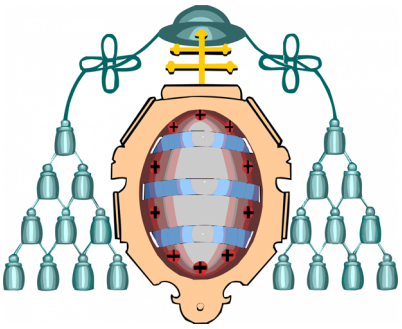


Universidad de Oviedo

Departamento de Ingeniería Eléctrica, Electrónica, de Computadores y Sistemas

Programa de Doctorado: Control de Procesos, Electrónica Industrial e Ingeniería Eléctrica

TESIS DOCTORAL

Supervisión de la energía eléctrica en edificios públicos de uso docente basada en técnicas de minería de datos visual

Serafín Alonso Castro

Enero 2012

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: SUPERVISION DE LA ENERGÍA ELÉCTRICA EN EDIFICIOS PÚBLICOS DE USO DOCENTE BASADA EN TÉCNICAS DE MINERÍA DE DATOS VISUAL	Inglés: ELECTRICAL POWER MONITORING BASED ON VISUAL DATA MINING TECHNIQUES IN PUBLIC BULDINGS USED FOR TEACHING

2.- Autor	
Nombre: SERAFIN ALONSO CASTRO	
Programa de Doctorado: CONTROL DE PROCESOS, ELECTRÓNICA INDUSTRIAL E INGENIERÍA ELÉCTRICA	
Órgano responsable: DEPARTAMENTO DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, DE COMPUTADORES Y SISTEMAS	

RESUMEN (en español)

En la actualidad, el consumo de energía eléctrica ha aumentado considerablemente y cada vez es mayor nuestra dependencia de esta energía. Según los expertos, esta tendencia ascendente se mantendrá en el futuro, lo que implica que su coste evolucionará al alza. También la cantidad de emisiones contaminantes se verá incrementada, por lo que los gobiernos han comenzado a promulgar leyes que favorecen consumos de energía eléctrica racionales y eficientes. En este sentido, son numerosas las políticas que potencian la eficiencia energética de equipos e instalaciones en los edificios, así como un uso racional y hábitos adecuados por parte de los consumidores. Estas directrices están orientadas a conseguir un ahorro energético y económico. Por otra parte, en los últimos años se han producido importantes cambios en el sector eléctrico con el fin de alcanzar un mercado libre y competitivo, donde los consumidores de energía eléctrica tengan nuevas oportunidades de ahorro.

Los edificios públicos deberían ser los primeros en adaptarse a esta nueva situación para dar ejemplo al resto, mejorando su eficiencia energética y aprovechando las ventajas que ofrece el nuevo mercado. Para esto, es vital disponer de un sistema de medida y supervisión de la energía eléctrica, que permita conocer y analizar las variables de las instalaciones eléctricas en los edificios. Gracias a la supervisión es posible gestionar el consumo de energía eléctrica, detectar fallos, sugerir y comprobar las medidas de ahorro energético, etc. Los sistemas de medida y supervisión tradicionales son generalmente cerrados y propietarios, tienen un periodo de muestreo bajo y una capacidad de almacenamiento pequeña. Además, proporcionan escasas y limitadas herramientas de visualización y análisis de datos. En cambio, las herramientas de visualización y análisis avanzadas permiten explotar la información intrínseca contenida en el gran volumen de datos almacenados para extraer conocimiento acerca de la instalación eléctrica en el edificio y poder tomar decisiones.

El desarrollo de este tipo de herramientas de visualización y análisis avanzadas se puede basar en técnicas de minería de datos visual. La minería de datos ofrece un enorme potencial, ya que combina métodos estadísticos, inteligencia artificial, aprendizaje automático, reconocimiento de patrones, gestión de las bases de datos y visualización de la información con el fin de encontrar relaciones desconocidas, extraer patrones y resumir los datos de forma novedosa y fácilmente entendible por el usuario. Estas técnicas se emplean en la reducción de la dimensionalidad, cuantificación de vectores, agrupamiento y visualización. Algunas técnicas usadas en esta tesis son el mapa auto-organizado (SOM), proyección de Sammon y agrupamiento *k-means*. No

obstante, existen multitud de métodos similares que se revisan a lo largo de la tesis, y que también podrían ser útiles, como por ejemplo los métodos de proyección lineal y no lineal, *manifold learning*, agrupamiento o *clustering*, etc.

En esta tesis se define e implementa el sistema de medida y supervisión de energía eléctrica en los edificios de la Universidad de León, dedicados a docencia e investigación. La arquitectura lógica de este sistema se basa en una estructura de triple capa, que hace uso de un patrón cliente-servidor modificado con una capa intermedia. La capa servidor contiene el conjunto de medidores eléctricos instalados en cada edificio, una estación meteorológica común, una red de comunicaciones basada en el protocolo Modbus que interconecta el sistema de medida distribuido y un servidor de adquisición que ejecuta el servicio de toma de datos eléctricos y ambientales. La capa intermedia está formada por una base de datos y su sistema gestor dedicados al almacenamiento de datos crudos, conocimiento extraído y parámetros del sistema, el servidor de explotación de datos encargado de ejecutar los algoritmos de minería de datos y un servidor web que proporciona a los usuarios las herramientas de visualización tradicionales y avanzadas de los datos eléctricos. En la capa cliente, los usuarios acceden a los interfaces web para supervisar y analizar la energía eléctrica, ya sea *on-line* u *off-line*. En el diseño de este sistema se han utilizado tecnologías estándar, modernas y no propietarias, facilitando futuras ampliaciones, el mantenimiento, la integración con otros sistemas, etc.

En este trabajo también se propone incorporar técnicas de minería de datos visual para obtener herramientas de visualización avanzadas. Estas nuevas herramientas complementan a las tradicionales, mejorando la supervisión y análisis del consumo de energía eléctrica en los edificios de la Universidad de León. Las variables meteorológicas y temporales que influyen en el consumo eléctrico se tienen en cuenta, con el fin de descubrir las relaciones y dependencias entre estas variables y las eléctricas. Dado que el entorno ambiental es común, una supervisión conjunta de todos los edificios permite una extracción de conocimiento más profunda, habilitando la comparación entre ellos. Los algoritmos de minería de datos utilizados se encargan de explotar los datos almacenados para extraer conocimiento acerca del comportamiento eléctrico de los edificios, de forma conjunta y condicionada por el ambiente.

El algoritmo propuesto para explotar los datos eléctricos es una modificación del algoritmo SOM, denominado envSOM, que permite obtener modelos eléctricos de cada edificio, condicionados por un conjunto de variables ambientales. Estos modelos contienen información acerca del comportamiento eléctrico de los edificios, dadas unas condiciones ambientales. La metodología de explotación de los datos combina el algoritmo envSOM junto con una proyección no lineal de Sammon y un agrupamiento *k-means* para comparar los edificios entre sí y averiguar el número de perfiles eléctricos similares existentes en el Campus de la Universidad de León.
Por otra parte, la generalización del algoritmo envSOM a *n* fases permite construir modelos eléctricos condicionados jerárquicamente por variables ambientales comunes, como por ejemplo las temporales. De esta forma, es posible generar modelos eléctricos de los edificios que tengan en cuenta la periodicidad diaria, semanal y anual en el consumo eléctrico, de modo que puedan ser empleados en la supervisión de valores instantáneos, detección de desviaciones o incluso en la previsión de futuros comportamientos. Se han definido herramientas de visualización avanzadas como por ejemplo los planos de componentes eléctricas, mapas de variables temporales y gráficos de comparación y grupos, de forma que el usuario pueda interpretar el conocimiento extraído en el proceso de explotación de los datos. La comparación entre edificios puede ser global teniendo en cuenta todas las variables eléctricas, individual para cada variable eléctrica y condicionada por las variables ambientales.

Las herramientas de supervisión avanzadas basadas en técnicas de minería de datos visual que se proponen en esta tesis permiten descubrir, ampliar o confirmar conocimiento acerca del consumo de energía eléctrica en los edificios de la Universidad de León. Esto hace posible la toma de decisiones encaminadas a reducir o gestionar el

consumo, optimizar el contrato, negociar la tarifa eléctrica, etc. Las herramientas avanzadas permiten descubrir patrones y relaciones entre variables eléctricas y ambientales, detectar fallos y desviaciones eléctricas provocadas por situaciones extraordinarias o puntuales, supervisar en línea utilizando mapas de colores más intuitivos, encontrar el número de grupos de edificios con perfiles eléctricos similares, comparar los edificios en base a su comportamiento eléctrico, prever futuros consumos, etc. Los resultados obtenidos verifican que el uso de técnicas de minería de datos visual es muy útil en el análisis y supervisión de grandes instalaciones eléctricas en edificios, como son los del Campus de la Universidad de León.

RESUMEN (en Inglés)

Nowadays, electricity consumption has increased considerably and our dependence on this energy is rising. According to the experts, this upward trend will continue in the future, involving a rise of its price. The pollutants ratio will also increase, so governments have begun to promulgate laws which promote a rational and efficient use of electricity. In this sense, there are many policies which pursue energy efficiency of electrical equipment and building facilities, as well as a rational use of energy and appropriate consumer habits. The aim of these guidelines is to achieve energy and economic savings. Moreover, in the last years, the electricity sector has undergone major changes aimed to reach a free and competitive market where consumers have new saving opportunities.

Public buildings should be the first ones which adapt to this new situation in order to set an example to the remaining ones, improving their energy efficiency and taking advantage of the opportunities offered by the new electricity market. For that, it is vital to develop a system for measuring and monitoring the electricity, which allows us to know and analyze the variables from the electrical facilities in the buildings. Thanks to monitoring it is possible to manage electricity consumption, detect faults, suggest and check energy saving measures, etc. The traditional systems for electrical measuring and monitoring are generally closed and proprietary, having a low sampling period and a small storage capacity. Furthermore, they provide few and limited visualization and data analysis tools. In contrast, advanced visualization and data analysis tools allow us to exploit the intrinsic information contained in the large volume of stored data in order to extract knowledge about the electrical building facilities and make decisions.

The development of this kind of advanced visualization and analysis tools can be based on visual data mining techniques. Data mining offers a huge potential combining statistical methods, artificial intelligence, machine learning, pattern recognition, database management and information visualization to find unknown relationships, extract patterns and summarize data in an innovative way so that users can easily understand them. These techniques are used for dimensionality reduction, vector quantization, clustering and visualization. Some techniques used in this thesis are self-organizing map (SOM), Sammon's mapping and k-means clustering. However, there are many similar methods, which are reviewed throughout the thesis and may also be useful, such as linear and nonlinear projection, manifold learning or clustering methods.

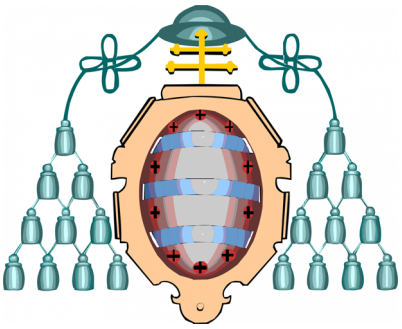
In this thesis, it is defined and implemented the system for measuring and monitoring the electricity in the buildings of the University of León, which are used for teaching and research. The logical architecture of this system is based on a three-layer structure, which makes use of a client-server pattern modified with an intermediate layer. The server layer comprises a set of electrical meters installed in each building, a common meteorological station, a communication network based on the Modbus protocol to interconnect the distributed measuring system and an acquisition server running a service to collect electrical and environmental data. The middle layer consists of a database and its management system dedicated to store raw data, extracted knowledge and system parameters, a data exploitation server responsible for executing the data mining algorithms and a web server providing users with traditional and advanced visualization tools for electrical data. In the client layer, users access to the web

interfaces for monitoring and analyzing the electricity, either on-line or off-line. In the design of this system, standard, modern and non-proprietary technologies have been used, facilitating future extensions, maintenance, integration with other systems, etc.

In this work, it is also proposed to incorporate visual data mining techniques to obtain advanced visualization tools. These new tools complement the traditional ones, improving the monitoring and analysis of electricity consumption in the buildings of the University of León. The meteorological and temporal variables which influence on the electricity consumption are taken into account to discover the relationships and dependences between these variables and electrical ones. Since the environment is common for all buildings, a joint monitoring for all of them permits a deeper knowledge extraction, enabling comparison among them. The data mining algorithms used in this work deal with exploiting the stored data to extract knowledge about the electrical behavior of the buildings, jointly and conditioned on the environment.

The algorithm proposed to exploit the electrical data is a modification of the SOM algorithm, called envSOM, which allows us to obtain electrical models for each building, conditioned on several environmental variables. These models contain information about the electrical behavior of the buildings, given certain environmental conditions. The methodology for exploiting data combines the envSOM algorithm with a nonlinear Sammon's projection and k-means clustering in order to compare buildings each other and find out the number of similar electrical profiles existing in the Campus of the University of León. On the other hand, the extension of the envSOM algorithm to n phases allows us to build electrical models conditioned hierarchically on common environmental variables, such as time-related ones. In this way, it is possible to generate electrical models of buildings which take into account the daily, weekly and annual periodicity of the electricity consumption, so that they can be used for monitoring, detecting deviations or even forecasting future behaviors. Advanced visualization tools such as planes of electrical components, maps of time variables and comparison and group graphs are defined, so that users can interpret the knowledge extracted in the data exploitation process. The comparison among buildings can be global (considering all electrical variables), individual for each electrical variable and conditioned on the environmental variables.

Advanced monitoring tools based on the visual data mining techniques proposed in this thesis let us discover, extend or confirm knowledge about the electricity consumption in the buildings of the University of León. This enables making decisions leading to reduce or manage the consumption, optimize the contract, negotiate electricity tariffs, etc. Advanced tools allow us to discover patterns and relationships between electrical and environmental variables, detect electrical faults and deviations caused by extraordinary or specific situations, monitor on-line using more intuitive color maps, find the number of building groups with similar electrical profiles, compare the buildings with regard to their electrical behavior, forecast future consumption, etc. The obtained results verify that the use of visual data mining techniques is very useful in the analysis and monitoring of large electrical facilities in buildings, such as those in the Campus of the University of León.



Universidad de Oviedo

Departamento de Ingeniería Eléctrica, Electrónica, de Computadores y Sistemas

Programa de Doctorado: Control de Procesos, Electrónica Industrial e Ingeniería Eléctrica

TESIS DOCTORAL

Supervisión de la energía eléctrica en edificios públicos de uso docente basada en técnicas de minería de datos visual

Memoria presentada para la obtención del título de Doctor por la Universidad de Oviedo

Autor: **D. Serafín Alonso Castro**

Directores: **Dr. D. Alberto B. Díez González**

Dr. D. Manuel Domínguez González

Gijón, Enero de 2012

A mi padre, in memoriam

A mi madre

A mis hermanos

Agradecimientos

El desarrollo de esta tesis es fruto de muchas horas de esfuerzo y dedicación por mi parte. Sin embargo, este trabajo no habría sido posible sin la ayuda, consejos, apoyo y ánimos de muchas personas e instituciones que me rodean. De una forma u otra, todos ellos me han ayudado a alcanzar esta meta. Por tanto, me gustaría expresarle mi más sincera gratitud a continuación.

A Manuel Domínguez y Alberto Díez, directores de esta tesis, por su excelente labor de orientación, guía y motivación, que han hecho posible este trabajo. Gracias a su experiencia, sus sabios consejos y sus críticas, siempre positivas y constructivas, la calidad de este trabajo ha mejorado considerablemente. Además, ellos me han brindado la oportunidad de formarme en el campo de la investigación.

A Juan José Fuertes, Miguel Ángel Prada y Antonio Morán por su inestimable ayuda, sus acertados consejos y sugerencias, que han hecho más llano y corto este camino para mí. Con ellos he compartido muchas horas de intenso trabajo en el laboratorio, viajes a congresos, descansos para el café y momentos distendidos e informales que ayudan a despejar la mente. Gracias por vuestra predisposición, compañerismo y calidad humana.

A mis compañeros españoles: Perfecto Reguera, Pablo Barrientos, Roberto García, Diego Fernández, Roberto Álvarez, Carlos del Canto, Sergio García, Álvaro Cantoral, Juan Manuel Ramón, David Álvarez, etc. Espero no olvidarme de nadie y si lo hago, mis más sinceras disculpas. En mayor o menor medida, todos han colaborado en este trabajo y me han aportado nuevos conocimientos e ideas, no sólo relativas a la tesis.

A mis compañeros fineses: Jaakko Hollmén, Mika Sulkava, Mikko Korpela, Janne Toivola, Prem Raj Adhikari y Miguel Ángel Prada (estos dos últimos fineses de adopción), con los que tuve la gran oportunidad de profundizar en el análisis de datos y compartir experiencias y vivencias durante el Otoño de 2010, meteorológicamente duro, pero personal y profesionalmente muy enriquecedor. A ellos, *kiitos*.

Al grupo de investigación Supervisión de Procesos Industriales (SUPPRESS) de la Universidad de León, en el cual he desarrollado este trabajo y donde me he sentido útil y totalmente integrado. En especial, a su líder Manuel Domínguez por fomentar el interés por aprender y el trabajo en grupo y además, ser una fuente constante de ideas.

Al grupo de investigación Parsimonious Modeling de Aalto University School of Science, en el cual me sentí perfectamente acogido. En especial, a su líder Jaakko Hollmén por el tiempo dedicado y las ideas compartidas durante las fructíferas discusiones de los Jueves,

que me ayudaron a ampliar el horizonte.

Al grupo de investigación Supervisión y Diagnóstico de Procesos Industriales de la Universidad de Oviedo, con el que he compartido ideas, impresiones y conocimientos, debido a varios proyectos de investigación conjuntos. En especial, a su líder Ignacio Díaz con el que siempre es un placer discutir y reflexionar acerca de cuestiones relacionadas con la temática de esta tesis.

Al Área de Automática y Control del Instituto de Automática y Fabricación (AAC-IAF) de la Universidad de León por ofrecerme interesantes proyectos de automatización, control y supervisión, que me ayudan a mantener vivo el instinto de ingeniero. En especial, a su responsable Manuel Domínguez que incansablemente impulsa este tipo de proyectos.

A la Universidad de León por el soporte económico, que aunque pequeño, vital para llevar a cabo trabajos de investigación como este. También, por el beneplácito para trabajar en sus instalaciones eléctricas, adquirir y analizar sus datos eléctricos y publicar esta información, con lo que seguramente todos saldremos beneficiados.

A la Universidad de Oviedo por proporcionarme una perspectiva diferente de la investigación al interaccionar con el mundo industrial y aportarme nuevos conocimientos durante los cursos de doctorado.

Al Departamento de Ingeniería Eléctrica y de Sistemas y Automática de la Universidad de León, al que he pertenecido durante estos años. A todos los miembros de este departamento, en especial, a María José Rodriguez por facilitarme los trámites administrativos.

Al Departamento de Ingeniería Eléctrica, Electrónica de Computadores y Sistemas de la Universidad de Oviedo, en el que se inscribe esta tesis. A todos los miembros de este departamento, en especial, a Gemma Hernández por facilitarme los trámites administrativos.

A la Unidad de Mantenimiento de la Universidad de León por facilitar la tareas de ingeniería en campo, el acceso a las instalaciones eléctricas y las gestiones administrativas. En especial al jefe de dicha unidad, Gerardo Hidalgo, por aportar su dilatada experiencia en todas las gestiones.

A mi madre y hermanos por su apoyo incondicional, su inmenso cariño, su inagotable ánimo y su infinita comprensión. En los momentos difíciles siempre han estado presentes. Por tanto, a ellos GRACIAS, con mayúsculas.

A mi padre, que aunque ya no está, nunca me olvidaré de él y sé que estaría muy orgulloso de mi, al igual que yo siempre lo estaré de él. A pesar de que este trabajo comenzó cuando él se fue y siempre me designó como técnico, sé que me habría apoyado para hacerlo realidad.

Al resto de mi familia, demasiado extensa para nombrarla aquí, y a mis amigos con los que me olvido y desconecto de estos temas. No todo en la vida es trabajo.

A TODOS: directores, compañeros, instituciones, familia, amigos, etc. MI MÁS PROFUNDO AGRADECIMIENTO.

Resumen

En la actualidad, el consumo de energía eléctrica ha aumentado considerablemente y cada vez es mayor nuestra dependencia de esta energía. Según los expertos, esta tendencia ascendente se mantendrá en el futuro, lo que implica que su coste evolucionará al alza. También la cantidad de emisiones contaminantes se verá incrementada, por lo que los gobiernos han comenzado a promulgar leyes que favorecen consumos de energía eléctrica racionales y eficientes. En este sentido, son numerosas las políticas que potencian la eficiencia energética de equipos e instalaciones en los edificios, así como un uso racional y hábitos adecuados por parte de los consumidores. Estas directrices están orientadas a conseguir un ahorro energético y económico. Por otra parte, en los últimos años se han producido importantes cambios en el sector eléctrico con el fin de alcanzar un mercado libre y competitivo, donde los consumidores de energía eléctrica tengan nuevas oportunidades de ahorro.

Los edificios públicos deberían ser los primeros en adaptarse a esta nueva situación para dar ejemplo al resto, mejorando su eficiencia energética y aprovechando las ventajas que ofrece el nuevo mercado. Para esto, es vital disponer de un sistema de medida y supervisión de la energía eléctrica, que permita conocer y analizar las variables de las instalaciones eléctricas en los edificios. Gracias a la supervisión es posible gestionar el consumo de energía eléctrica, detectar fallos, sugerir y comprobar las medidas de ahorro energético, etc. Los sistemas de medida y supervisión tradicionales son generalmente cerrados y propietarios, tienen un periodo de muestreo bajo y una capacidad de almacenamiento pequeña. Además, proporcionan escasas y limitadas herramientas de visualización y análisis de datos. En cambio, las herramientas de visualización y análisis avanzadas permiten explotar la información intrínseca contenida en el gran volumen de datos almacenados para extraer conocimiento acerca de la instalación eléctrica en el edificio y poder tomar decisiones.

El desarrollo de este tipo de herramientas de visualización y análisis avanzadas se puede basar en técnicas de minería de datos visual. La minería de datos ofrece un enorme potencial, ya que combina métodos estadísticos, inteligencia artificial, aprendizaje automático, reconocimiento de patrones, gestión de las bases de datos y visualización de la información con el fin de encontrar relaciones desconocidas, extraer patrones y resumir los datos de forma novedosa y fácilmente entendible por el usuario. Estas técnicas se emplean en la reducción de la dimensionalidad, cuantificación de vectores, agrupamiento y visualización. Algunas técnicas usadas en esta tesis son el mapa auto-organizado (SOM), proyección de Sammon y agrupamiento *k-means*. No obstante, existen multitud de métodos similares que

se revisan a lo largo de la tesis, y que también podrían ser útiles, como por ejemplo los métodos de proyección lineal y no lineal, *manifold learning*, agrupamiento o *clustering*, etc.

En esta tesis se define e implementa el sistema de medida y supervisión de energía eléctrica en los edificios de la Universidad de León, dedicados a docencia e investigación. La arquitectura lógica de este sistema se basa en una estructura de triple capa, que hace uso de un patrón cliente-servidor modificado con una capa intermedia. La capa servidor contiene el conjunto de medidores eléctricos instalados en cada edificio, una estación meteorológica común, una red de comunicaciones basada en el protocolo Modbus que interconecta el sistema de medida distribuido y un servidor de adquisición que ejecuta el servicio de toma de datos eléctricos y ambientales. La capa intermedia está formada por una base de datos y su sistema gestor dedicados al almacenamiento de datos crudos, conocimiento extraído y parámetros del sistema, el servidor de explotación de datos encargado de ejecutar los algoritmos de minería de datos y un servidor web que proporciona a los usuarios las herramientas de visualización tradicionales y avanzadas de los datos eléctricos. En la capa cliente, los usuarios acceden a los interfaces web para supervisar y analizar la energía eléctrica, ya sea *on-line* u *off-line*. En el diseño de este sistema se han utilizado tecnologías estándar, modernas y no propietarias, facilitando futuras ampliaciones, el mantenimiento, la integración con otros sistemas, etc.

En este trabajo también se propone incorporar técnicas de minería de datos visual para obtener herramientas de visualización avanzadas. Estas nuevas herramientas complementan a las tradicionales, mejorando la supervisión y análisis del consumo de energía eléctrica en los edificios de la Universidad de León. Las variables meteorológicas y temporales que influyen en el consumo eléctrico se tienen en cuenta, con el fin de descubrir las relaciones y dependencias entre estas variables y las eléctricas. Dado que el entorno ambiental es común, una supervisión conjunta de todos los edificios permite una extracción de conocimiento más profunda, habilitando la comparación entre ellos. Los algoritmos de minería de datos utilizados se encargan de explotar los datos almacenados para extraer conocimiento acerca del comportamiento eléctrico de los edificios, de forma conjunta y condicionada por el ambiente.

El algoritmo propuesto para explotar los datos eléctricos es una modificación del algoritmo SOM, denominado envSOM, que permite obtener modelos eléctricos de cada edificio, condicionados por un conjunto de variables ambientales. Estos modelos contienen información acerca del comportamiento eléctrico de los edificios, dadas unas condiciones ambientales. La metodología de explotación de los datos combina el algoritmo envSOM junto con una proyección no lineal de Sammon y un agrupamiento *k-means* para comparar los edificios entre sí y averiguar el número de perfiles eléctricos similares existentes en el Campus de la Universidad de León. Por otra parte, la generalización del algoritmo envSOM a n fases permite construir modelos eléctricos condicionados jerárquicamente por variables ambientales comunes, como por ejemplo las temporales. De esta forma, es posible generar modelos eléctricos de los edificios que tengan en cuenta la periodicidad diaria, semanal y anual en el consumo eléctrico, de modo que puedan ser empleados en la supervisión de valores instantáneos, detección de desviaciones o incluso en la previsión de futuros comportamientos.

Se han definido herramientas de visualización avanzadas como por ejemplo los planos de componentes eléctricas, mapas de variables temporales y gráficos de comparación y grupos, de forma que el usuario pueda interpretar el conocimiento extraído en el proceso de explotación de los datos. La comparación entre edificios puede ser global teniendo en cuenta todas las variables eléctricas, individual para cada variable eléctrica y condicionada por las variables ambientales.

Las herramientas de supervisión avanzadas basadas en técnicas de minería de datos visual que se proponen en esta tesis permiten descubrir, ampliar o confirmar conocimiento acerca del consumo de energía eléctrica en los edificios de la Universidad de León. Esto hace posible la toma de decisiones encaminadas a reducir o gestionar el consumo, optimizar el contrato, negociar la tarifa eléctrica, etc. Las herramientas avanzadas permiten descubrir patrones y relaciones entre variables eléctricas y ambientales, detectar fallos y desviaciones eléctricas provocadas por situaciones extraordinarias o puntuales, supervisar en línea utilizando mapas de colores más intuitivos, encontrar el número de grupos de edificios con perfiles eléctricos similares, comparar los edificios en base a su comportamiento eléctrico, prever futuros consumos, etc. Los resultados obtenidos verifican que el uso de técnicas de minería de datos visual es muy útil en el análisis y supervisión de grandes instalaciones eléctricas en edificios, como son los del Campus de la Universidad de León.

Abstract

Nowadays, electricity consumption has increased considerably and our dependence on this energy is rising. According to the experts, this upward trend will continue in the future, involving a rise of its price. The pollutants ratio will also increase, so governments have begun to promulgate laws which promote a rational and efficient use of electricity. In this sense, there are many policies which pursue energy efficiency of electrical equipment and building facilities, as well as a rational use of energy and appropriate consumer habits. The aim of these guidelines is to achieve energy and economic savings. Moreover, in the last years, the electricity sector has undergone major changes aimed to reach a free and competitive market where consumers have new saving opportunities.

Public buildings should be the first ones which adapt to this new situation in order to set an example to the remaining ones, improving their energy efficiency and taking advantage of the opportunities offered by the new electricity market. For that, it is vital to develop a system for measuring and monitoring the electricity, which allows us to know and analyze the variables from the electrical facilities in the buildings. Thanks to monitoring it is possible to manage electricity consumption, detect faults, suggest and check energy saving measures, etc. The traditional systems for electrical measuring and monitoring are generally closed and proprietary, having a low sampling period and a small storage capacity. Furthermore, they provide few and limited visualization and data analysis tools. In contrast, advanced visualization and data analysis tools allow us to exploit the intrinsic information contained in the large volume of stored data in order to extract knowledge about the electrical building facilities and make decisions.

The development of this kind of advanced visualization and analysis tools can be based on visual data mining techniques. Data mining offers a huge potential combining statistical methods, artificial intelligence, machine learning, pattern recognition, database management and information visualization to find unknown relationships, extract patterns and summarize data in an innovative way so that users can easily understand them. These techniques are used for dimensionality reduction, vector quantization, clustering and visualization. Some techniques used in this thesis are self-organizing map (SOM), Sammon's mapping and k-means clustering. However, there are many similar methods, which are reviewed throughout the thesis and may also be useful, such as linear and nonlinear projection, manifold learning or clustering methods.

In this thesis, it is defined and implemented the system for measuring and monitoring

the electricity in the buildings of the University of León, which are used for teaching and research. The logical architecture of this system is based on a three-layer structure, which makes use of a client-server pattern modified with an intermediate layer. The server layer comprises a set of electrical meters installed in each building, a common meteorological station, a communication network based on the Modbus protocol to interconnect the distributed measuring system and an acquisition server running a service to collect electrical and environmental data. The middle layer consists of a database and its management system dedicated to store raw data, extracted knowledge and system parameters, a data exploitation server responsible for executing the data mining algorithms and a web server providing users with traditional and advanced visualization tools for electrical data. In the client layer, users access to the web interfaces for monitoring and analyzing the electricity, either on-line or off-line. In the design of this system, standard, modern and non-proprietary technologies have been used, facilitating future extensions, maintenance, integration with other systems, etc.

In this work, it is also proposed to incorporate visual data mining techniques to obtain advanced visualization tools. These new tools complement the traditional ones, improving the monitoring and analysis of electricity consumption in the buildings of the University of León. The meteorological and temporal variables which influence on the electricity consumption are taken into account to discover the relationships and dependences between these variables and electrical ones. Since the environment is common for all buildings, a joint monitoring for all of them permits a deeper knowledge extraction, enabling comparison among them. The data mining algorithms used in this work deal with exploiting the stored data to extract knowledge about the electrical behavior of the buildings, jointly and conditioned on the environment.

The algorithm proposed to exploit the electrical data is a modification of the SOM algorithm, called envSOM, which allows us to obtain electrical models for each building, conditioned on several environmental variables. These models contain information about the electrical behavior of the buildings, given certain environmental conditions. The methodology for exploiting data combines the envSOM algorithm with a nonlinear Sammon's projection and k-means clustering in order to compare buildings each other and find out the number of similar electrical profiles existing in the Campus of the University of León. On the other hand, the extension of the envSOM algorithm to n phases allows us to build electrical models conditioned hierarchically on common environmental variables, such as time-related ones. In this way, it is possible to generate electrical models of buildings which take into account the daily, weekly and annual periodicity of the electricity consumption, so that they can be used for monitoring, detecting deviations or even forecasting future behaviors. Advanced visualization tools such as planes of electrical components, maps of time variables and comparison and group graphs are defined, so that users can interpret the knowledge extracted in the data exploitation process. The comparison among buildings can be global (considering all electrical variables), individual for each electrical variable and conditioned on the environmental variables.

Advanced monitoring tools based on the visual data mining techniques proposed in

this thesis let us discover, extend or confirm knowledge about the electricity consumption in the buildings of the University of León. This enables making decisions leading to reduce or manage the consumption, optimize the contract, negotiate electricity tariffs, etc. Advanced tools allow us to discover patterns and relationships between electrical and environmental variables, detect electrical faults and deviations caused by extraordinary or specific situations, monitor on-line using more intuitive color maps, find the number of building groups with similar electrical profiles, compare the buildings with regard to their electrical behavior, forecast future consumption, etc. The obtained results verify that the use of visual data mining techniques is very useful in the analysis and monitoring of large electrical facilities in buildings, such as those in the Campus of the University of León.

Índice general

Índice de tablas

Índice de figuras

Lista de símbolos

$\boldsymbol{m}_i$	Vector prototipo o codebook
$\boldsymbol{g}_i$	Coordenadas de las neuronas en la malla de salida
$\boldsymbol{x}_i$	Vector de entrada
$\boldsymbol{\omega}$	Máscara aplicada en el cálculo de la neurona ganadora
$\boldsymbol{\Omega}$	Máscara aplicada en la adaptación
$c(t)$	Neurona ganadora o BMU en el instante de tiempo t
$\boldsymbol{m}_c$	Vector prototipo de la neurona ganadora
$\boldsymbol{g}_c$	Coordenadas de la neurona ganadora en la malla de salida
$h_{ci}(t)$	Función de vecindad en el instante de tiempo t
$\alpha(t)$	Velocidad o tasa de aprendizaje en el instante de tiempo t
$\sigma(t)$	Radio de vecindad en el instante de tiempo t
σ_0	Radio de vecindad inicial
M	Número de neuronas
N	Número de muestras de datos
$E(p)$	Edificio número p
f	Número de subconjuntos de variables ambientales
VA_l	Variable ambiental l
$VA_{l_f}^f$	Variable ambiental l_f perteneciente al subconjunto f
VE_k^e	Variable eléctrica k del edificio e
n	Número de fases de entrenamiento basadas en SOM
$\boldsymbol{\omega}^{(n)}$	Máscara aplicada en el cálculo de la neurona ganadora en la fase n

$\mathbf{\Omega}^{(n)}$	Máscara aplicada en la adaptación en la fase n
$L_1\left(E(p), E(q)\right)$	Distancia L_1 o cityblook entre los edificios p y q
SG_{pq}	Similitud global entre los edificios p y q
SV_{pq}^{k}	Similitud individual para la variable eléctrica k entre los edificios p y q
SC_{pq}^{k}	Similitud condicionada para la variable eléctrica k entre los edificios p y q
Tx	Coordenada X (seno) de la variable temporal T
Ty	Coordenada Y (coseno) de la variable temporal T
$X_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$	Distancia entre dos puntos de datos $\boldsymbol{x}_i$ y $\boldsymbol{x}_j$
$\boldsymbol{c}_k$	Centroide del grupo k
$\boldsymbol{y}_i$	Vector en el espacio de salida
$\overline{x} = x_m$	Promedio de la variable x
$\boldsymbol{x}^T$	Vector traspuesto de $\boldsymbol{x}$
$\|^\circ\|$	Distancia Euclídea
a, b, c	Fases en un sistema eléctrico trifásico

Lista de acrónimos

A/D	Analógico/Digital
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
BD	Base de Datos
BMS	Building Management System
BMU	Best Matching Unit
BT	Baja Tensión
CCA	Curvilinear Component Analysis
CDA	Curvilinear Distance Analysis
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSMA/CD	Carrier Sense Multiple Access Collision Detection
CSV	Comma-Separated Values
DM	Data Mining
DMZ	De-Militarized Zone
DNP	Distributed Network Protocol
DR	Dimensionality Reduction
DSP	Digital Signal Processing
EDA	Exploratory Data Analysis
EEM	Enterprise Energy Management
EIS	Energy Information System
EMS	Energy Management System

E/S	Entrada/Salida
FIFO	First In First Out
FMS	Facilities Management System
GPS	Global Positioning System
GRNN	General Regression Neural Network
GTM	Generative Topographic Mapping
hLLE	Hessian Locally Linear Embedding
HMI	Human-Machine Interface
HTML	Hypertext Markup Language
HVAC	Heating, Ventilating and Air Conditioning Systems
IA	Inteligencia Artificial
ICA	Independent Component Analysis
IIS	Internet Information Server
IP	Internet Protocol
IrDA	Infrared Data Association
ISOMAP	Isometric Feature Mapping
KDD	Knowledge Discovery in Databases
kNN	k-Nearest Neighbor
LAN	Local Area Network
LBG	Linde-Buzo-Gray method
LE	Laplacian Eigenmap
LED	Light Emitting Diode
LLE	Locally Linear Embedding
ℓSDE	Landmark Semidefinite Embedding
LTSA	Local Tangent Space Alignment
MAPE	Mean Absolute Percentage Error
MDS	Multidimensional Scaling
MLP	Multilayer Perceptron

MT	Media Tensión
OSI	Open System Interconnection
PCA	Principal Component Analysis
PDA	Personal Digital Assistant
PLC	Programmable Logic Controller
RBF	Radial Basis Function Network
RMSE	Root Mean Squared Error
RTU	Remote Terminal Unit
SAI	Sistema de Alimentación Ininterrumpida
SCADA	Supervisory Control and Data Acquisition
SDE	Semidefinite Embedding
SGBD	Sistema Gestor de Base de Datos
SNE	Stochastic Neighbor Embedding
SOM	Self-Organizing Map
SVR	Support Vector Regression
TCP	Transport Control Protocol
THD	Total Harmonic Distortion
t-SNE	t-Distributed Stochastic Neighbor Embedding
TS-SOM	Tree-Structured Self-Organizing Map
TUR	Tarifa de Último Recurso
U-matrix	Unified Distance Matrix
UTP	Unshielded Twisted Pair
VDM	Visual Data Mining
VI	Virtual Instrument
VLAN	Virtual Local Area Network
VQ	Vector Quantization
2D	Dos Dimensiones
3D	Tres Dimensiones

Capítulo 0

Introducción

En este capítulo introductorio, se motiva el trabajo que se lleva a cabo a lo largo de esta tesis, se plantean los objetivos que se pretenden conseguir y se presenta la estructura de capítulos en los que se divide este documento.

0.1. Introducción y motivación

En los últimos años se han producido importantes cambios en el sector eléctrico motivados principalmente por la liberalización del mercado eléctrico y el esfuerzo en mejorar la eficiencia energética en las instalaciones. Uno de los objetivos de estos cambios es reducir el consumo de energía eléctrica, que ha experimentado un importante aumento, el cual continuará en el futuro. El principal responsable de este incremento es el consumo eléctrico originado por los edificios del sector terciario, el cual ha aumentado exponencialmente debido a mayores exigencias relativas al confort térmico y visual, seguridad, comodidad, etc., que requieren multitud de equipamiento eléctrico. Además, un consumo elevado provoca efectos negativos en el medio ambiente debido al incremento de las emisiones de CO_2 y otros gases y partículas contaminantes. Por este motivo, los gobiernos están instaurando políticas que fomentan el ahorro energético y el uso eficiente y racional de la energía, lo que se traduce en un ahorro económico en la factura eléctrica.

En esta situación, los consumidores de electricidad deberían adoptar estrategias encaminadas a mejorar la eficiencia energética y a gestionar la demanda de energía eléctrica de sus equipos e instalaciones. Una tarea imprescindible previa a la gestión y toma de decisiones es la supervisión del consumo de energía eléctrica. Es imposible gestionar adecuadamente una instalación eléctrica si previamente no se tiene información acerca de la misma. Por tanto, es necesario medir, adquirir, almacenar, visualizar y analizar el conjunto de variables eléctricas de la instalación. En definitiva, dotar a la instalación de un sistema de supervisión de la energía eléctrica que permita conocer la curva de carga actual, observar picos o caídas de tensión, averiguar la distribución del consumo por tramos horarios, detectar fallos o averías en los equipos, estimar consumos futuros, verificar facturas, etc. Los sistemas de supervisión de la energía eléctrica comerciales se centran básicamente en la visualización de valores puntuales y evoluciones temporales actuales y pasadas, manejo de eventos y avisos, recuperación de datos históricos para su análisis externo y realización de análisis estadísticos

sencillos. Estas funcionalidades pueden ser escasas o muy limitadas en el estudio de multitud de información de forma simultánea. Además, este tipo de sistemas son propietarios, cerrados y poco flexibles, la frecuencia de muestreo en la adquisición de los datos suele ser baja, las ampliaciones software están limitadas y son muy costosas y presentan ciertas dificultades para trabajar con grandes cantidades de datos y/o variables, lo que dificulta la tarea de análisis al usuario.

La incorporación de herramientas avanzadas al sistema de supervisión para la extracción de conocimiento implícito en los datos puede ayudar al usuario en la tarea de análisis y toma de decisiones. Por tanto, un módulo software dedicado a la explotación de datos que integre estas herramientas puede ser de gran utilidad. Los algoritmos avanzados usados en la reducción de la dimensionalidad ofrecen un enorme potencial en el estudio del elevado número de variables eléctricas procedentes de una instalación con el objetivo de mejorar su eficiencia energética y gestionar la demanda. Con frecuencia estos algoritmos se utilizan junto con técnicas de visualización, dando lugar a métodos de minería de datos visual (*Visual Data Mining*, VDM). Estos métodos aprovechan la capacidad visual del usuario para descubrir patrones en el consumo, detectar desviaciones y fallos, estimar comportamientos futuros, etc., es decir, permiten razonar en base a mapas de visualización en 2D. Un algoritmo que combina perfectamente las características de reducción de la dimensionalidad y visualización, es el mapa auto-organizado (*Self-Organizing Map*, SOM). El SOM ha sido aplicado satisfactoriamente en el campo de los sistemas eléctricos de potencia para resolver problemas aislados de diagnóstico de fallos, clasificación de consumidores, predicción de cargas, etc.

0.2. Objetivos de la tesis

Como hipótesis de partida, en esta tesis se sostiene que las técnicas de minería de datos visual resultan útiles en la supervisión de la energía eléctrica. Los objetivos planteados se centran en:

1. El diseño y desarrollo conceptual y funcional de una plataforma tecnológica para la medida, adquisición, almacenamiento y supervisión de todas las variables eléctricas involucradas en el suministro de energía eléctrica a los edificios que constituyen el Campus de la Universidad de León, así como las variables meteorológicas que caracterizan el entorno ambiental de los edificios. Esto implica la definición y especificación de los medidores eléctricos y meteorológicos, la red de comunicación entre medidores, el servicio de adquisición, la estructura de almacenamiento, el tipo de supervisión, la gestión de usuarios, etc.

2. La supervisión y análisis de los perfiles de consumo de energía eléctrica de los edificios utilizando herramientas avanzadas basadas en técnicas de minería de datos visual y herramientas tradicionales de supervisión, de forma que la extracción de conocimiento y razonamiento en base a los datos sea más sencilla y eficaz. Esto implica la elección, definición y uso de técnicas avanzadas de análisis de datos que permitan

el reconocimiento de patrones eléctricos comunes influenciados por las condiciones meteorológicas, la búsqueda de dependencias entre variables, análisis por tramos horarios, la agrupación de los edificios en base a su comportamiento eléctrico, la comparación entre edificios respecto a propiedades comunes, etc.

0.3. Estructura de la tesis

Esta tesis se estructura de la siguiente forma:

- En el capítulo 0 se introduce y se motiva este trabajo, se exponen los objetivos que se pretenden alcanzar y se indica la estructura de la tesis.
- En el capítulo 1 se sitúa al lector en la temática de la tesis. Se revisan la situación actual del consumo de energía eléctrica, los sistemas de supervisión de la energía eléctrica en edificios y los conceptos básicos de minería de datos visual.
- En el capítulo 2 se presenta el estado actual del arte relativo a técnicas de reducción de la dimensionalidad y cuantificación de vectores que tienen por objetivo la visualización de grandes volúmenes de datos, haciendo hincapié en aquellas empleadas en la supervisión y análisis de la energía eléctrica.
- En el capítulo 3 se explica en detalle la arquitectura del sistema de medida y supervisión diseñado e implementado para llevar a cabo la supervisión de la energía eléctrica de los edificios de la Universidad de León.
- En el capítulo 4 se propone la metodología basada en técnicas minería de datos visual para resolver el problema que surge en la supervisión conjunta y condicionada por un ambiente común de gran cantidad de edificios y en la comparación de su comportamiento eléctrico.
- En el capítulo 5 se explican los experimentos realizados para verificar la metodología adoptada y se muestran los resultados obtenidos en esta tesis.
- En el capítulo 6 se exponen las conclusiones extraídas en esta tesis, se enumeran las aportaciones científicas y técnicas y se sugieren las líneas futuras de investigación que se derivan de este trabajo.

Capítulo 1

Antecedentes

En este capítulo, se introduce al lector en la temática de la tesis. Inicialmente, se realiza un recorrido por la situación actual del consumo de energía eléctrica a nivel mundial y en España. A continuación, se introducen y revisan los sistemas de supervisión como medio para gestionar el consumo de electricidad y lograr la eficiencia energética en las instalaciones. Finalmente, se expone el concepto de minería de datos visual (VDM) que se utiliza en la supervisión y análisis del consumo de energía eléctrica.

1.1. Consumo de energía eléctrica

1.1.1. Situación actual y gestión de la demanda

La electricidad es una fuente de energía limpia, cómoda y cada vez más utilizada a nivel mundial por los consumidores finales. Actualmente, esta fuente de energía se sitúa en segundo lugar en orden de importancia y uso, por delante del gas y el carbón y sólo superada por el petróleo, dado su gran uso en el transporte (International Energy Agency, IEA, 2010). Por otra parte, los combustibles utilizados en la generación de electricidad son el carbón (40 %), el gas natural (20 %), renovables (18 %), nuclear (16 %) y el petróleo (6 %) (US Energy Information Administration, US EIA, 2007). La generación basada en ciclos combinados con gas natural, cogeneración y energías renovables ha sido la que ha experimentado una mayor expansión en la última década, debido al encarecimiento del petróleo, el agotamiento de los combustibles fósiles, las exigencias medioambientales, el avance de la tecnología, etc. En los últimos años, el consumo de energía eléctrica ha experimentado un aumento a nivel mundial del 2.4 %. Uno de los sectores que más influye en este crecimiento es el que agrupa a los edificios. Por ejemplo, los edificios de tipo residencial suponen el 28 % del consumo total, los comerciales el 30 %, mientras que el sector industrial y el transporte se quedan en el 24 % y 18 %, respectivamente. Esta tendencia de crecimiento se mantendrá en el futuro (ver figura 1.1 [1]), principalmente en los países desarrollados, cuyas infraestructuras eléctricas están sufriendo una fuerte expansión para adaptarse a esta situación y garantizar el correcto suministro de energía eléctrica a todos los consumidores. Este es el caso de España, donde el consumo de energía eléctrica se incrementó un 3.3 % en el año 2010, situándose en 260.61

[1] Fuente: (US Energy Information Administration, US EIA, 2007).

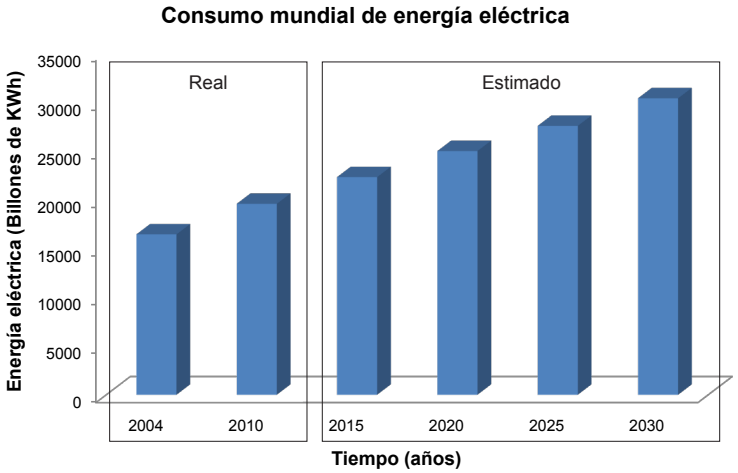


Figura 1.1: Tendencia ascendente del consumo de energía eléctrica.

TWh.

Un mayor consumo de electricidad provoca inconvenientes, como por ejemplo el aumento de emisiones de gases y partículas contaminantes. En España, se estima que el 60 % de las emisiones de SO_2, el 20 % de NO_x, el 95 % de los residuos radioactivos y el 33 % de CO_2 se deben directamente al consumo de electricidad (Pérez *et al.*, 2005). El gas CO_2 es el principal causante del efecto invernadero, por lo que ya se están adoptando estrategias para reducir sus emisiones de acuerdo con el protocolo de Kioto (Oberthür y Ott, 1999). Otro efecto negativo provocado por la tendencia ascendente del consumo de electricidad es el aumento de la demanda punta. Los consumos elevados en situaciones concretas y puntuales provocan sobrecargas en las redes de distribución y transporte, así como en generación, llegando incluso a la desconexión por seguridad. Los factores que caracterizan estas situaciones e influyen en el consumo de energía eléctrica son la temperatura ambiental, el nivel de actividad (día laborable o festivo) y la situación económica (Red Eléctrica de España, REE, 2011), siendo este último el más decisivo en el consumo final. En este sentido, el crecimiento del consumo de electricidad se ha suavizado debido a la actual crisis económica.

Las anteriores circunstancias hacen necesario adoptar políticas y estrategias que potencien el consumo racional y eficiente de la energía eléctrica y penalicen el derroche de energía, que garanticen un suministro fiable de energía en las mejores condiciones de calidad y precio, que cumplan los requerimientos ambientales en cuanto a la reducción de las emisiones contaminantes y la mayor participación de fuentes renovables en la generación. El objetivo final de estas directrices es lograr un sistema energético sostenible que concuerde con el desarrollo económico del país (España. Ministerio de Industria, Turismo y Comercio, 2010).

El suministro de energía eléctrica es esencial para el funcionamiento de cualquier sociedad moderna y por lo tanto, su precio es un factor decisivo en la competitividad de la economía de un país. El desarrollo tecnológico y funcionamiento de la industria eléctrica determinan la evolución del resto de sectores productivos. Una perfecta coordinación entre todos los

agentes involucrados, desde la generación hasta el consumo, pasando por las instituciones reguladoras, favorece una gestión eficaz de la demanda de energía eléctrica. Para ello, se debe tener presente la alta variabilidad de la demanda en el tiempo y la imposibilidad de almacenar la energía eléctrica, por lo que la oferta tiene que ajustarse perfectamente a la demanda.

Se entiende por gestión de la demanda de electricidad el conjunto de acciones cuyo fin es influir sobre el uso que los consumidores hacen de la electricidad, de forma que se fomente el uso racional y eficiente y se logre un ahorro energético considerable (Pérez *et al.*, 2005). En España, las principales líneas de actuación son las siguientes:

- Facilitar la respuesta de la demanda a los precios de la electricidad promoviendo tarifas avanzadas acordes con la demanda esperada.
- Fomentar el ahorro y la eficiencia energética en el consumo eléctrico incrementando las ayudas destinadas a los consumidores domésticos, las pequeñas y medianas empresas (PYMEs) y el sector de los servicios públicos.
- Apoyo transversal a las dos acciones anteriores mediante la implantación gradual de mediciones y registros avanzados del consumo en los diferentes tramos horarios, control remoto de la potencia demandada y comunicación bilateral entre consumidores y productores. También, se deberían impulsar actividades específicas de I+D en pro de la gestión de la demanda.

1.1.2. Demanda eléctrica en edificios del sector terciario

El sector terciario o servicios es el sector más influyente en el aumento del consumo de energía eléctrica producido en los últimos años. Este sector incluye subsectores como el comercio, transportes, comunicaciones, finanzas, turismo, hostelería, ocio, cultura, espectáculos, administración pública y servicios públicos (sanidad, educación, etc.). El sector terciario es responsable del 46.6 % del consumo de electricidad a nivel mundial (International Energy Agency, IEA, 2010). El consumo de electricidad de este sector alcanzó los 760 TWh en el año 2007 en la Unión Europea, lo que representa un crecimiento del 29 % desde el año 1999 (Bertoldi y Atanasiu, 2009). En España, el consumo de este sector se sitúa en 77 TWh, lo que supone un incremento del 72 %, bastante superior a la media europea.

Dentro del sector terciario, el consumo de electricidad originado en el interior de los edificios, ya sean residenciales o comerciales, públicos o privados, tiene una gran importancia ya que ha experimentado un rápido crecimiento en los últimos años (US Energy Information Administration, US EIA, 2007). Esto se debe principalmente al mayor número de equipamiento eléctrico instalado en los edificios con el fin de aumentar los servicios y el nivel de confort. Además, este equipamiento no suele pertenecer a la gama alta de eficiencia energética, ya que existe una escasa formación y cultura relacionada con el ahorro energético. Por tanto, es necesario adoptar directrices e instaurar acciones concretas dirigidas a aprovechar el gran potencial de ahorro de energía eléctrica en los edificios, que aún está sin explotar (Unión Europea, 2010).

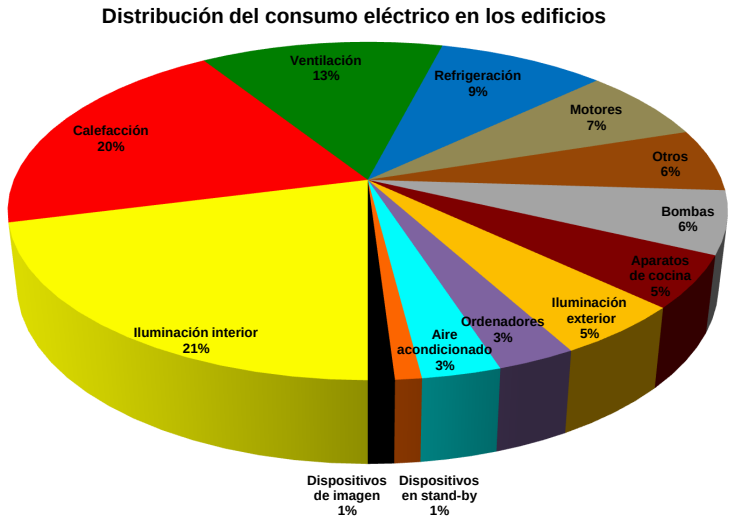


Figura 1.2: Distribución del consumo eléctrico en los edificios del sector terciario.

Entre todos los tipos de edificios, aquellos que son gestionados por las distintas administraciones e instituciones públicas deben ser un ejemplo para la sociedad, en lo que se refiere al ahorro energético y económico. Las distintas normativas deben fomentar el uso moderado, racional y eficiente de la electricidad, así como el uso de la energía renovable como fuente de electricidad en el sector de la edificación. Por ejemplo, todos los edificios públicos de nueva construcción deben incorporar algún tipo de energía renovable en sus instalaciones para autoconsumo (España, 2006).

Aunque esta tesis se puede extender a toda clase de edificios públicos, se centra principalmente en los edificios dedicados a uso docente e investigación. La adopción de políticas y estrategias encaminadas a la reducción del consumo y la eficiencia energética por parte de centros docentes podría contribuir a una buena educación energética de los jóvenes, con el fin de lograr un uso racional de la energía y por consiguiente, un ahorro energético en el futuro (Dias *et al.*, 2004).

Por otra parte, los sistemas de iluminación (interior y exterior), los sistemas de calefacción, ventilación y aire acondicionado (*Heating, Ventilating and Air Conditioning Systems*, HVAC) y los sistemas de refrigeración o frío industrial son los responsables de la mayor parte del consumo de energía eléctrica en el sector terciario (Bertoldi y Atanasiu, 2009). La distribución del consumo eléctrico dentro de este sector se puede observar en la figura 1.2 [2]. Actualmente, todos los edificios incorporan este tipo de sistemas en sus instalaciones, destacando por encima de todos los sistemas HVAC. El incremento en el consumo de energía eléctrica por parte de estos sistemas es muy significativo, llegando a representar entre el 40 % y 50 % del consumo total dentro del edificio (Pérez-Lombard *et al.*, 2008).

[2]Fuente: (Bertoldi y Atanasiu, 2009).

1.1.3. Eficiencia energética y políticas de ahorro

Generalmente, el desarrollo económico de un país lleva asociado un incremento del consumo eléctrico. Además, el aumento del número de consumidores, cuyas instalaciones, equipos y hábitos distan mucho de los ideales, favorece el crecimiento del consumo. Gracias a una adecuada gestión de la demanda de energía eléctrica es posible invertir esta tendencia ascendente. En este sentido, una de las principales líneas de actuación es el ahorro energético y la mejora de la eficiencia en las instalaciones. A nivel mundial, se estima que se malgasta el 30 % de la energía, existiendo un ahorro potencial entre el 25 % y 35 %. En España, existe un potencial para mejorar la eficiencia energética de forma económicamente rentable de al menos el 20 % (Pérez *et al.*, 2005).

Actualmente, la mejora de la eficiencia energética en las instalaciones eléctricas está siendo perseguida con gran empeño por todos los países desarrollados. Las políticas adoptadas intentan lograr la reducción del consumo de energía eléctrica, pero sin afectar a la calidad del producto o servicio obtenido por medio de esa energía. Las posibles acciones a ejecutar están orientadas a introducir mejoras en el equipamiento, mejoras en la gestión y uso de la energía o bien una combinación de ambas. A la hora de tomar medidas dirigidas a mejorar la eficiencia energética, es necesario valorar los aspectos técnicos, económicos, sociales y de mercado, no sólo los energéticos (Janssen, 2005).

Debido a su gran influencia en el consumo eléctrico total, los edificios del sector terciario son objeto de numerosas políticas, estrategias, planes o acciones, que tienen como finalidad mejorar su eficiencia energética. Las actuaciones propuestas para lograr esta mejora se centran en fomentar una cultura de ahorro y uso racional de la energía, ofrecer información y formación a los usuarios, proporcionar incentivos económicos para mejoras, favorecer las inversiones, apoyar la investigación y desarrollo de nuevas tecnologías, establecer los requisitos mínimos de eficiencia energética, instaurar medidas de control y seguimiento de las actuaciones, promover auditorías energéticas, revisar la normativa aplicable, etc. (Unión Europea, 2010). Aunque en el subsector de los edificios públicos existe un importante potencial de ahorro de energía, en la práctica este es difícil de aprovechar debido a la presencia de una serie de barreras de tipo administrativo, legal y económico que lo dificultan.

El ahorro de energía en edificios se puede lograr mediante una mejora de su gestión energética, equipos e instalaciones, es decir, supervisando y analizando los consumos energéticos y sus desviaciones, definiendo un protocolo de puesta en marcha y parada de los sistemas de climatización e iluminación, incorporando equipos de bajo consumo y alto rendimiento, etc. Pero no se deben descuidar las campañas de formación, información y concienciación a los usuarios (Instituto para la Diversificación y Ahorro de la Energía, IDAE, 2007). Otro aspecto a promover es la rehabilitación de los edificios existentes de forma que adquieran una envolvente térmica con alta calificación energética (España, 2006). Respecto a los edificios de nueva construcción, se debe seguir un proceso de certificación energética en el que se verifica la conformidad de la calificación de eficiencia energética obtenida por el proyecto y por el edificio construido (España, 2007). Esto hará que el gasto de energía en climatización del edificio, el cual supone un porcentaje elevado, sea

menor. Las actuaciones convencionales dirigidas a lograr un ahorro energético deben ser complementadas con tecnologías innovadoras y el uso de la energía renovable para mantener la sostenibilidad de los edificios (Chwieduk, 2003). Se debe alcanzar un equilibrio óptimo entre las inversiones realizadas y los costes energéticos ahorrados a lo largo del ciclo de vida del edificio.

Una herramienta muy útil para detectar oportunidades de ahorro energético o demostrar la rentabilidad económica de las inversiones en mejora de la eficiencia es la auditoria energética. Una auditoría energética es una inspección, estudio y análisis de los flujos de energía en el edificio. Las auditorías permiten conocer la situación energética actual, conocer el funcionamiento y eficiencia de los equipos e instalaciones, inventariar los principales equipos e instalaciones existentes, realizar mediciones y registros eléctricos y térmicos, analizar las posibilidades de optimización del suministro de energía, analizar la posibilidad de instalar energías renovables, proponer mejoras y realizar su evaluación técnica y económica (Ente Regional de la Energía de Castilla y León, EREN, 2009).

La innovación y desarrollo tecnológico es otra pieza clave en el proceso de mejora de la eficiencia energética, de forma que la eficacia de las medidas dependerá de la rapidez en la incorporación de nuevos equipos al mercado, produciendo un aumento del conocimiento, de la competitividad de las empresas y de los servicios que prestan.

1.1.4. Liberalización del mercado eléctrico y tarifas.

El proceso de liberalización del mercado eléctrico fue promovido por la Unión Europea, que en el año 1997 estableció las bases iniciales mediante la directiva 96/92/CE (Unión Europea, 1997). El objetivo era establecer un marco jurídico común para la gestión integral del sistema eléctrico europeo, alcanzar un mercado libre y abierto en cuanto a competencias, garantizar el suministro de energía y fomentar un consumo racional y eficiente de la energía eléctrica en los estados miembros (Jamasb y Pollitt, 2005). El proceso de liberalización del sector eléctrico se llevó a cabo de forma paralela en cada país miembro, que fueron transponiendo las directrices europeas a su legislación. En España se formuló la ley 54/1997 del sector eléctrico (España, 1997), en la que se estableció un proceso gradual para la entrada de consumidores en el mercado libre de la electricidad. Primeramente, sólo se habilitó este acceso a usuarios con gran consumo de energía, luego a aquellos con suministros en media y alta tensión y finalmente, a partir del año 2003, todos los usuarios se consideran cualificados y tienen derecho a contratar la electricidad en el mercado libre (España, 2000).

Antes de 1997, el gobierno controlaba totalmente las tarifas eléctricas y regulaba el precio de la energía en cada periodo horario. En la figura 1.3 se muestra, a modo de ejemplo, una tarifa eléctrica regulada típica de 3 periodos. A partir de 2003, se inicia un periodo de transición, en el cual coexisten tarifas reguladas y libres. Es el propio usuario el que debe elegir el tipo de mercado en el que operar. En Julio de 2009, se instaura definitivamente el mercado libre para la energía eléctrica. A partir de esta fecha, los precios y las tarifas son totalmente libres y pueden ser negociados con la compañía eléctrica. También aparece la tarifa de último recurso (TUR) para pequeños usuarios que no quieran o expresen

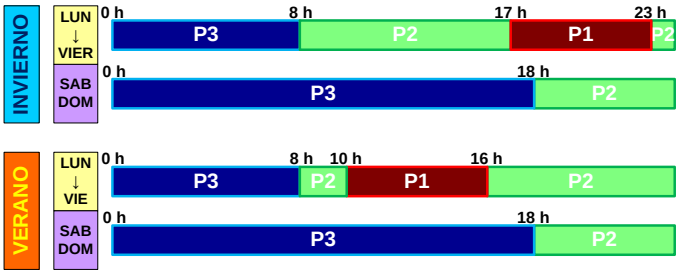


Figura 1.3: Tarifa eléctrica regulada típica de 3 periodos horarios (P1 o punta, P2 o llano y P3 o valle).

explícitamente su deseo de negociar con la compañía. El gobierno establece los precios máximos aplicables para estas tarifas TUR (España, 2009).

El proceso de liberalización del mercado eléctrico distingue claramente tres negocios dentro del sector: la generación, el transporte y distribución y la comercialización. Una nueva figura aparece en el sector, la comercializadora de electricidad. La tarea de comercialización incluye la compra de electricidad a los productores, el pago de peajes a los distribuidores por el uso de sus redes y la venta de esa energía a los consumidores finales.

Por tanto, a partir de ahora sólo existirán dos posibilidades de contratación de energía eléctrica:

- **Acudir al mercado liberalizado**: el cliente contrata la electricidad con una empresa comercializadora a un precio y condiciones libremente pactadas, acude al mercado mayorista o *pool* de energía o bien suscribe un contrato directamente con un productor.
- **Acogerse al mercado regulado**: el cliente contrata la electricidad con una empresa comercializadora de último recurso a un precio fijado por el gobierno. Solamente los usuarios con una potencia contratada igual o inferior a 10 KW y suministro en baja tensión se pueden acoger a la TUR correspondiente.

Las ventajas que se pueden obtener en el mercado liberalizado son interesantes desde el punto de vista económico. El consumidor puede negociar el precio de la energía eléctrica con una empresa comercializadora, o bien solicitar ofertas a varias compañías. El usuario también puede contratar directamente con una empresa distribuidora el acceso a sus redes eléctricas. Todo esto implica cambios sustanciales en el tipo de contrato establecido, en la forma de medida, tipo de contador, etc., lo que conlleva un coste adicional relacionado con el cambio de contratación. Por tanto, los usuarios deben valorar los precios y servicios ofrecidos por las empresas comercializadoras en uno y otro tipo de mercado antes de tomar la decisión.

La mayor parte de los edificios públicos, concretamente los destinados a la docencia e investigación, poseen suministros de energía eléctrica en media o baja tensión y generalmente superan los 10 KW de potencia contratada. Por este motivo, las administraciones o instituciones públicas deben acudir al mercado liberalizado para la contratación de

electricidad. Esto implica que previamente se debe adquirir un perfecto conocimiento del comportamiento eléctrico de los edificios para, posteriormente, proceder de forma favorable en la negociación con la empresa comercializadora.

1.2. Supervisión de la energía eléctrica

1.2.1. Supervisión básica

Las grandes instalaciones eléctricas en los edificios requieren un sistema de supervisión básico que permita medir, registrar, visualizar y analizar sus variables. Los sistemas de supervisión son vitales en la gestión eficiente del consumo eléctrico ya que es difícil gestionar o controlar lo que no se puede medir (Rossiter, 2005). Por tanto, es necesario conocer y disponer de información acerca de la instalación eléctrica en un edificio, lo que aporta una serie de ventajas tales como el seguimiento de las mejoras en la eficiencia, comprobación del ahorro energético y económico, verificación y optimización de facturas eléctricas, ayuda en la realización de auditorías energéticas, previsión de posibles penalizaciones, predicción del consumo futuro, detección y diagnóstico de fallos, obtención de informes, etc.

Las funciones principales de un sistema de supervisión básico son la medida de todas las variables eléctricas, adquisición de los datos, supervisión mediante interfaces hombre-máquina (*Human-Machine Interface*, HMI), visualización de valores instantáneos e históricos, gestión de alarmas, comunicación con otros sistemas, almacenamiento de los datos y análisis de toda la información. Los sistemas SCADA (*Supervisory Control and Data Acquisition*) se utilizan con cierta frecuencia en la supervisión de energía eléctrica (Qiu *et al.*, 2002; Yao y Ku, 2003). En este caso, la funcionalidad de control no existe o es mínima y el sistema SCADA realiza simplemente funciones de supervisión, tomando datos de uno o varios medidores eléctricos situados en el nivel de campo y representándolos convenientemente en diversas pantallas de visualización. En la figura 1.4, se muestra una pantalla de visualización básica de un software SCADA comúnmente utilizado en la supervisión de energía eléctrica.

Aunque la supervisión básica de la energía eléctrica cubre la mayor parte de las necesidades, presenta una serie de limitaciones o inconvenientes. Generalmente, este tipo de sistemas son propietarios de la empresa que los desarrolla y por consiguiente, poseen una estructura compleja y cerrada, de forma que la integración con otros sistemas de terceros es más bien escasa o nula. Además, se centran en la medida, registro y supervisión de la energía consumida, despreciando el resto de variables eléctricas. Por otra parte, no incorporan herramientas avanzadas para el análisis de los datos históricos y solamente permiten realizar estudios estadísticos sencillos y preestablecidos.

1.2.2. Evolución de la supervisión

Los sistemas de supervisión básicos han ido evolucionando en el tiempo, gracias en parte al avance y desarrollo de la tecnología. Actualmente, los instrumentos de medida son capaces de muestrear a una frecuencia elevada, son configurables, pueden almacenar

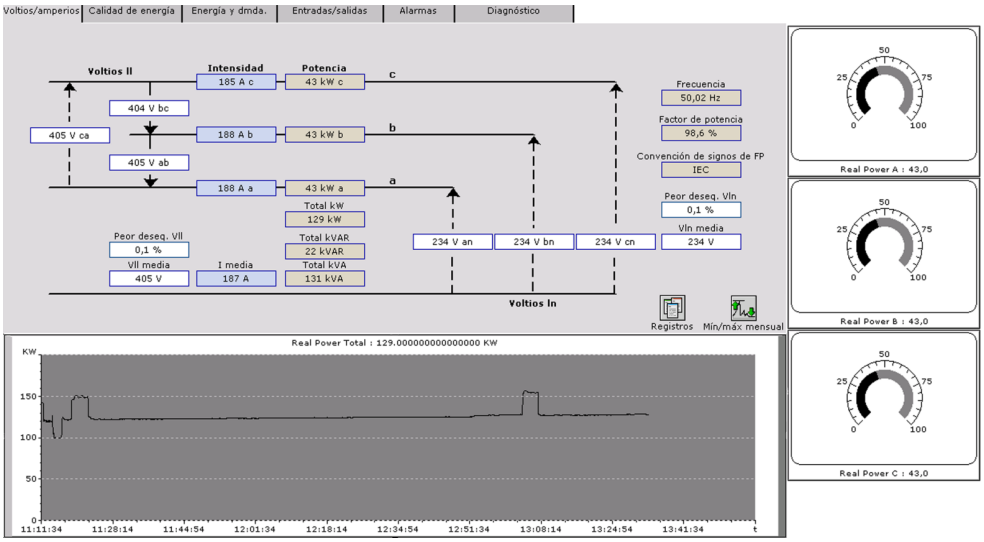


Figura 1.4: Pantalla de visualización de un software SCADA utilizado en la supervisión de energía eléctrica.

datos en memoria interna de forma temporal, soportan varios protocolos de comunicación, etc. Por otro lado, las aplicaciones software para adquirir y analizar datos utilizan bases de datos abiertas, incluyen tareas automatizadas para la obtención de informes, permiten exportar texto, imágenes o datos en cualquier formato, son accesibles de forma remota a través de Internet, notifican las alarmas o avisos a través del móvil o correo electrónico, etc. Las herramientas avanzadas adquieren un papel esencial en la supervisión y análisis ya que deben ser capaces de trasladar la información contenida en la totalidad de los datos a la toma de decisiones (Khan, 2001). Además, un sistema de supervisión exhaustivo no debe olvidar la medida y análisis de la calidad de la energía, cuya disminución puede causar un funcionamiento ineficiente de los equipos o incluso llegar a dañarlos (Cristaldi *et al.*, 2002).

La necesidad de obtener información acerca de la energía total consumida dentro de los edificios y no sólo la eléctrica, ha impulsado el desarrollo de los denominados sistemas de información de la energía (*Energy Information Systems*, EIS) (Motegi y Piette, 2002). Estos sistemas proveen a través de Internet información acerca del uso y demanda de la energía, las características energéticas del edificio, datos de los sistemas HVAC, condiciones meteorológicas, precio de la energía, etc. La gestión conjunta de todos los consumos energéticos dentro de un edificio genera sinergias relativas al ahorro energético que se deben aprovechar. Los sistemas dedicados a la gestión de energía en la empresa (*Enterprise Energy Management*, EEM) son un tipo de EIS con características avanzadas para el análisis, que determinan en tiempo real las necesidades energéticas del edificio y transmiten esta información a los agentes encargados de la generación, transporte y distribución de la energía (Forth y Tobin, 2002).

La extensión de las redes de comunicación y la aparición de diversos protocolos estándar han contribuido a la mejora en la integración de los sistemas de supervisión con otro tipo de sistemas dentro de los edificios (control de la climatización, iluminación, accesos, etc.),

ya sean del mismo o distinto fabricante. La coordinación entre todos estos servicios es posible gracias a los denominados sistemas de gestión del edificio (*Building Management System*, BMS) y de las instalaciones (*Facilities Management System*, FMS). Además, estos sistemas de gestión hacen uso de Internet como medio de acceso, lo que facilita la integración de diferentes sistemas dentro de la misma arquitectura (Wang y Xie, 2002). Los edificios equipados con este tipo de sistemas de gestión se denominan con cierta frecuencia edificios inteligentes (*Smart buildings*).

La tendencia en los sistemas de gestión de la energía (*Energy Management System*, EMS) es incorporar herramientas avanzadas que permitan realizar la explotación del gran volumen de datos históricos almacenados. Debido a la existencia de interacciones no lineales entre las variables de un sistema eléctrico de potencia, las técnicas de inteligencia artificial (IA) son muy útiles para resolver problemas en el control, operación y planificación del mismo (Dahhaghchi *et al.*, 1997). Dentro del campo de la inteligencia artificial, los algoritmos de aprendizaje automático (*Machine learning*) (Alpaydin, 2004) se emplean para reducir la dimensionalidad de los datos, lo que facilita la visualización y la extracción de información implícita. Concretamente, las redes neuronales artificiales (*Artificial Neural Networks*, ANNs) tienen una gran aceptación en la resolución de problemas complejos dentro de los sistemas eléctricos de potencia, tales como la clasificación de perfiles de consumo, evaluación de interrupciones en el suministro, predicción de la demanda, diagnóstico de fallos, supervisión de la calidad de la energía, análisis de maquinaria y equipamiento eléctrico, planificación y toma de decisiones, etc. Cabe destacar que es necesario elegir bien el tipo de red neuronal, las características o variables, los datos de entrenamiento y los parámetros de ajuste para obtener buenos resultados (Vankayala y Rao, 1993).

1.3. Minería de datos visual

1.3.1. Descubrimiento de conocimiento en bases de datos y minería de datos

El concepto de descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Databases*, KDD) se refiere al proceso automatizado, complejo e iterativo que combina un conjunto de métodos para extraer conocimiento a partir de grandes volúmenes de datos (Fayyad *et al.*, 1996a). Este proceso implica una serie de tareas (Fayyad *et al.*, 1996b):

- Gestión de la base de datos.
- Preparación y transformación de los datos.
- Extracción de las características más relevantes de los datos.
- Minería de datos: encontrar patrones, reglas o modelos interesantes y desconocidos.
- Evaluar los resultados.
- Interpretar y utilizar el conocimiento extraído.

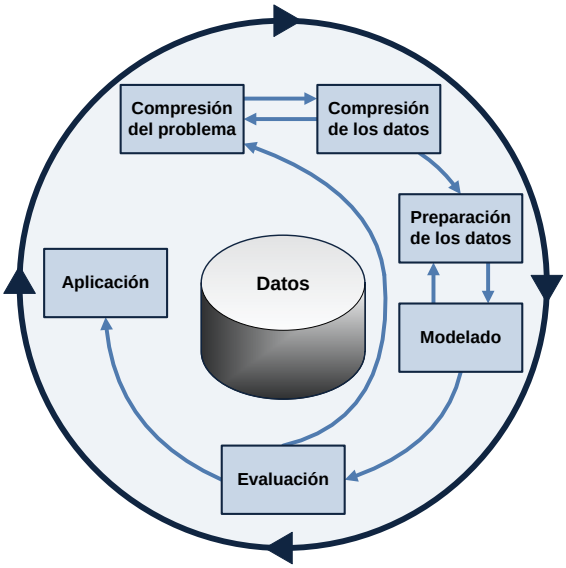


Figura 1.5: Estándar CRISP-DM para la minería de datos.

La minería de datos (*Data Mining*, DM) es una tarea dentro del proceso KDD centrada en la aplicación de algoritmos específicos en el análisis de grandes volúmenes de datos con el fin de encontrar relaciones desconocidas, extraer patrones y resumir los datos de forma novedosa y fácilmente entendible por el usuario (Hand *et al.*, 2001). La minería de datos es una tarea interdisciplinar del campo de la ciencia de computadores que combina métodos estadísticos, inteligencia artificial, aprendizaje automático, reconocimiento de patrones, gestión de las bases de datos y visualización. El aumento de la potencia de cálculo de procesamiento, la capacidad de almacenamiento y la conectividad entre computadores han provocado la proliferación de las aplicaciones de minería de datos para transformar simples datos en conocimiento, lo que proporciona siempre una ventaja cualitativa.

El término KDD está asociado al conjunto de tareas que requiere el descubrimiento de conocimiento en bases de datos, mientras que el término DM se refiere al conjunto de técnicas o métodos utilizados para extraer patrones y relaciones de los datos. No obstante, el concepto DM es con frecuencia usado en un contexto más amplio que se corresponde con el proceso KDD, como por ejemplo en el campo de la estadística, análisis de datos y bases de datos. Por el contrario, el término KDD es más popular en el campo de la inteligencia artificial y el aprendizaje automático (Fayyad *et al.*, 1996c). Por tanto, no es extraño encontrar en la literatura el término DM para referirse al proceso en conjunto.

El creciente interés en el desarrollo y uso de aplicaciones basadas en técnicas de minería de datos ha provocado la aparición de varios estándares relativos a las tareas (definición de las entradas y salidas), a la tecnología utilizada en las aplicaciones (protocolos para el acceso e intercambio de datos), al proceso en conjunto (secuencia de tareas o etapas) y a la arquitectura (capas o partes del sistema) (Clifton y Thuraisingham, 2001). Un ejemplo de estos estándares, ampliamente difundido por los proveedores de aplicaciones de minería de datos y utilizado en la industria, es el modelo CRISP-DM (*Cross-Industry Standard Process*

for Data Mining) (Chapman *et al.*, 2000). El modelo CRISP-DM estandariza y amplia la visión del proceso KDD, de forma que no sólo se centra en la preparación de los datos y modelado (Laine, 2003). Este estándar será seguido en la implementación de las tareas de minería de datos que se proponen a lo largo de esta tesis. La metodología seguida por el modelo CRISP-DM divide el proceso de minería de datos en las siguientes tareas o etapas (ver figura 1.5):

1. **Comprensión del problema**: determinar los objetivos, evaluar la situación, determinar las metas de la minería y planificar el proyecto.

2. **Comprensión de los datos**: adquirir, describir, explorar y verificar la calidad de los datos.

3. **Preparación de los datos**: seleccionar, preprocesar, calcular, integrar y dar forma a los datos.

4. **Modelado**: seleccionar la técnica de modelado, crear un diseño para el *test*, construir un modelo y evaluar el modelo.

5. **Evaluación**: evaluar los resultados, revisar el proceso de minería y determinar los pasos siguientes.

6. **Aplicación**: puesta en funcionamiento de la aplicación, planificar el seguimiento y mantenimiento, generar un informe final y revisar el proyecto.

La minería de datos visual (*Visual Data Mining*, VDM) combina las técnicas de DM junto con las técnicas de visualización para realizar la exploración y análisis de grandes volúmenes de datos (Keim, 2002; Ferreira de Oliveira y Levkowitz, 2003). VDM realiza una transformación de los datos con el fin de lograr una representación visual que permita una interpretación fácil, intuitiva y rápida por parte de cualquier usuario sin conocimientos específicos en DM. Por tanto, se integra al usuario en el proceso de análisis, explotando la habilidad humana en la percepción visual y aprovechando el escaso esfuerzo que requiere el razonamiento en base a objetos visuales. VDM convierte el proceso de análisis de los datos en la resolución de un problema visual, lo que aporta mayor eficacia en la obtención de resultados y reduce el esfuerzo (Wong, 1999).

Cualquier sistema moderno de supervisión de la energía eléctrica es capaz de adquirir gran cantidad de información y almacenarla en su base de datos. La extracción no trivial de la información implícita, desconocida y potencialmente útil que se encuentra oculta en los datos puede ser una labor relativamente sencilla, rápida y automática si se emplean técnicas de DM y se integran dentro de las herramientas de supervisión. En la literatura se pueden encontrar numerosos trabajos que emplean técnicas de DM para la extracción de conocimiento de grandes bases de datos propiedad de empresas eléctricas y así poder caracterizar el tipo de clientes o consumidores (Sforna, 2000; Figueiredo *et al.*, 2005).

1.3.2. Análisis exploratorio de datos y reconocimiento de patrones

Los objetivos perseguidos en DM se pueden clasificar de acuerdo al tipo de actividad o tarea y algoritmos requeridos (Hand *et al.*, 2001). A continuación se enumeran estos objetivos:

1. Análisis exploratorio de los datos.
2. Modelado descriptivo.
3. Modelado predictivo: clasificación y regresión.
4. Descubrimiento de patrones y reglas.
5. Recuperación por contenido.

El análisis exploratorio de los datos (*Exploratory Data Analysis*, EDA) es la primera actividad dentro de DM y su misión es examinar de forma global el conjunto de datos (Tukey, 1977). Es decir que, los datos se exploran sin una idea clara y concreta de lo que se está buscando. Normalmente, EDA se emplea para obtener un resumen de los datos, comprobar la calidad de los mismos y/o formular hipótesis que sean interesantes de verificar. Por tanto, EDA permite complementar a las técnicas estadísticas convencionales que se emplean en la verificación de hipótesis antes de un análisis de datos confirmatorio (*Confirmatory Data Analysis*) (Tukey, 1980). El análisis confirmatorio se puede aplicar una vez que se dispone de hipótesis estadísticas para el problema bajo estudio.

EDA hace uso de un conjunto de técnicas estadísticas y gráficas, como por ejemplo gráficos de caja, gráficos dispersos, análisis de componentes principales, análisis de correlación, análisis de factores, etc. (Tikka, 2008). Los métodos de visualización pueden servir de gran ayuda en la generación efectiva de hipótesis. Asimismo, las técnicas de proyección y reducción de la dimensionalidad son herramientas útiles en EDA, ya que transforman los datos multidimensionales a una baja dimensión, facilitando su visualización y exploración (Sulkava, 2008). En Hoaglin *et al.* (2000) se pueden revisar varias herramientas útiles en EDA.

Otro de los objetivos de DM es el descubrimiento de patrones y reglas, es decir, la búsqueda de características locales interesantes en los datos, en lugar de construir un modelo global. El reconocimiento de patrones (*Pattern recognition*) es una parte del aprendizaje automático, cuya función es descubrir automáticamente regularidades en los datos utilizando algoritmos ejecutados en un computador (Bishop, 2006). En DM, se usan diferentes métodos del campo del aprendizaje automático. Estos algoritmos son capaces de aprender a partir de datos, siguiendo estrategias de optimización para predecir en base a la muestra de datos y encontrar generalizaciones o reducciones de los datos. El tipo de aprendizaje se puede dividir en supervisado, es decir, aquel que requiere especificaciones de las salidas e información conocida del entorno y no supervisado, donde las salidas del modelo son desconocidas y se intenta representar la estructura interna de los datos que minimiza un criterio de error. Como ejemplo de aprendizaje supervisado, podemos citar la clasificación y

regresión. El agrupamiento, estimación de la densidad y métodos de proyección son ejemplos de aprendizaje no supervisado (Cherkassky y Mulier, 2007).

1.3.3. Visualización de la información

La visualización de la información (*Information visualization*) es la tarea que tiene por objetivo la representación de toda clase de información abstracta, ya sean datos de tipo numérico o no, de forma visual, interactiva y soportada en un computador, ampliando el conocimiento acerca de la misma (Card *et al.*, 1999). Las técnicas de visualización de la información implican la representación de los datos en gráficos (*Data visualization*), así como la revelación de la estructura e información implícita de los mismos (*Statistical and scientific visualization*). Estas técnicas de visualización tratan de exponer toda clase de información, destacando sus aspectos más significativos, de modo claro, efectivo y con un significado gráfico que, generalmente, resulta más intuitivo y sencillo de comprender (Fayyad *et al.*, 2002). La calidad de las visualizaciones obtenidas viene determinada por la precisión en la representación de los datos y por una buena legibilidad.

La tendencia en la visualización es alcanzar un compendio entre la funcionalidad y la vistosidad, de modo que las herramientas permitan visualizar la información claramente, pero que a la vez estimulen la atención del usuario. No obstante, el uso decorativo de las representaciones gráficas puede conllevar a una desinformación, si esto es realizado por usuarios inexpertos (Tufte, 1983).

Como ya se ha comentado, VDM requiere el uso de técnicas de visualización orientadas a la exploración y análisis visual de los resultados producidos por los métodos de DM. En este caso, las técnicas de visualización persiguen transformar la información, sin pérdida significativa, hacia representaciones gráficas que permitan al usuario razonar de forma visual, en base a imágenes, gráficos, iconos, etc. (Keim, 2001). El usuario participa activamente en el proceso de DM, aportando creatividad, pericia, flexibilidad y conocimiento en la resolución del problema. En esta tesis, se hace uso de las técnicas de visualización para supervisar la energía eléctrica en edificios.

Según Keim y Kriegel (1996), las técnicas de visualización orientadas al análisis de grandes conjuntos de datos se pueden clasificar en:

- **Técnicas orientadas a píxeles**: cada dato es asignado a un píxel de una figura.
- **Técnicas geométricas**: los datos son proyectados sobre un espacio de visualización.
- **Técnicas basadas en iconos**: cada dato es asignado a un icono que describe los datos.
- **Técnicas jerárquicas**: las variables son ordenadas en un conjunto jerárquico.
- **Técnicas basadas en gráficos**: los datos se representan en gráficos que facilitan la comprensión.

Los aspectos a tener en cuenta a la hora de visualizar la información durante el proceso de VDM son los siguientes (Keim, 2002):

- **Tipo de datos a ser visualizados**: datos de una (1D) o dos dimensiones (2D), datos multidimensionales, texto, jerarquías, gráficos y algoritmos o software.
- **Técnica de visualización empleada**: gráficos 2D/3D, transformaciones geométricas, gráficos basados en iconos, píxeles compactos y gráficos apilados.
- **Técnica de interacción disponible**: proyección, filtrado, zoom, búsqueda, solicitud, enlazado, separado y distorsión.

La tarea principal de las técnicas de visualización es, pues, presentar un gran volumen de datos al usuario para agilizar su razonamiento, facilitar la exploración de los datos y ayudar en el reconocimiento de estructuras, patrones, novedades, anomalías, tendencias o correlaciones. También deben facilitar la recuperación y búsqueda de información, la comparación entre valores y el descubrimiento de errores o desviaciones. Las capacidades de las representaciones gráficas de información se pueden ampliar, integrándolas dentro de un entorno computacional que soporte el uso de interacciones y distorsiones sobre la información para proporcionar diferentes vistas de los resultados (Himberg, 2004).

La expansión de la red Internet, que es en sí misma una enorme fuente de información compleja, diversa y dinámica, ha hecho posible el desarrollo de numerosas técnicas de visualización de la información destinadas a facilitar la navegación de los usuarios, localizar la información de forma precisa y rápida y analizar datos de acceso y tráfico en la red (Zhang, 2007). Herramientas de visualización similares se podrían integrar en aplicaciones web con el fin de proporcionar diferentes vistas interactivas de la información a varios usuarios simultáneamente, independientemente del lugar donde se localicen (Rohrer y Swing, 1997).

Generalmente, las técnicas de visualización tratan de resumir toda la información en la pantalla de un computador o en un espacio visualizable 2D ó 3D. Cuando se desea visualizar información multidimensional compleja, es necesario aplicar previamente otro tipo de técnicas que realicen la reducción de la dimensionalidad, de modo que los datos puedan ser proyectados en un espacio de baja dimensión visualizable. De igual manera, es necesario simplificar el conjunto de datos realizando una compresión de los mismos para obtener sus vectores prototipo. Esta simplificación del conjunto de datos a información estrictamente relevante hace posible la abstracción necesaria que requiere la tarea de visualización.

1.3.4. Reducción de la dimensionalidad, cuantificación de vectores y proyección

Como se ha expuesto anteriormente, la visualización efectiva de la información contenida en un gran conjunto de datos con multitud de variables requiere previamente reducir su dimensionalidad. La reducción de la cantidad de datos es también muy importante antes del procesamiento de grandes volúmenes de datos por ciertos algoritmos, minimizando la carga computacional que requieren.

La reducción de la dimensionalidad (*Dimensionality Reduction*, DR) es el proceso que tiene como objetivo representar de forma compacta el conjunto de datos de entrada, en un

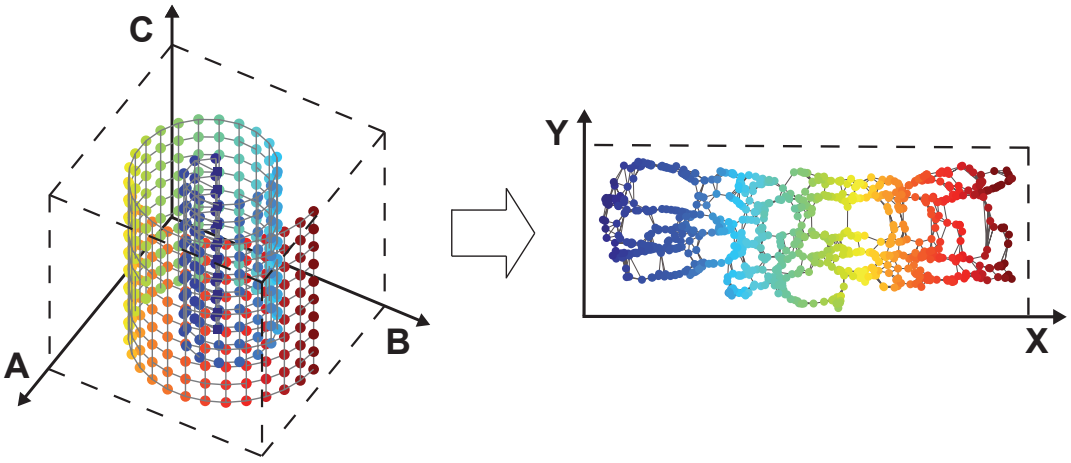


Figura 1.6: Reducción de la dimensionalidad.

espacio de salida de baja dimension visualizable, conservando la estructura de los datos y su información más relevante (Carreira-Perpiñán, 1996). Desde el punto de vista del reconocimiento de patrones, los datos de entrada se transforman a una representación útil para su comprensión y utilización, mientras que desde el criterio geométrico se trata de una representación en un sistema de coordenadas diferente. La reducción de la dimensionalidad trata de paliar el problema existente en el manejo de datos de alta dimensionalidad, es decir, con un número elevado de variables, conocido como maldición de la dimensionalidad (*Curse of dimensionality*) (Bellman, 1961). Este término se refiere a que el número de datos necesarios para estimar una función de varias variables con un cierto grado de precisión crece exponencialmente con el número de variables, siendo los espacios de alta dimensionalidad intrínsecamente dispersos. La reducción del número de variables es posible gracias a que la dimensión intrínseca o latente de los datos suele ser menor que la real, debido a la existencia de variables correlacionadas y con ruido e información redundante. Por tanto, las técnicas de reducción de la dimensionalidad tratan de encontrar el número de variables independientes que caracterizan el conjunto de datos, mejoran la capacidad de generalización en clasificadores y reducen la carga computacional en la clasificación de patrones. Estas técnicas se pueden clasificar en tres tipos siguiendo el orden de aplicación en el proceso de minería de datos visual (Cuadrado, 2002):

- **Selección de variables**: métodos estadísticos, conocimiento a priori, etc.
- **Extracción de características**: técnicas de análisis espectral, normalización, etc. Se suele utilizar conocimiento a priori.
- **Técnicas de proyección**: proyecciones lineales, escalado multidimensional, modelos autoasociativos, mapas auto-organizados, etc. Estas técnicas no suelen aplicar conocimiento previo, sólo consideraciones en la geometría de la distribución de los datos.

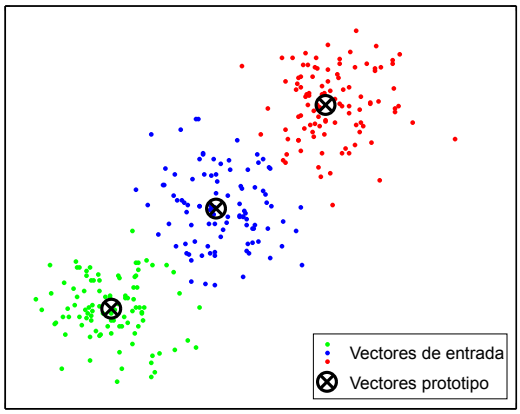


Figura 1.7: Cuantificación de vectores.

Esta tesis se centra en el uso de técnicas de proyección destinadas a la reducción de la dimensionalidad. Las técnicas de proyección establecen una correspondencia no biyectiva de tipo "muchos a uno" entre cada punto del espacio de datos o entrada y un punto en el subespacio o variedad topológica (*Manifold*[3]) de menor dimensión intrínseca dentro del espacio de entrada. A continuación se podrá establecer, una correspondencia biyectiva entre este subespacio topológico y el espacio de visualización o salida que tiene la misma dimensión y es normalmente 2D (Venna, 2007). Por tanto, el objetivo de la reducción de la dimensionalidad es encontrar ese subespacio o variedad topológica que permita la proyección de los datos a una forma compacta y una reconstrucción no singular y derivable, aunque contenga pequeños errores (Carreira-Perpiñán, 2001). La proyección debe evitar la pérdida de información significativa y conservar la topología de los datos, es decir, que puntos adyacentes o vecinos en el espacio de entrada deben serlo también en el espacio de salida (ver figura 1.6). Existen muchas calificaciones y clasificaciones asociadas a las técnicas de proyección, como por ejemplo lineal-no lineal (el uso de modelos no lineales que conectan las variables intrínsecas con las observadas suele ser más eficaz), continua-discreta (generalizar la proyección a otros puntos es directo cuando el modelo es continuo y es necesaria la interpolación si es discreta), fuerte-suave (se requiere una reducción drástica cuando el espacio es de alta dimensionalidad, mientras que la leve se utiliza en espacios de menor dimensionalidad), etc. (Lee y Verleysen, 2007).

Por otra parte, la compresión o compactación de los datos facilita su manejo y procesamiento. Para lograr esto, se emplean técnicas de cuantificación de vectores (*Vector Quantization*, VQ), cuyo objetivo es modelar la función de densidad de probabilidad de los datos, es decir, aproximar los vectores de entrada utilizando un número pequeño de vectores prototipo que los representan (Gray y Neuhoff, 1998). Una cuantificación de vectores es una transformación del espacio de datos o entrada de alta dimensión en un conjunto pequeño de datos prototipo, que minimizan la función de distorsión, es decir, el coste asociado a

[3]Espacio topológico que en una escala lo suficientemente pequeña se asemeja al espacio Euclídeo de una dimensión específica, llamada dimensión de la variedad, es decir, cada punto k-dimensional de un manifold tiene una vecindad homeomorfica en un subespacio k-dimensional abierto perteneciente a $\mathbb{R}^k$.

la sustitución del dato de entrada por su prototipo correspondiente (ver figura 1.7). La cuantificación de vectores posee gran similitud con las técnicas de agrupamiento o *clustering*, ya que ambas dividen el conjunto de datos de entrada en diferentes grupos, los cuales se representan por su centroide. No obstante, existe una diferencia fundamental que estriba en el objetivo perseguido. La cuantificación de vectores intenta representar a los datos mediante sus vectores prototipo, mientras que las técnicas de *clustering* buscan agrupaciones interesantes de los datos de entrada para su interpretación.

Entre todas las técnicas utilizadas en la reducción de la dimensionalidad y cuantificación de vectores, destaca el mapa auto-organizado (*Self-Organizing Map*, SOM). El SOM permite lograr ambos objetivos simultáneamente a la vez que ofrece numerosas herramientas para la visualización, donde la información se representa codificada por colores (Kaski, 1997; Vesanto, 2002). En el siguiente capítulo se revisará el mapa auto-organizado en detalle, junto con otras técnicas que tienen por objetivos la reducción de la dimensionalidad, la cuantificación de vectores y la visualización.

Capítulo 2

Técnicas de minería de datos en la supervisión de energía eléctrica

En este capítulo se realiza una revisión del conjunto de técnicas presentes en la literatura que se utilizan en el proceso de minería de datos y cuyos objetivos principales son la reducción de la dimensionalidad, la cuantificación de vectores, el agrupamiento o *clustering* y la visualización de los datos en un espacio de salida. Se hará especial énfasis en aquellos métodos que se han utilizado en la supervisión y análisis de la energía eléctrica. Inicialmente, se revisa ampliamente el mapa auto-organizado (SOM), ya que esta tesis gira entorno a él. También se citan algunas variantes y algoritmos similares. A continuación, se explican los métodos de proyección tradicionales, ya sean de tipo lineal o no lineal. Posteriormente, se describen algunos de los métodos novedosos para el aprendizaje de variedades topológicas o *manifolds*. Finalmente, se exponen los principales métodos de agrupamiento.

2.1. El mapa auto-organizado (SOM)

2.1.1. Descripción del SOM

El mapa auto-organizado (*Self-Organizing Map*, SOM) (Kohonen, 1982, 2001) es una red neuronal no supervisada, basada en un proceso de aprendizaje competitivo y cooperativo, que realiza una proyección desde un espacio de entrada, normalmente de alta dimensionalidad, a un espacio de salida de baja dimensionalidad, que se corresponde con una malla o retícula discreta y topológicamente ordenada. La estructura básica de la red SOM se representa en la figura 2.1. Normalmente esta malla de salida es bidimensional, lo que facilita la visualización de los datos de entrada, aunque también se podrían utilizar mallas de salida de una o tres dimensiones. Cada neurona o unidad de la malla de salida está conectada a todos los nodos situados en el espacio de entrada y se describe por medio de su vector de pesos o vector prototipo (*codebook vector*) $k-$dimensional $\mathbf{m}_i$ del espacio de entrada y su posición $\mathbf{g}_i$ en la malla de baja dimensionalidad del espacio de salida (ver figura 2.2). Los vectores prototipo cuantifican los vectores de entrada $\mathbf{x}$, dividiendo el

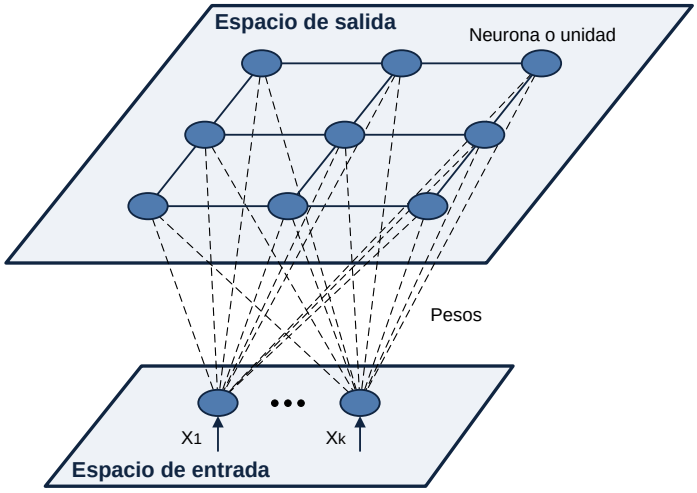


Figura 2.1: Esquema del mapa auto-organizado (SOM).

espacio en una colección finita de regiones de Voronoi[1], mientras que las coordenadas de las neuronas preservan la topología de la distribución de los datos de entrada, es decir, los vectores cercanos en el espacio de entrada mantienen la vecindad en el espacio de salida. Las neuronas interaccionan lateralmente con sus adyacentes por medio de la función de vecindad h_{ci}.

El proceso de aprendizaje necesario para llevar a cabo la proyección requiere dos etapas. En primer lugar, se selecciona la neurona ganadora $c(t)$ (*Best Matching Unit*, BMU), es decir, aquella que mejor representa el vector de entrada $\mathbf{x}(t) = [x_1, x_2, \ldots, x_k] \in \mathbb{R}^k$. La búsqueda de la neurona ganadora se realiza por medio de un proceso competitivo según la ecuación 2.1, que permite seleccionar aquella neurona cuya distancia al vector de entrada es la menor.

$$c(t) = \arg\min_i \|\mathbf{x}(t) - \mathbf{m}_i(t)\|,\ i = 1, 2, \ldots, M \tag{2.1}$$

M es el número de neuronas, t indica el instante de tiempo y $\|^\circ\|$ representa una medida de distancia entre el vector de entrada $\mathbf{x}(t)$ y los vectores prototipo $\mathbf{m}_i(t)$, típicamente la norma Euclídea que viene definida por la ecuación 2.2.

$$\|\mathbf{x}(t) - \mathbf{m}_i(t)\| = \sqrt{\sum_k [x_k(t) - m_{ik}(t)]^2} \tag{2.2}$$

En la segunda etapa tiene lugar un proceso cooperativo y adaptativo mediante el cual los pesos de las neuronas ganadoras y sus vecinas se adaptan en función del nuevo vector de entrada según la ecuación 2.3.

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t) h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)] \tag{2.3}$$

$\alpha(t)$ es la velocidad o tasa de aprendizaje y $h_{ci}(t)$ es la función de vecindad, generalmente una función Gaussiana de la distancia entre las coordenadas de las neuronas en el espacio

[1]La región de Voronoi de un vector prototipo $\mathbf{m}_i$ contiene a todos los vectores de entrada $\mathbf{x}$ que se encuentran más próximos a dicho vector prototipo y no a otro.

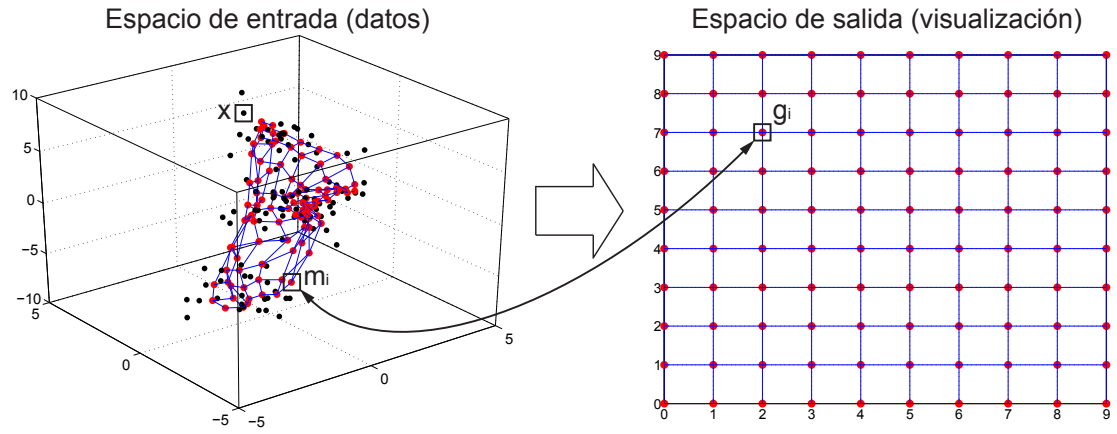


Figura 2.2: Proyección del mapa auto-organizado (SOM).

de salida, que viene definida por la ecuación 2.4.

$$h_{ci}(t) = exp\left(-\frac{\|\mathbf{g}_i - \mathbf{g}_c\|^2}{2\sigma(t)^2}\right) \tag{2.4}$$

Aunque la función de vecindad más típica es la Gaussiana, se pueden utilizar otros tipos de funciones de vecindad para modificar la región de influencia que un vector de entrada tiene en la malla del SOM, como por ejemplo la función burbuja, la función *cut-Gauss* o la función Epanechikov (Vesanto *et al.*, 2000). Cualquier función de vecindad debe satisfacer los siguientes requisitos:

- Su mayor valor se encuentra en la neurona ganadora $c(t)$, para la que la distancia $\|\mathbf{g}_i - \mathbf{g}_c\|$ es cero.
- Decrece a cero cuando la distancia tiende a infinito.

$\sigma(t)$ es el radio de vecindad, que varía en el tiempo. Esta variación puede ser de tipo exponencial (Ritter *et al.*, 1992), expresada por una de las ecuaciones 2.5.

$$\begin{cases} \sigma(t) = \sigma_0 \cdot exp\left(-\frac{t}{\tau_1}\right) & t = 0, 1, 2, \ldots \\ \\ \sigma(t) = \frac{a}{1+d\tau} \end{cases} \tag{2.5}$$

σ_0 es el valor inicial del radio de vecindad, τ_1 y τ son constantes de tiempo y a y d son parámetros numéricos.

La relación de vecindad entre neuronas depende del tipo de malla, que puede ser rectangular o hexagonal. La forma de la malla puede influir en la facilidad de la red elástica formada por los vectores prototipo $\mathbf{m}_i$ para orientarse y estabilizarse de acuerdo a la función de densidad de probabilidad $p(x)$. Por otra parte, la topología del mapa también influye en la vecindad, ya que neuronas situadas en los extremos de la malla tienen menos vecinos que las neuronas del centro del mapa. Esto provoca el efecto borde, es decir, que las neuronas

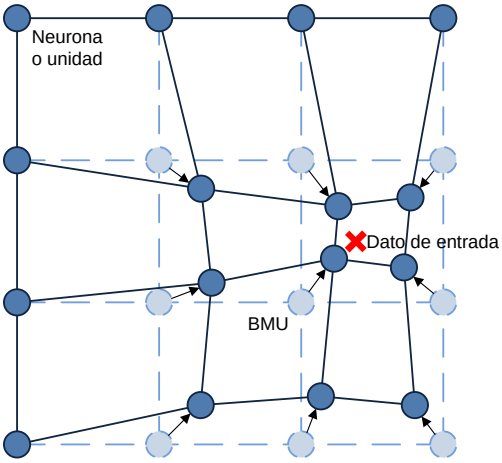


Figura 2.3: Esquema del proceso de adaptación en el mapa auto-organizado (SOM).

de los extremos sólo son atraídas hacia los nodos del interior de la malla. Para evitar este efecto no deseado, se pueden emplear otro tipo de topologías como la cilíndrica o la toroidal (Ultsch, 2003a), de modo que no existan extremos y todas las neuronas tengan el mismo número de vecinos.

Los parámetros que rigen la adaptación (la velocidad de aprendizaje $\alpha(t)$ y el radio de vecindad $\sigma(t)$) son críticos para el éxito o fracaso de la proyección y, en general, su valor disminuye con el tiempo (*annealing*). No obstante, para el ajuste de los mismos solamente se cuenta con criterios heurísticos. Asimismo, el número de neuronas de la red o escala del modelo afecta a la capacidad de precisión y generalización, que son objetivos contrapuestos. Un SOM con gran número de neuronas permite una regresión más precisa, mientras que un SOM con pocas neuronas logra una mayor generalización (Alhoniemi *et al.*, 1999). Además, es conveniente que las dimensiones de la red se adecúen a la distribución de los datos.

El proceso de entrenamiento crea una especialización de las neuronas a ciertas áreas del espacio de entrada. Para conseguirlo, las neuronas de la red se deben exponer a un número suficiente de patrones que garantice que el proceso de auto-organización se desarrolla adecuadamente (Haykin, 1994). La regla de aprendizaje arrastra a la neurona ganadora y a sus vecinas hacia cada nuevo vector de entrada, como una red flexible que se pliega sobre el conjunto de datos de entrada. En la figura 2.3 se muestra de forma esquemática este proceso. Las líneas discontinuas claras representan el estado original de la malla, mientras que las líneas continuas oscuras indican el estado posterior a la adaptación originada por la entrada de un nuevo dato. Esta es la razón por la que el mapa preserva la topología del espacio de entrada mediante su ordenación espacial, lo que proporciona una representación fiel de las características importantes de la entrada (Kohonen, 1990; Haykin, 1994). Los vectores k-dimensionales de las neuronas de la red se modifican de tal manera que en las zonas donde hay una mayor densidad de datos se localiza una cantidad mayor de neuronas. Es decir, las regiones inducidas en el espacio de visualización son tanto mayores cuanto más intenso es el estímulo al que representan, por lo que se obtiene una cierta preservación de la distribución

de probabilidad de entrada. A lo largo del entrenamiento se distinguen claramente dos fases:

1. **Fase de auto-organización u ordenación**: es la primera fase del proceso y debe ser relativamente corta. Tanto la tasa de aprendizaje $\alpha(t)$ como el radio de vecindad $\sigma(t)$ deben tomar inicialmente valores relativamente altos y luego, disminuir gradualmente. El objetivo es que la vecindad incluya inicialmente casi todas las neuronas, disminuyendo lentamente (Kangas, 1994).

2. **Fase de convergencia**: es la fase final que tiene por objetivo afinar el mapa. Para ello, la tasa de aprendizaje $\alpha(t)$ y el radio de vecindad $\sigma(t)$ deben tomar valores pequeños y decrecientes, pero sin llegar a cero. Esta fase puede emplear la mayor parte del tiempo de computación.

Dependiendo del tipo de entrenamiento del SOM, se pueden distinguir dos algoritmos diferentes:

- **Entrenamiento secuencial**: el SOM se entrena iterativamente, de forma que en cada paso se introduce un único vector del conjunto de datos de entrada elegido aleatoriamente y se calculan las distancias entre este y todos los vectores prototipo para seleccionar la neurona ganadora según la ecuación 2.1. A continuación, se lleva a cabo la adaptación por medio de la ecuación 2.3. Este proceso se repite con el resto de vectores del conjunto de datos de entrada y un determinado número de veces o épocas para facilitar la convergencia del mapa. El principal inconveniente de este tipo de entrenamiento es el elevado tiempo requerido cuando el número de datos y/o neuronas es elevado. El entrenamiento secuencial es adecuado para su ejecución en tiempo real.

- **Entrenamiento por lotes o *batch***: el SOM se entrena iterativamente un número de épocas, de forma que todos los vectores que constituyen el conjunto de datos de entrada se introducen simultáneamente al algoritmo para calcular las correspondientes neuronas ganadoras, utilizando la ecuación 2.6.

$$c(j) = \arg\min_i \|\mathbf{x}_j - \mathbf{m}_i(t)\|,\ i = 1, 2, \ldots, M,\ j = 1, 2, \ldots, N \tag{2.6}$$

  La adaptación se realiza mediante la media ponderada de los vectores de entrada, de forma que el nuevo peso de cada vector es la función de vecindad $h_{c(j)i}(t)$ correspondiente a su neurona ganadora $c(j)$. Para ello, se divide el conjunto de datos de entrada según las regiones de Voronoi establecidas por los vectores prototipo, es decir, que un vector de entrada pertenece al conjunto de vectores asociados a la neurona más próxima. La ecuación 2.7 define el proceso de adaptación que tiene lugar en el entrenamiento de tipo *batch*, que es mucho más eficiente computacionalmente y por lo tanto, más rápido.

$$\mathbf{m}_i(t+1) = \frac{\sum_{j=1}^{N} h_{c(j)i}(t)\mathbf{x}_j}{\sum_{j=1}^{N} h_{c(j)i}(t)} \tag{2.7}$$

Un proceso previo al entrenamiento y de vital importancia en la obtención de un resultado satisfactorio, es la inicialización de los pesos de las neuronas. Para ello, se pueden utilizar diferentes métodos:

- **Inicialización aleatoria**: los vectores de pesos de las neuronas toman inicialmente valores diferentes, pequeños y totalmente aleatorios (Haykin, 1994).
- **Inicialización con muestras iniciales**: los vectores de pesos toman valores correspondientes a alguna de las muestras de datos, por lo que se encuentran en la misma zona del espacio que los datos de entrada (Hollmén, 1996).
- **Inicialización lineal**: los vectores de pesos se inicializan de forma ordenada y lineal, a lo largo de los mayores autovectores (Kohonen, 2001).

La convergencia del SOM a un estado estacionario solamente está probada para el caso de una dimensión (Cottrell *et al.*, 1998) y la complejidad computacional del algoritmo es $\mathcal{O}(NMK)$, donde K es la dimensión de los datos de entrada, M el número de neuronas y N el número de datos (Vesanto *et al.*, 2000).

El mapa auto-organizado presenta una serie de propiedades que lo hacen idóneo para la visualización y exploración de grandes volúmenes de datos. A continuación, se citan estas propiedades del algoritmo SOM:

- **Aproximación de los datos de entrada**: los vectores prototipo $\mathbf{m}_i$ proporcionan una buena aproximación de los vectores de entrada $\mathbf{x}$, de forma similar a la cuantificación de vectores (VQ) (Luttrell, 1989). Por tanto, el error cometido al sustituir cada vector de entrada por el vector prototipo más próximo, denominado error de cuantificación (Kohonen *et al.*, 1996), es mínimo. No obstante, la función de vecindad impone una restricción en la representación de los datos por vectores prototipo, ya que impide la ruptura de la estructura topológica de los datos.
- **Aproximación de la función de densidad de los datos**: el SOM refleja variaciones en la distribución de probabilidad de los datos de entrada $p(x)$, ya que las regiones con muchos datos se representan con una mayor resolución, es decir, un mayor número de neuronas en el espacio de salida. El factor de magnificación describe la relación entre los datos y la densidad de vectores prototipo en la salida. No obstante, el SOM tiende a sobrerepresentar regiones de baja densidad y subrepresentar regiones de alta densidad (Haykin, 1994).
- **Preservación de la topología de los datos**: de forma intuitiva, la preservación de la topología indica similitudes en la estructura de vecindad de los conjuntos de datos de entrada y salida. Es decir, que neuronas cercanas en la malla de salida, se corresponden con vectores prototipo también próximos en el espacio de entrada. Las proyecciones que transforman los datos en las coordenadas de sus neuronas BMU y las coordenadas en los datos deben ser continuas con respecto a cada topología (Villmann *et al.*, 1997). Para preservar la topología, el SOM lleva a cabo una proyección local

correcta, pero no optimiza la proyección de forma global como otros métodos (Kaski, 1997).

- **Reducción de la dimensionalidad**: es una consecuencia directa de las propiedades anteriores. La capacidad de aproximar la geometría de la función de densidad de los datos, conservando a la vez la topología definida por una malla de salida de baja dimensionalidad (2D), implica que el mapa resultante puede resumir en las posiciones de la malla 2D todas las características del espacio de entrada de alta dimensionalidad. Además, al mismo tiempo que se reduce el número de vectores, se realiza una proyección de los mismos a la malla de salida.

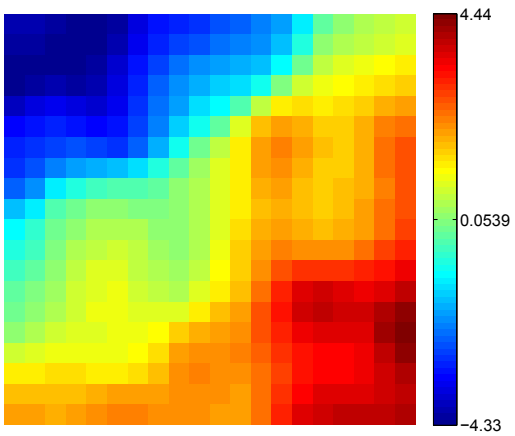
- **Capacidad de visualización**: el SOM proporciona una visualización compacta y ordenada de la información debido a que los vectores se proyectan en un espacio delimitado y fijo en la malla de salida (Kaski, 1997). Además, el uso de una representación homogénea facilita y agiliza la interpretación de la información. El SOM permite visualizar la estructura de grupos o *clusters* de los datos de entrada utilizando las distancias entre vectores prototipo. Aunque dicho análisis visual es subjetivo e implica riesgos, no es necesario suponer de antemano el número y forma de los *clusters*, como en muchos algoritmos específicos de agrupamiento. La combinación de un algoritmo SOM y uno de escalado multidimensional proporciona mejores resultados en la visualización (Flexer, 2001). Las cualidades del SOM hacen que proporcione fiabilidad (*Trustworthiness*) en la visualización de los datos de entrada (Venna, 2007).

Uno de los principales inconvenientes del SOM radica en la ausencia de una función de energía, lo que dificulta el análisis y la optimización del algoritmo (Erwin *et al.*, 1992). La definición de BMU, que determina la probabilidad de asignación de una muestra a una neurona, no es una parte intrínseca de la función de error, lo que impide obtener una función de energía válida. Por otra parte, no se puede garantizar la convergencia del algoritmo cuando la dimensión de la malla de salida es mayor de uno.

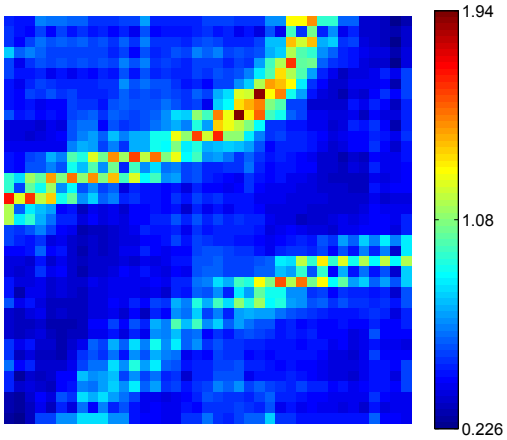
2.1.2. Visualización basada en SOM

Como se comentó en el punto anterior, una de las propiedades del SOM es su capacidad para visualizar la información sobre el espacio reticular de salida que es fijo, homogéneo y topológicamente ordenado. Por medio de un código de color, se pueden mostrar propiedades escalares o variables del espacio de entrada, distancias entre neuronas, etc. A continuación, se presentan brevemente las principales herramientas de visualización basadas en el SOM que se pueden encontrar en la bibliografía (ver figura 2.4).

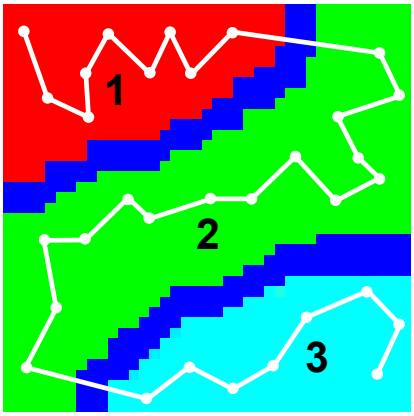
Plano de componentes. Este mapa, conocido también como **mapa de características**, permite describir variables o componentes del vector de entrada mediante los valores que toman sus vectores prototipo (Tryba *et al.*, 1989). Los valores escalares de la componente se representan mediante una escala de colores para cada neurona (ver figura 2.4a).



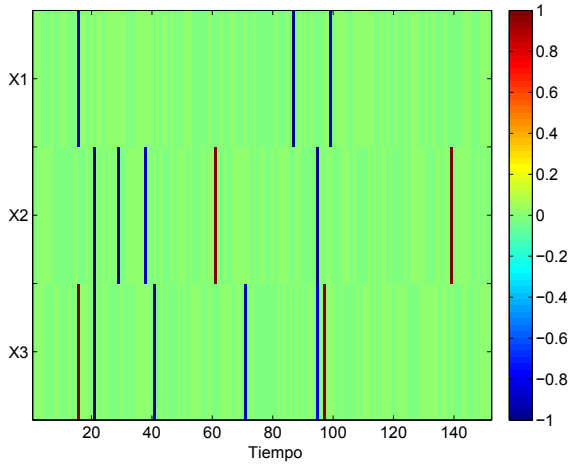
(a) Plano de componentes



(b) Matriz de distancias



(c) Proyección de la trayectoria sobre el mapa de estados



(d) Visualización de los residuos

Figura 2.4: Ejemplos de herramientas de visualización basadas en el mapa auto-organizado (SOM).

Existen tantos planos de componentes como variables de entrada. Comparando planos de componentes se puede detectar si dos componentes están correlacionadas.

Matriz de distancias. También conocida como matriz de distancias unificada (***u-matrix***) (Ultsch y Siemon, 1990), es un mapa que representa por medio de colores la distancia promedio de cada neurona a sus vecinas, según la función de vecindad. La distancia promedio interneuronal guarda una estrecha relación con la densidad de neuronas en una región determinada del espacio de entrada, por lo que este mapa permite descubrir visualmente la estructura de *clusters* de los datos. Densidades elevadas (distancias pequeñas) representan regiones muy pobladas de datos, es decir, *clusters*. Al contrario, las zonas con baja densidad (distancias grandes) pueden verse como separadores de grupos. En la figura 2.4b se muestra un ejemplo de la matriz de distancias, donde se pueden visualizar

claramente 3 grupos. En la literatura se puede encontrar métodos automáticos de etiquetado de los grupos sobre este mapa (Domínguez, 2003). Existen algunas variantes de la matriz de distancias como por ejemplo la ***p-matrix*** (Ultsch, 2003b) que asocia a cada neurona una estimación de la densidad de datos y la ***u*-matrix*** (Ultsch, 2005) que tiene en cuenta la densidad de datos para representar las distancias.

Mapa de estados. Estos mapas se construyen etiquetando o coloreando las neuronas de acuerdo al grupo al que pertenecen (Fuertes *et al.*, 2005). De forma similar que la matriz de distancias, los mapas de estados permiten visualizar claramente los grupos de datos. Cada uno de ellos indica un estado o situación del sistema objeto de análisis.

Proyección de la trayectoria. La BMU correspondiente al vector de datos actual se puede proyectar sobre la malla reticular para indicar la situación o punto de operación de un proceso. Conectando la secuencia de puntos de operación a lo largo del tiempo, se genera una trayectoria de la evolución del proceso sobre la malla (Kasslin *et al.*, 1992), a partir de la cual es posible determinar qué *clusters* o condiciones de proceso son accesibles desde una dada y cuál es la probabilidad de transición entre ellas (Fuertes *et al.*, 2007). Los vectores definidos por la trayectoria se pueden utilizar para construir modelos dinámicos de evolución del sistema (Fuertes *et al.*, 2007). En la figura 2.4c se muestra un ejemplo de la trayectoria sobre un mapa de estados.

Visualización de los residuos. El concepto de residuo se puede relacionar con el error de cuantificación (Kohonen *et al.*, 1996) entre un vector de entrada y el vector prototipo correspondiente a la BMU de un SOM entrenado (ver ecuación 2.8).

$$\mathbf{r}(t) = \mathbf{x}(t) - \mathbf{m}_c(t). \tag{2.8}$$

Los residuos se pueden utilizar en la detección de desviaciones o fallos. Cabe mencionar que la utilización de información sobre la evolución dinámica del proceso puede mejorar el cálculo de los residuos (Fuertes, 2006). El residuo $\mathbf{r}(t)$ tiene k componentes que representan las desviaciones individuales de cada variable respecto a sus valores esperados. La representación gráfica del residuo vectorial en el tiempo es muy útil cuando la dimensión del espacio de entrada es alta (Díaz y Hollmén, 2002). Los valores de las desviaciones se codifican por medio de diferentes colores, dependiendo de si es positiva (rojo) o negativa (azul). En la figura 2.4d se puede observar la representación del residuo correspondiente a 3 componentes a lo largo del tiempo.

Histograma. Este mapa, conocido también como **mapa *hit***, representa la frecuencia de repetición de las neuronas ganadoras (BMUs), es decir, la proporción de datos que codifica cada neurona del mapa (Vesanto, 1999).

Mapa de etiquetas. Este mapa muestra identificadores de los datos de entrada en el lugar en que se proyecta su BMU asociada. Etiquetas significativas o clave pueden resultar

útiles para analizar el mapa (Rauber y Merkl, 1999).

Mapa de modelos. Estos mapas permiten evaluar el grado de cumplimiento de un modelo analítico para cada neurona del espacio de visualización. El modelo puede ser expresado por una ecuación explícita (Díaz *et al.*, 2005) o por reglas borrosas, en cuyo caso se denomina **mapa borroso** (Cuadrado, 2002).

Mapa de activación. Esta representación permite conocer que zona del espacio de visualización corresponde a un conjunto de datos de entrada, asemejándose a su función de densidad (Díaz *et al.*, 2002b).

Mapa de correlación. Este mapa es una representación de las correlaciones locales entre variables de entrada en el espacio de visualización (Díaz *et al.*, 2002a). Estos mapas ayudan a detectar relaciones no lineales entre las variables.

Mapa de diferencias. Estos mapas permiten la comparación entre procesos que se rigen por el mismo patrón de funcionamiento, con el fin de conocer las desviaciones existentes entre ambos (Fuertes, 2006). Para ello, se representa la diferencia entre los vectores prototipo del mapa del proceso B y los vectores prototipo de las neuronas ganadoras correspondientes a esos vectores en el mapa del proceso A. Se puede añadir información acerca de las transiciones para mejorar la visualización, obteniendo en este caso el mapa de diferencias dinámico.

Mapa dinámico de transiciones. Este mapa combina el mapa de estados con las probabilidades de transición de un estado a otro, que se pueden representar mediante flechas con un grosor proporcional (Fuertes *et al.*, 2007). Esta visualización facilita la supervisión en línea de un proceso y permite detectar fallos de forma sencilla cuando los cambios de estado producidos tienen una probabilidad muy baja o nula.

2.1.3. Variantes y algoritmos similares al SOM

Muchos han sido los algoritmos similares y extensiones del SOM propuestos en la literatura desde la aparición de las primeras variantes (Kangas *et al.*, 1990). Debido a la potencia y versatilidad del algoritmo SOM para cuantificar vectores, reducir la dimensionalidad y visualizar estructuras de datos, numerosos autores han tratado de ampliar su funcionalidad, mejorar sus prestaciones o eliminar sus limitaciones. Por ejemplo, es posible definir nuevas funciones de vecindad, modificar el proceso de búsqueda de la BMU y/o el de adaptación, utilizar diferentes métricas, optimizar el tiempo de entrenamiento, definir de forma dinámica la estructura del mapa, construir mapas jerárquicos, introducir información a priori acerca de las clases, etc. Una visión general de las principales ideas para modificar el SOM básico se puede encontrar en Kohonen (2001). A continuación, se presentan algunas de las principales variantes y algoritmos similares al SOM.

Soft Topographic Vector Quantization (STVQ). Este variante del SOM tradicional tiene origen probabilístico e intenta paliar el problema encontrado a la hora de optimizar el algoritmo básico debido a la ausencia de una función de energía sobre la que se pueda aplicar un descenso de gradiente (Graepel *et al.*, 1998). Para ello, se realiza un cambio en la definición de la neurona ganadora, de forma que esta tome parte en la función de error. Existe una generalización de este algoritmo, denominado **Kernel-based Soft Topographic Mapping** (STMK), que utiliza funciones de núcleo (Graepel *et al.*, 1998). Asimismo, se han propuesto algoritmos que reproducen exactamente la distribución de probabilidad de los datos de entrada, es decir, que poseen un factor de magnificación igual a la unidad. Como ejemplos, podemos citar **Infomax** (Linsker, 1989) y **Winner Relaxing SOM** (Claussen, 2005).

SOM Supervisado. A diferencia del SOM tradicional, que es no supervisado, el SOM supervisado utiliza información acerca de la identidad de las clases en el proceso de aprendizaje (Kohonen, 2001). Por tanto, los vectores de entrada constan de dos partes bien diferenciadas, los datos y un identificador de la clase a la que pertenecen. La segunda parte se elimina para visualizar los resultados. Esta variante del SOM mejora la precisión en tareas de clasificación de datos ya que la segunda parte de los vectores de entrada es idéntica para todos aquellos que pertenecen a la misma clase (Melssen *et al.*, 2006). También se ha aplicado en el reconocimiento de estructuras de datos (Hagenbuchner y Tsoi, 2005).

Tree-Structured SOM (TS-SOM). Este algoritmo introduce el concepto de búsqueda jerárquica de la neurona ganadora y utiliza varios SOMs tradicionales estructurados en forma de pirámide con el fin de reducir el tiempo de entrenamiento (Koikkalainen y Oja, 1990). El entrenamiento comienza en el nivel más alto de la pirámide donde se construye un SOM con una sola neurona. Cuando finaliza el entrenamiento, su vector prototipo se mantiene fijo y se continua de forma similar en los niveles inferiores donde los SOMs construidos poseen cada vez un mayor número de neuronas. La búsqueda de la neurona ganadora se lleva a cabo dentro de los vectores prototipo descendientes situados en la misma rama de la estructura de árbol. También se incluyen en la búsqueda los vecinos a la neurona ganadora situados en los niveles superiores para mantener la vecindad. El **PIC-SOM** (Laaksonen *et al.*, 2000) utiliza varios TS-SOMs separados para extraer imágenes similares a una de referencia de las bases de datos. Las salidas de los TS-SOMs se combinan automáticamente en paralelo en función de las preferencias del usuario.

SOM por capas. Este variante del SOM tradicional tiene como objetivo mejorar el análisis exploratorio de los datos, el cual en muchas ocasiones resulta ineficaz porque las especificaciones de la red SOM son inadecuadas para una determinada estructura de los datos de entrada. Para ello, se entrena un SOM dinámico por cada capa siguiendo una jerarquía. Cada SOM adapta su estructura y sus dimensiones en función de las necesidades, dedicando mayor número de neuronas donde se proyectan más datos de entrada (Dittenbach *et al.*, 2000). Las principales ventajas obtenidas con este tipo de algoritmos que utilizan

SOMs dinámicos son la reducción del tiempo de entrenamiento debido al concepto de búsqueda jerárquica por capas, la habilidad para descubrir estructuras de datos jerárquicas y la visualización detallada de los grupos de datos (Rauber *et al.*, 2002).

SOM temporal. El SOM se puede emplear en el procesamiento de secuencias temporales para explotar la información dinámica de los datos. Para ello, el SOM necesita incorporar algún mecanismo de memoria a corto plazo que capture la información del contexto, como por ejemplo las líneas de retardo, los integradores *leaky*, la recurrencia o mecanismos de reacción-difusión (Guimarães *et al.*, 2003; Prada, 2009). Algunos algoritmos que implementan estas extensiones temporales son **RSOM** (Koskela *et al.*, 1998), **VQTAM** (Barreto y Araújo, 2001), **SOMTAD** (Principe *et al.*, 2002), **KSOM** (Barreto *et al.*, 2004), **Merge SOM** (Hammer *et al.*, 2004), etc.

Kernel Regression SOM (KR-SOM). Esta extensión constituye una versión continua o interpolada del SOM que intenta paliar las desventajas debidas al carácter discreto de este (Díaz *et al.*, 2001). La interpolación puede ser exacta mediante una red *Radial Basis Function* (RBF) (Park y Sandberg, 1991) o bien no exacta utilizando *General Regression Neural Network* (GRNN) (Specht, 1991). El principal problema del KR-SOM es la desigualdad existente entre la proyección directa (desde el espacio de entrada al de visualización) y la inversa (desde el espacio de visualización al de entrada).

SOM Parametrizado (PSOM). Este algoritmo es una extensión continua del SOM tradicional que muestra excelentes capacidades de generalización a partir de pequeños conjuntos de datos de entrenamiento (Walter *et al.*, 2000). El PSOM resuelve el problema de la desigualdad entre proyecciones del KR-SOM utilizando sucesivas iteraciones, lo que se traduce en un coste computacional relativamente elevado.

ViSOM. Este algoritmo tiene como objetivo mejorar la visualización en el espacio de salida. Para ello, restringe la contracción lateral del SOM, regularizando las distancias entre neuronas para preservarlas lo más posible (Yin, 2002). La malla generada permite una medida cuantitativa, directa y visual de las distancias en el mapa.

SOM hiperbólico. Este algoritmo utiliza un espacio de salida hiperbólico para la visualización, a diferencia del SOM tradicional que emplea uno Euclídeo (Helge y Ritter, 1999). Esto permite visualizar con precisión grandes estructuras jerárquicas de datos ya que la vecindad alrededor de un punto se incrementa de forma exponencial con su radio.

WEBSOM. Este algoritmo basado en el SOM tradicional permite organizar documentos de texto heterogéneos a través de mapas de visualización bidimensionales que contienen información significativa para su exploración y búsqueda (Honkela *et al.*, 1996). Los documentos relacionados aparecen próximos entre sí en el espacio de visualización.

Neural gas. Este algoritmo se asemeja al SOM, pero no existe la restricción en la vecindad impuesta por una malla fija de salida y el proceso de adaptación se basa en un criterio de ordenación de las distancias entre el dato de entrada y los vectores prototipo (Martinetz *et al.*, 1993). *Neural gas* se utiliza con espacios de entrada complejos cuando el objetivo es cuantificar o agrupar los datos de entrada, ya que la falta de un espacio de salida limita su uso para visualización.

Generative Topographic Mapping (GTM). Este método es semejante al SOM, pero tiene base estadística. El GTM es un modelo no lineal de variables latentes en el que se trata de encontrar una representación para una distribución de los datos en el espacio de entrada en función de las variables latentes (Bishop *et al.*, 1998). Dado que el modelo GTM se define como una transformación del espacio latente al de datos, es necesario invertir la proyección mediante el teorema de Bayes[2], lo que da lugar a la distribución a posteriori en el espacio latente, de la cual se suele representar su media. Si el objetivo es la visualización, se utilizan espacios de 2 dimensiones.

Manifold-SOM (M-SOM). Esta extensión del SOM tradicional es de tipo supervisado y tiene como objetivo la visualización eficiente de espacios de entrada con alta dimensionalidad en los cuales los datos están contenidos en una variedad topológica o *manifold* (Similä, 2007). La búsqueda de la neurona ganadora se lleva a cabo en base a las coordenadas del *manifold*, mientras que el proceso de actualización se realiza en función de las coordenadas del *manifold* y las del espacio de entrada.

Isotop. Este algoritmo similar al SOM está centrado básicamente en la reducción no lineal de la dimensionalidad de los datos. En el SOM tradicional esta tarea se encuentra íntimamente relacionada con la cuantificación de vectores, mientras que Isotop permite separar ambas (Lee *et al.*, 2003). No obstante, existe la posibilidad de cuantificar los datos cuando su número es elevado, evitando cualquier relación de vecindad entre ellos en esta primera etapa. En la siguiente etapa los datos vecinos se conectan por medio de un grafo. En la fase final, los vectores prototipo de alta dimensionalidad se transforman a coordenadas de baja dimensionalidad y se actualizan iterativamente siguiendo una regla similar al SOM. Isotop se puede considerar un método de *manifold learning*, los cuales se describirán más adelante en este capítulo.

2.1.4. El SOM en la supervisión y análisis de la energía eléctrica

Las aplicaciones del SOM son numerosas y variadas, como se desprende del gran número de artículos publicados relativos a este algoritmo (Oja *et al.*, 2003). El SOM se ha aplicado en áreas muy diversas, como por ejemplo la supervisión de procesos,

[2]El teorema de Bayes expresa la probabilidad condicional de un evento aleatorio A dado B en términos de la distribución de probabilidad condicional del evento B dado A y la distribución de probabilidad marginal de sólo A, es decir, $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$.

tratamiento de imágenes, organización de documentos, reconocimiento del habla, acústica, telecomunicaciones, robótica, biomedicina, neurofisiología, etc. (Kohonen, 2001).

Dentro del campo de la supervisión y análisis de la energía eléctrica, el SOM se ha utilizado principalmente para resolver dos tipos de problemas:

- **Clasificar a los consumidores en función de sus perfiles de carga**: en un mercado libre y competitivo, la clasificación de consumidores proporciona ventajas interesantes a las compañías de generación y distribución de energía eléctrica, como por ejemplo en la toma de decisiones estratégicas, adaptación del número y tipo de tarifas ofertadas, reducción del coste y aumento de la satisfacción del cliente, etc. (Chicco *et al.*, 2004; Figueiredo *et al.*, 2005). Asimismo, los propios consumidores de electricidad pueden emplear la información obtenida en la clasificación para cambiar su equipamiento o hábitos, de forma que disminuya la diferencia con el consumidor prototipo de su clase (Räsänen *et al.*, 2008). Generalmente, la agrupación de consumidores en clases es el resultado de un proceso de minería de datos sobre la gran cantidad de información que registran las compañías eléctricas (Sforna, 2000). El SOM es la base de la metodología de minería de datos y lleva a cabo un agrupamiento no supervisado, utilizando los perfiles diarios de potencia expresados en el tiempo (Chicco *et al.*, 2004; Verdú *et al.*, 2006), o bien en el dominio de la frecuencia para detectar comportamientos periódicos (Verdú *et al.*, 2006). Otras técnicas de agrupamiento clásicas (Figueiredo *et al.*, 2005; Räsänen *et al.*, 2008) o lógica borrosa (Sforna, 2000) se aplican con cierta frecuencia a partir de los vectores prototipo del SOM para mejorar la clasificación de consumidores.

- **Predecir el consumo eléctrico futuro en función de situaciones pasadas**: por otra parte, el mercado cambiante actual requiere una predicción precisa del consumo eléctrico a corto plazo (para el día siguiente) con el fin de ajustar la capacidad de generación, analizar la fiabilidad del equipamiento, planificar actividades extraordinarias, prever posibles penalizaciones por exceso de potencia, modificar el precio, etc. (Carpinteiro *et al.*, 2004; Fan *et al.*, 2006; Tafreshi y Farhadi, 2007). Sin embargo, la predicción del consumo eléctrico es una tarea compleja debido a las relaciones no lineales que existen entre la demanda de potencia y las condiciones meteorológicas, tipos de día, etc. (Beccali *et al.*, 2004; Fan *et al.*, 2006; Tafreshi y Farhadi, 2007). El SOM se emplea para construir modelos eléctricos basados en datos de situaciones pasadas que capturen estas relaciones no lineales entre variables. Algunos modelos de predicción explotan la memoria auto-asociativa del SOM (Lendasse *et al.*, 2002; Tafreshi y Farhadi, 2007), mientras que otros se basan en la capacidad del SOM para agrupar los datos en clases (Beccali *et al.*, 2004; Fan *et al.*, 2006). Generalmente, varios modelos individuales para cada posible situación (días de la semana, fiestas, vísperas, días posteriores, etc.) son preferibles a un único modelo para todas las situaciones conjuntas (Tafreshi y Farhadi, 2007). Además, la combinación del SOM con otras técnicas no lineales de reducción de la dimensionalidad (Lendasse *et al.*, 2002; Beccali *et al.*, 2004) o regresión (Fan *et al.*, 2006) proporciona

predicciones más precisas. Incluso la concatenación de dos SOMs consecutivos de forma jerárquica mejora los resultados de la predicción (Carpinteiro *et al.*, 2004).

2.2. Métodos tradicionales lineales

Los métodos tradicionales lineales empleados en la reducción de la dimensionalidad son adecuados para representar estructuras de datos sencillas que presentan relaciones de tipo lineal. Su objetivo es encontrar una combinación lineal de las variables de entrada que permita su visualización en un espacio de salida de menor dimensión. Estos métodos desempeñan una importante función en la minería de datos, ya que son fáciles de comprender, aplicar e interpretar, son muy eficientes computacionalmente y pueden servir como base para obtener representaciones más complejas. Algunas de las aplicaciones de los métodos tradicionales lineales en el campo de la supervisión y análisis de energía eléctrica se centran en la predicción a corto plazo de la demanda de electricidad utilizando una regresión basada en las componentes principales que trata de capturar las variaciones en la demanda durante el día (Taylor *et al.*, 2006). Otro tipo de aplicaciones hacen uso de las componentes independientes para estimar los perfiles de carga en redes eléctricas sin necesidad de conocer sus parámetros y topología (Liao y Niebur, 2003).

2.2.1. Análisis de componentes principales (PCA)

El análisis de componentes principales (*Principal Component Analysis*, PCA) (Hotelling, 1933), también denominado transformación de Karhunen-Loéve, consiste en la proyección lineal de los vectores $\mathbf{x}_i$ del espacio de entrada en un espacio de salida de menor dimensionalidad. Este espacio de salida tiene como base la matriz $\mathbf{U}$ que contiene los vectores propios correspondientes a los valores propios mayores de la matriz de covarianzas $\mathbf{\Sigma}$, llamados componentes principales (ver ecuaciones 2.9).

$$\begin{aligned} \mathbf{\Sigma} &= \tfrac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \\ \mathbf{\Sigma} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \\ \mathbf{y}_i &= \mathbf{U}^T\mathbf{x}_i \end{aligned} \tag{2.9}$$

$\overline{\mathbf{x}}$ es la media de los vectores y $\mathbf{\Lambda}$ es una matriz diagonal que contiene los valores propios. Puesto que la matriz de covarianzas $\mathbf{\Sigma}$ es real y simétrica, sus vectores propios son ortogonales. En la figura 2.5 se puede observar la proyección de un conjunto de datos en un espacio de entrada 3D a lo largo de los dos primeros autovectores (espacio de salida 2D).

Este método no supervisado de reducción de la dimensionalidad preserva aquellas direcciones en las cuales los datos poseen la mayor varianza (Alpaydin, 2004). A medida que se toman menos vectores propios como base del espacio de salida (mayor reducción de la dimensionalidad), el error de aproximación a los datos de entrada es mayor. PCA es un

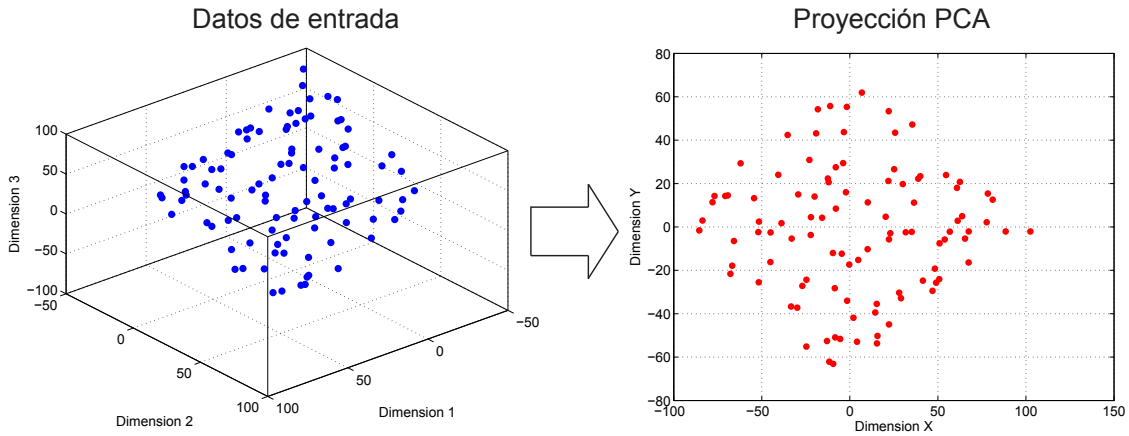


Figura 2.5: Proyección establecida por el análisis de componentes principales (PCA).

método sencillo y muy popular, que ha sido aplicado en multitud de campos de estudio, incluido el campo de la energía eléctrica (Ndiaye y Gabriel, 2011). PCA requiere que los datos de entrada pertenezcan a un subespacio lineal, lo que se convierte en su principal limitación, de modo que PCA no es útil cuando los datos presentan relaciones no lineales, dada su naturaleza estadística.

2.2.2. Projection pursuit

Projection pursuit (Friedman y Tukey, 1974) es un método no supervisado que implica encontrar una proyección lineal "interesante" de los datos multidimensionales de entrada. Para ello, se realiza la optimización de una cierta función objetivo, llamada índice de proyección. La proyección se considera "interesante" cuando los datos proyectados poseen alguna estructura que difiere de una distribución normal. Para reducir la dimensionalidad de los datos, se elimina la componente a lo largo de la proyección encontrada y se repite el proceso de búsqueda de nuevas proyecciones "interesantes". Este método se ve afectado en menor medida que otros por el problema de la maldición de la dimensionalidad, pero tiene como desventajas ser una técnica lineal, requerir una gran carga computacional y preservar *outliers*[3]. PCA es un caso particular de *projection pursuit* cuando el índice de proyección es la varianza de los datos.

2.2.3. Análisis de componentes independientes (ICA)

El análisis de componentes independientes (*Independent Component Analysis*, ICA) (Comon, 1994) es un método que intenta revelar factores ocultos que subyacen en un conjunto de datos, de forma que las proyecciones lineales sean estadísticamente lo más independientes posibles. Se supone que los datos son una mezcla lineal de las variables latentes que son desconocidas, al igual que la matriz o sistema de mezcla (Hyvärinen *et al.*,

[3]Muestras de datos que poseen un valor fuera de rango, es decir, muy dispares del resto.

2001). Se asume que las variables latentes, denominadas componentes independientes o factores de los datos, son mutuamente independientes y no Gaussianas. ICA puede encontrar una nueva codificación más interesante de los datos que revele ciertos patrones. Este método es más potente que PCA para encontrar factores subyacentes en los datos, ya que la condición de independencia es más fuerte que la de correlación. ICA se puede considerar como caso particular de *projection pursuit*.

2.3. Métodos tradicionales no lineales

Los métodos tradicionales no lineales empleados en la reducción de la dimensionalidad son adecuados para representar estructuras de datos contenidos en subespacios fuertemente curvados y grupos con formas arbitrarias. La visualización de datos que presentan relaciones no lineales y tienen una alta dimensionalidad es compleja cuando se emplean métodos de proyección lineal. En cambio, los métodos tradicionales no lineales producen visualizaciones sencillas, basándose principalmente en la preservación de las distancias mutuas entre los datos. Es decir, tratan de representar los datos de entrada en un espacio de salida de menor dimensión, de modo que las distancias entre datos sean tan similares como sea posible a las distancias originales. Estos métodos difieren en la ponderación de las distancias y en la optimización de las representaciones. Como ejemplos de aplicaciones de los métodos tradicionales no lineales en el campo de la supervisión y análisis de energía eléctrica, podemos citar la construcción del regresor y la estimación de su dimensión óptima, es decir, el número mínimo de variables necesarias, para predecir futuras demandas de potencia (Lendasse *et al.*, 2002) y la reducción del volumen de datos que las compañías eléctricas llevan a cabo de forma previa a la clasificación de sus clientes (Chicco *et al.*, 2006).

2.3.1. Generalizaciones no lineales del PCA

Las transformaciones lineales (giros y desplazamientos) no siempre son útiles a la hora de proyectar un conjunto de datos debido a las estructuras complejas no lineales de los datos y a la optimización en la separación de grupos. Los métodos no lineales se adaptan mejor a estas estructuras complejas de los datos de entrada. Como ya se comentó anteriormente, el principal defecto de los métodos lineales y concretamente del PCA, es que no son útiles cuando los datos presentan relaciones no lineales. Para resolver este problema, se han propuesto algunas generalizaciones no lineales del algoritmo PCA que se indican a continuación.

El algoritmo **PCA de núcleo** (*Kernel* PCA) (Schölkopf *et al.*, 1997) tiene como principal ventaja que convierte el problema de autovalores en productos escalares, evitando de este modo un proceso de optimización no lineal. Los *kernels* permiten llevar a cabo el cálculo sin necesidad de realizar explícitamente la transformación al espacio de alta dimensión. Este método evita mínimos locales, pero en cambio es muy dependiente del *kernel* utilizado. Si el conjunto de datos es elevado, este algoritmo requiere un esfuerzo computacional alto.

Otro tipo de generalización no lineal de PCA son las **curvas principales** (Hastie y Stuetzle, 1989). Cada punto que forma una curva principal es el promedio de los datos para los cuales dicho punto es el más cercano en la curva. Por tanto, las curvas principales son aquellas que pasan por el centro de un conjunto de datos k-dimensional, proporcionando un resumen no lineal de ese conjunto. El concepto se puede extender a dimensiones o variedades más altas y entonces reciben el nombre de superficies o variedades principales. Las curvas principales guardan gran similitud con el SOM (Mulier y Cherkassky, 1995).

Un **perceptrón multicapa** (*Multilayer Perceptron*, MLP) se comporta de forma similar al PCA cuando se utiliza de forma autoasociativa (Kramer, 1991). Para ello, la red debe poseer una capa de entrada y de salida con un número de neuronas igual a la dimensión de los datos de entrada y la capa oculta lineal debe tener un número de neuronas menor. El entrenamiento se lleva a cabo con el algoritmo tradicional de *backpropagation*, pero presentando cada patrón como la entrada y salida deseadas. Se pueden introducir otras capas ocultas con una función de activación no lineal antes y después de la capa lineal. MLP posee cierta tendencia a acabar en mínimos locales.

2.3.2. Escalado multidimensional (MDS)

Los métodos de escalado multidimensional (*Multidimensional Scaling*, MDS) tienen como objetivo representar los datos de entrada, preservando su distancia mutua. Para ello, llevan a cabo un proceso de minimización de una función de coste, de forma que las distancias en el espacio de baja dimensión de salida representen las disimilitudes del espacio de entrada. La matriz de distancias mutuas se puede calcular a partir de los datos, o bien puede proceder de la evaluación de disimilitudes entre objetos.

Existen diversos tipos de algoritmos que se incluyen dentro del escalado multidimensional. Uno de los primeros en aparecer en la literatura fue el **MDS clásico**, MDS lineal o escalado clásico (Torgerson, 1952). El MDS clásico es un algoritmo de tipo lineal, similar al PCA, que se resuelve como un problema de autovalores. Sin embargo, actualmente el término MDS es comúnmente asociado a métodos de proyección no lineal ya que aportan mejores resultados.

Otro tipo de algoritmo es el denominado **MDS métrico** (Kruskal y Wish, 1978), que utiliza la función de coste indicada en la ecuación 2.10.

$$E = \sum_i \sum_{j \neq i} (X_{ij} - Y_{ij})^2 \tag{2.10}$$

$X_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ e $Y_{ij} = d(\mathbf{y}_i, \mathbf{y}_j)$ son las matrices de distancias mutuas de los datos de entrada $\mathbf{x} \in \mathbb{R}^n$ y los de salida $\mathbf{y} \in \mathbb{R}^p$, respectivamente. El MDS métrico se centra en preservar largas distancias, lo que resulta en una buena conservación de la estructura global.

El **MDS no métrico** (Kruskal y Wish, 1978) es otro tipo de algoritmo de escalado multidimensional, motivado por la necesidad de tratar datos ordinales. La matriz de distancias mutuas contiene información cualitativa acerca de las similitudes o disimilitudes entre objetos. En este caso, la función de coste a optimizar se modifica introduciendo

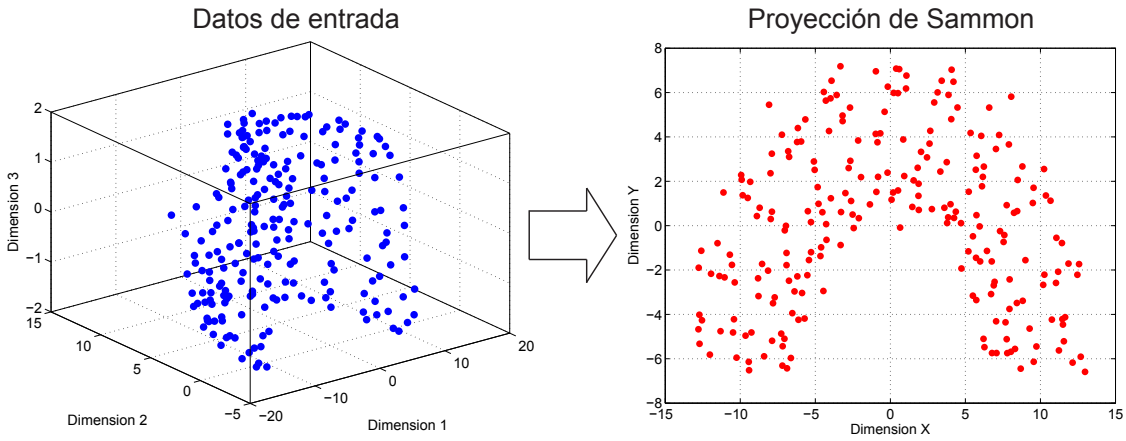


Figura 2.6: Ejemplo de la proyección de Sammon (*Sammon's mapping*).

una función monótona creciente que actúa sobre las distancias originales, de forma que la proyección preserve el orden de distancias entre vectores, pero no su valor absoluto.

Todos estos métodos son intuitivos pero presentan algunas desventajas como su alta carga computacional, tendencia a converger a mínimos locales o la necesidad de que los datos estén distribuidos de forma uniforme sobre su variedad.

2.3.3. Proyección de Sammon

La proyección de Sammon (*Sammon's mapping*) (Sammon, Jr., 1969) es un tipo de algoritmo MDS que utiliza una función de coste normalizada en función de las distancias en el espacio de entrada, que se expresa según la ecuación 2.11.

$$E = \frac{1}{\sum_i \sum_{j<i} X_{ij}} \sum_i \sum_{j<i} \frac{(X_{ij} - Y_{ij})^2}{X_{ij}} \tag{2.11}$$

Al igual que otros métodos MDS, la proyección de Sammon intenta preservar las distancias relativas entre los datos, de forma que la estructura de los datos en el espacio de baja dimensión sea lo más similar posible a la de los datos de entrada. La introducción del factor de distancias originales $X_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ en el denominador de la función de coste, permite dar más énfasis a las distancias pequeñas. De esta forma, se preserva la topología local de los datos, ya que la proximidad normalmente refleja similitud o pertenencia a un grupo. A modo de ejemplo, en la figura 2.6 se muestra la proyección 2D de un conjunto de datos que forman parte de una estructura semi-toroidal en 3D.

Las coordenadas en el espacio de salida se determinan minimizando la función de coste E expresada en la ecuación 2.11. Este proceso de minimización se lleva a cabo de forma iterativa, aplicando técnicas de optimización estándar, como por ejemplo el método cuasi-Newton. En la versión *batch* del algoritmo de minimización, los pasos seguidos son los siguientes:

1. Calcular todas las distancias mutuas en el espacio de entrada $X_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$. Este paso requiere un gran esfuerzo computacional cuando el conjunto de datos es grande.

2. Inicializar las coordenadas $y_{k(i)}$ de los datos en el espacio de salida $\mathbf{y}_i$. En esta etapa, se puede emplear el método PCA para hallar las componentes principales del conjunto de datos o bien, proporcionar valores aleatorios. Hay que tener en cuenta que el resultado final puede variar en función del tipo de inicialización.

3. Calcular el término $\Delta y_{k(i)}(t) = \frac{\left(\frac{\partial E}{\partial y_{k(i)}}\right)}{\left(\frac{\partial^2 E}{\partial {y_{k(i)}}^2}\right)}$ para las coordenadas de todos los datos en el espacio de salida $\mathbf{y}_i$.

4. Actualizar las coordenadas $y_{k(i)}$ de todos los datos en el espacio de salida $\mathbf{y}_i$ de acuerdo a la ecuación 2.12.

$$y_{k(i)}(t+1) = y_{k(i)}(t) - \alpha \Delta y_{k(i)}(t) \tag{2.12}$$

   El parámetro α se denomina factor mágico y se recomienda que sus valores oscilen entre 0,3 y 0,4.

5. Repetir los pasos 3 y 4 hasta que el valor de la función de error E no disminuya más.

La proyección de Sammon no incluye un estimador de la dimensión intrínseca, la cual es impuesta por el usuario y tampoco posee la habilidad de generalizar la proyección de nuevos datos. Por tanto, este método debe ser ejecutado de forma separada para cada dimensión especificada y/o cada conjunto de datos.

Este método ofrece un buen compendio entre la robustez y la repetibilidad, es simple, discreto y no lineal, lo que facilita un manejo eficiente de estructuras de datos ligeramente curvadas. Además, la proyección de Sammon se puede combinar con métodos de cuantificación de vectores, como por ejemplo el SOM, cuando el número de datos es elevado. En este caso, se logra una proyección más rápida ya que el SOM realiza una sustancial compresión de los datos, sin pérdida significativa de información. La proyección de Sammon se aplica a los vectores *codebook* procedentes del SOM, en lugar de a los datos originales (Simula *et al.*, 1999; König, 2000). Lo expuesto anteriormente justifica su elección y uso en el desarrollo de esta tesis.

2.3.4. Análisis de componentes curvilíneas (CCA)

El análisis de componentes curvilíneas (*Curvilinear Component Analysis*, CCA) (Demartines y Hérault, 1997) es un método variante del MDS, que minimiza una función de coste centrada en las distancias entre puntos en el espacio de salida, definida por la ecuación 2.13.

$$E = \frac{1}{2} \sum_i \sum_{j \neq i} (X_{ij} - Y_{ij})^2 F(Y_{ij}, \lambda_Y) \tag{2.13}$$

Generalmente, F es una función acotada y monótona decreciente, tal y como la que se indica en la ecuación 2.14.

$$F(Y_{ij}, \lambda_Y) = \begin{cases} 1 & si \quad Y_{ij} \leq \lambda_Y \\ 0 & si \quad Y_{ij} > \lambda_Y \end{cases} \tag{2.14}$$

El parámetro λ_Y decrece a lo largo del entrenamiento y permite controlar la escala a la que trabaja el algoritmo.

CCA es un algoritmo de menor complejidad computacional que otros métodos MDS, pues en cada etapa se fija un vector $\mathbf{y}_i$, adaptándose el resto $\mathbf{y}_j$. De este modo, en cada ciclo de adaptación no es necesario calcular $N(N-1)/2$ distancias, sino solamente las distancias entre el dato i y los demás. CCA permite la interpolación y extrapolación de nuevos datos. Para ello, se utiliza la misma función de coste, pero solamente se calcula con respecto al punto $\mathbf{y}_0$ correspondiente al dato $\mathbf{x}_0$. El resto de datos se mantienen fijos en la adaptación. CCA permite, al igual que el SOM, realizar una proyección no lineal y una cuantificación de vectores, y además posee buenas características para la visualización (Venna y Kaski, 2006). Asimismo, tiene la ventaja de que el espacio de salida es continuo y se puede adaptar mejor a la estructura topológica de los datos.

Existen varias modificaciones del algoritmo CCA, como por ejemplo el **análisis de distancias curvilíneas** (*Curvilinear Distance Analysis*, CDA) (Lee *et al.*, 2000), que utiliza distancias geodésicas[4] en lugar de Euclídeas, o el **escalado multidimensional local** (*Local Multidimensional Scaling*, Local MDS) (Venna y Kaski, 2006), que permite parametrizar la fiabilidad y continuidad de la visualización.

2.4. Métodos de manifold learning

Los métodos de *manifold learning* tratan de representar las variedades topológicas (*Manifolds*) existentes en los datos de entrada (Cayton, 2005). Para ello, asumen que los datos pertenecen a un subespacio no lineal de baja dimensión que se encuentra dentro de un espacio de entrada, cuya dimensionalidad es normalmente elevada. El objetivo de estos métodos es descubrir y desplegar este subespacio topológico o *manifold* para su correcta visualización e interpretación. Lo ideal sería que la dimensionalidad del espacio de salida se correspondiese con la dimensionalidad intrínseca o latente del subespacio. Sin embargo, la tarea de visualización impone una dimensionalidad de salida 2D ó 3D, lo que puede causar problemas en este tipo de métodos (Venna, 2007). Las aplicaciones de los métodos de *manifold learning* en el campo de la supervisión y análisis de energía eléctrica son escasas debido, probablemente, a que estos métodos son relativamente recientes. No obstante, existe un creciente interés por parte de los investigadores en explotar la estructura intrínseca de la variedad topológica que contiene la distribución de los datos. Como ejemplo, se puede citar su empleo satisfactorio en el modelado y predicción del precio de la electricidad (Chen *et al.*, 2008).

[4]Distancia del camino mínimo a lo largo de una variedad topológica que une dos puntos pertenecientes a la misma.

En la figura 2.7a se muestra un ejemplo de *manifold* muy típico en la literatura (Lee y Verleysen, 2007; Cayton, 2005), denominado '*Swiss roll*', que consiste en un conjunto de puntos discretos situados sobre una capa que se va enrollando sobre sí misma. El color de los puntos denota el radio creciente de la espiral generada. Este *manifold*, cuya dimensionalidad intrínseca es 2D, está incrustado dentro de un espacio 3D. Por tanto, los métodos de *manifold learning* deberían ser capaces de encontrar un subespacio 2D que permita "desenrollar" la estructura, evitar superposiciones de los sucesivos giros de la espiral y obtener una proyección biyectiva entre el espacio de entrada y el de salida. En la figura 2.7 [5] se pueden observar las proyecciones del *manifold* de referencia '*Swiss roll*', utilizando algunos de los diferentes métodos de *manifold learning* que se describirán a continuación.

2.4.1. Isometric Feature Mapping (ISOMAP)

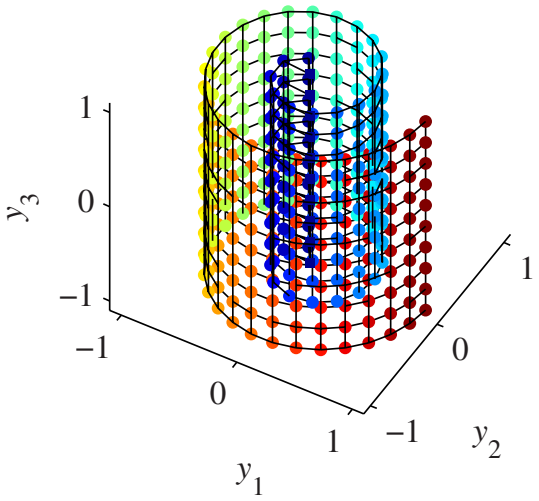
Isometric Feature Mapping (ISOMAP) (Tenenbaum *et al.*, 2000) es un método que se puede considerar como una variante del MDS, ya que emplea las distancias geodésicas en lugar de las Euclídeas a la hora de proyectar los datos. Se denomina distancia geodésica al camino mínimo a lo largo de un *manifold*, que une dos puntos pertenecientes al mismo. El uso de la distancia Euclídea en el escalado multidimensional ignora la existencia de variedades topológicas, ya que dos puntos cercanos en el espacio de entrada podrían estar situados muy lejos dentro de un *manifold* de baja dimensionalidad definido por los datos. En este caso, la distancia geodésica puede ser más útil. ISOMAP intenta preservar la topología de los datos de forma global, a la vez que las distancias mutuas entre puntos. Este método consta de 3 etapas principales:

1. Construir los grafos con los k vecinos más próximos para todos los puntos.
2. Calcular la rutas más cortas utilizando el algoritmo de Dijkstra o de Floyd.
3. Realizar una proyección MDS al espacio Euclídeo de baja dimensionalidad.

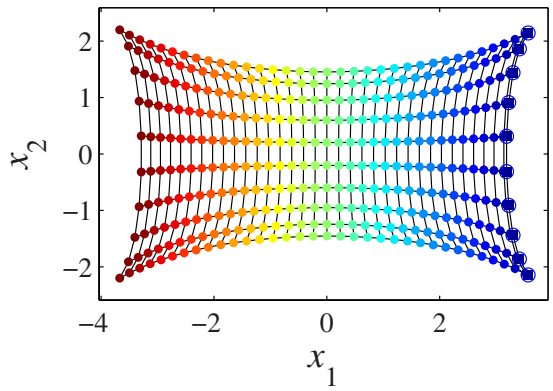
El objetivo de las dos primeras etapas es obtener las distancias geodésicas entre todos los pares de puntos de datos, las cuales se suponen lineales cuando estos se encuentran cercanos. MDS realiza la optimización global a partir de la matriz de distancias mínimas. En la figura 2.7b se puede observar el resultado de aplicar este método al *manifold* de referencia '*Swiss roll*'. ISOMAP proporciona una estimación de la dimensionalidad del *manifold* por medio del número de autovalores distintos de cero encontrados por el algoritmo. Sin embargo, la escasez de datos de entrenamiento, la dependencia de la vecindad y la complejidad del algoritmo dificultan el correcto aprendizaje de la variedad topológica.

Se han propuesto algunas modificaciones del ISOMAP con el fin de mejorar la representación de variedades y la eficiencia computacional (De Silva y Tenenbaum, 2003). **Conformal ISOMAP** (C-ISOMAP) es capaz de aprender la estructura de variedades topológicas curvadas. Por otra parte, **Landmark ISOMAP** (L-ISOMAP) permite reducir la gran carga computacional global requerida.

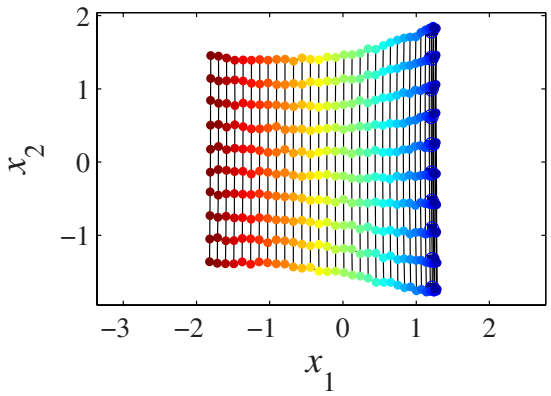
[5] Fuente: (Lee y Verleysen, 2007).



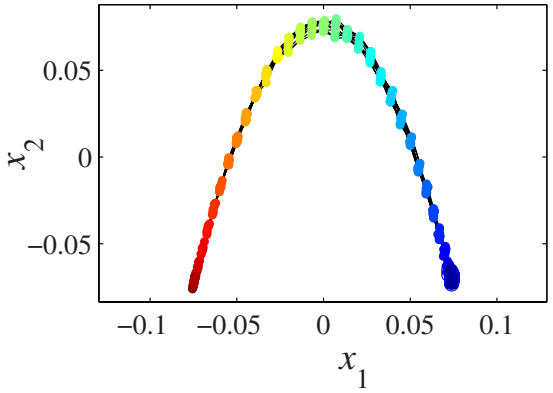
(a) Manifold de referencia: '*Swiss roll*'



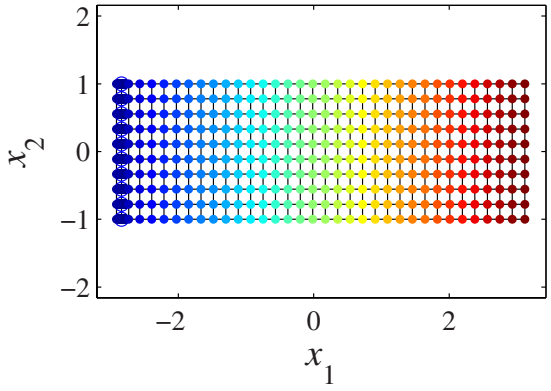
(b) Proyección con ISOMAP



(c) Proyección con LLE



(d) Proyección con LE



(e) Proyección con SDE

Figura 2.7: Proyecciones 2D de la estructura '*Swiss roll*' obtenidas con métodos de *manifold learning*.

2.4.2. Locally Linear Embedding (LLE)

Locally Linear Embedding (LLE) (Roweis y Saul, 2000) es un método que permite proyectar datos contenidos en variedades topológicas complejas a partir de sus aproximaciones localmente lineales, ya que supone que cada dato y sus vecinos se encuentran en un subespacio lineal de la variedad o muy próximos a ella cuando el muestreo del *manifold* es uniforme. LLE elimina la necesidad de calcular las distancias relativas entre puntos de datos, a diferencia de los métodos MDS e ISOMAP presentados anteriormente. La preservación de la topología se realiza de forma local, mediante las relaciones de vecindad en los datos. Este método consta de 3 etapas:

1. Asignar los vecinos a cada punto de datos de entrada $\mathbf{x}_i$ usando, por ejemplo, los k vecinos más próximos.

2. Calcular los pesos w_{ij} que mejor reconstruyen de forma lineal los datos $\mathbf{x}_i$ a partir de sus vecinos. Para ello, se minimiza el error de reconstrucción dado por la ecuación 2.15 utilizando mínimos cuadrados con las restricciones $w_{ij} = 0$ (si el dato $\mathbf{x}_j$ no es vecino de $\mathbf{x}_i$) y $\sum_j w_{ij} = 1$.

$$\varepsilon(\mathbf{W}) = \sum_i \left\| \mathbf{x}_i - \sum_j w_{ij}\mathbf{x}_j \right\|^2 \tag{2.15}$$

3. Realizar una proyección en un espacio de baja dimensionalidad, de forma que los datos de salida $\mathbf{y}_i$ mejor reconstruidos por w_{ij}, minimicen la función de coste indicada en la ecuación 2.16.

$$\Phi(\mathbf{Y}) = \sum_i \left\| \mathbf{y}_i - \sum_j w_{ij}\mathbf{y}_j \right\|^2 \tag{2.16}$$

   Este problema se puede resolver mediante un cálculo de autovalores con la restricción que $\mathbf{Y}$ tenga media cero y covarianza unitaria.

En la figura 2.7c se puede observar el resultado de aplicar este método al *manifold* de referencia '*Swiss roll*'.

El método **Hessian LLE** (hLLE) (Donoho y Grimes, 2003) es una modificación del método LLE basada en el marco teórico del método Laplacian Eigenmap, el cual se presentará en el punto siguiente, donde se sustituye el estimador Laplaciano[6] del grafo por uno Hessiano[7]. hLLE aplica PCA al conjunto de datos vecinos de $\mathbf{x}_i$ para estimar su espacio tangente, el cual se proyecta linealmente. De forma similar, el método **Local Tangent Space Alignment** (LTSA) (Zhang y Zha, 2004) estima primeramente los espacios tangentes. En el siguiente paso, se realiza un alineado de los mismos, minimizando una

[6]Denominada también matriz de admitancia o matriz de Kirchhoff, la matriz Laplaciana se define como $L = D - A$, donde D es la matriz de grados de los vértices del grafo y A es la matriz de adyacencia del grafo.

[7]Matriz cuadrada definida por las derivadas parciales de segundo orden de una función.

función de coste para encontrar las coordenadas globales de los puntos en el espacio de baja dimensionalidad.

2.4.3. Laplacian Eigenmap (LE)

El método *Laplacian Eigenmap* (LE) (Belkin y Niyogi, 2001) es similar al algoritmo LLE, pues también considera solamente los k vecinos más cercanos, pero difiere en que la relación de vecindad es simétrica entre pares de datos. La justificación geométrica de este algoritmo se basa en la correspondencia entre el Laplaciano de un grafo, el operador Laplace-Beltrami de una variedad y las conexiones a la función de difusión. LE preserva localmente la topología de los datos y tiene una conexión natural al agrupamiento o *clustering*. Este método consta de 3 etapas principales:

1. Construir el grafo adyacente con los k vecinos más próximos o utilizando un radio de vecindad ϵ, de forma que dos nodos i y j se conectan por medio de una arista si $\|\mathbf{x}_i - \mathbf{x}_j\|^2 < \epsilon$.

2. Elegir los pesos de las aristas utilizando una función de difusión de núcleo de la forma $W_{ij} = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t}\right)$ cuando los nodos i y j estén conectados y $W_{ij} = 0$ en caso contrario. Otra forma simple de elegir los pesos evitando el uso del parámetro t es $W_{ij} = 1$ si los nodos i y j están conectados por una arista y $W_{ij} = 0$ en caso contrario.

3. Obtener los *eigenmaps* resolviendo la ecuación 2.17.

$$\mathbf{L}\mathbf{y} = \lambda \mathbf{D}\mathbf{y}, \tag{2.17}$$

   $\mathbf{D}$ es una matriz diagonal y sus elementos son sumas de $\mathbf{W}$ $\left(D_{ii} = \sum_j W_{ij}\right)$. La matriz Laplaciana $\mathbf{L} = \mathbf{D} - \mathbf{W}$ es simétrica, positiva y semidefinida. La solución son los autovectores correspondientes a los autovalores más pequeños distintos de cero.

En la figura 2.7d se puede observar el resultado de aplicar este método al *manifold* de referencia '*Swiss roll*'.

2.4.4. Stochastic Neighbor Embedding (SNE)

Stochastic Neighbor Embedding (SNE) (Hinton y Roweis, 2002) es un método probabilístico que permite la proyección de datos, o bien de sus disimilitudes, preservando las probabilidades de que los puntos de datos sean vecinos. Para ello, se centran funciones Gaussianas en cada punto del espacio de entrada y se determina la distribución de probabilidad sobre todos los potenciales vecinos. El objetivo es encontrar una configuración de puntos que aproxime esas distribuciones en el espacio de salida. Este método consta de 4 etapas principales:

1. Seleccionar los vecinos mediante los k vecinos más próximos o utilizando un radio de vecindad ϵ.

2. Calcular las probabilidades p_{ij} y q_{ij}. La primera de ellas (p_{ij}) viene dada por la ecuación 2.18 y representa la probabilidad de que el dato $\mathbf{x}_i$ escoja a $\mathbf{x}_j$ como su vecino.

$$p_{ij} = \frac{\exp\left(-d_{ij}^2\right)}{\sum_{k \neq i} \exp\left(-d_{ik}^2\right)} \tag{2.18}$$

   d_{ij}^2 indica las disimilitudes entre dos puntos de datos de entrada $\mathbf{x}_i$ y $\mathbf{x}_j$, que se pueden calcular por medio de la distancia Euclídea escalada, es decir, $d_{ij}^2 = \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2}$, donde σ_i es el ancho del núcleo Gaussiano. La segunda (q_{ij}) representa la probabilidad inducida en el espacio de baja dimensionalidad de que el punto $\mathbf{y}_i$ escoja a $\mathbf{y}_j$ como su vecino y viene dada por la ecuación 2.19.

$$q_{ij} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{y}_i - \mathbf{y}_k\|^2\right)} \tag{2.19}$$

3. Minimizar la función de coste indicada en la ecuación 2.20, la cual es la suma de las divergencias de Kullback-Leibler[8] entre las probabilidades p_{ij} y q_{ij}. El objetivo de la proyección es igualar ambas probabilidades lo máximo posible.

$$\Im = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.20}$$

4. Proyectar y actualizar los datos en el espacio de baja dimensionalidad a través de un descenso de gradiente según la ecuación 2.21.

$$\mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} - \eta^{(k)} \nabla \Im^{(k)} \tag{2.21}$$

   $\eta^{(k)}$ es la tasa de aprendizaje y el gradiente $\nabla\Im$ viene dado por las derivadas parciales de primer orden de $\Im$ respecto a todos los puntos $\mathbf{y}_i$.

Una modificación del SNE, denominada **t-Distributed Stochastic Neighbor Embedding** (t-SNE) (Van der Maaten y Hinton, 2008), ha sido propuesta para mejorar la optimización de la función de coste y aliviar las fuerzas de atracción entre datos cuando existen muchos cercanos. t-SNE utiliza la función de coste simétrica a SNE con gradientes más simples y una distribución t-Student[9] en lugar de una Gaussiana para calcular las similitudes entre puntos de datos en el espacio de baja dimensionalidad.

2.4.5. Semidefinite Embedding (SDE)

El método *Semidefinite Embedding* (SDE) (Weinberger y Saul, 2004), denominado también *Maximum Variance Embedding*, usa la programación semidefinida (Vandenberghe

[8] Medida de diferencia no conmutativa y no simétrica entre dos distribuciones de probabilidad.

[9] Distribución de probabilidad continua que surge cuando se estima la media de una población normalmente distribuida con una desviación estándar desconocida y un conjunto de muestras pequeño.

y Boyd, 1996) para encontrar una matriz de Gram o Gramiana[10] apropiada, tal que los datos y sus vecinos en el subespacio de baja dimensionalidad son localmente isométricos a sus correspondientes en el espacio de entrada, dada una relación de vecindad. Este método consta de 3 etapas principales:

1. Calcular los k vecinos más próximos a cada dato de entrada y construir el grafo que conecta cada punto a sus vecinos y estos entre sí.

2. Calcular la matriz de Gram $\mathbf{K}$ de la proyección de máxima varianza que preserva las distancias de todos los bordes del grafo de vecindad, es decir, que maximiza las distancias relativas entre los puntos no conectados al grafo. Mediante programación semidefinida, se maximiza la función objetivo $Tr(\mathbf{K})$, es decir, la traza[11] de la matriz de Gram, de forma que se maximiza la varianza de los puntos de salida. Las restricciones impuestas para todos los vecinos de $\mathbf{x}_i$ y $\mathbf{x}_j$ se indican en la expresión 2.22.

$$\begin{cases} \mathbf{K} \geq 0 \\ \\ \sum_{ij} \mathbf{K}_{ij} = 0 \\ \\ \mathbf{K}_{ii} + \mathbf{K}_{jj} - \mathbf{K}_{ij} - \mathbf{K}_{ji} = \mathbf{G}_{ii} + \mathbf{G}_{jj} - \mathbf{G}_{ij} - \mathbf{G}_{ji} \end{cases} \tag{2.22}$$

   $\mathbf{K}_{ij} = \mathbf{y}_i \cdot \mathbf{y}_j$ es la matriz de Gram de salida, mientras que $\mathbf{G}_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j$ es la matriz de Gram de entrada.

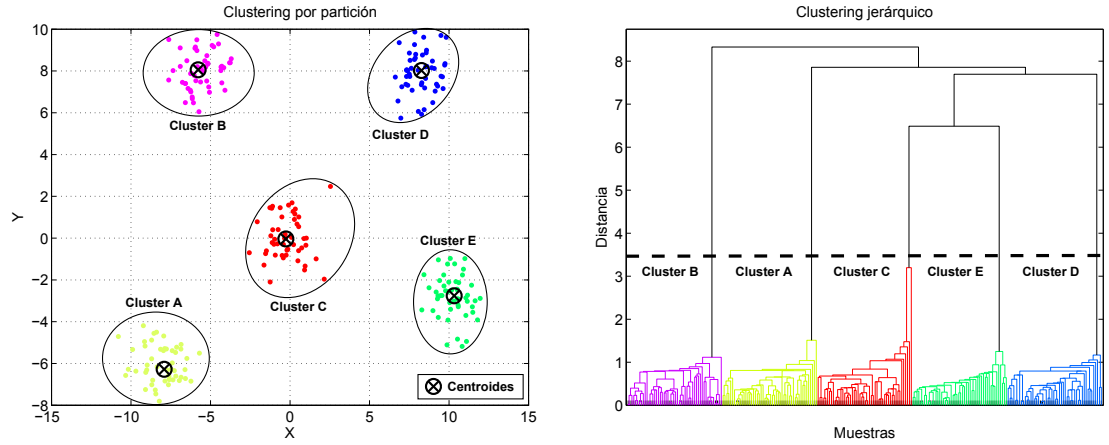
3. Encontrar la proyección a un espacio de baja dimensionalidad a partir de los autovectores dominantes de la matriz de Gram $\mathbf{K}$. Para ello, se calcula la descomposición espectral de la matriz como en el método MDS, es decir, $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ y las coordenadas del espacio de salida vienen dadas por $\mathbf{Y} = \mathbf{U}\mathbf{\Lambda}^{1/2}$.

En la figura 2.7e se puede observar el resultado de aplicar este método al *manifold* de referencia '*Swiss roll*'.

2.5. Métodos de agrupamiento

Los métodos de agrupamiento o *clustering* tienen como objetivo asignar o clasificar de forma no supervisada los datos de entrada en diferentes grupos o *clusters* que contienen datos de entrada similares (Jain *et al.*, 1999). Para ello, se detecta el número óptimo de grupos y se determina la pertenencia de todos los datos a los grupos encontrados en función de alguna medida de similitud. El agrupamiento puede ser muy útil en el análisis exploratorio de datos para el reconocimiento de patrones y clasificación de objetos. Las aplicaciones de los métodos de agrupamiento en el campo de la supervisión y análisis de energía eléctrica se centran en el reconocimiento de patrones de consumo diario de los consumidores, con el

[10] Matriz que define el producto escalar entre vectores.
[11] Suma de los elementos de la diagonal principal de una matriz cuadrada.



(a) Método por partición

(b) Método jerárquico

Figura 2.8: Tipos de métodos de agrupamiento.

fin de identificar y clasificar aquellos perfiles de carga más representativos y modificar las tarifas eléctricas (Chicco *et al.*, 2006; Tsekouras *et al.*, 2008). A continuación, se describirán los dos tipos básicos de métodos de agrupamiento que se pueden encontrar en la bibliografía: por partición y jerárquicos (ver figura 2.8).

2.5.1. Métodos por partición

Los métodos de agrupamiento por partición dividen el conjunto de datos de entrada en un número de grupos determinado, de forma que se minimiza una de las funciones de coste o criterios de dispersión locales o globales indicados en la ecuaciones 2.23.

$$\begin{cases} D_W = \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \mathbf{c}_k)(\mathbf{x}_i - \mathbf{c}_k)^T \\ D_B = \sum_{k=1}^{K} |C_k| (\mathbf{c}_k - \mathbf{c})(\mathbf{c}_k - \mathbf{c})^T \\ D_T = D_W + D_B = \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{c})(\mathbf{x}_i - \mathbf{c})^T \end{cases} \tag{2.23}$$

D_W mide la dispersión de los datos dentro de su propio grupo, D_B tiene en cuenta la dispersión de los datos en *clusters* diferentes y D_T es una medida de la dispersión total de los datos. k es el número de grupos, $\mathbf{c}_k$ es el promedio de los datos contenidos en el grupo C_k y $\mathbf{c}$ es el promedio de todos los datos de entrada.

El criterio de dispersión D_W dado en el ecuación 2.23, se puede expresar también como la suma de errores cuadráticos entre un vector de datos $\mathbf{x}_i$ y el centroide del grupo más próximo $\mathbf{c}_k$, es decir, el vector prototipo que mejor representa y caracteriza al conjunto

de datos agrupados dentro del mismo *cluster*. Por tanto, los vectores prototipo separan los datos de entrada en regiones que contienen los vecinos más próximos, de forma similar a la cuantificación de vectores (VQ), proporcionando una representación compacta y reducida de los datos de entrada.

El método de agrupamiento por partición más divulgado en la bibliografía es el **método *k-means*** (MacQueen, 1967). *k-means* busca de forma iterativa el número de grupos óptimo que realiza una partición de mínima varianza en el espacio de datos de entrada, minimizando la función de coste dada en la ecuación 2.24.

$$SSE = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2 \tag{2.24}$$

Este método consta de los siguientes pasos:

1. Inicializar todos los centroides $\mathbf{c}_k$ de forma aleatoria o bien aplicando conocimiento a priori.

2. Asignar cada dato de entrada $\mathbf{x}_i$ al grupo C_l más próximo, es decir, $\mathbf{x}_i \in C_l$ *si* $\|\mathbf{x}_i - \mathbf{c}_l\| < \|\mathbf{x}_i - \mathbf{c}_k\|,\ i = 1, 2, \ldots, N,\ k = 1, 2, \ldots, K,\ i \neq l$.

3. Calcular de nuevo los centroides, teniendo en cuenta la partición actual. Esto es, $\mathbf{c}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i$, donde $|C_k|$ representa el número de datos dentro del grupo k.

4. Repetir los pasos 2 y 3 hasta que los centroides no se modifiquen nunca más.

El método de Linde-Buzo-Gray (LBG) (Linde *et al.*, 1980) es muy similar al *k-means*, pero en lugar de las agrupaciones, busca los vectores prototipo que minimizan el error cuadrático medio de cuantificación, definido como la distancia entre un dato de entrada $\mathbf{x}_i$ y el vector prototipo que mejor lo representa $\mathbf{c}_k$. En la figura 2.8a se puede observar el resultado de aplicar el método *k-means* a un conjunto de puntos bastante separados.

Paradójicamente, el número de grupos es un parámetro que normalmente se debe proporcionar a este tipo de métodos. Este es uno de sus principales inconvenientes, ya que se necesita conocer de antemano el número de grupos, cuando es lo que se está buscando. Con el fin de obtener resultados aceptables, se puede ejecutar el algoritmo varias veces con diferentes inicializaciones y se selecciona la mejor configuración en base a alguna medida o índice de validación (Bezdek y Pal, 1998; Halkidi *et al.*, 2001). Otros problemas del método *k-means* son su limitación para separar grupos con formas no convexas y su gran coste computacional cuando el número de datos de entrada es elevado. Para paliar estos inconvenientes, se puede aplicar un SOM previo al método *k-means*, lo que mejora considerablemente el resultado del agrupamiento y reduce el tiempo empleado (Kuo *et al.*, 2002).

2.5.2. Métodos jerárquicos

Los métodos de agrupamiento jerárquico (Gordon, 1987) buscan construir una potencial estructura de grupos mediante una secuencia de particiones anidadas a partir de los datos

de entrada. Por tanto, se genera un árbol de grupos donde los datos se organizan siguiendo una estructura jerárquica, en función de algún criterio de enlace o vínculo (*Linkage*). Los criterios de enlace permiten caracterizar la disimilitud o similitud entre *clusters* mediante alguna medida de distancia. Los principales criterios se indican en las ecuaciones 2.25.

$$\begin{cases} \delta_{SL} = \min\left\{d(\mathbf{x}_i, \mathbf{x}_j)\right\},\ i \in C_k,\ j \in C_l \\ \\ \delta_{CL} = \max\left\{d(\mathbf{x}_i, \mathbf{x}_j)\right\},\ i \in C_k,\ j \in C_l \\ \\ \delta_{AL} = \frac{1}{|C_k||C_l|} \sum_{i \in C_k} \sum_{j \in C_l} d\left(\mathbf{x}_i, \mathbf{x}_j\right) \end{cases} \tag{2.25}$$

$d(\mathbf{x}_i, \mathbf{x}_j)$ es una medida de distancia entre los datos de entrada $\mathbf{x}_i$ y $\mathbf{x}_j$. C_k y C_l son dos grupos determinados, mientras que $|C_k|$ y $|C_l|$ indican el número de datos dentro del grupo k y l, respectivamente.

Según el criterio de enlace unitario (*Single Linkage*), la disimilitud entre dos grupos δ_{SL} viene dada por la distancia más pequeña entre todos los datos pertenecientes a ambos grupos. En cambio, en el criterio de enlace completo (*Complete Linkage*), la disimilitud entre dos grupos δ_{CL} viene determinada por la distancia más grande entre todos los datos pertenecientes a ambos grupos. Si se elige el criterio de enlace promedio (*Average Linkage*), la disimilitud entre dos grupos δ_{AL} se define como la distancia promedio entre todos los pares de datos de ambos grupos. Otros criterios de enlace se pueden revisar en Xu y Wunsch (2005).

El resultado del agrupamiento jerárquico se visualiza normalmente mediante un árbol binario o dendrograma, cuya raíz representa a todos los datos y las hojas a cada dato individual. La altura del dendrograma expresa la distancia entre datos o grupos. El número de grupos se obtiene "cortando" el dendrograma en un determinado nivel de acuerdo a algún índice de validación. En la figura 2.8b se puede observar el dendrograma obtenido a partir de un conjunto de puntos bastante separados.

El agrupamiento jerárquico es útil para la visualización de la estructura potencial de grupos cuando existen relaciones jerárquicas en los datos de entrada. Como inconvenientes se pueden destacar su gran coste computacional y su mayor sensibilidad a datos con ruido, respecto a los métodos por partición como el *k-means*.

Los métodos de agrupamiento jerárquico se pueden clasificar en dos tipos. El **agrupamiento jerárquico por aglomeración** busca aglutinar datos en grupos cada vez más grandes (sentido ascendente). Este método consta de los siguientes pasos:

1. Construir inicialmente tantos grupos como datos de entrada, de forma que cada grupo contenga un solo dato.

2. Calcular las disimilitudes o similitudes entre grupos $d\left(C_k, C_l\right)$, de acuerdo a algún criterio de enlace indicado en las ecuaciones 2.25.

3. Fusionar los 2 grupos más similares para construir un grupo mayor, situado más arriba en la jerarquía.

4. Repetir los pasos 2 y 3 hasta obtener un solo grupo.

Por el contrario, el **agrupamiento jerárquico por escisión** procede en sentido descendente, ya que inicialmente un solo grupo contiene a todos los datos de entrada. Este tipo de agrupamiento divide sucesivamente el grupo inicial en grupos más pequeños y los sitúa en niveles inferiores en la jerarquía, hasta obtener tantos grupos como número de datos. Estos métodos no son comunes en la práctica porque inicialmente consideran $2^{(N-1)} - 1$ posibles divisiones, lo que supone un coste computacional mayor que en el caso del agrupamiento jerárquico por aglomeración.

2.5.3. Otros métodos de agrupamiento

Se han propuesto multitud de métodos de agrupamiento en la literatura (Xu y Wunsch, 2005). Por ejemplo, se pueden citar algoritmos de agrupamiento borroso (*clustering fuzzy*), basados en la densidad, basados en núcleo, basados en redes neuronales, etc. Cabe destacar que el SOM es una red neuronal que se puede utilizar para descubrir y visualizar la estructura de *clusters* de un conjunto de datos, como se ha expuesto anteriormente. Además, se ha demostrado que la aplicación de métodos de agrupamiento de tipo aglomerativo o partitivo a partir de los vectores prototipo del SOM produce resultados comparables a la aplicación de los mismos métodos a partir de los datos originales, siendo el *clustering* de las neuronas del SOM considerablemente más rápido y estable (Vesanto y Alhoniemi, 2000).

Un método relativamente reciente y que produce resultados muy aceptables cuando la distribución de los datos no se asemeja a estructuras uniformes y bien separadas es el **agrupamiento espectral** (*Spectral clustering*) (Ng *et al.*, 2001). Este método proyecta los datos de entrada originales en un espacio donde los *clusters* son más obvios y a continuación aplica un método de agrupamiento tradicional, como por ejemplo el *k-means*. El agrupamiento espectral consta de las siguientes etapas:

1. Construir un grafo a partir de la matriz de similitud $\mathbf{S} \in \mathbb{R}^{n \times n}$ entre los datos de entrada $\mathbf{x}_i$.

2. Calcular la matriz Laplaciana $\mathbf{L}$ que puede ser normalizada o no.

3. Hallar los k primeros autovectores $\mathbf{u}_i,\ i = 1, 2, \ldots, k$ de la matriz $\mathbf{L}$.

4. Normalizar la matriz $\mathbf{U} \in \mathbb{R}^{n \times k}$ que contiene a los autovectores $\mathbf{u}_i,\ i = 1, 2, \ldots, k$, de forma que la norma sea la unidad. Como resultado se obtiene la matriz $\mathbf{T} \in \mathbb{R}^{n \times k}$, cuyas filas definen los puntos $\mathbf{y}_i \in \mathbb{R}^k$. Este paso no se tiene en cuenta cuando el agrupamiento es no normalizado.

5. Aplicar el método *k-means* a los puntos $\mathbf{y}_i$ para obtener los *clusters* $C_1, C_2, \ldots, C_k$.

Capítulo 3

Definición del sistema de medida y supervisión de la energía eléctrica

En este capítulo se plantea la necesidad de supervisar la energía eléctrica en los edificios del Campus de la Universidad de León, una vez conocido su sistema de suministro eléctrico y su distribución geográfica. A continuación, se define la arquitectura del sistema de medida y supervisión de la energía eléctrica en los edificios, la cual implementa un patrón cliente-servidor con una capa intermedia adicional. Esta arquitectura de triple capa se describe en detalle a lo largo de este capítulo. En la capa servidor, se localizan el sistema de medida formado por los medidores eléctricos distribuidos en cada edificio y una estación meteorológica común y el servidor de adquisición. Una red de comunicación basada en un protocolo estándar conecta entre sí todos los equipos de medida de esta capa. La capa intermedia está constituida por el servidor de almacenamiento, el servidor de explotación y el servidor web. Finalmente, los usuarios situados en la capa cliente pueden acceder al sistema de forma remota a través de Internet para llevar a cabo la supervisión.

3.1. Suministro de energía eléctrica a los edificios y necesidad de la supervisión

El Campus de la Universidad de León está formado por un conjunto de edificios, enumerados en la tabla 3.1, que se localizan principalmente en tres zonas: noroeste, centro y suroeste de la ciudad de León (ver figura 3.1). Como se puede observar en esta figura, la mayor parte de los edificios se encuentran ubicados en la zona noroeste, conocida como el Campus de Vegazana.

Estos edificios están destinados generalmente a uso docente e investigación, excepto aquellos que prestan servicios complementarios de apoyo, como por ejemplo las cafeterías, la sucursal bancaria, el centro de procesamiento de datos, etc. Todos estos edificios públicos de la Universidad de León se pueden encuadrar dentro del sector terciario debido a su utilidad. Aunque en principio puede parecer que todos los edificios tienen características similares, esto no es así ya que por ejemplo, la superficie útil de cada edificio es muy diferente. El principal consumo de energía eléctrica en este tipo de edificios se debe a

Número	Etiqueta	Edificio	Superficie útil (m^2)
1	E1	Centro de Procesamiento de Datos (CRAI-TIC)	4034
2	E2	Radio Universitaria	218
3	E3	Industriales e Informática	8470
4	E4	Aulario	2040
5	E5	Tecnológico	11500
6	E6	Filosofía y Letras	14700
7	E7	Derecho	8050
8	E8	Enfermería	4280
9	E9	Ciencias del Trabajo	4280
10	E10	Clínica Veterinaria	5315
11	E11	Veterinaria	17100
12	E12	Cafetería I	1470
13	E13	Ciencias Biológicas	12480
14	E14	Servicios	1800
15	E15	Desarrollo Ganadero y Sanidad Animal	2000
16	E16	Animalario	1245
17	E17	Ciencias de la Actividad Física y del Deporte	8475
18	E18	Pabellón Deportivo	3630
19	E19	Frontón	1090
20	E20	Biblioteca San Isidoro	5479
21	E21	Sucursal Bancaria	244
22	E22	Cafetería II	681
23	E23	Molecular	2680
24	E24	Complejo Agrícolas	10204
25	E25	Colegio Mayor	5500
26	E26	Complejo Rectorado	9020
27	E27	Complejo la Serna	1265
28	E28	Minas	4640
29	E29	Centro de Idiomas	6100
30	E30	Ciencias Económicas y Empresariales	11665

Tabla 3.1: Lista de edificios del Campus de la Universidad de León.

sus sistemas de iluminación, refrigeración y calefacción, ventilación y aire acondicionado (*Heating, Ventilating and Air Conditioning Systems*, HVAC) (Pérez-Lombard *et al.*, 2008). Estos sistemas engloban prácticamente la totalidad del consumo eléctrico en los edificios de la Universidad de León, excepto en aquellos donde existe un equipamiento de tipo industrial empleado en los diversos laboratorios de investigación.

La compañía suministra la energía eléctrica a los edificios del Campus desde 13 puntos de conexión a la red eléctrica de distribución, que se corresponden con los puntos de facturación. El esquema unifilar del sistema de suministro eléctrico a los edificios se puede observar en la figura 3.2. Los 30 edificios se agrupan en los diferentes puntos de facturación, de forma que algunos (P1) agrupan un único edificio (E1), mientras que otros (P2) aglutinan varios edificios (E3, E4, E5, E17, E18, E19). En el sistema de suministro eléctrico a los edificios se pueden distinguir dos tipos de puntos de facturación, dependiendo del nivel de tensión. Los edificios localizados en la zona noroeste (Campus de Vegazana) y en el suroeste de la ciudad se alimentan de la red eléctrica de distribución en media tensión de 21 KV. Por lo tanto, se necesita un transformador eléctrico para reducir la tensión hasta un valor adecuado para su

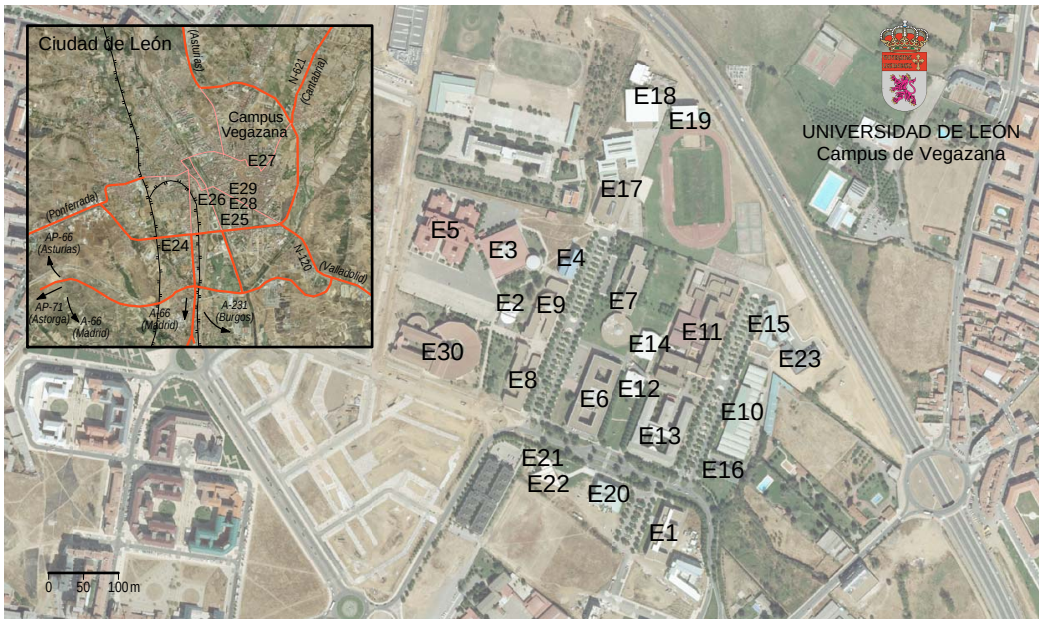


Figura 3.1: Localización de los edificios de la Universidad de León.

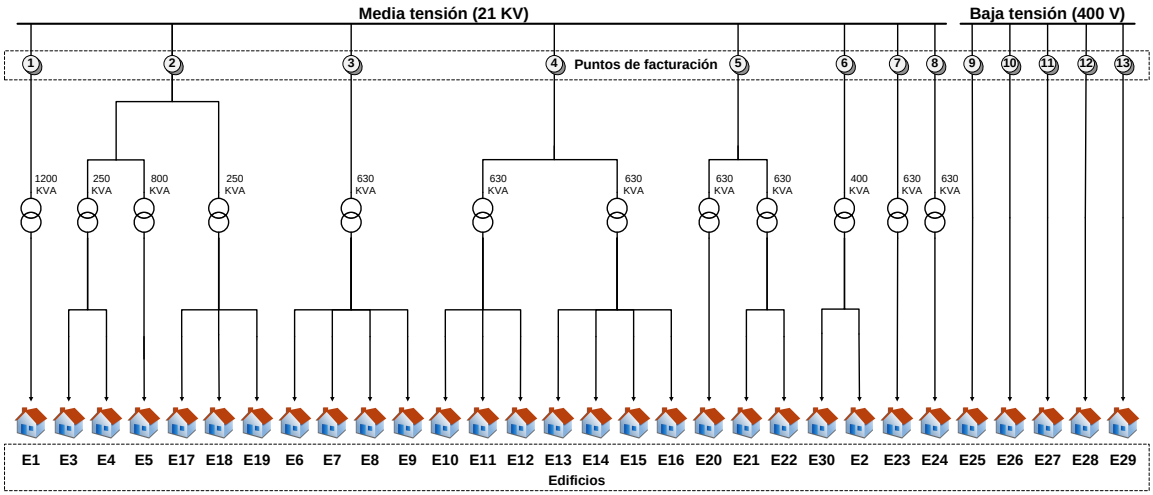


Figura 3.2: Esquema unifilar del suministro de energía eléctrica a los edificios.

uso en los edificios (400 V). Por otra parte, los edificios localizados en el centro de la ciudad de León reciben energía directamente de la red eléctrica de distribución en baja tensión de 400 V.

Como se ha comentado anteriormente, la compañía eléctrica mide la energía consumida en 13 puntos diferentes. No obstante, sería necesario medir y registrar datos del consumo eléctrico individual en cada edificio para mejorar su eficiencia energética. Esto permitiría analizar datos históricos, detectar anomalías o bien conocer la energía consumida por las empresas concesionarias de servicios, como las cafeterías, evitando tener que distribuir el consumo eléctrico en base a la superficie útil de los edificios. En este sentido, en algunos edificios se instalaron medidores propios de la Universidad de León, que son capaces de medir un amplio número de variables eléctricas y no sólo aquellas implicadas en la facturación (potencia pico, energías activa y reactiva). La supervisión del consumo de energía eléctrica registrado por estos medidores se realizaba de forma manual, anotando periódicamente el valor indicado en sus pantallas o *displays*. Este proceso de lectura de medidores era lento y requería desplazamientos de personal a los edificios. Además, al no realizar un registro continuo, se desperdiciaban todos los datos intermedios que podrían ser útiles para analizar tendencias históricas y explotar su información intrínseca con el fin de conseguir un ahorro energético y económico.

En esta situación, se planteó la posibilidad de definir un sistema de medida y supervisión moderno de la energía eléctrica en los edificios del Campus de la Universidad de León. Este sistema debe capturar el mayor número de variables eléctricas posible en todos los edificios, almacenar los datos de forma continua, permitir su exportación para un tratamiento posterior y facilitar la explotación de datos históricos para extraer el conocimiento implícito. El problema de ingeniería planteado se basa en diseñar un sistema de medida y supervisión de la energía eléctrica con una estructura abierta, escalable y robusta. Esta estructura se debe apoyar en tecnologías estándar para integrar fácilmente los medidores de nuevos edificios o incluso otros sistemas de supervisión o gestión. El sistema de supervisión debe incorporar herramientas tradicionales de visualización, que proporcionan valores instantáneos, tendencias, alarmas o avisos, informes, valores máximos y mínimos y consumos en diferentes periodos de tiempo, así como herramientas avanzadas de visualización basadas en técnicas de minería de datos, que aportan un concepto innovador con mucho potencial desde el punto de vista de la explotación de datos. Como consecuencia, la capacidad de almacenamiento debe ser lo suficientemente grande como para no imponer un límite a la minería de datos. Un requisito vital de los sistemas de supervisión modernos es el acceso remoto a través de Internet. Esto hace posible la monitorización desde cualquier parte, sin necesidad de que exista un puesto central de supervisión.

Dado que algunos edificios poseen ya instalado un medidor eléctrico, la primera tarea es comprobar su configuración y funcionamiento, sustituyendo aquellos que se encuentran defectuosos, instalando nuevos equipos donde sea necesario y parametrizando los medidores adecuadamente. Es imprescindible conocer si los equipos de medida disponen de una interfaz de comunicación estándar y compatible con el sistema de medida y supervisión para el intercambio de datos, debido a la distribución geográfica de los edificios.

3.2. Arquitectura de medida y supervisión

La arquitectura adoptada por el sistema de medida y supervisión de la energía eléctrica en los edificios de la Universidad de León se basa en una estructura de triple capa (Eckerson, 1995). Esta estructura plantea disponer de tres capas bien diferenciadas con funciones independientes, aunque todas ellas relacionadas entre sí. La estructura de triple capa se basa en el patrón cliente-servidor, muy utilizado en las aplicaciones que utilizan Internet como medio de acceso, aunque presenta una gran innovación que mejora su modularidad y escalabilidad. La introducción de una nueva capa intermedia entre el cliente y el servidor constituye una interfaz genérica entre el usuario y los equipos de medida, lo que proporciona una importante abstracción para el usuario. Esta capa intermedia está constituida por el sistema que gestiona la información suministrada por el servidor, teniendo como soporte de la misma generalmente una base de datos.

La arquitectura de triple capa adoptada se encuentra a medio camino entre un patrón con 2 capas (cliente-servidor) y uno con múltiples capas. Los avances en equipos, sistemas operativos, bases de datos y redes de comunicación han permitido desarrollar sistemas SCADA basados en arquitecturas cada vez con menos capas, donde el procesamiento se encuentra repartido en varios servidores (Marcuse *et al.*, 1997). En este sentido, algunas arquitecturas de medida y supervisión utilizan una única capa o nivel y distribuyen las funciones entre un conjunto de equipos (Nagata, 2006). Por otra parte, el uso de equipos diferentes de los servidores, como por ejemplo los autómatas programables (*Programmable Logic Controllers*, PLCs), puede proporcionar ventajas en el control de cargas eléctricas, en función de los datos medidos. No obstante, un PLC posee una capacidad de cálculo inferior y poca memoria interna de almacenamiento, por lo que sólo sería útil en pequeños sistemas de medida y supervisión (para un edificio) (Bayindir *et al.*, 2011).

La utilización de esta arquitectura permite separar las funciones de acceso de usuarios al sistema de supervisión en el nivel superior de las de acceso a los equipos de medida situados en el nivel inferior, proporcionando la característica de modularidad. El uso de esta estructura de triple capa en entornos industriales proporciona además una mejora en la integridad y accesibilidad a los datos, garantizando su recuperación ante posibles fallos.

Como se ha comentado anteriormente, el sistema de medida y supervisión consta de tres capas o niveles bien diferenciados, pero a su vez relacionados entre sí (Fuertes *et al.*, 2011):

- **Capa servidor**: es el nivel inferior donde se sitúan los equipos de medida eléctrica y meteorológica que llevan a cabo la captura de las variables eléctricas y ambientales. Una red de comunicación dedicada a la medida se emplea como elemento de interconexión entre todos los medidores, los cuales se distribuyen a lo largo de los edificios. Un servidor de adquisición se encarga de recolectar los datos procedentes de los medidores y enviarlos al nivel intermedio para su almacenamiento.

- **Capa intermedia**: en este nivel se lleva a cabo el almacenamiento y la minería de datos. Aquí se sitúan el servidor de base de datos, el servidor de explotación de datos y el servidor web, el cual proporciona los interfaces de usuario a los clientes. Todos

los servidores de esta capa se comunican mediante la red de explotación, excepto el servidor web, que por motivos de seguridad se aísla dentro de una subred denominada DMZ (*De-Militarized Zone*).

- **Capa cliente**: este nivel superior está formado por los equipos de los usuarios conectados a la red de supervisión, generalmente Internet, que solicitan mediante un navegador web los interfaces de usuario para visualizar el consumo de energía eléctrica.

Cabe mencionar que se puede establecer cierta analogía entre la arquitectura de triple capa y la pirámide de automatización, agrupando los niveles de supervisión y gestión. La capa servidor se correspondería con el nivel de campo, la capa intermedia con el nivel de control y la capa cliente con los niveles de supervisión y gestión.

Para implementar el concepto de triple capa, se ha creado una arquitectura de red compleja que consta de 3 redes de comunicación independientes, pero relacionadas entre sí por los correspondientes equipos de enlace. En este sentido, el servidor de adquisición enlaza la red de medida con la de explotación, mientras que el servidor web enlaza la red de explotación con la de supervisión. La plataforma tecnológica que soporta el Laboratorio Remoto de Automática de la Universidad de León (LRA-ULE)[1] utiliza una estructura de red muy similar a la expuesta (Domínguez *et al.*, 2005). En la figura 3.3 se puede ver la arquitectura del sistema de medida y supervisión de la energía eléctrica en los edificios de la Universidad de León.

En la red de supervisión (Internet), existen diferentes equipos cliente (PCs, PDAs, teléfonos móviles, etc.) que se pueden conectar al sistema de supervisión mediante un navegador web estándar. Cabe destacar que solamente los usuarios registrados tienen acceso al sistema de supervisión. Un conjunto de *routers* proveen seguridad al sistema ante posibles intrusos. Los usuarios autorizados solicitan páginas web con los interfaces de usuario al servidor web y este consulta el almacén correspondiente en la base de datos, donde reside la información requerida. El servidor de almacenamiento devuelve la información al servidor web para ser presentada adecuadamente en la página web de forma que el usuario pueda interpretarla y analizarla.

Por otra parte, el servidor de explotación de los datos ejecuta los algoritmos de minería de datos para extraer conocimiento acerca de los edificios. Para ello, este equipo toma datos crudos eléctricos, ambientales, temporales, etc. del correspondiente almacén de la base de datos, lleva a cabo un preprocesado, una exploración y un modelado y finalmente almacena el conocimiento obtenido sobre los edificios, de nuevo en la base de datos. Generalmente, se utiliza el entorno Matlab como motor de cálculo en la minería de datos, dadas sus excelentes cualidades en el procesamiento masivo de datos. Este proceso puede estar controlado localmente por un experto en técnicas de minería de datos, que decide en base a su experiencia los datos a utilizar y los parámetros de los algoritmos. No obstante, transcurridas las fases iniciales de la minería, es posible automatizar el proceso de forma que se ejecute periódicamente.

[1] www.lra.unileon.es

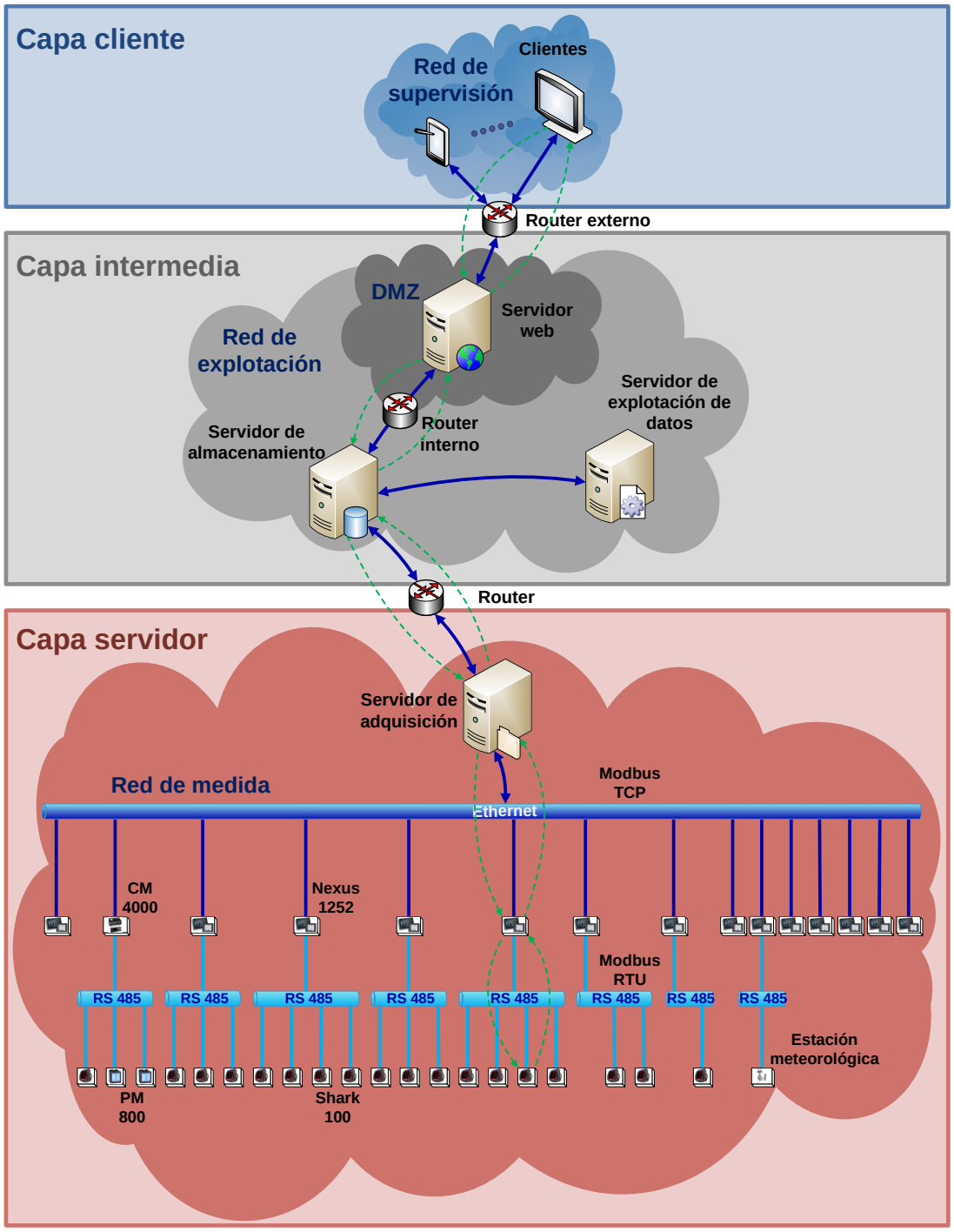


Figura 3.3: Arquitectura del sistema de medida y supervisión.

El servidor de adquisición se emplea para recolectar y almacenar continuamente los datos eléctricos y ambientales crudos procedentes de los equipos de medida. Este equipo realiza peticiones de lectura a todos los medidores y guarda los datos obtenidos en el almacén correspondiente de la base de datos. Los medidores eléctricos y ambientales se comunican mediante la red de medida, que está constituida por segmentos Ethernet y serie RS-485. Ambos tipos de segmentos están enlazados mediante pasarelas de comunicación. El protocolo de comunicación empleado en la red de medida es Modbus. Cada edificio posee instalado un medidor eléctrico con el fin de capturar sus variables eléctricas. En algunas ocasiones existe un medidor eléctrico adicional que agrupa un conjunto de edificios y que, además de medidor, se utiliza como pasarela de comunicación para un segmento serie. Los modelos de medidores utilizados son Nexus 1252, CM 4000, Shark 100 y PM 800.

A continuación se explicará con más detalle la arquitectura del sistema de medida y supervisión desarrollado.

3.3. Capa servidor

El nivel inferior o capa servidor está formado por todos los equipos de medida eléctrica, la estación meteorológica, la red de comunicación y el servidor de adquisición. Los medidores eléctricos junto con los sensores meteorológicos se encargan de capturar las variables físicas reales, constituyendo el sistema de medida propiamente dicho. Este sistema de medida se puede considerar distribuido debido a la deslocalización geográfica de los edificios y a las distancias entre ellos. Por tanto, una red de comunicación rápida y fiable que enlace todos los medidores es imprescindible para el correcto funcionamiento del sistema de medida y supervisión de la energía eléctrica.

3.3.1. Equipos de medida eléctrica

Los medidores eléctricos son los equipos encargados de capturar las variables eléctricas que caracterizan a cada edificio. Todos los edificios poseen un equipo de medida en la entrada del suministro de energía eléctrica (ver figura 3.4). La existencia de equipos de diferentes fabricantes y modelos se debe a los requerimientos técnicos y económicos en la fecha de instalación y confiere aún más complejidad al sistema de medida. Todos los medidores son compatibles entre sí, pero existen pequeñas diferencias en cuanto al número de variables capturadas, el número y tipo de interfaces de comunicación que poseen, la capacidad de cálculo, etc. Aunque la variedad de equipos siempre dificulta el mantenimiento, la posibilidad de instalar cualquier medidor compatible confiere flexibilidad y otras ventajas a la hora de sustituir o incorporar nuevos equipos de medida eléctrica. Actualmente, los medidores que existen en el mercado incorporan interfaces de comunicación estándar, lo que los hace ideales para el desarrollo de sistemas abiertos. Los edificios de nueva construcción se pueden incluir fácilmente en el sistema de medida siempre que se instale un medidor de características similares a los existentes.

A continuación se describen los diferentes modelos de los equipos que realizan la medida

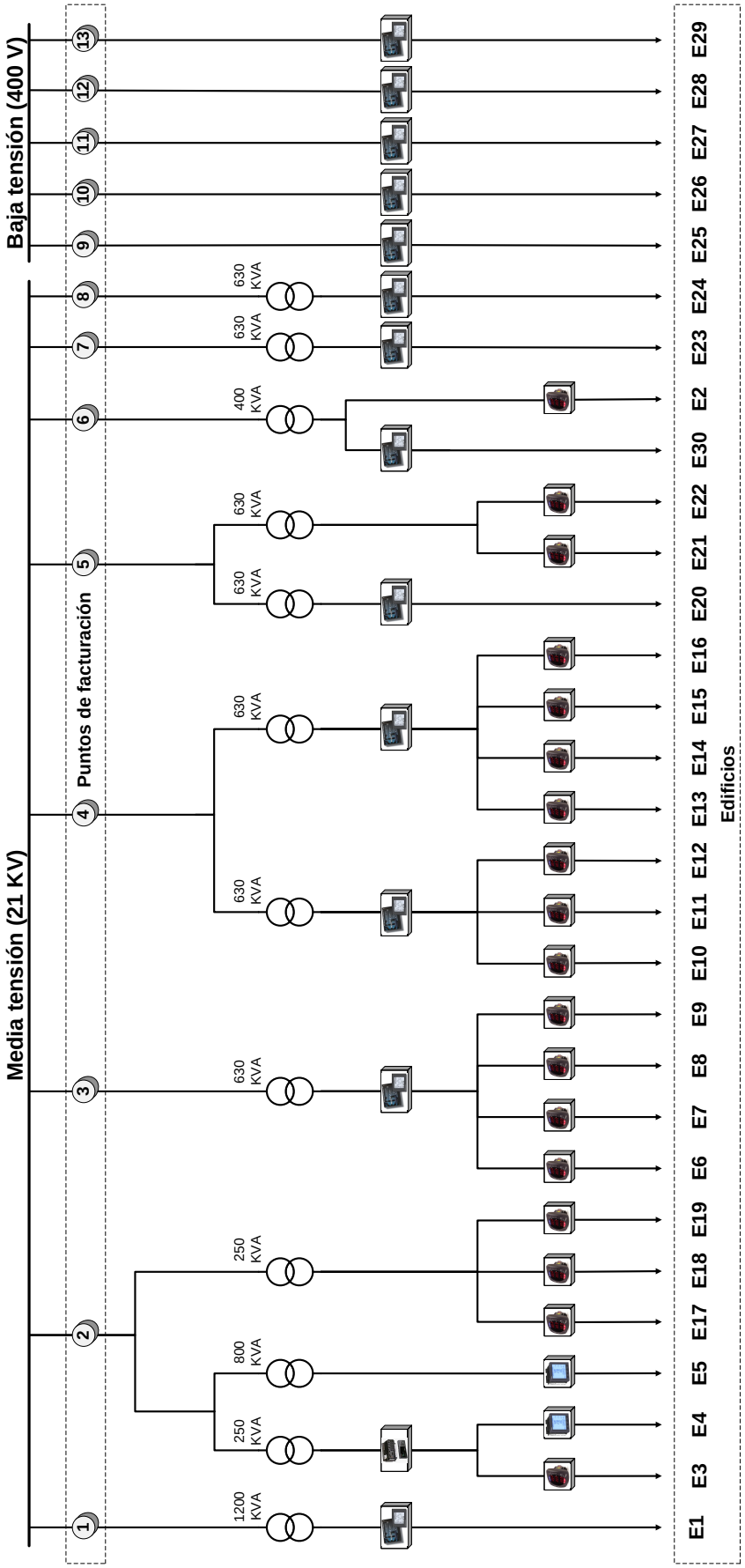


Figura 3.4: Distribución de los medidores eléctricos.

eléctrica en los edificios:

- **Medidores Nexus 1252**: este modelo fabricado por Electro Industries/GaugeTech[2] no sólo mide las variables eléctricas sino que además permite realizar un análisis de las mismas. Realiza la medida y registro de picos, cortes y otros efectos transitorios como el *flicker*[3], lo que facilita el análisis de la calidad de la energía. El análisis de armónicos también es posible. Su tecnología de medida trabaja con 512 muestras por ciclo de la onda de tensión y corriente, lo que otorga gran precisión en las medidas. Si existe un transformador en la línea trifásica de suministro, el equipo tiene en cuenta en la medida la pérdida de energía provocada. Este equipo puede almacenar datos en una pequeña memoria interna durante un corto periodo de tiempo para una evaluación posterior. Posee 4 puertos serie RS-485 y RS-232 de alta velocidad para conectar tarjetas de E/S digitales o analógicas, entradas de contaje, etc. Otra utilidad de estos puertos es la creación de una red serie con varios esclavos en la que el equipo Nexus 1252 es el maestro, es decir que puede funcionar como pasarela de comunicación entre diferentes redes. Además, un puerto Ethernet está disponible junto con un módem 56K. Los protocolos de comunicación soportados son Modbus y DNP e incorpora servidor web integrado para la lectura remota de variables. Más características del medidor Nexus 1252 se pueden revisar en Electro Industries/GaugeTech (2010a).

- **Medidores CM 4000**: este modelo del fabricante Schneider Electric[4] se utiliza en la medida y análisis de instalaciones eléctricas grandes y críticas que no pueden sufrir interrupciones en el suministro. Este equipo puede detectar y registrar los transitorios de corta duración, los huecos y las deformaciones en la onda de tensión. Asimismo, realiza mediciones del efecto *flicker*. Incorpora filtros *antialiasing* para reducir el error de medición, lo que garantiza una gran precisión en la medida. Este equipo realiza un análisis de la calidad de la energía en base a un análisis de armónicos. Además, detecta la dirección de las perturbaciones, es decir, si estas se producen aguas arriba o aguas abajo del medidor. Las variables eléctricas se registran cada segundo, 100 ms o cada ciclo de la onda (20 ms) en una memoria interna, que a su vez se puede utilizar para almacenar datos de facturación, eventos, etc. El medidor CM 4000 ofrece tendencias y previsiones del consumo de energía y permite definir alarmas o valores de ajuste para señales de aviso. Dispone de 25 E/S flexibles para conectar instrumentación adicional. Incorpora una interfaz de comunicaciones Ethernet de alta velocidad y RS-485 para comunicación serie con equipos esclavos. El medidor CM 4000 puede funcionar como pasarela entre ambas redes y como cualquier equipo de Schneider Electric soporta el protocolo Modbus. Incorpora servidor web para páginas HTML y notificación automática de alarmas por correo electrónico. Información detallada acerca de este equipo se puede encontrar en Schneider Electric (2005).

- **Medidores Shark 100**: este modelo fabricado por Electro Industries/GaugeTech

[2] www.electroind.com
[3] Fluctuación rápida y repetitiva de la tensión que produce el parpadeo de la iluminación.
[4] www.schneider-electric.com

realiza únicamente funciones de medida eléctrica. Las principales aplicaciones son en el campo de la facturación y medidas en el nivel secundario de distribución dentro de una instalación eléctrica. Una de las ventajas de este modelo es su bajo coste. Este modelo permite modificar su configuración para decidir las variables a medir, todo ello sin sustituir el equipo físico. La configuración por defecto permite medir solamente tensiones y corrientes, mientras que la configuración avanzada ofrece tensiones, corrientes, potencias, energías y armónicos. La tecnología de medida proporciona una precisión de 0.1 % en tensiones y corrientes. El equipo incorpora una pantalla LED para visualizar localmente los valores de las variables eléctricas, pero también es posible utilizar una versión sin pantalla, en la que funciona únicamente como transductor. Respecto a las interfaces de comunicación, posee un puerto de infrarrojos IrDA para lectura y/o programación remota desde una PDA o portátil. También tiene un puerto serie RS-485 que soporta los protocolos Modbus y DNP. Cabe destacar que existe la opción de añadir un módulo externo con una interfaz Ethernet. Más especificaciones del medidor Shark 100 se pueden encontrar en Electro Industries/GaugeTech (2010b).

- **Medidores PM 800**: la serie PM 800 del fabricante Schneider Electric consta de tres modelos físicamente idénticos (PM 800, PM 820 y PM 850), aunque se diferencian entre sí en las variables medidas y en las funciones disponibles. Estos modelos tienen unas capacidades de medida de gran rendimiento, útiles para supervisar una instalación eléctrica con un coste bajo. Todos los equipos poseen una pantalla para la visualización local de las variables en las tres fases y el neutro al mismo tiempo. Ofrecen la posibilidad de definir alarmas personalizadas e incorporan una pequeña memoria interna no volátil para realizar registros de datos y/o eventos. Aunque en un número reducido, poseen E/S digitales para conectar señales externas. Incluyen un puerto de comunicación serie RS-485 y soportan el protocolo Modbus. Un módulo externo con una interfaz Ethernet puede ser añadido de forma opcional. El modelo superior de la serie mide valores THD (*Total Harmonic Distortion*) y armónicos (magnitudes y ángulos) en tensiones y corrientes de forma individual para cada fase. Además ofrece una captura de la forma de onda, curvas de tendencias y previsiones a corto plazo. Por estos motivos, los medidores utilizados en el sistema de medida y supervisión son PM 850. Una descripción más detallada de este medidor se puede revisar en Schneider Electric (2006).

- **Medidor ION 7650**: este modelo del fabricante Schneider Electric ha salido al mercado recientemente y posee avanzadas prestaciones. Incorpora la medición de componentes simétricas, detección rápida de oscilaciones, picos y dirección de las perturbaciones, captura de formas de ondas a 1024 muestras por cada ciclo y supervisión de la calidad de la energía con una medida de hasta 63 armónicos. Dispone de una pantalla de cristal líquido que permite visualizar localmente no sólo valores sino que también gráficas temporales. Realiza una medición en los 4 cuadrantes y las 3 fases con una precisión de clase 0.2. Con este equipo es posible aplicar una corrección del transformador de medida y una compensación de la pérdida producida por el transformador de la línea de suministro eléctrico. Puede registrar una secuencia

de eventos, tendencias históricas y capturas de pantalla. Posee 4 E/S analógicas, 16 entradas digitales, 7 salidas digitales y 65 puntos para alarmas y control. Los puertos de comunicación disponibles son Ethernet, módem, RS-232/485, RS-485 y óptico. Los protocolos soportados son Modbus, DNP y MV-90. Este equipo puede funcionar como pasarela entre los puertos Ethernet y RS-485 e incluye servidor web y envíos por correo electrónico para alarmas y datos. Más información de este equipo de medida se puede encontrar en Schneider Electric (2010). El equipo ION 7650 se utiliza en la verificación del resto de medidores, análisis puntuales de una instalación eléctrica en un edificio y en la toma de datos concretos con gran precisión para construir modelos.

Podemos decir que los modelos Nexus 1252 y CM 4000 son equipos que pertenecen a la gama alta, es decir, que poseen unas mejores prestaciones, a diferencia de los modelos Shark 100 y PM 800 que se encuadran dentro de la gama media. El modelo ION 7650 estaría dentro de una gama superior, ya que posee una tecnología moderna totalmente diferente. Los medidores del fabricante Electro Industries/GaugeTech son más empleados en América, mientras que los equipos de Schneider Electric son más comunes en Europa, concretamente en España.

Las variables básicas medidas por los equipos son la tensión y la corriente. El resto de variables eléctricas que ofrecen los equipos se calculan a partir de muestras procedentes de las ondas de tensión y corriente por medio de un DSP (*Digital Signal Processor*) interno. Los DSP están especialmente diseñados para ejecutar a gran velocidad algoritmos costosos como puede ser una transformada de Fourier para obtener los armónicos. Estos algoritmos que calculan el resto de variables eléctricas pueden variar ligeramente, dependiendo del fabricante.

Todos los equipos miden tensiones hasta 400 V respecto al neutro y 700 V entre líneas. Por tanto, los medidores se conectan directamente a la línea trifásica de suministro de energía eléctrica de cada edificio. No obstante, es conveniente colocar un fusible que proteja al equipo ante picos de tensión y cortocircuitos. Por el contrario, para medir la corriente siempre es necesario colocar un transformador de medida que reduzca su valor hasta valores comprendidos entre 0 y 5 amperios de forma que la electrónica del medidor no se dañe, ya que el consumo de corriente de los edificios puede ser elevado (cientos de amperios). Este rango de 0-5 A es estándar y todos los equipos lo utilizan en sus entradas de medida de corriente (IEEE Association, 2008). La relación de los transformadores de corriente varía en función del consumo del edificio y puede oscilar desde 20/5 hasta 1000/5. Por tanto, los medidores se conectan en el punto de medida de la instalación eléctrica trifásica mediante la correspondiente instrumentación de medida, la cual consta de un transformador de corriente por cada fase, incluido el neutro.

En la figura 3.5 se puede observar el esquema de conexión típico de un medidor en el punto de medida de la red trifásica de suministro eléctrico. Normalmente, al primario del transformador de línea llegan 3 fases y del secundario se obtienen las 3 fases y el neutro. Por tanto, la distribución de la energía eléctrica al edificio se realiza en 4 hilos para que sea posible alimentar, tanto a receptores trifásicos, como monofásicos (fase y neutro). Del secundario del transformador se obtiene una o varias líneas que alimentan a uno o varios

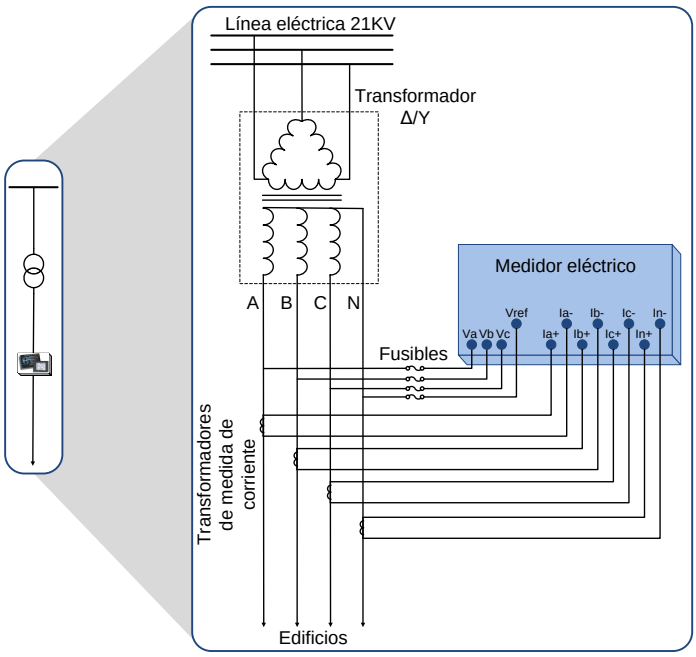


Figura 3.5: Conexión de los medidores eléctricos en el punto de medida.

edificios. Dentro de cada edificio, la línea principal se divide en varias líneas secundarias de distribución para cada zona. Según el punto donde se instalen los medidores, es posible conocer el consumo de un conjunto de edificios, de un solo edificio o de una zona dentro de un edificio.

Todos los medidores instalados ofrecen una serie de variables comunes, que en definitiva, son las variables más significativas que caracterizan a los edificios y por eso son almacenadas por el sistema de medida y supervisión. No obstante, los equipos miden otras variables eléctricas complementarias, dependiendo del fabricante y modelo. Por ejemplo, los modelos de la gama alta miden un mayor número de armónicos. Las variables eléctricas comunes medidas por los equipos se exponen en la tabla 3.2.

3.3.2. Estación meteorológica

La estación meteorológica se encarga de capturar las variables ambientales externas que caracterizan el entorno de los edificios e influyen en el consumo eléctrico. Estas condiciones ambientales varían mínimamente de unos edificios a otros, por lo que se pueden considerar comunes a todos ellos. Por tanto, la medida de las condiciones meteorológicas se realiza en un solo punto (ver figura 3.6). Aunque las estaciones meteorológicas comerciales proveen multitud de variables atmosféricas y aplicaciones de gran utilidad, en este caso se ha optado por desarrollar una estación propia para la toma de datos ambientales. En la implementación de esta estación se ha utilizando una tarjeta de adquisición y varios sensores, que lógicamente tienen un coste bastante inferior a una estación comercial.

Como se comentó anteriormente, los equipos de medida eléctrica modelo Nexus 1252

Número	Etiqueta	Variable	Unidad
1	Va	Tensión entre fase a y neutro	Voltios (V)
2	Vb	Tensión entre fase b y neutro	Voltios (V)
3	Vc	Tensión entre fase c y neutro	Voltios (V)
4	Vab	Tensión entre fases a-b	Voltios (V)
5	Vbc	Tensión entre fases b-c	Voltios (V)
6	Vca	Tensión entre fases c-a	Voltios (V)
7	Ia	Corriente en la fase a	Amperios (A)
8	Ib	Corriente en la fase b	Amperios (A)
9	Ic	Corriente en la fase c	Amperios (A)
10	In	Corriente en el neutro	Amperios (A)
11	F	Frecuencia de red	Hercios (Hz)
12	P	Potencia activa en las tres fases	Kilovatios (KW)
13	Q	Potencia reactiva en las tres fases	Kilovoltioamperios reactivos (KVAR)
14	S	Potencia aparente en las tres fases	Kilovoltioamperios (KVA)
15	FP	Factor de potencia en las tres fases	-
16	EA	Energía activa en las tres fases	Kilovatios hora (KWh)
17	ER	Energía reactiva en las tres fases	Kilovoltioamperios reactivos hora (KVARh)
18	THD Va	Distorsión armónica total en la tensión Va	Porcentaje del valor fundamental (%)
19	THD Vb	Distorsión armónica total en la tensión Vb	Porcentaje del valor fundamental (%)
20	THD Vc	Distorsión armónica total en la tensión Vc	Porcentaje del valor fundamental (%)
21	THD Ia	Distorsión armónica total en la corriente Ia	Porcentaje del valor fundamental (%)
22	THD Ib	Distorsión armónica total en la corriente Ib	Porcentaje del valor fundamental (%)
23	THD Ic	Distorsión armónica total en la corriente Ic	Porcentaje del valor fundamental (%)
24	H1 Va	Primer armónico en la tensión Va	Porcentaje del valor fundamental (%)
25	H2 Va	Segundo armónico en la tensión Va	Porcentaje del valor fundamental (%)
26	H3 Va	Tercer armónico en la tensión Va	Porcentaje del valor fundamental (%)
27	H1 Ia	Primer armónico en la corriente Ia	Porcentaje del valor fundamental (%)
28	H2 Ia	Segundo armónico en la corriente Ia	Porcentaje del valor fundamental (%)
29	H3 Ia	Tercer armónico en la corriente Ia	Porcentaje del valor fundamental (%)
30	H4 Ia	Cuarto armónico en la corriente Ia	Porcentaje del valor fundamental (%)
31	H5 Ia	Quinto armónico en la corriente Ia	Porcentaje del valor fundamental (%)
32	H6 Ia	Sexto armónico en la corriente Ia	Porcentaje del valor fundamental (%)
33	H7 Ia	Séptimo armónico en la corriente Ia	Porcentaje del valor fundamental (%)
34	H1 Vb	Primer armónico en la tensión Vb	Porcentaje del valor fundamental (%)
35	H2 Vb	Segundo armónico en la tensión Vb	Porcentaje del valor fundamental (%)
36	H3 Vb	Tercer armónico en la tensión Vb	Porcentaje del valor fundamental (%)
37	H1 Ib	Primer armónico en la corriente Ib	Porcentaje del valor fundamental (%)
38	H2 Ib	Segundo armónico en la corriente Ib	Porcentaje del valor fundamental (%)
39	H3 Ib	Tercer armónico en la corriente Ib	Porcentaje del valor fundamental (%)
40	H4 Ib	Cuarto armónico en la corriente Ib	Porcentaje del valor fundamental (%)
41	H5 Ib	Quinto armónico en la corriente Ib	Porcentaje del valor fundamental (%)
42	H6 Ib	Sexto armónico en la corriente Ib	Porcentaje del valor fundamental (%)
43	H7 Ib	Séptimo armónico en la corriente Ib	Porcentaje del valor fundamental (%)
44	H1 Vc	Primer armónico en la tensión Vc	Porcentaje del valor fundamental (%)
45	H2 Vc	Segundo armónico en la tensión Vc	Porcentaje del valor fundamental (%)
46	H3 Vc	Tercer armónico en la tensión Vc	Porcentaje del valor fundamental (%)
47	H1 Ic	Primer armónico en la corriente Ic	Porcentaje del valor fundamental (%)
48	H2 Ic	Segundo armónico en la corriente Ic	Porcentaje del valor fundamental (%)
49	H3 Ic	Tercer armónico en la corriente Ic	Porcentaje del valor fundamental (%)
50	H4 Ic	Cuarto armónico en la corriente Ic	Porcentaje del valor fundamental (%)
51	H5 Ic	Quinto armónico en la corriente Ic	Porcentaje del valor fundamental (%)
52	H6 Ic	Sexto armónico en la corriente Ic	Porcentaje del valor fundamental (%)
53	H7 Ic	Séptimo armónico en la corriente Ic	Porcentaje del valor fundamental (%)

Tabla 3.2: Variables eléctricas capturadas por los medidores para cada edificio.

y CM 4000 pueden funcionar como maestros en una red serie. Aprovechando esta funcionalidad, se ha conectado un módulo externo de entradas analógicas a uno de los puertos RS-485 del equipo Nexus 1252 instalado en el edificio E24. Este módulo de adquisición del fabricante Electro Industries/GaugeTech requiere alimentación externa y posee 8 canales analógicos de 0-20 mA. Inicialmente sólo se utilizan 3 canales, uno para cada sensor conectado, parametrizados entre 4 y 20 mA. Los sensores cableados a las entradas analógicas del módulo son:

- Sensor de temperatura ambiente exterior de tipo resistivo cuyo rango de medida se encuentra entre -35 ºC y 70 ºC.
- Sensor de humedad relativa de tipo capacitivo con el rango de medida [0, 100] %.
- Sensor de radiación solar de tipo termoeléctrico con el rango de medida [0, 1000] W/m^2.

No obstante, otros sensores meteorológicos (presión atmosférica, precipitación, velocidad del viento, etc.) se pueden conectar si es necesario.

El módulo analógico adquiere continuamente los valores que proporcionan los sensores meteorológicos, mientras que el equipo Nexus 1252 registra esos datos en las correspondientes direcciones de memoria. De esta forma, es posible acceder a las variables meteorológicas al mismo tiempo que a las eléctricas como se comentará posteriormente.

3.3.3. Red de medida

Los medidores se encuentran distribuidos ya que los edificios se localizan en varias zonas de la ciudad de León. Por tanto, es necesario una red de comunicación que enlace a todos los equipos de medida. Esta red de campo está formada por diversos segmentos, los cuales se diferencian en el soporte físico que utilizan para la transmisión de la información. Los estándares de comunicación que constituyen el nivel físico de los segmentos de la red de medida se agrupan en dos clases:

- **Segmentos Ethernet**: el estándar IEEE 802.3 (IEEE Association, 2002) define las características de cableado y señalización del nivel 1 o nivel físico y los formatos de las tramas de datos del nivel 2 o nivel de enlace de datos del modelo OSI (*Open System Interconnection*) (Zimmermann, 1980). El medio físico de los segmentos Ethernet está formado por 4 pares de hilos trenzados de categoría 6 UTP, la velocidad puede llegar a ser de 1 Gbps y la distancia máxima entre dos nodos adyacentes es de 100 m. Pueden existir segmentos Ethernet cuyo medio físico es la fibra óptica monomodo, en cuyo caso la velocidad de transmisión se mantiene, pero la distancia máxima puede alcanzar hasta los 5000 m. Múltiples segmentos Ethernet se pueden enlazar mediante estaciones repetidoras de señal, formando una red. Además, el uso de dispositivos de interconexión como *switches* y *routers* permiten conectar diferentes segmentos y redes entre sí, haciendo posible la expansión. Las tramas o paquetes Ethernet están constituidos por un campo de preámbulo, inicio, dirección de destino, dirección de

origen, tipo y longitud de los datos, información o datos y finalmente el control de errores. El mecanismo de acceso al medio empleado por este estándar es CSMA/CD (*Carrier Sense Multiple Access/Collision Detection*). Las redes Ethernet permiten la transmisión de grandes volúmenes de datos a altas velocidades.

- **Segmentos serie RS-485**: el estándar RS-485 ó EIA-485 (Electronic Industries Alliance, EIA, 2003) indica las especificaciones en el nivel 1 o nivel físico del modelo OSI. El medio físico utilizado en los segmentos serie está formado por 2 hilos trenzados y apantallados. Este estándar define un sistema en bus de transmisión multipunto diferencial, el cual permite transmitir información a velocidades de 100 Kbps y una distancia de 1200 m. Se admiten hasta 32 estaciones emisoras-receptoras en cada segmento serie. Estos segmentos emplean la configuración *half-duplex*, con la que sólo es posible o bien enviar o recibir información al mismo tiempo. El mecanismo maestro-esclavo evita que se produzcan transmisiones de datos simultáneas, de forma que sólo hay un equipo transmitiendo a la vez en cada segmento serie. Los segmentos serie poseen una topología en bus, siendo siempre necesario colocar una resistencia de 120 Ω para atenuar la señal al final del bus. Las tramas serie poseen un campo de inicio, dirección, datos, control de paridad, detección de errores y parada. Los parámetros típicos de configuración empleados en la transmisión de información son 8 bits de datos, sin paridad, velocidad 19200 baudios, detección de errores, un bit de inicio y otro de parada. Las redes serie permiten transmitir pequeños volúmenes de datos a una distancia media, la cual se reduce considerablemente si se incrementa la velocidad de transmisión.

En la figura 3.6 se puede ver la estructura de la red de comunicación utilizada en el sistema de medida de la energía eléctrica en los edificios de la Universidad de León. Los segmentos serie RS-485 enlazan medidores muy próximos entre sí, mientras que los segmentos Ethernet permiten conectar los medidores que se encuentran en edificios muy alejados. Los medidores modelo Nexus 1252 y CM 4000 poseen una interfaz de comunicaciones Ethernet, por lo que se conectan directamente a los segmentos Ethernet. Además de medir las variables eléctricas, estos equipos funcionan como pasarela entre el segmento Ethernet y el serie RS-485. Cuando las distancias físicas entre los medidores no son muy elevadas es posible crear un segmento serie que comunica los medidores con este tipo de interfaz (modelo Shark 100 y PM 800). Cada equipo en el segmento serie posee un identificador único y el número máximo de medidores en el segmento serie no puede superar 32. Los medidores modelo Nexus 1252 y CM 4000 funcionan como maestros en un segmento serie, siendo los equipos Shark 100 y PM 800 los esclavos. La principal ventaja de esta estructura de medida que combina segmentos Ethernet y serie RS-485 es el ahorro económico en equipos de medida, aunque se pueden producir cuellos de botella en la comunicación simultánea de varios medidores conectados en un mismo segmento serie. Cabe destacar que los medidores de los edificios localizados en el centro y suroeste de la ciudad (E24-E29) deben disponer de una interfaz Ethernet, ya que es la única opción de integrarlos en el sistema de medida y supervisión. En este caso es físicamente imposible cablear un segmento serie que enlace varios edificios debido a la distancia existente entre ellos. Otra opción sería disponer de una pasarela de comunicación

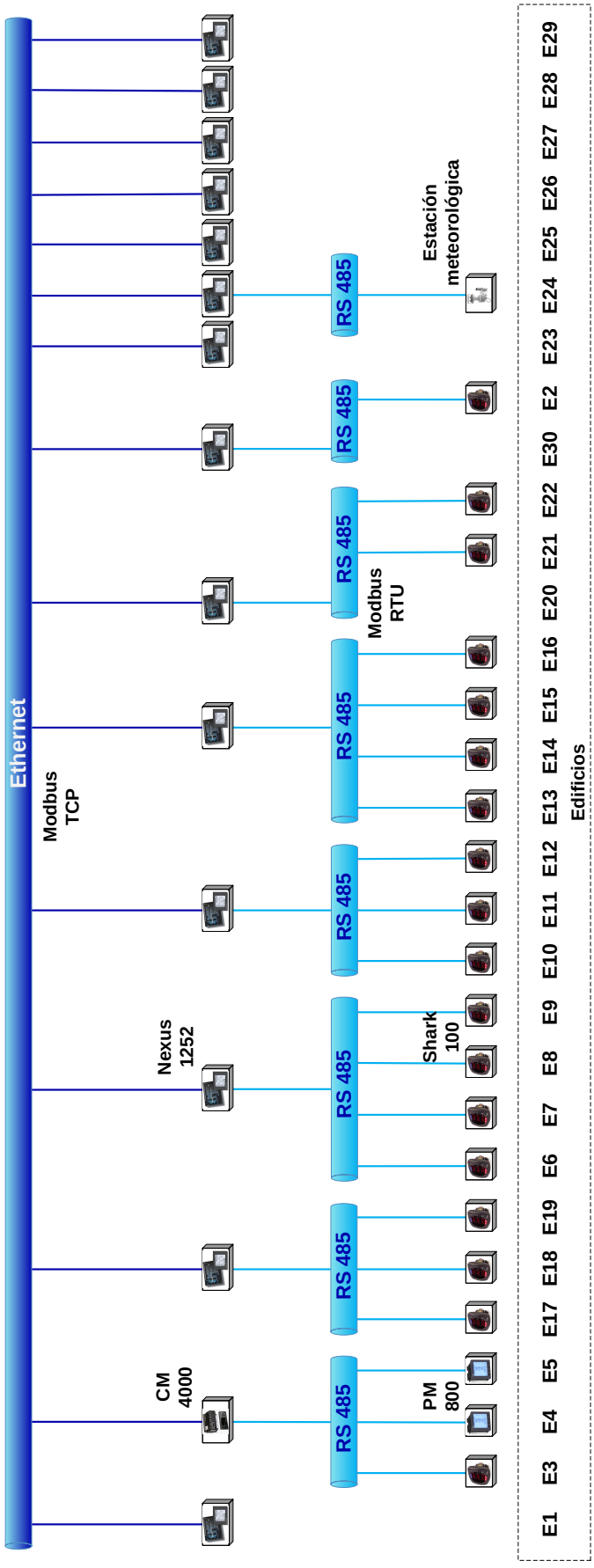


Figura 3.6: Red de comunicación que enlaza los medidores eléctricos y la estación meteorológica.

a Ethernet para uno o varios medidores con interfaz serie RS-485.

La red de medida hace uso de la red de área local que la Universidad de León posee en todos sus edificios. El objetivo es aprovechar los recursos disponibles y minimizar las tareas de cableado de la red. Dentro de esta red LAN (*Local Area Network*) basada en el estándar Ethernet, se ha definido una red de área local virtual lógicamente independiente y destinada únicamente al sistema de medida, la cual permite comunicar entre sí los medidores eléctricos de todos los edificios. Gracias a la red VLAN (*Virtual Local Area Network*) es posible separar segmentos lógicos dentro de la red LAN de la Universidad de León y así se evita intercambiar información directamente con otros segmentos dedicados a otros fines como la secretaria virtual, el profesorado, etc. La red VLAN funciona en el nivel 3 o nivel de red del modelo OSI y es fácilmente configurable mediante software, lo que aporta flexibilidad y garantiza la escalabilidad del sistema de medida y supervisión. Para agregar nuevos medidores, simplemente es necesario indicar la pertenencia de un nuevo segmento de red a la VLAN (Freeman y Passmore, 1996). La red VLAN está constituida por varios segmentos Ethernet, generalmente uno por cada edificio, excepto donde existe un segmento serie. Cuando los edificios se encuentran situados en zonas alejadas entre sí, como por ejemplo la zona del centro de la ciudad y el Campus de Vegazana, se utiliza la red de la compañía operadora de comunicaciones para enlazar los dos segmentos Ethernet, los cuales pertenecen a la misma VLAN. Desde el punto de vista práctico es como si solamente existiese un segmento Ethernet, debido a que la compañía de telecomunicaciones garantiza ese servicio, independientemente de la distancia existente. Por tanto, el sistema de medida y supervisión podría integrar cualquier edificio, incluso aquellos que la Universidad de León posee en otras ciudades.

3.3.4. Protocolo de comunicación

El protocolo de comunicación utilizado en el sistema de medida está basado en el estándar Modbus (Modbus Organization, 2006). Modbus es un protocolo de comunicación situado en el nivel 7 o nivel de aplicación del modelo OSI y se basa en una arquitectura de tipo maestro-esclavo. Este protocolo fue creado en 1979 por Modicon (ahora Schneider Electric) para su uso con autómatas programables en la industria. Modbus es un protocolo ampliamente utilizado por los sistemas SCADA y en concreto, por los sistemas de supervisión de la energía eléctrica (Mohagheghi *et al.*, 2009; Bayindir *et al.*, 2011). Además es un protocolo abierto que puede facilitar una futura integración con otro tipo de sistemas de automatización y control en los edificios (Kastner *et al.*, 2005).

Los campos requeridos en un paquete de datos de tipo Modbus son la dirección física del equipo dentro de la red, el código de la función que indica la acción a realizar por el equipo, los datos a trasmitir y finalmente una comprobación de errores. Los códigos de función soportados por Modbus son variados y podemos encontrar funciones de lectura/escritura de valores analógicos de 16 bits (*Read Holding Registers/Write Single Register*) o valores digitales de 1 bit (*Read Coils/Write Single Coil*), entre otras.

En el sistema de medida y supervisión de la energía eléctrica, las peticiones que se llevan

a cabo son siempre de lectura (*Read Holding Registers*) de una o más variables eléctricas procedentes de los medidores. El servidor de adquisición actúa siempre como maestro en la red de medida y realiza esas peticiones de lectura. Los medidores eléctricos son esclavos en la red y responden a las peticiones con los datos solicitados. En los segmentos Ethernet el servidor de adquisición realiza una solicitud y a continuación el medidor responde con los valores. En el caso de un segmento serie RS-485, el servidor realiza la petición al medidor, pero es la pasarela de comunicación la que gestiona esta petición, actuando como maestro en la red serie y como esclavo del segmento Ethernet.

Podemos distinguir dos versiones de Modbus, dependiendo del medio físico sobre el cual se implementa (ver figura 3.6):

- **Modbus TCP**: esta versión de Modbus, donde la transmisión de los datos se realiza mediante paquetes TCP/IP, se emplea en los segmentos Ethernet.
- **Modbus RTU**: esta versión de Modbus se utiliza en los segmentos serie RS-485. Modbus RTU (*Remote Terminal Unit*) es una representación binaria compacta de los datos y es más eficiente que otras variantes serie como por ejemplo Modbus ASCII.

Las pasarelas de comunicación realizan una labor fundamental en la traducción del protocolo en ambos sentidos entre un segmento de red con protocolo Modbus TCP y otro que utiliza Modbus RTU.

3.3.5. Servidor de adquisición

El servidor de adquisición es el equipo destinado a la toma de datos procedentes de todos los medidores eléctricos y la estación meteorológica y al envío de los mismos hacia el servidor de almacenamiento donde se insertan de forma ordenada. De esta forma se realiza la función de enlace entre el nivel inferior, donde se encuentran los equipos de medida eléctrica y ambiental, y el nivel intermedio, donde se sitúa el servidor de almacenamiento. Un *router* gestiona el tráfico entre las redes de medida y de explotación, permitiendo solamente conexiones iniciadas por el servidor de adquisición. Por tanto, la actividad en la red de explotación es totalmente independiente del tráfico en la red de medida. Ahora bien, cualquier fallo en este servidor es crítico para el funcionamiento del sistema de medida y supervisión, produciendo un paro en la toma de datos. La instalación de un servidor redundante que entre en funcionamiento cuando surja un fallo solucionaría el problema, aunque encarecería el sistema.

La primera misión de este equipo es ejecutar continuamente un servicio que lleva a cabo la adquisición de datos. Para este propósito se hace uso de la red de medida que interconecta todos los equipos de medida. El protocolo de comunicación utilizado en la toma de datos es Modbus como ya se explicó anteriormente. El servicio de adquisición actúa como un maestro en la red de medida y es el que realiza las peticiones de tipo Modbus a los medidores eléctricos de acuerdo con el periodo de muestreo seleccionado (normalmente 1 ó 2 minutos). La adquisición se lleva a cabo de forma paralela para todos los medidores, es decir, que el servicio realiza tantas peticiones Modbus como medidores existen. A continuación, los

medidores responden enviando todos los datos solicitados al maestro. En cada petición se solicitan todas las variables, por lo que los datos se envían en bloque, dentro de una misma respuesta. El orden de recepción de respuestas procedentes de los medidores no es el mismo que el orden de las peticiones ya que dependen de las demoras introducidas por la red de comunicación y por cada medidor. Una adquisición de tipo secuencial, en la que las solicitudes se realizan una tras otra, una vez recibida la respuesta del medidor, hace que los retrasos se sumen. La consecuencia es que el periodo de muestreo mínimo posible será de unos pocos minutos, o incluso mayor si se añaden nuevos medidores. Por tanto, la adquisición en paralelo se revela como la más apropiada debido al elevado número de medidores existente. De este modo se garantiza una total independencia en la elección del periodo de muestreo, que podría llegar a ser incluso de unos pocos segundos. El límite está impuesto por la red serie RS-485, donde el maestro produce cuellos de botella en la comunicación.

La segunda misión del servidor de adquisición es enviar los datos recibidos hacia el servidor de almacenamiento e insertarlos de forma ordenada en la base de datos. En el proceso de inserción se utiliza un sistema de gestión de colas de tipo FIFO (*First In, First Out*), es decir, que la primera muestra de datos recibida procedente de los medidores, será la primera en ser insertada en la base de datos. En la inserción se añade de forma automática el instante de tiempo correspondiente.

Tanto en la conexión con los medidores como con el servidor de almacenamiento se ha implementado una redundancia software de forma que el servicio de adquisición realiza una re-conexión automática en el caso de perder la conectividad con estos equipos. De forma complementaria, también se realizan hasta 3 intentos en las peticiones y en las inserciones en el caso de no recibir la correspondiente respuesta una vez transcurrido un cierto tiempo de espera. Esto evita posibles fallos en la transmisión de información y otorga fiabilidad al sistema de adquisición.

El funcionamiento del servicio de adquisición es configurable, lo que aumenta su versatilidad. No sólo es posible modificar el periodo de muestreo, sino que también es posible seleccionar el número de medidores activos, su tipo o modelo, sus direcciones físicas, etc. De esta forma, cuando se modifican los parámetros, el servicio de adquisición adapta su funcionamiento a los nuevos valores.

3.4. Capa intermedia

El nivel intermedio introducido entre las capas cliente y servidor soporta el almacenamiento de los datos procedentes de todos los equipos de medida y el conocimiento extraído de los edificios. Este conocimiento se obtiene como resultado de aplicar técnicas de minería de datos que ejecuta el servidor de explotación de datos, el cual también reside en esta capa. Tanto el proceso de almacenamiento como el de extracción de conocimiento, se llevan a cabo de forma transparente para el usuario del sistema de medida y supervisión. No obstante, el proceso de minería de datos puede ser guiado por un experto, sobre todo en las fases iniciales de introducción de nuevos métodos. El servidor web sirve de enlace entre este nivel intermedio y el nivel superior, proporcionando a los usuarios los interfaces

de visualización para la supervisión.

3.4.1. Servidor de almacenamiento

El servidor de almacenamiento es el equipo destinado a guardar y gestionar toda la información relativa al sistema de medida y supervisión, y no sólo los datos eléctricos. Por tanto, se pueden distinguir 3 tipos de almacenes de información diferentes dentro de este servidor:

- **Almacén de datos**: en este almacén se guardan todos los datos crudos procedentes de los medidores eléctricos de los edificios y la estación meteorológica.
- **Almacén de conocimiento**: en esta zona de almacenamiento residen los resultados de los métodos de minería de datos que ejecuta el servidor de explotación. Estos métodos generan abundante conocimiento de los edificios, el cual se puede emplear en el análisis y supervisión avanzados.
- **Almacén de parámetros**: aquí se guardan los parámetros de configuración del sistema de medida y supervisión, tales como el periodo de muestreo, las direcciones físicas de los medidores dentro de la red de medida, el tipo o modelo de los medidores, su permiso de activación, etc.

El servidor de almacenamiento está constituido básicamente por una base de datos (BD) y su sistema gestor (SGBD) de tipo Microsoft SQL Server. La elección de una base de datos como medio de almacenamiento garantiza aspectos como el control en el acceso a la información, la fiabilidad en su compartición y la seguridad ante posibles pérdidas de información. Los objetivos del sistema gestor son proporcionar al usuario abstracción en la información, independencia en las modificaciones, consistencia en las búsquedas e inserciones, seguridad en los permisos de acceso, control de transacciones y minimización del tiempo de acceso (Date, 2003).

La base de datos se encuentra estructurada en diferentes tablas de acuerdo con los tres tipos de almacenes mencionados, que se utilizan para contener toda la información. El conjunto de tablas está relacionado por un identificador que es único para cada edificio. De esta forma, es posible conocer en una sola operación de acceso los datos, el conocimiento y los parámetros relativos a un edificio en concreto.

Otro elemento destacado en la base de datos son las tareas programadas que se ejecutan de forma automática según el intervalo de tiempo indicado. Por ejemplo, se ha definido una tarea que realiza llamadas al módulo de explotación para iniciar el proceso de minería de datos, otra tarea realiza copias de seguridad de forma continua para evitar pérdidas de información, etc.

El sistema gestor controla cualquier acceso a la base de datos. Por ejemplo, las inserciones de datos realizadas por el servicio de adquisición se almacenan de forma ordenada según son recibidas. El sistema gestor de la base de datos es capaz de manejar sin problemas altas velocidades en las inserciones, lo que garantiza la escalabilidad y robustez del sistema. El

SGBD gestiona también las selecciones y búsquedas de información, como por ejemplo, aquellas requeridas por el proceso de minería de datos, o por la supervisión mediante herramientas de visualización.

Una tarea de mantenimiento es imprescindible para garantizar un funcionamiento ágil y rápido del sistema gestor de la base de datos en las búsquedas y liberar espacio para un almacenamiento futuro. Por ejemplo, una extracción masiva de datos se puede llevar a cabo tras un proceso de minería de datos.

3.4.2. Servidor de explotación de datos

El servidor de explotación de datos es el equipo destinado a ejecutar los servicios de minería de datos que extraen conocimiento de los edificios en base a datos reales de instantes de tiempo pasados. Este conocimiento se puede emplear en la supervisión de la energía eléctrica, aportando un valor añadido a la supervisión tradicional. El módulo de explotación consta de 3 servicios de minería de datos diferentes, en función del objetivo perseguido:

- **Servicio de preprocesado**: la función de este servicio es realizar un tratamiento de las muestras erróneas, vacantes, *outliers* y otros valores que provocan efectos perjudiciales en etapas posteriores de la explotación de datos. Por este motivo, siempre se debe ejecutar en primer lugar y complementa a los otros dos servicios.

- **Servicio de exploración**: este servicio está dedicado a la búsqueda de patrones de comportamiento eléctrico de los edificios y relaciones entre variables eléctricas y ambientales y a la comparación entre edificios. Este proceso normalmente se lleva a cabo una vez cada año.

- **Servicio de modelado**: el servicio de modelado es el encargado de obtener un modelo que represente fielmente el comportamiento eléctrico de los edificios. Este modelado se podría realizar una vez cada mes por ejemplo.

Para ejecutar estos servicios se emplea Matlab, software ampliamente reconocido en tareas de procesamiento masivo de datos. Una tarea programada en la base de datos realiza una llamada al módulo de explotación y este inicia la ejecución de los servicios. Previamente se procede a la consulta de los datos necesarios que se encuentran almacenados en la base de datos. Siempre es imprescindible ejecutar en primer lugar el servicio de preprocesamiento y a continuación el de exploración o el de modelado, dependiendo del objetivo perseguido. También se podrían ejecutar ambos a la vez, aunque esto requiere una gran carga computacional y es preferible hacerlo por separado. La ejecución de estos servicios puede durar varias horas o incluso días, en función de la cantidad de datos utilizados. Una vez finalizado el proceso de minería de datos, los resultados se envían hacia la base de datos, donde se almacenan. Posteriormente, se pueden emplear en la supervisión de la energía eléctrica en los edificios o simplemente consultar para realizar un análisis *off-line*.

3.4.3. Servidor web

El servidor web es el equipo cuya función es publicar la información de modo que un usuario remoto pueda supervisar el sistema de energía eléctrica. La información se encuentra almacenada en el servidor de almacenamiento situado dentro de la red de explotación, mientras que un cliente remoto accede al sistema a través de la red de supervisión, por lo que este equipo realiza el enlace entre ambas redes, y en consecuencia entre la capa cliente y la intermedia. Dado que las redes de comunicación en estas dos capas son diferentes e independientes, la actividad en la red de explotación no afecta a la capa cliente y lo que es más importante, que el tráfico en la red de supervisión no influya en la capa intermedia.

Dentro de la red de explotación, se ha definido una subred denominada zona desmilitarizada (*De-Militarized Zone*, DMZ) cuyo objetivo es fortalecer la seguridad del sistema ante posibles intrusos (Dubrawsky *et al.*, 2006). De esta forma, sólo se puede establecer una conexión desde la red de supervisión con el servidor web, restringiendo cualquier otra posible conexión entrante hacia la red de explotación. Para ello, se emplean dos *routers* que actúan como *firewall*, uno de ellos externo que controla el tráfico entre la red de supervisión y la subred DMZ y otro interno que bloquea el tráfico entre la subred DMZ y la red de explotación.

El servidor web de tipo Windows IIS (*Internet Information Server*) ejecuta un servicio web que publica las diferentes herramientas de visualización incrustadas dentro de una página web cuando un usuario las direcciona desde un navegador. Una vez que se detecta el acceso de un usuario habilitado, el servicio web inicia la solicitud de información al servidor de almacenamiento. La transmisión de información entre ambos servidores se produce de forma continua para que la visualización en la interfaz de usuario correspondiente sea correcta. Cabe destacar que cualquier fallo en el servidor web es crítico para el proceso de supervisión. Un servidor web redundante solucionaría el problema como ya se comentó para el servidor de adquisición.

3.5. Capa cliente

En el nivel superior se encuadran todos los equipos dedicados a la supervisión de la energía eléctrica. No existe un puesto centralizado y dedicado a la supervisión, ya que esta se realiza de forma remota a través de la red de supervisión. Cualquier equipo con acceso a esta red, ya sea un PC, PDA, teléfono móvil, etc., se puede utilizar como medio de supervisión. Normalmente, la red de supervisión se extiende a Internet con el fin de dotar al sistema de mejores prestaciones, habilitando la supervisión desde cualquier parte (Qiu *et al.*, 2002; Nagata, 2006). Por tanto, el sistema de supervisión es accesible vía web, lo que permite supervisar el consumo de energía eléctrica desde cualquier institución perteneciente a la Universidad de León y desde el exterior de la misma.

El servidor web publica la información contenida en la base de datos y hace posible la supervisión remota. De esta forma, la información fluye entre las redes de explotación y supervisión, totalmente independientes entre sí. Las aplicaciones o contenidos web

desarrollados para llevar a cabo la supervisión residen y se ejecutan en el servidor web. Estas aplicaciones están constituidas por elementos dinámicos comunes embebidos en páginas HTML, como gráficas, mapas, controles, indicadores, etc., los cuales definen la interfaz de usuario para la supervisión. Las tecnologías empleadas para desarrollar las interfaces gráficas de supervisión son ASP.NET (Liberty y Hurwitz, 2003) y LabVIEW (National Instruments, 2003).

ASP.NET es un marco de trabajo para aplicaciones web dinámicas desarrollado por Microsoft que permite la creación de formularios web utilizando diferentes tipos de controles sencillos. La mayor parte de aplicaciones implican, como en este caso, el acceso a datos y su presentación al cliente en un navegador, para lo que ASP.NET ofrece multitud de funciones. Por otra parte, LabVIEW es un lenguaje de programación gráfico que tiene un gran potencial en el desarrollo de interfaces gráficas. Las aplicaciones creadas en este lenguaje se denominan *Virtual Instruments* (VI) y su panel frontal se puede visualizar fácilmente de forma remota. En el desarrollo de los contenidos web para el sistema de supervisión se han utilizado indistintamente ambas tecnologías. Cabe destacar que nuevas aplicaciones web basadas en otras tecnologías también se podrían integrar en el sistema.

Las aplicaciones web desarrolladas para la supervisión son clasificadas en dos grupos, atendiendo al tipo de información que se emplea en la visualización. Así se puede distinguir entre supervisión tradicional y avanzada. A su vez, dentro de cada uno de estos dos grupos se puede realizar una nueva división, dependiendo del instante de tiempo en el que se realiza la visualización. Entonces, surge el concepto de supervisión *on-line* y *off-line*. A continuación se detallan los tipos de supervisión citados.

3.5.1. Supervisión tradicional

Dentro de la supervisión tradicional se incluyen todas las aplicaciones web que emplean herramientas clásicas en la supervisión. Estas herramientas permiten la visualización de valores instantáneos, valores máximos y mínimos, tendencias de las variables, alarmas o avisos, etc. en tiempo real, es decir, visualizan la muestra actual que ha sido tomada por el medidor. Por tanto, podemos definir este tipo de supervisión como tradicional, la cual se lleva a cabo *on-line*. Por otra parte, las herramientas clásicas de supervisión también permiten recuperar datos históricos para visualizarlos y analizarlos. En este caso, podemos definir la supervisión como tradicional, que se lleva a cabo *off-line*. Además, existe una herramienta que facilita la descarga de datos en formato texto (TXT ó CSV) para su tratamiento posterior mediante una aplicación externa, utilizando por ejemplo técnicas estadísticas. En definitiva, el sistema de supervisión soporta cualquier funcionalidad comúnmente proporcionada por las aplicaciones de supervisión que existen en el mercado.

En la figura 3.7 se muestra un entorno de supervisión desarrollado para llevar a cabo una visualización clásica de la energía eléctrica en los edificios de la Universidad de León. La página web está estructurada en diferentes pestañas (tensiones, corrientes, potencias, armónicos, etc.), dentro de las cuales se muestra una gráfica temporal con las variables seleccionadas, los valores actuales, los máximos y los mínimos indicando el instante de

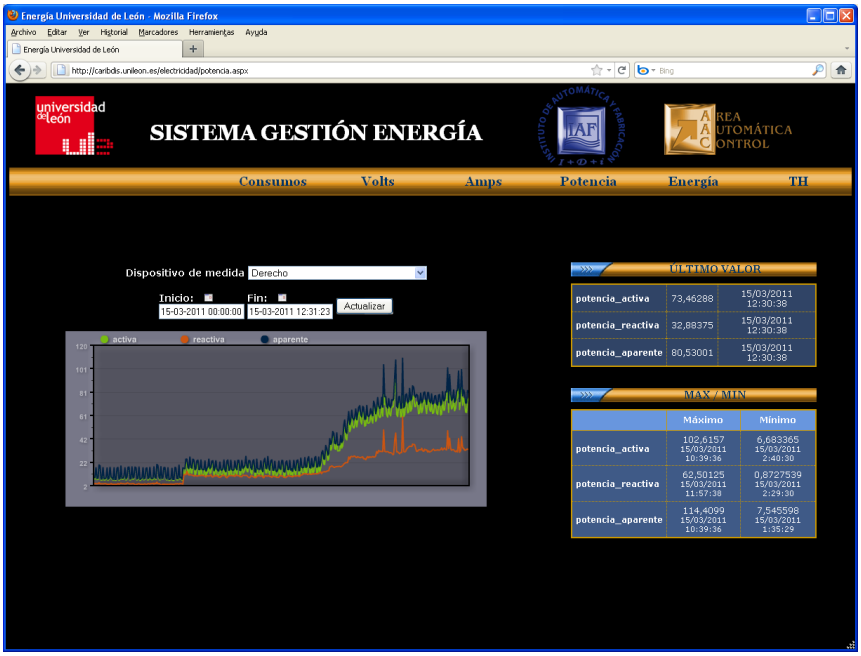


Figura 3.7: Supervisión tradicional de la energía eléctrica en los edificios de la Universidad de León.

tiempo en el que se produjeron. Mediante un menú desplegable se elige el medidor fuente de datos, es decir, el edificio a supervisar.

3.5.2. Supervisión avanzada

Este tipo de supervisión engloba todas las aplicaciones web que emplean herramientas avanzadas que permiten la visualización del conocimiento extraído en el proceso de minería de datos. Cuando se emplea el conocimiento obtenido en el modelado eléctrico de los edificios en una supervisión *on-line*, es posible estimar comportamientos eléctricos futuros, detectar desviaciones, etc., además de visualizar los valores actuales. En cambio, si se utiliza el conocimiento extraído en la exploración de los datos para realizar un análisis visual del comportamiento eléctrico de los edificios basado en gráficos de comparación y grupos, podemos afirmar que la supervisión es de tipo *off-line*, ya que el análisis exploratorio de los datos corresponde a un determinado intervalo de tiempo pasado.

Un entorno de supervisión avanzada aporta un concepto diferente en la visualización, con la ventaja de permitir al usuario razonar y obtener conclusiones en base al conocimiento extraído de instantes pasados y a la muestra actual. Por este motivo, el usuario puede tomar decisiones en tiempo real, durante la supervisión, sin tener que llevar a cabo un análisis de los datos posterior.

En la figura 3.8 se muestra un entorno de supervisión avanzada desarrollado para llevar a cabo una visualización del conocimiento extraído de los edificios. La aplicación web está formada por un conjunto de mapas de colores que tienen en cuenta la periodicidad en

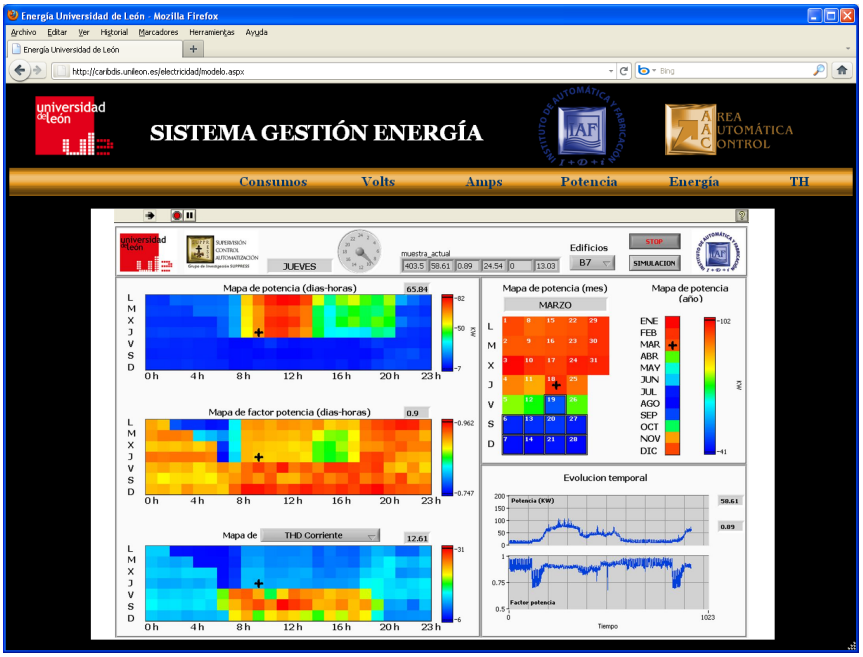


Figura 3.8: Supervisión avanzada de la energía eléctrica en los edificios de la Universidad de León.

el consumo eléctrico y proporcionan valores actuales y estimados por el modelo, valores esperados para un determinado instante, etc., facilitando el razonamiento del usuario. El medidor fuente de datos correspondiente al edificio que se quiere supervisar se puede seleccionar mediante un menú desplegable.

Capítulo 4

Metodología aplicada en la explotación de datos

En este capítulo se expone la metodología aplicada en la explotación de los datos eléctricos procedentes de los edificios del Campus de la Universidad de León. La implementación de la minería de datos sigue el modelo CRISP-DM (*Cross Industry Standard Process for Data Mining*). Inicialmente, se plantea y estudia el problema que surge en el análisis y supervisión de la energía eléctrica en los edificios, de forma conjunta y condicionada por un entorno ambiental común. El sistema de medida desarrollado adquiere y almacena continuamente los datos eléctricos y ambientales, los cuales son estudiados e interpretados con detalle. Además, los datos requieren un preprocesamiento previo al modelado para excluir las muestras erróneas, seleccionar las variables más significativas y evitar que valores con diferentes rangos distorsionen los modelos. En la etapa de modelado, se realiza una exploración de los datos que facilita el examen o análisis general de los mismos y a continuación, se construyen los modelos eléctricos de los edificios. Para ello, en esta tesis se propone el algoritmo envSOM, una variante del SOM básico que permite generar modelos eléctricos condicionados por un entorno ambiental común y comparar los edificios. La metodología adoptada será evaluada mediante una serie de experimentos y finalmente, los resultados obtenidos se aplicarán en la supervisión de la energía eléctrica, haciendo uso de las herramientas avanzadas de visualización propuestas en esta tesis.

4.1. Planteamiento del problema y metodología propuesta

El sistema de medida y supervisión de la energía eléctrica de la Universidad de León permite adquirir y almacenar las variables eléctricas correspondientes a todos los edificios, así como las variables ambientales que caracterizan el entorno común que los rodea. Por tanto, se dispone de un gran volumen de información que reside en la base de datos del sistema. El estudio y la visualización simultánea de dicha información por medio de herramientas de supervisión tradicionales es una tarea ardua y laboriosa, que requiere gran cantidad de recursos y mucho tiempo. Además, la incorporación de nuevos edificios al sistema de medida y supervisión incrementará aún más la complejidad en el manejo y visualización de toda la información. Por otra parte, en los últimos años se han introducido importantes cambios

en el sector eléctrico, tales como la desregulación de las tarifas eléctricas y el impulso de la eficiencia energética en las instalaciones, que hacen necesario disponer de herramientas modernas que permitan la explotación ágil y eficiente de los datos eléctricos para aprovechar las ventajas que ofrece el mercado eléctrico actual. El conocimiento extraído mediante estas herramientas debe ser útil en la negociación con las compañías comercializadoras con el fin de lograr un ahorro energético y económico.

Para explotar el gran volumen de datos disponible y extraer la información intrínseca de los datos eléctricos utilizando estas nuevas herramientas, será necesario tomar como punto de partida el uso de métodos que combinen las características de reducción de la dimensionalidad, cuantificación de vectores y visualización. Los métodos de minería de datos visual basados en SOM pueden ser útiles para esta tarea. El algoritmo SOM es una excelente herramienta para el modelado, agrupamiento y visualización de información cuando se tiene una gran cantidad de datos. Además, el SOM se ha aplicado satisfactoriamente en la supervisión y análisis de sistemas eléctricos (Domínguez *et al.*, 2011). Sin embargo, cuando se requiere descubrir patrones eléctricos condicionados por el entorno, el SOM puede presentar ciertas dificultades a la hora de conseguir una organización de las variables eléctricas de forma condicionada por el ambiente. Además, la organización establecida por el SOM será aún más compleja, si el número de edificios y/o variables eléctricas es elevado como en este caso. Por tanto, la comparación del comportamiento eléctrico de varios edificios no será sencilla, lo que complicará la supervisión y análisis de todos los edificios de forma conjunta y global.

Para resolver este problema, es necesario emplear un método que complemente las cualidades del SOM relativas a la reducción de la dimensionalidad, cuantificación de vectores y visualización con la propiedad de proporcionar mapas condicionados por ciertas variables ambientales que caracterizan el entorno. Dado que el entorno es común para todos los edificios, este algoritmo debe facilitar la comparación entre sus comportamientos eléctricos. En esta tesis se propone una modificación del SOM, denominada envSOM, para extraer patrones de comportamiento eléctrico y generar modelos basados en datos, pero condicionados por variables de tipo ambiental. El algoritmo envSOM debe permitir la supervisión y análisis de la energía eléctrica en todos los edificios de forma conjunta, teniendo en cuenta la influencia del entorno común. La comparación entre perfiles eléctricos de los edificios también debe ser posible. Para ello, se pueden definir herramientas de supervisión avanzadas que permitan visualizar de forma sencilla e intuitiva la información extraída por el envSOM.

Con el fin de facilitar al personal de la Universidad de León las tareas de supervisión y análisis de los datos y proporcionar conocimiento acerca del comportamiento eléctrico de los edificios, se incorporará al sistema de medida y supervisión un módulo de explotación de datos en la capa intermedia. Este módulo debe implementar los métodos de minería de datos basados en el algoritmo envSOM y extraer conocimiento que permita la supervisión avanzada de los edificios. En la figura 4.1 se representa el proceso de explotación de datos. Las variables eléctricas y ambientales contenidas en el almacén de datos, son utilizadas, previo preprocesamiento, por los servicios de exploración y modelado para generar conocimiento

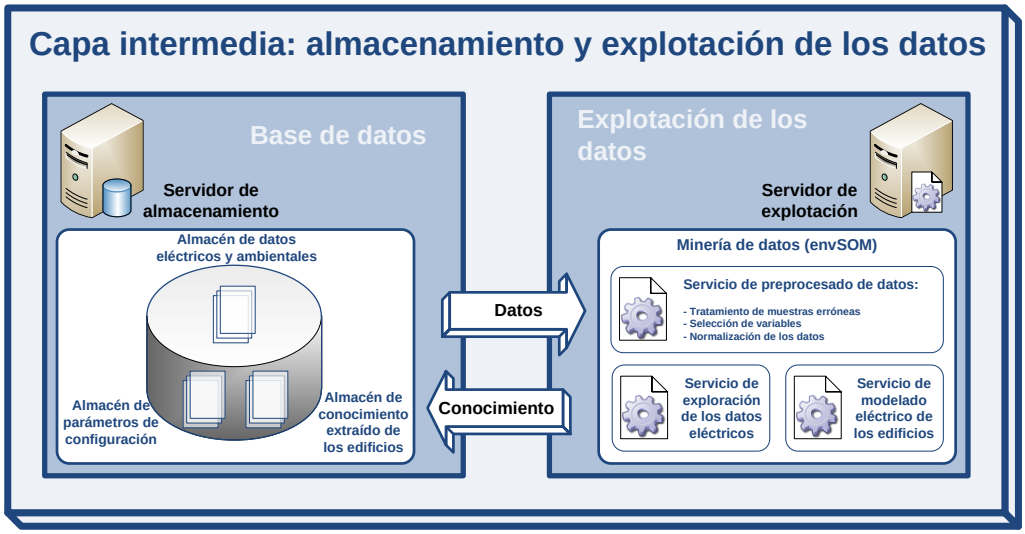


Figura 4.1: Esquema de la explotación de los datos.

acerca de los edificios, el cual se guarda en el correspondiente almacén de la base de datos. Estos servicios implementan la metodología de explotación de datos que se explicará a lo largo del presente capítulo.

4.2. El algoritmo envSOM

El algoritmo envSOM (*environmental SOM*) es una modificación del SOM tradicional, adecuado para la extracción y análisis de información procedente de grandes volúmenes de datos correspondientes a diferentes procesos o sistemas, que son condicionados por un entorno o ambiente común (Alonso *et al.*, 2011c). Este algoritmo conserva las mismas características y ventajas que el SOM tradicional. En este sentido, el envSOM permite obtener mapas ordenados espacialmente, preservando la topología de los datos de entrada. El principal inconveniente es su mayor coste computacional ya que se requiere la ejecución de dos SOMs consecutivos. Además, hay que tener en cuenta que las variables que caracterizan el entorno o ambiente común y que influyen en los procesos o sistemas se deben seleccionar de antemano. En esta tarea, es necesario tener un profundo conocimiento de los procesos o sistemas para realizar una elección adecuada de aquellas variables que realmente ejercen influencia.

La diferencia principal con el SOM tradicional estriba en que este algoritmo tiene en cuenta las condiciones (tiempo, temperatura, humedad, presión, velocidad del viento, radiación luminosa, etc.) que definen el entorno o ambiente común que rodea a varios procesos. Cabe mencionar que cualquier otra variable que influya en todos los procesos se puede considerar también como variable de entorno. Se puede afirmar que el envSOM representa en gran medida la función de probabilidad de los datos de entrada, dadas las condiciones ambientales. Para ello, se han introducido innovaciones en la búsqueda de la neurona ganadora y en la adaptación en las dos fases consecutivas basadas en el SOM

tradicional.

Este algoritmo se puede emplear en el reconocimiento de los patrones ambientales que existen en un conjunto de datos, aunque también se podría utilizar en el descubrimiento de otro tipo de patrones bien distintos de los ambientales. Nuevas muestras de datos o variables que definen los procesos o sistemas se pueden incorporar en la segunda fase, lo que confiere al entrenamiento del envSOM una característica incremental. En este caso, se supone que las condiciones ambientales no sufren variaciones significativas en su rango durante un mismo periodo de tiempo.

4.2.1. Primera fase

En la primera fase, se entrena un SOM tradicional usando todas las variables del conjunto de datos con la peculiaridad que únicamente aquellas que caracterizan el entorno o ambiente se utilizan en el proceso de cálculo de la neurona ganadora. Con esto se logra obtener un preciso modelo de las condiciones ambientales comunes. La inicialización puede ser lineal o aleatoria, dependiendo de las preferencias del usuario. Es necesario conocer a priori las variables ambientales que influyen en el comportamiento del proceso o sistema, ya que solamente estas intervendrán en el cálculo de la neurona ganadora. Esto se logra mediante una máscara binaria $\boldsymbol{\omega}$ que se aplica en el proceso de búsqueda de la neurona ganadora. Al igual que en el SOM básico, la neurona ganadora c se obtiene por medio de la ecuación 4.1.

$$c(t) = \arg\min_i \|\mathbf{x}(t) - \mathbf{m}_i(t)\|_{\boldsymbol{\omega}},\ i = 1, 2, \ldots, M, \tag{4.1}$$

donde $\mathbf{x}$ representa el dato de entrada actual y $\mathbf{m}$ indica los vectores *codebook*. M y t son el número de neuronas o unidades del mapa SOM y el tiempo, respectivamente. El uso de una máscara binaria $\boldsymbol{\omega}$ en la búsqueda de la neurona ganadora marca la diferencia con el SOM básico. Normalmente, la distancia Euclídea $\|^\circ\|$ se emplea en el cálculo de la neurona ganadora (ver ecuación 4.2).

$$\|\mathbf{x}(t) - \mathbf{m}_i(t)\|^2_{\boldsymbol{\omega}} = \boldsymbol{\omega}\|\mathbf{x}(t) - \mathbf{m}_i(t)\|^2 = \sum_k \omega_k \left[x_k(t) - m_{ik}(t)\right]^2 \tag{4.2}$$

La máscara $\boldsymbol{\omega}$ es un vector de dimensión k (número de componentes o variables) cuyos valores ω_k son 1 ó 0, dependiendo si la componente corresponde a una variable ambiental o no. El proceso de adaptación se mantiene similar al SOM tradicional, por lo que todas las variables se actualizan convenientemente. Las componentes ambientales son perfectamente organizadas en esta primera fase. El resto de variables tomarán valores aleatorios porque no se produce ninguna organización en estas componentes. El objetivo de esta primera fase es conseguir un modelo que represente lo mejor posible al entorno o ambiente común. Además, permite obtener una adecuada inicialización para la segunda fase ya que ahora los valores del resto de variables se encuentran en el rango apropiado.

4.2.2. Segunda fase

En la segunda fase, se entrena un nuevo SOM usando todas las variables. Los vectores *codebook* procedentes de la primera fase se utilizan en la inicialización, facilitando la

convergencia del algoritmo. Cabe destacar que las componentes ambientales ya han sido completamente organizadas en la primera fase, logrando un preciso modelo del entorno que condicionará la organización del resto de variables (aquellas que caracterizan el sistema o proceso). En esta fase, todas las componentes se usan en la búsqueda de la neurona ganadora y, a diferencia de la primera fase, no se aplica ninguna máscara. Sin embargo, en esta fase sí se modifica el proceso de adaptación. Sólo es necesario actualizar las variables del proceso o sistema debido a que las ambientales ya han sido organizadas correctamente. Para ello, una máscara binaria $\mathbf{\Omega}$ es introducida en el proceso de adaptación que se lleva a cabo de acuerdo a la ecuación 4.3.

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t)h_{ci}(t)\mathbf{\Omega}[\mathbf{x}(t) - \mathbf{m}_i(t)] \tag{4.3}$$

La máscara $\mathbf{\Omega}$ es un vector de dimensión k (número de componentes o variables) cuyos valores Ω_k son o bien 0 si la componente es de tipo ambiental o 1 en el otro caso. La tasa de aprendizaje $\alpha(t)$ y la función de vecindad $h_{ci}(t)$ no se modifican, de modo que se pueden utilizar un valor decreciente en el tiempo y una función de tipo Gaussiano respectivamente, o bien otro tipo de funciones. Al final de esta segunda fase todas las variables son organizadas correctamente. El objetivo de esta fase final y, por tanto, del algoritmo propuesto es lograr un modelo preciso del conjunto de datos procedentes de uno o varios sistemas, dada la información ambiental común.

4.2.3. Ejemplo sencillo con datos binarios

A continuación, se muestra un pequeño ejemplo con un conjunto de datos binarios creados de forma artificial con el fin de mostrar el funcionamiento del algoritmo envSOM. Esto nos permitirá comprobar la utilidad de este algoritmo para detectar grupos o patrones en los datos de entrada y comparar las componentes o variables entre sí, dadas unas condiciones de entorno comunes.

El conjunto de datos artificial $\mathbf{X}$ está formado por 4 variables (X1, X2, X3, X4) que contienen 16000 muestras cada una (ver figura 4.2). Estas variables han sido creadas de forma estructurada siguiendo un patrón binario, es decir, desde (0,0,0,0) hasta (1,1,1,1). Este patrón se corresponde con los números desde 0 hasta 15, representados en el sistema binario. Un pequeño nivel de ruido (10 %) ha sido añadido a todas las variables de modo que los valores oscilan entre -0.1 y 1.1. Cada patrón binario contiene en total 1000 muestras. La elección de estos datos de prueba ha sido motivada por su sencilla visualización, lo que se traduce en una mejor comprensión del funcionamiento del algoritmo. Utilizando el algoritmo envSOM, se deberían encontrar 16 grupos en el conjunto de datos.

Inicialmente, se entrena un SOM tradicional con los datos artificiales $\mathbf{X}$ como entrada. El número de épocas es suficientemente elevado para garantizar una perfecta organización (500 épocas). Las dimensiones del SOM utilizadas son 16×20 (320 neuronas), la vecindad es de tipo Gaussiano y la tasa de aprendizaje decrece con el tiempo. Como es esperado, el SOM tradicional produce un agrupamiento de los datos en 16 zonas, lo que se puede observar en la matriz de distancias unificada. Sin embargo, cada componente ha sido organizada de

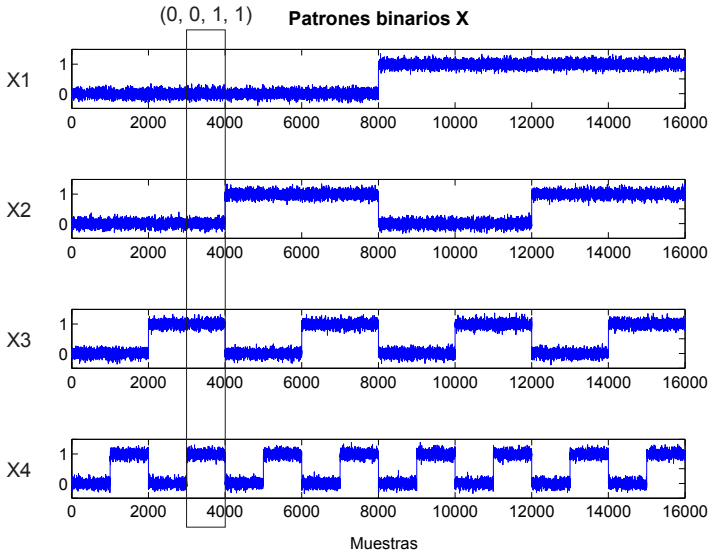
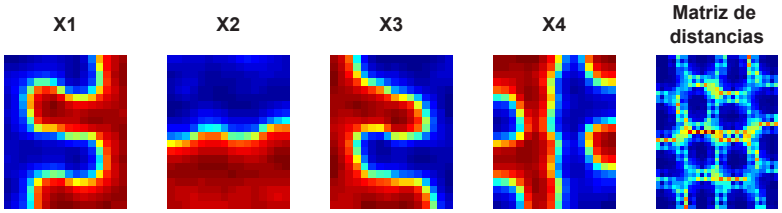


Figura 4.2: Patrones binarios **X**.

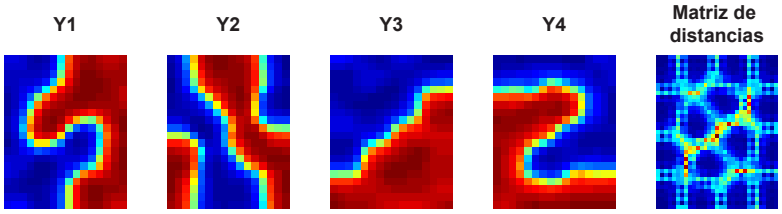
una forma diferente sin seguir ninguna distribución determinada. Si un nuevo SOM fuese entrenado con otro conjunto de datos diferente, pero que sigue los mismos patrones binarios que en el primer caso, por ejemplo (Y1, Y2, Y3, Y4), la organización de sus variables podría ser totalmente diferente. Por tanto, sería muy difícil comparar los resultados de los conjuntos de datos **X** e **Y**. En la figura 4.3 se pueden observar ambos resultados obtenidos aplicando el SOM básico.

Cuando los datos contienen información de las condiciones ambientales que influyen en el proceso o sistema, podría ser útil en el descubrimiento de patrones que la organización de los mapas este condicionada por las variables ambientales. Para ello, se supone que las variables X1 y X2 son de tipo ambiental y las otras dos, X3 y X4, son características del proceso o sistema que se desea analizar. Con estos datos de entrada, se entrena un envSOM cuyos parámetros en ambas fases son idénticos (500 épocas, 320 neuronas, vecindad Gaussiana y tasa de aprendizaje decreciente). En la primera fase, sólo X1 y X2 se usan en el proceso de búsqueda de la neurona ganadora. A continuación, todas las variables de los datos **X** participan por igual en el proceso de adaptación o actualización. Con el fin de mostrar las capacidades del envSOM y comparar los procesos **X** e **Y**, se entrena un nuevo envSOM similar pero utilizando los datos **Y**. El resultado obtenido en esta primera fase para los conjuntos de datos **X** e **Y** se puede observar en la figura 4.4. Se puede ver como únicamente las variables ambientales X1 y X2 han sido correctamente organizadas ya que X3 y X4 no toman parte en el cálculo de la neurona ganadora. La matriz de distancias unificada muestra claramente cuatro patrones, los cuales corresponden a las posibles combinaciones binarias entre las variables X1 y X2. Lo mismo sucede para el conjunto de datos **Y**.

En la segunda fase del envSOM, las cuatro variables de los datos **X** se usan en la búsqueda de la neurona ganadora. A continuación, X1 y X2 se mantienen fijas, mientras que X3 y X4 se adaptan convenientemente. Al final de esta fase, se logra una completa organización en



(a) Patrones binarios **X**



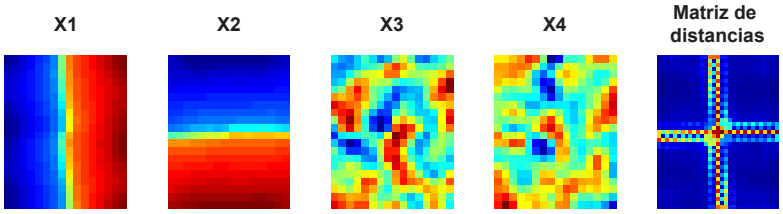
(b) Patrones binarios **Y**

Figura 4.3: Resultado del SOM tradicional para dos conjuntos de datos formados por patrones binarios.

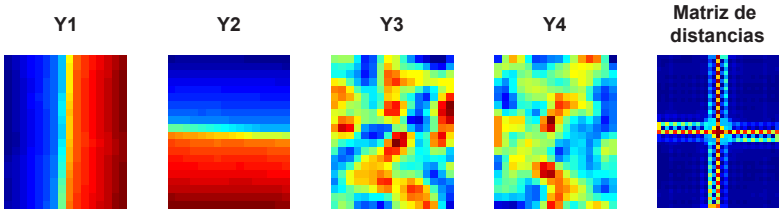
todas las componentes **X**. Para los datos **Y**, también se completa la organización de todas sus variables en esta segunda fase (ver figura 4.5). Al igual que cuando se entrena un SOM básico, ahora se pueden distinguir los 16 patrones binarios existentes en los datos de entrada. Si se compara el error de cuantificación de los mapas construidos por medio del envSOM (0,1717) y el SOM tradicional (0,1364), se tiene que ambos modelos poseen una precisión similar en la aproximación de los datos. La ventaja del envSOM es que permite obtener una organización de las características X3 y X4 condicionada por las variables ambientales X1 y X2, facilitando la visualización de las interacciones entre variables. De la misma forma, en el envSOM que ha sido entrenado utilizando los datos **Y**, como es esperado, se tiene que las componentes Y3 e Y4 están condicionadas por Y1 e Y2.

No obstante, la organización de los mapas con los datos **X** e **Y** aún podría haber resultado diferente, dependiendo del tipo de inicialización, lo que implica una difícil comparación entre ambos conjuntos de datos. Para evitar esto, se puede utilizar el resultado obtenido después de la primera fase del envSOM con los datos **X** para la inicialización de la segunda fase del envSOM con los datos **Y**. Con esto se consigue que la organización de Y3 e Y4 sea similar a X3 y X4, respectivamente. Ahora la comparación entre los resultados procedentes de ambos conjuntos de datos **X** e **Y** será más fácil de llevar a cabo.

Además, cuando las componentes Y1 e Y2 sean exactamente las mismas que X1 y X2 porque representen condiciones ambientales comunes, la primera fase del envSOM se puede entrenar conjuntamente con las variables X1, X2, X3, X4, Y3 e Y4. En este caso, la segunda fase del envSOM se puede realizar mediante un SOM individual para cada conjunto de datos o, por el contrario, un SOM con todas las variables, es decir, X1, X2, X3, X4, Y3 e Y4 (ver figura 4.6). La primera propuesta se puede aplicar siempre, pero cuando el número de variables y/o conjunto de datos es suficientemente pequeño, la segunda propuesta produce resultados similares y por consiguiente será preferida porque requiere un coste computacional

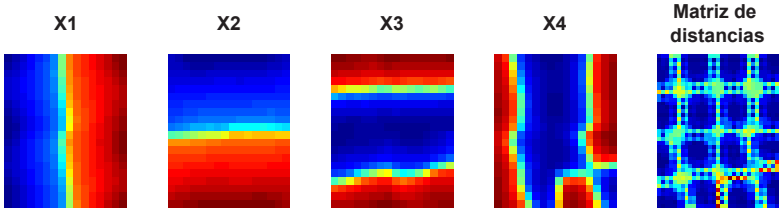


(a) Patrones binarios **X**

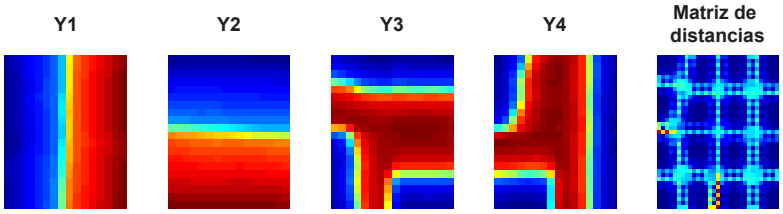


(b) Patrones binarios **Y**

Figura 4.4: Resultado de la primera fase del envSOM para dos conjuntos de datos formados por patrones binarios.



(a) Patrones binarios **X**



(b) Patrones binarios **Y**

Figura 4.5: Resultado de la segunda fase del envSOM para dos conjuntos de datos formados por patrones binarios.

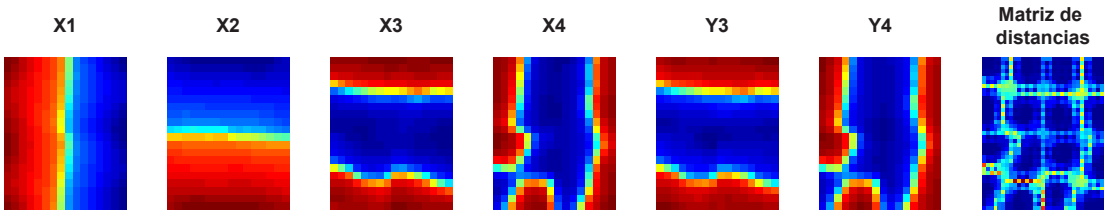


Figura 4.6: Resultado del envSOM para los datos binarios **X** e **Y** de forma conjunta.

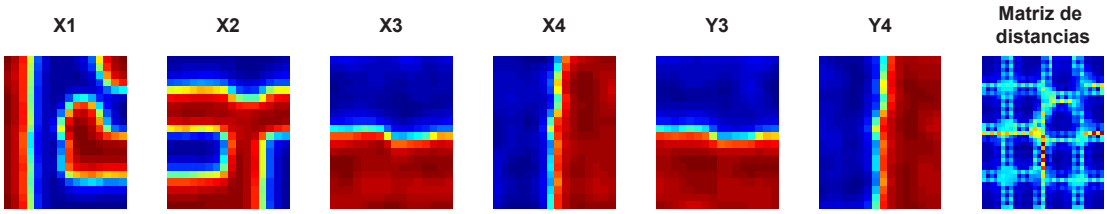


Figura 4.7: Resultado del SOM tradicional para los datos binarios **X** e **Y** de forma conjunta.

menor.

Finalmente, también es interesante comparar el resultado del envSOM con el de un SOM básico que considere los datos **X** e **Y** de forma conjunta. El algoritmo SOM tradicional se puede utilizar en el descubrimiento de patrones correspondientes a varios procesos o sistemas, que están influenciados por el mismo entorno. En este sentido, se puede entrenar un SOM tradicional con todas las variables de los datos **X** e **Y** (X1, X2, X3, X4, Y3 e Y4), siendo X1 e X2 aquellas que influyen en ambos procesos o sistemas. El resultado obtenido se puede observar en la figura 4.7. El SOM básico permite descubrir los 16 patrones binarios y comparar ambos procesos o sistemas **X** e **Y** perfectamente. Pero la diferencia fundamental es que las variables X3, X4, Y3 e Y4 no son organizadas de forma condicionada por las ambientales X1 e X2, comunes a ambos procesos. Por tanto, es difícil visualizar y comparar el comportamiento de los procesos o sistemas, dadas unas condiciones ambientales. Además, en un caso real con un número de procesos y/o variables elevado, las componentes no serán organizadas tan bien y la comparación será más difícil.

4.3. Preprocesamiento de los datos

La metodología propuesta para el preprocesamiento de los datos comprende tres tareas: el tratamiento de las muestras erróneas, la selección de variables y la normalización de los datos (ver figura 4.8). Estas tareas se llevan a cabo siempre de forma secuencial antes de aplicar los métodos de minería de datos. A continuación se explicará cada una de ellas en detalle.

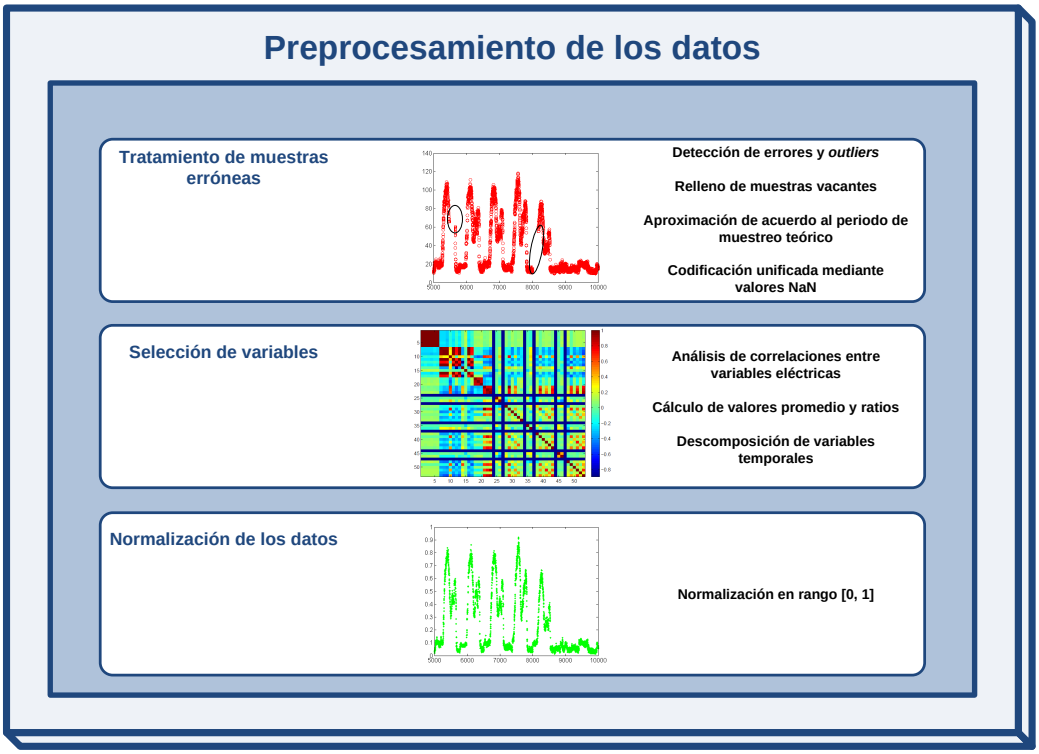


Figura 4.8: Esquema de la metodología adoptada en el preprocesamiento de los datos.

4.3.1. Tratamiento de muestras erróneas

En el proceso de adquisición y almacenamiento de los datos surgen con frecuencia problemas que provocan la existencia de muestras erróneas en la base de datos. Los datos erróneos son siempre un inconveniente en la supervisión debido a la pérdida de información real acerca del proceso o sistema. Cuando se emplean métodos de minería de datos basados en SOM, se requiere información real y fiable en el entrenamiento para producir resultados válidos ya que un número elevado de muestras erróneas provoca efectos negativos. No obstante, se ha demostrado que el SOM puede modelar adecuadamente datos que contienen un pequeño número de valores erróneos o desconocidos (Samad y Harp, 1992). Para ello, estos valores se deben etiquetar previamente con un código de error para que sean ignorados durante el entrenamiento.

Las muestras erróneas adquiridas y almacenadas en la base de datos tienen su origen principalmente en la capa servidor del sistema de medida y supervisión. A continuación, se describen las principales fuentes de error:

- **Medidores**: los fallos en los medidores son muy dispares y dependen del modelo del medidor. En general, los errores procedentes de esta fuente son escasos y extraños. Principalmente, están causados por desconexiones accidentales provocadas por las protecciones que evitan daños en los equipos, colapsos en el microprocesador y la memoria interna, bloqueos en los armónicos durante su cálculo (sólo en los modelos inferiores) y configuraciones erróneas.

- **Red de comunicación**: la red de comunicación introduce pequeñas demoras en la transmisión de los datos, provocando retardos que se suman al periodo de muestreo teórico (1 ó 2 minutos). En los segmentos de red Ethernet este retardo es mínimo, mientras que en los segmentos serie puede ser elevado. Como consecuencia, el tiempo entre dos muestras consecutivas es generalmente un poco mayor que el periodo de muestreo teórico. Con el fin de evitar que el periodo de muestreo real sobrepase en exceso al valor teórico, se establece un tiempo de espera o *timeout* máximo desde que se realiza la petición de datos hasta que se reciben. Si se supera este tiempo, el servicio de adquisición continúa y la muestra se codifica como errónea. También se pueden producir desconexiones de red (fallos en *routers*, *switches*, etc.), lo que se traduce en muestras inexistentes o vacantes. En este caso, la transmisión de datos entre los medidores y el equipo de almacenamiento no es posible, produciendo huecos en los datos. Esta es la fuente más frecuente de errores ya que los tiempos son variables y dependen del tráfico en la red de comunicación.

- **Servidores de adquisición y almacenamiento**: cualquier fallo en los equipos de adquisición y almacenamiento provoca muestras vacantes ya que, o bien no es posible la ejecución del servicio de adquisición, o la base de datos no se encuentra operativa. Este tipo de fallos no es común, aunque cuando se produce suele tener graves consecuencias que implican grandes pérdidas de información procedente de todos los medidores.

En esta tesis se propone llevar a cabo un tratamiento de errores previo a la minería de datos. Para ello, se han desarrollado algoritmos que permiten detectar y codificar todos los tipos de error de forma unificada mediante valores NaN, lo que permite identificarlos y evita su procesamiento en muchas herramientas de tratamiento masivo de datos. Estos algoritmos se ejecutan en modo *batch* de forma que se procesan a la vez todos los datos correspondientes a un periodo de tiempo determinado.

Algunos errores son etiquetados por el propio sistema de adquisición. Por ejemplo, valores -1 indican un exceso en el tiempo de respuesta del medidor, valores -2 denotan un fallo interno del medidor y valores 65535 se registran si el medidor tiene una alta carga computacional (cálculo de armónicos). En este caso, el proceso de detección de errores es sencillo y los anteriores valores se sustituyen por valores NaN. Las muestras vacantes o simplemente huecos también se descubren fácilmente calculando la diferencia de tiempo entre dos muestras consecutivas y comparándola con el periodo de muestro teórico y a continuación, se rellenan con valores NaN. Para eliminar las pequeñas variaciones en el periodo de muestreo, se propone un algoritmo que aproxima la muestra real a la teórica más cercana en el tiempo. El método de interpolación lineal se podría utilizar, aunque requiere un mayor coste computacional y no aporta grandes mejoras. De este modo, se logran datos muestreados con un periodo constante. Finalmente, se propone el uso de un algoritmo que realiza un filtrado paso banda con el fin reemplazar valores atípicos o *outliers* por valores NaN. Para ello, se definen valores umbrales lógicos para las variables a procesar. Por ejemplo, la tensión entre fases oscila entorno a 400 V, considerando como valores atípicos aquellos superiores a 450 V e inferiores a 350 V. Los valores de las variables expresadas en porcentaje

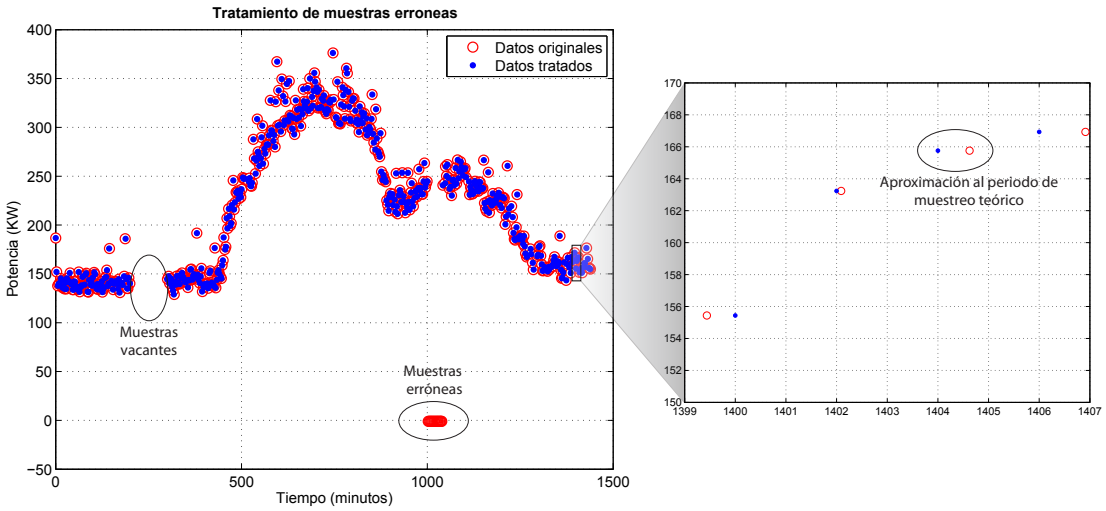


Figura 4.9: Tratamiento de muestras erróneas.

(armónicos, distorsiones armónicas totales, etc.) deben pertenecer al intervalo [0, 100]. En la figura 4.9 se puede observar el proceso de tratamiento de muestras erróneas para la variable potencia.

4.3.2. Selección de variables

El número de variables eléctricas almacenadas es elevado (53) y dispar ya que se dispone de tensiones, corrientes, potencias, factor de potencia, frecuencia, energías, distorsiones armónicas totales y armónicos. Esto genera problemas en el análisis de los datos debido a su dimensionalidad (*Curse of dimensionality*) (Bellman, 1961). Por tanto, una selección de variables es conveniente para reducir inicialmente la dimensión de los datos, sin pérdida significativa de información y proseguir con la minería de datos. Se propone el uso de un análisis de correlación lineal entre variables eléctricas para elegir las más adecuadas para la minería de datos. Este método estadístico utiliza el coeficiente de correlación lineal de Pearson y se basa en medir la cantidad de información no redundante que contiene cada variable (Kendall, 1948). Además, se tiene en cuenta el conocimiento apriorístico experto sobre los edificios y sus variables eléctricas. Por ejemplo, la potencia demandada y el factor de potencia son variables muy importantes en una instalación eléctrica y no se debe prescindir de ellas.

Una vez realizado el análisis de correlaciones, se utiliza un procedimiento de búsqueda en todos los edificios para detectar aquellos pares de variables que poseen valores altos de correlación. Si la relación es directa, indica que ambas variables tienen una evolución similar y será suficiente seleccionar una de ellas. Cuando la relación es inversa, las variables tienen una evolución contrapuesta, es decir, si una aumenta, la otra disminuye. En este caso, una o ambas variables se podrían seleccionar para la minería de datos. Por ejemplo, la potencia estará fuertemente correlacionada con la energía ya que esta es la suma de las potencias

instantáneas a lo largo del tiempo.

Cabe destacar que se pueden calcular nuevas variables a partir de las variables eléctricas almacenadas por el sistema de medida y supervisión. Los valores promedio entre fases para la variable tensión (ver ecuación 4.4) o corriente pueden ser más útiles en ciertas ocasiones que los propios valores de cada fase.

$$V_m = \overline{V} = \frac{(Va + Vb + Vc)}{3} \tag{4.4}$$

Por otra parte, un gran número de factores definidos para medir la calidad de la energía eléctrica utilizan variables eléctricas básicas de cada fase (IEEE Association, 1995). Por ejemplo, la ecuación 4.5 se utiliza con frecuencia para estimar el desequilibrio en corriente (Xu *et al.*, 2010).

$$DC = \left[\frac{max\left(|Ia - Ib|, |Ib - Ic|, |Ic - Ia|\right)}{I_m}\right] * 100 \tag{4.5}$$

I_m indica la corriente promedio entre fases, es decir, $I_m = \overline{I} = \frac{(Ia+Ib+Ic)}{3}$. Las variables que caracterizan la calidad de la energía (distorsiones de la onda y desequilibrios) son muy importantes en el análisis y supervisión de una instalación eléctrica, ya que una baja calidad provoca defectos en el funcionamiento de los equipos o incluso daños irreversibles, así como un exceso en la demanda de potencia total. En un futuro no muy lejano, las compañías eléctricas tendrán en cuenta en la facturación, además de la cantidad de energía consumida, la calidad de la misma (Roncero *et al.*, 2011). Otras relaciones o ratios entre variables con significado físico real se pueden emplear en la minería de datos. En este sentido, el factor de cresta definido como la relación entre la potencia máxima y la media durante un cierto tiempo (ver ecuación 4.6), caracteriza la curva de carga de una instalación eléctrica.

$$FC = \frac{P_{max}}{\overline{P}} \tag{4.6}$$

Como es bien conocido, las variables de tipo meteorológico influyen en el consumo de energía eléctrica. Por este motivo, variables como la temperatura se utilizan siempre en el modelado de datos eléctricos (Beccali *et al.*, 2004; Fan *et al.*, 2006; Tafreshi y Farhadi, 2007). Es lógico pensar que el consumo de los sistemas HVAC y de refrigeración en los edificios varía en función de la temperatura exterior. En la figura 4.10 se puede observar una relación inversa entre la temperatura y la demanda de potencia a lo largo de un año. Esta relación puede ser también directa, es decir, a mayor temperatura mayor consumo, dependiendo del tipo de equipamiento eléctrico en el edificio. Otras variables meteorológicas, tales como la humedad relativa, velocidad del viento, precipitación, radiación solar, radiación luminosa, presión barométrica, etc., se pueden utilizar igualmente en la minería de datos.

No debemos olvidar la característica estacional que posee el consumo de energía eléctrica. La demanda de potencia está condicionada por el mes del año, por el día de la semana y por la hora del día (ver figura 4.11). Las compañías eléctricas tienen en cuenta en la facturación la periodicidad del consumo de electricidad, por lo que las tarifas eléctricas distribuyen la energía consumida en varios periodos (3 ó 6) con precios diferentes, teniendo en cuenta si es

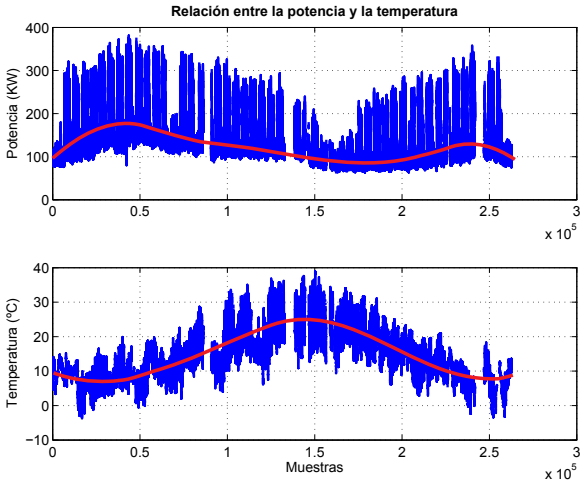


Figura 4.10: Influencia de la temperatura en el consumo de energía eléctrica.

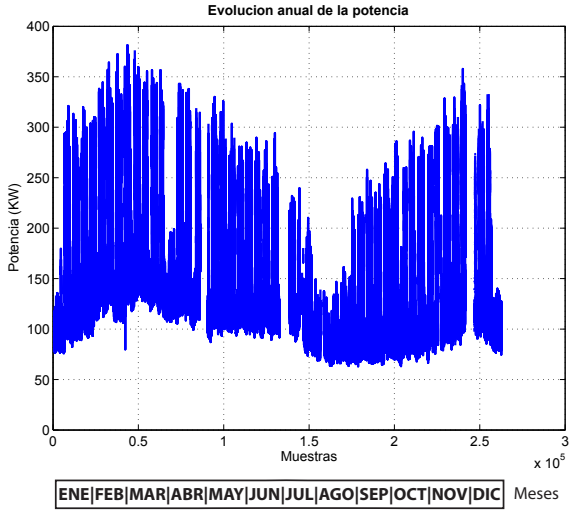
invierno o verano, si es fin de semana u otro día y si es mañana, tarde o noche. Por tanto, el tiempo absoluto se puede desglosar en las correspondientes variables (año, mes, día, hora, minutos y segundos) para considerar eficazmente la periodicidad del consumo eléctrico en la minería de datos.

Cabe destacar que las variables temporales son continuas y sus valores acumulativos se repiten periódicamente. Por ejemplo, el valor de la hora del día comienza en 0, se incrementa hasta llegar a 23 y a continuación vuelve a empezar en 0, pero correspondiendo al día siguiente. Cuando se incluyen variables con este tipo de discontinuidad en el modelado de datos eléctricos basado en SOM, se producen efectos negativos en la organización del mapa. La discontinuidad existente entre los valores extremos de la variable hora (0, 23) genera una distancia Euclídea máxima, lo que se traduce en zonas alejadas en el mapa, cuando en realidad solamente están separadas 1 hora. Además, los modelos deberían agrupar comportamientos similares entre la hora 0 y 23. Para solucionar este problema, se propone descomponer todas las variables temporales en dos nuevas variables que se corresponden con las coordenadas X e Y de la manecilla de horas de un reloj (Carpinteiro *et al.*, 2004; González y Zamarreño, 2005; Fan *et al.*, 2009). Las variables Hx y Hy obtenidas con esta transformación son continuas y periódicas, ya que se corresponden a las funciones seno y coseno del ángulo α que forma la manecilla con la vertical (ver ecuaciones 4.7).

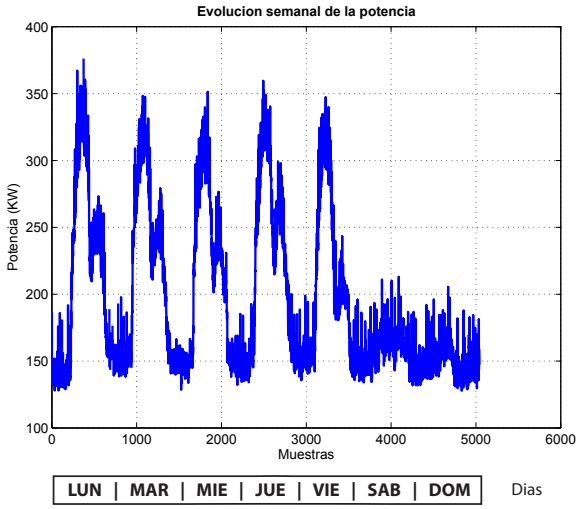
$$\begin{cases} Hx = \sin(\alpha) = \sin\left(\frac{2\pi H}{24}\right) \\ \\ Hy = \cos(\alpha) = \cos\left(\frac{2\pi H}{24}\right) \end{cases} \tag{4.7}$$

De forma análoga, se podría definir esta transformación para otras variables temporales como por ejemplo el día de la semana, obteniendo un reloj para los días (ver ecuaciones 4.8).

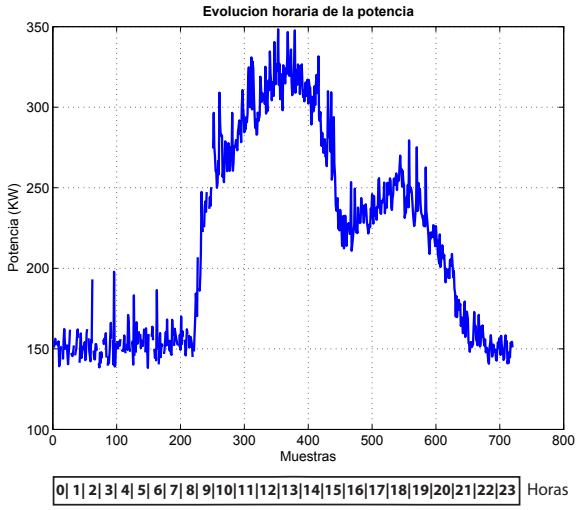
$$\begin{cases} Dx = \sin(\beta) = \sin\left(\frac{2\pi D}{7}\right) \\ \\ Dy = \cos(\beta) = \cos\left(\frac{2\pi D}{7}\right) \end{cases} \tag{4.8}$$



(a) Anual



(b) Semanal



(c) Horaria

Figura 4.11: Periodicidad en el consumo de energía eléctrica.

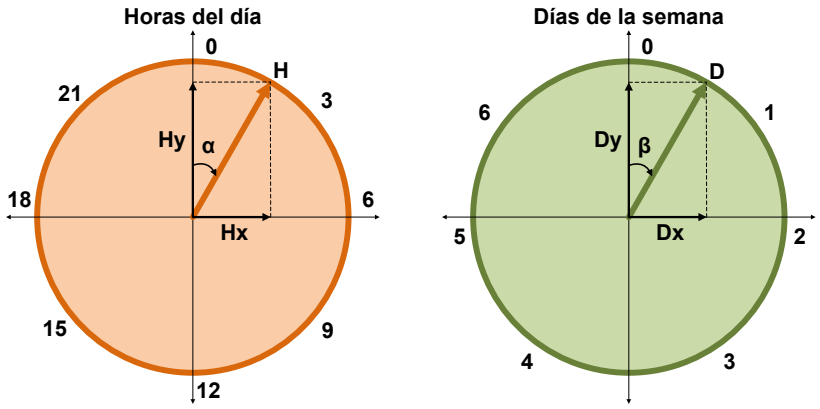


Figura 4.12: Transformación aplicada a las variables temporales (hora y día).

β es el ángulo que forma la manecilla del reloj con la vertical (día 0). Los días son codificados desde 0 (Domingo) hasta 6 (Sábado) y se descomponen en dos coordenadas, Dx y Dy. En la figura 4.12 se representan de forma gráfica las transformaciones indicadas anteriormente.

Independientemente del tiempo absoluto, el consumo de energía eléctrica puede variar bruscamente en función del tipo de día. Por ejemplo, los días festivos o vísperas tendrán un consumo totalmente diferente a los días laborables. Por otra parte, los periodos de vacaciones también se deberían tener en cuenta ya que la actividad de personas en los edificios será casi nula. Una variable que distinga entre un tipo de día y otro es necesaria en el modelado de los datos eléctricos (Fan *et al.*, 2006; Tafreshi y Farhadi, 2007).

4.3.3. Normalización de los datos

El preprocesamiento de los datos incluye como paso final la normalización de los mismos. La normalización es una transformación que se aplica a los datos con anterioridad a los métodos de minería de datos para evitar que las variables con mayor valor relativo tengan más influencia o peso en el entrenamiento. Concretamente en los métodos basados en SOM, la distancia Euclídea utilizada en el proceso de cálculo de la neurona ganadora magnifica las magnitudes grandes. Por ejemplo, el factor de potencia pertenece al intervalo [0, 1], mientras que la potencia puede oscilar entre 0 y 300 KW. Por tanto, el peso de la potencia será 300 veces superior al del factor de potencia en el cálculo de la neurona ganadora. Este efecto no deseado se elimina con una normalización en rango [0, 1] para todas las variables. Este tipo de normalización realiza un escalado lineal entre el valor máximo y el mínimo de cada variable x según la ecuación 4.9.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{4.9}$$

Otros límites en el rango tales como [-1, 1] también se podrían utilizar, al igual que otros tipos de normalización que tengan en cuenta la varianza o la frecuencia de aparición de los datos (Vesanto *et al.*, 2000). Se supone que todos los errores han sido detectados y

codificados adecuadamente en la etapa previa de tratamiento de errores. De lo contrario, un valor atípico podría distorsionar el rango de la variable y, por consiguiente, el resultado de este tipo de normalización.

4.4. Exploración de los datos eléctricos procedentes de todos los edificios

La exploración tiene como objetivo realizar un análisis general de los datos eléctricos procedentes de todos los edificios. El uso de variables ambientales comunes a todos los edificios permite reconocer fácilmente los patrones eléctricos y encontrar las relaciones entre variables eléctricas y ambientales. Además, un análisis conjunto basado en una comparación entre todos los edificios permite detectar similitudes en el comportamiento eléctrico de los edificios (Alonso *et al.*, 2011a). Para llevar a cabo la exploración, se utilizarán métodos de minería de datos que combinan la propiedades de reducción de la dimensionalidad, cuantificación de vectores y visualización. La metodología adoptada se puede observar en la figura 4.13. El algoritmo envSOM se utiliza para reducir la dimensión de los datos de entrada, obtener mapas topológicamente ordenados para todos los edificios y así descubrir los patrones eléctricos en base al entorno común mediante herramientas de visualización. La comparación tiene como punto de partida las matrices de similitud entre edificios que se calculan a partir de los vectores *codebook* procedentes del envSOM, logrando una nueva reducción de la dimensión. Las similitudes entre edificios se visualizan aplicando una proyección de Sammon a partir de las matrices. De forma adicional, un agrupamiento basado en el método *k-means* se emplea para conocer los edificios que tienen un comportamiento eléctrico similar.

4.4.1. Búsqueda de patrones eléctricos condicionados por un ambiente común

En esta etapa de la exploración, se propone el uso del algoritmo envSOM para descubrir patrones eléctricos influenciados por un entorno común en los datos procedentes de todos los edificios. El SOM se podría utilizar perfectamente en la búsqueda de patrones, pero la ventaja del envSOM es que proporciona una organización similar en todos los mapas correspondientes a los edificios, dado un entorno ambiental común. Por tanto, se puede llevar a cabo una comparación visual utilizando, por ejemplo, los planos de componentes, ya que las neuronas localizadas en una misma zona representan a datos similares. Esto permite dar un paso más en la exploración de los datos, haciendo posible el cálculo de las matrices de similitud entre edificios a partir de los vectores prototipo de las neuronas.

En la primera fase del envSOM, se logra un modelo que caracteriza el entorno ambiental que influye en el consumo eléctrico de los edificios. En los datos de entrada utilizados en esta fase se distinguen dos tipos de variables, las ambientales VA y las eléctricas VE para cada edificio e, de forma que el vector de datos de entrada $\mathbf{x}$ se puede expresar mediante la

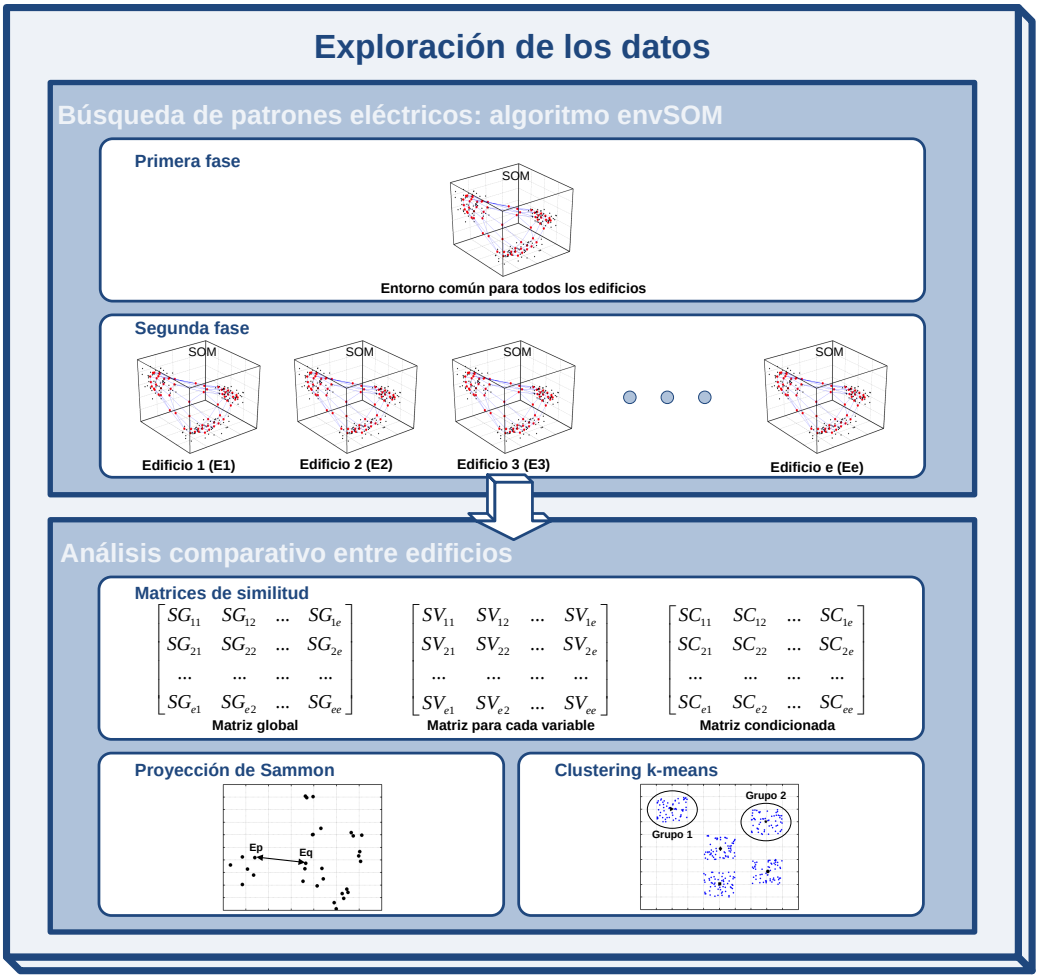


Figura 4.13: Esquema de la metodología adoptada en la exploración de los datos.

ecuación 4.10.

$$\begin{aligned}\mathbf{x} = \Big[& VA_1\ VA_2\ \ldots\ VA_l \,\vdots\!\vdots \\ & VE_1^1\ VE_2^1\ \ldots\ VE_k^1 \,\vdots\, VE_1^2\ VE_2^2\ \ldots\ VE_k^2 \,\vdots\, \ldots \,\vdots\, VE_1^e\ VE_2^e\ \ldots\ VE_k^e\Big]\end{aligned} \tag{4.10}$$

El número de variables ambientales l y eléctricas k utilizadas en la búsqueda de patrones puede variar, dependiendo de la selección de variables realizada. En el entrenamiento de esta primera fase se debe aplicar una máscara $\boldsymbol{\omega}$ para excluir a las variables eléctricas del proceso de cálculo de la neurona ganadora. Esta máscara es la siguiente:

$$[1\ 1\ \ldots\ 1 \,\vdots\!\vdots\, 0\ 0\ \ldots\ 0 \,\vdots\, 0\ 0\ \ldots\ 0 \,\vdots\ \ldots\ \vdots\, 0\ 0\ \ldots\ 0]$$

Una vez logrado el modelo ambiental común, se requiere obtener un modelo de los datos eléctricos para cada edificio. Para ello, se propone una segunda fase del envSOM de forma individual para cada edificio. En esta segunda fase, se logra organizar los datos eléctricos procedentes de todos los edificios, pero de forma condicionada por el entorno ambiental común que influye en ellos. En este caso, el vector de datos de entrada $\mathbf{x}$ que es usado para obtener el modelo eléctrico para el edificio e viene dado por la ecuación 4.11.

$$\mathbf{x}^e = \Big[VA_1\ VA_2\ \ldots\ VA_l \,\vdots\!\vdots\, VE_1^e\ VE_2^e\ \ldots\ VE_k^e\Big] \tag{4.11}$$

Las variables ambientales VA son el resultado de la primera fase, mientras que las variables eléctricas VE se adaptan convenientemente por medio de la máscara $\boldsymbol{\Omega}$ que se muestra a continuación:

$$[0\ 0\ \ldots\ 0 \,\vdots\!\vdots\, 1\ 1\ \ldots\ 1]$$

El envSOM no sólo permite descubrir los patrones ambientales existentes en los datos eléctricos, sino que además localiza aquellos patrones idénticos en zonas o posiciones similares en los mapas correspondientes a cada edificio. Si existen patrones temporales que caracterizan el entorno ambiental, estos también serán capturados por el envSOM sin tener que recurrir a métodos específicos de procesamiento de series temporales. Esto facilita la minería de datos y comparación visual entre todos los edificios.

4.4.2. Análisis comparativo entre edificios

El análisis comparativo entre edificios consta de tres etapas: cálculo de las matrices de similitud, proyección de similitudes y agrupamiento de los edificios según su comportamiento eléctrico.

Similitud entre edificios. Utilizando el resultado del envSOM, es posible comparar de forma visual todos los edificios. Por ejemplo, se podrían utilizar los planos de componentes de las variables eléctricas para esta tarea. No obstante, la dimensionalidad de la información puede ser todavía considerable ya que depende del número de variables eléctricas y edificios, lo que implica un esfuerzo y tiempo elevados durante el proceso de comparación. Por tanto,

una nueva reducción de la dimensionalidad es necesaria para facilitar la exploración de los datos. Para ello, se propone calcular matrices de similitud entre edificios a partir de los vectores *codebook* normalizados procedentes del envSOM. Estas matrices, cuya dimensión es $e \times e$, contienen información acerca de la similitudes o diferencias entre comportamientos eléctricos de los edificios. La métrica utilizada en el cálculo de las matrices de similitud es la distancia L_1 o *cityblock*. Otras métricas, como por ejemplo la distancia Euclídea, no son recomendadas debido a que resaltan las grandes disimilitudes y atenúan las pequeñas. La correspondencia establecida por el envSOM entre componentes eléctricas permite calcular distancias directamente entre neuronas localizadas en la misma posición en los mapas que agrupan datos con condiciones ambientales idénticas. A continuación se indican los tres tipos de matrices calculadas:

- **Matriz de similitud global**: esta matriz aglutina las distancias L_1 entre todas las componentes eléctricas de los vectores *codebook* de los edificios. Por tanto, la matriz global contiene información de las similitudes o diferencias entre los edificios en base a un compendio de variables eléctricas. La similaridad entre un edificio p y otro q se calcula mediante la ecuación 4.12, donde m_{ik} corresponde al peso de la neurona i y la componente o variable eléctrica k.

$$SG_{pq} = L_1\Big(E(p), E(q)\Big) = \sum_k \sum_i \Big(m_{ik}^p - m_{ik}^q\Big) \tag{4.12}$$

- **Matriz de similitud para cada variable**: esta matriz posee información de las similitudes o diferencias entre los edificios en base a una variable eléctrica individual. Por tanto, las distancias L_1 se obtienen sólo para una determinada componente eléctrica. Cabe destacar que es posible calcular tantas matrices de este tipo como componentes o variables eléctricas existan utilizando la ecuación 4.13.

$$SV_{pq}^k = L_1\Big(E^k(p), E^k(q)\Big) = \sum_i \Big(m_{ik}^p - m_{ik}^q\Big) \tag{4.13}$$

  En este caso, k corresponde a una variable determinada que ha sido elegida para obtener las similitudes.

- **Matriz de similitud condicionada**: esta matriz proporciona información de las similitudes o diferencias entre edificios en base a una o todas las componentes eléctricas, pero condicionadas por las variables ambientales. En el cálculo de las distancias L_1 no intervienen los pesos de todas las neuronas ya que algunas quedarán excluidas por la condición impuesta. Dependiendo de la variable ambiental y límites que definen la condición, se pueden obtener infinitas matrices de este tipo. En la ecuación 4.14 se muestra una matriz de similitud para la variable eléctrica k, condicionada por la variable ambiental l.

$$SC_{pq}^k = L_1\Big(E^k(p), E^k(q)\Big) \mid l = \sum_i \Big(m_{ik}^p - m_{ik}^q\Big) \;\mid\; \nu_1 < l < \nu_2 \tag{4.14}$$

  ν_1 y ν_2 son los límites inferior y superior de la condición.

Proyección de las similitudes. Los tres tipos de matrices indicados anteriormente contienen información de las similitudes o diferencias y se pueden utilizar para comparar los edificios entre sí. Sin embargo, es relativamente difícil interpretar la información de estas matrices de forma visual. Por tanto, se propone proyectar esta información en un espacio de visualización 2D de forma que sea sencilla su comprensión. Para ello, se aplica la proyección de Sammon (Sammon, Jr., 1969) que intenta mantener las distancias reales de los datos originales en el espacio de salida, dando énfasis a pequeñas diferencias entre edificios. No obstante, otras técnicas de proyección no lineal se podrían aplicar en esta etapa (Lee y Verleysen, 2007). Con esto se logra una nueva reducción de la dimensionalidad de manera que cada edificio se representa por un punto en el espacio 2D de visualización. La comparación entre edificios se basa en las distancias relativas entre puntos. La proximidad de dos puntos en el espacio de salida indica que los edificios a los que representan poseen un comportamiento eléctrico similar.

Agrupamiento de edificios. Una etapa complementaria de agrupamiento podría ser útil en el análisis comparativo para dividir los edificios en grupos de acuerdo a su comportamiento eléctrico y detectar el número de perfiles eléctricos distintos que existen. Se propone aplicar el método *k-means* (MacQueen, 1967) a partir de las matrices de similitud o directamente de los vectores *codebook* procedentes del envSOM, que realiza una sustancial compresión de los datos (Vesanto y Alhoniemi, 2000). La aplicación de este algoritmo a partir de los datos eléctricos originales sería inviable debido al elevado tiempo y coste computacional que requiere. Además, los datos originales no poseen una organización condicionada común, a diferencia de los pesos de las neuronas del envSOM. Cabe destacar que son posibles tres tipos de agrupamientos, dependiendo de la matriz de similitud utilizada. Con el fin de lograr el número de grupos o *clusters* que mejor se adapte a los datos de entrada, se propone realizar varias ejecuciones del algoritmo con diferentes valores para este número. A continuación, se debe realizar una evaluación de las diferentes ejecuciones para elegir el número óptimo de grupos. El error medio entre los datos y el centroide del grupo correspondiente se puede emplear en esta evaluación. Generalmente, a mayor número de grupos, menor será el error y mejor el agrupamiento.

4.5. Modelado del comportamiento eléctrico de los edificios

El modelado basado en datos tiene como objetivo capturar el comportamiento eléctrico de los edificios de forma que se pueda utilizar en la supervisión *on-line*. Si los modelos son construidos con datos eléctricos adquiridos en situaciones normales, entonces se pueden usar en la detección de desviaciones o incluso en la predicción de futuros comportamientos. El algoritmo envSOM generalizado a n fases se puede utilizar para construir los modelos con cierto nivel de detalle. Este método permite construir modelos eléctricos condicionados por varios subconjuntos de variables ambientales, estableciendo una jerarquía en la organización de los mapas (Alonso *et al.*, 2011b). La metodología seguida para lograr el modelado eléctrico de los edificios se expone gráficamente en la figura 4.14.

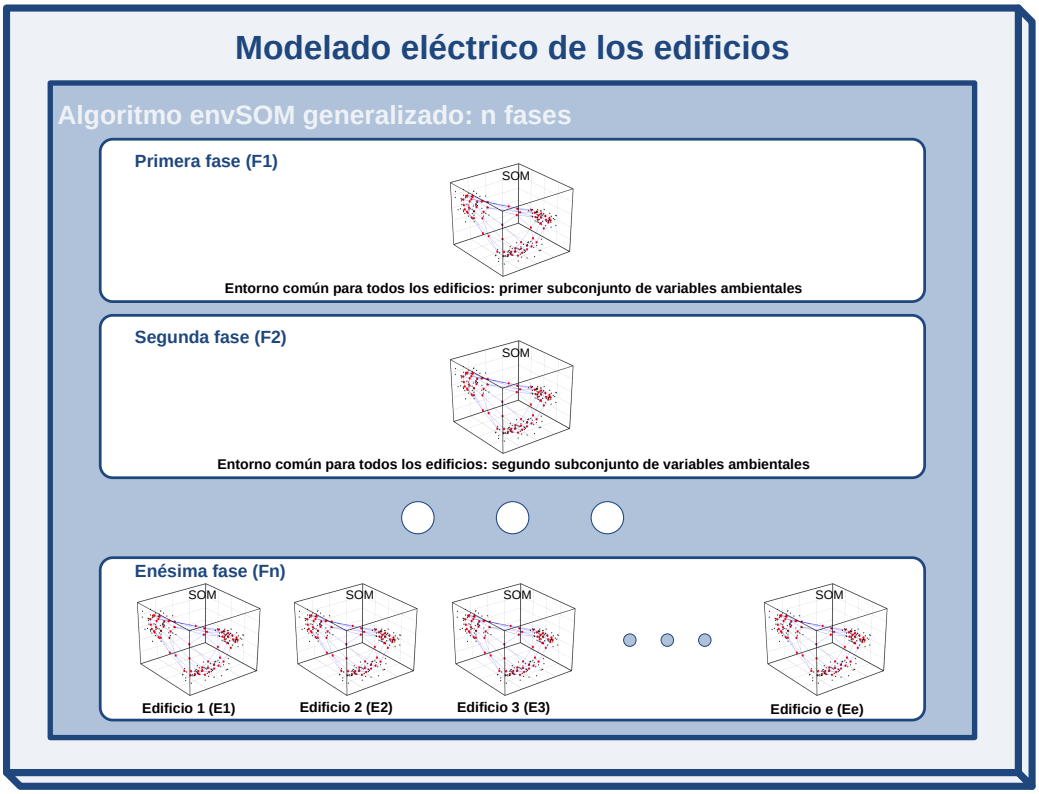


Figura 4.14: Esquema de la metodología adoptada en el modelado eléctrico de los edificios.

4.5.1. El algoritmo envSOM generalizado a n fases

El entrenamiento de dos o más envSOMs consecutivos puede ser interesante cuando existen varios subconjuntos de variables ambientales que definen un entorno común. Por ejemplo, en el entorno común se pueden distinguir variables temporales y meteorológicas. A su vez, las temporales se pueden dividir en año, mes, día, hora, etc. y así sucesivamente. Esta clasificación puede ayudar a mejorar la organización condicionada de las variables que caracterizan un proceso o sistema, adoptando una estructura jerárquica de forma que los niveles superiores condicionan a los inferiores. En el caso de dos envSOMs, el número total de fases basadas en SOM utilizadas es 4 ya que cada envSOM consta de dos fases. Si el número de subconjuntos de variables ambientales es elevado, el número de fases SOM también lo será, implicando un mayor tiempo y coste computacional en el entrenamiento. Si se tienen f subconjuntos de variables, el número de fases SOM será el doble, $n = 2f$. Con el objetivo de reducir el número de fases SOM requeridas y optimizar recursos, se propone aglutinar la segunda fase del primer envSOM y la primera fase del segundo envSOM, de forma que en una única fase se realice la misma función. Esta idea se puede extender a concatenaciones de más de dos envSOMs. El algoritmo envSOM generalizado es más eficiente debido a que el número de fases SOM que requiere es $n = f + 1$. A continuación, se explican en detalle las n fases del algoritmo envSOM generalizado:

- **Primera fase**: el objetivo de esta primera fase es conseguir un modelo que represente

a un entorno común caracterizado por el primer subconjunto de variables ambientales. Para ello, se entrena un SOM tradicional usando todas las variables, pero únicamente aquellas que definen este primer entorno se utilizan en el cálculo de la neurona ganadora. Esto se logra mediante una máscara binaria[1] $\boldsymbol{\omega}^{(1)}$ que excluye al resto de variables en la búsqueda de la neurona ganadora. Los valores $\omega_k^{(1)}$ son 1 ó 0, dependiendo si la componente corresponde a una variable ambiental del primer subconjunto o no. Esta fase del envSOM generalizado es similar a la primera fase de un envSOM original de dos fases, es decir, que se utilizan las ecuaciones 4.1 y 4.2 en el proceso de búsqueda de la neurona ganadora, mientras que el proceso de adaptación de las variables no se modifica y por consiguiente, todas ellas se actualizan convenientemente.

- **Segunda fase y sucesivas**: el objetivo de la segunda fase es lograr un modelo que defina un entorno común caracterizado por el segundo subconjunto de variables ambientales. Cabe destacar que el modelo construido en esta fase se encuentra en un nivel jerárquico inferior al modelo del primer entorno obtenido en la fase anterior. De acuerdo a esta jerarquía, la organización de este subconjunto de variables ambientales será condicionada por las variables pertenecientes al primer subconjunto. Un nuevo SOM básico se entrena usando todas las variables. Los vectores *codebook* procedentes de la primera fase se utilizan en su inicialización. En el proceso de búsqueda de la neurona ganadora intervienen únicamente las variables ambientales de este segundo subconjunto, mientras que en el proceso de adaptación todas las variables se actualizan, excepto aquellas que pertenecen al primer subconjunto, que ya han sido perfectamente organizadas en la fase anterior. Por tanto, los valores binarios de la máscara $\boldsymbol{\omega}^{(2)}$ son 1 cuando la componente corresponde a una variable ambiental del segundo subconjunto y 0 en cualquier otro caso. En cambio, los valores binarios de la máscara $\boldsymbol{\Omega}^{(2)}$ son 0 si la componente pertenece al primer subconjunto de variables y 1 en cualquier otro caso. Las ecuaciones 4.1 y 4.2 definen el cálculo de la neurona ganadora y la ecuación 4.3 la adaptación. Se puede decir que esta fase es el resultado de combinar la segunda fase del primer envSOM y la primera fase del segundo envSOM, suponiendo que se entrenan dos envSOMs consecutivos. El efecto que se produce es idéntico, con la ventaja de reducir el número de fases SOM lo que implica una disminución del tiempo de entrenamiento y del coste computacional. Las sucesivas fases (tercera, cuarta, quinta, etc.) hasta la penúltima fase $(n-1)$ son similares a esta segunda fase, por lo que se omite su descripción.

- **Última fase**: el objetivo de la enésima fase (n) es lograr un modelo preciso del conjunto de datos procedentes de uno o varios procesos o sistemas, dadas las condiciones del entorno común que han sido divididas en $f = n - 1$ subconjuntos con variables ambientales. El modelo construido en esta fase se encuentra en el nivel más bajo de la jerarquía establecida en el organización. Por tanto, las variables que caracterizan al proceso o sistema son condicionadas jerárquicamente por todos los

[1] Los superíndices de las máscaras $\boldsymbol{\omega}$ y $\boldsymbol{\Omega}$ indican la fase del envSOM a la que corresponden.

subconjuntos de variables ambientales definidos. En esta fase final, se entrena un nuevo SOM usando todas las variables. Igualmente que en otras fases intermedias, la inicialización se lleva a cabo con los vectores *codebook* procedentes de la fase anterior. En el proceso de búsqueda de la neurona ganadora intervienen todas las variables y no se enmascara ninguna, pero el proceso de adaptación si se modifica, siguiendo la ecuación 4.3. Dado que las variables ambientales de todos los subconjuntos ya han sido organizadas correctamente, sólo es necesario actualizar las variables propias del sistema. Para ello, se utiliza la máscara $\mathbf{\Omega}^{(n)}$, cuyos valores binarios son 0 si la componente corresponde a una variable ambiental y 1 si es una variable del sistema. Esta fase es similar a la segunda fase de un envSOM original de dos fases.

4.5.2. Generación de los modelos eléctricos

En esta tesis se propone construir modelos eléctricos de los edificios por medio del algoritmo envSOM generalizado, cuyo número de fases n depende del número de subconjuntos de variables ambientales f creados ($n = f + 1$). Primeramente, es necesario clasificar en diferentes subconjuntos las variables que caracterizan el entorno común. Generalmente, cuanto más fina sea esta división, es decir, mayor número de subconjuntos de variables ambientales se tengan, mayor será la resolución de los modelos. En contra, se tiene que el número de neuronas y fases del envSOM deberán ser mayores, lo que implica un tiempo y coste computacional elevados en el entrenamiento.

El algoritmo envSOM generalizado proporciona modelos eléctricos que están condicionados por todos los subconjuntos de variables ambientales de forma jerárquica. La organización de las variables ambientales pertenecientes a un determinado subconjunto está condicionada por todas las variables de los subconjuntos superiores. Por tanto, es muy importante el orden de los subconjuntos ya que de él depende la organización final de las variables eléctricas. Cabe destacar que las variables temporales (año, mes, día, hora, minutos, segundos, etc.) mantienen rigurosamente este orden ya que 60 segundos son 1 minuto, 60 minutos son 1 hora, 24 horas son 1 día y así sucesivamente. Esto permite pensar en la posibilidad de clasificarlas en diferentes subconjuntos para considerar la periodicidad en el consumo eléctrico. Otras variables, como por ejemplo la temperatura media de cada año, mes, día, hora, etc., también se pueden añadir a los subconjuntos de variables temporales.

No se debe olvidar que los modelos eléctricos se generan a partir de datos reales procedentes de situaciones o estados pasados. Los métodos de minería de datos requieren un conjunto de datos de entrada que contenga información relativa a las situaciones eléctricas más significativas. Siempre que el conjunto de datos de entrada sea grande, se puede reducir su tamaño seleccionando solamente algunos datos, pero sin prescindir de aquellos más representativos para el modelado. Tanto el tamaño como la riqueza de los datos son cruciales en la generación de los modelos eléctricos. En el modelado es muy común dividir el conjunto de datos en un subconjunto de entrenamiento y en otro de prueba. El primer subconjunto se emplea en la generación de los modelos eléctricos y el segundo en su evaluación (Hand *et al.*, 2001; Bishop, 2006). Medidas de error como por ejemplo el MAPE (*Mean Absolute*

Percentage Error) y el RMSE (*Root Mean Squared Error*), mostradas en la expresión 4.15, se utilizan comúnmente para evaluar el grado de ajuste o aproximación de los modelos eléctricos (Carpinteiro *et al.*, 2004; Fan *et al.*, 2009).

$$\begin{cases} MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{A_i - B_i}{A_i}\right| \\ \\ RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - B_i)^2} \end{cases} \tag{4.15}$$

A_i corresponde a los datos reales de prueba, B_i son los datos del modelo y N indica el número total de datos. El objetivo final es obtener para cada edificio un modelo eléctrico que posea una buena interpretabilidad, y a su vez una precisión aceptable.

Modelo común para el primer subconjunto de variables ambientales. El primer subconjunto de variables ambientales $VA^f,\ f = 1$ se modela en esta primera fase del envSOM generalizado. El número de variables que forman el subconjunto puede variar entre 1 y l_1. El vector de datos de entrada $\mathbf{x}$ para el entrenamiento se expresa de forma general en la ecuación 4.16.

$$\begin{aligned} \mathbf{x} = \Big[& VA_1^1\ VA_2^1\ \ldots\ VA_{l_1}^1 \vdots\vdots VA_1^2\ VA_2^2\ \ldots\ VA_{l_2}^2 \vdots\vdots \ldots \vdots\vdots VA_1^f\ VA_2^f\ \ldots\ VA_{l_f}^f \vdots\vdots \\ & VE_1^1\ VE_2^1\ \ldots\ VE_k^1 \vdots VE_1^2\ VE_2^2\ \ldots\ VE_k^2 \vdots \ldots \vdots VE_1^e\ VE_2^e\ \ldots\ VE_k^e \Big] \end{aligned} \tag{4.16}$$

La máscara $\boldsymbol{\omega}^{(1)}$ utilizada es la siguiente:

$$[1\ 1\ \ldots\ 1 \vdots\vdots 0\ 0\ \ldots\ 0 \vdots\vdots \ldots \vdots\vdots 0\ 0\ \ldots\ 0 \vdots\vdots 0\ 0\ \ldots\ 0 \vdots 0\ 0\ \ldots\ 0 \vdots \ldots \vdots 0\ 0\ \ldots\ 0]$$

El resultado obtenido es un modelo parcial del entorno ya que sólo representa a las variables ambientales pertenecientes al primer subconjunto definido.

Modelo común para el segundo subconjunto de variables ambientales y sucesivos. El segundo subconjunto de variables ambientales $VA^f,\ f = 2$ se modela en esta segunda fase del envSOM generalizado. El número de variables que forman el subconjunto puede variar entre 1 y l_2. El vector de datos de entrada $\mathbf{x}$ para el entrenamiento viene dado también por la ecuación 4.16. En cambio, en este caso la máscara $\boldsymbol{\omega}^{(2)}$ es la siguiente:

$$[0\ 0\ \ldots\ 0 \vdots\vdots 1\ 1\ \ldots\ 1 \vdots\vdots \ldots \vdots\vdots 0\ 0\ \ldots\ 0 \vdots\vdots 0\ 0\ \ldots\ 0 \vdots 0\ 0\ \ldots\ 0 \vdots \ldots \vdots 0\ 0\ \ldots\ 0]$$

Con el fin de preservar el modelo común para el primer subconjunto de variables ambientales obtenido en la fase anterior, se utiliza la máscara $\boldsymbol{\Omega}^{(2)}$ siguiente:

$$[0\ 0\ \ldots\ 0 \vdots\vdots 1\ 1\ \ldots\ 1 \vdots\vdots \ldots \vdots\vdots 1\ 1\ \ldots\ 1 \vdots\vdots 1\ 1\ \ldots\ 1 \vdots 1\ 1\ \ldots\ 1 \vdots \ldots \vdots 1\ 1\ \ldots\ 1]$$

El resultado obtenido es un modelo parcial del entorno ya que sólo representa a las variables ambientales pertenecientes al segundo subconjunto definido. La jerarquía establecida hace

que la organización de las variables ambientales pertenecientes a este subconjunto sea condicionada por las variables ambientales del primer subconjunto. De forma análoga a esta segunda fase de modelado, se pueden generar tantos modelos parciales del entorno como subconjuntos de variables ambientales se hayan definido.

Modelos eléctricos de los edificios. En la enésima fase del envSOM generalizado, se modelan los variables eléctricas para todos los edificios VE^e, con respecto a un entorno común definido por varios subconjuntos de variables ambientales. El número de variables eléctricas puede oscilar entre 1 y k. La ecuación 4.17 expresa el vector de datos de entrada $\mathbf{x}$ para el edificio e.

$$\mathbf{x}^e = \Big[VA_1^1 \, VA_2^1 \, \ldots \, VA_{l_1}^1 \vdots\vdots VA_1^2 \, VA_2^2 \, \ldots \, VA_{l_2}^2 \vdots\vdots \ldots \vdots\vdots VA_1^f \, VA_2^f \, \ldots \, VA_{l_f}^f \vdots\vdots VE_1^e \, VE_2^e \, \ldots \, VE_k^e \Big] \tag{4.17}$$

La máscara utilizada en la adaptación $\mathbf{\Omega}^{(n)}$ es la siguiente:

$$[0\,0\,\ldots\,0 \vdots\vdots 0\,0\,\ldots\,0 \vdots\vdots \ldots \vdots\vdots 0\,0\,\ldots\,0 \vdots\vdots 1\,1\,\ldots\,1]$$

Con esto se consiguen los modelos eléctricos individuales para cada edificio, dado un modelo ambiental común que ha sido construido jerárquicamente por niveles. Estos modelos contienen información del comportamiento eléctrico de los edificios, que son influenciados por un entorno ambiental común. La organización de las variables eléctricas que caracterizan a los edificios, es condicionada por las variables ambientales. Dado que las variables ambientales son comunes para todos los edificios, es factible pensar en herramientas de visualización similares para todos ellos.

4.6. Herramientas de visualización

4.6.1. Gráficos de comparación

Los planos de componentes procedentes del envSOM que corresponden a las variables eléctricas se pueden utilizar como herramientas para comparar de forma visual el comportamiento eléctrico de los edificios. El número total de planos de componentes eléctricas es $e \times k$, es decir, el número de edificios e multiplicado por el número de variables eléctricas k. Normalmente, este tipo de comparación requiere un gran esfuerzo visual y tiempo, dado el elevado número de mapas a examinar. En cambio, la proyección de Sammon proporciona un conjunto de puntos que se corresponden con los edificios e, utilizando un espacio de salida 2D (ver figura 4.15). La distancia relativa entre los puntos indica la diferencia que existe en el comportamiento eléctrico de los edificios. Es decir que, si dos puntos se encuentran muy próximos (distancia pequeña), entonces el comportamiento eléctrico de los edificios que representan será similar. Por el contrario, puntos alejados (distancia grande) indican que los edificios presentan diferencias significativas en sus perfiles eléctricos. La visualización de los puntos en un gráfico 2D permite realizar una comparación

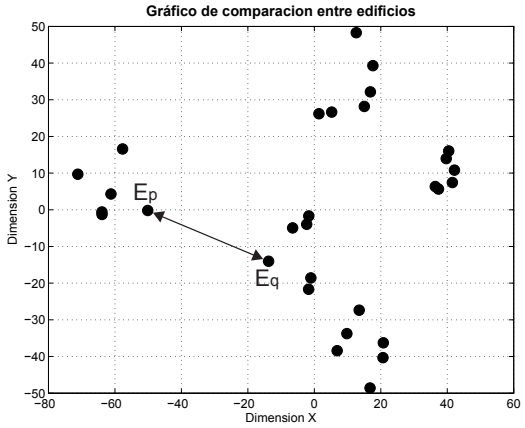


Figura 4.15: Ejemplo de un gráfico de comparación.

sencilla entre edificios en base a sus variables eléctricas, por lo que será el tipo de comparación preferida. En esta tesis, se definen tres tipos de gráficos de comparación en función de la matriz de similitud empleada para la proyección de Sammon:

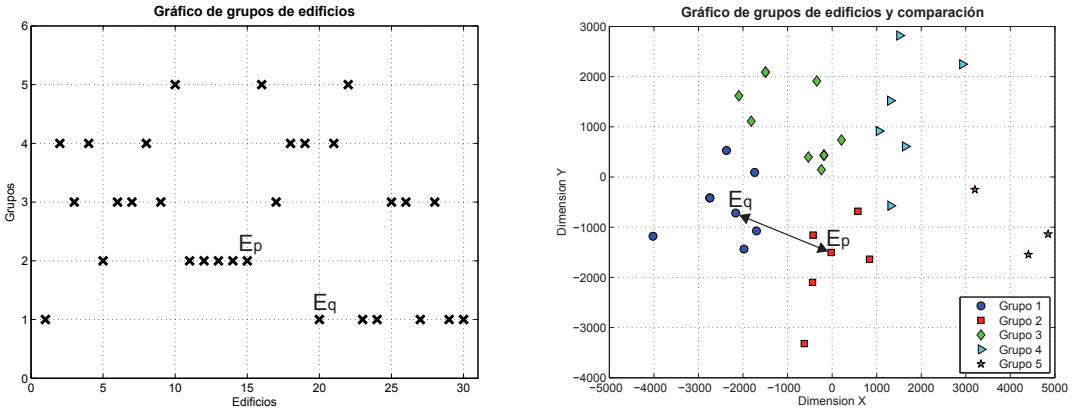
- **Gráfico de comparación global**: un solo gráfico 2D permite visualizar y comparar de forma global el comportamiento eléctrico de todos los edificios. Para ello, la matriz de similitud utilizada en la proyección de Sammon se calcula mediante la ecuación 4.12. Se puede considerar que este tipo de visualización y comparación corresponde a un compendio de todas las variables eléctricas.

- **Gráfico de comparación para cada variable**: este gráfico 2D permite visualizar y comparar el comportamiento eléctrico de todos los edificios para una determinada variable eléctrica. En este caso, la matriz de similitud utilizada en la proyección de Sammon se obtiene por medio de la ecuación 4.13. Cabe destacar que existen tantos gráficos de comparación de este tipo como variables eléctricas empleadas.

- **Gráfico de comparación condicionada**: en este gráfico 2D se visualiza y compara, bien de forma global o individualmente por cada variable, el comportamiento eléctrico condicionado por las variables comunes de tipo ambiental para todos los edificios. La matriz de similitud utilizada en la proyección de Sammon se calcula mediante la ecuación 4.14. El número de gráficos de comparación depende del número de condiciones y de los límites impuestos. Esto puede ser útil para detectar relaciones entre variables eléctricas y ambientales en los edificios.

4.6.2. Visualización de los grupos de edificios

El método *k-means* clasifica los edificios en diferentes grupos, dependiendo de sus perfiles eléctricos. El número total de grupos indica el número de perfiles eléctricos diferentes que existen entre todos los edificios. Los edificios que poseen un comportamiento eléctrico similar



(a) Grupos

(b) Grupos y comparación

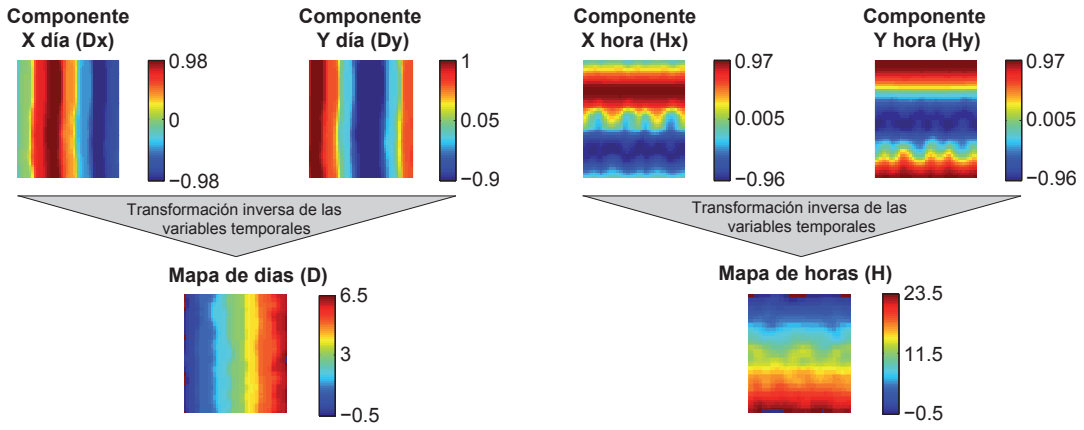
Figura 4.16: Visualización de los grupos de edificios.

pertenecen al mismo grupo, mientras que aquellos con perfiles dispares se aglomeran en grupos diferentes. Para visualizar los grupos de edificios, se podría utilizar un gráfico en el que se representan en el eje de abscisas los edificios y en el eje de ordenadas el grupo al que pertenecen (ver figura 4.16a). En este caso, no interesa representar los centroides de los grupos ni los datos de entrada, ya que no aportan información relevante.

Aprovechando los gráficos de comparación expuestos anteriormente, se propone incluir en ellos la información de los grupos. Por ejemplo, cada punto se puede representar con un color o símbolo diferente, dependiendo del grupo al que pertenezca el edifico correspondiente (ver figura 4.16b). La unificación de la visualización de las distancias relativas y de los grupos en un solo gráfico permite concentrar el conocimiento y facilita el análisis conjunto de ambos resultados, por lo que será preferida. Teniendo en cuenta el tipo de agrupamiento realizado, podemos visualizar grupos de edificios similares de forma global, individualmente para cada variable eléctrica y de forma condicionada por las variables ambientales.

4.6.3. Mapas de variables temporales

La transformación de las variables temporales (año, mes, día, hora, minutos, segundos, etc.) en las coordenadas X e Y correspondientes a la manecilla horaria de un reloj es útil en la exploración y modelado de los datos basado en el envSOM. Esta descomposición en variables sinusoidales evita discontinuidades en la organización de los mapas. Sin embargo, resulta difícil visualizar la variable temporal original una vez que el mapa ha sido entrenado, ya que es necesario visualizar simultáneamente y establecer correspondencia entre los dos planos de componentes correspondientes a las coordenadas X e Y. Además, la relación entre las dos coordenadas X e Y y la variable temporal original es sinusoidal lo que complica aún más la visualización. Para solucionar este problema, se propone un nuevo mapa o plano de visualización que se basa en deshacer la transformación inicial aplicada a las variables temporales. Esta transformación inversa se aplica una vez que ha finalizado el entrenamiento



(a) Mapa de días

(b) Mapa de horas

Figura 4.17: Ejemplos de los mapas de variables temporales.

del envSOM. Por ejemplo, la ecuación 4.18 permite deshacer la transformación inicial para la variable hora dada por la expresión 4.7.

$$m_{i(H)} = \arctan\left(\frac{m_{i(Hx)}}{m_{i(Hy)}}\right) \tag{4.18}$$

$m_{i(H)}$ corresponde al peso de la neurona i y la componente temporal hora de día H, $m_{i(Hx)}$ denota el peso de la neurona i y la componente X (función seno) de la manecilla del reloj de horas y $m_{i(Hy)}$ es el peso de la neurona i y la componente Y (función coseno). El nuevo mapa de visualización para la variable hora, o simplemente mapa de horas, indica la hora de 0 a 23 codificada mediante colores, lo que facilita su visualización. De forma análoga, la transformación inversa dada por la ecuación 4.19 deshace la descomposición inicial dada por la expresión 4.8 y permite obtener un mapa de visualización de la variable temporal día de la semana, denominado mapa de días. Los días entre 0 (Domingo) y 6 (Sábado) se codifican mediante colores.

$$m_{i(D)} = \arctan\left(\frac{m_{i(Dx)}}{m_{i(Dy)}}\right) \tag{4.19}$$

$m_{i(D)}$ corresponde al peso de la neurona i y la componente temporal día de la semana D, $m_{i(Dx)}$ denota el peso de la neurona i y la componente X (función seno) de la manecilla del reloj de días y $m_{i(Dy)}$ es el peso de la neurona i y la componente Y (función coseno). En la figura 4.17 se puede observar dos ejemplos de la generación de los mapas de variables temporales a partir de las componentes X e Y correspondientes. Dependiendo del tipo de variable temporal utilizada en la exploración o modelado, se podrían definir otros mapas de visualización similares, tales como el mapa de meses, el mapa de años, etc.

Conviene destacar que el cálculo de la matriz de distancias considerando solamente las coordenadas X e Y de una variable temporal, permite también visualizar adecuadamente esta variable. No obstante, se requiere etiquetar previamente el valor de la variable temporal en las zonas correspondientes del mapa.

Debido a que el tiempo es una variable común a todos los edificios, los mapas de variables temporales permiten realizar una visualización idéntica para todos ellos. Además, estos

mapas de visualización se pueden usar en la supervisión en línea de la energía eléctrica en los edificios junto con los planos de componentes eléctricas. Para ello, se puede representar la neurona ganadora en los mapas o incluso la trayectoria que describe la neurona ganadora a lo largo del tiempo (Kasslin *et al.*, 1992).

Se puede establecer una analogía entre los mapas de variables temporales y los mapas de estados. Un proceso industrial posee una serie de estados que son recorridos durante su operación normal. Cuando surge un fallo o funcionamiento anómalo en el proceso, se produce una transición entre estados indebida o poco probable que se puede visualizar en el mapa de estados o en la matriz de distancias (Díaz, 2000; Fuertes *et al.*, 2005). Igualmente, los edificios adoptan diferentes estados eléctricos normales a lo largo del tiempo. Si se produce un comportamiento eléctrico extraño o anormal con respecto al instante de tiempo actual, se podría visualizar en uno o varios mapas de variables temporales.

4.6.4. Visualización de las desviaciones

La visualización de desviaciones habilita la detección y cuantificación de situaciones anormales y tiene como objetivo final averiguar la causa que provoca ese desvío. La generación de modelos eléctricos para los edificios hace posible prever su comportamiento eléctrico en el tiempo. Cualquier variación del comportamiento actual con respecto al esperado indica la existencia de una desviación o residuo. No obstante, pueden aparecer falsas desviaciones de pequeña magnitud, cuando el modelo es burdo o se utilizan pocos datos pasados en su construcción. Lo ideal sería que los datos contengan la mayoría de estados o situaciones eléctricas para que estas sean capturadas por el modelo.

Para la visualización de desviaciones eléctricas se puede utilizar una gráfica donde se codifica mediante colores la magnitud del residuo vectorial en el tiempo. El residuo vectorial se define como la diferencia entre todas las componentes de los datos de entrada y los valores predichos por el modelo que mejor se aproximan al dato de entrada (Díaz y Hollmén, 2002; Domínguez *et al.*, 2007). Esta visualización individual permite detectar que componente o componentes eléctricas provocan la desviación. Pero, si el número de variables eléctricas es elevado, esta visualización puede llegar a ser confusa. Por otra parte, en grandes edificios existe una cierta relación entre las variables eléctricas por lo que, generalmente, todas se verán afectadas en mayor o menor medida ante una situación anormal.

Por tanto, es lícito pensar en la utilización de una herramienta que permita visualizar y detectar desviaciones eléctricas de forma simultánea en todas las variables. Para ello, se propone el uso de los mapas de variables temporales descritos anteriormente. Esta propuesta se basa en establecer una correspondencia directa entre una desviación eléctrica y una desviación temporal. Esta correspondencia resulta de la organización de las variables eléctricas condicionada de forma jerárquica por las temporales, que establece el algoritmo envSOM generalizado a n fases. Representando la neurona ganadora calculada utilizando todas las componentes del vector de entrada, o bien su trayectoria sobre estos mapas es posible conocer el valor esperado de las variables temporales. La trayectoria debe recorrer todos los estados temporales ordenadamente, excepto cuando surgen situaciones anormales.

Las desviaciones se pueden detectar comprobando que los valores obtenidos del modelo no coinciden con el valor real de las variables temporales, que es siempre conocido. Es decir que existe una desviación cuando los valores actuales de las variables eléctricas se corresponden a otro instante de tiempo distinto al actual. Cuanto mayor es la desviación temporal, generalmente mayor será la desviación eléctrica. Cuando se dispone de varias variables temporales en el modelo, las desviaciones significativas en magnitud y/o número de variables implicadas se muestran en el mapa correspondiente a la variable temporal superior en la jerarquía, mientras que aquellas cuya magnitud es menor y/o afectan a pocas variables se indican en el mapa de la variable temporal inferior en la jerarquía. Por ejemplo, si coexisten los mapas de días y horas, los primeros representan las desviaciones destacadas y los segundos aquellas menos significativas.

Además de detectar y cuantificar las desviaciones, en muchas ocasiones es conveniente averiguar la dirección de las mismas. En este caso, la dirección viene dada por el valor de la variable tiempo que mejor representa a los valores de las variables eléctricas actuales. Una vez detectada la desviación mediante las variables temporales, se puede realizar una nueva búsqueda de la neurona ganadora utilizando conocimiento sobre el tiempo, que es siempre conocido. En la nueva búsqueda solamente se tienen en cuenta aquellas neuronas que poseen valores en sus componentes temporales distintos del actual. Con esto se logra obtener una nueva neurona ganadora que representa al comportamiento eléctrico actual e indica un instante de tiempo distinto del real, al que corresponde dicho comportamiento. Por ejemplo, una situación en la que la neurona ganadora indica que es Domingo, pero el día de la semana real es Martes, es síntoma de una desviación eléctrica. También se puede utilizar el conocimiento apriorístico del tiempo real para evitar pequeñas o falsas desviaciones. Dado que el número de neuronas que intervienen en el cálculo de la neurona ganadora se ve reducido considerablemente al incorporar conocimiento, el tiempo que emplea este nuevo proceso de búsqueda es mínimo y se puede implementar sin inconvenientes en la supervisión en línea.

Capítulo 5

Experimentación y resultados

En este capítulo se explican con detalle los experimentos realizados y se presentan los resultados obtenidos. El objetivo es validar la metodología propuesta en esta tesis y demostrar la utilidad de las técnicas de minería de datos visual en la supervisión de la energía eléctrica en todos los edificios de la Universidad de León de forma conjunta. Para explotar los datos eléctricos procedentes de los edificios, adquiridos y almacenados por el sistema de medida y supervisión desarrollado, se proponen dos experimentos. En el primero se realiza un análisis exploratorio de los datos eléctricos y una comparación entre edificios, mientras que en el segundo se lleva a cabo una supervisión en línea basada en modelos del consumo de energía eléctrica en los edificios. Además, se presentan las herramientas de visualización que tienen por objeto facilitar la supervisión y extraer conocimiento de los edificios.

5.1. Definición de los experimentos

La metodología de explotación de datos propuesta en el capítulo 4 busca extraer conocimiento, que posteriormente pueda ser utilizado por las herramientas de supervisión avanzadas de la energía eléctrica. Cualquier usuario debería ser capaz de analizar y razonar de forma visual en base a esa información. Con objeto de validar la metodología de explotación de datos y las herramientas de supervisión avanzadas propuestas en esta tesis, se han llevado a cabo dos tipos de experimentos:

- **Experimento de exploración**: el objetivo del primer experimento es comprobar la metodología de exploración propuesta, basada en la combinación del algoritmo envSOM, la proyección de Sammon y el método *k-means*. La exploración de datos permite realizar un análisis comparativo de los edificios en base a su comportamiento eléctrico, el cual está condicionado por las variables ambientales. Asimismo, este experimento trata de comprobar la utilidad de nuevas herramientas de visualización, las cuales son utilizadas por las aplicaciones web desarrolladas para la supervisión avanzada *off-line*.
- **Experimento de modelado**: el objetivo del segundo experimento es validar la metodología de modelado propuesta, la cual se basa en un algoritmo envSOM de

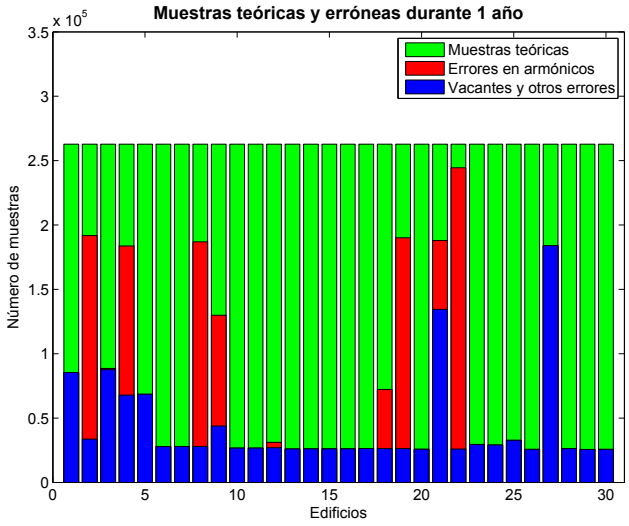


Figura 5.1: Número de muestras teóricas y erróneas tratadas. Los números de los edificios se indican en la tabla 3.1.

tres fases. La obtención de un modelo de comportamiento eléctrico para cada edificio es la base para una supervisión avanzada *on-line*. Este experimento también permite comprobar la utilidad de nuevas herramientas de visualización, las cuales son utilizadas por las aplicaciones web desarrolladas para la supervisión avanzada.

En la ejecución de ambos experimentos se ha empleado el software Matlab, junto con la librería SOM *toolbox* (Vesanto *et al.*, 2000) desarrollada en el *Department of Information and Computer Science, Aalto University*. Las funciones de esta librería se han modificado convenientemente para implementar el algoritmo envSOM.

5.1.1. Tareas previas a los experimentos

La salida del preprocesamiento es la base para la realización de los experimentos. Ambos experimentos utilizan datos que previamente han sido procesados adecuadamente con el fin de evitar que muestras erróneas, vacantes o fuera de rango, variables insignificantes o con información redundante y valores de componentes con diferentes rangos influyan negativamente en los resultados de la minería de datos. El servicio de preprocesado toma datos crudos almacenados en la base de datos, realiza su correspondiente procesamiento y guarda los datos preprocesados de nuevo. La tarea de preprocesamiento se lleva a cabo cada mes para todos los edificios.

Tratamiento de muestras erróneas. En este primer paso, las muestras que no aportan información válida se codifican mediante valores NaN. En la figura 5.1 se resume la salida del tratamiento de muestras erróneas y vacantes a lo largo de un año (desde Marzo del 2010 hasta Febrero del 2011) para todos los edificios. El número de muestras correctas teóricas, teniendo en cuenta que el periodo de muestreo es de 2 minutos, debería ser 262800

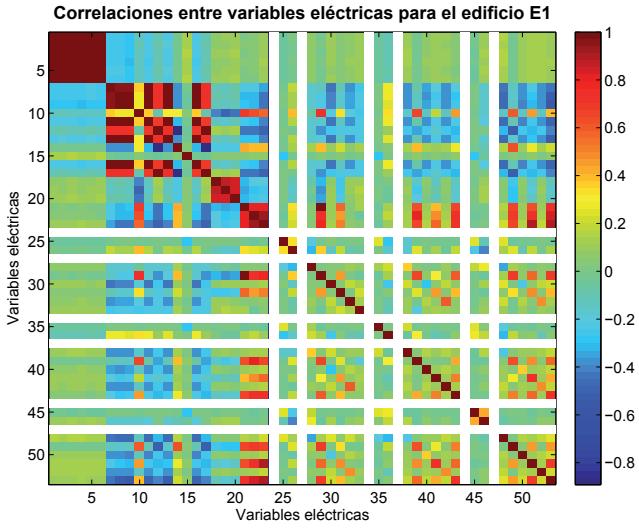


Figura 5.2: Correlaciones entre las variables eléctricas para el edificio E1. Los números de las variables se indican en la tabla 3.2.

por cada edificio (barras en verde). Debido a diferentes errores, ya sea en los equipos de medida, la red de comunicación o en el sistema de adquisición y almacenamiento, el número de muestras correctas es inferior. Mención especial requieren los errores existentes en las variables armónicos y THDs tanto en tensión como en corriente (barras en rojo) procedentes de los medidores instalados en los edificios E2, E4, E8, E9, E18, E19, E21 y E22. Este tipo de errores pueden ser debidos a un fallo, descalibración o error en el *firmware* del medidor. Actualmente, se está estudiando el origen de estos errores para lo cual se ha introducido un nuevo medidor en el edificio E4 con objeto de verificar la medida de estas variables. La mayor parte del resto de errores (barras en azul) se producen por desconexiones puntuales de los medidores, o bien por problemas técnicos en el sistema de medida y supervisión. El edificio E27 posee un gran número de muestras vacantes ya que su medidor se desconectó del sistema para acometer reformas estructurales en su interior. En el edificio E21, una avería en el medidor provocó la ruptura del fusible de protección en una de las fases, con lo que la medida de tensiones y en consecuencia del resto de variables, son erróneas. En los edificios E1, E3, E4 y E5 se llevó a cabo una modificación en la estructura de medida eléctrica, por lo que el número de muestras vacantes es ligeramente superior al resto de los edificios. Este primer paso permite localizar y tratar los datos erróneos para evitar que afecten de forma negativa a los resultados de los experimentos.

Selección de variables. Este segundo paso implica una primera reducción de la dimensionalidad en el proceso de minería de datos. Por tanto, la selección de variables es de vital importancia para la ejecución de los experimentos. No se debe introducir variables redundantes, que no aportan información relevante y aumentan el tiempo de ejecución de los experimentos, pero por el contrario tampoco se debe prescindir de variables significativas que poseen información característica de los edificios. Por esta razón, se ha llevado a cabo

Etiqueta	Variable	Definición	Unidad
$\overline{V}$	Tensión promedio de las 3 fases	$\overline{V} = \frac{Va+Vb+Vc}{3}$	Voltios (V)
P	Potencia activa en las 3 fases	P	Kilovatios (KW)
FP	Factor de potencia en las 3 fases	FP	-
DC	Desequilibrio entre las corrientes de fase	$DC = \left[\frac{max(\lvert Ia-Ib\rvert,\lvert Ib-Ic\rvert,\lvert Ic-Ia\rvert)}{I_m}\right] * 100$	Porcentaje (%)
$\overline{THDv}$	Distorsión armónica total promedio en la tensión	$\overline{THDv} = \frac{THD\ Va+THD\ Vb+THD\ Vc}{3}$	Porcentaje (%)
$\overline{THDi}$	Distorsión armónica total promedio en la corriente	$\overline{THDi} = \frac{THD\ Ia+THD\ Ib+THD\ Ic}{3}$	Porcentaje (%)
ES	Energía activa en las 3 fases por superficie del edificio	$ES = \frac{EA}{S}$	Vatios hora por metro cuadrado (Wh/m^2)

Tabla 5.1: Variables eléctricas seleccionadas para la minería de datos.

un análisis de correlaciones lineales entre todas las variables eléctricas, que permite detectar relaciones entre ellas y seleccionar aquellas más representativas y comunes en todos los edificios. Además, se ha tenido en cuenta el conocimiento técnico que se posee acerca de los edificios y sus instalaciones eléctricas.

En la figura 5.2 se muestran de forma gráfica las correlaciones entre variables eléctricas para el edificio E7. Un valor 1 (rojo) indica que la correlación existente es fuerte y directa y un valor -1 (azul) corresponde a una correlación fuerte e inversa. Como se puede observar las tensiones de fase (Va, Vb, Vc) están fuertemente relacionadas entre sí, de igual forma que las tensiones compuestas (Vab, Vbc, Vca). A su vez, tensiones de fase y compuestas están correlacionadas. En cuanto a las corrientes de fase (Ia, Ib, Ic) y neutro (In) están relacionadas entre sí, de forma análoga que las potencias (P, Q, S) y las energías (EA, ER). Asimismo, corrientes, potencias y energías están correlacionadas en mayor o menor medida. Por otra parte, tanto las distorsiones armónicas totales en tensión (THD Va, THD Vb, THD Vc) como en corriente (THD Ia, THD Ib, THD Ic) poseen una gran correlación entre sí. Los armónicos impares en corriente (H3 Ia, H5 Ia, H7 Ia, H3 Ib, H5 Ib, H7 Ib, H3 Ic, H5 Ic, H7 Ic) están íntimamente relacionados entre sí, y a su vez con la distorsión armónica total correspondiente (THD Ia, THD Ib, THD Ic). Con los armónicos impares en tensión (H3 Va, H3 Vb, H3 Vc) sucede algo idéntico. Cabe destacar que los armónicos fundamentales en tensión y corriente (H1 Va, H1 Vb, H1 Vc, H1 Ia, H1 Ib, H1 Ic) poseen siempre el mismo valor (100 %) y por lo tanto las correlaciones resultantes corresponden a valores NaN que no se pueden interpretar (color blanco).

Las variables eléctricas elegidas para realizar la explotación de datos se indican en la tabla 5.1. En este tipo de edificios, las variables eléctricas más importantes son la potencia (P) y el factor de potencia (FP), por lo que se descartaron las variables de energía al estar

Etiqueta	Variable	Definición	Unidad
Hx	Hora del día (coordenada X)	$Hx = \sin(\alpha) = \sin\left(\frac{2\pi H}{24}\right)$	-
Hy	Hora del día (coordenada Y)	$Hy = \cos(\alpha) = \cos\left(\frac{2\pi H}{24}\right)$	-
Dx	Día de la semana (coordenada X)	$Dx = \sin(\beta) = \sin\left(\frac{2\pi D}{7}\right)$	-
Dy	Día de la semana (coordenada Y)	$Dy = \cos(\beta) = \cos\left(\frac{2\pi D}{7}\right)$	-
TE	Temperatura exterior	TE	Grados Celsius (ºC)
HR	Humedad relativa	HR	Porcentaje (%)
RS	Radiación solar	RS	Vatios por metro cuadrado (W/m^2)
FL	Tipo de día (Festivo/Laborable)	FL	-

Tabla 5.2: Variables ambientales seleccionadas para la minería de datos.

fuertemente relacionadas con la potencia. Los valores tensión promedio, THDs promedio en tensión y en corriente ($\overline{V}$, $\overline{THDv}$ y $\overline{THDi}$) de las tres fases aportan una información básica, evitando utilizar las variables de cada fase que poseen una fuerte correlación entre sí. Las corrientes de fase están correlacionas entre sí, pero el cálculo del desequilibrio en corriente (DC) a partir de ellas caracteriza el balance de cargas en la distribución de receptores monofásicos en el edificio lo que influye en la corriente del neutro o incluso en el tipo de armónicos existentes. Además, el cálculo de la relación energía por superficie (ES) del edificio se utiliza habitualmente para redistribuir costes de la energía eléctrica consumida entre los diferentes edificios del mismo punto de facturación o instituciones dentro de un mismo edificio. Un análisis espectral específico en armónicos sería conveniente, aunque se sale de los objetivos planteados en esta tesis relativos a la supervisión general de todos los edificios.

Respecto a las variables ambientales que caracterizan el entorno común que rodea a los edificios, se eligieron variables de tipo meteorológico y temporal. Cualquier variable relacionada con el tiempo se descompone en dos coordenadas X e Y de tipo sinusoidal correspondientes a la manecilla de un reloj para evitar discontinuidades entre los valores máximo y mínimo de la variable temporal. Por ejemplo, la variable hora del día (H) se representa por dos coordenadas (Hx, Hy). De forma análoga, la variable día de la semana (D) se sustituye por dos coordenadas (Dx, Dy). Por otra parte, la variable tipo de día (FL) permite distinguir entre días festivos y laborables, para los que se puede suponer un comportamiento eléctrico muy diferente. Además, se asume que las variables meteorológicas que más influyen en el consumo eléctrico son la temperatura exterior (TE), la humedad relativa (HR) y la radiación solar (RS), ya que están directamente relacionadas con el funcionamiento de los sistemas HVAC de los edificios. En la tabla 5.2, se indican las variables

ambientales seleccionadas para la explotación de datos.

Normalización de los datos. En este último paso previo a la ejecución de los experimentos, se realiza una transformación de todas las variables (eléctricas y ambientales), de forma que el valor mínimo de cada variable pasa a ser 0 mientras que el valor máximo se convierte en 1. Como salida de la normalización, se obtienen datos en rango [0, 1] preparados para el entrenamiento de los algoritmos de minería de datos.

5.1.2. Experimento de exploración

Este primer experimento demuestra la utilidad de la metodología de exploración en la supervisión de la energía eléctrica en los edificios del Campus de la Universidad de León. Esta metodología está basada en el algoritmo envSOM, la proyección de Sammon y el método *k-means*. En este experimento se utilizan datos registrados por el sistema de medida y supervisión para explorar, visualizar, analizar y comparar las variables eléctricas de todos los edificios mediante planos de componentes y gráficos de comparación.

El conjunto de datos empleado en este experimento corresponde a muestras capturadas durante 1 año (desde Marzo de 2010 hasta Febrero de 2011) con un periodo de muestreo de 2 minutos. Esto implica que se dispone de 262800 muestras por cada edificio. Las variables implicadas en el experimento son 13 y se detallan en la tabla 5.3. Se puede distinguir entre variables ambientales (coordenada X de la hora, coordenada Y de la hora, temperatura, humedad, radiación solar y tipo de día) y variables eléctricas (tensión promedio, potencia, factor de potencia, desequilibrio en corriente, THD promedio en tensión, THD promedio en corriente y relación energía por superficie). Teniendo en cuenta el número de edificios (30), variables (13) y muestras (262800), el volumen de datos de entrada es superior a 56 millones.

Con estos datos de entrada se entrena en modo *batch* un envSOM de 2 fases. La primera fase está constituida por un SOM común para los 30 edificios, por lo que la dimensión del espacio de entrada es 216 (6 variables ambientales comunes para todos los edificios y 210 eléctricas). El vector de datos de entrada $\mathbf{x}$ se define mediante la expresión 5.1.

$$\begin{aligned}\mathbf{x} = \Big[& Hx\ Hy\ TE\ HR\ RS\ FL \,\vdots\vdots \\ & \overline{V}^1\ P^1\ FP^1\ DC^1\ \overline{THD}v^1\ \overline{THD}i^1\ ES^1 \,\vdots \\ & \overline{V}^2\ P^2\ FP^2\ DC^2\ \overline{THD}v^2\ \overline{THD}i^2\ ES^2 \,\vdots \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \,\vdots \\ & \overline{V}^e\ P^e\ FP^e\ DC^e\ \overline{THD}v^e\ \overline{THD}i^e\ ES^e \Big]\end{aligned} \tag{5.1}$$

Por tanto, la máscara aplicada en la búsqueda de la neurona ganadora $\boldsymbol{\omega}$ es:

$$\Big[1\,1\,1\,1\,1\,1 \vdots\vdots\, 0\,0\,0\,0\,0\,0\,0 \vdots\, 0\,0\,0\,0\,0\,0\,0 \vdots\ \ldots\ \vdots\, 0\,0\,0\,0\,0\,0\,0 \Big]$$

	Etiqueta	Variables	Unidad
Ambientales	Hx	Hora del día (coordenada X)	-
	Hy	Hora del día (coordenada Y)	-
	TE	Temperatura exterior	Grados Celsius (ºC)
	HR	Humedad relativa	Porcentaje (%)
	RS	Radiación solar	Vatios por metro cuadrado (W/m^2)
	FL	Tipo de día (Festivo/Laborable)	-
Eléctricas	$\overline{V}$	Tensión promedio de las 3 fases	Voltios (V)
	P	Potencia activa en las 3 fases	Kilovatios (KW)
	FP	Factor de potencia en las 3 fases	-
	DC	Desequilibrio entre las corrientes de fase	Porcentaje (%)
	$\overline{THD}v$	Distorsión armónica total promedio en la tensión	Porcentaje (%)
	$\overline{THD}i$	Distorsión armónica total promedio en la corriente	Porcentaje (%)
	ES	Energía activa en las 3 fases por superficie útil del edificio	Vatios hora por metro cuadrado (Wh/m^2)

Tabla 5.3: Variables empleadas en la exploración de los datos eléctricos.

El resto de parámetros del SOM se han elegido utilizado reglas heurísticas. Las dimensiones del espacio de salida son 70×50 (3500 neuronas o unidades). Es decir, una neurona por cada 75 muestras de entrada aproximadamente. Teniendo en cuenta que existen datos erróneos que no intervienen en la organización del SOM, realmente el número de muestras agrupadas por cada neurona será ligeramente inferior a 75 (entre 50 y 60). La tasa de aprendizaje $\alpha(t)$ decrece en el tiempo y la función de vecindad $h_{ci}(t)$ es de tipo Gaussiano. La topología utilizada es de tipo hoja y la malla es rectangular. La inicialización es lineal a lo largo de los mayores autovalores, mientras que el número de épocas de entrenamiento es 100.

En la segunda fase del envSOM se aplica un SOM individual para cada edificio. La dimensión del espacio de entrada para cada SOM es 13 (6 variables ambientales y 7 eléctricas). En este caso, el vector de datos de entrada $\mathbf{x}$ usado en el entrenamiento para cada edificio e está formado según la expresión 5.2.

$$\mathbf{x}^e = \left[Hx\ Hy\ TE\ HR\ RS\ FL \vdots \overline{V}^e\ P^e\ FP^e\ DC^e\ \overline{THD}v^e\ \overline{THD}i^e\ ES^e \right] \tag{5.2}$$

La máscara aplicada durante el proceso de adaptación $\mathbf{\Omega}$ es la misma para todos los edificios:

$$\left[0\,0\,0\,0\,0\,0 \vdots 1\,1\,1\,1\,1\,1\,1 \right]$$

Se toman los vectores prototipo resultado de la primera fase como inicialización de los 30 SOMs, convenientemente extraídos del primer SOM. El resto de parámetros son idénticos al SOM de la primera fase. Las dimensiones del espacio de salida son 70×50, la tasa de aprendizaje $\alpha(t)$ decrece en el tiempo, la función de vecindad $h_{ci}(t)$ es de tipo Gaussiano, la topología es de tipo hoja, la malla es rectangular y el número de épocas de entrenamiento es 100.

La duración del entrenamiento del envSOM con las características indicadas es de 35 horas aproximadamente en el servidor de explotación (equipo con procesador de 3 núcleos AMD Phenom(tm) 8600B, 2.30 GHz y 5 GB de RAM). Teniendo en cuenta el volumen de datos de entrada (mayor de 56 millones), este tiempo no es excesivo. Además, no es necesario realizar el entrenamiento cada vez que un usuario accede al sistema de supervisión y se repite con baja frecuencia (1 vez al año).

A partir de los vectores prototipo obtenidos como resultado del envSOM, se calculan las diferentes matrices de similitud entre edificios (global, individuales por cada variable eléctrica y condicionadas por las variables ambientales) utilizando la distancia L_1 o *cityblock*, cuyas dimensiones son 30×30. Estas matrices contienen información acerca de las similitudes y diferencias existentes entre edificios. La proyección de Sammon se utiliza para representar estas similitudes o diferencias en un espacio de visualización 2D fácilmente interpretable. Cada edificio se visualiza como un punto en el gráfico. Cabe mencionar que la proyección de Sammon requiere que todos los elementos de las matrices de similitud sean estrictamente mayores que 0 y no se puede aplicar en caso contrario.

De forma adicional, el algoritmo *k-means* se utiliza para etiquetar cada punto en el gráfico, dependiendo del grupo al que pertenece el edificio correspondiente. El número óptimo de grupos de edificios se elige entre 2 y 10 siguiendo los criterios de mínimo error medio de los datos de entrada con respecto al centroide y mínimo índice de Davies & Bouldin (Davies y Bouldin, 1979). Con cada conjunto de datos de entrada se realiza una única repetición del algoritmo. En el proceso de minimización del error entre los datos de entrada y el centroide correspondiente se emplea la distancia L_1.

5.1.3. Experimento de modelado

El segundo experimento demuestra la utilidad de la metodología de modelado en la supervisión de la energía eléctrica en los edificios del Campus de la Universidad de León. Esta metodología está basada en el algoritmo envSOM generalizado de 3 fases, que permite obtener mapas de visualización de las variables eléctricas. Para considerar la periodicidad en el consumo eléctrico, las variables temporales que definen el entorno común se han dividido en dos subconjuntos que se corresponden con el día y la hora. Aplicando la transformación propuesta para variables temporales, se obtienen las correspondientes coordenadas X e Y para cada una de ellas. En este experimento se utilizan datos registrados por el sistema de medida y supervisión para visualizar y analizar las variables eléctricas de todos los edificios mediante planos de componentes y mapas de días y horas.

El conjunto de datos empleado en este experimento corresponde a muestras capturadas durante 1 mes (Marzo de 2010) con un periodo de muestreo de 2 minutos. Con el fin de tener muestras del mismo número de días para cada día de la semana, se han empleado datos de 4 semanas completas, prescindiendo de las muestras correspondientes a los últimos 3 días del mes de Marzo. Por tanto, el número de muestras utilizadas es 20160 para cada uno de los 30 edificios.

El conjunto de datos de entrada se ha dividido en dos subconjuntos, uno destinado al

	Etiqueta	Variables	Unidad
Ambientales	Dx	Día de la semana (coordenada X)	-
	Dy	Día de la semana (coordenada Y)	-
	Hx	Hora del día (coordenada X)	-
	Hy	Hora del día (coordenada Y)	-
Eléctricas	$\overline{V}$	Tensión promedio de las 3 fases	Voltios (V)
	P	Potencia activa en las 3 fases	Kilovatios (KW)
	PF	Factor de potencia en las 3 fases	-
	DC	Desequilibrio entre las corrientes de fase	Porcentaje (%)
	$\overline{THDv}$	Distorsión armónica total promedio en la tensión	Porcentaje (%)
	$\overline{THDi}$	Distorsión armónica total promedio en la corriente	Porcentaje (%)

Tabla 5.4: Variables empleadas en el modelado del comportamiento eléctrico de los edificios.

entrenamiento del envSOM y obtención de los modelos eléctricos y el otro para su evaluación. El subconjunto de entrenamiento consta de 15120 muestras correspondientes a tres semanas, mientras que el subconjunto de prueba posee 5040 muestras correspondientes a una semana y contiene un día festivo (Viernes), que será útil para comprobar la detección de desviaciones.

Las 10 variables implicadas en el experimento de modelado se detallan en la tabla 5.4. Se puede distinguir entre variables ambientales (coordenadas X e Y del día y coordenadas X e Y de la hora) y variables eléctricas (tensión promedio, potencia, factor de potencia, desequilibrio en corriente, THD promedio en tensión y THD promedio en corriente).

Teniendo en cuenta el número de edificios (30), variables (10) y muestras del subconjunto de entrenamiento (15120), el volumen de datos de entrada es superior a 2 millones. A este subconjunto de datos de entrada se aplica un envSOM de 3 fases. La primera fase está constituida por un SOM común para los 30 edificios, por lo que la dimensión del espacio de entrada es 184 (4 variables ambientales y 180 eléctricas). El vector de datos de entrada $\mathbf{x}$ viene dado por la expresión 5.3.

$$\begin{aligned}\mathbf{x} = \Big[& Dx\ Dy \vdots\vdots Hx\ Hy \vdots\vdots \\ & \overline{V}^1\ P^1\ FP^1\ DC^1\ \overline{THDv}^1\ \overline{THDi}^1 \vdots \\ & \overline{V}^2\ P^2\ FP^2\ DC^2\ \overline{THDv}^2\ \overline{THDi}^2 \vdots \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \vdots \\ & \overline{V}^e\ P^e\ FP^e\ DC^e\ \overline{THDv}^e\ \overline{THDi}^e \Big]\end{aligned} \tag{5.3}$$

La máscara aplicada en la búsqueda de la neurona ganadora $\boldsymbol{\omega}^{(1)}$ es:

$$\Big[1\ 1 \vdots\vdots 0\ 0 \vdots\vdots 0\ 0\ 0\ 0\ 0\ 0 \vdots 0\ 0\ 0\ 0\ 0\ 0 \vdots\ \ldots\ \vdots 0\ 0\ 0\ 0\ 0\ 0\Big]$$

A las variables ambientales se aplica un bajo nivel de ruido (1 %) para conseguir una pequeña variabilidad en los vectores prototipo de las correspondientes componentes y mejorar así la inicialización de la siguientes fases. El resto de parámetros del SOM se

han elegido mediante reglas heurísticas. Las dimensiones del espacio de salida son 30×20 (600 neuronas o unidades). En este caso, se tiene aproximadamente una neurona por cada 25 muestras de entrada para obtener modelos con precisiones aceptables. La tasa de aprendizaje $\alpha(t)$ decrece en el tiempo y la función de vecindad $h_{ci}(t)$ es de tipo Gaussiano. La topología usada es de tipo hoja y la malla es rectangular. La inicialización es lineal a lo largo de los mayores autovalores y el número de épocas de entrenamiento es 500.

La segunda fase del envSOM es muy similar a la primera, aplicándose un SOM común para todos los edificios. El vector de datos de entrada $\mathbf{x}$ viene dado por la expresión 5.3 y su dimensión es 184 igualmente. La máscara aplicada en la búsqueda de la neurona ganadora $\boldsymbol{\omega}^{(2)}$ es:

$$\left[0\,0 \vdots\vdots 1\,1 \vdots\vdots 0\,0\,0\,0\,0\,0 \vdots 0\,0\,0\,0\,0\,0 \vdots \;\ldots\; \vdots 0\,0\,0\,0\,0\,0\right]$$

La novedad está en la máscara utilizada en la adaptación $\boldsymbol{\Omega}^{(2)}$, la cual es:

$$\left[0\,0 \vdots\vdots 1\,1 \vdots\vdots 1\,1\,1\,1\,1\,1 \vdots 1\,1\,1\,1\,1\,1 \vdots \;\ldots\; \vdots 1\,1\,1\,1\,1\,1\right]$$

Esta máscara es necesaria para evitar que se modifique la variable día de la semana, la cual ya ha sido organizada adecuadamente en la fase anterior. El resto de parámetros son idénticos al SOM de la primera fase, tomando los vectores prototipo resultado de la primera fase como inicialización.

En la tercera fase del envSOM se aplica un SOM individual para cada edificio e. En este caso, el vector de datos de entrada $\mathbf{x}$ para el entrenamiento tiene una dimensión de 10 (4 variables ambientales y 6 eléctricas) y se define mediante la expresión 5.4.

$$\mathbf{x}^e = \left[Dx\ Dy \vdots\vdots Hx\ Hy \vdots\vdots \overline{V}^e\ P^e\ FP^e\ DC^e\ \overline{THD}v^e\ \overline{THD}i^e\right] \tag{5.4}$$

La máscara aplicada durante el proceso de adaptación $\boldsymbol{\Omega}^{(3)}$ es la misma para todos los edificios:

$$\left[0\,0 \vdots\vdots 0\,0 \vdots\vdots 1\,1\,1\,1\,1\,1\right]$$

Como inicialización de los 30 SOMs se usan los vectores prototipo resultado de la segunda fase. En número de épocas de entrenamiento se duplica (1000) para conseguir una perfecta organización, mientras que el resto de parámetros son idénticos a los SOM de fases anteriores.

La duración del entrenamiento del envSOM con las características indicadas fue de 5 horas aproximadamente en el servidor de explotación (equipo con procesador de 3 núcleos AMD Phenom(tm) 8600B, 2.30 GHz y 5 GB de RAM). Cabe destacar que no es necesario realizar el entrenamiento cada vez que un usuario accede al sistema de supervisión y se procesan simultáneamente más de 2 millones de datos procedentes de todos los edificios con una frecuencia media (1 vez cada mes).

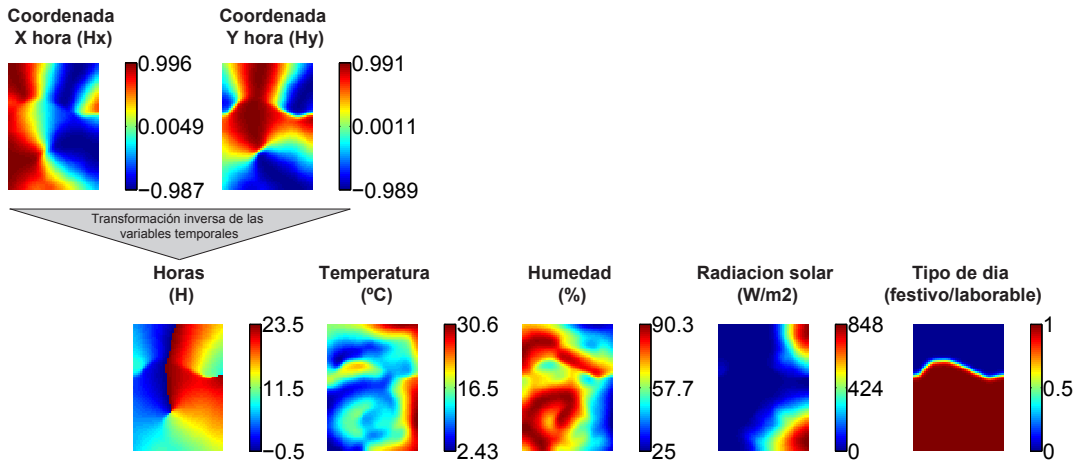


Figura 5.3: Planos de componentes de las variables que caracterizan el entorno ambiental común.

5.2. Resultados de la exploración

5.2.1. Supervisión y análisis basados en los planos de componentes

El primer resultado obtenido en el experimento de exploración de datos es un conjunto de planos de componentes para cada edificio, procedentes de la segunda fase del envSOM. La supervisión y análisis de la energía eléctrica en todos los edificios del Campus se puede realizar a partir de estos planos de componentes. Básicamente, este resultado de la exploración permite al usuario visualizar de forma general la distribución y rango de las variables eléctricas, las cuales son condicionadas por las variables ambientales para todos los edificios a lo largo de un año. La comparación visual entre los planos de componentes de variables eléctricas de distintos edificios es posible gracias a la organización establecida por el envSOM. Cabe recordar que neuronas situadas en la misma posición en los mapas agrupan datos de entrada con características ambientales idénticas. No obstante, esta comparación se puede convertir en una tarea costosa en tiempo y esfuerzo dado el elevado número de planos (30 edificios y 7 variables eléctricas).

Análisis de las componentes ambientales.

En la figura 5.3 se muestran los planos correspondientes a las componentes que caracterizan el entorno ambiental. Estos mapas definen un modelo meteorológico y temporal común para todos los edificios, que condiciona la organización de las componentes eléctricas. Tal y como se explicó anteriormente, este modelo viene dado por las coordenadas X e Y de la hora (Hx, Hy), la temperatura (TE), la humedad (HR), la radiación solar (RS) y el tipo de día (FL).

El modelo meteorológico y temporal presenta dos zonas bien diferenciadas por la componente tipo de día. Se puede ver como la zona correspondiente a los días laborables es

mayor ya que este tipo de días predomina sobre los días festivos. Gracias a esta variable se separan comportamientos eléctricos muy diferentes entre ambos tipos de días. La radiación solar oscila entre 0 W/m^2 durante la noche y 848 W/m^2 en días soleados, predominando valores en el mapa donde esta variable es baja. La temperatura y la humedad se encuentran correlacionadas de forma inversa, de modo que cuando una es alta, la otra es baja y viceversa. Los valores de temperatura oscilan entre 2 y 31 ^{o}C y los de humedad relativa entre 25 y 90 %. Son valores típicos de un clima continental a lo largo de un año y una localización geográfica como la de la ciudad de León.

En principio, parece difícil conocer la distribución de las horas mediante las coordenadas X e Y en el modelo ambiental. Sin embargo, aplicando la transformación inversa para variables temporales definida por la ecuación 4.18, se obtiene el plano o mapa de horas, donde si se puede interpretar fácilmente la distribución horaria. Como se puede intuir, en este plano existen dos zonas donde se repiten las mismas horas, en función del tipo de día (laborable o festivo).

Analizando detenidamente los planos de las componentes ambientales, se comprueba que en horas nocturnas (azul) la temperatura y la radiación solar son bajas y la humedad alta, independientemente del tipo del día. Durante las horas correspondientes a la tarde, la temperatura y la radiación son altas y la humedad mínima. El modelo del entorno no aporta ninguna novedad, aunque es conveniente comprobar que este modelo refleja fielmente las condiciones ambientales que influyen en el consumo eléctrico de los edificios.

Análisis de las componentes eléctricas.

La visualización de los planos de las componentes eléctricas permite supervisar y analizar la energía eléctrica en cada edificio. Las relaciones entre variables eléctricas y ambientales se pueden descubrir fácilmente observando simultáneamente los planos correspondientes. Además, es posible realizar una comparación visual específica entre dos edificios mediante los planos de la misma componente eléctrica.

En las figuras 5.4, 5.5 y 5.6 se presentan los planos de componentes correspondientes a las variables eléctricas de los 30 edificios. En una vista general de todos los planos de componentes, se puede decir que la mayor parte de los edificios tienen un comportamiento eléctrico muy similar. No obstante, existen algunas diferencias que se mencionarán a continuación para cada variable eléctrica.

Variable tensión promedio. La tensión promedio en las tres fases oscila entre 376 V en el edificio E18 y 418 V en los edificios E4 y E26. Generalmente, se consideran valores normales de tensión aquellos que se sitúan en el rango $400 \pm 5\,\%$ V. Por tanto, los edificios E12, E18 y E19 poseen un valor mínimo de tensión inferior al límite normal de trabajo (380 V). Los edificios E18 y E19 se encuentran muy alejados del transformador de línea, existiendo una caída de tensión en la línea de alimentación. En el caso del edificio E12, el nivel de tensión se puede ver reducido debido a la elevada demanda de potencia que tienen los edificios alimentados por el mismo transformador (E10, E11 y E12).

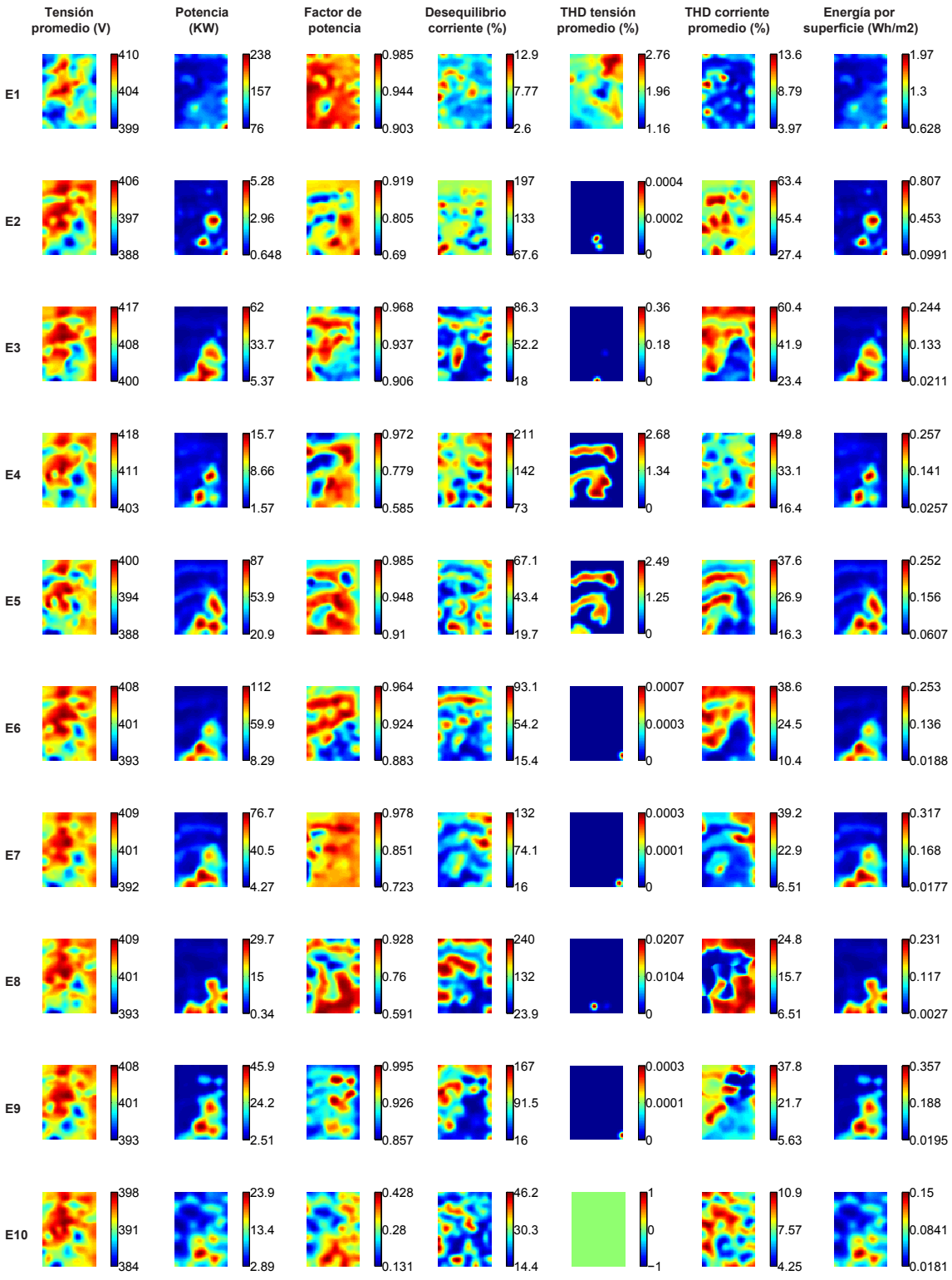


Figura 5.4: Planos de componentes correspondientes a las variables eléctricas de los edificios E1-E10.

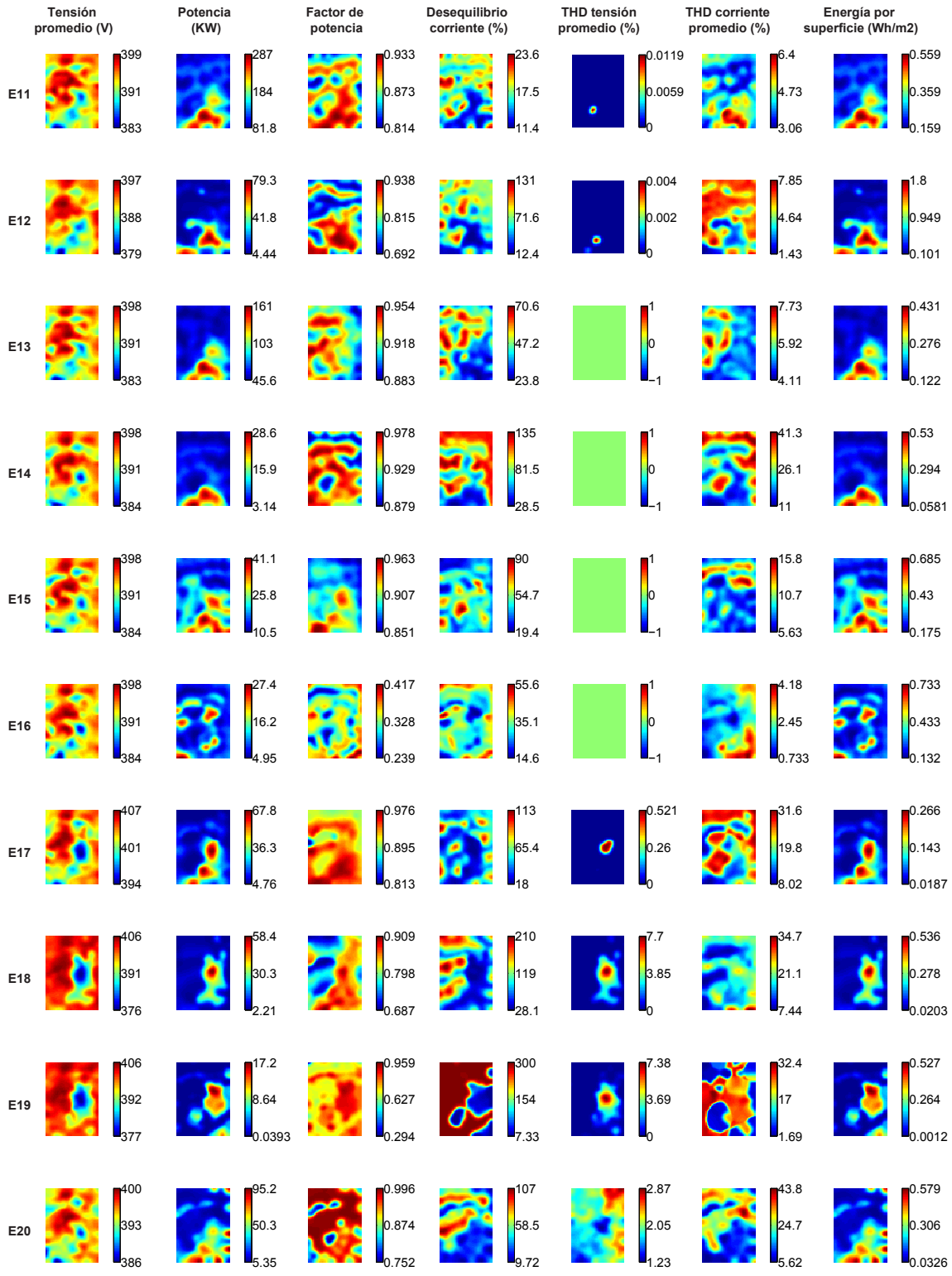


Figura 5.5: Planos de componentes correspondientes a las variables eléctricas en los edificios E11-E20.

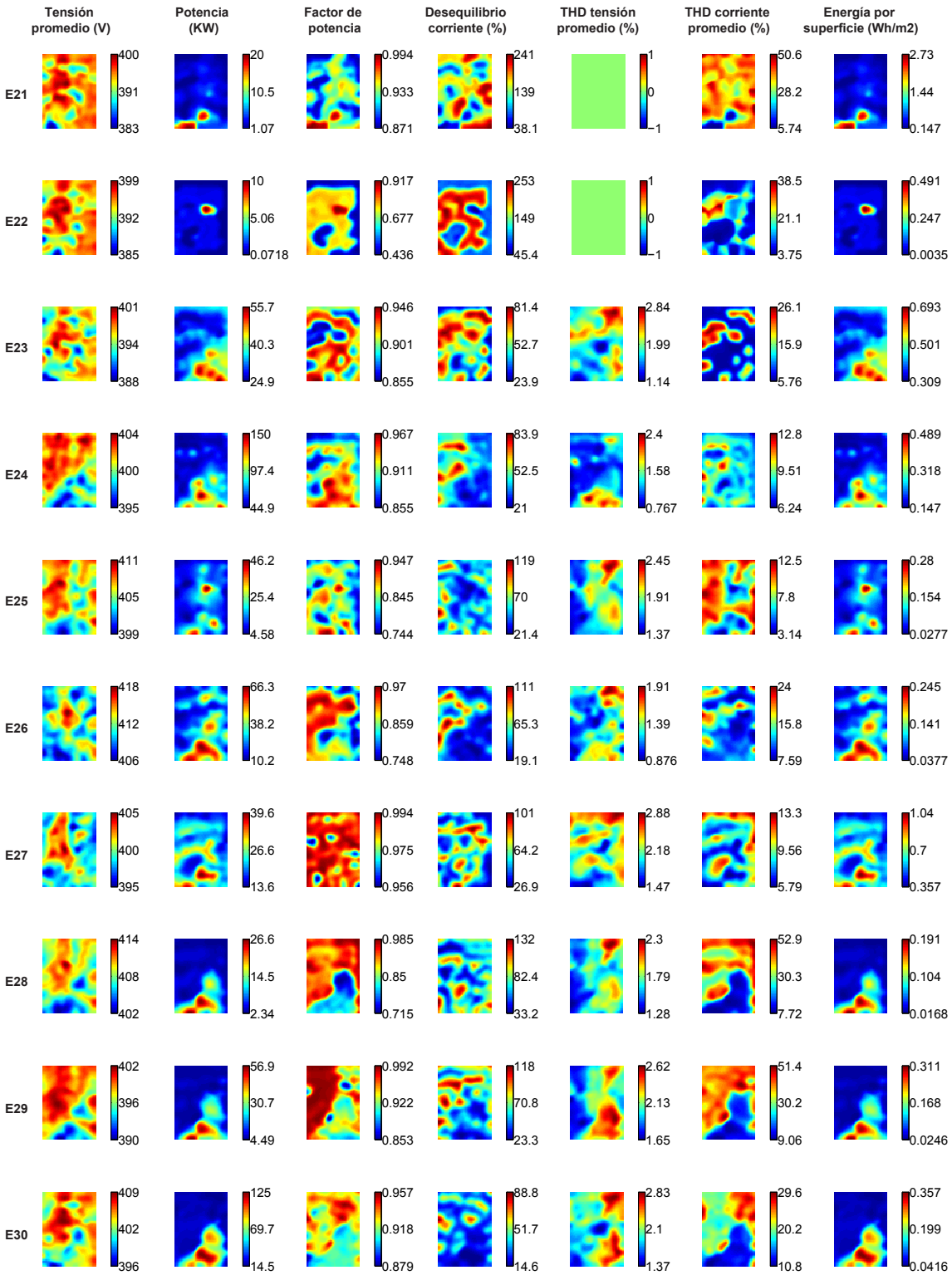


Figura 5.6: Planos de componentes correspondientes a las variables eléctricas en los edificios E21-E30.

No obstante, los planos de componentes de las tensiones presentan grandes similitudes, lo que indica que las variaciones de tensión adoptan el mismo patrón en todos los edificios. En general, en los mapas se observa que predominan los valores de tensión altos (rojo), situados en las zonas correspondientes a las horas nocturnas (entre las 23 y las 5 horas) o días festivos. También, se aprecia una clara relación entre la tensión y la potencia. El aumento de la demanda de potencia en el edificio, provoca una pequeña caída de tensión, excepto en el edificio E1, donde la potencia es bastante constante durante el año ya que el funcionamiento de servidores y sistemas de enfriamiento es continuo.

Variable potencia. La potencia varía en un amplio rango, dependiendo del edificio. Por ejemplo, el edificio E11 demanda una potencia que oscila entre 82 y 287 KW y en el edificio E2 esta variable toma valores entre 1 y 5 KW. El análisis de la potencia es muy importante ya que representa el consumo de energía instantáneo y permite conocer el perfil de carga del edificio.

Observando las zonas de elevada potencia (rojo) en todos los planos, se puede afirmar que la mayor parte de los edificios concentran su demanda de potencia en las horas de la mañana (8-14 horas) de los días laborables. Durante ciertas horas de la tarde de los días laborables (16-19 horas) existe una potencia demandada ligeramente inferior. Por el contrario, las zonas en los planos con una potencia mínima (azul) en el edificio corresponden a horas nocturnas o días festivos, como es lógico en este tipo de edificios dedicados a docencia, investigación y servicios complementarios. Existen algunos edificios (E2, E18, E19 y E22) dedicados a actividades de ocio de los estudiantes (radio, deporte y eventos musicales en la cafetería) en los cuales la potencia más alta se corresponde con las horas de la tarde (19-22 horas). Además, los edificios E10, E16, E20, E23, E25, E26 y E27 demandan potencia durante los días festivos debido al funcionamiento continuo del equipamiento de investigación, la apertura ininterrumpida de los centros de estudio en periodos de exámenes y la actividad en las residencias de estudiantes.

Por otra parte, en algunos edificios se puede apreciar una relación directa entre la temperatura y la potencia, es decir, la potencia demandada aumenta cuando la temperatura es alta, provocado por los sistemas de aire acondicionado, ventilación y refrigeración del edificio. Como ejemplo se pueden citar los edificios E1, E8, E15, E20, E23 y E24. Por el contrario, en otros edificios (E10, E12, E16, E25, E26 y E27) se observa una relación inversa entre estas dos variables, es decir, la potencia aumenta cuando la temperatura es baja, probablemente debido al sistema de calefacción del edificio.

Respecto a la radiación solar, se puede afirmar que la potencia disminuye cuando la radiación solar es máxima, debido a la desconexión de los sistemas de iluminación. No obstante, el tramo horario de máxima radiación solar coincide con un periodo de escasa actividad en los edificios (14-16 horas), lo que influye también en la reducción de la potencia. Cabe destacar que la potencia en el edificio E20 posee un alto valor cuando la radiación es alta en los días festivos, lo que corresponde a épocas de exámenes durante las cuales el horario de apertura de las salas de estudio de la biblioteca central es continuo.

Variable factor de potencia. El factor de potencia se puede considerar aceptable cuando varía entre 0.9 y 1. No obstante, actualmente la compañía eléctrica establece penalizaciones en la facturación si esta variable es inferior a 0.95, por lo que es recomendable que su valor sea lo más proximo posible a 1.

La mayor parte de los edificios poseen un factor de potencia bastante alejado de la unidad, lo que implica que serán objeto de recargos adicionales en la factura por exceso de energía reactiva. Solamente los edificios E1, E3, E5 y E27 poseen un rango de valores en el factor de potencia aceptables. En cambio, los edificios E2, E4, E8, E12, E18, E19 y E22 poseen un factor de potencia pésimo e inadmisible. Mención especial requieren los edificios E10 y E16, cuyo factor de potencia presenta un valor irreal inferior a 0.5, provocado por un error en el cableado del medidor que ya ha sido subsanado. El resto de los edificios presentan un factor de potencia que se debe mejorar, siempre y cuando el coste de las penalizaciones sea elevado.

En los edificios E2, E4, E5, E7, E8, E9, E10, E21, E24 y E25 se observa una clara correlación directa entre el factor de potencia y la potencia demandada. Por el contrario, el factor de potencia en el resto de edificios empeora cuando aumenta la potencia, debido principalmente a cargas de tipo inductivo. El factor de potencia no tiene un patrón ambiental marcado ya que depende en mayor medida de la potencia. Por lo tanto, existen edificios en los que el factor de potencia posee valores altos durante las noches (E29), las noches y festivos (E28), el día (E18) o el día y festivos (E7).

Variable desequilibrio en corriente. El desequilibrio en corriente proporciona información acerca del tipo de cargas en el edificio ya que la conexión no balanceada de cargas monofásicas en una instalación trifásica provoca un aumento de esta variable. Una distribución de cargas no equitativa entre las tres fases provoca efectos perjudiciales en los devanados de los transformadores y en las líneas de distribución de potencia. El desequilibrio en corriente toma valores bajos en el edificio E1 (3-13 %) y altos en E19 (7-300 %). El valor máximo posible (300 %) indica que únicamente circula corriente por una de las tres fases.

Los edificios E4, E8, E18, E19, E21 y E22 poseen un pésimo desequilibrio en corriente y sería imprescindible revisar su instalación eléctrica. Los edificios E2, E7, E9, E12, E14, E17, E20, E25, E26, E28, E29 superan en determinados instantes el valor 100 % y sería conveniente la revisión de sus instalaciones eléctricas. El desequilibrio en corriente del resto de edificios se considera aceptable.

Cabe destacar que el edificio E19 presenta un desequilibrio en corriente máximo cuando la potencia es casi nula, aunque su valor disminuye considerablemente cuando existe demanda de potencia. De hecho, el desequilibrio en corriente tiende a aumentar cuando disminuye la potencia. Como consecuencia, los valores más altos de esta variable se producen en horario nocturno y días festivos.

Variable THD promedio en tensión. La distorsión armónica total promedio en la tensión posee valores que oscilan entre 0 y 8 %. La mayoría de los edificios tienen un valor THD promedio en tensión nulo o muy próximo a 0. Por el contrario, el máximo valor de esta

variable se encuentra en los edificios E18 y E19, superando el límite de 5 % que la compañía eléctrica establece como normal, lo que podría provocar un funcionamiento deficiente en los receptores trifásicos. La causa de este elevado valor THD promedio en tensión puede ser la línea de alimentación desde el transformador hasta el edificio, de gran longitud y en paralelo con otras líneas.

Los edificios E1, E4, E5, E20, E23, E24 y E30 también poseen valores THD promedio en tensión altos, que se deben estudiar detenidamente. En este caso, las causas son múltiples y variadas dependiendo del edificio. Así por ejemplo, en E1 existen grandes sistemas de alimentación ininterrumpida (SAIs) para los servidores, mientras que en E20 existen amplias zonas iluminadas con lámparas de descarga.

Por otra parte, es conveniente destacar los edificios conectados a la red eléctrica en baja tensión (E25, E26, E27, E28 y E29), los cuales poseen un valor THD promedio en tensión muy similar entre sí, próximo a 2 %. Esto puede ser debido a la existencia de multitud de consumidores heterogéneos, que distorsionan la onda de tensión en la red eléctrica de distribución donde se encuentran conectados. Atendiendo a esta variable, se puede afirmar que las perturbaciones introducidas en la red eléctrica en baja tensión son normalmente más numerosas que en la red eléctrica en media tensión.

No obstante, es complicado establecer una patrón común para la variable THD en tensión. En los edificios alimentados desde la red eléctrica en baja tensión (E25, E26, E27, E28 y E29), esta variable toma generalmente valores altos en tramos diurnos. En otros casos, esta variable está fuertemente correlacionada con la demanda de potencia (E17, E18, E19 y E24).

Variable THD promedio en corriente. La distorsión armónica total promedio en la corriente posee valores entre 1 % y 63 %. Actualmente muchos receptores eléctricos incorporan circuitos de potencia, los cuales introducen gran cantidad de componentes armónicas en la red eléctrica. Valores muy elevados de esta variable provocan calentamiento de los conductores y receptores eléctricos, disminución del factor de potencia, aumento de la corriente útil, disparos erróneos de interruptores automáticos, aumento de las pérdidas por histéresis y corrientes parásitas en los transformadores, campos giratorios de secuencia inversa en motores, etc. Por tanto, la variable THD en corriente debe tomar valores bajos para evitar este tipo de problemas en los sistemas eléctricos.

Los edificios E1, E10, E11, E12, E13, E15, E16, E24, E25 y E27 poseen valores de THD en corriente adecuados (menores de 15 %). En cambio, los edificios E2, E3, E4, E21, E28 y E29 tienen valores elevados (mayores de 50 %) y es probable que surja algún tipo de anomalía en sus instalaciones. En el resto de edificios, la variable THD en corriente toma valores significativos comprendidos entre 15 % y 50 %, lo cual puede provocar calentamiento en los conductores.

Todos los edificios, excepto E11 y E19, presentan una relación inversa entre esta variable y la potencia, es decir que cuando aumenta la potencia demandada, disminuye el THD en corriente. En los edificios E2, E4, E9, E10, E12, E14, E15, E17, E21, E23 y E24, el factor de potencia disminuye cuando el THD en corriente aumenta, mientras que en los

edificios E3, E6, E8, E11, E13, E16, E19, E22, E28, E29 y E30 ambas variables están relacionadas directamente. El desequilibrio en corriente está directamente relacionado con el THD en corriente en los edificios E7, E13, E14, E20, E23 y E24. Esto indica la existencia de componentes homopolares en las corrientes de fase, las cuales circulan con mayor facilidad por el neutro de la instalación.

Por otra parte, el THD en corriente aumenta generalmente en los periodos diurnos y días laborables, excepto en los edificios E3, E6, E12, E28, E29 y E30.

Variable energía por superficie. La relación entre la energía y la superficie útil del edificio es un buen indicador del número de receptores eléctricos en la instalación. El valor máximo de esta variable es 2,73 Wh/m^2 y se produce en el edificio E21, mientras que el valor mínimo es 0,0012 Wh/m^2 y corresponde al edificio E19. Los edificios E1 y E12 también poseen un alto consumo de energía por unidad de superficie. El resto de edificios poseen un valor comprendido entre 0 y 1 Wh/m^2. En general, pequeños edificios que utilizan sistemas de climatización autónomos eléctricos o bien incorporan hornos y congeladores industriales como las cafeterías, tienen un valor elevado en este ratio. En el caso del edificio E1, existen multitud de ordenadores y servidores, así como sistemas de refrigeración, calefacción, ventilación y aire acondicionado, lo que hace que este ratio también alcance valores elevados.

Como ya se comentó anteriormente, la energía consumida está íntimamente relacionada con la potencia y por tanto, el valor máximo del ratio energía por superficie se produce cuando la demanda de potencia es máxima.

Comparación específica entre dos edificios.

Los planos de componentes procedentes del envSOM también se pueden utilizar para una comparación y análisis específicos entre dos edificios concretos. Para ello, nos centraremos en los edificios E1 y E2 situados de forma consecutiva en la parte superior de la figura 5.4. En una comparación visual general, se puede afirmar que todas las variables eléctricas de ambos edificios presentan grandes diferencias, tanto en rangos como en su distribución.

E1 tiene un nivel de tensión promedio superior a E2 y los correspondientes planos presentan una distribución completamente diferente, corroborando que estos edificios se alimentan desde puntos distintos de la red eléctrica en media tensión. E1 demanda una potencia constante, que aumenta cuando la temperatura y radiación solar son altas, mientras que E2 tiene una demanda de potencia puntual, centrada principalmente en dos tramos horarios. Respecto al factor de potencia, en E1 se tienen habitualmente valores excelentes y en E2 aparecen valores muy bajos durante las noches y días festivos. Lo mismo sucede con las variables desequilibrio en corriente y THD promedio en corriente. Sin embargo, E2 tiene mejores valores que E1 en la variable THD promedio en tensión. El ratio energía por superficie toma valores mayores en el edificio E1, debido a su numeroso equipamiento eléctrico.

Otras comparaciones específicas de este tipo se podrían realizar para el resto de edificios, aunque este proceso visual es lento y costoso porque la información está aún muy dispersa

en un número elevado de planos de componentes (210), correspondientes a los 30 edificios y 7 variables eléctricas. Por este motivo, es necesario condensar más la información por medio de una proyección en un espacio de visualización 2D, donde sea rápido y sencillo analizar y comparar los edificios entre sí.

Los gráficos de comparación se obtienen aplicando una proyección de Sammon a las matrices de similitud, las cuales se calculan a partir de los vectores *codebook* de la segunda fase del envSOM. De forma complementaria, se representan los grupos de edificios con un comportamiento eléctrico similar que se obtienen al aplicar el método *k-means*. La validación del agrupamiento para todas las comparaciones (global, individuales y condicionadas) se resume en la tabla 5.5. El número de grupos k óptimo se ha elegido en base a los valores mínimos del índice de Davies & Bouldin y/o del error medio de los datos con respecto al centroide, los cuales se encuentran resaltados en negrita en la tabla.

A continuación se muestran los resultados obtenidos en las comparaciones entre edificios a partir de las diferentes matrices de similitud indicadas en el punto 4.4.2 y el correspondiente agrupamiento.

5.2.2. Comparación global

La primera comparación sencilla entre edificios se obtiene como resultado de la proyección de Sammon a partir de la matriz de similitud global, calculada mediante la ecuación 4.12. El resultado de esta primera proyección se puede observar en la figura 5.7. En este gráfico de comparación, cada edificio se representa por un punto, de tal forma que edificios próximos en el plano 2D presentan un comportamiento eléctrico muy similar, mientras que edificios proyectados muy distantes se comportan de manera muy diferente. Esta primera comparación está basada en un compendio de todas las variables eléctricas que definen el comportamiento de los edificios. Además, en el mismo gráfico 2D se han etiquetado los grupos de edificios que resultaron de aplicar el método *k-means*.

En el gráfico de comparación global, se puede observar que los edificios E1, E2, E3, E4, E10, E11, E16, E19, E21, E22 y E28 destacan del resto, proyectados en el centro del gráfico. Las causas son múltiples, por ejemplo, E1 y E11 poseen un rango de potencia muy superior al resto, E2, E3, E4, E21, E22 y E28 tienen valores altos en el desequilibrio y THD promedio en corriente, E19 alcanza valores máximos en el desequilibrio en corriente y E10 y E16 toman valores muy bajos en el factor de potencia. En ambos edificios (E10 y E16) se ha detectado un error en el cableado de los medidores que provocaba estos valores pésimos.

Los edificios E1 y E2 presentan una gran disimilitud en su comportamiento eléctrico porque se proyectan muy alejados en el gráfico y pertenecen a grupos diferentes (grupo 3, o rombos en rojo, y grupo 5, o estrellas en magenta, respectivamente). Esto coincide con el resultado de la comparación visual entre planos de componentes del envSOM, expuesto anteriormente. El número óptimo de grupos de edificios es 6, es decir, que existen 6 perfiles de comportamiento eléctrico diferentes en el Campus de la Universidad de León. La mayoría de edificios se comportan de forma similar ya que pertenecen a los grupos 2 (cuadrados en verde) y 3 (rombos en rojo). El grupo 4 (triángulos hacia la derecha en cian) contiene

		k=2		k=3		k=4		k=5		k=6		k=7		k=8		k=9		k=10	
		Índice	Error	Índice	Error	Índice	Error	Índice	Error	Índice	Error	Índice	Error	Índice	Error	Índice	Error	Índice	Error
Global	Todas las variables eléctricas	1.6221	14986	1.6665	12602	1.1933	9486	1.2888	8376	**1.1525**	**7041**	1.1439	6225	1.0559	5853	1.0213	4962	0.9666	4677
Individual	Tensión	0.6830	756	0.6507	366	0.5700	236	0.8550	206	**0.5740**	**153**	0.6799	186	0.6191	182	0.7038	115	0.5488	93
	Potencia	0.5710	413	0.7491	282	0.9082	251	**0.7737**	**125**	0.7788	103	0.9655	223	0.8896	218	0.8035	79	0.7980	46
	Factor de potencia	0.2255	1131	0.5647	498	0.7319	413	0.8247	375	**0.6339**	**242**	0.9011	206	0.8954	193	0.8085	179	0.7329	156
	Desequilibrio en corriente	0.8686	1751	1.0042	1423	1.0070	1291	1.3272	1259	**0.9541**	**674**	0.9375	520	0.8111	560	0.8564	385	0.7245	493
	THD en corriente	0.8278	1955	0.9011	1250	0.9707	1026	0.9124	830	0.9593	744	**0.7965**	**633**	1.0225	571	0.8421	474	0.8031	494
	Energía por superficie	0.8666	311	1.0285	232	**0.7939**	**148**	0.8248	121	0.8644	113	0.9726	108	0.8134	85	0.7731	80	0.7403	82
Condicionado	Potencia (7-15h)	0.6141	204	0.6159	130	**0.4677**	**56**	0.5581	40	0.6727	34	0.6382	30	0.7197	27	0.7305	26	0.7257	28
	Potencia (15-22h)	0.5643	230	0.7593	156	0.7380	133	**0.6795**	**65**	0.7358	62	0.8353	51	0.8416	49	0.8307	47	0.5826	23
	Potencia (22-7h)	0.4993	267	0.5806	188	**0.4364**	**64**	0.7951	169	0.6362	45	0.6983	165	0.4956	19	0.5021	32	0.5197	32
	Potencia (16-31°C)	0.5406	134	0.6269	74	**0.5715**	**47**	0.7053	35	0.7642	33	0.8362	31	0.7436	28	0.7432	25	0.7872	27
	Potencia (2-15°C)	0.5880	274	0.6445	185	0.9439	167	0.8183	159	**0.6924**	**57**	0.7302	48	0.7904	54	0.6617	41	0.7689	40
	Potencia (laborables)	0.5867	339	0.7301	222	0.7978	197	0.9212	187	**0.6567**	**72**	0.9130	176	0.8968	175	0.7803	59	0.7594	52
	Potencia (festivos)	0.5163	239	0.5371	111	**0.4357**	**54**	0.5754	43	0.5572	35	0.7605	33	0.5026	30	0.7401	30	0.6216	25
	Potencia (8-21h, 0-200W/m^2 y laborables)	0.6074	129	0.6181	80	**0.4805**	**35**	0.5067	24	0.7935	22	0.7401	18	0.8451	16	0.7808	18	0.7158	15
	Factor de potencia (18-24h y festivos)	0.2296	128	**0.4733**	**44**	0.5422	30	0.7592	28	0.7149	19	0.8114	17	0.7269	15	0.6233	14	0.6849	13
	Desequilibrio en corriente (7-22h y laborables)	0.8443	504	1.1369	429	0.8309	256	**0.6019**	**221**	0.6766	202	0.8513	192	0.6527	93	0.7679	110	0.6698	73
	THD en corriente (7-22h y laborables)	0.9100	606	0.8220	301	0.9849	263	0.9989	260	**0.7109**	**155**	0.9649	195	0.8215	173	0.7719	110	0.7725	83

Tabla 5.5: Validación del agrupamiento realizado con el método *k-means*.

solamente el edificio E19, que posee un gran desequilibrio en corriente. Los edificios E10 y E16 constituyen el grupo 6 (triángulos hacia la izquierda en amarillo) debido al pésimo factor de potencia que poseen.

5.2.3. Comparación para cada variable eléctrica

La segunda comparación sencilla posible se obtiene como resultado de las proyecciones de Sammon a partir de las matrices de similitud para cada variable eléctrica, calculadas mediante la ecuación 4.13. Como resultado se obtienen 6 gráficos de comparación, uno por cada variable, excepto para la variable THD promedio en tensión que toma valores nulos en varios edificios, provocando que la matriz de similitud tenga elementos nulos y por tanto, es imposible aplicar la proyección de Sammon.

Variable tensión promedio. La figura 5.8 muestra la comparación entre edificios con respecto a la variable tensión promedio de las tres fases. Se puede observar que el nivel de tensión suministrado por la compañía eléctrica es muy similar en la mayoría de edificios. Destacan por su elevada tensión los edificios E3, E4, E26 y E28 (grupo 6, o triángulos hacia la izquierda en amarillo) y por su bajo nivel E18 y E19 (grupo 3, o rombos en rojo). Los edificios E1 y E2 se proyectan no muy alejados, aunque pertenecen a grupos distintos (5, o estrellas en magenta, y 4, o triángulos hacia la derecha en cian). El número de perfiles de tensión en los edificios del Campus de la Universidad de León es 6.

Comentar que los edificios alimentados desde el mismo transformador de línea se encuentran muy próximos entre sí ya que la tensión es idéntica para todos ellos. Por ejemplo los edificios E13, E14, E15 y E16 están conectados al mismo punto y se proyectan cercanos. Lo mismo sucede con los edificios E6, E7, E8 y E9.

Variable potencia. La figura 5.9 muestra la comparación entre edificios con respecto a la variable potencia. En esta imagen se puede observar que el perfil de potencia en los edificios E1, E11, E13 y E24 difiere claramente del resto ya que se proyectan bastante alejados, sobresaliendo de la mayoría. Además, estos 4 edificios constituyen el grupo 5 (estrellas en magenta). Todos poseen características muy similares (superficie elevada, grandes sistemas de refrigeración y aire acondicionado y numeroso equipamiento industrial para investigación).

Los edificios E1 y E2 se proyectan muy alejados entre sí en el gráfico debido a una gran diferencia en la demanda de potencia, perteneciendo a grupos distintos (grupo 4, o triángulos hacia la derecha en cian, y grupo 5, o estrellas en magenta). Los edificios pertenecientes al grupo 3 (rombos en rojo) poseen una demanda de potencia significativa y se deben tener en cuenta a la hora de reducir el consumo.

Existen 5 perfiles distintos de demanda de potencia en el Campus de la Universidad de León. Esto induce a pensar en la agrupación en un mismo punto de facturación de todos los edificios pertenecientes al mismo grupo con el fin de aprovechar las similitudes que poseen en cuanto a la variable potencia.

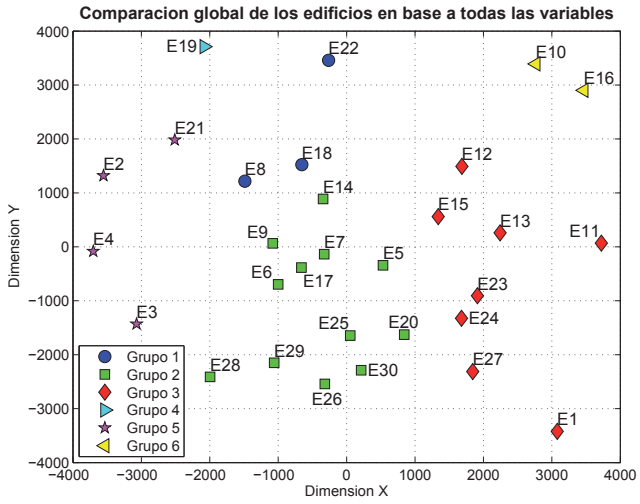


Figura 5.7: Comparación y grupos de edificios en base a todas las variables eléctricas.

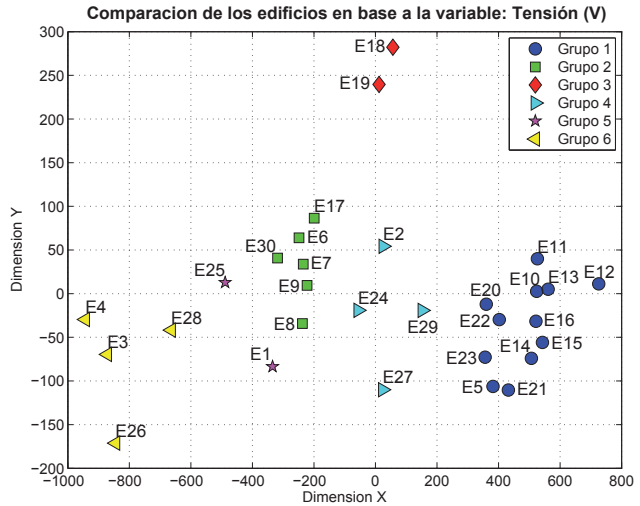


Figura 5.8: Comparación y grupos de edificios en base a la variable tensión promedio.

Variable factor de potencia. En la figura 5.10 se presenta la comparación entre edificios con respecto a la variable factor de potencia. Aquí se puede observar claramente como los edificios E10 y E16 se representan muy alejados del resto porque presentan valores pésimos en el factor de potencia y por tanto, muy diferentes de la mayoría. Como ya se comentó anteriormente, esto ha dado pie a revisar ambos medidores, encontrando un error en su cableado que ya se ha corregido.

Los edificios E8 y E22 destacan en la figura ya que ambos han permanecido sin actividad un largo periodo de tiempo durante el año tomado para las pruebas, lo que ha provocado valores muy bajos en el factor de potencia. En general, el factor de potencia mejora cuando la demanda de potencia aumenta, excepto en E28. Este edificio tiene una gran demanda de potencia provocada por receptores inductivos y carece de equipos para la corrección del factor de potencia, por lo que también sobresale en el gráfico. E4 y E12 toman valores bajos en el factor de potencia, aunque sólo durante las noches y días festivos. Respecto a los edificios E1 y E2, se puede decir que poseen ciertas diferencias en su factor de potencia ya que están alejados y pertenecen a grupos distintos (grupo 2, o cuadrados en verde, y grupo 3, o rombos en rojo, respectivamente).

El número de grupos teniendo en cuenta esta variable es 6. Los edificios de los grupos 1, o círculos en azul (E10 y E16), 3, o rombos en rojo (E2, E25, E26 y E28), 4, o triángulos hacia la derecha en cian (E4, E8, E12, E18 y E19), y 6, o triángulos hacia la izquierda en amarillo (E22), así como el edificio E20 (grupo 2, o cuadrados en verde) tienen un factor de potencia inadecuado. Por tanto, se debería actuar urgentemente para corregir esta variable y así evitar penalizaciones en la facturación.

Variable desequilibrio en corriente. En la figura 5.11 se presenta la comparación entre edificios con respecto a la variable desequilibrio en corriente. En esta figura podemos ver que el edificio E19 se proyecta distante del resto ya que toma valores máximos en el desequilibrio en corriente. Además, constituye en solitario el grupo 3 (rombos en rojo). Los edificios E2, E4, E8, E9, E14, E18, E21 y E22 también se alejan del resto. En general, cuando el consumo de corriente es bajo, la medida del desequilibrio es alta (E2, E8, E19, E21 y E22).

Existen 6 tipos de perfiles en cuanto al desequilibrio en corriente. La mayoría de los edificios pertenecen a los grupos 1 (círculos en azul) y 2 (cuadrados en verde). Analizando los edificios E1 y E2, se tiene que se proyectan de nuevo bastante alejados y son contenidos en el grupo 1, o círculos en azul (E1), y en el 5, o estrellas en magenta (E2).

Variable THD promedio en corriente. La figura 5.12 muestra la comparación entre edificios con respecto a la variable THD promedio en corriente. Como se puede observar, existen comportamientos muy dispares por lo que respecta a esta variable. Los edificios E1 y E2 se representan muy alejados entre sí, lo que indica una gran disimilitud entre ambos en cuanto al THD promedio en corriente.

El número de perfiles relativos a la variable THD promedio en corriente es 7. Los edificios E2, E3 (grupo 5, o estrellas en magenta de cinco puntas) y E19 (grupo 2, o cuadrados en verde) destacan del resto debido a sus altos valores, por lo que se debería actuar

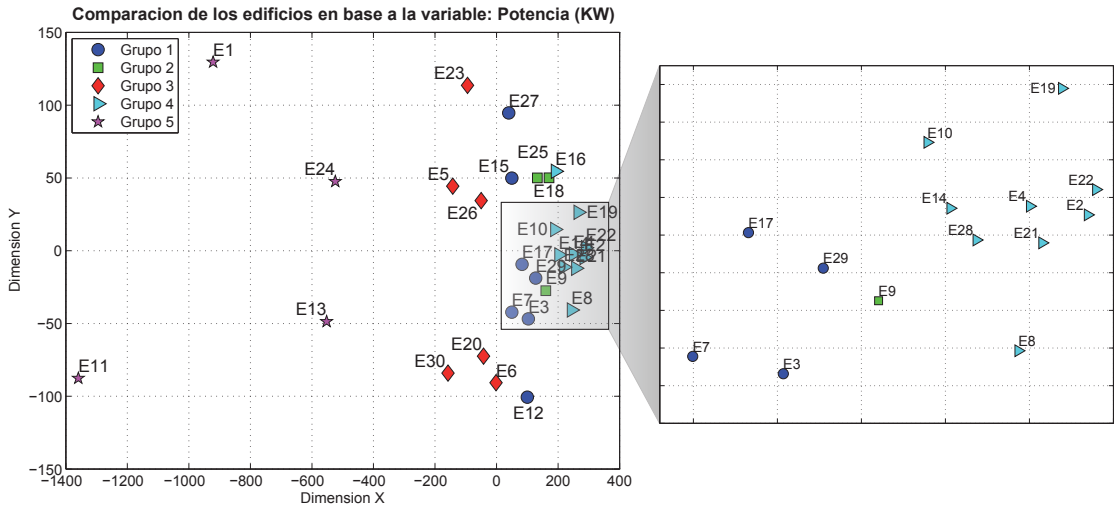


Figura 5.9: Comparación y grupos de edificios en base a la variable potencia.

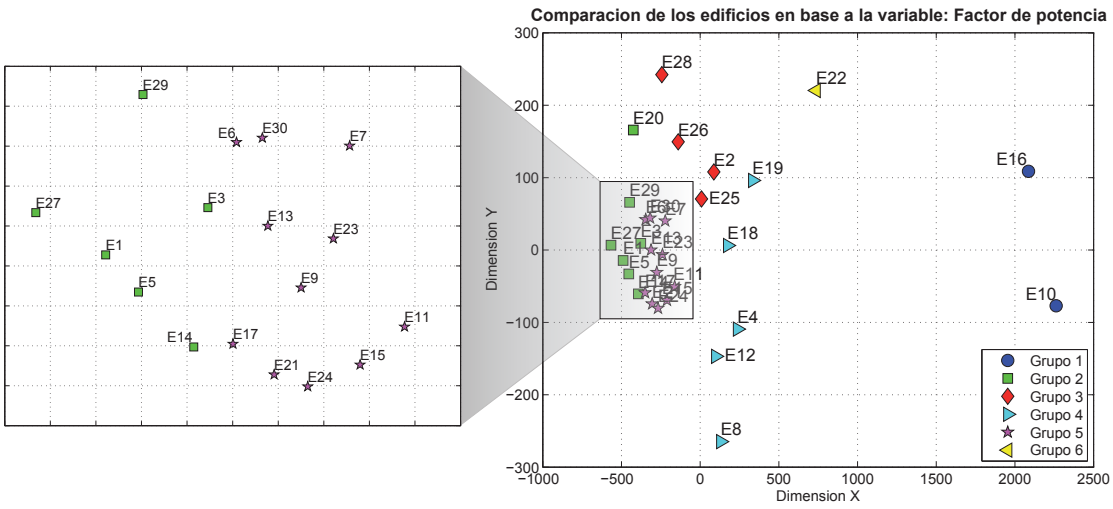


Figura 5.10: Comparación y grupos de edificios en base a la variable factor de potencia.

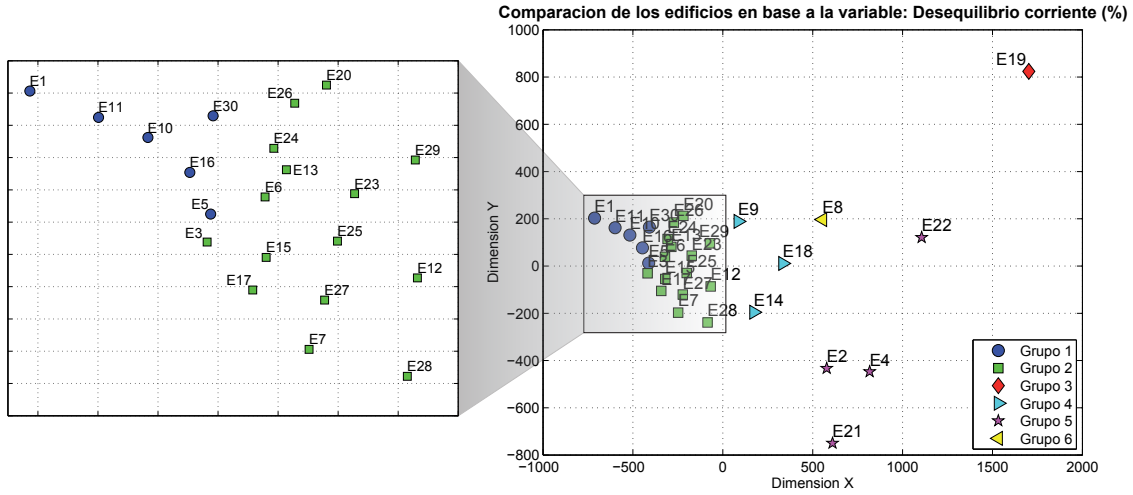


Figura 5.11: Comparación y grupos de edificios en base a la variable desequilibrio en corriente.

urgentemente en su instalación para evitar daños. Analizando la variable THD promedio en corriente junto con el factor de potencia y el desequilibrio en corriente, se puede observar que los edificios E2, E4, E8, E19 y E22 destacan del resto en los tres gráficos simultáneamente, lo que indica que existe una relación entre estas tres variables. Además, la potencia es baja en todos ellos, lo que puede condicionar al resto de variables eléctricas.

Variable energía por superficie. Finalmente, la figura 5.13 muestra la comparación entre edificios con respecto a la variable energía consumida por superficie útil. Aquí se pueden observar los edificios que poseen un consumo de energía elevado en función de su superficie, lo que indica la densidad y tipo de receptores eléctricos. Por ejemplo, los edificios E1 y E27 (grupo 2, o cuadrados en verde) junto con E12 y E21 (grupo 3, o rombos en rojo) tienen un valor elevado para este ratio. Una vez más, E1 y E2 se proyectan alejados entre sí.

Existen 4 perfiles distintos en todos los edificios del Campus. La mayoría de ellos poseen un ratio medio-bajo, admisible en este tipo de edificios, perteneciendo a los grupos 1 (círculos en azul) y 4 (triángulos hacia la derecha en cian).

5.2.4. Comparación condicionada por las variables ambientales

La tercera comparación sencilla posible se obtiene como resultado de las proyecciones de Sammon a partir de las matrices de similitud para una o todas las variables eléctricas condicionadas por el entorno ambiental, calculadas mediante la ecuación 4.14. Como resultado se pueden obtener multitud de gráficos de comparación, dependiendo de las variables eléctrica y ambiental elegidas y los umbrales que definen la condición. A continuación, se presentan algunos resultados que revelan claras influencias de las variables

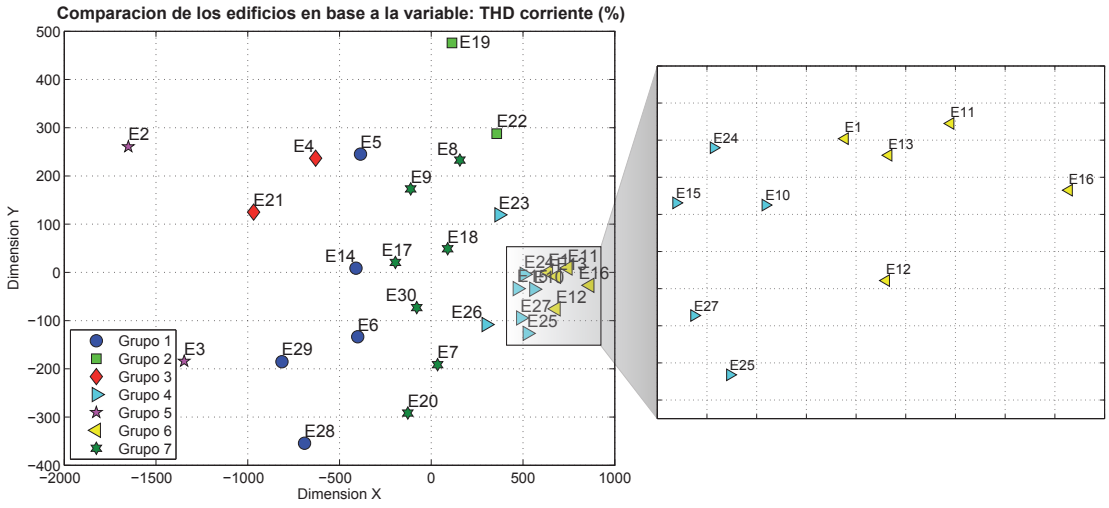


Figura 5.12: Comparación y grupos de edificios en base a la variable THD promedio en corriente.

ambientales en las eléctricas.

Potencia condicionada por la variable hora. En la figura 5.14 se muestra la comparación entre edificios con respecto a la variable potencia condicionada por tres tramos horarios. Un análisis conjunto de las tres proyecciones condicionadas también aporta conocimiento de los edificios.

En la figura 5.14a se puede comparar los edificios con respecto a la potencia demandada durante la mañana. La mayor parte de los edificios se proyectan próximos entre sí, excepto E1, E13, E24 (grupo 2, o cuadrados en verde) y E11 (grupo 3, o rombos en rojo). El número de perfiles de potencia durante la mañana es 4.

En la figura 5.14b se comparan los edificios con respecto a la potencia demandada durante la tarde. Se puede observar que los edificios se proyectan más dispersos que durante la mañana, lo que indica que existen mayores diferencias en los perfiles de potencia. Por lo tanto, el número de perfiles es mayor (5 grupos).

En la figura 5.14c se presenta la comparación de los edificios con respecto a la potencia demandada durante la noche. Las distancias entre la mayoría de edificios se reducen considerablemente, aunque aumentan con respecto a E1, E11, E13 y E24 (grupo 4, o triángulos hacia la derecha en cian), lo que revela que estos edificios demandan una elevada potencia residual de mantenimiento durante periodos sin actividad debido al funcionamiento continuo de algunos receptores eléctricos (servidores, equipamiento de investigación, etc.). Como es lógico, el número de perfiles es ahora menor (4 grupos), debido a la escasa actividad en la mayoría de los edificios.

Comparando las tres proyecciones entre sí, se puede afirmar que la potencia demandada en la mayor parte de los edificios durante la mañana y la noche es similar, mientras que

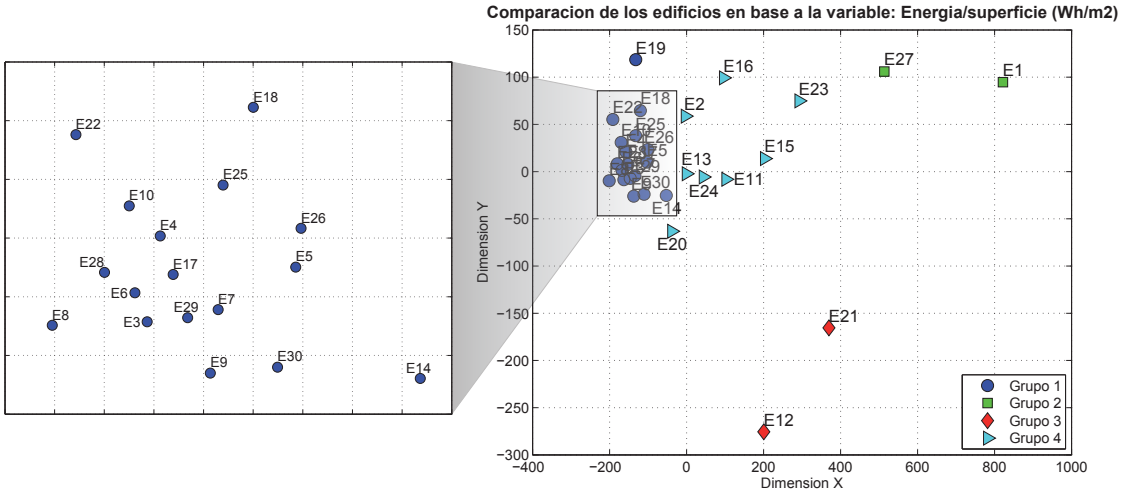


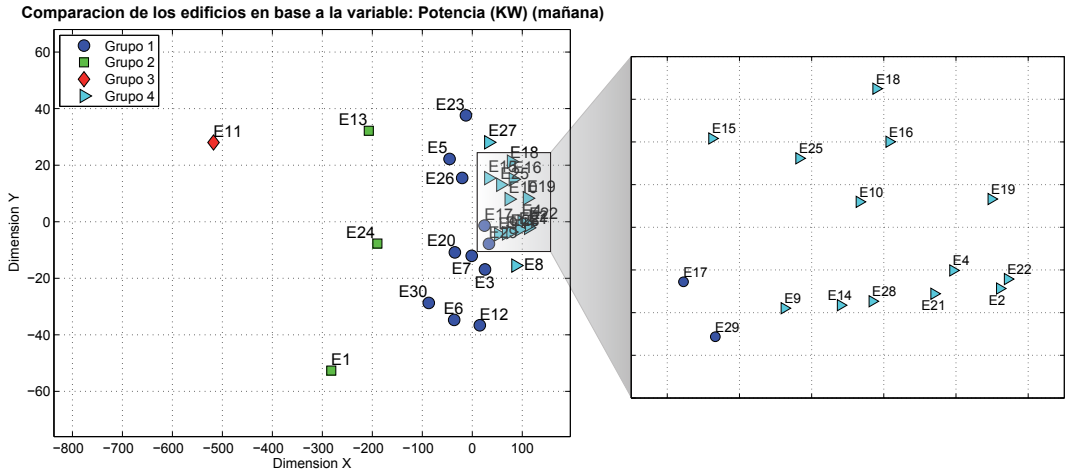
Figura 5.13: Comparación y grupos de edificios en base a la variable energía por superficie.

por la tarde difiere un poco. Observando los edificios E18 y E19 en las tres proyecciones, podemos ver que se proyectan un poco más alejados del resto durante la tarde, lo que indica que su demanda de potencia es mayor en este tramo horario. Estos edificios corresponden a los centros deportivos de la Universidad de León y la mayor actividad en ellos tiene lugar durante las horas de la tarde.

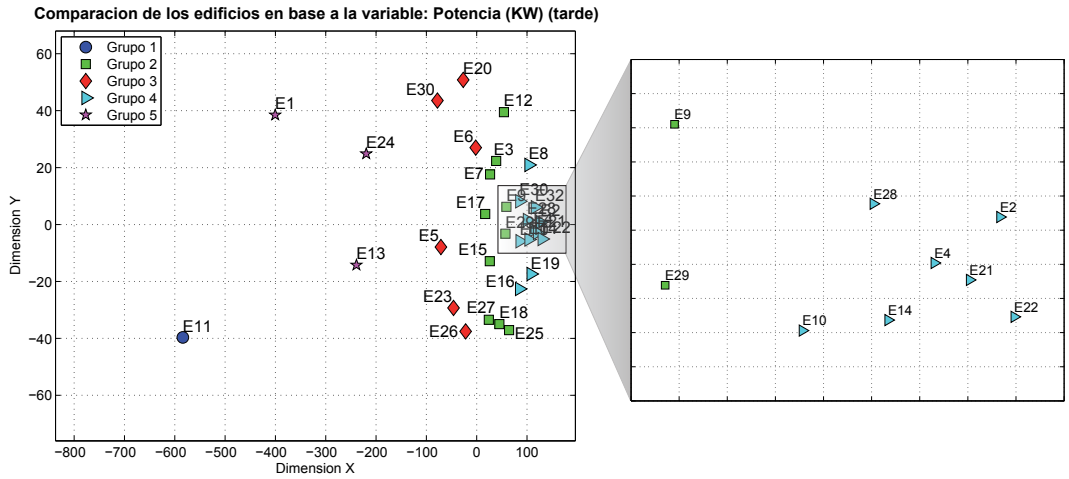
Potencia condicionada por la variable temperatura. En la figura 5.15 se presentan la comparaciones entre edificios con respecto a la variable potencia condicionada por la temperatura, lo que permite detectar los edificios donde la temperatura influye en su demanda de potencia.

En la figura 5.15a se puede comparar los edificios con respecto a la potencia demandada cuando la temperatura es alta (16-31 ºC). El número total de perfiles diferentes es 4 y la mayoría de edificios se representan muy próximos entre sí, perteneciendo al grupo 1 (círculos en azul). El motivo es que temperaturas altas corresponden principalmente a los meses de verano, en los cuales la actividad en los centros es baja (periodo de vacaciones). En los edificios E1, E11, E13, E20, E23 y E24 existe cierta actividad durante el verano y además, los equipos destinados a investigación y los sistemas de refrigeración, ventilación y aire acondicionado tienen mucho peso en la demanda total de potencia.

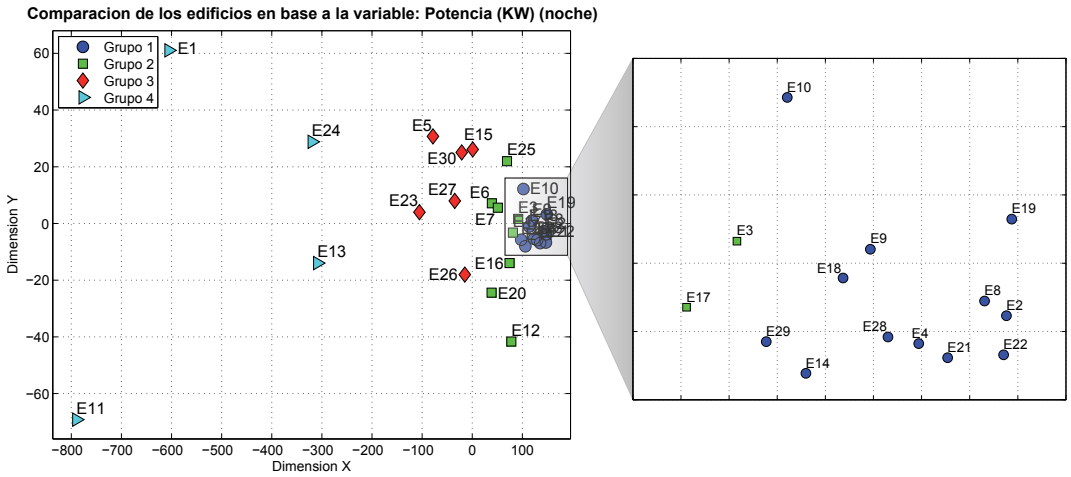
En la figura 5.15b se comparan los edificios con respecto a la potencia demandada cuando la temperatura es baja (2-15 ºC). Los edificios se proyectan dispersos en el gráfico, lo que indica que existen perfiles de potencia muy diferentes. De hecho, el número de perfiles es mayor en este caso (6). Los rangos de temperatura bajos corresponden a los meses de invierno, en los cuales el número de personas y actividades en los edificios es elevado, provocando un aumento de la potencia demandada. Este incremento de potencia enmascara el consumo de energía de los sistemas de calefacción eléctricos, por lo que es complicado detectar los edificios en los que se hace un uso abusivo de estos sistemas de aporte de calor



(a) Mañana (7-15 h)



(b) Tarde (15-22 h)



(c) Noche (22-7 h)

Figura 5.14: Comparación y grupos de edificios en base a la variable potencia condicionada por la hora.

extra en invierno.

Potencia condicionada por la variable tipo de día. La figura 5.16 muestra las comparaciones entre edificios con respecto a la variable potencia condicionada por el tipo de día, lo que permite detectar aquellos edificios donde se demanda potencia durante los días festivos.

En la figura 5.16a se puede comparar los edificios con respecto a la potencia demandada en los días laborables. Los edificios se proyectan dispersos, lo que indica que existen diferencias en su demanda de potencia. En el gráfico se han etiquetado 6 perfiles distintos. De nuevo, los edificios E1, E11, E13 y E24 (grupo 3, o rombos en rojo) destacan del resto debido a su intensa actividad en los días de trabajo.

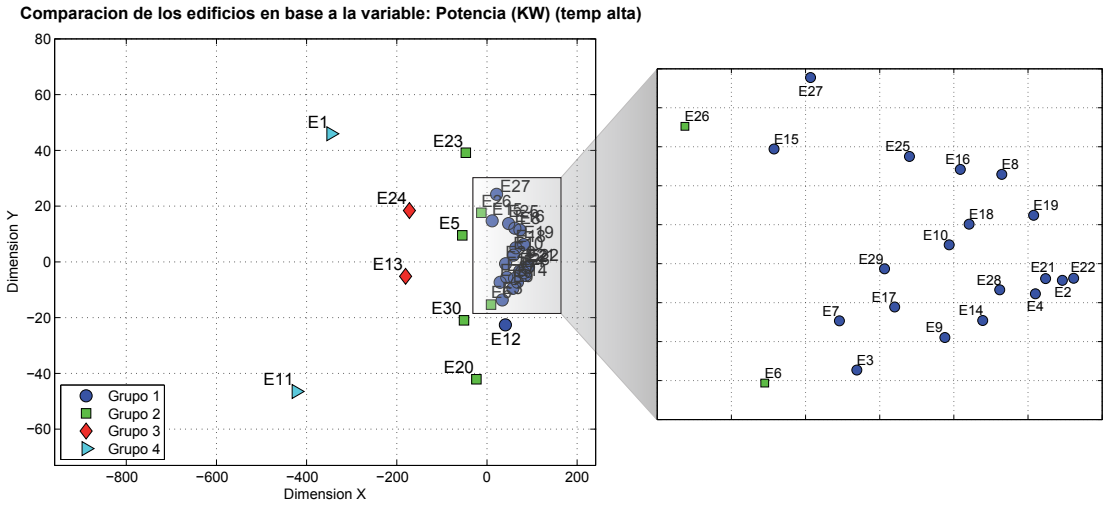
En la figura 5.16b se puede comparar los edificios con respecto a la potencia demandada en los días festivos. Los edificios se representan muy próximos entre sí, excepto E1 y E11 (grupo 1, o círculos en azul). En este caso, el número de perfiles es menor (4) debido a la escasa actividad. Cabe destacar el edificio E20, donde hay actividad en determinados días festivos que coinciden con los periodos de exámenes finales.

Potencia condicionada por tres variables: hora, tipo de día y radiación solar. En todas las comparaciones condicionadas anteriores, la condición se define mediante una única variable ambiental. No obstante, se pueden combinar diferentes variables ambientales para definir una condición múltiple que pueda influir en una variable eléctrica. Como ejemplo, en la figura 5.17 se muestra una comparación en base a la potencia condicionada por 3 variables: la hora (8-21 horas), el tipo de día (laborables) y la radiación solar (0-200 W/m^2). El objetivo de este análisis es conocer en que edificios destaca el consumo eléctrico relativo a su sistema de iluminación. Para ello, se presupone que la iluminación artificial es necesaria durante las horas y días de apertura del edificio, siempre que la radiación solar sea baja.

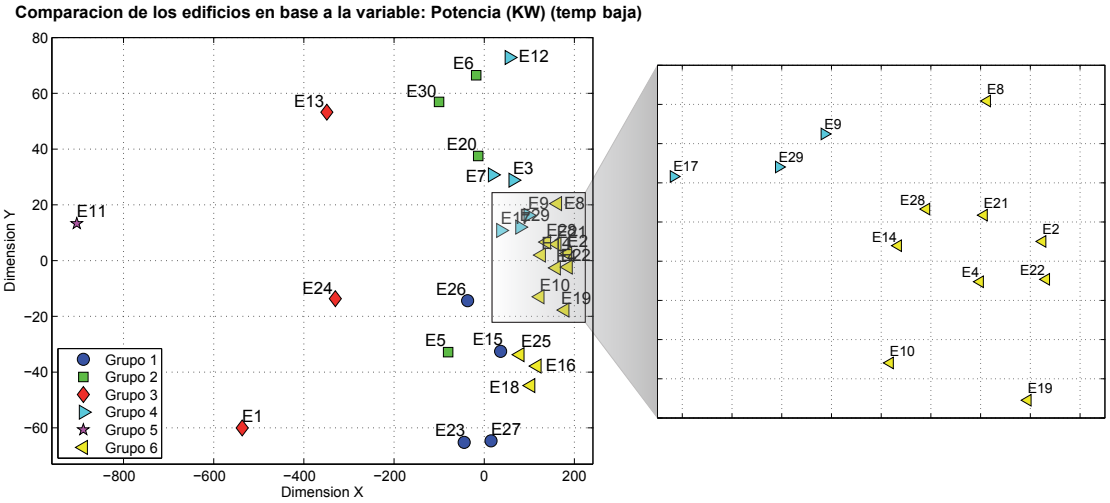
Los edificios contenidos en los grupos 3, o rombos en rojo (E11), y 4, o triángulos hacia la derecha en cian (E1, E13 y E24), sobresalen en el gráfico de comparación. Asimismo, los edificios E5, E6 y E30 poseen una demanda de potencia alta en esta situación. La característica común en todos ellos es la gran superficie a iluminar. En cambio, otros edificios con una superficie inferior (E12, E18 y E23) también tienen una potencia alta que destaca en el gráfico. En el caso del edificio E18, la causa puede ser el gran número de focos halógenos que iluminan el pabellón deportivo.

Cabe destacar que en el edificio E20 se instalaron luminarias con mejor eficiencia energética, por lo que se encuentra proyectado un poco más cerca de la mayoría de edificios. Recientemente, también se ha modernizado el sistema de iluminación en E5 y por lo tanto, cabe esperar que la distancia respecto a la mayoría de edificios será menor en futuras comparaciones, corroborándose el ahorro energético.

Factor de potencia condicionado por dos variables: hora y tipo de día. Las comparaciones condicionadas mostradas hasta ahora se han centrado en la variable potencia. Sin embargo, podría ser interesante que otras variables eléctricas, tales como el factor

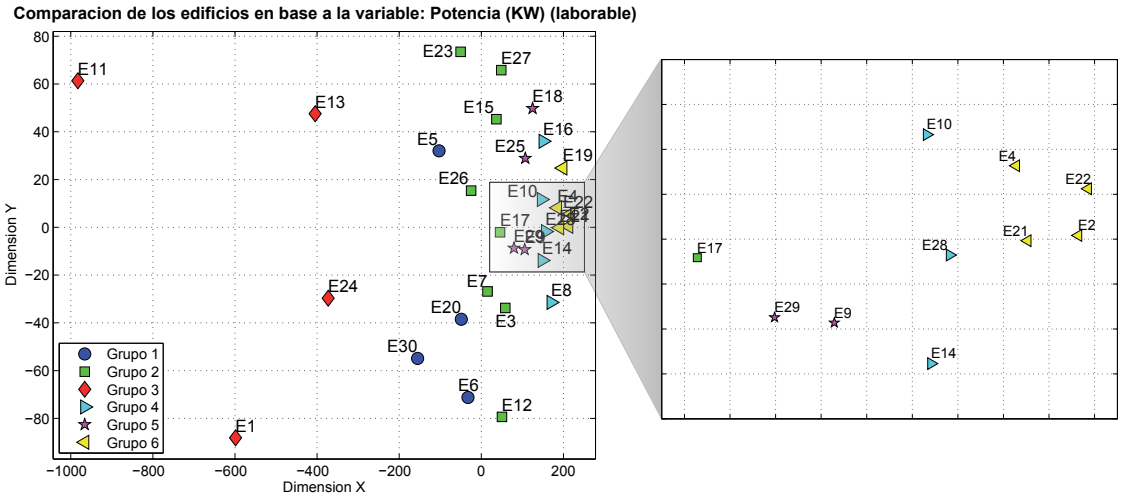


(a) Temperatura alta (16-31 °C)

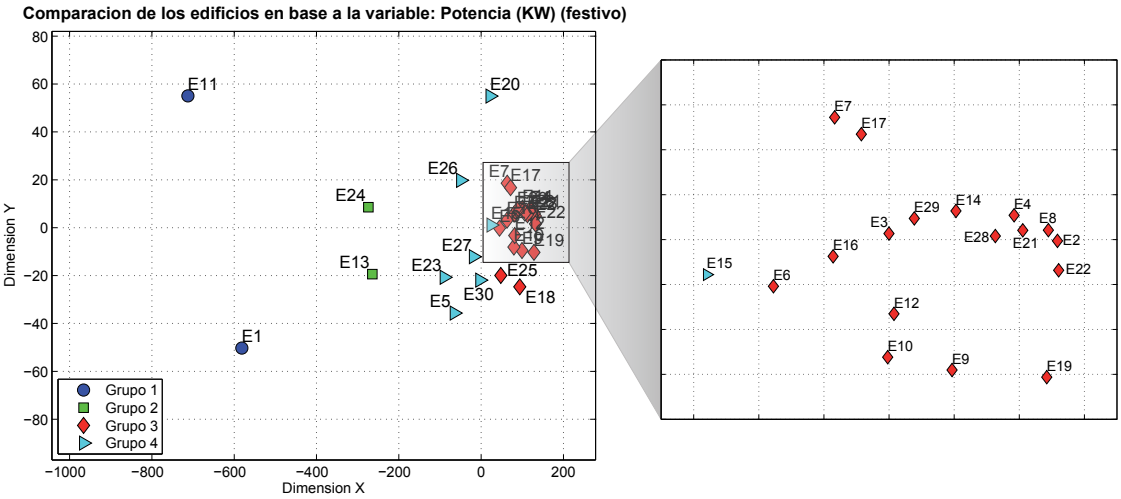


(b) Temperatura baja (2-15 °C)

Figura 5.15: Comparación y grupos de edificios en base a la variable potencia condicionada por la temperatura.



(a) Días laborables



(b) Días festivos

Figura 5.16: Comparación y grupos de edificios en base a la variable potencia condicionada por el tipo de día.

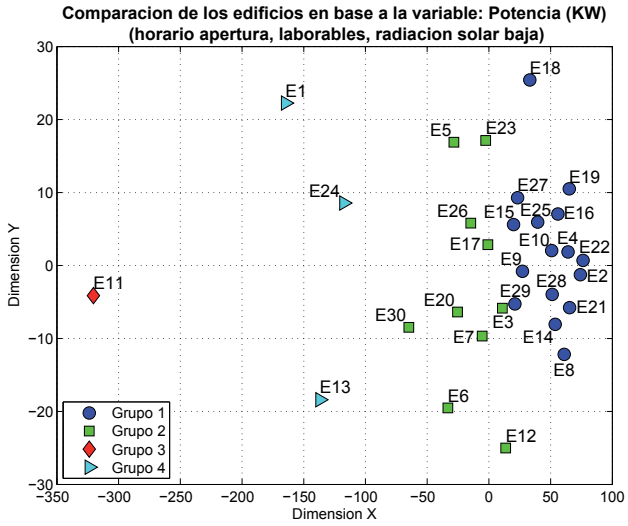


Figura 5.17: Comparación y grupos de edificios en base a la variable potencia condicionada por tres variables ambientales (hora, tipo de día y radiación solar).

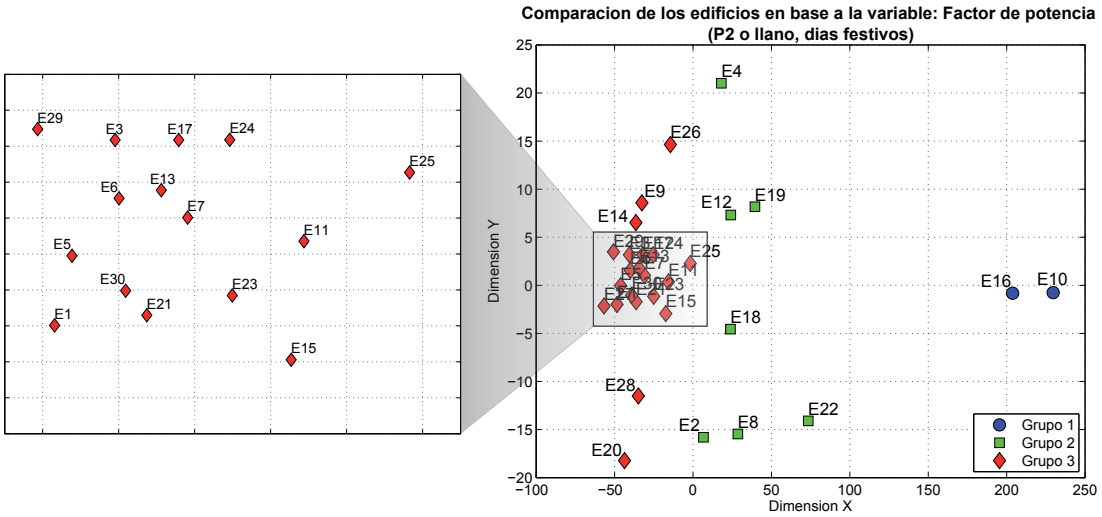


Figura 5.18: Comparación y grupos de edificios en base a la variable factor de potencia condicionada por la hora (18-24 h) y el tipo de día (festivos), correspondientes al periodo de facturación P2 o llano.

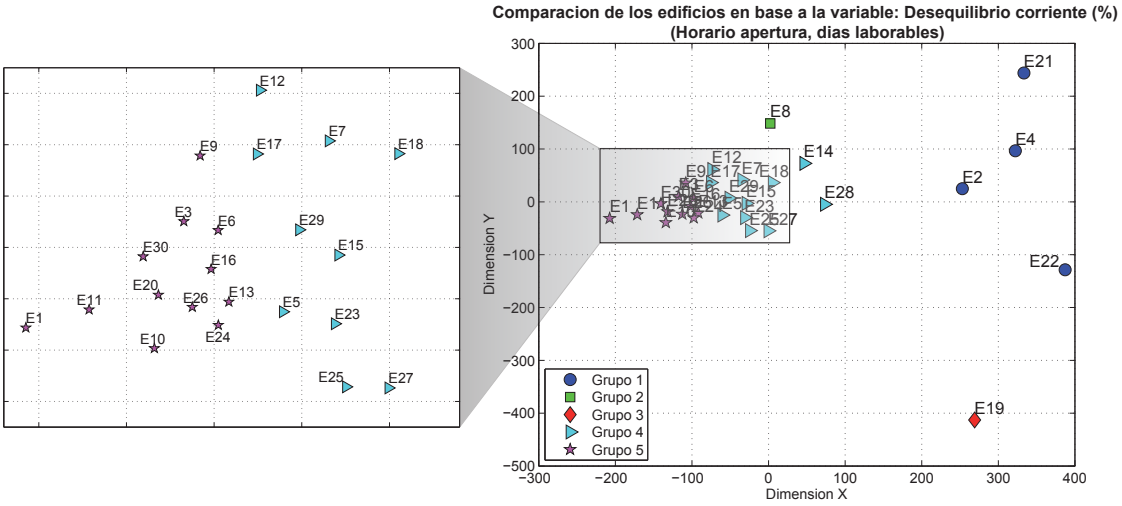
de potencia, desequilibrio en corriente o THD promedio en corriente, se involucren en la comparación. Como ejemplo, en la figura 5.18 se muestra una comparación en base al factor de potencia condicionado por 2 variables: la hora (18-24 horas) y el tipo de día (festivos). Esta condición corresponde a las horas del periodo de facturación P2 o llano durante los días festivos, en los que apenas hay consumo de energía, pero las compañías eléctricas penalizan factores de potencia bajos. Por tanto, en la facturación correspondiente a días festivos puede tener más peso el término de energía reactiva que el de activa. El objetivo de este análisis es detectar en que edificios se producen estas penalizaciones relativas al factor de potencia.

Los grupos 1, o círculos en azul (E10 y E16), y 2, o cuadrados en verde (E2, E4, E8, E12, E18, E19 y E22), así como los edificios E9, E14, E20, E26 y E28 tienen un factor de potencia bajo en esta situación. Por el contrario, el resto de edificios se encuentran muy próximos entre sí y no son penalizados en este periodo.

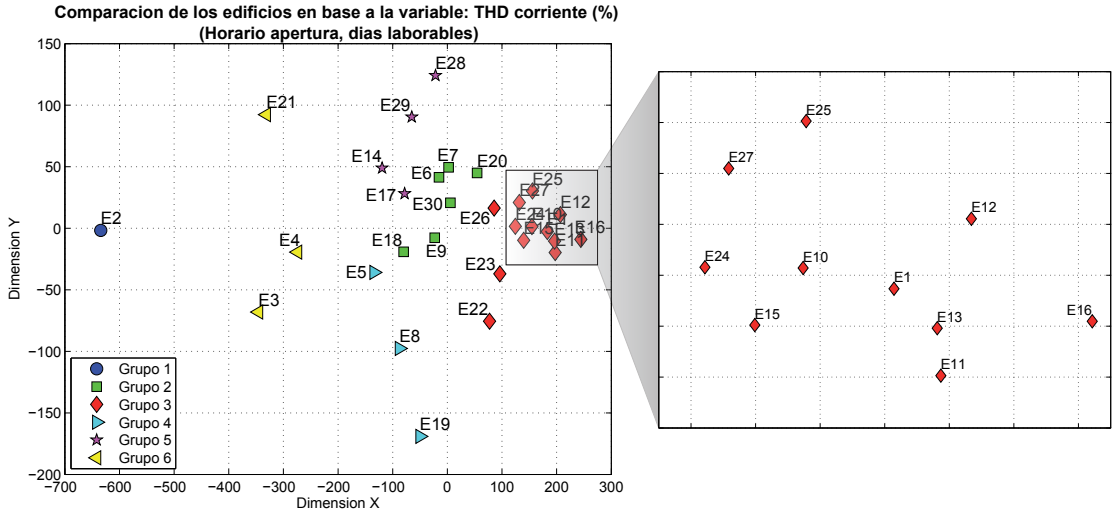
Observando simultáneamente este gráfico de comparación con el individual para la variable factor de potencia (figura 5.10), se puede afirmar que E9 y E14 tienen un factor de potencia bajo en el periodo P2 durante los días festivos, que mejora en otros periodos de tiempo, de forma que se compensa y no son objeto de penalización. En cambio, los edificios E2, E4, E8, E10, E12, E16, E18, E19, E20, E22, E26 y E28 tienen un factor de potencia bajo en todos los periodos de facturación. A pesar de que E25 es objeto de penalización, este edificio tiene un factor de potencia aceptable en el periodo P2 analizado, proyectándose próximo a la mayoría de edificios.

Desequilibrio y THD promedio en corriente condicionados por dos variables: hora y tipo de día. Finalmente, la visualización simultánea de dos gráficos de comparación, dada una condición común, puede ayudar a descubrir y corroborar relaciones entre variables eléctricas. Como ejemplo, en la figura 5.19 se muestra una comparación en base a las variables desequilibrio y THD promedio en corriente, cuya condición común viene definida por la hora (8-21 horas) y el tipo de día (laborables). Esta situación corresponde con los periodos de actividad en los edificios, cuando la demanda de potencia es elevada. El desequilibrio en corriente afecta a los devanados de los transformadores, sobrecargando unas fases más que otras. A su vez, el THD en corriente provoca que la corriente útil necesaria en una instalación sea superior a la nominal. La combinación de ambos efectos puede provocar serios daños en las instalaciones eléctricas.

Los edificios E2, E4, E8, E19, E21 y E28 destacan en los gráficos de comparación correspondientes al desequilibrio en corriente y el THD promedio en corriente, revelando que existe una relación entre ambas variables eléctricas en estos edificios. En cambio, E3 tiene un alto THD promedio en corriente, pero un bajo desequilibrio en corriente, mientras que en E22 ocurre lo contrario. El número de grupos de edificios es similar para ambas variables (5 y 6). Los edificios contenidos en los grupos 1, o círculos en azul (E2, E4, E21 y E22), y 3, o rombos en rojo (E19), tienen un alto desequilibrio en corriente. Por otra parte, los edificios de los grupos 1, o círculos en azul (E2), y 6, o triángulos hacia la izquierda en amarillo (E3, E4 y E21), poseen un elevado THD promedio en corriente.



(a) Desequilibrio en corriente



(b) THD promedio en corriente

Figura 5.19: Comparación y grupos de edificios en base a la variables desequilibrio y THD promedio en corriente condicionadas por la hora (8-21 h) y el tipo de día (laborables), correspondientes a los periodos de actividad.

5.3. Resultados del modelado

5.3.1. Modelos eléctricos de los edificios

En este punto se presentan los 30 modelos del comportamiento eléctrico de los edificios con respecto al tiempo, obtenidos como resultado del experimento de modelado mediante un envSOM de tres fases. Los resultados corresponden a un mes del año tipo (Marzo), en el que existe una cierta variabilidad en las condiciones ambientales y los edificios presentan un nivel de actividad constante sin excesivas interrupciones en el tiempo. Los modelos eléctricos permiten al usuario realizar una supervisión *on-line* de la energía eléctrica en todos los edificios del Campus, utilizando como herramientas de visualización los planos de componentes de cada variable. Además, es posible comparar la evolución actual de las variables eléctricas con la esperada, que ha sido capturada por el modelo. Las diferencias entre los valores actuales y los del modelo, es decir, las desviaciones eléctricas, se deben analizar detenidamente. Estas pueden ser debidas a comportamientos que no se han modelado correctamente, situaciones nuevas y/o anomalías.

En las figuras 5.20, 5.21 y 5.22 se muestran los planos de componentes que definen el modelo de comportamiento eléctrico para cada edificio durante el mes de Marzo. Las componentes eléctricas de los modelos están condicionadas por las variables temporales comunes día y hora. Gracias al envSOM, neuronas situadas en la misma posición en todos los planos corresponden a instantes de tiempo idénticos. Cabe destacar que de nuevo es posible llevar a cabo una comparación y análisis visual entre los planos de las componentes eléctricas de distintos edificios. Sin embargo, esta tarea es costosa y se aleja del objetivo final del modelado: la supervisión *on-line* de cada edificio.

Para ilustrar una comparación basada en los planos de las componentes eléctricas condicionadas por el tiempo durante el mes de Marzo, se toman de nuevo los edificios E1 y E2 (situados en la parte superior de la figura 5.20). En este caso, podemos decir que la tensión promedio de las 3 fases es máxima en horas nocturnas en E1 y durante los fines de semana y noches en E2. La demanda de potencia es alta de Lunes a Viernes entre las 12 y 14 horas en E1, mientras que esta es puntual entorno a las 14 horas y a las 20 horas en E2. En cuanto al factor de potencia, E1 presenta un excelente valor, excepto en horas nocturnas y el Domingo, y por el contrario E2 tiene un bajo factor de potencia, que sólo mejora con el aumento de la potencia. Respecto al desequilibrio en corriente, E1 tiene normalmente valores muy bajos, excepto por las noches. E2 posee valores bajos pero aumentan cuando la demanda de potencia es elevada. La variable THD promedio en tensión es totalmente diferente ya que toma valores nulos en E2 y valores bajos en E1 que aumentan durante el Domingo. Finalmente, la variable THD promedio en corriente presenta grandes diferencias en rango y en su distribución. E1 posee valores bajos que aumentan ligeramente durante las horas nocturnas y E2 tiene valores altos que disminuyen cuando se demanda potencia.

Los modelos eléctricos se han evaluado utilizando un conjunto de datos correspondiente a una semana y los resultados se muestran en la tabla 5.6. La medida de error elegida para evaluar los modelos es el RMSE, ya que el MAPE genera valores infinito en aquellas variables

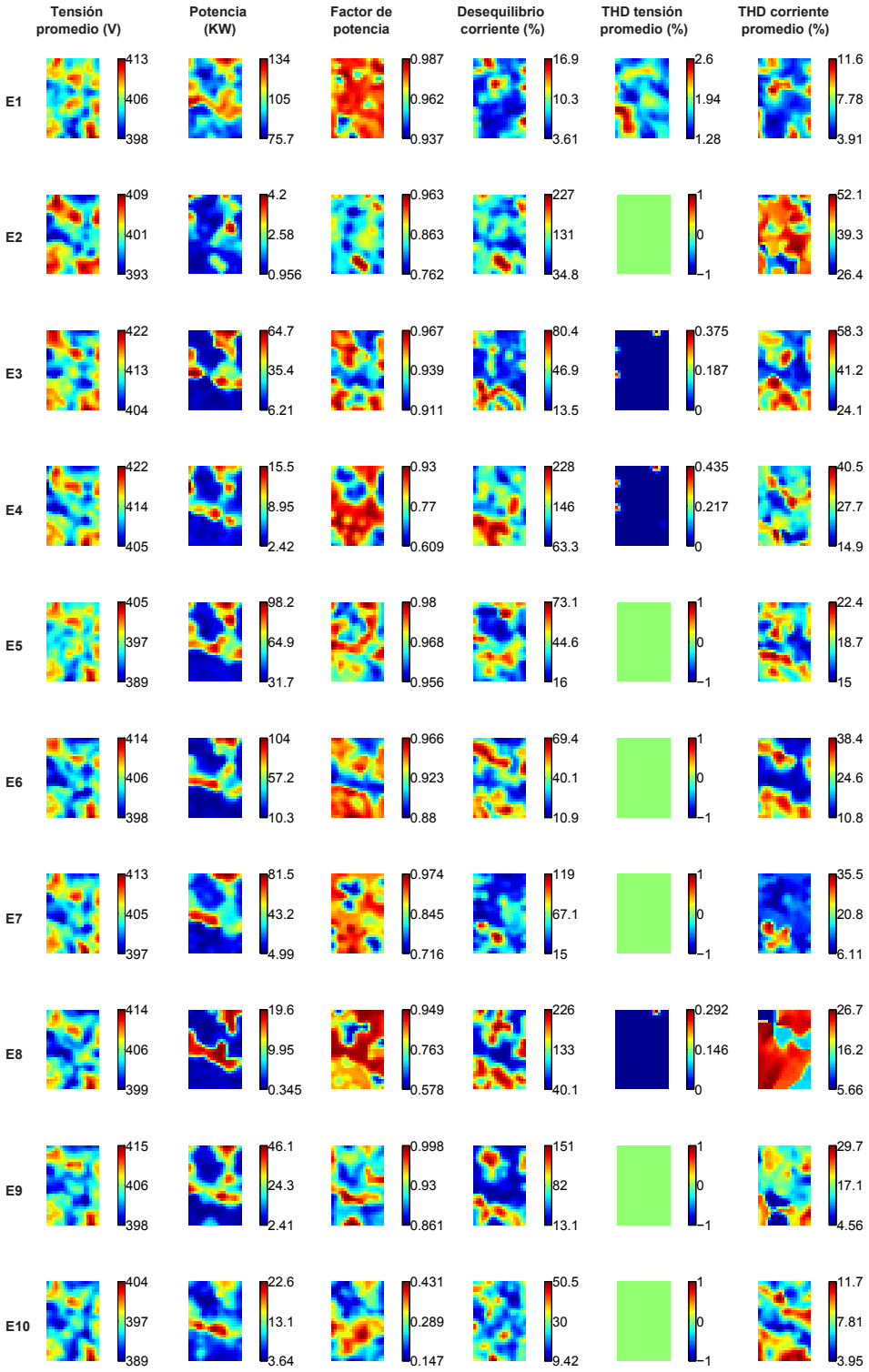


Figura 5.20: Planos de componentes que definen el modelo de comportamiento eléctrico en los edificios E1-E10.

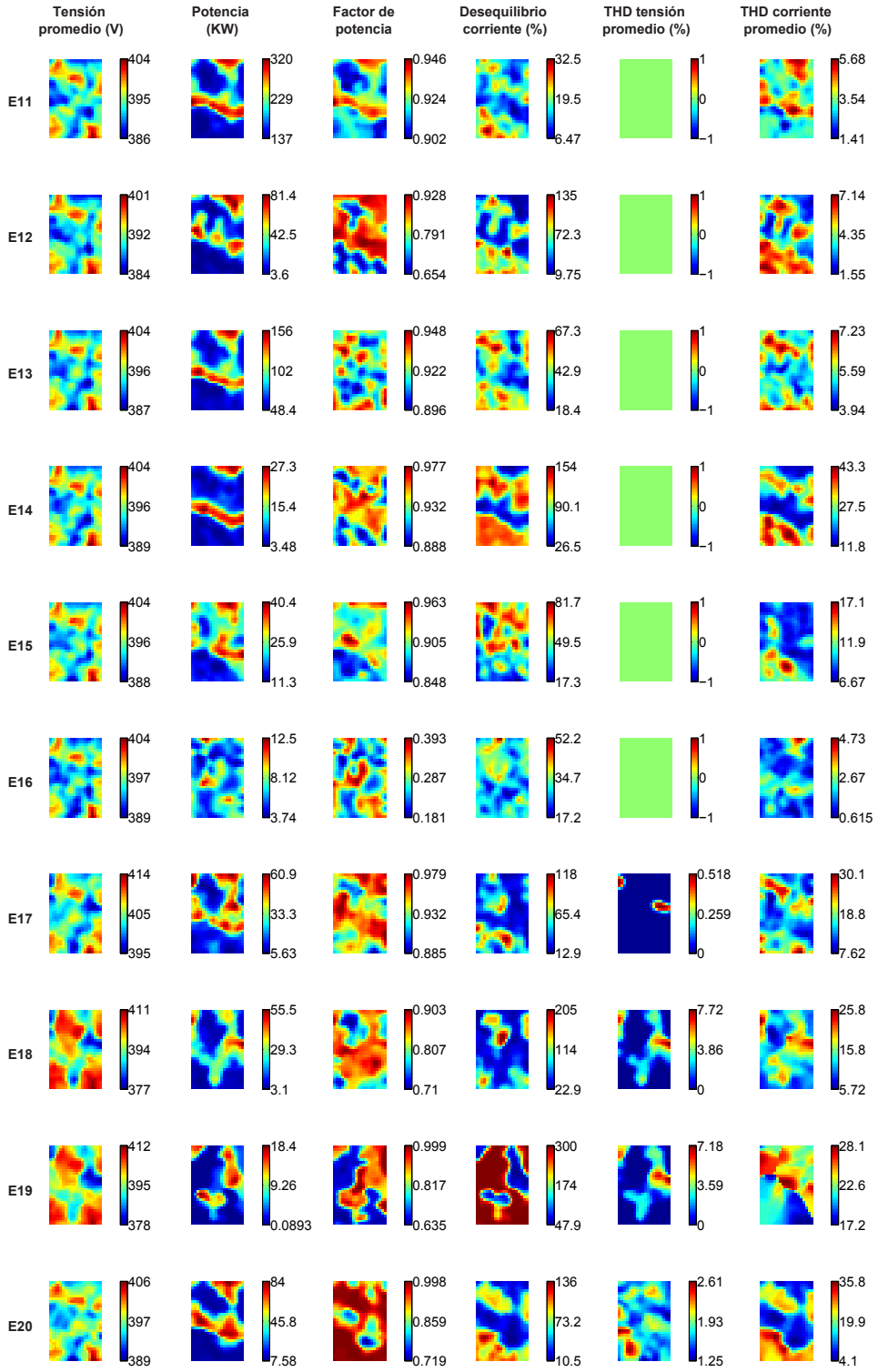


Figura 5.21: Planos de componentes que definen el modelo de comportamiento eléctrico en los edificios E11-E20.

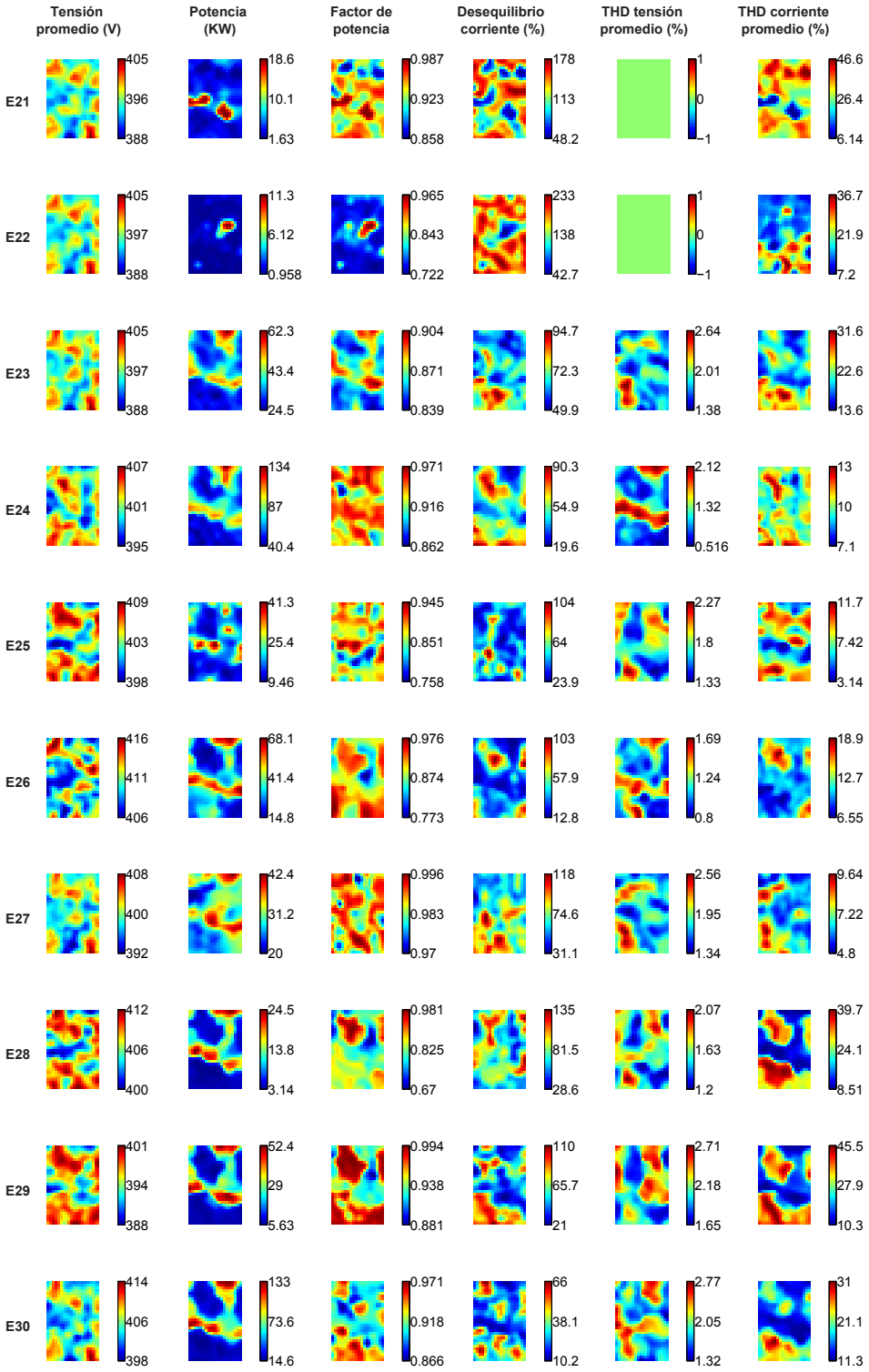


Figura 5.22: Planos de componentes que definen el modelo de comportamiento eléctrico en los edificios E21-E30.

	RMSE (%)						
Edificio	**Tensión promedio** ($\overline{V}$)	**Potencia** (P)	**Factor de potencia** (FP)	**Desequilibrio en corriente** (DC)	**THD promedio en tensión** ($\overline{THDv}$)	**THD promedio en corriente** ($\overline{THDi}$)	**Error medio**
E1	10.77	8.23	11.91	11.11	9.58	9.26	10.14
E2	9.32	7.73	9.76	13.12	0	12.09	8.67
E3	8.97	6.31	12.18	10.77	2.50	9.28	8.33
E4	9.56	5.95	9.88	11.87	2.21	10.30	8.30
E5	9.96	8.54	2.83	15	0	13.75	8.35
E6	9.73	4.53	10.53	11.81	0	9.21	7.63
E7	9.78	5.49	9.15	10.60	0	7.72	7.12
E8	10.18	5.07	10.35	18.50	3.18	9.32	9.43
E9	9.88	5.02	7.79	12.24	0	7.89	7.14
E10	10.13	10.33	8.68	10.21	0	12.99	8.72
E11	9.41	4.97	9.14	11.99	0	8.85	7.39
E12	9.73	8.78	12.09	12.81	0	9.05	8.74
E13	9.57	4.79	8.70	10.71	0	7.13	6.82
E14	10.32	4.65	7.20	11.43	0	8.73	7.06
E15	9.90	8.21	7.52	10.86	0	10.59	7.85
E16	10.35	7.37	12.76	10.37	0	7.93	8.13
E17	8.86	5.44	5.73	10.06	1.87	10.54	7.08
E18	6.48	6.19	8.42	9.35	4.86	8	7.22
E19	7.53	7.50	15.73	10.16	9.29	13.26	10.58
E20	10.17	4.41	5.13	9.36	10.25	6.41	7.62
E21	10.62	7.87	6.57	11.84	0	13.02	8.32
E22	9.85	4.77	9.21	12.43	0	12.55	8.14
E23	10.08	7.12	9.05	9.78	9.03	9.50	9.09
E24	10.04	7.18	8.76	10.74	7.65	9.65	9.01
E25	9.82	6.91	10.11	11.89	7.97	5.06	8.62
E26	10.65	10.54	16.16	9.66	19.48	16.12	13.77
E27	10.19	9.71	9.40	13.88	11.26	11.58	11
E28	10.02	4.40	8.87	13.62	10.29	5.57	8.80
E29	9.87	4.30	9.04	13.10	7.22	8.28	8.63
E30	9.94	4.53	9.34	11.07	9.67	7.92	8.75

Tabla 5.6: Evaluación del grado de aproximación de los modelos eléctricos.

eléctricas que toman valores nulos (THD promedio en tensión). Los errores permiten conocer el grado de ajuste o aproximación de los modelos eléctricos, lo cual es necesario a la hora de realizar estimaciones. El error RMSE máximo es 13.77 % en el edificio E26, mientras que el mínimo es 6.62 % en E13. Estos errores se podrían reducir utilizando en el entrenamiento una mayor cantidad de datos, correspondientes al mismo mes de otros años, que contengan nuevas situaciones eléctricas. En este caso, el número de neuronas empleadas en la construcción de los modelos eléctricos debería ser mayor para mantener su resolución.

5.3.2. Mapas o planos de días y horas

Los modelos eléctricos han sido construidos en base a variables temporales ya que el comportamiento eléctrico de los edificios está muy influenciado por la actividad de personas en su interior, que a su vez depende de las horas del día y los días de la semana para

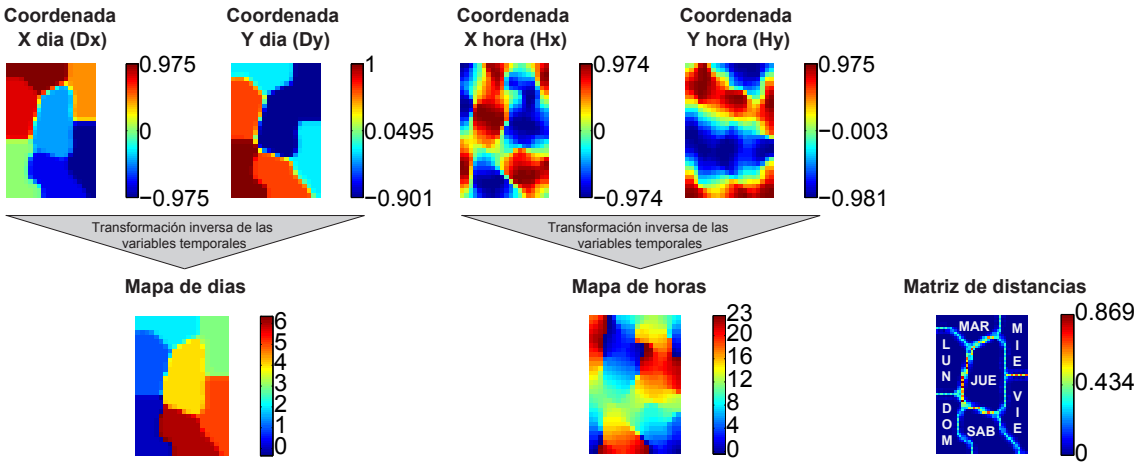


Figura 5.23: Mapas de días, mapa de horas y matriz de distancias, comunes para todos los edificios.

cada mes. Cada una de estas dos variables temporales se descompone en sus respectivas coordenadas X e Y, de acuerdo a las ecuaciones 4.7 y 4.8, evitando discontinuidades en los modelos. Las coordenadas X e Y de las variables temporales comunes condicionan la organización de las componentes eléctricas que definen el modelo eléctrico para cada edificio y además sirven como herramientas de visualización.

En principio, es difícil interpretar adecuadamente los planos de componentes correspondientes a las coordenadas X e Y de la hora y el día, por lo que estos no se utilizan directamente en la supervisión. En cambio, se pueden emplear los mapas o planos generados a partir de las correspondientes coordenadas X e Y de la hora y el día, aplicando la transformación inversa de las variables temporales (ecuaciones 4.18 y 4.19). La información proporcionada por los mapas de horas y días no aporta por si sola una gran ventaja en la supervisión, pero si se combina con la información contenida en los modelos eléctricos y la proyección de la neurona ganadora, puede ser útil para supervisar el estado de actividad en los edificios. Estos mapas se pueden asimilar a mapas de estados de un proceso, suponiendo que los edificios tienen una serie de niveles de actividad a lo largo de un mes, que se corresponden a las horas del día y los días de la semana.

En la figura 5.23 se muestran el mapa de días y el mapa de horas obtenidos a partir de las coordenadas X e Y de las dos variables temporales para el mes de Marzo. En esta figura también se puede observar la matriz de distancias unificada resultante. El mapa de días presenta 7 zonas independientes y claramente definidas, que denotan los días de la semana desde Domingo (0) a Sábado (6). Por otra parte, el mapa de horas presenta 24 zonas dentro de cada día, representado a las horas del día desde las 0 horas hasta las 23 horas. En la matriz de distancias se distinguen perfectamente las 7 zonas correspondientes a los días de la semana, las cuales se han etiquetado convenientemente. La matriz de distancias etiquetada se puede utilizar en la supervisión, aunque en este caso se debería omitir al contener información similar al mapa de días.

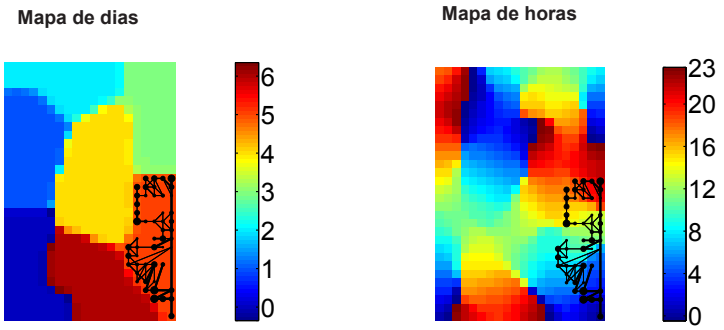
5.3.3. Supervisión de las desviaciones

Los mapas de días y horas se pueden emplear para supervisar las desviaciones del comportamiento eléctrico esperado y detectar periodos de tiempo con un comportamiento eléctrico anómalo. Representando la neurona ganadora o su trayectoria sobre los mapas de días y horas, se puede detectar y visualizar comportamientos extraños no capturados por el modelo, es decir, posibles desviaciones eléctricas. Si la neurona ganadora se proyecta en una zona del mapa de días distinta a aquella que corresponde al día actual, entonces existe una desviación de gran magnitud en una variable eléctrica o bien involucra a muchas variables. Por el contrario, si la neurona ganadora coincide con el día, pero no con la hora actual, entonces la desviación es pequeña en magnitud o sólo afecta a pocas variables.

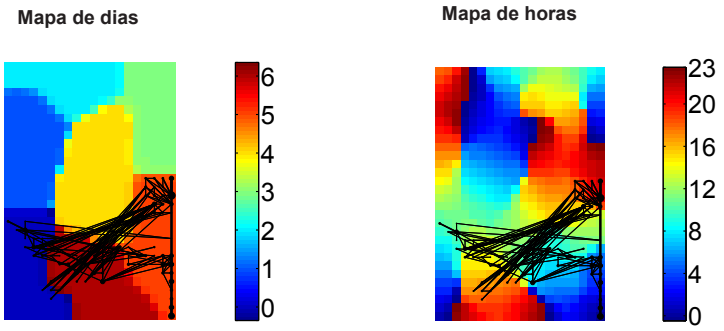
En la figura 5.24 se muestran los mapas de días y horas, junto con la proyección de la trayectoria, como herramientas para la supervisión de desviaciones en el comportamiento eléctrico de los edificios. La figura 5.24a corresponde a una situación normal de funcionamiento en el edificio E7. Se puede observar como la trayectoria descrita por la neurona ganadora a lo largo de un Viernes laborable recorre toda la zona naranja (5 o Viernes) del mapa de días. Las 24 horas correspondientes dentro del mapa de horas también son recorridas uniformemente. Por el contrario, la figura 5.24b muestra una situación de gran desviación en el comportamiento eléctrico provocada por un Viernes festivo no esperado o previsto por el modelo. La trayectoria de la neurona ganadora recorre ocasionalmente la zona naranja (5 o Viernes) en el mapa de días, "saltando" frecuentemente a las zonas roja (6 o Sábado) y azul (0 o Domingo). Esto indica que el comportamiento eléctrico actual es similar a uno de esos días, debido a la escasa actividad en E7. Respecto al mapa de horas, la neurona ganadora se proyecta en las zonas roja y azul (horas nocturnas) dentro del Viernes y en horas cercanas a la actual, pero de otros días (Sábado y Domingo).

Mencionar que la neurona ganadora se calcula utilizando todas las componentes. Ahora bien, cuando se detecta que el día de la neurona ganadora obtenida inicialmente no se corresponde con el actual, se realiza una nueva búsqueda de la neurona ganadora, excluyendo aquellas neuronas que se corresponden con el día actual. Con esto se consigue una mayor precisión en la detección de desviaciones y además es posible conocer el día que mejor se aproxima al comportamiento eléctrico actual.

Para completar la supervisión de las desviaciones en los edificios, se propone visualizar de forma vectorial los residuos con respecto al modelo, es decir, un gráfico de color por cada componente eléctrica a lo largo del tiempo que indique la diferencia entre el dato actual y el valor del modelo que mejor se aproxime él. En la figura 5.25 se muestran las desviaciones a lo largo de una semana, incluyendo un Viernes festivo como perturbación en el comportamiento normal del edificio E7. La desviación se localiza entre la muestras 2880 y 3600, donde el residuo respecto al modelo de la variable potencia es negativo, mientras que el del factor de potencia, desequilibrio y THD promedio en corriente son positivos. En cambio, el residuo para la variable tensión unas veces es positivo y otras negativo, sin un patrón claro. Cabe destacar que continuamente aparecen falsos residuos de pequeña magnitud que se podrían aminorar representado el cuadrado de sus valores o, por supuesto, construyendo un modelo



(a) Normal (Viernes laborable)



(b) Desviación (Viernes festivo)

Figura 5.24: Supervisión de las desviaciones eléctricas utilizando los mapas de días y horas y la trayectoria de la neurona ganadora.

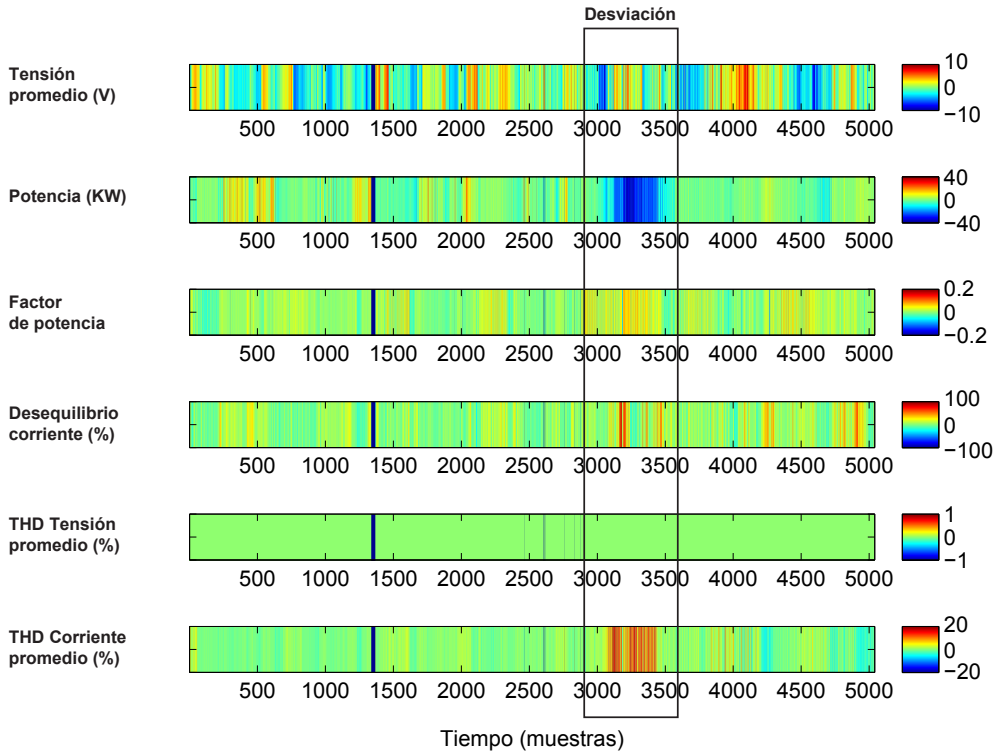


Figura 5.25: Supervisión de las desviaciones eléctricas en el tiempo.

más preciso con datos que contengan una amplia variedad de situaciones.

Estos gráficos de color en el tiempo permiten visualizar las desviaciones y además detectar la variable o variables que las provocan. No obstante, en general las variables eléctricas presentan una relación en sus evoluciones, de forma que si aumenta la potencia, disminuyen la tensión, el factor de potencia y el desequilibrio en corriente y rara vez evolucionan independientemente.

5.3.4. Aplicación para la supervisión avanzada *on-line* de la energía eléctrica

La aplicación desarrollada para la supervisión *on-line* de la energía eléctrica en los edificios de la Universidad de León debe incorporar las herramientas de visualización avanzadas, obtenidas en el experimento de modelado. Estas herramientas son los planos de componentes que definen el modelo eléctrico de cada edificio, los mapas de días y horas comunes en todos los edificios y la proyección de la neurona ganadora o su trayectoria. De forma adicional, se puede añadir un gráfico en el tiempo de los residuos vectoriales respecto al modelo o las evoluciones temporales de las variables, siguiendo los conceptos de la supervisión tradicional.

Los planos de componentes y la trayectoria de la neurona ganadora se pueden emplear para conocer el estado actual o intuir cuál será el siguiente estado esperado en los edificios. Observando los planos de las componentes eléctricas se puede prever cuando se demandará más potencia, si el factor de potencia sobrepasará el límite de penalización, etc. Los mapas de días y horas junto con la trayectoria son útiles para supervisar el nivel de actividad y la magnitud y dirección de las desviaciones eléctricas en los edificios.

En la figura 5.26 se muestran las herramientas de visualización avanzadas utilizadas en la supervisión de la energía eléctrica en los edificios del Campus de la Universidad de León. Los planos de componentes representan el modelo eléctrico para el edificio E7 y la trayectoria sobre ellos corresponde a la evolución de las variables eléctricas durante un Miércoles normal. Se puede observar como la trayectoria recorre toda la zona verde (3 o Miércoles) en el mapa de días y la neurona ganadora no se ha proyectado nunca fuera de esta zona, lo que indica que no se han producido desviaciones importantes durante ese día. En el mapa de horas se comprueba que la neurona ganadora recorre las 24 horas del día uniformemente, indicando la inexistencia de pequeñas desviaciones. El grosor del punto denota el número de veces que se repite la neurona ganadora. Hay que tener en cuenta que es muy probable que el comportamiento del edificio no se vea alterado entre dos o más muestras consecutivas capturadas cada 2 minutos, siendo la neurona ganadora la misma. Se puede prever para cualquier Miércoles de Marzo que la demanda de potencia será máxima entorno a las 12 horas, que la penalización por energía reactiva (bajo factor de potencia) es máxima entorno a las 7 horas, debido probablemente al arranque del sistema de calefacción y que las variables desequilibrio y THD promedio en corriente tomarán valores bajos durante todo el día.

Como resultado final, se ha creado una aplicación software que integra las herramientas

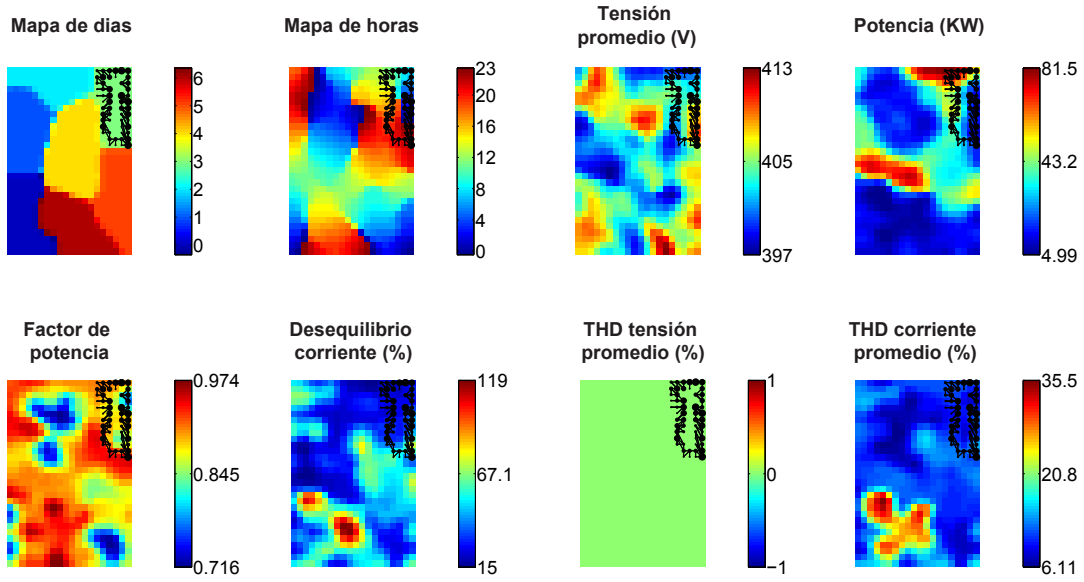


Figura 5.26: Supervisión de la energía eléctrica en el edificio E7 utilizando los planos de componentes eléctricas, los mapas de días y horas y la trayectoria de la neurona ganadora.

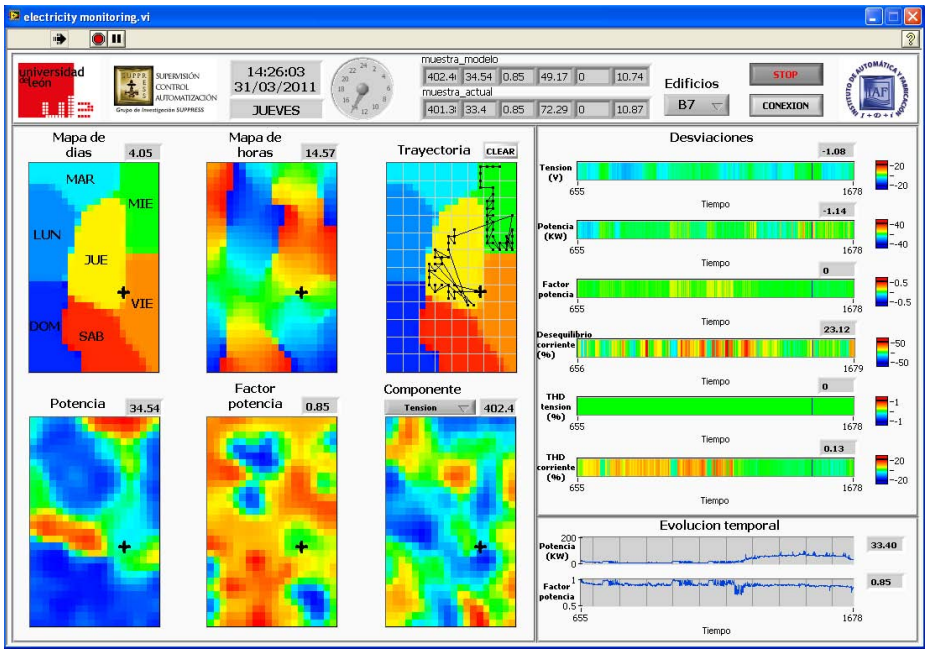


Figura 5.27: Aplicación de supervisión *on-line* basada en las herramientas de visualización avanzadas.

de visualización avanzadas expuestas (planos de las componentes eléctricas, mapas de días y horas y trayectoria de la neurona ganadora), habilitando la supervisión *on-line* de todos los edificios del Campus de la Universidad de León (ver figura 5.27). Esta aplicación ha sido programada en LabVIEW, un lenguaje de programación gráfica que permite desarrollar interfaces de usuario y publicarlos vía web de forma sencilla. La aplicación es accesible de forma remota a través de Internet y ha sido puesta en marcha y validada durante el mes de Marzo de 2011. La ventana de la aplicación está estructurada en 4 zonas bien diferenciadas:

- **Zona de controles**: esta zona está situada en la parte superior de la ventana y en ella se localizan controles como la ejecución de la aplicación, el selector de edificios, la conexión con el edificio correspondiente e indicadores como la fecha, el día y la hora actuales y la muestra actual y la predicha por el modelo.
- **Zona de planos**: esta zona se sitúa en la parte izquierda de la ventana y contiene los mapas de días y horas, la trayectoria de la neurona ganadora sobre un mapa de días, los planos de componentes de las variables eléctricas más importantes (potencia y factor de potencia) y otro plano adicional donde se puede elegir la componente a mostrar. Se han omitido las escalas de color en los planos, pero siempre se indica el valor numérico al lado del nombre del plano correspondiente.
- **Zona de desviaciones**: esta zona se encuentra en la parte superior derecha de la ventana y aquí se puede observar cualquier desviación desglosada por componentes a lo largo del tiempo y codificada en una escala de color. La magnitud de la desviación también se indica mediante valores numéricos.
- **Zona de evoluciones temporales**: en la parte izquierda inferior se mantiene la representación en el tiempo de las variables más importantes (potencia y factor de potencia), utilizada en la supervisión tradicional.

Capítulo 6

Conclusiones y líneas futuras

En este capítulo se exponen las conclusiones obtenidas en el presente trabajo. También se enumeran las aportaciones que han resultado de esta tesis. Las contribuciones se clasifican en científicas y técnicas. Finalmente, se mencionan las posibles líneas futuras de investigación que se desprenden de la tesis.

6.1. Conclusiones

En esta tesis se ha propuesto una metodología basada en técnicas de minería de datos visual para la supervisión de la energía eléctrica en edificios públicos de uso docente, como son aquellos que forman el Campus de la Universidad de León. Primeramente, se ha diseñado y desarrollado un sistema que permite adquirir y almacenar los datos eléctricos procedentes de todos los edificios y las condiciones ambientales que los rodean. A continuación, se ha probado que los métodos de explotación de datos son útiles para llevar a cabo un análisis exploratorio de las variables eléctricas en todos los edificios, las cuales son influenciadas por un entorno ambiental común. El procesamiento conjunto de los datos eléctricos procedentes de todos los edificios y los datos ambientales comunes facilita el análisis comparativo entre los edificios. Esto hace posible la supervisión de la energía eléctrica de forma simultánea y condicionada por el entorno común en todos los edificios. Las herramientas de visualización presentadas ayudan a comprender el comportamiento eléctrico de los edificios de forma cualitativa, a prever el siguiente estado eléctrico y a detectar desviaciones en el caso que la situación actual no se corresponda con la esperada. Se ha demostrado que las aplicaciones de supervisión avanzadas proporcionan de forma innovadora y fácilmente interpretable un mayor conocimiento acerca del comportamiento eléctrico de los edificios, complementando a las aplicaciones de supervisión tradicionales.

Las conclusiones que se pueden extraer de esta tesis son las siguientes:

- Se ha puesto en marcha un sistema de medida y supervisión basado en una estructura de triple capa, que permite la adquisición, almacenamiento y visualización de las variables eléctricas procedentes de todos los edificios de la Universidad de León y de las variables ambientales que los rodean. Su estructura garantiza robustez, modularidad e independencia de las partes, escalabilidad, integración con nuevos

sistemas y flexibilidad. Además, la gran capacidad de almacenamiento de este sistema es vital desde el punto de vista de la explotación de los datos históricos.

- Se ha demostrado la utilidad de los métodos de minería de datos visual en la supervisión y análisis de la energía eléctrica en varios varios edificios, de forma conjunta y condicionada por el ambiente común. Estos métodos permiten reducir la dimensionalidad de los datos sin pérdida significativa de información, lo que facilita la visualización, la extracción de conocimiento y la posterior toma de decisiones. Además, estos métodos permiten tratar de forma conjunta y eficaz datos de varios procesos o sistemas condicionados por un entorno común, como por ejemplo el consumo de energía eléctrica en varios edificios, lo que favorece el estudio de relaciones entre ellos y de relaciones entre variables eléctricas y ambientales.

- La explotación de datos engloba un preprocesamiento, una exploración y un modelado de los datos. La metodología adoptada en el preprocesamiento consiste en un tratamiento de muestras erróneas, una selección de variables y un normalizado de los datos. La metodología de exploración combina el algoritmo envSOM junto con una proyección de Sammon para comparar el comportamiento eléctrico de los edificios. Una agrupación de edificios similares llevada a cabo por medio del método *k-means* completa la exploración. La metodología de modelado propone el uso de un algoritmo envSOM de n fases para construir los modelos eléctricos jerárquicos.

- El algoritmo envSOM, una modificación del SOM básico propuesta en esta tesis, ha demostrado ser útil en el descubrimiento y visualización de perfiles de comportamiento de varios procesos o sistemas que están condicionados por variables ambientales comunes, como es el caso del consumo de energía eléctrica en un conjunto de edificios cercanos.

- Los mapas de variables temporales definidos a partir de las coordenadas X e Y han demostrado ser herramientas de visualización útiles en la supervisión de la energía eléctrica en edificios públicos. En este sentido, los mapas de días y horas se han utilizado para visualizar el estado eléctrico actual y detectar desviaciones en el comportamiento eléctrico de los edificios.

- Los resultados han permitido descubrir o confirmar conocimiento acerca de la distribución del consumo de energía eléctrica en los edificios de la Universidad de León. Gran parte de la energía se consume en los edificios que poseen una gran superficie, numeroso equipamiento eléctrico dedicado a investigación o una elevada actividad de personas en su interior. No hay que olvidar pequeños edificios, como las cafeterías o la sucursal bancaria, que tienen un consumo elevado por unidad de superficie. Por otra parte, los edificios que poseen sistemas de refrigeración y aire acondicionado consumen mucha energía durante los meses de verano. Los edificios donde se practica deporte consumen más energía durante las últimas horas de la tarde y los fines de semana (festivos).

- Los resultados obtenidos revelan que existen 6 perfiles diferentes de comportamiento eléctrico en los edificios de la Universidad de León. Considerando la potencia, el número óptimo de perfiles es 5. Teniendo en cuenta el factor de potencia, hay 6 grupos. Se puede concluir que edificios con un mismo perfil eléctrico se deberían agrupar, si es posible, dentro del mismo punto de facturación con el fin de aprovechar los comportamientos similares. Además, se debería aplicar una tarifa eléctrica idéntica a aquellos edificios con un mismo perfil.

- Los resultados también han facilitado la detección de fallos y errores en los medidores eléctricos. El tratamiento de un número elevado de muestras anómalas en los armónicos y de valores anormales en el factor de potencia puso de manifiesto la existencia de errores en la medida eléctrica.

6.2. Aportaciones

Las contribuciones de esta tesis se pueden clasificar en dos grupos: las aportaciones en el campo de la ciencia e investigación y las aportaciones técnicas en el campo de la ingeniería de control, supervisión y automatización de procesos.

Aportaciones científicas. Como aportaciones al mundo de la ciencia e investigación podemos citar las siguientes:

- Se ha comprobado como los métodos de minería de datos basados en SOM ofrecen un enorme potencial en el análisis del consumo de energía eléctrica en edificios públicos. Las herramientas de visualización avanzadas permiten entender mejor el consumo eléctrico y verificar la facturación. Este trabajo ha sido presentado en *18th World Congress of the International Federation of Automatic Control*, IFAC 2011.

- Se ha desarrollado un nuevo algoritmo, el envSOM, una variante del SOM tradicional que permite la búsqueda de patrones condicionados por el entorno ambiental en grandes volúmenes de datos. El algoritmo envSOM es adecuado para el análisis y comparación visual de datos procedentes de varios procesos o sistemas que están influenciados por un entorno común. Este trabajo ha sido presentado en *8th Workshop on Self-Organizing Maps*, WSOM 2011.

- Se ha propuesto una metodología que permite realizar un análisis exploratorio y comparativo de la energía eléctrica en un conjunto de edificios. El algoritmo envSOM facilita el estudio conjunto de todos los edificios y las técnicas de proyección como Sammon habilitan el análisis y comparación visual. Este trabajo ha sido presentado en *10th International Symposium on Intelligent Data Analysis*, IDA 2011.

- Se ha propuesto una metodología para la supervisión en línea basada en modelos del consumo de energía eléctrica en los edificios. Un algoritmo envSOM de 3 fases se emplea para construir los modelos eléctricos. Los mapas de días y horas, junto con los planos de

las componentes eléctricas se utilizan como herramientas avanzadas de visualización. Este trabajo ha sido presentado en *5th International Scientific Conference on Physics and Control*, PHYSCON 2011.

- Se ha definido un sistema de medida y supervisión de la energía eléctrica para todos los edificios de la Universidad de León. Dicho sistema se basa en una arquitectura de red compleja de triple capa. Este trabajo ha sido presentado en *XXXII Jornadas de Automática*, JA 2011.

Aportaciones técnicas. Las principales aportaciones técnicas a nivel de ingeniería para la Universidad de León son las siguientes:

- Se ha desarrollado una plataforma tecnológica para la medida y supervisión de la energía eléctrica en todos los edificios de la Universidad de León. Esta plataforma permite medir las variables eléctricas, almacenar los datos históricos y visualizar de forma remota las evoluciones temporales de las variables. Además, incorpora un módulo de explotación de los datos históricos.

- Se ha verificado el funcionamiento de todos los medidores eléctricos ya instalados. En este sentido, se han detectado 5 equipos de medida eléctrica dañados que se han tenido que sustituir. Concretamente, 4 medidores modelo Nexus 1252 presentaban el mismo problema, un fallo en el puerto de comunicación serie RS-485. Esto está siendo analizado con la ayuda del fabricante.

- Se ha definido la estructura de medida e instalado medidores eléctricos en aquellos edificios que carecían de ellos, con el fin de ser integrados en el sistema de medida y supervisión. Se ha optado siempre por la colocación de medidores independientes en cada edificio, lo que permite disponer de todas las variables eléctricas y no sólo las potencias y energías calculadas por diferencia.

- Se ha detectado una conexión errónea en los transformadores de medida de corriente de 2 medidores, lo que provocaba valores en la medida extraños. Este problema se ha solucionado fácilmente intercambiado las conexiones de los transformadores de medida de corriente.

- Se ha detectado un numero elevado de muestras erróneas en las componentes armónicas medidas por varios medidores modelo Shark 100. Actualmente se está estudiando la causa de este fallo.

- Se han desarrollado aplicaciones que permiten a los usuarios supervisar la energía eléctrica en todos los edificios desde cualquier parte a través de Internet. Esto aporta ventajas al personal de mantenimiento y gestión económica de la Universidad de León, ya que ahora es rápido y sencillo consultar todos los consumos de energía eléctrica de los edificios para posteriormente imputarlos a las distintas instituciones y empresas externas.

- Se ha adquirido conocimiento acerca de los perfiles eléctricos de los edificios, el cual se usa en el proceso de negociación de las tarifas y contratos con las compañías eléctricas.

6.3. Líneas futuras

Esta tesis es el inicio de muchas y variadas líneas futuras de trabajo que pueden generar tanto aportaciones científicas como grandes beneficios de tipo energético y económico para la Universidad de León. En este trabajo quedan abiertas las siguientes líneas de investigación:

- Plantear nuevas modificaciones e innovaciones del SOM que permitan resolver futuros problemas que surjan en la supervisión de la energía eléctrica. Por ejemplo, desde el punto de vista de la facturación de energía eléctrica, sería interesante un algoritmo que procese información de la tarifa eléctrica (periodos estacionales, diarios y horarios, precios, límites de penalizaciones, etc.), junto con las variables eléctricas y permita conocer la tarifa óptima durante un determinado periodo.

- Emplear otros métodos de minería de datos para analizar y supervisar el consumo de energía eléctrica. En este sentido, se podría profundizar en el uso de los métodos de *manifold learning*, los cuales realizan de forma novedosa la reducción de la dimensionalidad y visualización de los datos. Además, la aplicación de este tipo de métodos en el campo de la energía eléctrica es una línea poco explotada y podría arrojar resultados muy interesantes y prometedores.

- Construir modelos eléctricos anuales con una mayor resolución. Para ello, se podría emplear un envSOM de 4 fases, incorporando el mes del año. Esto implica la definición del mapa de meses que se podría utilizar en la supervisión. Para aumentar la precisión del modelado, se deberían tener en cuenta variables adicionales como las meteorológicas, la actividad en los edificios, etc. Estos modelos se podrían emplear en la predicción de futuros comportamientos eléctricos de los edificios.

- Desarrollar un módulo de predicción que permita prever el consumo de energía eléctrica y su coste asociado en determinadas condiciones e instantes de tiempo. Esto sería de gran utilidad para la Universidad de León ya que se podría estimar por adelantado el gasto energético y económico o incluso posibles penalizaciones correspondientes a situaciones concretas y actividades extraordinarias. Además, una predicción a largo plazo sería interesante para negociar el contrato del siguiente año. Este módulo podría incorporar varios métodos de predicción del consumo eléctrico basados en series temporales como por ejemplo *Support Vector Regression* (SVR), *Radial Basis Function* (RBF), *Autoregressive Integrated Moving Average* (ARIMA), series de Fourier, SOM, etc. o basados en métodos híbridos que combinen el envSOM con alguna de las anteriores técnicas.

- Proponer nuevas herramientas avanzadas de visualización que complementen a las existentes. Anteriormente se mencionaron los mapas de meses. Por otra parte, la

generación de mapas de diferencias entre edificios podría ser útil para profundizar en la comparación. En esta misma línea, se podrían definir mapas aditivos entre edificios. Dado que un punto de facturación agrupa varios edificios, este se podría analizar en base a un mapa obtenido como adición, por ejemplo de los planos de componentes de los edificios correspondientes. No obstante, desde el punto de vista eléctrico, esta idea solamente sería válida para corrientes, potencias y energías.

- Realizar un análisis específico de las componentes armónicas. La descomposición espectral de tensiones y principalmente corrientes puede proporcionar interesantes resultados desde el punto de vista de la detección de fallos. Además, puede ayudar a conocer mejor el tipo de receptores eléctricos en cada edificio, en función del armónico predominante.

- Incorporar el resto de edificios de la Universidad de León en el sistema de medida y supervisión, principalmente los edificios localizados en el Campus de Ponferrada. Cualquier nuevo edificio proyectado debería tener como requisito técnico la colocación de un equipo de medida eléctrica compatible con el sistema de actual.

- Mejorar el sistema de medida y supervisión actual. Las futuras ampliaciones del sistema requerirán una gestión más eficiente de los clientes o usuarios de supervisión así como de las aplicaciones de supervisión. Además, la integración en el sistema de medida y supervisión de otros consumos energéticos como por ejemplo el gas y el agua, permitirá realizar una supervisión del gasto energético integral y más exhaustiva.

- Establecer un plan de actuación que aproveche el conocimiento extraído para optimizar el consumo eléctrico. Las líneas maestras de este plan se deben dirigir hacia la reorganización del sistema de suministro de electricidad, supresión de puntos de facturación o conexión a la red eléctrica, eliminación de transformadores con pequeñas cargas, control de los sistemas de refrigeración, aire acondicionado e iluminación de zonas de tránsito comunes y elaboración de informes anuales para la negociación con las compañías eléctricas.

Bibliografía

Alhoniemi, E., J. Hollmén, O. Simula y J. Vesanto. "Process monitoring and modeling using the self-organizing map". *Integrated Computer-Aided Engineering*, 6(1), 3–14, 1999.

Alonso, S., M. Domínguez, M. A. Prada, M. Sulkava y J. Hollmén, "Comparative analysis of power consumption in university buildings using envSOM". En J. Gama, E. Bradley y J. Hollmén (editores), *Advances in Intelligent Data Analysis X*, tomo 7014 de *Lecture Notes in Computer Science*, 10–21, Springer Berlin / Heidelberg, 2011a. ISBN 978-3-642-24799-6.

Alonso, S., A. Morán, M. A. Prada, P. Barrientos y M. Domínguez, "A variant of self-organizing map (SOM) for monitoring power consumption". En *5th International Scientific Conference on Physics and Control, PHYSCON2011*, León, España, 2011b.

Alonso, S., M. Sulkava, M. A. Prada, M. Domínguez y J. Hollmén, "EnvSOM: a SOM algorithm conditioned on the environment for clustering and visualization". En J. Laaksonen y T. Honkela (editores), *Advances in Self-Organizing Maps*, tomo 6731 de *Lecture Notes in Computer Science*, 61–70, Springer Berlin / Heidelberg, 2011c. ISBN 978-3-642-21565-0.

Alpaydin, E. *Introduction to machine learning.* MIT Press, Cambridge, MA, 2004. ISBN 0-262-01211-1.

Barreto, G. A. y A. F. R. Araújo. "Time in self-organizing maps: An overview of models". *International Journal of Computer Research*, 10(2), 139–179, 2001.

Barreto, G. A., J. C. M. Mota, L. G. M. Souza y R. A. Frota, "Nonstationary time series prediction using local models based on competitive neural networks". En B. Orchard, C. Yang y M. Ali (editores), *Innovations in Applied Artificial Intelligence*, tomo 3029 de *Lecture Notes in Computer Science*, 1146–1155, Springer Berlin / Heidelberg, 2004. ISBN 978-3-540-22007-7.

Bayindir, R., E. Irmak, I. Colak y A. Bektas. "Development of a real time energy monitoring platform". *International Journal of Electrical Power & Energy Systems*, 33(1), 137–146, 2011.

Beccali, M., M. Cellura, V. Lo Brano y A. Marvuglia. "Forecasting daily urban electric load

profiles using artificial neural networks". *Energy Conversion and Management*, 45(18-19), 2879–2900, 2004.

Belkin, M. y P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering". En *Advances in Neural Information Processing Systems*, tomo 14, 585–591, MIT Press, 2001.

Bellman, R. *Adaptive control processes: a guided tour*. Princeton University Press, 1961. ISBN 978-0-691-07901-1.

Bertoldi, P. y B. Atanasiu, "Electricity consumption and efficiency trends in European Union". Informe Técnico EUR 24005 EN, European Commission. Joint Research Centre. Institute for Energy, 2009.

Bezdek, J. y N. Pal. "Some new indexes of cluster validity". *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 28(3), 301–315, 1998.

Bishop, C. M. *Pattern recognition and machine learning*, 1ª edición. Information science and statistics. Springer New York, 2006. ISBN 978–0–387–31073–2.

Bishop, C. M., M. Svensen y C. K. I. Williams. "GTM: The generative topographic mapping". *Neural Computation*, 10(1), 215–234, 1998.

Card, S. K., J. D. Mackinlay y B. Shneiderman. *Readings in information visualization: Using vision to think*. Morgan Kaufmann Publishers, 1999. ISBN 1-55860-533-9.

Carpinteiro, O. A., A. J. Reis y A. P. da Silva. "A hierarchical neural model in short-term load forecasting". *Applied Soft Computing*, 4(4), 405–412, 2004.

Carreira-Perpiñán, M. A., *Continuous latent variable models for dimensionality reduction and sequential data reconstruction*. Tesis Doctoral. Department of Computer Science, University of Sheffield, UK, 2001.

Carreira-Perpiñán, M. A., "A review of dimension reduction techniques". Informe Técnico CS–96–09, Deparment of Computer Science, University of Sheffield, 1996.

Cayton, L., "Algorithms for manifold learning". Informe Técnico CS2008-0923, University of California, San Diego (UCSD), 2005.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth, *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS, 2000.

Chen, J., S.-J. Deng y X. Huo. "Electricity price curve modeling and forecasting by manifold learning". *IEEE Transactions on Power Systems*, 23(3), 877–888, 2008.

Cherkassky, V. y F. Mulier. *Learning from data – Concepts, theory and methods*, 2ª edición. John Wiley & Sons, New York, 2007. ISBN 978-0-471-68182-3.

Chicco, G., R. Napoli y F. Piglione. "Comparisons among clustering techniques for electricity customer classification". *IEEE Transactions on Power Systems*, 21(2), 933–940, 2006.

Chicco, G., R. Napoli, F. Piglione, P. Postolache, M. Scutariu y C. Toader. "Load pattern-based classification of electricity customers". *IEEE Transactions on Power Systems*, 19(2), 1232–1239, 2004.

Chwieduk, D. "Towards sustainable-energy buildings". *Applied Energy*, 76(1-3), 211–217, 2003.

Claussen, J. C. "Winner-relaxing self-organizing maps". *Neural Computation*, 17(5), 996–1009, 2005.

Clifton, C. y B. Thuraisingham. "Emerging standards for data mining". *Computer Standards & Interfaces*, 23(3), 187–193, 2001.

Comon, P. "Independent component analysis - a new concept?." *Signal Processing*, 36, 287–314, 1994.

Cottrell, M., J.-C. Fort y G. Pagès. "Theoretical aspects of the SOM algorithm." *Neurocomputing*, 21(1-3), 119–138, 1998.

Cristaldi, L., A. Ferrero y S. Salicone. "A distributed system for electric power quality measurement". *IEEE Transactions on Instrumentation and Measurement*, 51(4), 776–781, 2002.

Cuadrado, A. A., *Supervisión de procesos complejos mediante técnicas de data mining con incorporación de conocimiento previo.* Tesis Doctoral. Universidad de Oviedo, Oviedo, España, 2002.

Dahhaghchi, I., R. Christie, G. Rosenwald y C.-C. Liu. "AI application areas in power systems". *IEEE Expert*, 12(1), 58–66, 1997.

Date, C. *An introduction to database systems*, 8ª edición. Addison-Wesley, 2003. ISBN 978-0-201-38590-8.

Davies, D. L. y D. W. Bouldin. "A cluster separation measure". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224–227, 1979.

De Silva, V. y J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction". En *Advances in Neural Information Processing Systems*, tomo 15, 705–712, MIT Press, 2003.

Demartines, P. y J. Hérault. "Curvilinear component analysis: a self organizing neural network for non linear mapping of data sets". *IEEE Transactions on Neural Networks*, 8, 148–154, 1997.

Dias, R. A., C. R. Mattos y J. A. P. Balestieri. "Energy education: breaking up the rational energy use barriers". *Energy Policy*, 32(11), 1339–1347, 2004.

Díaz, I., *Detección e identificación de fallos en procesos industriales mediante técnicas de procesamiento digital de señal y redes neuronales. Aplicación al mantenimiento predictivo de accionamientos eléctricos.* Tesis Doctoral. Universidad de Oviedo, Oviedo, España, 2000.

Díaz, I., A. A. Cuadrado y A. B. Díez, "Correlation visualization of high dimensional data using topographic maps." En J. Dorronsoro (editor), *Artificial Neural Networks–ICANN 2002*, tomo 2415 de *Lecture Notes in Computer Science*, 1005–1012, Springer Berlin / Heidelberg, 2002a. ISBN 978-3-540-44074-1.

Díaz, I., A. A. Cuadrado, A. B. Díez y G. Ojea. "Modelado visual de procesos industriales." *Revista Iberoamericana de Automática e Informática Industrial*, 2(4), 101–112, 2005.

Díaz, I., A. B. Díez y A. A. Cuadrado, "Complex process visualization through continuous feature maps using radial basis functions." En G. Dorffner, H. Bischof y K. Hornik (editores), *Artificial Neural Networks–ICANN 2001*, tomo 2130 de *Lecture Notes in Computer Science*, 443–449, Springer Berlin / Heidelberg, 2001. ISBN 978-3-540-42486-4.

Díaz, I., A. B. Díez, A. A. Cuadrado y M. Domínguez, "Prior knowledge integration in self-organizing maps for complex process supervision". En *15th World Congress of the International Federation of Automatic Control, IFAC2002*, Barcelona, España, 2002b.

Díaz, I. y J. Hollmén, "Residual generation and visualization for understanding novel process conditions." En *International Joint Conference on Neural Networks, IJCNN2002*, tomo 3, 2070–2075, Piscataway, USA, 2002.

Dittenbach, M., D. Merkl y A. Rauber, "The growing hierarchical self-organizing map". En *International Joint Conference on Neural Networks, IJCNN2000*, tomo 6, 15–19, Piscataway, USA, 2000.

Domínguez, M., *Supervisión remota de procesos complejos vía Internet mediante técnicas de data mining visual. Aplicación a una planta piloto industrial.* Tesis Doctoral. Universidad de Oviedo, Oviedo, España, 2003.

Domínguez, M., J. Fuertes, I. Díaz, A. Cuadrado, S. Alonso y A. Morán, "Analysis of electric power consumption using self-organizing maps." En *18th World Congress of the International Federation of Automatic Control, IFAC2011*, Milán, Italia, 2011.

Domínguez, M., P. Reguera y J. J. Fuertes. "Laboratorio remoto para la enseñanza de la automática en la Universidad de León (España)". *Revista Iberoamericana de Automática e Informática Industrial*, 2(2), 36–45, 2005.

Domínguez, M., P. Reguera, J. J. Fuertes, I. Díaz y A. A. Cuadrado. "Internet-based remote supervision of industrial processes using self-organizing maps." *Engineering Applications of Artificial Intelligence*, 20(6), 757–765, 2007.

Donoho, D. L. y C. Grimes. "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data". *Proceedings of the National Academy of Sciences, PNAS*, 100(10), 5591–5596, 2003.

Dubrawsky, I., C. T. Baumrucker, J. Caesar, M. Krishnamurthy, T. W. Shinder, B. Pinkard, E. Seagren *et al. Designing and building enterprise DMZs*. Elsevier Inc., 2006. ISBN 978-1-59749-100-6.

Eckerson, W. W. "Three tier client/server architectures: Achieving scalability, performance, and efficiency in client/server applications". *Open Information Systems*, 3(20), 46–50, 1995.

Electro Industries/GaugeTech, *Nexus 1250/1252 meter. Installation & operation manual*, January 2010a.

Electro Industries/GaugeTech, *Shark 100 & 100T meter. Installation & operation manual*, March 2010b.

Electronic Industries Alliance, EIA, *Electrical characteristics of generators and receivers for use in balanced multipoint systems, TIA/EIA 485-A*, 2003.

Ente Regional de la Energía de Castilla y León, EREN, *Manual de procedimiento para la realización de auditorías energéticas en edificios*, 2009.

Erwin, E., K. Obermayer y K. Schulten. "Self-organizing maps: Order, convergence properties and energy functions". *Biological Cybernetics*, 67, 47–55, 1992.

España. Ministerio de Industria, Turismo y Comercio, *La energía en España 2009*, 2010.

España, *LEY 54/1997, de 27 de Noviembre, del Sector Eléctrico.*, 1997.

España, *REAL DECRETO-LEY 6/2000, de 23 de Junio, de Medidas Urgentes de Intensificación de la Competencia en Mercados de Bienes y Servicios.*, 2000.

España, *REAL DECRETO 314/2006, de 17 de marzo, por el que se aprueba el Código Técnico de la Edificación.*, 2006.

España, *REAL DECRETO 47/2007, de 19 de enero, por el que se aprueba el Procedimiento básico para la certificación de eficiencia energética de edificios de nueva construcción.*, 2007.

España, *REAL DECRETO 485/2009, de 3 de Abril, por el que se regula la puesta en marcha del suministro de último recurso en el sector de la energía eléctrico.*, 2009.

Fan, S., L. Chen y W.-J. Lee. "Short-term load forecasting using comprehensive combination based on multimeteorological information". *IEEE Transactions on Industry Applications*, 45(4), 1460–1466, 2009.

Fan, S., C. Mao y L. Chen. "Electricity peak load forecasting with self-organizing map and support vector regression". *IEEJ Transactions on Electrical and Electronic Engineering*, 1(3), 330–336, 2006.

Fayyad, U., G. Grinstein y A. Wierse. *Information visualization in data mining and knowledge discovery.* Morgan Kaufmann, 2002. ISBN 978-1558606890.

Fayyad, U., G. Piatetsky-Shapiro y P. Smyth. "The KDD process for extracting useful knowledge from volumes of data". *Communications of the ACM*, 39(11), 27–34, 1996a.

Fayyad, U. M., G. Piatetsky-Shapiro y P. Smyth, "From data mining to knowledge discovery: An overview." En U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth y R. Uthurusamy (editores), *Advances in Knowledge Discovery and Data Mining*, 1–34, AAAI Press/MIT Press, Menlo Park, 1996b.

Fayyad, U. M., G. Piatetsky-Shapiro y P. Smyth. "From data mining to knowledge discovery in databases". *AI Magazine*, 17(3), 37–54, 1996c.

Ferreira de Oliveira, M. C. y H. Levkowitz. "From visual data exploration to visual data mining: A survey". *IEEE Transactions on Visualization and Computer Graphics*, 9(3), 378–394, 2003.

Figueiredo, V., F. Rodrigues, Z. Vale y J. Gouveia. "An electric energy consumer characterization framework based on data mining techniques". *IEEE Transactions on Power Systems*, 20(2), 596–602, 2005.

Flexer, A. "On the use of self-organizing maps for clustering and visualization". *Intelligent Data Analysis*, 5(5), 373–384, 2001.

Forth, B. y T. Tobin. "Right power, right price". *IEEE Computer Applications in Power*, 15(2), 22–27, 2002.

Freeman, J. y D. Passmore, "The virtual LAN technology". Informe Técnico 200374-001, Decisys, Inc., Sterling, VA, 1996.

Friedman, J. H. y J. W. Tukey. "A projection pursuit algorithm for exploratory data analysis." *IEEE Transactions on Computers*, 23(9), 881–890, 1974.

Fuertes, J. J., *Modelado y representación de la dinámica de sistemas complejos mediante técnicas de data mining visual para la supervisión remota de procesos industriales vía Internet.* Tesis Doctoral. Universidad de Oviedo, Oviedo, España, 2006.

Fuertes, J. J., M. Domínguez, S. Alonso, M. A. Prada, A. Morán y P. Barrientos, "Sistema de supervisión de energía eléctrica en la Universidad de León". En *XXXII Jornadas de Automática, JA2011*, Sevilla, España, 2011.

Fuertes, J. J., M. A. Prada, M. Domínguez, P. Reguera, I. Díaz y A. A. Cuadrado, "Modeling of dynamics using process state projection on the self-organizing map". En J. de Sá,

L. Alexandre, W. Duch y D. Mandic (editores), *Artificial Neural Networks–ICANN 2007*, tomo 4668 de *Lecture Notes in Computer Science*, 589–598, Springer Berlin / Heidelberg, 2007. ISBN 978-3-540-74689-8.

Fuertes, J. J., P. Reguera, M. Domínguez, I. Díaz y A. A. Cuadrado, "Industrial supervision system based on visual data mining and motion trajectory analysis." En *16th World Congress of the International Federation of Automatic Control, IFAC2005*, Prague, Czech Republic, 2005.

González, P. y J. Zamarreño. "Prediction of hourly energy consumption in buildings based on a feedback artificial neural network". *Energy and Buildings*, 37, 595–601, 2005.

Gordon, A. "A review of hierarchical classification". *Journal of the Royal Statistical Society (A)*, 150(2), 119–137, 1987.

Graepel, T., M. Burger y K. Obermayer. "Self-organizing maps: generalizations and new optimization techniques". *Neurocomputing*, 21(1-3), 173–190, 1998.

Gray, R. M. y D. L. Neuhoff. "Quantization". *IEEE Transactions on Information Theory*, 44(6), 1–63, 1998.

Guimarães, G., V. Sousa-Lobo y F. Moura-Pires. "A taxonomy of self-organizing maps for temporal sequence processing." *Intelligent Data Analysis*, 7(4), 269–290, 2003.

Hagenbuchner, M. y A. C. Tsoi. "A supervised training algorithm for self-organizing maps for structures." *Pattern Recognition Letters*, 26, 1874–1884, 2005.

Halkidi, M., Y. Batistakis y M. Vazirgiannis. "On clustering validation techniques". *Intelligent Information Systems*, 17(2), 107–145, 2001.

Hammer, B., A. Micheli, A. Sperduti y M. Strickert. "Recursive self-organizing network models." *Neural Networks*, 17, 1061–1085, 2004.

Hand, D. J., P. Smyth y H. Mannila. *Principles of data mining.* MIT Press Cambridge, 2001. ISBN 0-262-08290-X.

Hastie, T. y W. Stuetzle. "Principal curves". *Journal of the American Statistical Association*, 84(406), 502–516, 1989.

Haykin, S. *Neural networks: A comprehensive foundation.* Macmillan Co., New York, 1994. ISBN 0-13-273350-1.

Helge y Ritter, "Self-organizing maps on non-euclidean spaces". En E. Oja y S. Kaski (editores), *Kohonen Maps*, 97–109, Elsevier Science B.V., 1999. ISBN 978-0-444-50270-4.

Himberg, J., *From insights to innovations: Data mining, visualization, and user interfaces.* Tesis Doctoral. Helsinki University of Technology, Espoo, Finland, 2004.

Hinton, G. E. y S. Roweis, "Stochastic neighbor embedding." En T. S. Becker y K. Obermayer (editores), *Advances in Neural Information Processing Systems*, tomo 15, 833–840, MIT Press, Cambridge, MA, 2002.

Hoaglin, D. C., F. Mosteller y J. W. Tukey. *Understanding robust and exploratory data analysis*, 1ª edición. Wiley-Interscience, 2000. ISBN 978-0471384915.

Hollmén, J., *Process modeling using the self-organizing map*. Tesis de Master. Helsinki University of Technology, Espoo, Finland, 1996.

Honkela, T., S. Kaski, K. Lagus y T. Kohonen, "Newsgroup exploration with WEBSOM method and browsing interface". Technical Report A32, Helsinki University of Technology, Espoo, Finland, 1996.

Hotelling, H. "Analysis of a complex of statistical variables into principal components". *Journal of Educational Psychology*, 15, 417–441, 1933.

Hyvärinen, A., J. Karhunen y E. Oja. *Independent component analysis*. John Wiley & Sons, 2001. ISBN 0-471-40540-X.

IEEE Association, *IEEE recommended practice for monitoring electric power quality*, 1995.

IEEE Association, *IEEE standard 802.3*, 2002.

IEEE Association, *IEEE standard requirements for instrument transformers*, 2008.

Instituto para la Diversificación y Ahorro de la Energía, IDAE, *Plan de acción 2008-2012 de la estrategia de ahorro y eficiencia energética en España.*, 2007.

International Energy Agency, IEA, *Key world energy statistics*, 2010.

Jain, A., M. N. Murty y P. J. Flynn. "Data clustering: A review". *ACM Computing Surveys*, 31(3), 264–323, 1999.

Jamasb, T. y M. G. Pollitt, "Electricity market reform in the European Union: review of progress towards liberalisation and integration". Informe Técnico 05-003WP, MIT Center for Energy and Environmental Policy Research, 2005.

Janssen, R., "Towards energy efficient buildings in Europe". Informe Técnico July 2005, The European Alliance of Companies for Energy Efficiency in Buildings, London, UK, 2005.

Kangas, J., *On the analysis of pattern sequences by self-organizing maps*. Tesis Doctoral. Helsinki University of Technology, Espoo, Finland, 1994.

Kangas, J., T. Kohonen y J. Laaksonen. "Variants of self-organizing maps". *IEEE Transactions on Neural Networks*, 1, 93–99, 1990.

Kaski, S., *Data exploration using self-organizing maps*. Tesis Doctoral. Helsinki University of Technology, Espoo, Finland, 1997.

Kasslin, M., J. Kangas y O. Simula, "Process state monitoring using self-organizing maps". En I. Aleksander y J. Taylor (editores), *Artificial Neural Networks*, tomo 2, 1531–1534, North-Holland, 1992.

Kastner, W., G. Neugschwandtner, S. Soucek y H. Newmann. "Communication systems for building automation and control". *Proceedings of the IEEE*, 93(6), 1178–1203, 2005.

Keim, D. y H. P. Kriegel. "Visualization techniques for mining large databases: a comparison." *IEEE transactions on knowledge and data engineering*, 8(6), 923–938, 1996.

Keim, D. A. "Visual exploration of large data sets." *Communications of the ACM*, 44(8), 38–44, 2001.

Keim, D. A. "Information visualization and visual data mining." *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 1–8, 2002.

Kendall, M. G. *Rank correlation methods.* Charles Griffin and Company, 1948. ISBN 978-0-195-20837-5.

Khan, A. "Monitoring power for the future". *Power Engineering Journal*, 15(2), 81–85, 2001.

Kohonen, T. "Self-organizing formation of topologically correct feature maps". *Biological Cybernetics*, 43(1), 59–69, 1982.

Kohonen, T. "The self-organizing map." *Proceedings of the IEEE*, 78(9), 1464–1480, 1990.

Kohonen, T. *Self-organizing maps*, 3ª edición. Springer-Verlag New York, Inc., Secaucus, USA, 2001. ISBN 3–5406–7921–9.

Kohonen, T., E. Oja, O. Simula, A. Visa y J. Kangas. "Engineering applications of the self-organizing map". *Proceedings of the IEEE*, 84(10), 1358–84, 1996.

Koikkalainen, P. y E. Oja, "Self-organizing hierarchical feature maps". En *International Joint Conference on Neural Networks, IJCNN1990*, tomo 2, 279–284, San Diego, USA, 1990.

König, A. "Interactive visualization and analysis of hierarchical neural projections for data mining". *IEEE Transactions on Neural Networks*, 11(3), 615–624, 2000.

Koskela, T., M. Varsta, J. Heikkonen y K. Kaski, "Temporal sequence processing using Recurrent SOM". En *2nd International Conference on Knowledge-Based Intelligent Electronic Systems, KES1998*, tomo 1, 290–297, Adelaide, Australia, 1998.

Kramer, M. "Nonlinear principal component analysis using autoassociative neural networks". *The American Institute of Chemical Engineers, AIChE Journal*, 37, 233–243, 1991.

Kruskal, J. B. y M. Wish. "Multidimensional scaling". *Sage University Paper Series on Quantitative Application in the Social Sciences*, 7–11, 1978.

Kuo, R., L. Ho y C. Hu. "Integration of self-organizing feature map and k-means algorithm for market segmentation". *Computers & Operations Research*, 29(11), 1475–1493, 2002.

Laaksonen, J., M. Koskela, S. Laakso y E. Oja. "PicSOM - content-based image retrieval with self-organizing maps". *Pattern Recognition Letters*, 21, 1199–1207, 2000.

Laine, S., *Using visualization, variable selection and feature extraction to learn from industrial data.* Tesis Doctoral. Helsinki University of Technology, Espoo, Finland, 2003.

Lee, J. A., C. Archambeau y M. Verleysen, "Locally linear embedding versus Isotop." En *11th European Symposium on Artificial Neural Networks, ESANN2003*, 527–534, Bruges, Belgium, 2003.

Lee, J. A., A. Lendasse, N. Donckers y M. Verleysen, "A robust nonlinear projection method." En *8th European Symposium on Artificial Neural Networks, ESANN2000*, 13–20, Bruges, Belgium, 2000.

Lee, J. A. y M. Verleysen. *Nonlinear dimensionality reduction.* Information Science and Statistics. Springer, 2007. ISBN 978-0387393506.

Lendasse, A., J. Lee, V. Wertz y M. Verleysen. "Forecasting electricity consumption using nonlinear projection and self-organizing maps". *Neurocomputing*, 48(1), 299–311, 2002.

Liao, H. y D. Niebur. "Load profile estimation in electric transmission networks using independent component analysis". *IEEE Transactions on Power Systems*, 18(2), 707–715, 2003.

Liberty, J. y D. Hurwitz. *Programming ASP.NET*, 2ª edición. O'Reilly Media, USA, 2003. ISBN 0–59–600487–7.

Linde, Y., A. Buzo y R. Gray. "An algorithm for vector quantizer design". *IEEE Transactions on Communications*, 28(1), 84–95, 1980.

Linsker, R. "How to generate ordered maps by maximizing the mutual information between input and output signals". *Neural Computation*, 1(3), 402–411, 1989.

Luttrell, S. P., "Self-organisation: a derivation from first principles of a class of learning algorithms". En *International Joint Conference on Neural Networks, IJCNN1989*, tomo 2, 495–498, 1989.

MacQueen, J., "Some methods for classification and analysis of multivariate observations". En L. LeCam y J. Neyman (editores), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I: Statistics*, 281–297, University of California Press, Berkeley and Los Angeles, USA, 1967.

Marcuse, J., B. Menz y J. Payne. "Servers in SCADA applications". *IEEE Transactions on Industry Applications*, 33(5), 1295–1299, 1997.

Martinetz, T. M., S. G. Berkovich y K. J. Schulten. ""Neural-Gas" network for vector quantization and its application to time-series prediction". *IEEE Transactions on Neural Networks*, 4(4), 558–569, 1993.

Melssen, W., R. Wehrens y L. Buydens. "Supervised Kohonen networks for classification problems". *Chemometrics and Intelligent Laboratory Systems*, 83, 99–113, 2006.

Modbus Organization, *Modbus application protocol specification*, 2006.

Mohagheghi, S., J. Stoupis y Z. Wang, "Communication protocols and networks for power systems-current status and future trends". En *IEEE/PES Power Systems Conference and Exposition, PSCE2009*, 1–9, Seattle, USA, 2009.

Motegi, N. y M. A. Piette, "Web-based energy information systems for large commercial buildings". En *10th National Conference on Building Commissioning, NCBC2002*, Lawrence Berkeley National Laboratory, Chicago, USA, 2002.

Mulier, F. y V. Cherkassky. "Self-organization as an iterative kernel smoothing process". *Neural Computation*, 7(6), 1165–1177, 1995.

Nagata, T., "An electric power energy monitoring system in campus using an Internet". En *IEEE Power Engineering Society General Meeting*, Montreal, Canada, 2006.

National Instruments, *LabVIEW user manual, no. 320999E-01*, 2003.

Ndiaye, D. y K. Gabriel. "Principal component analysis of the electricity consumption in residential dwellings". *Energy and Buildings*, 43(2-3), 446–453, 2011.

Ng, A. Y., M. I. Jordan y Y. Weiss, "On spectral clustering: Analysis and an algorithm". En *Advances in Neural Information Processing Systems*, 849–856, MIT Press, 2001.

Oberthür, S. y H. E. Ott. *The Kyoto protocol: International climate policy for the 21st century*, 1ª edición. Springer-Verlag, 1999. ISBN 3-540-66470-X.

Oja, M., S. Kaski y T. Kohonen. "Bibliography of self-organizing map (SOM) papers: 1998–2001 addendum". *Neural Computing Surveys*, 3, 1–156, 2003.

Park, J. y I. W. Sandberg. "Universal approximation using Radial-Basis-Function networks". *Neural Computation*, 3(2), 246–257, 1991.

Pérez, J., L. Sánchez y M. Pardo. *La gestión de la demanda de la electricidad*, tomo 1 y 2. Fundación alternativas, 2005. ISBN 84–96204–65–0.

Pérez-Lombard, L., J. Ortiz y C. Pout. "A review on buildings energy consumption information". *Energy and Buildings*, 40(3), 394–398, 2008.

Prada, M. A., *Técnicas de extracción del conocimiento basadas en data mining visual para la supervisión de procesos industriales. Análisis de la dinámica basado en mapas auto-organizados.* Tesis Doctoral. Universidad de León, León, España, 2009.

Principe, J., N. Euliano y S. Garani. "Principles and networks for self-organization in space-time." *Neural Networks*, 15(8-9), 1069–1083, 2002.

Qiu, B., H. B. Gooi, Y. Liu y E. K. Chan. "Internet-based SCADA display system". *IEEE Computer Applications in Power*, 15(1), 14–19, 2002.

Räsänen, T., J. Ruuskanen y M. Kolehmainen. "Reducing energy consumption by using self-organizing maps to create more personalized electricity use information". *Applied Energy*, 85(9), 830–840, 2008.

Rauber, A. y D. Merkl, "Automatic labeling of self-organizing maps: Making a treasure-map reveal its secrets". En N. Zhong y L. Zhou (editores), *Methodologies for Knowledge Discovery and Data Mining*, tomo 1574 de *Lecture Notes in Computer Science*, 228–237, Springer Berlin / Heidelberg, 1999. ISBN 978-3-540-65866-5.

Rauber, A., D. Merkl y M. Dittenbach. "The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data". *IEEE Transactions on Neural Networks*, 13(6), 1331–1341, 2002.

Red Eléctrica de España, REE, *El sistema eléctrico español 2010*, 2011.

Ritter, H., T. Martinetz y K. Schulten. *Neural computation and self-organizing maps: An introduction*. Addison Wesley Longman Publishing Co., Redwood City, USA, 1992. ISBN 0-201-55442-9.

Rohrer, R. M. y E. Swing. "Web-based information visualization". *IEEE Computer Graphics and Applications*, 17(4), 52–59, 1997.

Roncero, C., M. I. Milanés, E. Romero, E. González y F. Barrero. "Medida de energía en condiciones de distorsión y desequilibrio". *DYNA Ingeniería e Industria*, 86(5), 567–574, 2011.

Rossiter, A. "Make your plant more energy efficient". *Chemical Engineering Progress, CEP*, 101(12), 31–36, 2005.

Roweis, S. T. y L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding". *Science*, 290, 2323–2326, 2000.

Samad, T. y S. A. Harp. "Self-organization with partial data". *Network: Computation in Neural Systems*, 3, 205–212, 1992.

Sammon, Jr., J. W. "A non-linear mapping for data structure analysis". *IEEE Transactions on Computers*, 18, 401–409, 1969.

Schneider Electric, *PowerLogic series 4000 Circuit Monitor reference manual*, 2005.

Schneider Electric, *PowerLogic series 800 Power Meter reference manual*, 2006.

Schneider Electric, *PowerLogic ION 7550/7650 power and energy meters user guide*, 2010.

Schölkopf, B., A. Smola y K.-R. Müller, "Kernel principal component analysis". En W. Gerstner, A. Germond, M. Hasler y J.-D. Nicoud (editores), *Artificial Neural*

Networks–ICANN1997, tomo 1327 de *Lecture Notes in Computer Science*, 583–588, Springer Berlin / Heidelberg, 1997. ISBN 978-3-540-63631-1.

Sforna, M. "Data mining in a power company customer database". *Electric Power Systems Research*, 55(3), 201–209, 2000.

Similä, T., *Advances in variable selection and visualization methods for analysis of multivariate data.* Tesis Doctoral. Helsinki University of Technology, Espoo, Finland, 2007.

Simula, O., J. Vesanto, E. Alhoniemi y J. Hollmén, "Analysis and modeling of complex systems using the self-organizing map." En N. Kasabov y R. Kozma (editores), *Neuro-Fuzzy Techniques for Intelligent Information Systems.*, tomo 1, 3–22, Physica-Verlag, 1999. ISBN 978-3-790-81187-2.

Specht, D. F. "A general regression neural network". *IEEE Transactions on Neural Networks*, 2(6), 568–576, 1991.

Sulkava, M., *Learning from environmental data: methods for analysis of forest nutrition time series.* Tesis Doctoral. Helsinki University of Technology, Espoo, Finland, 2008.

Tafreshi, S. M. M. y M. Farhadi, "Improved SOM based method for short term load forecast of Iran power network". En *8th International Power Engineering Conference, IPEC2007*, 1377–1384, Singapore, 2007.

Taylor, J. W., L. M. de Menezes y P. E. McSharry. "A comparison of univariate methods for forecasting electricity demand up to a day ahead". *International Journal of Forecasting*, 22(1), 1–16, 2006.

Tenenbaum, J. B., V. de Silva y J. C. Langford. "A global geometric framework for nonlinear dimensionality reduction". *Science*, 290, 2319–2323, 2000.

Tikka, J., *Input variable selection methods for construction of interpretable regression models.* Tesis Doctoral. Helsinki University of Technology, Espoo, Finland, 2008.

Torgerson, W. S. "Multidimensional scaling: I. theory and method". *Psychometrika*, 17(4), 401–419, 1952.

Tryba, V., S. Metzen y K. Goser, "Designing basic integrated circuits by self-organizing feature maps". En *Neuro-Nîmes'89. International Workshop on Neural Networks and their Applications*, 225–235, Nanterre, France, 1989.

Tsekouras, G., P. Kotoulas, C. Tsirekis, E. Dialynas y N. Hatziargyriou. "A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers". *Electric Power Systems Research*, 78(9), 1494–1510, 2008.

Tufte, E. *The visual display of quantitative information.* Graphics Press, 1983. ISBN 978-0-961-39210-9.

Tukey, J. W. *Exploratory data analysis.* Addison-Wesley, 1977. ISBN 978-0-201-07616-5.

Tukey, J. W. "We need both exploratory and confirmatory". *The American Statistician*, 34(1), 23–25, 1980.

Ultsch, A., "Maps for the visualization of high-dimensional data spaces". En *4th Workshop on Self Organizing Maps, WSOM2003*, 225–230, Kitakyushu, Japan, 2003a.

Ultsch, A., "U-Matrix: A tool to visualize clusters in high dimensional data". Informe Técnico 36, Department of Mathematics of Computer Science, University of Marburg, 2003b.

Ultsch, A., "U*C: Self-organized clustering with emergent feature maps". En *Lernen Wissensentdeckung Adaptivität, LWA2005*, 240–244, Saarbrücken, Germany, 2005.

Ultsch, A. y H. P. Siemon, "Kohonen's self-organizing feature maps for exploratory data analysis". En *International Neural Network Conference, INNC1990*, 305–308, Dordrecht, Netherlands, 1990.

Unión Europea, *DIRECTIVA 96/92/CE DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 19 de Diciembre sobre normas comunes para el mercado interior de la electricidad.*, 1997.

Unión Europea, *DIRECTIVA 2010/31/CE DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 19 de Mayo relativa a la eficiencia energética de los edificios.*, 2010.

US Energy Information Administration, US EIA, *International Energy Outlook 2007 (IEO2007)*, 2007.

Van der Maaten, L. y G. Hinton. "Visualizing data using t-SNE". *Journal of Machine Learning Research*, 9, 2579–2605, 2008.

Vandenberghe, L. y S. Boyd. "Semidefinite programming". *SIAM (Society for Industrial and Applied Mathematics) Review*, 38(1), 49–95, 1996.

Vankayala, V. S. S. y N. D. Rao. "Artificial neural networks and their applications to power systems: a bibliographical survey". *Electric Power Systems Research*, 28(1), 67–79, 1993.

Venna, J., *Dimensionality reduction for visual exploration of similarity structures.* Tesis Doctoral. Helsinki University of Technology, Espoo, Finland, 2007.

Venna, J. y S. Kaski. "Local multidimensional scaling." *Neural Networks*, 19(6-7), 889–899, 2006.

Verdú, S. V., M. O. García, C. Senabre, A. G. Marín y F. G. Franco. "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps". *IEEE Transactions on Power Systems*, 21(4), 1672–1682, 2006.

Vesanto, J. "SOM-based data visualization methods." *Intelligent Data Analysis*, 3(2), 111–126, 1999.

Vesanto, J., *Data exploration process based on the self-organizing map.* Tesis Doctoral. Helsinki University of Technology, Espoo, Finland, 2002.

Vesanto, J. y E. Alhoniemi. "Clustering of the self-organizing map". *IEEE Transactions on Neural Networks*, 11(3), 586–600, 2000.

Vesanto, J., J. Himberg, E. Alhoniemi y J. Parhankangas, "SOM toolbox for Matlab 5". Informe Técnico A57, Helsinki University of Technology, 2000.

Villmann, T., R. Der, M. Herrmann y T. M. Martinetz. "Topology preservation in self-organizing feature maps: exact definition and measurement". *IEEE Transactions on Neural Networks*, 8(2), 256–266, 1997.

Walter, J., C. Nölker y H. Ritter, "The PSOM algorithm and applications". En *International ICSC Symposium on Neural Computertion, NC2000*, 758–764, Berlin, Germany, 2000.

Wang, S. y J. Xie. "Integrating building management system and facilities management on the Internet". *Automation in Construction*, 11(6), 707–715, 2002.

Weinberger, K. Q. y L. K. Saul, "Unsupervised learning of image manifolds by semidefinite programming". En *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR2004*, tomo 2, 988–995, Philadelphia, USA, 2004.

Wong, P. C. "Visual data mining". *IEEE Computer Graphics and Applications*, 19(5), 20–21, 1999.

Xu, R. y D. Wunsch. "Survey of clustering algorithms". *IEEE Transactions on Neural Networks*, 16(3), 645–678, 2005.

Xu, Y., L. Tolbert, J. Kueck y D. Rizy. "Voltage and current unbalance compensation using a static var compensator". *IET Power Electronics*, 3(6), 977–9888, 2010.

Yao, A. W. y C. Ku. "Developing a PC-based automated monitoring and control platform for electric power systems". *Electric Power Systems Research*, 64(2), 129–136, 2003.

Yin, H. "Data visualisation and manifold mapping using ViSOM". *Neural Networks*, 15, 1005–1016, 2002.

Zhang, J. *Visualization for information retrieval*, 1ª edición. Springer, 2007. ISBN 978-3-540-75147-2.

Zhang, Z. y H. Zha. "Principal manifolds and nonlinear dimension reduction via local tangent space alignment". *SIAM (Society for Industrial and Applied Mathematics) Journal of Scientific Computing*, 26, 313–338, 2004.

Zimmermann, H. "OSI reference model–the ISO model of architecture for open systems interconnection." *IEEE Transactions on Communications*, 28(4), 425–432, 1980.