the remainder sensors. It is therefore necessary to work with ten-minute scale data, which makes it a challenge.

The rest of the paper is organized as follows: Section 2 includes a geographical description of the chosen study area and the dataset, and also gives the main characteristics of the sensors. Section 3, describes the three interpolation/imputation methods used and the validation criteria applied. Section 4 presents a comparison of the results achieved with each method. Finally the conclusions are drawn.

2. Experimental Section

2.1. Description of Study Area and Data

The current study uses ten-minute data collected from the MeteoGalicia network [29], a regional meteorological service with more than 100 locations located in Galicia (Spain) providing global radiation data. The network integrates stations with both meteorological and agro-climatic purposes and the regional government openly offers observations from its stations on the Internet [35].

The study area extends over a small area located in northwest Spain (Figure 1), between 42 °24'–42 °34' northern parallels and 8 °42'–8 °52' western meridians, covering an area of approximately 254 km^2. Bordered by the Atlantic Ocean to the West, and situated between two coastal bays, called "rias" (R á de Villagarc á and R á de Pontevedra), it has a temperate maritime climate and is one of the Galician areas that receives more solar radiation [36].

Figure 1. Map of the area studied.

Grapevine cultivation has a widespread presence in the region. This is the main reason for the high density of meteorological stations in the area. The dataset collected for this study comes from nine closely spaced meteorological stations. The greatest distance between them is less than 20 km. Table 1 shows the distance and the correlation coefficients matrix of the variable solar radiation for all the meteorological stations involved in the present study.

Geographical distribution and some climatological parameters (yearly mean values for 2013) of the studied radiometric stations are presented in Table 2. The data collected represents the global horizontal solar radiation ($W \cdot m^{-2}$) and is available on a 10-min scale basis, therefore, each station supplies 144 values per day.

The nine series ran from 21 December 2012 to 20 January 2014. The period of measurements was chosen to take into account seasonal variability [37]. Each station registered a maximum of 56,880 observations during this period, and the dataset includes a total of 511,345 observations. During the period of study, data missing for each station were: A Armenteira (1.00%), A Lanzada (0.004%), Sanxenxo (0.004%) and Corón (0.002%). No data was missing in the other stations.

Table 1. Distances among the studied stations (km) and correlation coefficients (CC).

Distances (Km)								
Torrequintáns	8.84	12.18	7.95	3.43	15.84	8.70	16.36	8.70
0.91	Castrelo	6.56	5.71	6.60	7.01	8.93	9.80	9.87
0.90	0.89	Simes	4.44	8.81	8.66	6.32	4.19	16.16
0.89	0.87	0.95	A Armenteira	4.52	11.08	3.25	8.60	13.46
0.93	0.90	0.89	0.89	P é Redondo	13.46	5.64	13.0	10.32
0.89	0.92	0.89	0.86	0.88	A Lanzada	14.12	8.76	14.83
0.90	0.88	0.90	0.90	0.89	0.88	Castrove	9.97	15.75
0.90	0.91	0.92	0.87	0.89	0.93	0.90	Sanxenxo	19.66
0.90	0.91	0.87	0.85	0.88	0.90	0.87	0.89	Corón
Correlation Coefficients								

Table 2. Geographical and climatological parameters for the stations studied.

Station	**Lat.**	**Long.**	**Elev**	**Temperature (°C)**			**Humidity**	**Precipitation**	**Irradiation G**	**Sunshine Hours**
	(°)	**(°)**	**(m)**	**Min.**	**Max**	**Mean**	**(%)**	**(L/m²)**	**(10 kJ/m²day)**	**(h)**
Castrelo	42.49	−8.80	32	5.0	24.8	14.6	84.1	128.8	1346.1	184.5
Simes	42.44	−8.77	97	4.8	26.0	14.7	80.4	147.7	1510.2	221.2
A Armenteira	42.47	−8.74	256	3.4	24.5	13.1	82.3	211.9	1135.7	152.8
P é Redondo	42.51	−8.73	150	6.0	25.5	14.6	80.4	175.4	1246.1	166.1
A Lanzada	42.46	−8.88	9	5.7	25.0	14.7	80.1	128.1	1425.4	201.1
Castrove	42.46	−8.70	424	5.5	23.5	12.9	80.6	206.3	1338.2	180.8
Sanxenxo	42.41	−8.80	34	6.6	25.5	15	77.3	147.5	1464.8	188.3
Corón	42.58	−8.80	3	7.4	23.8	14.9	78.5	149.7	1453.5	199.3
Torrequintáns	**42.54**	**−8.72**	**52**	**4.5**	**25.7**	**14.5**	**73.9**	**121.0**	**1318.8**	**175.5**

2.2. Sensors for the Measurement of the Solar Radiation Flux Density: Pyranometers

The nine stations employed in the present research have three different pyranometers. Table 3 shows the main properties which are of concern when evaluating the quality of these instruments. Figure 2 shows one pyranometer model and how it is placed in the meteorological station.

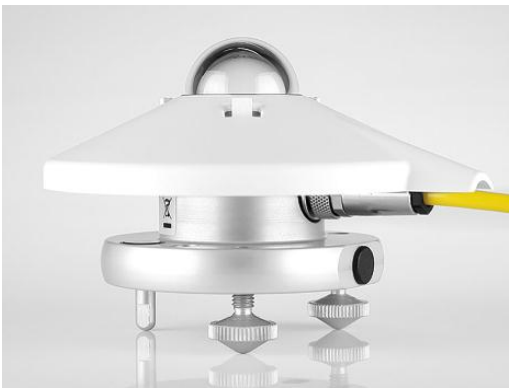
The thermopile detectors are sensitive to the whole shortwave spectrum in contrast to the solid-state silicon photodiodes. The basic uncertainties in the best practical solar radiation data available are roughly 5% in total global horizontal [38].

Table 3. Characteristics of the network pyranometers.

	SKS 1110 Skye	SP-LITE 2 Kipp & Zonnen	CMP-3 Kipp & Zonnen
Sensor	**Silicon Photocell**	**Photodiode**	**Thermopile**
Maximum operational irradiance	5000 W/m^2	2000 W/m^2	2000 W/m^2
Spectral range	0.4–1.1 μm	0.4 to 1.1 μm	0.3–2.8 μm
Sensitivity	1 mV/100 W/m^2	60–100 μV/W/m^2	5 to 20 μV/W/m^2
Directional response	5% max	<10 W/m^2	<20 W/m^2
Non-stability (Change/year)	±2%	<2%	<1%
Temperature response	±0.2%/°C	<0.15%/°C	<5% (−10 °C to +40 °C)
Response time	10 ns	<500 ns	<18 s
Non-linearity	<0.2%	<2.5% (0 to 1000 W/m^2)	<1.5% (100 to 1000 W/m^2)
Number of stations	5	3	1

Figure 2. (**a**) Corón meteorological station with its thermopile pyranometer inside a circle. The model is the CMP-3 (source:www.meteogalicia.es); (**b**) Picture of the CMP-3 sensor taken from Kipp-Zonnen pyranometers brochure.

(**a**)



(**b**)

3. Methodology

In order to facilitate modelling, only daytime (sunrise to sunset) solar irradiance readings were considered. No other filter was applied to remove outliers. The MeteoGalicia network provides a flag to identify the quality of each value measured (see Table 4). In the dataset there is no data with codes: 0, 4, or 5, and some with codes 2, 3 and 9, but only original validated data (code 1) has been considered. Therefore the dataset was divided into a training set, with two thirds of the samples (14,553 samples); and a test set, with the remaining one third of the samples (7276 samples) as usual. The training set is used to generate the models by different methods (MLR, MICE); the test set is used to validate the models with independent data, since the test set did not provide information to be able to build the models.

Table 4. Quality flags of the measurements provided by the network.

Quality Code	Description
0	No validated data
1	Original validated data
2	Suspect data
3	Erroneous data
4	Accumulated data
5	Interpolated data
9	**No registered data**

3.1. Inverse Distance Weighting (IDW)

IDW is a deterministic method of spatial interpolation. It is based on the distance between the location for which a value has to be interpolated and the locations of observations [28]. The point is that the solar radiation in a particular location presents a high correlation with the values registered in closed sites. Thus, it is possible to estimate solar radiation at any point through a linear combination of the values measured from neighboring sites. If G_E is the solar irradiance estimation at a site with no measurement, it can be calculated following the Equations (1) and (2) [23]:

$$G_E = \frac{\sum_{i=1}^{n} W(r_{i,E}) G_i}{\sum_{i=1}^{n} W(r_{i,E})} \quad (1)$$

$$W(r_{i,E}) = \frac{1}{r_{i,E}^{p}} \quad (2)$$

where G_i is the record of solar measured irradiation at site "*i*", with $i = 1, 2, \ldots, n$, $W(r_{i,E})$ is the weighting function between the *i*-th site, $r_{i,E}$ is the distance between the *i*-th solar irradiation measurement station and the estimation site, and "p" is the power parameter used in the interpolation. $p = 2$ is often chosen to provide even more weight to the closest locations [28]. We considered three cases: $p = 1, 1.5, 2$. Altitude was not taken into account for the distance calculations because the stations are all at practically the same height above sea level.

3.2. Multiple Linear Regressions Models (MLR Models)

The MLR models use a set of independent variables that helps to explain the independent variable; in this case, the measures of the neighboring stations were chosen as explanatory variables because of the high correlations between them (see Table 1). The correlation coefficient (CC) is expressed according to the following equation:

$$CC = \frac{\sum_{i=1}^{n} (Gx_i - \overline{Gx_i})(Gy_i - \overline{Gy_i})}{\sqrt{\sum_{i=1}^{n} (Gx_i - \overline{Gx_i})^2 \sum_{i=1}^{n} (Gy_i - \overline{Gy_i})^2}} \quad (3)$$

where Gx_i and Gy_i are the ten-minute measurements of global irradiation at the stations *x* and *y* respectively, and $\overline{Gx_i}$ or $\overline{Gy_i}$ are the mean of all the measures.

Multiple linear regression is a statistical method that accordingly models the relationship between a dependent variable (y) and a set of independent variables (x_1, x_2, …, x_p). The model can be represented as follows [39]:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (4)$$

where α is called the intercept, β_i are called the slopes or coefficients, ε is an error with zero mean and constant variance, and it is accepted that each independent variable has a linear relationship with the dependent variable.

Equation (4) can be rewritten in matrix form, *i.e.*:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (5)$$

In this study y_i are the ten-minute observations of one station, and x_{ij} are the ten-minute observations of the remaining eight. Therefore in this case $n = p = 9$ and β_p are the coefficient associated to location "*p*". In order to obtain the intercept and the coefficients we took the least square approach with a confidence interval of 95%. After attending to the values of *F* and *t* statistics, coefficients not significantly different from zero are set to zero for the model.

3.3. The MICE Algorithm

The Multiple Imputation by Chained Equations (MICE) algorithm developed by van Buuren and Groothuis-Oudshoorn [40] is a Markov Chain Monte Carlo Method where the state space is the collection of all imputed values. Like any other Markov Chain, in order to converge, the MICE algorithm needs to satisfy the three following properties [41–44]:

- *Irreducible*: The chain must be able to reach all parts of the state space;
- *Aperiodic*: The chain should not oscillate between different states;
- *Recurrence*: Any Markov chain can be considered as recurrent if the probability that the Markov chain starting from *i* will return to *i* is equal to one.

In practice, the convergence of the MICE algorithm is achieved after a relatively low number of iterations, usually somewhere between five and 20 [44]. According to the experience of the algorithm creator, in general five iterations are enough, but some special circumstances would require a greater number of iterations. In the case of the present research, and due to the performance of the results obtained when compared with the other methods applied, five iterations were considered to be enough. This number of iterations is much lower than in other applications of the Markov Chain Monte Carlo methods, which often require thousands of operations. In spite of these, and from a researcher's point of view and experience, it must be also remarked that in the most common of the applications each iteration of the MICE algorithm would take several minutes or even a few hours. Furthermore, the duration of each iteration is mainly linked with the number of variables involved in the calculus and not with the number of cases. It must be taken into consideration that imputed data can have a considerable amount of random noise, depending on the strength of the relations between the variables. So in those cases in which there are low correlations among variables or they are completely independent, the algorithm

convergence will be faster. Finally, high rates of missing data (20% or more) would slow down the convergence process work. The MICE algorithm [44] for the imputation of multivariate missing data consist on the following steps:

1. Specify an imputation model $P(Y_j^{mis}|Y_j^{obs}, Y_{-j}, R)$ for variable Y_j with $j = 1, \ldots, p$.
   The MICE algoritm obtains the posterior distribution of *R* by sampling interative from the above represented conditional formula. The parameters *R* are specific to the respective conditional densities and are not necessarily the product of a factorization of the true joint distribution.
2. For each j, fill in starting imputations Y_j^0 by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \ldots, T$ (iterations).
4. Repeat for $j = 1, \ldots, p$ (variables).
5. Define $Y_{-j}^t = (Y_1^t, \ldots, Y_{j-1}^t, Y_{j+1}^{t-1}, \ldots, Y_p^{t-1})$ as the currently complete data except Y_j.
6. Draw $\emptyset_j^t \sim P(\emptyset_j^t | Y_j^{obs}, Y_{-j}^t, R)$.
7. Draw imputations $Y_j^t \sim P(Y_j^{mis} | Y_j^{obs}, Y_{-j}^t, R, \emptyset_j^t)$.
8. End repeat j.
9. End repeat t.

In the algoritm referred to, *Y* represents a $n \times p$ matrix of partially-observed sample data, *R* is a $n \times p$ matrix, 0–1 response indicators of *Y*, and $\emptyset$ represents the parameters space. Please note that in MICE imputation [45], initial guesses for all missing elements are provided for the $n \times p$ matrix of partially observed sample. For each variable with missing elements, the data are divided into two subsets, one of them containing all the missing data. The subset with all available data is regressed on all other variables. Then, the missing subset is predicted from the regression and the missing values are replaced with those obtained from the regression. This procedure is repeated for all variables with missing elements. After this, all the missing elements are imputed according to the algorithm explained above, the regression and predictions are repeated until the stop criterion is reached. In this case, until a certain number of consecutive iterates fall within the specified tolerance for each of the imputed values.

3.4. Validation of the Models

Leave-one-out cross-validation has been used to analyze the spatial error of interpolated data [24,25]. This procedure involves using eight of the nine stations in the model to obtain the estimated value in the ninth station (this one is left out) in order to calculate RMSE and MAE for this station. The process is repeated nine times, once for each station.

The performance of the three methods has been evaluated using common statistics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) both expressed in W $\cdot$m^{-2}, and in percentage of the measured mean values, *i.e.*:

$$RMSE\ (Wm^{-2}) = \sum_{i=1}^{n} \sqrt{\frac{1}{n}(\widehat{G}_i - G_i)^2} \quad (6)$$

$$MAE(Wm^{-2}) = \frac{1}{n}\sum_{i=1}^{n} \left|\widehat{G}_i - G_i\right| \quad (7)$$

$$RMSE\ (\%) = \frac{RMSE}{\frac{1}{n}\sum_{i=1}^{n} G_i} x100 \quad (8)$$

$$MAE\ (\%) = \frac{MAE}{\frac{1}{n}\sum_{i=1}^{n} G_i} x100 \quad (9)$$

where G_i and $\widehat{G}_i$ are the measurements and the model-estimated values of global radiation respectively, and "*n*" is the number of ten-minute data points of the validation set. The RMSE weights large estimation errors more strongly than small errors and it is considered a very important model validation metric. Also, MAE is a useful complement of the measured-modeled scatter plot near the 1-to-1 line [45].

4. Results and Discussion

In this section the results of the different models tested are presented in order to compare their performance.

4.1. Results of IDW Method

Table 5 shows the RMSE and MAE for the IDW model. In spite of the short distance between the radiometric stations, the IDW model offers the poorest results. The influence of the power parameter "*p*" is barely noticeable, even though it is true that the small value of RMSE is obtained in most cases for $p = 2$. Exceptions are Castrelo, Castrove and Sanxenxo, where the lowest RMSE is obtained for $p = 1$; and Corón and Torrequintáns, where this value is obtained for $p = 1.5$. The average difference between the maximum and minimum values obtained for each station with different "*p*" was less than 1%.

Table 5. RMSE and MAE obtained with the IDW method.

	RMSE (%)			**MAE (%)**		
	***p* = 2**	***p* = 1.5**	***p* = 1**	***p* = 2**	***p* = 1.5**	***p* = 1**
Castrelo	29.90	29.49	29.13	23.03	22.93	22.89
Simes	31.55	32.09	32.87	35.81	37.54	39.66
A Armenteira	37.33	37.79	38.77	46.51	49.18	52.53
P é Redondo	29.67	29.98	30.56	37.53	38.44	39.77
A Lanzada	30.41	30.61	30.84	24.31	24.58	24.88
Castrove	35.84	34.87	34.34	44.05	46.90	50.29
Sanxenxo	34.88	33.93	33.25	31.51	31.89	32.54
Corón	32.70	32.69	32.73	29.13	29.12	29.17
Torrequintáns	27.21	26.73	26.87	28.50	28.69	29.32

4.2. Results of MLR Models

In Table 6 the multiple linear regression models for each location are presented. As expected, for each model the highest coefficient is related with the station that had shown the highest correlation (see CC in Table 1), and in some cases the stations with lower correlations have disappeared from the model. However, it is clear that the nearest locations do not always have the highest weight within the total of locations; in fact, this only occurs in four of the stations: P é Redondo, Castrove, Sanxenxo and

Torrequintáns. This explains the similar RMSE obtained for these MLR models (see Table 7) and the corresponding IDW models. An exception was Sanxenxo, whose MLR model offers a significant improvement, as with A Armenteira's. A detailed review of MLR models for Sanxenxo and A Armenteira shows that each of them has a relatively high negative coefficient, which tends to compensate for the high values given to the other stations. In the preliminary performance tests carried out with the dataset of one of the locations of this study, the air temperature was added to the regression model. However, the result, in terms of RMSE, was only slightly better (2%) with the addition of this variable, so finally no auxiliary variables were added to the MLR models.

Table 6. Multiple Linear Regression Models for each location.

Model Parameters	**Castrelo**	**Simes**	**A Armenteira**	**P é Redondo**	**A Lanzada**	**Castrove**	**Sanxenxo**	**Corón**	**Torrequintáns**
Interception	8.57	0	−8.41	7.45	13.30	−3.26	−2.50	23.23	2.10
β_1 (Castrelo)	-	0.06	0	0.16	0.31	0	0.16	0.27	0.11
β_2 (Simes)	0.09	-	0.69	−0.07	0	0.03	0.47	0	0.06
β_3 (A Armenteira)	0	0.60	-	0.21	0	0.37	−0.21	0	0.08
β_4 (P é Redondo)	0.17	−0.05	0.18	-	0.04	0.14	0.03	0.09	0.41
β_5 (A Lanzada)	0.27	0	0	0.03	-	0.08	0.36	0.24	0.03
β_6 (Castrove)	0.13	0.02	0.20	0.09	0.06	-	0.16	0.07	0.09
β_7 (Sanxenxo)	0.18	0.28	−0.15	0.03	0.36	0.21	-	0.09	0.07
β_8 (Corón)	0.12	0	0	0.06	0.19	0.07	0.06	-	0.15
β_9 (Torrequintáns)	**0.17**	**0.05**	**0.07**	**0.42**	**0.04**	**0.13**	**0.08**	**0.23**	-

Table 7. RMSE and MAE obtained with the MLR models.

Station	**RMSE (%)**	**MAE (%)**
Castrelo	26.80	28.40
Simes	25.55	22.73
A Armenteira	28.40	31.91
P é Redondo	28.38	41.73
A Lanzada	27.47	35.70
Castrove	33.81	48.30
Sanxenxo	26.98	32.93
Corón	29.78	45.63
Torrequintáns	**26.52**	**30.36**

4.3. Results of MICE Method

Table 8 shows the RMSE and MAE values obtained by means of the MICE algorithm for the nine meteorological stations in the study. Due to the random component of this algorithm, the procedure was applied five times for each of the stations. In order to verify that the results obtained for the five different iterations are better than those achieved by the other methods, the five iterations are presented. These tables also contain the average values of the five replications.

Table 8. RMSE and MAE obtained with MICE.

	Castrelo		Simes		A Armenteira		P éRedondo		A Lanzada		Castrove		Sanxenxo		Cor ón		Torrequint áns	
Iteration	**RMSE (%)**	**MAE (%)**	**RMSE (%)**	**MAE (%)**	**RMSE (%)**	**RMSE (%)**	**RMSE (%)**	**MAE (%)**	**RMSE (%)**	**MAE (%)**	**RMSE (%)**	**MAE (%)**	**RMSE (%)**	**MAE (%)**	**RMSE (%)**	**MAE (%)**	**RMSE (%)**	**MAE (%)**
1	12.74	11.03	11.93	10.26	13.29	12.57	13.39	16.71	12.57	11.98	15.86	21.11	12.41	14.22	13.79	13.94	12.86	13.36
2	12.60	10.70	11.79	10.46	13.33	12.46	13.26	16.56	12.46	12.08	15.58	20.57	12.43	14.19	14.00	13.85	12.32	13.25
3	12.60	10.72	12.27	10.59	13.46	12.67	13.41	17.03	12.67	12.12	15.72	20.62	12.43	14.16	14.08	14.71	12.44	13.03
4	12.68	10.72	12.00	11.12	13.64	12.59	13.12	16.20	12.59	12.21	15.69	20.41	12.64	14.75	14.03	13.92	12.63	13.40
5	12.55	10.67	11.80	10.33	13.15	12.35	13.56	16.97	12.35	11.69	15.66	20.26	12.45	14.49	13.87	13.61	12.53	13.35
Average	12.63	10.77	11.96	10.55	13.37	12.53	13.35	16.69	12.53	12.02	15.70	20.59	12.47	14.36	13.95	14.01	12.56	13.28

4.4. Comparison of the Three Methods

Finally, Tables 9 and 10 show the RMSE and MAE values of the average MICE algorithm in comparison with the MLR and IDW methods. As may be observed in Table 9, the RMSE results of the MICE algorithm are much lower than those for the other two methods: the average RMSE value of the nine meteorological stations for the MICE algorithm was 13.37% while that for the MLR it was 28.19% and 31.68% for the IDW. It must be highlighted that in all the stations the RMSE values using the MICE algorithm were lower than in the MLR and IDW methods. Similarly, Table 10 shows how MAE values of the results of MICE algorithm are lower in all the stations than in MLR and IDW methods. The average value of the MAE with the MICE algorithm for the nine stations was 14.14%, while the average values for the same stations obtained with MLR and IDW methods were of 35.30% and 34.19% respectively.

Table 9. Results of the three models in terms of RMSE.

Station	MICE (%)	MLR (%)	IDW (%)
Castrelo	12.74	26.80	29.13
Simes	12.27	25.55	31.55
A Armenteira	13.64	28.40	37.33
P é Redondo	13.56	28.38	29.67
A Lanzada	12.67	27.47	30.41
Castrove	15.86	33.81	34.34
Sanxenxo	12.64	26.98	33.25
Cor ón	14.08	29.78	32.69
Torrequint áns	**12.86**	**26.52**	**26.73**

Table 10. Results of the three models in terms of MAE.

Station	MICE (%)	MLR (%)	IDW (%)
Castrelo	11.03	28.40	22.89
Simes	10.59	22.73	35.81
A Armenteira	12.53	31.91	46.51
P é Redondo	16.97	41.73	37.53
A Lanzada	12.2	35.70	24.31
Castrove	21.11	48.30	50.29
Sanxenxo	14.75	32.93	32.54
Cor ón	14.71	45.63	29.12
Torrequint áns	**13.36**	**30.36**	**28.69**

All the calculi corresponding to the different algorithms were performed with a computer equipped with an Intel Xeon E5-1650 processor and 16 GB RAM. The average time of all the runs of the MICE algorithm was 334.66 s, with a standard deviation of 9.98 s. There were a few variations in the computer times, depending on the meteorological station estimated. The average time of the five MICE replications for each meteorological station was as follows: Castrelo 335.60 s, Simes 339.00 s, Armenteira 330.4 s, Pe Redondo 322.8 s, A Lanzada 328.6 s, Castrove 332.6 s, Sanxenxo 334.8 s, Cor ón 343.4 s and Torrequintans 334.7 s. The computational time required for the calculus of the results of IDW and MLR methods was always under one second in both cases.

5. Conclusions/Outlook

Solar radiation presents a very high variability at ten-minute scales and ostensibly random behavior in our geographical study area; hence data imputation is difficult when a datapoint or a set of data is lost. IDW and MLR methods show similar performance taking MAE and RMSE criteria. The use of auxiliary variables, such as temperature, does not represent a significant enhancement. The MICE algorithm performed better. In this paper the validity of the MICE method in imputing global solar radiation gaps has been demonstrated. In spite of its larger computational cost, better results were obtained by the MICE algorithm. It can therefore be stated that this algorithm is of great interest for all those applications not requiring an answer in real time.

The estimation of missing data is required for some statistical techniques such as time series analysis. These techniques enable prediction models to be created and improved. The interest in these kinds of models is multiple as they can be employed for the evaluation of the solar energy available in order to take technical decisions about the best solution for its exploitation, for the estimation of derivate variables such as evotranspiration and in combination with other parameters for the evaluation of harvest productivity.

Once they know the performance of the MICE for recovering lost data from nearby stations, the authors of the present research propose to evaluate the use of this method as an estimator of the ten-minute radiation values at those points placed in intermediate positions between stations where a solar radiation device is not permanently located.

In this case, the application of the MICE algorithm cannot be performed directly. The terrain must be taken into account to determine that horizon of the sun is visible where the estimate is made. By taking this into account, the MICE algorithm will impute only those values corresponding to a horizon clear of obstacles to the sun's trajectory.

This would make it possible to obtain historical series of solar radiation for any point where solar radiation is not directly measured. These historical series would be used by means of simulations in order to improve the calculus of the performance of photovoltaic facilities at any point it would be required. Finally, they could also be used to simulate the growth of crops, or in any other application in which the knowledge of the series of solar radiation is crucial.

Acknowledgments

Francisco Javier de Cos Juez and Fernando Sánchez Lasheras appreciate support from the Spanish Economics and Competitiveness Ministry, through grant AYA2010-18513. We would also like to thank the linguistics expert Anthony Ashworth and Philologist Javier Angel Rodríguez Gesto (EOI, Badajoz) for their revision of the English grammar and spelling of the manuscript.

Author Contributions

Francisco Javier de Cos Juez, José Luis Calvo Rollé, Concepción Crespo Turrado and Benigno Antonio Rodríguez Gómez conceived the study. María del Carmen Meizoso and Fernando Sánchez Lasheras programmed the required algorithms. María del Carmen Meizoso, Fernando Sánchez Lasheras, Benigno Antonio Rodríguez Gómez and Francisco Javier de Cos Juez

interpreted the results and drafted the manuscript; Concepción Crespo Turrado and José Luis Calvo Rollé supervised the experimental data analysis; They also contributed to the critical revision and improvement of the paper. All authors have approved the final version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Manual on the Global Observing System. Volume I (Anex V to the WMO Technical Regulations). *Global Aspects. 2003 (WMO-No. 544)*; World Meteorological Organization: Geneva, Switzerland, 2003; pp. III.1–III.20.
2. Şahin, A.D.; Şen, Z. Solar irradiation estimation methods from sunshine and cloud cover data. In *Modeling Solar Radiation at the Earth's Surface*; Badescu, V., Ed.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 145–173.
3. Badescu, V.; Dumitrescu, A. Simple models to compute solar global irradiance from the CMSAF product Cloud Fractional Coverage. *Renew. Energy* **2014**, *66*, 118–131.
4. Yang, H.; Kurtz, B.; Nguyen, D.; Urquhart, B.; Chow, C.W.; Ghonima, M.; Kleissl, J. Solar irradiance forecasting using a ground-based sky imager developed at UC San Diego. *Sol. Energy* **2014**, *103*, 502–524.
5. Badarinath, K.V.S.; Kharol, S.K.; Kaskaoutis, D.G.; Kambezidis, H.D. Influence of atmospheric aerosols on solar spectral irradiance in an urban area. *J. Atmos. Sol. Terr. Phys.* **2007**, *69*, 589–599.
6. Zajaczkowski, J.; Wong, K.; Carter, J. Improved historical solar radiation gridded data for Australia. *Environ. Model. Softw.* **2013**, *49*, 64–77.
7. World Meteorological Organization (WMO). *Guide to Meteorological Instruments and Methods of Observation*, 7th ed.; WMO: Geneva, Switzerland, 2008; pp. 157–190.
8. Bojanowski, J.S.; Vrieling, A.; Skidmore, A.K. A comparison of data sources for creating a long-term time series of daily gridded solar radiation for Europe. *Sol. Energy* **2014**, *99*, 152–171.
9. Vignola, F.; Michalsky, J.; Stoffel, T. *Solar and Infrared Radiation Measurements*; CRC: London, UK, 2012.
10. Muneer, T.; Fairooz, F. Quality control of solar radiation and sunshine measurements—Lessons learnt from processing worldwide databases. *Build. Serv. Eng. Res. Technol.* **2002**, *23*, 151–166.
11. Younes, S.; Claywell, R.; Muneer, T. Quality control of solar radiation data: Present status and proposed new approaches. *Energy* **2005**, *30*, 1533–1549.
12. Journée, M.; Bertrand, C. Quality control of solar radiation data within the RMIB solar measurements network. *Sol. Energy* **2011**, *85*, 72–86.
13. Kambezidis, H.D.; Psiloglou, B.E. The Meteorological Radiation Model (MRM): Advancements and Applications. In *Modeling Solar Radiation at the Earth's Surface*; Badescu, V., Ed.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 357–392.
14. Rigollier, C.; Bauer, O.; Wald, L. On the clear sky model of ESRA-European Solar Radiation Atlas-with respect to the Heliosat method. *Sol. Energy* **2000**, *68*, 33–48.

15. Gueymard, C.A. REST2: High performance solar radiation model for cloudless-sky irradiance, illuminance and photosynthetically active radiation-Validation with a benchmark dataset. *Sol. Energy* **2005**, *82*, 272–285.
16. Ceamanos, X.; Carrer, D.; Roujean, J.L. Improved retrieval of direct and diffuse downwelling surface shortwave flux in cloudless atmosphere using dynamic estimates of aerosol content and type: Application to the LSA-SAF project. *Atmos. Chem. Phys.* **2014**, *14*, 8209–8232.
17. Liang, S.; Zhao, X.; Liu, S.; Yuan, W.; Cheng, X.; Xiao, Z.; Zhang, X.; Liu, Q.; Cheng, J.; Tang, H.; *et al.* A long-term Global LAnd Surface Satellite (GLASS) data-set for environmental studies. *Int. J. Digit. Earth* **2013**, *6*, 5–33.
18. Chen, Y.; Xia, J.; Liang, S.; Feng, J.; Fisher, J.B.; Li, X.; Li, X.L.; Liu, S.; Ma, Z.; Miyata, A.; *et al.* Comparison of satellite-based evapotranspiration models over terrestrial ecosystems in China. *Remote Sens. Environ.* **2014**, *140*, 279–293.
19. Perdomo, R.; Banguero, E.; Gordillo, G. Statistical modeling for global solar radiation forecasting in Bogotá. In Proceedings of the 35th IEEE Photovoltaic Specialists Conference, Honolulu, HI, USA, 20–25 June 2010; pp. 2374–2379.
20. Reikard, G. Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Sol. Energy* **2009**, *83*, 342–349.
21. Glasbey, C.A. Imputation of missing values in spatio-temporal solar radiation data. *Environmetrics* **1995**, *6*, 363–371.
22. Perez, R.; Seals, R.; Zelenka, A. Comparing satellite remote sensing and ground network measurements for the production of site/time specific irradiance data. *Sol. Energy* **1997**, *60*, 89–96.
23. Şen, Z.; Şahİn, A.D. Spatial interpolation and estimation of solar irradiation by cumulative semivariograms. *Sol. Energy* **2001**, *71*, 11–21.
24. Gutierrez-Corea, F.V.; Manso-Callejo, M.A.; Moreno-Regidor, M.P.; Velasco-Gómez, J. Spatial estimation of sub-hour Global Horizontal Irradiance based on official observations and remote sensors. *Sensors* **2014**, *14*, 6758–6787.
25. Tiengrod, P.; Wongseree, W. A comparison of spatial interpolation methods for surface temperature in Thailand. In Proceedings of the International Computer Science and Engineering Conference (ICSEC), Nakorn Pathom, Thailand, 4–6 September 2013; pp. 174–178.
26. Ozelkan, E.; Bagis, S.; Ustundag, B.B.; Ozelkan, E.C.; Yucel, M.; Ormeci, C. Land Surface Temperature—Based Spatial Interpolation Using a Modified Inverse Distance Weighting Method. In Proceedings of the 2nd International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Fairfax, VA, USA, 12–16 August 2013; pp. 110–115.
27. Evrendilek. F.; Berberoglu, S.; Gulbeyaz, O.; Ertekin, C. Modeling Potential Distribution and Carbon Dynamics of Natural Terrestrial Ecosystems: A Case Study of Turkey. *Sensors* **2007**, *7*, 2273–2296.
28. World Meteorological Organization (WMO). *Guide to Climatological Practices 2011*; WMO: Geneva, Switzerland, 2011; pp. 5.1–5.13.
29. Mirás-Avalos, J.M.; Rodríguez-Gómez, B.A.; Meizoso-López, M.C.; Sande-Fouz, P.; González-García, M.Á.; Paz-González, A. Data quality assessment and monthly stability of ground solar radiation in Galicia (NW Spain). *Sol. Energy* **2012**, *86*, 3499–3511.
30. Ahmad, M.J.; Tiwari, G.N. Solar radiation models—A review. *Int. J. Energy Res.* **2011**, *35*, 271–290.

31. Daut, I.; Irwanto, M.; Irwan, Y.M.; Gomesh, N.; Ahmad, N.S. Combination of Hargreaves method and linear regression as a new method to estimate solar radiation in Perlis, Northern Malaysia. *Sol. Energy* **2011**, *85*, 2871–2880.
32. Meza, F.; Varas, E. Estimation of mean monthly solar global radiation as a function of temperature. *Agric. For. Meteorol.* **2000**, *100*, 231–241.
33. Dimas, F.; Gilani, S.; Aris, M. Hourly solar radiation estimation from limited meteorological data to complete missing solar radiation data. In Proceedings of the International Conference on Enviroment Science and Engineering IPCBEE, Singapore, 26–28 February 2011; Volume 8, pp. 4–8.
34. De Miguel, A.; Bilbao, J. Test reference year generation from meteorological and simulated solar radiation data. *Sol. Energy* **2005**, *78*, 695–703.
35. Meteogalicia. Estacións Meteoroloxicas. Available online: http://www2.meteogalicia.es/galego/observacion/estacions/listaEstacions.asp (accessed on 4 July 2014).
36. Pettazzi, A.; Salsón Casado, S. *Atlas de Radiación Solar de Galicia*; Xunta de Galicia. Consellería de Medio Ambiente, Territorio e Infraestructura (MeteoGalicia, Área de Observación e Climatología): Galicia, Spain, 2011; p. 26.
37. Muneer, T.; Younes, S.; Munawwar, S. Discourses on solar radiation modeling. *Renew. Sustain. Energy Rev.* **2007**, *11*, 551–602.
38. Gueymard, C.A.; Myers, D.R. Solar Radiation Measurement: Progress in Radiometry for Improved Modeling. In *Modeling Solar Radiation at the Earth's Surface*; Badescu, V., Ed.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 1–27.
39. Şahin, M.; Kaya, Y.; Uyar, M. Comparison of ANN and MLR models for estimating solar radiation in Turkey using NOAA/AVHRR data. *Adv. Space Res.* **2013**, *51*, 891–904.
40. Van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, 45, 1–67.
41. Roberts, G.O. Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman & Hall: London, UK, 1996; pp. 45–47.
42. Tierny, L. Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman & Hall: London, UK, 1996; pp. 59–71.
43. Van Buuren, S. *Flexible Imputation of Missing Data*; Chapman & Hall/CRC: London, UK, 2012.
44. Liu, Y.; Brown, S.D. Comparison of five iterative imputation methods for multivariate classification. *Chemom. Intell. Lab.* **2013**, *120*, 106–115.
45. Perez, R.; Lorenz, E.; Pelland, S.; Beauharnois, M.; van Knowe, G.; Hemker, K.; Heinemannb, D.; Remunde, J.; Müllere, S.C.; Traunmüllerf, W.; *et al.* Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Sol. Energy* **2013**, *94*, 305–326.

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

sensors

MDPI

Article

A New Missing Data Imputation Algorithm Applied to Electrical Data Loggers

Concepción Crespo Turrado [1], Fernando Sánchez Lasheras [2,*], José Luis Calvo-Rollé [3], Andrés José Piñón-Pazos [3] and Francisco Javier de Cos Juez [4]

Received: 7 October 2015; Accepted: 7 December 2015; Published: 10 December 2015
Academic Editor: Vittorio M. N. Passaro

1 Maintenance Department, University of Oviedo, San Francisco 3, Oviedo 33007, Spain; ccrespo@uniovi.es
2 Department of Construction and Manufacturing Engineering, University of Oviedo, Campus de Viesques, Gijón 33204, Spain
3 Departamento de Ingeniería Industrial, University of A Coruña, A Coruña 15405, Spain; jlcalvo@udc.es (J.L.C.-R.), andres.pinon@udc.es (A.J.P.-P.)
4 Prospecting and Exploitation of Mines Department, University of Oviedo, Oviedo 33004, Spain; fjcos@uniovi.es
* Correspondence: sanchezfernando@uniovi.es; Tel.: +34-984-833-135; Fax: +34-985-182-433

Abstract: Nowadays, data collection is a key process in the study of electrical power networks when searching for harmonics and a lack of balance among phases. In this context, the lack of data of any of the main electrical variables (phase-to-neutral voltage, phase-to-phase voltage, and current in each phase and power factor) adversely affects any time series study performed. When this occurs, a data imputation process must be accomplished in order to substitute the data that is missing for estimated values. This paper presents a novel missing data imputation method based on multivariate adaptive regression splines (MARS) and compares it with the well-known technique called multivariate imputation by chained equations (MICE). The results obtained demonstrate how the proposed method outperforms the MICE algorithm.

Keywords: missing data imputation; multivariate imputation by chained equations (MICE); Multivariate adaptive regression splines (MARS); quality of electric supply; voltage; current; power factor

1. Introduction

The presence of harmonics in an electrical system is associated with many problems in its performance. The main problems are overheating in conductors, especially in the neutral ones, due to the skin effect, and activating automatic breakers producing problems with supply continuity. Finally, the deterioration of the waveform of the voltage harmonic distortion associated would cause malfunctions of some devices.

As the existence of harmonics cannot be avoided, monitoring in real-time is necessary in order to control them within certain limits. Additionally, sometimes they can be transferred by acting on the installation in order to avoid its effects by means of filters either active or passive. In these cases, the use of isolation transformers, super-immunized differential breakers, *etc.*, must be studied.

Another problem frequently encountered in an electrical installation is the imbalance between phases. Although it is well known that balance is achieved by working at the highest levels of the installed capacity in order to take full advantage of the installation, sometimes this is not possible. An imbalance is usually caused by a bad load distribution between phases and provokes a high current return displayed by the neutral, as it has to compensate for the gap being at the center of the scheme vectors. These problems will increase if these charges are also producing linear and

harmonic distortion. In addition, imbalances may also cause the performance of the protection of the low voltage at the output of the transformer arise above its caliber in the overloaded phase currents.

In this context, the quality of electricity is a problem represented in all of its parameters: voltage, current, frequency anomalies, *etc*., that cause failures or disability of electrical or electronic devices [1]. Nowadays, the quality of electricity is a challenge in terms of efficiency, optimization, stability, fault prevention, and so on [2]. Science and technology have advanced, and continue to do so, significantly, with the aim of mitigating some of the consequent typical problems, which disturb electrical quality and, thus, the above mentioned challenges would be impossible to overcome, at least satisfactorily [3].

There are several different contributions with the above-mentioned aim. For example, in [4] a new power quality deviation index based on principal curves is proposed. [5] shows a complete review of signal processing and intelligent methods used for self-classification of power quality events and an influence of noise on recognition and classification of perturbations. A smart instrument used for recognition, labeling, and quantitation of power and energy quality disturbance is described in [6]. In the same way, [7] presents an intelligent instrument for instantaneous high-resolution frequency measurement in accordance with typical indicator values for the quality of electrical power control and monitoring, while [8] describes a communication infrastructure developed to obtain reliable data delivery with low cost, in order to avoid the problems in the provision of the power quality monitoring service.

In some buildings it would be of interest to monitor the main electrical parameters. This real-time monitoring and control is required in order to balance new loads and reduce the general consumption of the building by means of the assessment of the residual consumption (or consumption out of working hours). Such information is also useful to optimize the rates to be contracted. Additionally, this monitoring would be useful for studying supply problems due to lack of balance or harmonics, for analyzing the quality of the energy, and also for preventing incidents with the machinery due to poor signal quality. Finally, it would also be of interest to study the operation of the building and analyze its efficiency depending on parameters such us the number of people who use it, power installed, square meters in use, *etc.*

During the data collection process it is possible, due to different circumstances, for a small amount of the information retrieved to be lost. For these situations it is important to have missing data imputation. A process of missing data imputation consist of filling missing values in data series with estimated ones.

The quality of the electric supply of buildings is not only limited to the continuity of the supply, as concepts such as reliability, safety, and maintenance are also important indicators. It is also necessary that the available information be complete. The lack of information in some records, generally translated as zeros, distort the results.

There is also an important economic component of the data record, and that is the optimization of supply contracts, in other words, knowing the consumption of a building distributed over time. It is possible to associate the activities in the building so that we obtain a balanced installation, performing the most demanding activities at the most convenient hours of the day. To this end, it is necessary to collect information both before the decision and after the implementation of measures, in order to compare similar periods of expenditure. The latter would also serve in the event that energy-saving measures of another kind, such as replacing lighting by low consumption, placing detectors in corridors, placing inverters in circulation pumps, *etc*., were implemented.

This paper evaluates a new imputation method, which allows the system to fill in the missing data of any of the sensor devices that are used in this research for the recording of voltages, currents and power factors. The proposed algorithm is based on multivariate adaptive regression splines and outperforms the results obtained by a benchmark method, as it is the multivariate imputation by chained equations (MICE) [9].

Nowadays, the two major methods for missing data imputation are multiple imputation and maximum likelihood. The maximum likelihood chooses as parameter estimates those values which,

if true, would maximize the probability that have in fact been observed. The multiple imputation is based on different methodologies but all follow these steps: some random variation is introduced into the data set and several imputed data sets are generated. After that, those data sets are used for problem analysis and finally the combination of the results into a single set of parameter estimates, standard errors and test statistics is made. Since the missing at random (MAR) assumption cannot be checked from the data at hand, it is important to take into account if missing data can be considered as MAR. In those cases that cannot be considered, they called not missing at random (NMAR). Several models of NMAR data have been developed and its detailed analysis is beyond the scope of the present research.

The rest of the paper is organized as follows: Section 2 includes information about the measurement equipment employed and a description of the data recorded. Section 3 describes the new proposed algorithm and the benchmark technique employed detailing also the two metrics used for comparing their performance. A comparison of the results achieved with each method for different levels of missing data is presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Experimental Section

2.1. Measurement Equipment

In the present research, the devices employed are specific to the measurement of power quality variables, which are described in this section. They have some common measurement features in common, namely: Voltage Line/Neutral (V. L/N), Voltage Line/Line (V. L/L), Current by line (Current), Power Input/Output (+/− Watts), Energy Input/Output (+/− Wh), Reactive Power (+/− VARs), Reactive Power Input/Output (+/− VARh), Apparent Power (VA), Apparent Energy (VAh), Power Factor (PF), and Frequency (Frequency). Table 1 shows the accuracy for each device during the different electrical measurements. It should be noted that the values shown in percentage corresponds to the reading percentage.

Table 1. The variables accuracy for each device.

Variable	Units	S-100	S200	NEXUS 1252		MP200
				200 ms	1 s	
V. L/N	V, KV	0.1%	0.1%	0.1%	0.05%	0.3%
V. L/L	V, KV	0.1%	0.2%	0.1%	0.05%	0.5%
Current	A, KA	0.1%	0.1%	0.1%	0.025%	0.3%
+/− watts	W	0.2%	0.2%	0.1%	0.06%	0.5%
+/− wh	Wh	0.2%	0.2%	N/A	0.04%	0.5%
+/− VARs	VARs	0.2%	0.2%	0.1%	0.08%	1.0%
+/− VARh	VARh	0.2%	0.2%	N/A	0.08%	1.0%
VA	VA	0.2%	0.2%	0.1%	0.1%	1.0%
VAh	VAh	0.2%	0.2%	N/A	0.08%	1.0%
FP	+/−0.5 to 1	0.2%	0.2%	0.1%	0.08%	1.0%
Frequency	Hertz	1.10^{-2}	$+/-3.10^{-2}$	3.10^{-2}	1.10^{-2}	$+/-1.10^{-2}$

The four devices used in the present study can perform all the mentioned measurements [10]; also, each device has additional capabilities that are discussed below in the following subsections.

2.2. Shark 100 (S-100)

One of the options included for this equipment is the optical IrDA port, which allows the programing of the device from a laptop or personal digital assistant (PDA). Additionally, it incorporates V-Switch technology. This tool lets the users update and include the required functions using programing commands, even after the installation of the device installation. The offered VSwitches (VSw) offered are:

• VSw 1—Volts and Amperes Meter—Default.
• VSw 2—Volts, Amperes, kW, kVA, kVAR, Frequency, PF.
• VSw 3—Volts, Amperes, kW, kVA, kVAR, Frequency, PF, kVAh, kVARh, kWh and Distributed Network Protocol (DNP) v.3.0.
• VSw 4—Volts, Amperes, kW, kVA, kVAR, Frequency, PF, kVAh, kVARh, kWh, %THD (total harmonic distortion), Boundary Alerts and Distributed Network Protocol (DNP) v.3.0.

A RS485 Port can be added as an option. With it, communication is feasible by using Modbus or DNP 3.0 Protocols. In addition to the RS485, the device also incorporates a KYZ pulse, which is used to send instantaneous information regarding energy consumption to other devices. It is possible to add an Ethernet option with the INP10 module, which is a 10/100BaseT Ethernet with the Modbus TCP protocol.

2.3. Shark 200 (S-200)

The Shark 200 system is a small-size device used for power and energy measurements. It provides an invoicing measuring feature, in conjunction with an advanced data recording system, measurement of the electrical power quality, communication, and I/O capabilities. This equipment also includes V-Switch technology. The V-Switches in this case incorporate the features shown in the Table 2.

Table 2. Features of the V-Switches technology.

Feature	Vs1	Vs2	Vs3	Vs4	Vs5	Vs6
Input/Output Expansion and Multifunction Measurement	√	√	√	√	√	√
2 MB (Megabytes) datalogging (dl)		√	√	√		
3 MB -dl					√	
4 MB -dl						√
Harmonic Study			√	√	√	√
TLC (transformers line compensation) and CT (Current transformers) / PT (Power Current) Compensation	√	√	√	√	√	√
Functions for Control and Limits Configuration				√	√	√
64 SPC (samples per cycle) Waves Datalogger					√	
512 SPC Waves Datalogger						√

The Shark 200 device from feature V2 to V6, offers the possibility of data recording by using historic tendencies, limit alerts, input/output deviations, and events categorization. For the V5 and V6 models, the waveform can be recorded.

It is possible to make an independent CBEMA (Computer and Business Equipment Manufacturers Association's) log plotter: The system records an independent CBEMA and it makes an autonomous CBEMA record for size, as well as potential event times.

The S-200 model offers an on-line harmonic analysis from to the 40th up to the 255th order for current and voltage inputs.

Regarding communication, this model includes the following features:

- One port RS485 port allows communication using Distributed Network Protocol (DNP) v.3.0 or Modbus protocols.
- KYZ Pulse—this device incorporates Pulse Outputs mapped to total energy.
- Furthermore, it has an optical IrDA port with the same functions as the previously-explained model.

2.4. Nexus 1252

In general terms, this device has advanced features that offer a global view of power and energy usage and, of course, visualization of the quality of electrical power within a power network. The device is able to capture a maximum of 512 samples per working cycle by event. Additionally, this

device performs events analysis by 16 bits A/D converter, for electric voltage and electric current, which offers high-resolution. Furthermore, it is possible to activate a waveform datalogging by triggers that enable power quality surveys, fault detection, and the like, to be performed.

In terms of harmonic measurements, the device is capable of measuring up to the 255th order, in the case of current and voltage. If necessary, it can measure the harmonics in real time up to the 128th order. The device provides the THD percentage and the K-Factor with the harmonics. Additionally, it is possible to monitor switching noise from several elements of an installation. Like the previous device, the Nexus 1252 is able to make an independent CBEMA, and it makes an autonomous CBEMA log for size and time of potential events, which gives the consequent advantages mentioned previously.

In terms of communications, the device has four ports, and each one is able to communicate in several common protocols, with the aim of reading purposes and control simultaneously. Several peripherals are available for displaying or for external I/O options.

2.5. Shark MP200

The MP200 model measures and provides information of power usage from eight three-phase WYE circuits or from twenty-four single-phase systems. The MP200 system can create precise reports of power usage, analyze peak demand, and provide control signals to limit peak demand and billing based on usage and demand.

The MP200 offers communication possibilities like the previous models. One typical USB port and two standard RS485 ports, with optional RJ45 wired, or 802.11 WiFi, are provided. These ports support standard protocols as Modbus ASCII, RTU, and TCP/IP. By V-Switch options, the MP200 can be configured for basic sensors with real-time data (V1) to Advanced Logger up to 2400 Days (V3).

2.6. Description of the Data

The data set employed for the present research corresponds to measurements of the voltage phase to neutrum, (three variables) phase-to-phase voltage (three variables), current in each phase (three variables) and the average power factor (one variable) of a three-phase electrical supply of a building. The records were taken each 15 min from 27 November 2014 at 18:45 to 31 May 2015 at 23:45.

The building under study in the present research belongs to the University of Oviedo (Spain). This building is called Severo Ochoa after the Nobel Prize-winning scientist and has five floors and two basement levels that sum a total of 8150 m^2. This building holds the Information Technology Services of the University including their server rooms and some scientific laboratories that include equipment such as nuclear magnetic resonance spectrometers, electron microscopes, X-ray diffractometers, and the like. For all these kind of facilities it is essential to guarantee a good quality standard of electrical supply 24 h a day, every day of the week. A total of 78 employees work in this building, which has an average daily energy consumption of 190,572 kWh.

2.7. Harmonics and Harmonic Distortion

The large number of heterogeneous receivers in the building, such as computers, uninterruptible power supply devices, ballasts of fluorescent lighting systems, variable speed drives, induction ovens, and capacitors all create harmonic distortions in the net. All of these non-linear loads cause the flow of harmonic currents in the distribution system.

According to Fourier's theorem, a periodic continuous function $f(x)$ with a period of $2L$ may be expressed as the sum of a series of sine or cosine terms each of which has a specific amplitude and phase coefficients known as Fourier coefficients. This theorem can be expressed with the following formula [11]:

$$\mathrm{f(x)} = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} \left[a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right)\right] \quad (1)$$

where:

$$a_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{n\pi x}{L}\right)dx \tag{2}$$

$$b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin(\frac{n\pi x}{L})dx \tag{3}$$

Harmonic frequencies are multiples of the waveform's fundamental frequency. The harmonic distortion may be defined as the degree to which a waveform deviates from a pure sinusoidal wave. In the case of an ideal sine wave, its harmonic component is equal to zero. The total harmonic distortion (THD) is defined as the sum of all harmonic components of the voltage or current waveform compared to the fundamental component of the voltage or current wave. For the case of the current, the THD formula can be expressed as follows:

$$THD = \frac{\sqrt{\sum_{i=1}^{\infty} I_i^2}}{I_1} \tag{4}$$

where I_i represents the amplitude of the different harmonics.

3. Methodology

The data set is made up of a total of 17,763 samples that correspond to the period of time referred in the description of the data. It is used to test two different algorithms: multiple imputation by chained equations (MICE) and the proposed algorithm AAA (Adaptive Assignation Algorithm). The dataset is submitted to a process of random data deletion. This process consisted of supposing that the probability of an observation being missing does not depend on observed or unobserved measurements. It is called missing-completely-at-random (MCAR). The process of random data deletion was repeated five times for three different levels of missing data: 10%, 15%, and 20% of the total. After each deletion process, both algorithms were applied to the resulting data subset and the performance of the two methods compared.

3.1. Multivariate Adaptive Regression Splines (MARS)

The algorithm proposed in the present research is based on the computation of multivariate adaptive regression splines (MARS) models, for the prediction of the missing values. MARS is a multivariate nonparametric technique [12]. Its main purpose is to predict the values of a continuous dependent variable, $y\ (n \times 1)$, from a set of independent exploratory variables, $X\ (n \times p)$. This model can be represented by the following Equation [13,14]:

$$y = f(X) + e \tag{5}$$

where f is a balanced sum of basis functions that depend on X and e is the error vector. One of the main advantages of MARS models is that they do not require any *a priori* assumptions about the functional relationships between dependent and independent variables [15–17]. The reason is that this relation is driven by the basis function determined by the regression data (X, y).

MARS is a generalization of classification and regression trees [18] and is able to overcome some of the limitation of this method. The MARS regression model is constructed by means of basis functions called splines. These splines are defined as follows:

$$[-(x-t)]_+^q = \begin{cases} (t-x)^q & if\ x < t \\ 0 & otherwise \end{cases} \tag{6}$$

$$[-(x-t)]_+^q = \begin{cases} (t-x)^q & if\ x < t \\ 0 & otherwise \end{cases} \tag{7}$$

3.2. The Proposed Algorithm AAA

In order to introduce the new algorithm, let us assume that we have a dataset formed by n different variables v_1, v_2, ..., v_n. In order to calculate the missing values of the i-th column, all the rows with no missing value in the said column are employed. Then, a certain number of MARS models are calculated. It is possible to find rows with very different amounts of missing data from zero (no missing data) to n (all values are missing). Those columns with all values missing will be removed and will be neither used for the model calculation nor imputed. Therefore, any amount of missing data from 0 to $n - 2$ is feasible (all variables but one with missing values).

In other words, if the dataset is formed by variables v_1, v_2, ..., v_n. and we want to estimate the missing values in column v_i, then the maximum number of different MARS models that would be computed for this variable (and in general for each column) is as follows: $\sum_{k=1}^{n-1} \begin{pmatrix} n-1 \\ k \end{pmatrix}$. For the case of the data under study in this research, with 10 different variables, a maximum of 5110 distinct MARS models would be trained (511 for each variable).

Table 3 represents the 25 first rows of the dataset in which the algorithm will be applied. When the algorithm is applied to the third column of these datasets (variable v_3), all those rows with missing data (represented by means of the symbol 'o') in the third column are not employed for the calculus of the models (rows in red). If those rows were removed, different models would be trained for the prediction of v_3 using different subsets of variables. Continuing with the example of variable v_3 and taking into account the data missing in the 25 first rows, it would be possible to train the following models:

Model 1: a model that uses as output variable v_3 and the other nine as input variables (v_1, v_2, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}).

Model 2: a model that uses as output variable v_3 and as input variables v_2, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}.

Model 3: a model that uses as output variable v_3 and as input variables v_1, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}.

Model 4: a model that uses as output variable v_3 and as input variables v_1, v_2, v_4, v_6, v_7, v_8, v_9, v_{10}.

Model 5: a model that uses as output variable v_3 and as input variables v_1, v_2, v_4, v_5, v_7, v_8, v_9, v_{10}.

Model 6: a model that uses as output variable v_3 and as input variables v_4, v_5, v_6, v_7, v_8, v_9, v_{10}.

Model 7: a model that uses as output variable v_3 and as input variables v_1, v_5, v_6, v_7, v_8, v_9, v_{10}.

Model 8: a model that uses as output variable v_3 and as input variables v_1, v_4, v_5, v_6, v_8, v_9, v_{10}.

After the calculation of all the available models, the missing data of each row will be calculated using those models that employ all the available non-missing variables of the row. In those cases in which no model was calculated, the missing data will be replaced by the median of the column. Please note in that the case of large data sets with a not-too-high percentage of missing data, these will be an infrequent case. In the case of missing completely at random data, the probability, represented by letter Q, of not having at least two non-missing values in a certain row can be expressed by the following formula:

$$Q = p^n + (1 - p)\,(n - 1)\,p^{n-1} \tag{8}$$

where: N is the number of variables; P is the rate of missing data in a MCAR case.

In the case of our example, none of the rows was in this situation for the 10 and 15% of missing data, while in the case of 20% of missing data it happened only in one line (less than 0.006% of the total amount of lines). These results are in line with those expected by the formula.

As a general rule for the algorithm, it has been decided that when certain value can be estimated using more than one MARS model, it must be estimated using the MARS model with the largest

number of input variables; the value would be estimated by any of those models chosen at random. Finally, in those exceptional cases in which no model is available for estimation, the median value of the variable will be used for the imputation.

Table 3. Example of the dataset (25 first rows).

Row #	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	X	X	X	X	X	X	X	X	X	X	Yes	yes	yes	yes	yes	yes	yes	yes
2	X	o	X	o	X	X	X	X	X	X	No	no	yes	no	no	yes	yes	no
3	X	X	X	X	o	X	X	X	X	X	No	no	no	yes	no	no	no	no
4	X	X	o	o	X	X	X	X	X	X	No	no	no	no	no	no	no	no
5	X	X	X	X	X	X	X	X	X	X	Yes	yes	yes	yes	yes	yes	yes	yes
6	X	o	X	X	X	X	X	X	X	X	No	no	yes	no	no	yes	yes	yes
7	o	X	X	X	X	X	X	X	X	X	No	yes	no	no	no	yes	no	no
8	X	X	X	X	X	X	X	X	X	X	Yes	yes	yes	yes	yes	yes	yes	yes
9	o	o	o	X	X	X	X	X	X	X	No	no	no	no	no	no	no	no
10	X	X	o	X	X	X	X	X	X	X	No	no	no	no	no	no	no	no
11	X	X	o	X	X	X	X	X	X	X	No	no	no	no	no	no	no	no
12	X	o	o	X	X	X	X	X	X	X	No	no	no	no	no	no	no	no
13	X	X	X	X	X	X	X	X	X	X	Yes	yes	yes	yes	yes	yes	yes	yes
14	o	o	X	X	X	X	X	X	X	X	No	no	no	no	no	yes	no	no
15	o	X	X	X	X	X	X	X	X	X	No	yes	no	no	no	yes	no	no
16	X	X	X	X	X	X	X	X	X	X	Yes	yes	yes	yes	yes	yes	yes	yes
17	o	o	o	o	o	o	X	X	X	X	No	no	no	no	no	no	no	no
18	X	X	X	X	X	X	X	X	X	X	Yes	yes	yes	yes	yes	yes	yes	yes
19	X	o	X	X	X	X	o	X	X	X	No	no	yes	no	no	no	no	yes
20	X	X	X	X	X	o	X	X	X	X	No	no	no	no	yes	no	no	no
21	X	o	o	X	o	X	X	X	X	X	No	no	no	no	no	no	no	no
22	X	X	X	X	X	X	X	X	X	X	Yes	yes	yes	yes	yes	yes	yes	yes
23	X	X	o	o	X	X	X	X	X	X	No	no	no	no	no	no	no	no
24	X	X	X	X	X	o	X	X	X	X	No	no	no	no	yes	no	no	no
25	X	X	o	X	X	X	X	X	o	X	No	no	no	no	no	no	no	no

3.3. The Benchmark Rechnique: The MICE Algorithm

The algorithm called multiple imputation by chained equations (MICE) algorithm was developed by van Buuren and Groothuis-Oudshoorn [19]. This referred algorithm is a Markov Chain Monte Carlo Method in which the state space is the collection of all imputed values [9]. As with any other Markov Chain, the MICE algorithm has to accomplish three properties [20–23] in order to converge. The referred properties are as follows:

The chain must be able to reach all parts of the state space. This means that it is irreducible.

The chain should not oscillate between different states. In other words, the Markov Chain must be aperiodic.

Finally, the chain must be recurrent. This means, as in any other Markov Chain, that the probability of the chain of starting from i and returning to i will be equal to one.

According to the experience of the algorithm creator [19], and also from our own previous experience [9], the convergence of the MICE algorithm is achieved after a relatively low number of iterations, usually somewhere between five and 20 [23]. In the case of the present research, up to 20 iterations were considered but as not statistically significant improvements with respect to five iterations were achieved, the results for five iterations are presented.

The MICE algorithm [23] for the imputation of multivariate missing data consists of the steps that are listed in Algorithm 1. In this algorithm Y represents a $n \times p$ matrix of partially-observed sample data, R is a $n \times p$ matrix, $0-1$ response indicators of Y, and $\varnothing$ represents the parameters space. This methodology was already explained by the authors in previous research published in this journal [9]. For a more detailed explanation of the algorithm we recommend another look at the original research by van Buuren and Groothuis-Oudshoorn [23].

Algorithm 1: MICE algorithm for imputation of multivariate missing data [19].

1. Specify an imputation model $P(Y_j^{mis}|Y_j^{obs}, Y_{-j}, R)$ for variable Y_j with $j = 1, \ldots, p$.
2. For each j, fill in starting imputations Y_j^0 by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \ldots, T$ (iterations).
4. Repeat for $j = 1, \ldots, p$ (variables).
5. Define $Y_{-j}^t = \left(Y_1^t, \ldots, Y_{j-1}^t, Y_{j+1}^{t-1}, \ldots, Y_p^{t-1}\right)$ as the currently complete data except Y_j.
6. Draw $\varnothing_j^t \sim P(\varnothing_j^t \mid Y_j^{obs}, Y_{-j}^t, R)$.
7. Draw imputations $Y_j^t \sim P(Y_j^{mis} \mid Y_j^{obs}, Y_{-j}^t, R, \varnothing_j^t)$.
8. End repeat j.
9. End repeat t.

3.4. Performance of the Algorithms

The performance of the proposed algorithm in comparison with MICE has been evaluated using the mean absolute error (MAE) and the root mean square error (RMSE). MAE measures the average magnitude of the error in a set of forecasts without considering their direction. It is a linear score, which weights all the individual differences equally, while RMSE is a quadratic scoring rule, which measures the average magnitude of the error. In the case of the RMSE, as errors are squared before they are averaged, it gives a relatively higher weight to large errors. When results are analyzed using both variables, it should be noted that the greater the difference between them, the greater the variance in the individual errors in the sample, taking into account that the lower their values, the better the model.

The formulae for both kind of errors are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2} \tag{10}$$

where: n is the number of samples; e_i is the error of the i-th sample calculated as the difference of predicted value versus real value.

The present article uses both RMSE and MAE. The underlying assumption when presenting RMSE [24] is that the errors are unbiased and follow a normal distribution. The MAE is suitable to describe uniformly distributed errors. As model errors are likely to have a normal distribution, the RMSE is a better metric to present than the MAE for such kind of data. Although in the case of errors following a normal distribution RMSE is more appropriate to use than MAE, it is the preferred metric for the indication of the model average error.

4. Results and Discussion

In this section the results of the MICE algorithm and the proposed one AAA package are presented and their performances compared. As was already stated in the section describing the data, due to the random component of both algorithms, a process of MCAR data deletion of 10%, 15%, and 20% of the information was performed five times. The performance of both algorithms was compared by means of RMSE and MAE metrics. In order to verify that the results obtained with the proposed AAA package for the five different iterations were better than those achieved by other methods, the results of the five iterations are presented. Those tables also contain the average values of the five replications the iterations with the same number use the same database. Table 4 shows the

RMSE values of the MICE and the proposed AAA package when applied to a database with 10% of the data missing. As can be observed for the ten variables considered, the RMSE values obtained by the new algorithm are considerably lower than those obtained using the MICE method. On average they are 15 times lower, and in all cases the RMSE values of the proposed algorithm are considerably lower. The results obtained for missing data rates of 15% (Table 5) and 20% (Table 6) are similar to those obtained for the missing rates of 10%.

Table 4. RMSE obtained with a 10% of missing data MICE and new proposed AAA package.

RMSE MICE 10% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	19.6804	33.2038	20.8058	37.7042	39.6212	48.5060	0.2294	0.2501	1.8328	0.0031
2	17.5901	30.4712	22.9721	41.6324	28.1667	49.4147	0.3048	0.3419	1.7036	0.0032
3	17.7238	29.2717	22.8719	32.0352	37.9528	49.5132	0.2612	0.3322	1.8113	0.0030
4	16.1665	30.9164	20.1289	41.8040	28.4577	47.6496	0.2710	0.2344	1.8349	0.0033
5	18.8739	33.3065	20.9843	32.6096	40.9730	45.3324	0.2492	0.2768	1.6502	0.0031
Average	18.0069	31.4339	21.5526	37.1571	35.0343	48.0832	0.2631	0.2871	1.7666	0.0032
RMSE NEW ALGORITHM 10% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	1.5556	1.6047	0.9758	1.5621	2.1820	1.8054	0.1320	0.1446	0.1233	0.0020
2	1.1417	1.0847	1.0334	1.5623	2.0075	1.8990	0.1441	0.1397	0.1206	0.0020
3	1.0186	1.0077	0.8325	2.6758	1.6684	1.8550	0.1366	0.1458	0.1170	0.0020
4	1.0750	1.1247	1.1410	1.4569	1.7598	1.6958	0.1349	0.1558	0.1278	0.0017
5	1.1056	1.0992	0.9680	1.6783	1.9487	1.8209	0.1331	0.1317	0.1128	0.0021
Average	1.1793	1.1842	0.9901	1.7871	1.9133	1.8152	0.1361	0.1435	0.1203	0.0020

Table 5. RMSE obtained with a 15% of missing data MICE and new proposed AAA package.

RMSE MICE 15% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	17.0837	28.8975	23.1269	38.2302	26.7259	45.7781	0.3816	0.2352	1.8578	0.0031
2	19.7831	31.6292	21.6406	44.4176	31.1867	50.5978	0.2733	0.4289	1.6515	0.0030
3	16.8887	32.1573	23.2565	34.8709	36.4404	49.8771	2.0080	0.3768	0.4198	0.0032
4	18.9432	30.8065	21.1655	43.2729	32.0558	43.2723	0.3458	0.3326	1.7407	0.0028
5	19.0647	30.0262	23.5861	32.4402	28.9609	44.9738	0.5376	0.2517	1.8402	0.0034
Average	18.3527	30.7033	22.5551	38.6463	31.0739	46.8998	0.7092	0.3251	1.5020	0.0031
RMSE NEW ALGORITHM 15% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	1.0916	1.0625	0.9937	1.6562	1.7851	1.7874	0.1355	0.1314	0.1184	0.0021
2	1.1417	1.0061	0.9990	1.6799	1.9229	1.7843	0.1285	0.1446	0.1235	0.0019
3	1.1178	1.1311	1.0816	1.7174	2.3560	1.7111	0.1345	0.1460	0.1249	0.0021
4	1.5109	1.0043	1.1689	1.5117	1.9786	1.8057	0.1250	0.1394	0.1262	0.0018
5	1.1151	1.0109	1.0351	1.6381	2.5543	1.8637	0.1290	0.1364	0.1324	0.0019
Average	1.1954	1.0430	1.0556	1.6406	2.1194	1.7904	0.1305	0.1396	0.1250	0.0020

Something similar to the RMSE occurs with the values obtained for the MAE metric. In this case, also, the values obtained with the new AAA package are significantly lower than those obtained for the MICE algorithm in the three cases: 10% (Table 7), 15% (Table 8), and 20% (Table 9). Please also note that the average improvement of the MAE values in the case of the proposed algorithm is 15 times greater than the MAE values obtained by means of the MICE algorithm.

Table 6. RMSE obtained with a 20% of missing data MICE and new proposed AAA package.

RMSE MICE 20% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	16.5536	31.8986	22.5109	40.5584	32.8036	45.9128	0.2721	0.3250	1.8791	0.0032
2	18.5886	33.8541	23.3965	37.1860	28.7115	45.3711	1.9242	0.3857	0.4218	0.0031
3	18.0006	27.0257	22.4172	45.9451	29.7499	46.9085	0.4640	0.4657	1.8450	0.0031
4	18.4734	33.6455	22.6390	34.8455	42.4674	48.5727	0.2675	0.2951	1.6894	0.0032
5	19.0185	31.5739	22.9867	36.8936	30.8237	48.1872	0.2883	0.3618	1.8228	0.0029
Average	18.1269	31.5996	22.7900	39.0857	32.9112	46.9905	0.6432	0.3667	1.5316	0.0031
RMSE NEW ALGORITHM 20% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	1.0303	1.0031	1.0081	1.5567	1.7295	2.1989	0.1293	0.1535	0.1168	0.0018
2	1.3300	0.9645	1.4421	1.6901	1.9225	1.9970	0.1386	0.1430	0.1197	0.0019
3	1.4028	1.0751	1.0209	1.5444	1.6905	2.1216	0.1384	0.1462	0.1256	0.0018
4	1.1410	0.9442	0.9836	1.7262	1.8554	1.7554	0.1307	0.1391	0.1279	0.0019
5	1.0760	1.0285	1.0464	1.6981	1.8016	2.1696	0.1396	0.1351	0.1233	0.0019
Average	1.1960	1.0031	1.1003	1.6431	1.7999	2.0485	0.1353	0.1434	0.1226	0.0019

Table 7. MAE obtained with a 10% of missing data MICE and new proposed AAA package.

MAE MICE 10% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	15.3661	26.3015	16.4787	29.9178	30.6437	38.7431	0.1769	0.1910	1.3884	0.0026
2	13.4635	23.0373	19.0932	32.4891	23.0155	39.6558	0.2150	0.2234	1.2821	0.0026
3	13.9818	22.8360	18.2539	25.1375	29.5444	40.2728	0.2001	0.2107	1.3684	0.0024
4	12.9084	24.4469	16.2147	33.7170	22.3214	36.2061	0.2236	0.1892	1.3993	0.0026
5	14.9136	26.4011	16.7147	26.2535	32.1382	34.5887	0.1964	0.2052	1.2552	0.0025
Average	14.1266	24.6045	17.3510	29.5030	27.5326	37.8933	0.2024	0.2039	1.3387	0.0025
MAE NEW ALGORITHM 10% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.8954	0.8944	0.7444	1.1966	1.7099	1.4210	0.1053	0.1134	0.0971	0.0016
2	0.9017	0.8568	0.8358	1.2213	1.5561	1.5129	0.1120	0.1093	0.0960	0.0016
3	0.8236	0.7970	0.6533	1.4566	1.2741	1.4450	0.1123	0.1140	0.0881	0.0016
4	0.8617	0.8708	0.9240	1.1401	1.4034	1.3599	0.1045	0.1220	0.0975	0.0013
5	0.9135	0.8886	0.7926	1.3332	1.5911	1.4610	0.1057	0.1044	0.0920	0.0016
Average	0.8792	0.8615	0.7900	1.2696	1.5069	1.4400	0.1080	0.1126	0.0941	0.0016

For each of the ten variables involved in the present study, two-way ANOVA tests were performed in order to examine the influence of the kind of algorithm employed for the imputation (MICE versus proposed algorithm), the level of missing data (10%, 15% and 20%) and the interaction of both factors. These studies were carried out for the RMSE and MAE metrics. The influence of the model employed was found in all the variables for both metrics. Neither the percentage of missing data nor its interaction with the model employed for imputation were found to be significant in any of the variables. For the RMSE parameter the p-value was of $p < 0.001$ for the kind of model (MICE vs proposed algorithm) in the variables V_{an}, V_{bn}, V_{cn}, V_{ab}, V_{bc}, V_{ca}, I_c and PF, for I_a was $p = 0.044$ and for I_b $p = 0.001$. For the RMSE, when considering the variable percentage of missing data, there were no statistically significant differences between percentages (10, 15 and 20%) and the following p-values were obtained: 0.980 for V_{an}, 0.885 for V_{bn} , 0.106 for V_{cn}, 0.921 for V_{ab}, 0.591 for V_{bc} , 0.770 for V_{ca} , 0.523 for I_a, 0.168 for I_b, 0.800 for I_c, and 0.784 for PF.

Table 8. MAE obtained with a 15% of missing data MICE and new proposed AAA package.

MAE MICE 15% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	12.7819	21.9118	18.6842	30.7929	21.5323	35.7718	0.2421	0.1845	1.4006	0.0024
2	15.6837	24.3796	17.1327	34.6552	24.3977	40.5020	0.1959	0.2489	1.3048	0.0023
3	12.9838	25.1110	18.3944	28.7176	28.7719	38.0755	1.5148	0.2157	0.2240	0.0025
4	14.3735	23.8986	17.4651	34.6752	25.7770	33.3585	0.2164	0.2117	1.3023	0.0022
5	14.9162	23.6717	18.2033	25.8131	23.0634	35.0113	0.2675	0.1877	1.4250	0.0026
Average	14.1478	23.7945	17.9759	30.9308	24.7084	36.5438	0.4873	0.2097	1.1314	0.0024
MAE NEW ALGORITHM 15% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.8882	0.8196	0.7866	1.3280	1.4423	1.3884	0.1091	0.1053	0.0973	0.0017
2	0.9152	0.8158	0.7646	1.3088	1.4749	1.3937	0.1035	0.1117	0.0986	0.0016
3	0.8710	0.9064	0.8255	1.3577	1.5212	1.3480	0.1067	0.1121	0.0961	0.0016
4	0.9951	0.7757	0.9129	1.2065	1.5648	1.4341	0.1002	0.1114	0.0983	0.0014
5	0.8625	0.7992	0.7959	1.2726	1.7787	1.4867	0.1029	0.1043	0.1021	0.0015
Average	0.9064	0.8234	0.8171	1.2947	1.5564	1.4102	0.1045	0.1089	0.0985	0.0015

Table 9. MAE obtained with a 20% of missing data MICE and new proposed AAA package.

MAE MICE 20% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	12.6252	25.2004	18.4560	33.2591	25.8328	35.2414	0.2007	0.2133	1.4431	0.0025
2	15.0349	26.6725	18.2779	28.8681	22.4685	34.9493	1.4675	0.2465	0.2272	0.0024
3	14.1887	20.7764	18.1093	36.6143	23.2577	36.0519	0.2438	0.2499	1.4036	0.0025
4	14.1482	25.9181	18.7435	28.1440	33.7646	37.7483	0.1893	0.2033	1.2435	0.0025
5	14.7270	25.3376	18.5203	28.8294	24.8967	38.1530	0.1971	0.2105	1.3253	0.0023
Average	14.1448	24.7810	18.4214	31.1430	26.0441	36.4288	0.4597	0.2247	1.1285	0.0024
MAE NEW ALGORITHM 20% MISSING DATA										
Iteration	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.7814	0.8120	0.8039	1.2398	1.3721	1.4579	0.1012	0.1155	0.0926	0.0014
2	0.8745	0.7664	0.8475	1.3372	1.5238	1.5322	0.1096	0.1105	0.0950	0.0015
3	0.9177	0.8772	0.8167	1.2134	1.3097	1.4525	0.1083	0.1145	0.0971	0.0014
4	0.8624	0.7406	0.7719	1.3691	1.4522	1.3876	0.1046	0.1086	0.1028	0.0015
5	0.8350	0.8022	0.8032	1.3666	1.3894	1.4925	0.1096	0.1062	0.0967	0.0015
Average	0.8542	0.7997	0.8086	1.3052	1.4094	1.4646	0.1067	0.1110	0.0969	0.0015

In the case of the metric MAE, the p-value was of $p < 0.001$ for the kind of model in the variables V_{an}, V_{bn}, V_{cn}, V_{ab}, V_{bc}, V_{ca}, I_b , I_c, and PF; for I_a the p-value was of 0.038. Additionally, for the MAE metric, in the case of the variable percentage of missing data, there are no statistically significant differences between percentages (10%, 15% and 20%) obtaining the following p-values: 0.990 for V_{an}, 0.786 for V_{bn} , 0.113 for V_{cn}, 0.887 for V_{ab}, 0.655 for V_{bc} , 0.643 for V_{ca} , 0.686 for I_a, 0.315 for I_b, 0.796 for I_c, and 0.424 for PF.

Finally, all the calculi of both the MICE and the AAA algorithm was performed with a computer equipped with an Intel Xeon E5-1650 processor and 16 GB RAM. The average time of the MICE algorithm runs was of 123.54 s. The AAA algorithm average completion time was of 74.36 s with a standard deviation of 8.32 s. In both case the dataset was formed by 17,763 samples each on them with 10 variables.

5. Conclusions

The existence of harmonics in electrical installations is an unavoidable issue nowadays. The use of real-time data collection devices is indispensable. During the process collection, it is possible for some data to be missing and, in this context, the use of missing data imputation techniques is essential.

The algorithm proposed in this research greatly improves the results obtained by means of one of the most renowned and common techniques used today. From the point of view of the authors, this new algorithm is of great interest for applications like the one proposed in the present paper. In spite of the good performance of the proposed algorithm, it must be also be taken into account that the proposed algorithm, like many others, would have imputation problems in those cases in which most of the missing data belonged to the same column or to a reduced subset of columns. In future research the use of support vector machines (SVM) [23,25] and hybrid methods [26–28] will be explored by the authors in order to find a new algorithm with even higher performance. Furthermore, authors will try to study the nonlinear time varying systems and other power quality features, taken into account proposals like [29,30]. Finally, another research line that will be explored is the missing data imputation in the time-frequency domain. It consists on estimating missing regions of the time-frequency representation of signals [31]. In this kind of researches, the imputation methods also make use of harmonics for the imputation, considering for instance, that in a certain moment there is missing information but that not all the information of all the frequencies is necessary lost at the same time. The algorithms developed would be of interest for any kind of signals.

The estimation of missing data is required in many different applications, such as time series analysis. The use of missing data imputation techniques allows the creation of prediction models using incomplete datasets.

Acknowledgments: Francisco Javier de Cos Juez and Fernando Sánchez Lasheras appreciate support from the Spanish Economics and Competitiveness Ministry, through grant AYA2014-57648-P and the Government of the Principality of Asturias (Consejería de Economía y Empleo), through grant FC-15-GRUPIN14-017. We would also like to thank the linguistic expert Anthony Ashworth for his revision of the English grammar and spelling of the manuscript.

Author Contributions: Francisco Javier de Cos Juez, José Luis Calvo Rolle, Concepción Crespo Turrado and Fernando Sánchez Lasheras conceived the study. Andrés José Piñón Pazos and Francisco Javier de Cos Juez programmed the required algorithms. Fernando Sánchez Lasheras and Francisco Javier de Cos Juez interpreted the results and drafted the manuscript; Concepción Crespo Turrado, Andrés José Piñón Pazos and José Luis Calvo Rolle supervised the experimental data analysis; they also contributed to the critical revision and improvement of the paper. All of the authors have approved the final version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chattopadhyay, S.; Mitra, M.; Sengupta, S. Electric Power Quality. In *Electric Power Quality*; Springer: Dordrecht, The Netherlands, 2011; pp. 5–12.
2. Dixit, J.B.; Yadav, A. *Electrical Power Quality*; University Science Press: New Delhi, India, 2010.
3. Stones, J.; Collinson, A. Power quality. *Power Eng. J.* **2001**, *15*, 58–64. [CrossRef]
4. Ferreira, D.D.; de Seixas, J.M.; Cerqueira, A.S.; Duque, C.A.; Bollen, M.H.J.; Ribeiro, P.F. A new power quality deviation index based on principal curves. *Electr. Power Syst. Res.* **2015**. [CrossRef]
5. Mahela, O.P.; Shaik, A.G.; Gupta, N. A critical review of detection and classification of power quality events. *Renew. Sustain. Energy Rev.* **2015**. [CrossRef]
6. Granados-Lieberman, D.; Valtierra-Rodriguez, M.; Morales-Hernandez, L.; Romero-Troncoso, R.; Osornio-Rios, R. A Hilbert Transform-Based Smart Sensor for Detection, Classification, and Quantification of Power Quality Disturbances. *Sensors* **2013**, *13*, 5507–5527. [CrossRef] [PubMed]
7. Granados-Lieberman, D.; Romero-Troncoso, R.J.; Cabal-Yepez, E.; Osornio-Rios, R.A.; Franco-Gasca, L.A. A Real-Time Smart Sensor for High-Resolution Frequency Estimation in Power Systems. *Sensors* **2009**, *9*, 7412–7429. [CrossRef] [PubMed]
8. Lim, Y.; Kim, H.-M.; Kang, S. A design of wireless sensor networks for a power quality monitoring system. *Sensors* **2010**, *10*, 9712–9725. [CrossRef] [PubMed]
9. Turrado, C.; López, M.; Lasheras, F.; Gómez, B.; Rollé, J.; Juez, F. Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions. *Sensors* **2014**, *14*, 20382–20389. [CrossRef] [PubMed]
10. www.electroind.com. Available online: http://www.electroind.com/products/ (accessed on 8 December 2015).
11. Kammler, D.W. *A First Course in Fourier Analysis*; Cambridge University Press: Cambridge, UK, 2008.

12. Friedman, J.H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]
13. García Nieto, P.J.; Martínez Torres, J.; de Cos Juez, F.J.; Sánchez Lasheras, F. Using multivariate adaptive regression splines and multilayer perceptron networks to evaluate paper manufactured using Eucalyptus globulus. *Appl. Math. Comput.* **2012**, *219*, 755–763. [CrossRef]
14. Guzmán, D.; Juez, F.J.C.; Myers, R.; Guesalaga, A.; Lasheras, F.S. Modeling a MEMS deformable mirror using non-parametric estimation techniques. *Opt. Expr.* **2010**, *18*, 21356–21369. [CrossRef] [PubMed]
15. García Nieto, P.J.; Alonso Fernández, J.R.; Sánchez Lasheras, F.; de Cos Juez, F.J.; Díaz Muñiz, C. A new improved study of cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain) using the MARS technique. *Sci. Total Environ.* **2012**, *430*, 88–92. [CrossRef] [PubMed]
16. De Cos Juez, F.J.; Lasheras, F.S.; García Nieto, P.J.; Suarez, M.A.S. A new data mining methodology applied to the modelling of the influence of diet and lifestyle on the value of bone mineral density in post-menopausal women. *Int. J. Comput. Math.* **2009**, *86*, 1878–1887. [CrossRef]
17. Machon-Gonzalez, I.; Lopez-Garcia, H.; Calvo-Rolle, J.L. A hybrid batch SOM-NG algorithm. In Proceedings of the 2010 International Joint Conference on Neural Networks, Barcelona, Spain, 18–23 July 2010.
18. De Andrés, J.; Lorca, P.; de Cos Juez, F.J.; Sánchez-Lasheras, F. Bankruptcy forecasting: A hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). *Expert Syst. Appl.* **2011**, *38*, 1866–1875. [CrossRef]
19. Van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *1*, 1–67. [CrossRef]
20. Roberts, G.O. Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman and Hall: London, UK, 1996; pp. 45–47.
21. Tierney, L. Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman and Hall: London, UK, 1996; pp. 59–71.
22. Van Buuren, S. *Flexible Imputation of Missing Data*; Chapman & Hall/CRC Press: London, UK, 2012.
23. Liu, Y.; Brown, S.D. Comparison of five iterative imputation methods for multivariate classification. *Chemom. Intell. Lab. Syst.* **2013**, *120*, 106–115. [CrossRef]
24. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)? —Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
25. Álvarez Antón, J.C.; García Nieto, P.J.; de Cos Juez, F.J.; Sánchez Lasheras, F.; González Vega, M.; Roqueñí Gutiérrez, M.N. Battery state-of-charge estimator using the SVM technique. *Appl. Math. Model.* **2013**, *37*, 6244–6253. [CrossRef]
26. García Nieto, P.J.; Alonso Fernández, J.R.; de Cos Juez, F.J.; Sánchez Lasheras, F.; Díaz Muñiz, C. Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the Trasona reservoir (Northern Spain). *Environ. Res.* **2013**, *122*, 1–10. [CrossRef] [PubMed]
27. Quintian, H.; Calvo-Rolle, J.L.; Corchado, E. A hybrid regression system based on local models for solar energy prediction. *Informatica* **2014**, *25*, 265–282. [CrossRef]
28. Manuel Vilar-Martinez, X.; Montero-Sousa, J.A.; Calvo-Rolle, J.L.; Casteleiro-Roca, J.L. Expert system development to assist on the verification of "TACAN" system performance. *Dyna* **2014**, *89*, 112–121. (In Spanish)
29. Viveros, R.A.; Yuz, J.I.; Perez-Ibacache, R.R. Simultaneous State and Parameter Estimation for a Nonlinear Time-Varying System. *Rev. Iberoam. Autom. Inform. Ind.* **2014**, *11*, 263–274. [CrossRef]
30. Farias, G.; Dormido-Canto, S.; Vega, J.; Santos, M.; Pastor, I.; Fingerhuth, S.; Ascencio, J. Iterative noise removal from temperature and density profiles in the TJ-II Thomson scattering. *Fusion Eng. Des.* **2014**, *89*, 761–765. [CrossRef]
31. Smaragdis, P.; Raj, B.; Shashanka, M. Missing Data Imputation for Time-Frequency Representations of Audio Signals. *J. Signal Process. Syst.* **2011**, *65*, 361–370. [CrossRef]

© 2015 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons by Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).

sensors

MDPI

Article

A Hybrid Algorithm for Missing Data Imputation and Its Application to Electrical Data Loggers

Concepción Crespo Turrado [1], Fernando Sánchez Lasheras [2,*], José Luis Calvo-Rollé [3], Andrés-José Piñón-Pazos [3], Manuel G. Melero [4] and Francisco Javier de Cos Juez [5]

1 Maintenance Department, University of Oviedo, San Francisco 3, Oviedo 33007, Spain; ccrespo@uniovi.es
2 Department of Construction and Manufacturing Engineering, University of Oviedo, Campus de Viesques, Gijón 33204, Spain
3 Departamento de Ingeniería Industrial, University of A Coruña, A Coruña 15405, Spain; jlcalvo@udc.es (J.L.C.-R.), andres.pinon@udc.es (A.-J.P.-P.)
4 Electrical Engineering Department, University of Oviedo, Campus de Viesques, Gijón 33204, Spain; melero@uniovi.es
5 Prospecting and Exploitation of Mines Department, University of Oviedo, Oviedo 33004, Spain; fjcos@uniovi.es
* Correspondence: sanchezfernando@uniovi.es; Tel.: +34-984-833-135; Fax: +34-985-182-433

Academic Editor: Kemal Akkaya
Received: 13 July 2016; Accepted: 7 September 2016; Published: 10 September 2016

Abstract: The storage of data is a key process in the study of electrical power networks related to the search for harmonics and the finding of a lack of balance among phases. The presence of missing data of any of the main electrical variables (phase-to-neutral voltage, phase-to-phase voltage, current in each phase and power factor) affects any time series study in a negative way that has to be addressed. When this occurs, missing data imputation algorithms are required. These algorithms are able to substitute the data that are missing for estimated values. This research presents a new algorithm for the missing data imputation method based on Self-Organized Maps Neural Networks and Mahalanobis distances and compares it not only with a well-known technique called Multivariate Imputation by Chained Equations (MICE) but also with an algorithm previously proposed by the authors called Adaptive Assignation Algorithm (AAA). The results obtained demonstrate how the proposed method outperforms both algorithms.

Keywords: missing data imputation; multivariate imputation by chained equations (MICE); Mahalanobis distances; Self-Organized Maps Neural Networks (SOM); Adaptive Assignation Algorithm (AAA); Multivariate Adaptive Regression Splines (MARS); quality of electric supply; voltage; current; power factor

1. Introduction

Currently, the importance of problems due to harmonics in electric networks is growing. This fact is due to the increase in the amount of non-linear loads. The two main problems related to harmonics are the overheating of conductors due to the skin effect and the activation of automatic breakers, which produce problems for supply continuity. Additionally, distortion of the voltage waveform may cause the malfunction of some devices. The monitoring of harmonics in real time is required to control them.

Another common problem in electrical networks is the imbalance between phases. This is usually caused by a bad load distribution between phases and provokes a high current return displayed by the neutral, as it has to compensate for the gap existing at the centre of the scheme vectors.

Electricity quality is an important issue that is present in the following variables: voltage, current, frequency anomalies, etc. The quality affects all devices connected to the power network, causing failure of the systems or disability [1]. Currently, an electric system is analyzed in terms of efficiency,

stability and optimization to obtain better quality of the system [2]. With the aim of reducing the issues and improving electricity quality, science and technology are evolving to mitigate problems and overcome the problems mentioned above [3].

Different studies in this field have been performed: for instance, a novel power quality deviation parameter based on principal curves is presented in [4]. In [5], a review of the signal processing and intelligent techniques and methods employed in the self-classification of the events of power quality and the influence of noise on the recognition and classification of perturbations has been made. [6] describes a device capable of labelling, recognizing, and quantifying energy and power quality perturbation. An intelligent device for high-resolution frequency measuring that agrees with the common indicator standards is shown in [7]; it is used for electricity quality monitoring and control. Furthermore, [8] exposes a communication infrastructure created to obtain consistent data delivery at low cost, with the aim to prevent the difficulties of the power quality monitoring service.

Monitoring the main electrical variables in electric systems in some buildings might be interesting. Therefore, monitoring is useful for the control with the objective of balancing the loads of a building, thus reducing the consumption of the electric energy of the building by decreasing the remaining consumption (during non-working hours). The analysis of the electric system in buildings is useful for determining the optimized rates. Furthermore, it is also useful for the analysis of supply issues that can affect the different loads, which are caused by a lack of balance or harmonics, analysis of the energy quality and preventing incidents as a result of poor signal quality. Finally, the analysis of the building operation and its efficiency study might be of interest when accounting for the dependency of the people who use it, area in use, installed power, etc.

During the data-collection process, it is likely that a small amount of the information retrieved may be lost. In these cases, missing data imputation algorithms must be applied because the substitution of missing data with zeros is not acceptable. The present research evaluates a new imputation method that is able to predict the value of any missing data in the sensor devices that are used in this research for the recording of electrical variables. The new algorithm is based on Self-Organized Maps Neural Networks and Mahalanobis distances and hybridizes them with the algorithm called the Adaptive Assignation Algorithm (AAA). The results obtained are benchmarked with those given by the AAA and multivariate imputation by chained equations (MICE) [9].

The rest of the paper is arranged as follows: Section 2 describes the measurement equipment and the database, Section 3 details the proposed algorithm and how its performance is measured and compared. Section 4 presents the results obtained and its comparison with other algorithms. Finally, the conclusions are drawn in Section 5.

2. Materials and Methods

2.1. Measurement Equipment

In this section, the specific power quality measurement devices that are employed in this work are described (Figure 1). The next measurements are common to all them, namely: Energy Output or Input (ENERGY), Reactive Energy Output or Input (R-ENERGY), Apparent Energy (A-ENERGY), Power Factor (P-F), Power Output or Input (POW), Reactive Power (R-POW), Apparent Power (A-POW), Voltage from Line to Line (VLL), Voltage from Line to Neutral (VLN), Current by line (I), and Frequency (FQ). The accuracy of each electrical measurement for all devices used is shown in Table 1. Note that all indicated percentage values refer to the obtained percentage.

Figure 1. Equipment (SK-100/200 on the left, Nexus1250 on the right top and MP200 on the right bottom; source: Electro Industries/GaugeTech, Westbury, New York—USA) [10].

Table 1. Device precision. POW: power, R-POW: reactive power, A-POW: active power, A-ENERGY: active energy, VLN: voltage line to neuter, VLL: voltage line to line, I: current, PF: power factor, FQ: frequency.

Variable	Units	MP200	NEXUS 1252		SK-200	SK-100
		(%)	200 mili Seg (%)	1 s (%)	(%)	(%)
POW	W	0.5	0.1	0.06	0.2	0.2
ENERGY	W·h	0.5	N/A	0.04	0.2	0.2
R-POW	VARs	1.0	0.1	0.08	0.2	0.2
R-ENERGY	VAR·h	1.0	N/A	0.08	0.2	0.2
A-POW	VA	1.0	0.1	0.1	0.2	0.2
A-ENERGY	VA·h	1.0	N/A	0.08	0.2	0.2
VLN	V/KV	0.3	0.1	0.05	0.1	0.1
VLL	V/KV	0.5	0.1	0.05	0.2	0.1
I	A/KA	0.3	0.1	0.025	0.1	0.1
PF	0.5 to 1	1.0	0.1	0.08	0.2	0.2
FQ (*)	Hz	$\pm 10^{-2}$ *	3.10^{-2} *	1.10^{-2} *	$\pm 3.10^{-2}$ *	1.10^{-2} *

* Accuracy in Hz.

All mentioned measurements can be performed by the four devices used in the present work.

The four devices have additional features. Shark 100, Shark 200 and Shark MP200 incorporate V-Switch technology, which allows the operator to add new functions to the devices using programming commands at any time after its installation. In the case of the Nexus 1252 device, it is possible to add isolated input/output modules and software options for additional functions. All of them have communication capabilities (some optional) as Modbus or DNP 3.0 (Distributed Network Protocol) protocols by an RS485 port, 10/100BaseT Ethernet capabilities or IrDA port. A deep analysis of features of each device is made in [10].

2.2. The Data Description

In this paper, the next dataset, which includes measurements of variables from an electrical power supply of an edifice, has been used.

- Three variables of each phase current
- Three variables of voltage from phase to phase
- Three variables of voltage from phase to neutral
- Average power factor

Between the 27 November 2014 at 18:45 and the 31 May 2015 at 23:45, the data set was logged, with an interval of 15 min.

A building called Severo Ochoa, in honour of the Novel Prize winner, was used in this work for the dataset. The University of Oviedo (Spain) is the owner of this building, which has a total area of 8.150 m^2, distributed over two basement levels and five floors; a total of 78 employees work in it. The ITS (Information Technology Services) of the University of Oviedo are also located in this edifice. The equipment of the ITS is distributed across server rooms and scientific laboratories. This equipment has to be supplied by a good quality power network at all times. The laboratory equipment mentioned above includes electron microscopes, NMR spectrometers, X-ray diffractometers, etc. The energy consumption is 190.572 KWh per day, on average. The data set detailed here was already employed by the authors in previous research [9].

The equipment mentioned above, and the building services, incorporate devices such as UPS (Uninterruptible Power Supply), VSD (Variable Speed Drive) and inductive and capacitive loads in switching mode. These electronic circuits are nonlinear loads, and all of them can create harmonic distortion in the power line. The harmonic distortion in the distribution system is caused by the harmonic currents flowing in the electronic loads.

3. Methodology

The data set employed in this research has a total of 17,763 samples that correspond to the period of time referred to in the description of the data. A process of random data deletion was performed using this data set.

The new algorithm presented in this paper hybridizes the Self-Organized Maps Neural Networks methodology with the Mahalanobis distances. The hybrid method obtained is combined with an algorithm already presented in this journal by the authors, called AAA [9], based on Multivariate Adaptive Regression Splines. The proposed methodology is new and its performance is even better than the one referenced and presented in a previous paper when applied to the same database. This method is considered a hybrid method because it combines well-known pattern recognition and machine learning methodologies in a hybrid model that is able to impute missing data [11,12].

The performance of the proposed new methodology, in comparison with AAA and MICE, has been evaluated using the mean absolute error (MAE) and the root mean square error (RMSE). They are very common metrics in forecasting research [13,14]. The reason why, in the present research, both are employed is their complementarity. The purpose of the MAE is the measurement of the average magnitude of the error in a set of forecasts without considering their direction while the RMSE is employed for its ability to describe uniformly distributed errors [13]. A more detailed explanation including the formulas employed can be found in [9]

Let us assume that we have a dataset formed by c different variables v_1, v_2, ..., v_c that are the columns of a data matrix whose total number of rows is r. The algorithm is applied via the following steps.

3.1. Creation of a New Matrix with Missing Values from the Original Data Set

This step of the algorithm is not required when it is applied to a data set in which missing data are going to be imputed, but it is mandatory in the present research to validate the algorithm by using a complete data set.

Let A be the original matrix (rxc) of r rows and c columns. As a first step and to obtain a matrix with a certain amount of missing data, a proportion of p elements in the matrix is removed. Let B be

the new (rxc) matrix, with a proportion p of missing elements. The removal is performed completely at random; therefore, the type of imputation that is going to be tested to determine the performance of the algorithm is the one known as missing completely at random (MCAR).

3.2. Creation of the Reduced Matrix

A new matrix in which all the rows with missing data are removed is created. This new matrix is called B^{red}. Although the number of rows s ($s \leq r$) of this matrix will change depending on the matrix that is going to be imputed, in those cases like the one presented in this algorithm in which the removal of data has been performed completely at random and in a proportion p, the number of remaining rows u will be represented by the following formula:

$$u = r \times (1 - p)^c , \tag{1}$$

where:

p: proportion of missing data considered;
r: number of rows of the original matrix;
c: number of columns of the data matrix;
Afterwards the B^{red} matrix is normalized.

3.3. Determination of the Director Vectors by Means of Self-Organized Maps Neural Networks

The Self-Organized Maps (SOM) Neural Network is a type of unsupervised neural-network algorithm whose main application is related to the visualization and interpretation of large dimensional data sets [15].

These types of maps are used to represent all the available observations (data vectors), with an optimized accuracy, by means of a reduced set of models. This is the reason why this technique has been chosen in the present research.

Let N be the dimension of the n director vectors X(t) $\in R^n$, t = 1, 2, . . . , n, where each sample vector is identified by a label. The two-dimensional output layer of the SOM map contains a rectangular mesh of k = 1, . . . , $x_{dim} \times y_{dim}$ nodes. Each one of these nodes is employed as a codebook vector W_k of dimension N. The calculus of the weight vectors is performed by using the following algorithm [16].

For a certain amount of iterations, follow the steps detailed below:

1. Choose one sample vector $X(t)$ at random;
2. Search for the nearest weight vector $W_c : ||X - W_c|| = min_j||X - W_j||$;
3. Update the weights W_i by means of the following rule:

$$W_i(t+1) = W_i(t) + h_{ci}(t) \cdot [X(t) - W_i(t)] , \tag{2}$$

where $h_{ci}(t)$ is the neighbour function, which, in the case of the present research and is being very common in the literature [15], is of the Gaussian type:

$$h_{ci}(t) = \alpha(t) \cdot \exp\left(\frac{-||W_c - W_i||}{2 \cdot \sigma^2(t)}\right) . \tag{3}$$

Weight of neurons lying in the neighbourhood $h_{ci}(t)$ of the winning neuron is moved closer to $X(t)$. The learning rate $\alpha(t) \in [0, 1]$ decreases monotously as the number of iterations increases, $\sigma(t)$ determining that the radius of the neighbourhood also decreases monotonically. After many iterations and the slow reduction of $\alpha(t)$ and $\sigma(t)$, the neighbourhood covers only a single node and the map is formed. Please note that those neurons, whose weights are closer in the parameter space W, are also closer on the mesh. After this process, the director vectors obtained are denormalized. The number of

director vectors chosen to create the Self-Organized Map in the case of the present algorithm is related to the number of rows in the B^{red} matrix. Let u be the number of rows in the matrix B^{red}; the total amount of director vectors will be a range of values $d = e \cdot u \ / \ e \in [0.05,\ 0.8]$; the reason for this range of values, empirically found, will be explained in the results sections.

3.4. Finding the Closest Director Vectors by Means of Mahalanobis Distances

The Mahalanobis distance is a well-known, non-Euclidean distance measure based on correlations between variables [17]. These correlations allow for the identification and analysis of different patterns. This measure is a useful way of determining the similarity of an unknown sample set to a known one, and, in the present research, it is used to compare each one of the rows of the data matrix with missing data with all the director vectors. It can be defined by the following formula:

$$d_A\left(x_1, x_2\right) = \sqrt{\left(x_1 - x_2\right)^T \cdot A \cdot \left(x_1 - x_2\right)}, \tag{4}$$

where x_1 and x_2 represent the sets of variables of two different rows of the data matrix, and $A \ \in R^{nxn}$ is a positively semi-definite matrix that represents the inverse of the covariance matrix of class $\{I\}$. By means of the eigenvalue decomposition, A can be decomposed into $A = W \cdot W^T$.

In the case of the present algorithm, the Mahalanobis distance of each vector row with two or more missing data points to all the director vectors is calculated. Please note that, in order to make this operation possible, all those variables with missing data in the row that come from the data matrix are removed in the director vector. The director vector with the lowest Mahalanobis distance value is selected and those missing variables in this row of the data matrix are filled using the values present in the corresponding row of the director vector.

Finally, the original matrix is reconstructed and the value of the missing data of those rows with only one or two missing data points are imputed by means of the AAA algorithm. As it has already been stated, this algorithm was presented in a previous work [9] published in this journal. The referenced algorithm is based on a multivariate non-parametric technique called Multivariate Adaptive Regression Splines (MARS) [18–21].

4. Results and Discussion

In this section, the results of the Hybrid Adaptive Assignation Algorithm (HAAA) are presented and compared with those of the AAA and MICE. The test was performed using the MCAR methodology, deleting 10%, 15% and 20% of the information. This process was repeated five times. The performances of the three algorithms were compared based on the MAE and RMSE metrics. The results of all the interactions performed are presented. To simplify the comparisons, the results that use the same original MCAR subsets are presented in the same table. The way in which results are presented is the same as the one that was employed in previous research, in which the performance of the AAA algorithm was analysed [9]. Each table also contains the average values of the five replications. Table 2 contains the RMSE values of the MICE, AAA and HAAA algorithms when applied to a database with 10% of the data missing. As can be observed in this table, for the variables of voltage, intensity and power factor employed in this research, the RMSE values obtained by the new algorithm are considerably lower than those obtained by using the AAA and MICE methods. In the case of 10% missing data, Table 2, the variable in which the RMSE is reduced to a lesser amount receives a 15% reduction, while the average reduction of all variables is 62%. For the case of 15% missing data, Table 3, the results are very similar, obtaining at least a reduction of the RMSE of 12% and an average reduction of 46%. Additionally, for the case of 20% missing data, Table 4, the results are equivalent, with a minimum 18% reduction of the RMSE value and an average of 48%.

Table 2. RMSE obtained with 10% missing data using MICE, AAA and the newly proposed algorithm, HAAA. RMSE: root mean square error, MICE: multivariate imputation by chained equations, AAA: adaptive assignation algorithm, HAAA: hybrid adaptive assignation algorithm, Van: voltage line a to neuter, Vbn: voltage line b to neutre, Vcn: voltage line c to neutre, Vab: voltage line a to b, Vbc: voltage line b to c, Vca: voltage line c to a, Ia: current line a, Ib: curren line b, Ic: current line c.

RMSE MICE 10% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	20.8398	34.3633	21.9653	30.9830	36.0714	48.0547	0.2653	0.7807	1.6655	0.0030
2	18.7495	31.6307	24.1316	31.0080	37.0614	48.1173	0.7918	0.3262	1.5742	0.0017
3	18.8833	30.4312	24.0313	34.9034	30.8320	47.8822	0.1946	0.1122	1.6727	0.0027
4	17.3260	32.0759	21.2884	31.8996	31.5952	48.7585	0.9634	0.6172	1.8090	0.0030
5	20.0333	34.4660	22.1438	32.1480	32.9437	47.7413	0.7631	0.1324	1.4919	0.0028
average	19.1664	32.5934	22.7121	32.1884	33.7007	48.1108	0.5956	0.3937	1.6427	0.0026
RMSE AAA Algorithm 10% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	1.0583	1.7376	1.1078	1.6251	1.0612	1.8042	0.1318	0.1478	0.1514	0.0030
2	1.0228	1.6687	1.2186	1.4223	2.0596	1.7124	0.1307	0.1749	0.1304	0.0020
3	0.9641	1.5329	1.2044	2.0471	1.9560	1.7213	0.1365	0.1985	0.1172	0.0015
4	0.9328	1.6923	1.1531	1.8030	1.8338	1.8457	0.1845	0.1581	0.1340	0.0021
5	1.0473	1.7783	1.1100	1.7025	1.5521	1.8885	0.1368	0.1374	0.2158	0.0024
average	1.0050	1.6819	1.1588	1.7200	1.6925	1.7944	0.1441	0.1633	0.1498	0.0022
RMSE New Algorithm 10% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.6029	0.6657	0.6165	0.0384	0.0218	0.0265	0.0866	0.0632	0.0863	0.0009
2	0.5663	0.6128	0.5283	0.0270	0.0558	0.0503	0.0599	0.0896	0.0687	0.0008
3	0.5789	0.6526	0.5457	0.0690	0.0497	0.0479	0.0768	0.0652	0.0687	0.0061
4	0.5264	0.5965	0.6352	0.0344	0.0383	0.0374	0.0608	0.0624	0.0609	0.0007
5	0.5853	0.5110	0.5568	0.0603	0.0354	0.0316	0.0612	0.0745	0.0523	0.0009
average	0.5720	0.6077	0.5765	0.0458	0.0402	0.0387	0.0691	0.0710	0.0674	0.0019

Table 3. RMSE obtained with 15% missing data using MICE, AAA and the newly proposed algorithm, HAAA.

RMSE MICE 15% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	21.2798	34.9133	22.2953	36.3636	31.1154	44.9074	0.3129	1.2746	1.0503	0.0031
2	19.1895	31.9607	24.4616	37.2800	31.0974	45.0104	0.5060	1.3345	1.7062	0.0034
3	19.1033	30.7612	24.4713	37.1822	31.1052	44.6659	0.4039	1.2087	1.1608	0.0036
4	17.6560	32.4059	21.6184	37.1192	30.6655	45.9235	0.6803	1.9764	1.8724	0.0029
5	20.4733	35.1260	22.8038	37.4040	30.9877	44.5148	0.3754	1.1864	1.2194	0.0043
average	19.5404	33.0334	23.1301	37.0698	30.9942	45.0044	0.4557	1.3961	1.4018	0.0035
RMSE AAA Algorithm 15% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	1.1133	1.8036	1.1298	0.0591	0.0650	0.1153	2.0629	2.1338	2.5664	0.0009
2	1.0778	1.6907	1.2406	0.0620	0.0629	0.1176	2.2728	1.5485	2.6213	0.0009
3	1.0081	1.5549	1.2374	0.0493	0.0656	0.1014	1.8102	1.9783	2.6717	0.0012
4	0.9658	1.7583	1.2191	0.0523	0.0393	0.1501	2.2242	1.5741	2.5330	0.0012
5	1.1023	1.8333	1.1320	0.0678	0.0533	0.0929	1.7768	2.1714	2.4560	0.0010
average	1.0534	1.7281	1.1918	0.0581	0.0572	0.1155	2.0294	1.8812	2.5697	0.0011
RMSE New Algorithm 15% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.6689	0.6987	0.6715	0.0439	0.0262	0.0309	0.6936	0.6543	0.7210	0.0010
2	0.5883	0.6678	0.5943	0.0325	0.0580	0.0536	0.6579	0.7209	0.7350	0.0009
3	0.6449	0.6966	0.5787	0.0745	0.0541	0.0545	0.8358	0.7183	0.7421	0.0011
4	0.5814	0.6185	0.6792	0.0388	0.0427	0.0407	0.6739	0.6574	0.6531	0.0008
5	0.6403	0.5550	0.6008	0.0658	0.0409	0.0349	0.6336	0.7715	0.7050	0.0009
average	0.6248	0.6473	0.6249	0.0511	0.0444	0.0429	0.6990	0.7045	0.7112	0.0009

Table 4. RMSE obtained with 20% missing data using MICE, AAA and the newly proposed algorithm, HAAA.

RMSE MICE 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	23.3518	36.9853	24.0713	43.6987	33.1613	50.9935	0.5688	0.6465	1.1420	0.0941
2	21.8535	34.3287	25.3496	30.8402	36.5679	53.5682	0.6665	0.5931	1.2083	0.0906
3	19.9913	33.1292	25.9513	40.3302	27.7061	53.0747	0.5814	0.5970	1.0409	0.0550
4	20.0240	33.5899	23.3944	30.7252	35.7307	51.8370	0.5847	0.3919	1.4901	0.0843
5	22.8413	36.0140	25.4678	45.1604	28.2537	48.6319	0.6590	0.5827	1.0313	0.1027
average	21.6124	34.8094	24.8469	38.1509	32.2839	51.6210	0.6121	0.5622	1.1825	0.0853
RMSE AAA Algorithm 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	1.2317	1.4731	1.1482	2.2405	2.2522	2.8032	0.142	0.1537	0.1597	0.0058
2	1.2850	1.3387	1.2478	2.53918	1.6669	2.7101	0.1353	0.1772	0.1731	0.0038
3	1.1857	1.2352	1.2742	2.07656	2.1855	2.8789	0.1162	0.1459	0.155	0.0068
4	1.0546	1.2359	1.1375	2.4018	1.7221	2.6810	0.162	0.146	0.1259	0.0066
5	1.3687	1.4813	1.1392	1.98403	2.2898	2.6632	0.1196	0.1533	0.1304	0.0077
average	1.2251	1.3528	1.1894	2.2484	2.0233	2.7473	0.1350	0.1552	0.1488	0.0061
RMSE New Algorithm 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.8761	0.9651	0.8787	0.4692	0.4224	0.4128	0.0978	0.0929	0.1123	0.0014
2	0.7659	0.7566	0.7127	0.3311	0.5584	0.6419	0.1118	0.1012	0.1252	0.0017
3	0.9113	0.9038	0.7267	0.7379	0.5989	0.5070	0.1341	0.1157	0.1076	0.0018
4	0.7294	0.8849	0.9456	0.4763	0.4598	0.4680	0.0953	0.1044	0.1112	0.0016
5	0.7587	0.8214	0.7192	0.6213	0.5170	0.3739	0.1125	0.1186	0.1066	0.0017
average	0.8083	0.8664	0.7966	0.5272	0.5113	0.4807	0.1103	0.1066	0.1126	0.0016

The results obtained when the MAE metric is applied to the three algorithms are equivalent. Table 5 shows the results obtained using the MAE metric for 10% missing data, while Table 6 does the same for 15% and Table 7 for 20%. When the algorithm proposed is compared with AAA in the case of 10% missing data, the average of improvement regarding the MAE metric is 35%, with a minimum value of 10%. For the case of 15% missing data, the average improvement of the MAE is 29%, with a minimum of an 8% improvement in one of the variables. When the amount of missing data is 20%, the average improvement of the referenced metric is 42%, with a minimum amount of 13%.

Table 5. MAE (mean absolute error) obtained with 10% missing data using MICE, AAA and the newly proposed algorithm, HAAA.

MAE MICE 10% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	16.5255	27.4609	17.6382	31.0773	31.8032	39.9026	0.1725	0.2588	1.3358	0.0024
2	14.6229	24.1967	20.2526	33.6486	24.1749	40.8152	0.1744	0.2355	1.3055	0.0025
3	15.1412	23.9954	19.4134	26.2969	30.7039	41.4323	0.2113	0.2623	1.3377	0.0024
4	14.0678	25.6064	17.3741	34.8765	23.4809	37.3655	0.2512	0.2306	1.3785	0.0026
5	16.0730	27.5605	17.8741	27.4130	33.2976	35.7481	0.2334	0.1752	1.4786	0.0060
average	15.2861	25.7640	18.5105	30.6624	28.6921	39.0528	0.2086	0.2325	1.3672	0.0032
MAE AAA Algorithm 10% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.8120	0.8681	0.7423	0.4379	1.3656	2.5745	0.1117	0.1210	0.9243	0.0059
2	0.8358	0.8545	0.7946	1.5311	1.2375	0.8110	0.1210	0.1228	1.0954	0.0062
3	0.8274	0.8546	0.7923	1.1356	1.4964	1.3782	0.1215	0.1272	0.9566	0.0095
4	0.8653	0.8651	0.7562	1.3314	1.4677	1.3657	0.1186	0.1222	1.0425	0.0085
5	0.9052	0.8561	0.7563	1.2364	1.2115	0.8277	0.1145	0.1177	0.9595	0.0120
average	0.8491	0.8597	0.7684	1.1345	1.3557	1.3914	0.1175	0.1222	0.9957	0.0084
MAE New Algorithm 10% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.8192	0.7406	0.6883	0.3447	0.5175	1.0331	0.1047	0.1091	0.1338	0.0042
2	0.6789	0.6371	0.7580	0.5359	0.5140	1.0838	0.1163	0.0872	0.1410	0.0033
3	0.7080	0.6473	0.7381	0.4839	0.3278	1.0783	0.0923	0.1055	0.1387	0.0054
4	0.6946	0.7001	0.6387	0.5285	0.2821	0.8947	0.1224	0.0845	0.1253	0.0049
5	0.7616	0.7693	0.6695	0.1723	0.6028	1.0647	0.0926	0.1158	0.1217	0.0069
average	0.7325	0.6989	0.6985	0.4131	0.4488	1.0309	0.1057	0.1004	0.1321	0.0049

Table 6. MAE obtained with 15% missing data using MICE, AAA and the newly proposed algorithm, HAAA.

MAE MICE 15% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	17.1855	27.7909	18.2982	32.3532	40.3426	31.4073	0.7919	0.9918	2.0018	0.0149
2	15.2829	24.4167	20.6926	24.3949	41.2552	34.0886	1.0895	0.9905	2.0695	0.0141
3	15.5812	24.6554	19.6334	31.0339	41.9823	26.7369	0.3773	0.6813	2.0927	0.0171
4	14.6178	26.0464	17.8141	23.7009	37.6955	35.2065	0.5732	0.5966	1.8180	0.0189
5	16.2930	27.8905	18.5341	33.9576	35.9681	27.6330	0.4885	1.2168	2.1226	0.0205
average	15.7921	26.1600	18.9945	29.0881	39.4488	31.0144	0.6641	0.8954	2.0209	0.0171
MAE AAA Algorithm 15% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.9321	0.9516	0.8651	1.3788	1.4240	1.3866	0.0932	0.0952	0.0865	0.0063
2	0.9576	0.8613	0.9000	1.3569	1.3313	1.3687	0.0958	0.0861	0.09	0.0075
3	0.5774	0.8965	0.8652	1.3458	1.5144	1.3788	0.0577	0.0896	0.0865	0.0101
4	0.9626	0.8945	0.9513	1.3565	1.5227	1.3744	0.0963	0.0895	0.0951	0.0096
5	0.9342	0.9852	0.9806	1.3026	1.5278	1.3506	0.0934	0.0985	0.0981	0.0124
average	0.8728	0.9178	0.9125	1.3481	1.4640	1.3718	0.0873	0.0918	0.0912	0.0092
MAE New Algorithm 15% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.7185	0.5040	0.6644	0.4819	0.5835	1.0881	0.0797	0.0851	0.0842	0.0072
2	0.5999	0.4390	0.7368	0.5861	0.5800	1.1278	0.0882	0.0699	0.0881	0.0056
3	0.6245	0.4420	0.7256	0.2459	0.3938	1.1113	0.0696	0.0839	0.0881	0.0093
4	0.6131	0.4770	0.6181	0.3534	0.3371	0.9497	0.0927	0.0684	0.0789	0.0086
5	0.6783	0.5150	0.6615	0.2834	0.6578	1.1307	0.0709	0.0908	0.0771	0.0114
average	0.6469	0.4754	0.6813	0.3902	0.5104	1.0815	0.0802	0.0796	0.0833	0.0084

Table 7. MAE obtained with 20% missing data using MICE, AAA and the newly proposed algorithm HAAA.

MAE MICE 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	19.8495	29.2709	19.7782	33.5372	32.2953	41.2306	0.5143	0.2057	1.2582	0.0032
2	16.1709	26.7847	22.7646	26.1709	36.1606	43.6232	0.6040	0.2416	1.1681	0.0026
3	17.6532	26.4314	21.7054	33.1059	28.5129	42.8703	0.2775	0.1110	0.8885	0.0027
4	15.8018	28.1184	20.4781	25.7729	36.6865	38.5835	0.3310	0.1324	0.7742	0.0023
5	18.9570	29.9625	20.0141	35.7336	28.8170	36.8561	0.2886	0.1155	1.4832	0.0034
average	17.6865	28.1136	20.9481	30.8641	32.4944	40.6328	0.4031	0.1612	1.1144	0.0028
MAE AAA Algorithm 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.8996	0.9180	0.8043	1.0513	1.7261	1.4762	0.6694	0.8079	0.6984	0.0014
2	0.8944	0.8669	0.7321	1.4339	1.4942	1.5055	0.8636	0.8248	0.8678	0.0015
3	0.8441	0.8711	0.6416	1.5295	1.6559	1.6439	0.8366	0.8794	0.7254	0.0016
4	0.7184	0.8919	0.8057	1.4782	1.8988	1.4632	0.8848	0.8029	0.9249	0.0018
5	0.7525	0.8642	0.7273	0.8251	0.9897	1.4285	0.7865	0.8989	0.8785	0.0019
average	0.8218	0.8824	0.7422	1.2636	1.5529	1.5035	0.8082	0.8428	0.8190	0.0016
MAE NEW Algorithm 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.1137	0.1810	0.1239	0.6595	0.7907	1.2953	0.7436	0.7280	0.7254	0.0011
2	0.1130	0.1702	0.1327	0.7933	0.6688	1.3942	0.7912	0.6542	0.7816	0.0012
3	0.1132	0.1547	0.1472	0.4827	0.5122	1.2889	0.7063	0.7437	0.8055	0.0012
4	0.0984	0.1658	0.1326	0.4718	0.4555	1.1273	0.8911	0.6175	0.7862	0.0014
5	0.1177	0.1846	0.1300	0.5498	0.9242	1.3971	0.6435	0.8116	0.7326	0.0015
average	0.1112	0.1713	0.1333	0.5914	0.6703	1.3006	0.7551	0.7110	0.7663	0.0013

Although the overall performance of the new algorithm has already been evaluated using MCAR data, from the point of view of the authors, there are a couple of situations in which the information is not missing completely at random and are of great interest for electrical measurements. These are as follows:

- The case in which there is correlation in the missingness of data: one possible situation when working with electrical data would be when all the missing information corresponds to the same phase. In order to simulate this kind of failure, five new data sets with a 20% of missing data were created. Each phase is represented by means of four different variables: one variable of phase current, two variables of voltage from phase to phase and one variable of voltage from phase to neutral. It means that each row with missing incomplete information has four missing variables or, in other words, that only 5% of the total of rows will have missing data. In the referred rows, randomly selected, the information of the variables of one of the phases was removed. It means that, for example, when information for variable V_{an} is missing, it is also missing the information of variables, V_{ab}, V_{ca} and I_a. The results obtained are presented in Tables 9 and 10. As it can be observed, the performance of the HAAA algorithm is worse than in the MCAR case, but it outperforms both MICE and AAA.
- The case in which most of the missing data correspond to a certain subset of variables. In order to simulate this kind of failure, five new datasets with a 90% of missing data in a single variable were created. In each dataset, a proportion of 90% elements in one single column were removed, leaving the rest of the variables with their original values. As it can be seen in Table 8, the imputation accuracy for all the algorithms decreased significantly. This was expected in such an unfavourable situation; however, it is possible to ascertain, as both algorithms HAAA and AAA considerably outperform the algorithm of reference MICE, HAAA being the one with the best results.

Table 8. MAE and RMSE obtained with 90% missing data in a single column (case of missing information in Van) using MICE, AAA and the newly proposed algorithm HAAA.

ID#	90% Missing Data in One Single Column					
	MAE			RMSE		
	MICE	AAA	HAAA	MICE	AAA	HAAA
1	87.1375	10.2289	8.9960	103.8879	12.3165	8.7614
2	68.0226	10.1661	8.9438	99.2328	12.8503	8.6589
3	76.3340	10.1900	8.4414	86.4040	11.8568	9.1133
4	67.0027	8.8589	7.1839	90.5413	10.5458	7.9936
5	82.5461	10.5940	7.5248	102.5351	13.6865	9.5870
average	76.2086	10.0076	8.2180	96.5202	12.2512	8.8228

Table 9. RMSE obtained with 20% missing data using MICE, AAA and the newly proposed algorithm, HAAA for the case in which there is correlation in the missingness of data.

RMSE MICE 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	24.4908	43.2588	47.5089	65.2910	52.2601	68.8207	0.7967	1.1136	1.8646	0.1084
2	26.4326	61.1208	42.9646	57.2527	61.4969	71.0171	1.2437	0.9490	1.2511	0.1627
3	33.4717	61.9103	48.7703	55.0477	52.4416	106.0520	1.0616	0.8542	1.6592	0.0826
4	35.8859	50.2750	30.4530	54.3050	38.1316	68.8768	0.6942	0.6224	2.6681	0.1056
5	44.4516	52.6348	40.4844	90.2603	29.1994	58.1034	1.2995	0.7163	1.6545	0.1789
average	32.9465	53.8399	42.0362	64.4314	46.7059	74.5740	1.0191	0.8511	1.8195	0.1276
RMSE AAA Algorithm 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	1.9878	1.7252	1.8032	2.8163	4.1109	5.3343	0.1437	0.1631	0.2222	0.0103
2	2.4359	1.6420	2.4058	3.0857	3.1677	4.3184	0.2665	0.2600	0.2786	0.0056
3	1.3291	1.4607	2.4013	2.3697	2.7676	5.6870	0.1173	0.1481	0.2999	0.0073
4	1.8910	1.3804	1.1758	3.7502	2.4218	3.9166	0.3068	0.2228	0.1346	0.0070
5	1.3798	1.8577	1.5385	2.1441	2.9301	3.6357	0.1499	0.2220	0.1604	0.0108
average	1.8047	1.6132	1.8649	2.8332	3.0796	4.5784	0.1968	0.2032	0.2191	0.0082
RMSE New Algorithm 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	1.0966	1.6848	1.3832	0.5570	0.5171	0.5755	0.1876	0.1183	0.1214	0.0024
2	1.2825	1.1124	1.2688	0.4565	0.6296	1.1775	0.1510	0.1534	0.1903	0.0028
3	0.9783	1.2418	1.1965	0.7558	0.8273	0.8882	0.1609	0.1871	0.1724	0.0034
4	1.2532	1.1211	1.4008	0.6415	0.7131	0.5738	0.1775	0.1299	0.1208	0.0017
5	1.1888	1.4617	1.3264	1.2255	0.7568	0.4553	0.1965	0.2133	0.1078	0.0028
average	1.1599	1.3243	1.3151	0.7273	0.6888	0.7341	0.1747	0.1604	0.1425	0.0026

Table 10. MAE obtained with 20% missing data using MICE, AAA and the newly proposed algorithm HAAA for the case in which there is correlation in the missingness of data.

MAE MICE 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	38.1829	54.5223	23.8789	58.9845	38.1634	52.8209	0.8517	0.4084	2.2732	0.0062
2	24.1992	33.7322	43.3651	45.5847	65.1002	63.0445	0.8267	0.4078	2.1729	0.0051
3	27.6520	33.3112	26.5114	46.9331	43.6067	56.4079	0.3682	0.1129	1.0877	0.0053
4	22.4674	51.2556	26.8316	37.0569	45.6670	54.6107	0.4494	0.2114	1.2456	0.0029
5	23.7421	49.7850	32.9160	38.4813	55.2293	65.3112	0.3032	0.1294	2.1787	0.0066
average	27.2487	44.5212	30.7006	45.4081	49.5533	58.4390	0.5599	0.2540	1.7916	0.0052
MAE AAA Algorithm 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	1.3325	1.3119	0.9742	1.8596	2.9608	2.8876	0.9854	1.3046	0.7014	0.0022
2	1.4758	1.2017	1.4369	1.7713	2.1919	1.8424	1.4863	0.8386	1.3763	0.0023
3	0.8996	0.9740	0.7864	2.7649	3.0907	2.8334	1.4497	1.2269	1.3796	0.0029
4	0.7598	1.1519	0.8644	1.7245	3.1458	2.4398	1.1288	1.0362	1.2571	0.0028
5	1.4621	1.3658	0.7819	1.5175	1.3176	1.4345	0.8660	1.6700	1.0700	0.0023
average	1.1860	1.2011	0.9687	1.9276	2.5414	2.2876	1.1832	1.2152	1.1569	0.0025
MAE New Algorithm 20% Missing Data										
ID#	**Van**	**Vbn**	**Vcn**	**Vab**	**Vbc**	**Vca**	**Ia**	**Ib**	**Ic**	**PF**
1	0.1442	0.2838	0.1714	1.0236	1.5602	2.5453	0.9079	1.1350	1.0716	0.0015
2	0.1183	0.2924	0.1868	0.8085	0.9362	2.0889	1.5490	0.8778	1.0643	0.0021
3	0.1604	0.2698	0.2391	0.5750	0.9770	2.4446	0.9220	1.4670	0.9484	0.0023
4	0.1155	0.1715	0.2262	0.5195	0.7244	2.1240	1.2272	1.2134	1.2432	0.0022
5	0.1373	0.3592	0.1652	0.8841	1.3553	2.4234	0.9504	1.4772	0.7877	0.0023
average	0.1351	0.2754	0.1977	0.7622	1.1106	2.3252	1.1113	1.2341	1.0230	0.0021

5. Conclusions

The improvement of power quality has become a necessity as the presence of power electronics in today's grids has been increasing in the last decades. Due to this problem, network monitoring with the help of real-time data collection devices is helpful. In this context, the availability of missing data imputation techniques is required.

This research presents a new algorithm and compares it with another algorithm proposed in a previous paper by the authors and also with a well-known missing data imputation algorithm. Although the algorithm presented in this paper outperforms the others, as the previous methods to which it is compared, it also has some limitations that must be taken into account. As those proposed before, our algorithm would have imputation problems in those cases in which most of the missing data belonged to the same variable or were concentrated in a certain subset of variables instead of distributed among all the variables of the data set. Currently, the authors continue to develop hybrid algorithms that would improve the results of existing algorithms when they have to address this type of issue. Finally, the missing data imputation in the time-frequency domain will also be explored in future works.

Acknowledgments: Francisco Javier de Cos Juez and Fernando Sánchez Lasheras appreciate support from the Spanish Economics and Competitiveness Ministry, through grant AYA2014-57648-P and the Government of the Principality of Asturias (Consejería de Economía y Empleo), through grant FC-15-GRUPIN14-017.

Author Contributions: Francisco Javier de Cos Juez, Concepción Crespo Turrado and Fernando Sánchez Lasheras conceived the study. Andrés José Piñón Pazos and José Luis Calvo Rollé programmed the required algorithms. Fernando Sánchez Lasheras, Manuel G. Melero and Francisco Javier de Cos Juez interpreted the results and drafted the manuscript; Concepción Crespo Turrado, Andrés José Piñón Pazos, Manuel G. Melero and José Luis Calvo Rollé supervised the experimental data analysis; and they also contributed to the critical revision and improvement of the paper. All of the authors have approved the final version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chattopadhyay, S.; Mitra, M.; Sengupta, S. Electric power quality. In *Electric Power Quality*; Springer: Dordrecth, The Netherlands, 2011; pp. 5–12.
2. Dixit, J.B.; Yadav, A. *Electrical Power Quality*; University Science Press: New Delhi, India, 2010.
3. Stones, J.; Collinson, A. Power quality. *Power Eng. J.* **2001**, *15*. [CrossRef]
4. Ferreira, D.D.; de Seixas, J.M.; Cerqueira, A.S.; Duque, C.A.; Bollen, M.H.J.; Ribeiro, P.F. A new power quality deviation index based on principal curves. *Electr. Power Syst. Res.* **2015**, *125*, 8–14. [CrossRef]
5. Mahela, O.P.; Shaik, A.G.; Gupta, N. A critical review of detection and classification of power quality events. *Renew. Sustain. Energy Rev.* **2015**, *41*. [CrossRef]
6. Granados-Lieberman, D.; Valtierra-Rodriguez, M.; Morales-Hernandez, L.; Romero-Troncoso, R.; Osornio-Rios, R. A hilbert transform-based smart sensor for detection, classification, and quantification of power quality disturbances. *Sensors* **2013**, *13*, 5507–5527. [CrossRef] [PubMed]
7. Granados-Lieberman, D.; Romero-Troncoso, R.J.; Cabal-Yepez, E.; Osornio-Rios, R.A.; Franco-Gasca, L.A. A real-time smart sensor for high-resolution frequency estimation in power systems. *Sensors* **2009**, *9*, 7412–7429. [CrossRef] [PubMed]
8. Lim, Y.; Kim, H.-M.; Kang, S. A design of wireless sensor networks for a power quality monitoring system. *Sensors* **2010**, *10*, 9712–9725. [CrossRef] [PubMed]
9. Crespo Turrado, C.; Sánchez Lasheras, F.; Calvo-Rolle, J.L.; Piñón-Pazos, A.J.; de Cos Juez, F.J. A new missing data imputation algorithm applied to electrical data loggers. *Sensors* **2015**, *15*, 31069–31082. [CrossRef] [PubMed]
10. Electro Industries, Gauge Tech. Index of Products. Available online: http://www.electroind.com/products/ (accessed on 13 November 2015).
11. Sánchez Lasheras, F.; Nieto, P.; de Cos Juez, F.; Bayón, R.; Suárez, V. A hybrid PCA-CART-MARS-based prognostic approach of the remaining useful life for aircraft engines. *Sensors* **2015**, *15*, 7062–7083. [CrossRef] [PubMed]
12. De Andres, J.; Lorca, P.; Sánchez-Lasheras, F.; de Cos Juez, F.J. Bankruptcy prediction and credit scoring: A review of recent developments based on hybrid systems and some related patents. *Rec. Pat. Comp. Sci.* **2012**, *5*, 11–20. [CrossRef]
13. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
14. Turrado, C.; López, M.; Sánchez Lasheras, F.; Gómez, B.; Rollé, J.; de Cos Juez, F.J. Missing data imputation of solar radiation data under different atmospheric conditions. *Sensors* **2014**, *14*, 20382–20399. [CrossRef] [PubMed]
15. Kohonen, T. *Self-Organizing Maps*, 1st ed.; Springer: Berlin, Germany; Heidelberg, Germany, 1995.
16. De Andres, J.; Sánchez-Lasheras, F.; Lorca, P.; de Cos Juez, F.J. A hybrid device of self organizing maps (SOM) and multivariate adaptive regression splines (MARS) for the forecasting of firms' bankruptcy. *J. Account. Manag. Inf. Syst.* **2011**, *10*, 351–374.
17. Sánchez-Lasheras, F.; de Andrés, J.; Lorca, P.; de Cos Juez, F.J. A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy. *Expert Syst. Appl.* **2012**, *39*, 7512–7523. [CrossRef]
18. García Nieto, P.J.; Alonso Fernández, J.R.; Sánchez Lasheras, F.; de Cos Juez, F.J.; Díaz Muñiz, C. A new improved study of cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain) using the MARS technique. *Sci. Total Environ.* **2012**, *430*, 88–92. [CrossRef] [PubMed]
19. De Cos Juez, F.J.; Sánchez Lasheras, F.; García Nieto, P.J.; Suarez, M.A.S. A new data mining methodology applied to the modelling of the influence of diet and lifestyle on the value of bone mineral density in post-menopausal women. *Int. J. Comput. Math.* **2009**, *86*, 1878–1887. [CrossRef]

20. Suárez Sánchez, A.; Riesgo Fernández, P.; Sánchez Lasheras, F.; de Cos Juez, F.J.; García Nieto, P.J. Prediction of work-related accidents according to working conditions using support vector machines. *Appl. Math. Comput.* **2011**, *218*, 3539–3552. [CrossRef]
21. De Cos Juez, F.J.; Sánchez Lasheras, F.; Roqueñí, N.; Osborn, J. An ANN-based smart tomographic reconstructor in a dynamic environment. *Sensors* **2012**, *12*, 8895–8911. [CrossRef] [PubMed]

© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).

10. FACTOR DE IMPACTO DE LAS PUBLICACIONES

Los artículos que se refieren a continuación fueron publicados, en distintos años, en la revista *SENSORS*. Esta revista se encuentra indexada en el "Journal Citation Report" (JCR) dentro de la categoría "Instruments & Instrumentation".

1. Concepción Crespo Turrado, María del Carmen Meizoso López, Fernando Sánchez Lasheras, Benigno Antonio Rodríguez Gómez, José Luis Calvo Rollé, Francisco Javier de Cos Juez. **Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions.** *Sensors* **2014**, *14*(11), 20382-20399.doi:10.3390/s141120382.

   En el año de publicación de este artículo (2014), el factor de impacto de la revista fue de 2.245, lo que la posicionó de la siguiente forma dentro de la categoría citada: cuartil Q1, posición 10 de un total de 56 revistas.

2. Concepción Crespo Turrado, Fernando Sánchez Lasheras, José Luis Calvo-Rollé, Andrés José Piñón-Pazos, Francisco Javier de Cos Juez. **A New Missing Data Imputation Algorithm Applied to Electrical Data Loggers.** *Sensors* **2015**, *15*(12), 31069-31082.doi:10.3390/s151229842.

   En el año de publicación de este artículo (2015), el factor de impacto de la revista fue de 2.033, lo que la posicionó de la siguiente forma dentro de la categoría citada: cuartil Q1, posición 12 de un total de 56 revistas.

3. Concepción Crespo Turrado, Fernando Sánchez Lasheras, José Luis Calvo-Rollé, Andrés-José Piñón-Pazos, Manuel G. Melero, Francisco Javier de Cos Juez. **A Hybrid Algorithm for Missing Data Imputation and Its Application to Electrical Data Loggers.** *Sensors* **2016**, *16*(9), 1467. doi:10.3390/s16091467.

   En el año de publicación de este artículo (2016), el factor de impacto de la revista fue de 2.677, lo que la posicionó de la siguiente forma dentro de la categoría citada: cuartil Q1, posición 10 de un total de 58 revistas.

11. BIBLIOGRAFÍA

[Acock] A.C. Acock, "Working with missing values," *Journal of Marriage and Family*, vol. 67, no. 4, pp. 1012-1028. Noviembre 2005.

[Allison] P.D. Allison, *Missing data*, SAGE Publications. 2009.

[Alves 1] A.P. Alves da Silva, V.H. Quintana, G.K.H. Pang, "A pattern analysis approach for topology determination, bad data correction and missing measurement estimation in power systems," *Proc. of the Twenty-Second Annual North American Power Symposium*, pp. 363-372. 1990.

[Alves 2] A.P. Alves da Silva, V.H. Quintana, G.K.H. Pang, "A pattern recognition approach for data estimation and debugging in power systems," *Canadian Journal of Electrical and Computer Engineering*, vol. 18, no. 1, pp. 1-9. 1993.

[Alves 3] A.P. Alves da Silva, V.H. Quintana, "Pattern analysis in power system state estimation," *Electrical Power & Energy Systems*, vol. 17, no. 1, pp. 51-60. 1995.

[Black] A.C. Black, O. Harel, D.B. McCoach, "Missing data techniques for multilevel data: implications of model misspecification," vol. 38, no. 9, pp. 1845-1865. 2011.

[Blair] S.M. Blair, C.D. Booth, G. Williamson, A. Poralis, V. Turnham, "Automatically detecting and correcting errors in power quality monitoring data," *IEEE Transactions on Power Delivery*, vol. 32, no. 2, pp. 1005-1013. Abril 2017.

[Bur] A. Bur, A.G. Expósito, "Observability and bad data identification when using ampere measurements in state estimation," *Proc. 1993 IEEE International Symposium on Circuits and Systems*, vol. 4, pp. 2668-2671, 1993.

[Buuren 1] S. van Buuren, *Flexible imputation of missing data*, Chapman & Hall/CRC, Londres. 2012.

[Buuren 2] S. van Buuren, K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1-67. Diciembre 2011.

[Chen 1] J. Chen, A. Abur, "Improved bad data processing via strategic placement of PMUs," *Proc. IEEE Power Engineering Society General Meeting*, 2005, vol. 1, pp. 509-513. 2005.

[Chen 2] J. Chen, A. Abur, "Placement of PMUs to enable bad data detection in state estimation," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1608-1615. Noviembre 2006.

[Chen 3] J. Chen, A. Lau, J. Cao, K. Wang, "Automated load curve data cleansing in power systems," *IEEE Transactions on Smart Grid*, vol. 1, no. 2, pp. 213-221. Septiembre 2010.

[Coutto 1] M.B. do Coutto Filho, J. C. Stacchini de Souza, "Forecasting-aided state estimation—Part I: Panorama," *IEEE Transactions on Power Systems*, vol. 24, no. 4, pp. 1667-1677. Noviembre 2009.

[Coutto 2] M.B. do Coutto Filho, J.C. Stacchini de Souza, M.A. Ribeiro Guimaraens, "Enhanced bad data processing by phasor-aided state estimation," *IEEE Transactions on Power Systems*, vol. 29, no. 5, pp. 2200-2209. Septiembre 2014.

[Crespo 1] C. Crespo, M. C. Meizoso, F. Sánchez, B. A. Rodríguez, J. L. Calvo, F. J. de Cos, "Missing data imputation of solar radiation data under different atmospheric conditions," *Sensors*, vol. 14, no. 11, pp. 20382-20399. Noviembre 2014.

[Crespo 2] C. Crespo, F. Sánchez, J. L. Calvo-Rollé, A. J. Piñón-Pazos, F. J. de Cos, "A new missing data imputation algorithm applied to electrical data loggers," *Sensors*, vol. 15, no. 12, pp. 31069-31082. Diciembre 2015.

[Crespo 3] C. Crespo-Turrado, J. L. Casteleiro-Roca, F. Sánchez-Lasheras, J. A. López-Vázquez, F. J. de Cos, J. L. Calvo-Rolle, E. Corchado, "Student performance prediction applying missing data imputation in electrical engineering studies degree," Chapter: Hybrid Artificial Intelligent Systems, vol. 9648, *Lecture Notes in Computer Science*, pp. 126-135.

[Crespo 4] C. Crespo, F. Sánchez, J. L. Calvo-Rollé, A.-J. Piñón-Pazos, M. G. Melero, F. J. de Cos, "A hybrid algorithm for missing data imputation and its application to electrical data loggers," *Sensors*, vol. 16, no. 9, 1467. Septiembre 2016.

[Doraiswami] R. Doraiswami, J. L. R. Pereira, "A new bad data detection & identification algorithm," *IEEE Transactions on Reliability*, vol. R-29, no. 4, pp. 333-335. Octubre 1980.

[EEI 1] Edison Electric Institute, Association of Edison Illuminating Companies, Utilities Telecom Council, *Smart meters and smart meter systems: a metering industry perspective*, EEI. 2011.

[EEI 2] Edison Electric Institute, *Uniform business practices for unbundled electricity metering*. Volume two, EEI. Diciembre 2000.

[Enders] C.K. Enders, *Applied missing data analysis*. ISBN 9781606236390, Guilford Press, Nueva York. 2010.

[Friedman] J.H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1-67. 1991.

[García] A. Garcia, A. Monticelli, P. Abreu, "Fast decoupled state estimation and bad data processing," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-98, no. 5, pp. 1645-1652. Septiembre/octubre 1979.

[Ghiocel] S.C. Ghiocel, J.H. Chow, G. Stefopoulus, B. Fardanesh, D. Maragal, B. Blanchard, M. Razanousky, D.B. Bertagnolli, "Phasor-measurement-based state estimation

for synchrophasor data quality improvement and power transfer interface monitoring," *IEEE Transactions on Power Systems*, vol. 29, no. 2, pp. 881-888. Marzo 2014.

[Graham] J.W. Graham, D.L. Coffman, "Structural equation modeling with missing data," en R.H. Hoyle (Editor), *Handbook of structural equation modeling*, pp. 277-295. ISBN 9781462516797, Guilford Press, Nueva York. 2012.

[Handschin] E. Handschin, F.C. Schweppe, J. Kohlas, A. Fiechter, "Bad data analysis for power system state estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-94, no. 2, pp. 329-337. Marzo/abril 1975.

[Harvey] P.R. Harvey, B. Stephen, S. Galloway, "Classification of AMI residential load profiles in the presence of missing data," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 1944-1945. Julio 2016.

[Hoverstad] B.A. Hoverstad, A. Tidemann, H. Langseth, "Effects of data cleansing on load prediction algorithms," *Proc. 2013 IEEE Computational Intelligence Applications in Smart Grid*, pp. 93-100. 2013.

[Huang 1] Y. Huang, S. Werner, J. Huang, N. Kashyap, V. Gupta, "State estimation in electric power grids," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 33-43. Septiembre 2012.

[Huang 2] S.-J. Huang, J. Lin, "Enhancement of anomalous data mining in power system predicting-aided state estimation," *IEEE Transactions on Power Systems*, vol. 19, no. 1, pp. 610-619. Febrero 2004.

[IEA 1] International Energy Agency, How 2 guide for smart grids in distribution networks. Roadmap development and implementation, OECD/IEA. 2015.

[IEA 2] International Energy Agency, Energy technology roadmaps a guide to development and implementation, IEA. 2014.

[IEA 3] International Energy Agency, *Technology roadmap smart grids*, OECD/IEA. 2011.

[Halatchev] M. Halatchev, L. Gruenwald, "Estimating missing values in related sensor data streams," *Advances in Data Management 2005, Proc. of the Eleventh International Conference on Management of Data*, pp. 83-94, 2005.

[IEEE Std. 1159] IEEE Std. 1159-2009: IEEE Recommended practice for monitoring electric power quality, IEEE Power & Energy Society. Diciembre 2000.

[Jha] I.S. Jha, S. Sen, V. Agarwal, "Advanced metering infrastructure analytics - A case study," *Proc. 2014 Eighteenth National Power Systems Conference*, pp. 1-6. 2014.

[Jiang 1] N. Jiang, "A Data Imputation Model in Sensor Databases," High Performance Computing and Communications, Proc. of the Third International Conference, pp. 86-96. 2007.

[Jiang 2] N. Jiang, Z. Chen, "Model-driven data cleaning for signal processing system in sensor networks," *Proc. 2010 2nd International Conference on Signal Processing Systems*, vol. 1, pp. 237-242. 2010.

[Jones] K.D. Jones, A. Pal, J.S. Thorp, "Methodology for performing synchrophasor data conditioning and validation," *IEEE Transactions on Power Systems*, vol. 30, no. 3, pp. 1121-1130. Mayo 2015.

[Kotiuga] W.W. Kotiuga, M. Vidyasagar, "Bad data rejection properties of weighted least absolute value techniques applied to static state estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-101, no. 4, pp. 844-853. Abril 1982.

[Lankutis] J.D. Lankutis, "Verifying data integrity. If you cannot believe the data, how can you believe the analytics?," *Proc. 2013 IEEE Rural Electric Power Conference*, pp. C4-1 – C4-4. 2013.

[Leite] A.M. Leite da Silva, M.B. do Coutto Filho, J.F. de Queiroz, "State forecasting in electric power systems," *IEE Proceedings*, vol. 130, no. 5, pp. 237-244. Septiembre 1983.

[Lin 1] J. Lin, S. Huang, K. Shih, "Application of sliding surface-enhanced fuzzy control for dynamic state estimation of a power system," *IEEE Transactions on Power Systems*, vol. 18, no. 2, pp. 570-577. Mayo 2003.

[Lin 2] S. Lin, C. Xie, B. Tang, R. Liu, A. Pan, "The data mining application in the power quality monitoring data analysis," *Proc. 2016 IEEE 11th Conference on Industrial Electronics and Applications*, pp. 338-342. 2016.

[Little] R.J.A. Little, D.B. Rubin, *Statistical analysis with missing data*. ISBN: 9780471183860, John Wiley and Sons, Nueva York, 2002.

[Lo 1] K.L. Lo, P.S. Ong, R.D. McColl, A.M. Moffatt, J.L. Sulley, "Development of a static state estimator part I: Estimation and bad data suppression," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-102, no. 8, pp. 2486-2491. Agosto 1983.

[Lo 2] K.L. Lo, P.S. Ong, R.D. McColl, A.M. Moffatt, J.L. Sulley, "Development of a static state estimator part II: Bad data replacement and generation of pseudomeasurements," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-102, no. 8, pp. 2492-2500. Agosto 1983.

[Lu] N. Lu, P. Du, X. Guo, F.L. Greitzer, "Smart meter data analysis," *Proc. 2012 IEEE/PES Transmission & Distribution Conference and Exposition*, pp. 1-6, 2012.

[Majidpour] M. Majidpour, P. Chu, R. Gadh, H.R. Pota, "Incomplete data in smart grid: Treatment of missing values in electric vehicle charging data," *Proc. 2014 International Conference on Connected Vehicles and Expo*, pp. 1041-1042. 2014.

[Marwala] T. Marwala, Computational intelligence form missing data imputation, estimation and management: Knowledge optimization techniques. ISBN: 9781605663364, Universidad de Witwatersrand, Sudáfrica. 2009.

[Mateos 1] G. Mateos, G.B. Giannakis, "Robust nonparametric regression via sparsity control with application to load curve data cleansing," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp.1571-1584. Abril 2012.

[Mateos 2] G. Mateos, G.B. Giannakis, "Load curve data cleansing and imputation via sparsity and low rank," *IEEE Transactions on Smart Grid*, vol. 4, no. 4, pp. 2347-2355. Diciembre 2013.

[Medina] F. Medina, M. Galván, *Imputación de datos: teoría y práctica*. ISBN: 9789213231012, CEPAL. 2007.

[Merrill] H.M. Merrill, F. C. Schweppe, "Bad data suppression in power system static state estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-90, no. 6, pp. 2718-2725. 1971.

[Mili] L. Mili, Th. Van Cutsem, M. Ribbens-Pavella, "Bad data identification methods in power system state estimation – A comparative study," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-104, no. 11, pp. 3037-3049. Noviembre 1985.

[Monticelli] A. Monticelli, "Electric power system state estimation," *Proc. of the IEEE*, vol. 88, no. 2, pp. 262-282. Febrero 2000.

[Newman] D.A. Newman, "Missing data: Five practical guidelines," *Organizational Research Methods*, vol. 17, no. 4, pp. 372-411. Septiembre 2014.

[Peppanen 1] J. Peppanen, J. Grimaldo, M. J. Reno, S. Grijalva, R. G. Harley, "Increasing distribution system model accuracy with extensive deployment of smart meters," *Proc. 2014 IEEE Power and Energy Society General Meeting*, pp. 1-5. 2014.

[Peppanen 2] J. Peppanen, M.J. Reno, M. Thakkar, S. Grijalva, R.G. Harley, "Leveraging AMI data for distribution system model calibration and situational awareness," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 2050-2059. Julio 2015.

[Peppanen 3] J. Peppanen, X. Zhang, S. Grijalva, M.J. Reno, "Handling bad or missing smart meter data through advanced data imputation," *Proc. 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference*, pp. 1-5. 2016.

[Quilumba 1] F.L. Quilumba, W. Lee, H. Huang, D.Y. Wang, R. Szabados, "An overview of AMI data preprocessing to enhance the performance of load forecasting," *Proc. 2014 IEEE Industry Application Society Annual Meeting*, pp. 1-7. 2014.

[Quilumba 2] F.L. Quilumba, W. Lee, H. Huang, D.Y. Wang, R.L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911-918. Marzo 2015.

[Quintana] V.H. Quintana, A. Simoes-Costa, M. Mier, "Bad data detection and identification techniques using estimation orthogonal methods," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-101, no. 9, pp. 3356-3364. Septiembre 1982.

[Reikard] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342-349. Marzo 2009.

[Rey] P. Rey del Castillo, Modelo para el tratamiento de conjuntos complejos con datos ausentes de variables categóricas en un contexto de e-democracia. Aplicación a encuestas de opinión. Tesis Doctoral, Madrid 2012.

[Roberts] G.O. Roberts, "Markov chain concepts related to sampling algorithms," en W.R. Gilks, S. Richardson, D.J. Spiegelhalter (Editores) *Markov chain Monte Carlo in practice*, pp. 45–47. ISBN: 9780412055515, Chapman and Hall, Londres. 1996.

[Rubin] D.B. Rubin, *Multiple imputation for nonresponse in surveys*. ISBN: 9780471087052, John Wiley and Sons, Nueva York, 1987.

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

[Schweppe 1] F.C. Schweppe, J. Wildes, "Power system static-state estimation, part I: Exact model," *Transactions on Power Apparatus and Systems*, vol. PAS-89, no. 1, pp. 120-125. Enero 1970.

[Schweppe 2] F.C. Schweppe, D. B. Rom, "Power system static-state estimation, part II: Approximate model," *Transactions on Power Apparatus and Systems*, vol. PAS-89, no. 1, pp. 125-130. Enero 1970.

[Schweppe 3] F.C. Schweppe, "Power system static-state estimation, part III: Implementation," *Transactions on Power Apparatus and Systems*, vol. PAS-89, no. 1, pp. 130-135. Enero 1970.

[Shi] D. Shi, D.J. Tylavsky, N. Logic, "An adaptive method for detection and correction of errors in PMU measurements," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1575-1583. Diciembre 2012.

[Soltani] N.Y. Soltani, G.B. Giannakis, "Online learning of electric vehicle consumers' charging behavior with missing data," *Proc. 2014 IEEE Global Conference on Signal and Information Processing*, pp. 243-247. 2014.

[Tang] G. Tang, K. Wu, J. Lei, Z. Bi, J. Tang, "From landscape to portrait: A new approach for outlier detection in load curve data," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1764-1773. Julio 2014.

[Terzija] V. Terzija, V. Stanojevic, "Power quality indicators estimation using robust Newton-type algorithm," *IEE Proceedings - Generation, Transmission and Distribution*, vol. 151, no. 4, pp. 477-485. Julio 2004.

[Valverde] G. Valverde, V. Terzija, "Unscented Kalman filter for power system dynamic state estimation," *IET Generation, Transmission & Distribution*, vol. 5, no. 1, pp. 29-37. 2011.

[Yang] Z. Yang, J. Cao, Y. Xu, H. Zhang, P. Yu, S. Yao, "Data Cleaning for Power Quality Monitoring," *Proc. 2013 Fourth International Conference on Networking and Distributed Computing*, pp. 111-115. Diciembre 2013.

[Zarco] P. J. Zarco, A. Gómez-Expósito, *Estimación de estado y de parámetros en redes eléctricas*, Dpto. de Ingeniería Eléctrica - Universidad de Sevilla. 1999.

[Zhu] J. Zhu, A. Bur "Bad data identification when using phasor measurements," *Proc. 2007 IEEE Lausanne Power Tech*, pp. 1676-1681. 2007.

ANEXO 1:

TABLAS DE RESULTADOS

Estación	MICE (%)	MLR (%)	IDW (%)
Castrelo	12.74	26.80	29.13
Simes	12.27	25.55	31.55
A Armenteira	13.64	28.40	37.33
Pé Redondo	13.56	28.38	29.67
A Lanzada	12.67	27.47	30.41
Castrove	15.86	33.81	34.34
Sanxenxo	12.64	26.98	33.25
Corón	14.08	29.78	32.69
Torrequintáns	12.86	26.52	26.73

Tabla 1: Resultados de los tres métodos de imputación en términos de RMSE para las estaciones gallegas

Estación	MICE (%)	MLR (%)	IDW (%)
Castrelo	11.03	28.40	22.89
Simes	10.59	22.73	35.81
A Armenteira	12.53	31.91	46.51
Pé Redondo	16.97	41.73	37.53
A Lanzada	12.2	35.70	24.31
Castrove	21.11	48.30	50.29
Sanxenxo	14.75	32.93	32.54
Corón	14.71	45.63	29.12
Torrequintáns	13.36	30.36	28.69

Tabla 2: Resultados de los tres métodos de imputación en términos de MAE para las estaciones gallegas

RMSE MICE 15% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	17.0837	28.8975	23.1269	38.2302	26.7259	45.7781	0.3816	0.2352	1.8578	0.0031
2	19.7831	31.6292	21.6406	44.4176	31.1867	50.5978	0.2733	0.4289	1.6515	0.0030
3	16.8887	32.1573	23.2565	34.8709	36.4404	49.8771	2.0080	0.3768	0.4198	0.0032
4	18.9432	30.8065	21.1655	43.2729	32.0558	43.2723	0.3458	0.3326	1.7407	0.0028
5	19.0647	30.0262	23.5861	32.4402	28.9609	44.9738	0.5376	0.2517	1.8402	0.0034
PRO MED	18.3527	30.7033	22.5551	38.6463	31.0739	46.8998	0.7092	0.3251	1.5020	0.0031

Tabla 3: RMSE obtenido con un 15% de datos faltantes aplicando MICE

RMSE NUEVO ALGORITMO AAA CON 15% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	1.0916	1.0625	0.9937	1.6562	1.7851	1.7874	0.1355	0.1314	0.1184	0.0021
2	1.1417	1.0061	0.9990	1.6799	1.9229	1.7843	0.1285	0.1446	0.1235	0.0019
3	1.1178	1.1311	1.0816	1.7174	2.3560	1.7111	0.1345	0.1460	0.1249	0.0021
4	1.5109	1.0043	1.1689	1.5117	1.9786	1.8057	0.1250	0.1394	0.1262	0.0018
5	1.1151	1.0109	1.0351	1.6381	2.5543	1.8637	0.1290	0.1364	0.1324	0.0019
PRO MED	1.1954	1.0430	1.0556	1.6406	2.1194	1.7904	0.1305	0.1396	0.1250	0.0020

Tabla 4: RMSE obtenido con un 15% de datos faltantes aplicando el nuevo paquete propuesto AAA

MAE MICE 15% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	12.782	21.912	18.684	30.793	21.532	35.7718	0.2421	0.1845	1.4006	0.0024
2	15.683	24.380	17.133	34.655	24.397	40.5020	0.1959	0.2489	1.3048	0.0023
3	12.983	25.111	18.394	28.717	28.772	38.0755	1.5148	0.2157	0.2240	0.0025
4	14.373	23.899	17.465	34.675	25.777	33.3585	0.2164	0.2117	1.3023	0.0022
5	14.916	23.672	18.203	25.813	23.063	35.0113	0.2675	0.1877	1.4250	0.0026
PRO MED	14.147	23.795	17.976	30.931	24.708	36.5438	0.4873	0.2097	1.1314	0.0024

Tabla 5: MAE obtenido con un 15% de datos faltantes aplicando MICE

MAE ALGORITMO NUEVO CON 15% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	0.8882	0.8196	0.7866	1.3280	1.4423	1.3884	0.1091	0.1053	0.0973	0.0017
2	0.9152	0.8158	0.7646	1.3088	1.4749	1.3937	0.1035	0.1117	0.0986	0.0016
3	0.8710	0.9064	0.8255	1.3577	1.5212	1.3480	0.1067	0.1121	0.0961	0.0016
4	0.9951	0.7757	0.9129	1.2065	1.5648	1.4341	0.1002	0.1114	0.0983	0.0014
5	0.8625	0.7992	0.7959	1.2726	1.7787	1.4867	0.1029	0.1043	0.1021	0.0015
PRO MED	0.9064	0.8234	0.8171	1.2947	1.5564	1.4102	0.1045	0.1089	0.0985	0.0015

Tabla 6: MAE obtenido con un 15% de datos faltantes con el nuevo paquete propuesto AAA

RMSE										
	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	FP
p	<0.001	<0.001	0.001	<0.001	0.001	<0.001	0.044	0.001	0.001	0.001
p’	0.980	0.885	0.106	0.921	0.591	0.770	0.523	0.168	0.800	0.784

Tabla 7: Valores de *p* para la métrica RMSE para el tipo de modelo MICE versus AAA

MAE										
	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	FP
p	<0.001	<0.001	0.001	<0.001	0.001	<0.001	0.038	0.001	0.001	0.001
p’	0.990	0.786	0.113	0.887	0.655	0.643	0.686	0.315	0.796	0.424

Tabla 8: Valores de p para la métrica MAE para el tipo de modelo MICE versus AAA

RMSE MICE 20% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	23.3518	36.9853	24.0713	43.6987	33.1613	50.9935	0.5688	0.6465	1.1420	0.0941
2	21.8535	34.3287	25.3496	30.8402	36.5679	53.5682	0.6665	0.5931	1.2083	0.0906
3	19.9913	33.1292	25.9513	40.3302	27.7061	53.0747	0.5814	0.5970	1.0409	0.0550
4	20.0240	33.5899	23.3944	30.7252	35.7307	51.8370	0.5847	0.3919	1.4901	0.0843
5	22.8413	36.0140	25.4678	45.1604	28.2537	48.6319	0.6590	0.5827	1.0313	0.1027
Pr.	21.6124	34.8094	24.8469	38.1509	32.2839	51.6210	0.6121	0.5622	1.1825	0.0853

Tabla 12: RMSE para MICE cuando hay un 20% de datos faltantes

RMSE ALGORITMO AAA CON 20% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	1.2317	1.4731	1.1482	2.2405	2.2522	2.8032	0.142	0.1537	0.1597	0.0058
2	1.2850	1.3387	1.2478	2.53918	1.6669	2.7101	0.1353	0.1772	0.1731	0.0038
3	1.1857	1.2352	1.2742	2.07656	2.1855	2.8789	0.1162	0.1459	0.155	0.0068
4	1.0546	1.2359	1.1375	2.4018	1.7221	2.6810	0.162	0.146	0.1259	0.0066
5	1.3687	1.4813	1.1392	1.98403	2.2898	2.6632	0.1196	0.1533	0.1304	0.0077
Pr.	1.2251	1.3528	1.1894	2.2484	2.0233	2.7473	0.1350	0.1552	0.1488	0.0061

Tabla 9: RMSE para el algoritmo AAA cuando hay un 20% de datos faltantes

RMSE NUEVO ALGORITMO HAAA 20% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	0.8761	0.9651	0.8787	0.4692	0.4224	0.4128	0.0978	0.0929	0.1123	0.0014
2	0.7659	0.7566	0.7127	0.3311	0.5584	0.6419	0.1118	0.1012	0.1252	0.0017
3	0.9113	0.9038	0.7267	0.7379	0.5989	0.5070	0.1341	0.1157	0.1076	0.0018
4	0.7294	0.8849	0.9456	0.4763	0.4598	0.4680	0.0953	0.1044	0.1112	0.0016
5	0.7587	0.8214	0.7192	0.6213	0.5170	0.3739	0.1125	0.1186	0.1066	0.0017
Pr.	0.8083	0.8664	0.7966	0.5272	0.5113	0.4807	0.1103	0.1066	0.1126	0.0016

Tabla 10: RMSE para el nuevo algoritmo HAAA cuando hay un 20% de datos faltantes

MAE MICE 20% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	19.8495	29.2709	19.7782	33.5372	32.2953	41.2306	0.5143	0.2057	1.2582	0.0032
2	16.1709	26.7847	22.7646	26.1709	36.1606	43.6232	0.6040	0.2416	1.1681	0.0026
3	17.6532	26.4314	21.7054	33.1059	28.5129	42.8703	0.2775	0.1110	0.8885	0.0027
4	15.8018	28.1184	20.4781	25.7729	36.6865	38.5835	0.3310	0.1324	0.7742	0.0023
5	18.9570	29.9625	20.0141	35.7336	28.8170	36.8561	0.2886	0.1155	1.4832	0.0034
Pr.	17.6865	28.1136	20.9481	30.8641	32.4944	40.6328	0.4031	0.1612	1.1144	0.0028

Tabla 11: MAE para el algoritmo MICE cuando hay un 20 % de Datos Faltantes

MAE AAA ALGORITMO 20% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	0.8996	0.9180	0.8043	1.0513	1.7261	1.4762	0.6694	0.8079	0.6984	0.0014
2	0.8944	0.8669	0.7321	1.4339	1.4942	1.5055	0.8636	0.8248	0.8678	0.0015
3	0.8441	0.8711	0.6416	1.5295	1.6559	1.6439	0.8366	0.8794	0.7254	0.0016
4	0.7184	0.8919	0.8057	1.4782	1.8988	1.4632	0.8848	0.8029	0.9249	0.0018
5	0.7525	0.8642	0.7273	0.8251	0.9897	1.4285	0.7865	0.8989	0.8785	0.0019
Pr.	0.8218	0.8824	0.7422	1.2636	1.5529	1.5035	0.8082	0.8428	0.8190	0.0016

Tabla 12: MAE para el algoritmo AAA cuando hay un 20% de datos faltantes

MAE NUEVO ALGORITMO 20% DATOS FALTANTES										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	0.1137	0.1810	0.1239	0.6595	0.7907	1.2953	0.7436	0.7280	0.7254	0.0011
2	0.1130	0.1702	0.1327	0.7933	0.6688	1.3942	0.7912	0.6542	0.7816	0.0012
3	0.1132	0.1547	0.1472	0.4827	0.5122	1.2889	0.7063	0.7437	0.8055	0.0012
4	0.0984	0.1658	0.1326	0.4718	0.4555	1.1273	0.8911	0.6175	0.7862	0.0014
5	0.1177	0.1846	0.1300	0.5498	0.9242	1.3971	0.6435	0.8116	0.7326	0.0015
Pr.	0.1112	0.1713	0.1333	0.5914	0.6703	1.3006	0.7551	0.7110	0.7663	0.0013

Tabla 13: MAE para el algoritmo HAAA cuando hay un 20% de datos faltantes

RSME PARA MICE CON UN 20% DE DATOS PERDIDOS CUANDO EXISTE CORRELACIÓN										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	24.490	43.258	47.508	65.291	52.260	68.821	0.7967	1.1136	1.8646	0.1084
2	26.432	61.120	42.964	57.252	61.496	71.018	1.2437	0.9490	1.2511	0.1627
3	33.471	61.910	48.770	55.047	52.441	66.052	1.0616	0.8542	1.6592	0.0826
4	35.885	50.275	30.453	54.305	38.131	68.876	0.6942	0.6224	2.6681	0.1056
5	44.451	52.634	40.484	90.260	29.199	58.103	1.2995	0.7163	1.6545	0.1789
Pr.	32.946	53.840	42.036	64.431	46.706	74.574	1.0191	0.8511	1.8195	0.1276

Tabla 14: RSME para MICE con un 20% de datos perdidos cuando existe correlación entre los datos

RMSE ALGORITMO AAA CON 20% DATOS PERDIDOS CUANDO EXISTE CORRELACION										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	1.9878	1.7252	1.8032	2.8163	4.1109	5.3343	0.1437	0.1631	0.2222	0.0103
2	2.4359	1.6420	2.4058	3.0857	3.1677	4.3184	0.2665	0.2600	0.2786	0.0056
3	1.3291	1.4607	2.4013	2.3697	2.7676	5.6870	0.1173	0.1481	0.2999	0.0073
4	1.8910	1.3804	1.1758	3.7502	2.4218	3.9166	0.3068	0.2228	0.1346	0.0070
5	1.3798	1.8577	1.5385	2.1441	2.9301	3.6357	0.1499	0.2220	0.1604	0.0108
Promedio	1.8047	1.6132	1.8649	2.8332	3.0796	4.5784	0.1968	0.2032	0.2191	0.0082

Tabla 15: RSME para AAA con un 20% de datos perdidos cuando existe correlación entre los datos

RME NUEVO ALGORITMO HAAA CON 20% DATOS PERDIDOS CUANDO EXISTE CORRELACION										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	1.096z6	1.6848	1.3832	0.5570	0.5171	0.5755	0.1876	0.1183	0.1214	0.0024
2	1.2825	1.1124	1.2688	0.4565	0.6296	1.1775	0.1510	0.1534	0.1903	0.0028
3	0.9783	1.2418	1.1965	0.7558	0.8273	0.8882	0.1609	0.1871	0.1724	0.0034
4	1.2532	1.1211	1.4008	0.6415	0.7131	0.5738	0.1775	0.1299	0.1208	0.0017
5	1.1888	1.4617	1.3264	1.2255	0.7568	0.4553	0.1965	0.2133	0.1078	0.0028
Pr.	1.1599	1.3243	1.3151	0.7273	0.6888	0.7341	0.1747	0.1604	0.1425	0.0026

Tabla 16: RSME para el nuevo algoritmo HAAA con un 20% de datos perdidos cuando existe correlación entre los datos.

MAE DE MICE 20% DATOS PERDIDOS CUANDO EXISTE CORRELACION										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	38.1829	54.5223	23.8789	58.9845	38.1634	52.8209	0.8517	0.4084	2.2732	0.0062
2	24.1992	33.7322	43.3651	45.5847	65.1002	63.0445	0.8267	0.4078	2.1729	0.0051
3	27.6520	33.3112	26.5114	46.9331	43.6067	56.4079	0.3682	0.1129	1.0877	0.0053
4	22.4674	51.2556	26.8316	37.0569	45.6670	54.6107	0.4494	0.2114	1.2456	0.0029
5	23.7421	49.7850	32.9160	38.4813	55.2293	65.3112	0.3032	0.1294	2.1787	0.0066
Pr.	27.2487	44.5212	30.7006	45.4081	49.5533	58.4390	0.5599	0.2540	1.7916	0.0052

Tabla 17: MAE de MICE con 20% de datos perdidos cuando existe correlación entre los datos

MAE AAA ALGORITMO CON 20% DATOS PERDIDOS CUANDO EXISTE CORRELACION										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	1.3325	1.3119	0.9742	1.8596	2.9608	2.8876	0.9854	1.3046	0.7014	0.0022
2	1.4758	1.2017	1.4369	1.7713	2.1919	1.8424	1.4863	0.8386	1.3763	0.0023
3	0.8996	0.9740	0.7864	2.7649	3.0907	2.8334	1.4497	1.2269	1.3796	0.0029
4	0.7598	1.1519	0.8644	1.7245	3.1458	2.4398	1.1288	1.0362	1.2571	0.0028
5	1.4621	1.3658	0.7819	1.5175	1.3176	1.4345	0.8660	1.6700	1.0700	0.0023
Pr.	1.1860	1.2011	0.9687	1.9276	2.5414	2.2876	1.1832	1.2152	1.1569	0.0025

Tabla 18: MAE del algoritmo AAA con un 20% de datos perdidos cuando existe correlación

MAE NUEVO ALGORITMO HAAA CON 20% DATOS PERDIDOS CUANDO EXISTE CORRELACION										
It.	Van	Vbn	Vcn	Vab	Vbc	Vca	Ia	Ib	Ic	PF
1	0.1442	0.2838	0.1714	1.0236	1.5602	2.5453	0.9079	1.1350	1.0716	0.0015
2	0.1183	0.2924	0.1868	0.8085	0.9362	2.0889	1.5490	0.8778	1.0643	0.0021
3	0.1604	0.2698	0.2391	0.5750	0.9770	2.4446	0.9220	1.4670	0.9484	0.0023
4	0.1155	0.1715	0.2262	0.5195	0.7244	2.1240	1.2272	1.2134	1.2432	0.0022
5	0.1373	0.3592	0.1652	0.8841	1.3553	2.4234	0.9504	1.4772	0.7877	0.0023
Promedio	0.1351	0.2754	0.1977	0.7622	1.1106	2.3252	1.1113	1.2341	1.0230	0.0021

Tabla 19: MAE del nuevo algoritmo HAAA con 20% de datos perdidos cuando existe correlación

90% DE DATOS PERDIDOS EN UNA SOLA COLUMNA						
	MAE			RSME		
Iteración	MICE	AAA	HAAA	MICE	AAA	HAAA
1	87.1375	10.2289	8.9960	103.8879	12.3165	8.7614
2	68.0226	10.1661	8.9438	99.2328	12.8503	8.6589
3	76.3340	10.1900	8.4414	86.4040	11.8568	9.1133
4	67.0027	8.8589	7.1839	90.5413	10.5458	7.9936
5	82.5461	10.5940	7.5248	102.5351	13.6865	9.5870
Promedio	76.2086	10.0076	8.2180	96.5202	12.2512	8.8228

Tabla 20: MAE y RMSE obtenido para un 90% de DATOS PERDIDOS en una sola columna (información perdida en Van) usando MICE, AAA y el nuevo algoritmo propuesto HAAA

ANEXO 2:

ESQUEMA UNIFILAR DEL C.G.B.T. DEL EDIFICIO SEVERO OCHOA

(según memoria técnica Nº 44/F/2013 suscrita por Francisco Uría Nieto, Ingeniero Técnico Industrial, con motivo de la reforma del C.G.B.T llevada a cabo en octubre de 2013)

Applus
Applus Norcontrol, S.L.U.
ORGANISMO DE CONTROL
AUTORIZADO N° [illegible]
08/10/13

300 KVA
380V-50Hz
HIMOINSA
TDR HIW 210

4(1x240)+1x120
RZ1-0,6/1KV

CUADRO CONTROL
GRUPO ELECTROGENO
(SG+MPA)

4(1x240)+1x120
RZ1-0,6/1KV-LP_160x200

80KVAr

20KV/380V
1250 KVA
Dy11
Vcc=6%
UNE 20138-82

3[3(1x240)]+3(1x120)+T240
RZ1-0,6/1KV-LP_1100x300

C.G.B.T. (RED/GRUPO/SAI)

4x630A

4x630A

4x1600

C.G.B.T.
GRUPO

4x630A

C.G.B.T.
RED

CVM

4x40A 4x100A 4x125A 4x25A 4x160 4x63 2x25 2x25 2x25 2x25 2x25 2x25 2x25 DIF 4x100 0.3 REG. 4x63 2x25 2x25 4x25 4x25 4x63

4x63 300 R (x16) DIF

4x63A 0.3 REG DIF

4x25 4x20 4x25 4x80 4x63 4x50 4x40

4x40A 4x100A 4x125A 4x25A 2x25 300 SEL 4x63 300 4x25 300

4x25 300 4x25 300 4x25 300 4x80 300 4x63 300 4x80 300 4x40 300

SAI
10 KVAS

G.PRESION
INCENDIOS
15KW

C.S.SEMISOTANO 2KW | C.S.PL.PRIMERA 2KW | C.S.PL.TERCERA 2KW | ASCENSOR 20KW | C.S.PL.BAJOCUB. 3.3KW | RESERVA

C.S.SOTANO 2KW | C.S.BAJA 2KW | C.S.PL.SEGUNDA 2KW | C.S.PARKING 48.5KW | C.S.C.T. 3.5KW | Z. COM. 8.5KW | CASETA COMPRESOR

C.S.SOTANO ZONA-1 28.6KW | C.S.SEMISOTANO ZONA-1 27.5KW | C.S.PL.BAJA ZONA-1 16KW | C.S.PL. 1ª ZONA-1 28.2KW | C.S.PL.2ª ZONA-1 25.9KW | C.S.PL. 3ª ZONA-1 22.6KW | C.S.PARKING 27.6KW | C.S.INSTALACIONES 280KW | C.S.MONTACARGAS 22KW | BOMBAS ACHIQUE | A.A. SAIS | GRUPO PRESION BIE AZOTEAR

C.S.SOTANO ZONA-2 28.7KW | C.S.SEMISOTANO ZONA-2 28.5KW | C.S.PL.BAJA ZONA-2 28.9KW | C.S.PL. 1ª ZONA-2 25.9KW | C.S.PL. 2ª ZONA-2 24.2KW | C.S.PL.3ª ZONA-2 20.8KW | BATERIA CONDENSADORES 300KVAR | C.S.EXTERIORES 8.7KW | G.PRESION FONTANERIA 4KW | DEPOSITO NITROGENO | A.A. CPD

3(1x50)+1x25+T25
RZ1-0,6/1KV

SAI
87 KVA

SAI
INFORMATICA

37.8 KVA

3(1x50)+1x25+T25
RZ1-0,6/1KV

4x125 0.3 reg.

4x63 0.3 reg.

C.G.B.T.
SAI

4x25 (x10) | 4x25A (x3)

C.S.SEMISOTANO ZONA-1 9KW | C.S.PL.BAJA ZONA-1 8KW | C.S.PL.1ª ZONA-1 10.5KW | C.S.PL.2ª ZONA-2 8.4KW | C.S.PL.SOTANO ZONA-1 9KW | CENTRALITA CONTROL 1000W | CENTRALITA INCENDIOS 1000W | RESERVA EQUIPADA

C.S.SEMISOTANO ZONA-2 10.5KW | C.S.PL.BAJA ZONA-2 10.2KW | C.S.PL.1ª ZONA-2 9.6KW | C.S.PL.SOTANO ZONA-2 10.5KW

C.S.PL.2ª ZONA-1 14.4KW | C.S.PL.TERCERA ZONA-1 12.5KW

C.S.PL.TERCERA ZONA-2 10.5KW