Proyecto Fin de Máster

Sistemas de Energía Térmica

SISTEMA DE GESTIÓN ENERGÉTICA PARA EDIFICIOS EMS SIGUIENDO HOJA DE RUTA BS-ISO 50006:2014

Autor: María del Carmen Guerrero Delgado

Tutor: Jose Luis Molina Felix

Co-Director: Jose Sánchez Ramos

Dep. Ingeniería Energética

Escuela Técnica Superior de Ingeniería

Sevilla, 2016

Proyecto Fin de Máster

Sistemas de Energía Térmica

SISTEMA DE GESTIÓN ENERGÉTICA PARA EDIFICIOS EMS SIGUIENDO HOJA DE RUTA BS-ISO 50006:2014

Autor:

María del Carmen Guerrero Delgado

Tutor:

José Luis Molina Felix

Co-Director:

José Sánchez Ramos

Dep. Ingeniería Energética

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

Sevilla, 2016

Proyecto Fin de Máster: SISTEMA DE GESTIÓN ENERGÉTICA PARA EDIFICIOS EMS SIGUIENDO HOJA DE RUTA BS-ISO 50006:2014

Autor: María del Carmen Guerrero Delgado

Tutor: Jose Luis Molina Felix

Co-Director Jose Sánchez Ramos

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2016

El Secretario del Tribunal

A mis padres, hermana y familia

AGRADECIMIENTOS

El trabajo presentado en el presente proyecto final de Master, no hubiera sido posible sin el apoyo incondicional de mi seres queridos y los grandes profesionales que me rodean.

Quisiera primero agradecer a Jose Luis Molina Felix, para mi ha sido un placer trabajar bajo su orientación y aprender dia a día de sus experiencias y enseñanza, tanto en el campo académico como profesional.

A José Sánchez Ramos de forma especial por su gran apoyo incondicional y por proporcionarme toda la ayuda y orientación necesaria durante la realización del mismo.

Finalmente, y no menos importante a Servando Álvarez Dominguez por muchos y grandes motivos.

María del Carmen Guerrero Delgado

Sevilla 2016

RESUMEN

El presente proyecto final de máster tiene como objetivo proporcionar un método para la obtención de líneas base del consumo de edificios. Este método a diferencia del material publicado y existente, tiene dos grandes aportaciones:

- Es una combinación de un algoritmo de clustering con una metodología de caracterización inversa de edificio, lo que combina una capacidad de extrapolación a diferentes condiciones climáticas y operación del edificio. Este aporte, a diferencia de los resultados que aparecen publicados, garantiza calidad de resultaos en periodos de explotación diferentes a los de referencia. De esta manera se ofrece más robustez a un método de verificación de ahorros o diagnóstico de errores a largo plazo
- Posibilidad de automatizarlo y explotarlo en software para gestión energética o verificación de ahorros

Por tanto el Trabajo que aquí se presenta, vinculado con las normas UNE-EN 50001 y BS ISO 50006, da solución al elemento clave que da sentido a ambas: obtención de línea base y referencias de consumo.

A lo largo del documento, además de explicar el propio algoritmo y sus fundamentos, se exponen ejemplos de 6 edificios tipo gran terciario con usos de oficina, judiciales, y sanitarias. En todos ellos los resultados en el periodo de referencia y en el periodo de validación son aceptables y de alta eficiencia.

…

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1 INTRODUCCIÓN

La energía es el impulso y un requisito fundamental para el desarrollo tanto social como económico. Desde la Revolución Industrial, el consumo energético mundial no ha parado de aumentar. Hoy por hoy, la utilización de dicha energía está ligada al estilo de vida, se consume más energía cuanto más desarrollada está una sociedad. Es por ello que, durante los últimos veinte años el reclamo de productos energéticos se ha incrementado en un 3% anual a nivel mundial.

Este hecho lleva a pensar en la situación energética actual. Como se puede ver en la Figura 1 a nivel mundial, el consumo de energía primaria se basa en energías no renovables. Aún así, no existe un grave problema con el suministro ya que se puede producir lo que se demanda. Pero al ser fuentes no renovables, éstas tardan millones de años en crearse por lo que en un futuro no muy lejano, no se podrán ofertar o su precio será excesivamente elevado.

España es un país muy dependiente energéticamente del petróleo, superando la media tanto europea como mundial. Esto hace que el país tenga una gran vinculación en este ámbito con el exterior. Esto también sucede en Europa donde actualmente se importa el 50% de la energía consumida y de la que se prevé que alcance el 70% en 2030.

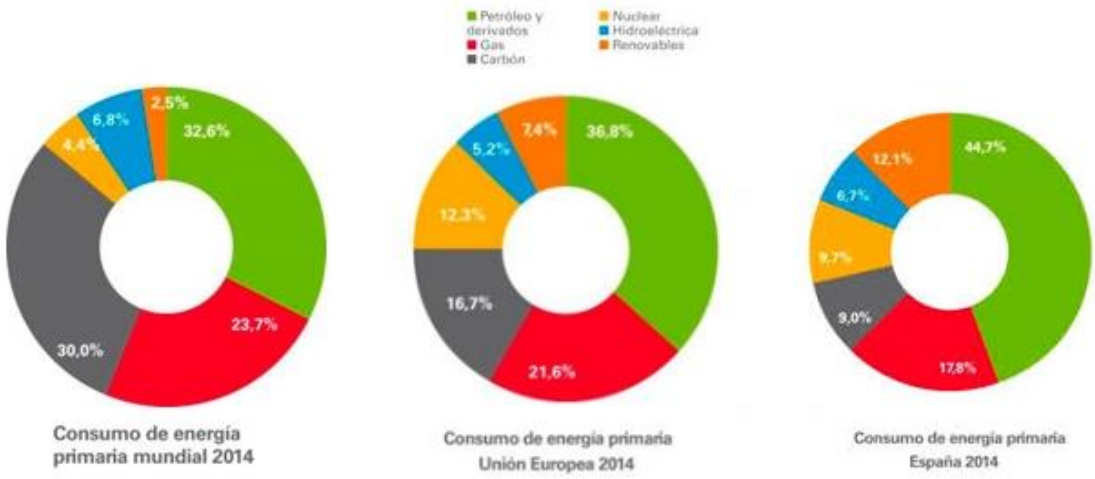

Figura 1. Consumos de energías primarias. Fuente: BP Statistical Review.

Otro problema que crea el consumo de energías no renovables son las emisiones y otros impactos medio ambientales que tanto su extracción como su transformación y transporte crean. Por ejemplo, la combustión del gas, del carbón o de los derivados petrolíferos generan emisiones tales como el CO_2 y el CO entre otros, que contribuyen a la generación y el aumento del efecto invernadero, la lluvia ácida o la contaminación del suelo, del aire y el agua.

Por estos problemas, la Unión Europea está llevando a cabo el plan Horizonte 2020 (H2020) [1]. Éste es el Programa Marco de Investigación e Innovación de la UE entre 2014-2020, que tiene como retos sociales, entre otros, la energía, el cambio climático y el uso eficaz de los recursos.

Además, otro dato importante es que, en Europa, el 40% de la energía total utilizada es consumida por los edificios y en éstos, más del 20% de dicha energía es desaprovechada. La mayor parte de ese consumo es debido a la calefacción y refrigeración de los edificios. Es por ello que la Comisión Europea en su Estrategia relativa a la calefacción y refrigeración [2], busca que el acondicionamiento, tanto en hogares como en empresas, sea más sostenible y eficiente. Asimismo, promueve la reducción de las importaciones y dependencias energéticas, de los costes, de las emisiones de gases de efecto invernadero además de cumplir con el acuerdo alcanzado en la Conferencia de París (COP21) sobre el clima.

Por todo eso, se plantea una necesidad inmediata de ahorro y reducción de la demanda. El método más sencillo y veraz de evaluación de ahorros, así como de medida de eficiencia para la certificación energética son las líneas base.

Según la norma ISO 50001·[3] la línea base es "*una referencia cuantitativa que proporciona la base de comparación del desempeño energético*". Además indica que ésta "*refleja un período especificado*" así como que "*puede normalizarse utilizando variables que afecten al uso y/o consumo de la energía*" y que es útil para el cálculo de ahorros.

A día de hoy se carece de un protocolo que proporcione a las empresas de los distintos sectores una ayuda para la obtención de dichas líneas. De esa carencia nace el presente proyecto.

Diversos estudios verifican que los sectores que abarcan mayor número de consumidores finales de energía, y por tanto tienen un mayor potencial de ahorro, son el sector de la vivienda y el terciario. Esto es debido fundamentalmente al uso de la calefacción y refrigeración, del alumbrado y de los aparatos eléctricos.

Como se observa en la Figura 2 en ambos casos el mayor consumo se debe a la climatización, siendo éste más acusado en el sector terciario. El consumo de climatización es el que más potencial de ahorro tiene motivado sobre todo por el mal uso que se hace de él.

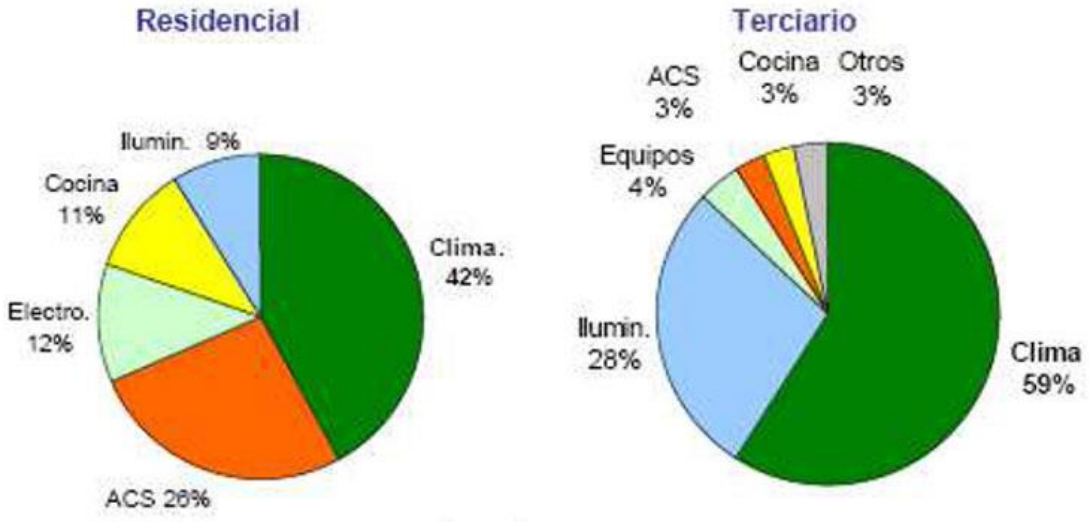

Figura 2. Consumo de energía en edificios residenciales y terciarios. Fuente: SGE

El ahorro en los edificios terciarios se ve promovido por la política energética existente en España. Éstos requieren un tratamiento diferente a la vivienda debido a la necesidad de retorno de la inversión, por su mayor tasa de rehabilitación y la importancia que en estos edificios tienen las medidas pasivas.

El principal problema en este sector nace de la demanda. Para los edificios construidos, la demanda está definida por la orientación y la envolvente térmica del mismo, las condiciones climáticas, así como de las cargas internas. La mayoría de estos factores son variables en el tiempo, lo que lleva a una demanda también variable.

Este sector tiene un potencial de ahorro muy elevado, es por ello que tanto la normativa actual como las empresas líderes en el sector energético, se centran en la evaluación energética y el cálculo de ahorros en edificios terciarios.

1.1 Estado del arte

Una vez comprobada que existe una necesidad, ¿qué se hace actualmente al respecto? En este apartado se va a describir cómo diferentes empresas u organizaciones utilizan las líneas base en diferentes aspectos relacionados con la eficiencia energética. Sin embargo, ninguno de ellos hace referencia a la obtención de las mismas corroborando así la necesidad del presente estudio.

1.1.1. EVO

Efficiency Valuation Organization, EVO, es una organización sin ánimo de lucro que busca ayudar a los ingenieros e invertir en productos de eficiencia energética. Esta organización es la única en el mundo dedicada a la creación de herramientas de medida y verificación (M&V) permitiendo así el progreso de la eficiencia. El plan de medida y verificación aporta un método sistemático para determinar el rendimiento energético para los usos finales de la instalación. The International Performance Measurement and Verification Protocol (IPMVP) [4] suministra un borrador de dicho plan.

El IPMVP distingue cuatro métodos de medida y verificación:

- A: Verificación Aislada de la Medida de la Eficiencia (MMEE): medición del parámetro clave. El ahorro se obtiene midiendo el parámetro clave en la instalación determinando así el consumo energético. En función de la variación que se prevea de dicho parámetro y de la duración del periodo de referencia, la medición puede ser puntual o continua.
  A partir de datos históricos, de especificaciones del fabricante o de suposiciones técnicas se realizan estimaciones de los parámetros que no han sido medidos, pudiendo así calcular el ahorro estimado a partir de dichos parámetros.
- B: Verificación Aislada de la MMEE: medición de todos los parámetros. En este caso, el ahorro se determina midiendo el consumo de energía de la instalación. Al igual que en la opción A, la medición puede ser continua o puntual.
- C: Verificación de toda la instalación. Esta opción se fundamenta en la monitorización del consumo energético de toda la instalación o de parte de la misma. Para esta alternativa, la medición ha de ser continua.
- D: Simulación calibrada. Consiste en la obtención del ahorro simulando el consumo de energía de parte o de toda la instalación. Esta simulación debe ser capaz de ajustar el rendimiento energético real de la instalación.

Para el caso de la evaluación integral de un edificio las opciones propias serían, o bien la opción C o bien la opción D. Todas las alternativas anteriores necesitan de una línea base energética para la estimación tanto de ahorros como de consumos, obteniendo así modelos que, una vez son calibrados, se asemejen lo máximo posible a la realidad.

En su página web [5], EVO describe la documentación que debe ser entregada para obtener un buen plan de M&V así como un buen cálculo de ahorros. Además, en ella hay ejemplos de M&V aplicados a diferentes equipos y situaciones en los cuales utiliza las líneas base, así como un ejemplo simplificado en Excel (ver Figura 6) donde obtiene la línea base de consumo y de demanda para un año así como éstas mismas respecto a los grados días de calefacción (en inglés, heating degree days, HDD). Es destacable que los cálculos realizados en dicha hoja Excel están en base mensual.

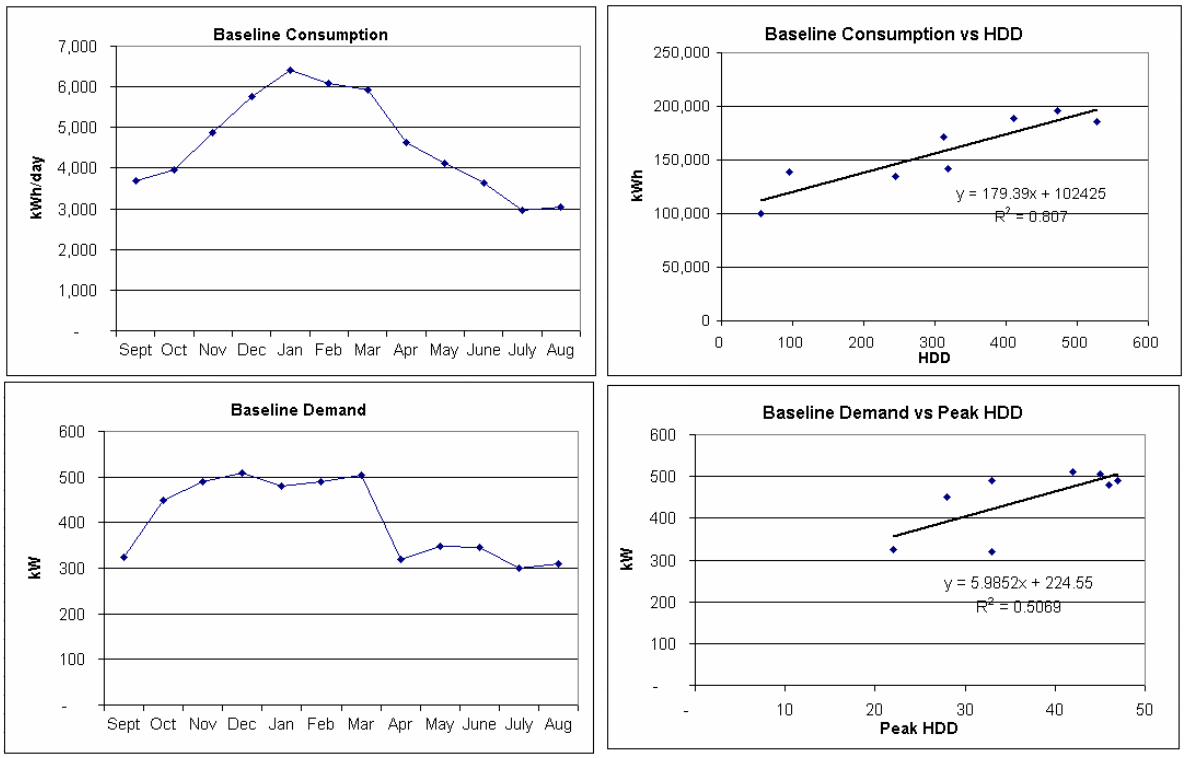

Figura 3. Ejemplo EVO.

1.1.2. DEXMA

DEXMA Energy Management es una empresa fundada en 2007 en Barcelona que ofrece un software flexible y rentable que integra herramientas que permite la visibilidad completa del consumo energético. Proponen como solución a la gestión energética la aplicación DEXCell Energy Manager (Figura 4). Ésta combina la monitorización avanzada, así como el análisis, las alertas, la creación de informes y de paneles de control en una plataforma de fácil uso.

En la página web [6] de esta compañía se encuentran diferentes ejemplos de actuación en gestión energética, todos referidos a edificios del sector terciario como son supermercados u oficinas. En todos ellos plantean una solución similar que consiste en medir y monitorizar los consumos que tiene la instalación. Un medidor irá a la acometida general de electricidad y dependiendo del uso del edificio y de los equipos que posea el resto de medidores se ubicarán en la acometida de frío industrial, de gas, de HVAC,… Esta información se obtendrá en tiempo real.

Una vez recogidos los datos, haciendo uso de la aplicación antes mencionada, se consigue analizar y gestionar la energía detectando así posibles ineficiencias energéticas.

Figura 4. DEXCell Energy Manager.

Esta herramienta permite obtener los consumos horarios de un período y compararlo con uno de la misma magnitud anterior (Figura 5), el consumo del mes presente y lo contrasta con el anterior así como hacer estimaciones mediante simulación del consumo anual, proporcionando el ahorro que se podría producir al reducir el consumo. Además esta aplicación genera informes sobre los costes producidos, documentando así, los incrementos o disminuciones en la potencia base que se producen semanalmente.

Una característica peculiar de este software es que crea alertas, avisando así al cliente, cuando los valores de consumo que se producen son superiores al máximo estimado.

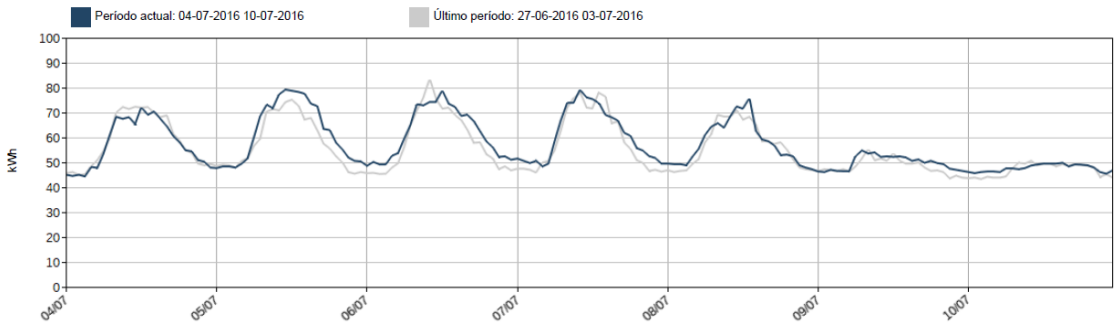

Figura 5. Ejemplo consumo eléctrico DEXMA.

Bien es cierto que esta empresa estima los consumos mediante un métido estadíatico en base a facturas anteriores.

1.1.3. AENOR

La Asociación Española de Normalización y Certificación (AENOR) es una entidad privada sin fines lucrativos. Su actividad contribuye a una mejora en la calidad y competitividad de las empresas, sus productos y servicios a través del desarrollo de normas técnicas y certificaciones. Un claro ejemplo es la contribución de AENOR en la elaboración de la Norma ISO 50001.

En el libro "Gestión de la eficiencia energética: cálculo del consumo, indicadores y mejora" [7] editado por esta asociación, se expone una forma de establecer la línea base a partir de la tabla representada en la Figura 6. Esta tabla representa una matriz donde las cuatro columnas situadas a la izquierda recogen el inventario de instalaciones, así como los equipos que la componen, la tipología de energía consumida y las áreas donde se ubican. Por su parte, las columnas situadas más a la derecha reúne el conjunto de indicadores de desempeño que la organización crea apropiado. Esta tabla se completa calculando los indicadores para un periodo base, que suele ser de un año.

Esta recopilación de información sirve como base para comparar otros periodos anteriores o posteriores al del cálculo pudiendo obtener correlaciones que evalúen el desarrollo del desempeño energético. Dichas correlaciones serán las líneas base buscadas.

Inventario instalaciones y equipos			Áreas de actividad	Indicadores de desempeño energético					
			Área 1	Intensidad energética (energía/unidad económica relevante)	Eficiencia energética	Otros 1	Otros 2	.....	Otros n
Instalación 1	equipo 1	combustible 1							
		combustible n							
		energía eléctrica 1							
		otros consumos							
	equipo 2								
	3								
	4								
	5								
	...								
Instalación 2	equipo 1								
	equipo 2								
	3								
	4								
	5								
	...								
Instalación n	equipo 1								
	equipo 2								
	3								
	4								
	5								
	...		Área 2						
Instalación 1	equipo 1	combustible 1							
		combustible n							
		energía eléctrica 1							
		otros consumos							
	equipo 2								
	3								
	4								
	5								
	...								
Instalación 2	equipo 1								
	equipo 2								
	3								
	4								
	5								
	...								
Instalación n	equipo 1								
	equipo 2								
	3								
	4								
	5								
			Área n						

Figura 6. Estructura elemental de una línea de base.

1.1.4. Departamento de Energía de EE.UU (DOE)

El Departamento de Energía de Estados Unidos (en inglés, Department of Energy, DOE), trabaja con un Programa Federal de Gestión de la Energía [8] (FEMP por sus siglas en inglés). Este programa permite la cuantificación y verificación de ahorros haciendo uso de seis pasos para obtener dicho objetivo.

1. Asignar los riesgos y responsabilidades del proyecto.
2. Desarrollar un plan de M&V del proyecto.
3. Definición de la línea base.
4. Instalación de equipos y sistemas.
5. Conducta posterior a la instalación de actividades de verificación.
6. Realizar en un intervalo regular actividades de M&V.

En este caso, para la definición de la línea base son necesarias unas condiciones físicas de las que se parte como son el inventario de equipos, la ocupación, los horarios de operación de los equipos, datos meteorológicos, etc. Estas características se obtienen a través de encuestas, mediciones e inspecciones. Es fundamental a la hora de

determinar la línea base, documentar todas las hipótesis que se hayan realizado, así como decidir qué necesita ser analizado y durante cuánto tiempo.

Una vez obtenida la línea base se puede ajustar y verificar según la precisión del método de M&V que se haya seleccionado. Con ello, al comparar el consumo obtenido por la línea base con el consumo de la instalación se puede obtener los ahorros. Igualmente, con las condiciones de referencia se puede observar posibles cambios producidos durante la fase de ejecución. Después de implementar la medida de conservación de la energía, se puede volver atrás y reevaluar la línea base.

Según DOE, el cálculo de ahorros se realizaría de la siguiente forma:

Ahorro= (Energía de la línea base- Energía posterior a la instalación) ± Ajustes

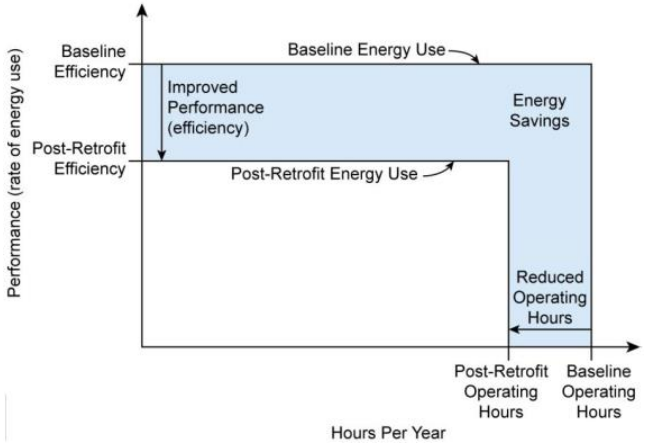

Figura 7. Ahorro en función del uso y desempeño.

Los métodos, al igual que los vistos en EVO, son:

- A: Verificación aislada con medición de parámetros clave. Es una combinación de factores medidos y estimados. Las mediciones se realizan a corto plazo, de forma puntual o continua, y se puede tomar a nivel de componente o de sistema. Los coeficientes estimados se deberán sustentar a partir de datos históricos o de los fabricantes. Los ahorros se calculan a partir de cálculo de línea base y de información del uso de la energía en el periodo de obtención de los valores medidos y estimados.
- B: Verificación aislada con medición de todos los parámetros. Esta opción se basa en mediciones de líneas base y de reacondicionamiento energético. Al igual que en el caso anterior, las mediciones se realizarán a corto plazo de forma continua o puntual, a nivel de componente o sistema. Los ahorros se determinan a partir del análisis de línea base y el consumo de energía en un periodo representativo.
- C: Verificación de toda la instalación. Esta opción se fundamenta en la monitorización continua del consumo energético de toda una instalación. Por ello, el ahorro se obtienen mediante un análisis de la línea base y del consumo en el periodo actual. Así se deben tener en cuenta variables independientes como lo son el clima y la ocupación. Esta opción requiere un inventario detallado de todos los equipos incluidos en las mediciones.
- D: Simulación calibrada. Consiste en el uso de un programa de simulación por ordenador para estimar el consumo energético de una instalación. Los modelos deben ser calibrados con datos reales de la instalación, ya sean horarios o mensuales. Dichas simulaciones deben incluir especificaciones de funcionamiento de los equipos, mediciones de consumo de componentes y los datos del medidor utilizado. Una vez calibrado el modelo, los ahorros se calculan comparando la simulación de línea base con una simulación del período de ejecución o bien con datos reales.

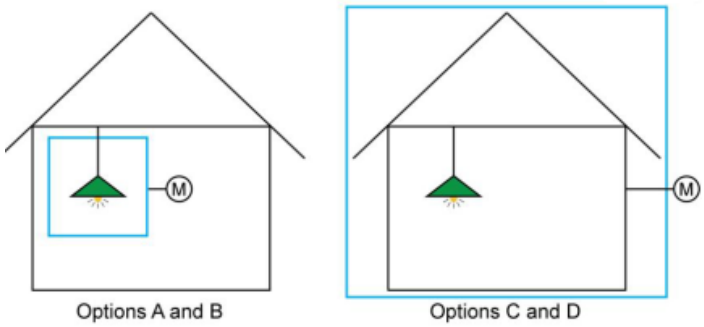

Figura 8. Métodos de M&V.

1.2 Análisis de la normativa

Ahora que se ha visto que las empresas no dan solución a la problemática, ¿existe alguna normativa que sirva de guía? Con este epígrafe se pretende hacer tanto una revisión de la normativa existente como verificar la inexistencia de una guía que facilite la obtención de la línea base energética aunque la mayor parte de esa reglamentación haga uso de ella.

1.2.1. UNE 50001-2011

Esta norma [3] tiene como objeto proveer los sistemas y procesos necesarios para que las distintas organizaciones mejoren su política energética. Está destinada a cualquier tipo de organismo.

En ella se especifican los requisitos de un sistema de gestión de energía que permite a la organización alcanzar los compromisos derivados de su política energética y establecer objetivos, metas y planes de acción para mejorar su rendimiento energético. Está basado en el ciclo de mejora continua Planificar – Hacer – Verificar – Actuar (PHVA) representado en la Figura 9. Con ello se consigue un uso más eficiente de las fuentes de energía así como la reducción del impacto ambiental.

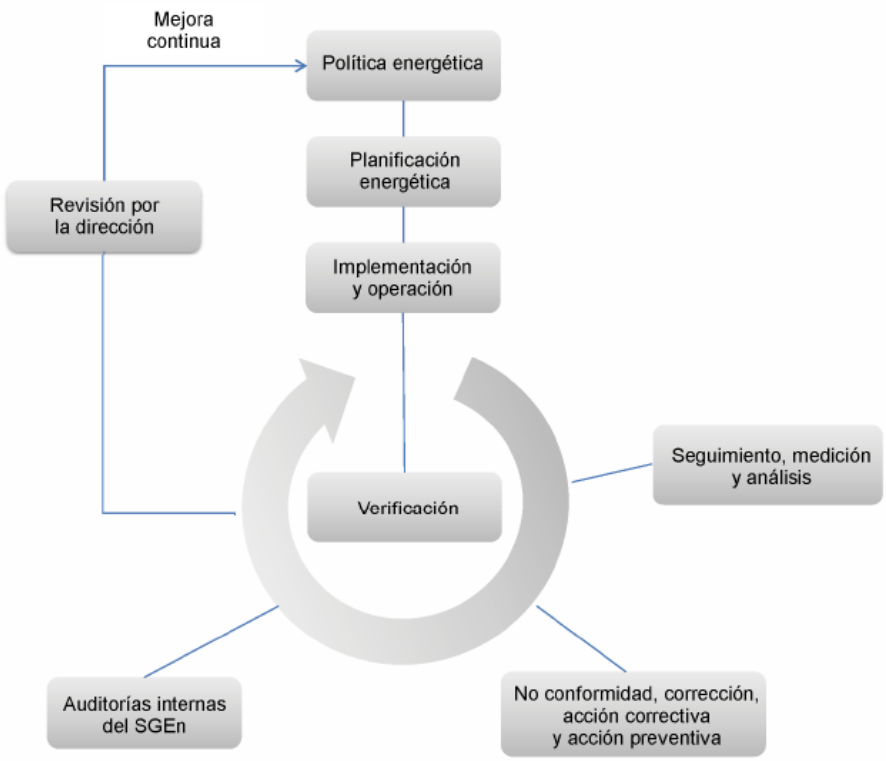

Figura 9. Modelo de sistema de gestión de la energía.

Además, esta norma puede utilizarse para la certificación, el registro y la autodeclaración del sistema de gestión de energía.

En cuanto a las líneas de base energética, están recogidas en el apartado de *Planificación energética* en el diagrama de la Figura 10 y son un requisito básico para la certificación de la ISO 50001. Estas deben ser establecidas por la organización a partir de información de revisiones energéticas anteriores y de la recopilación de datos del uso y consumo de la energía. Además, los diferentes cambios (cambios en el proceso, en los patrones de operación,…) que se produzcan en la eficiencia energética deben evaluarse en relación a dicha línea de base energética. También se deben evaluar los cambios cuando un método predeterminado así lo establezca o cuando dicha línea base no refleje el uso y consumo de energía de dicha organización.

1.2.2. ISO 50002

La norma ISO 50002 [9] habla de las auditorías energéticas estableciendo el procedimiento adecuado para realizar una buena auditoría.

Según esta norma, las líneas base se deben utilizar en el análisis de las mejoras del rendimiento energético de la organización. Éstas se emplean en los tres niveles de auditorías que diferencia la propia norma. También se incluyen las líneas base en la evaluación de oportunidades (cálculo de ahorros).

1.2.3. ISO 50003:2014

Esta norma [10] recoge los requisitos específicos para la competencia, consistencia e imparcialidad en la auditoría y la certificación de sistemas de gestión energéticos. Dichos sistemas de gestión energéticos determinan y ajustan líneas base con las que comparar los resultados para evaluar la eficiencia energética.

1.2.4. ISO 50004:2014

La norma ISO 50004 [11] proporciona una guía para el establecimiento, la implementación, el mantenimiento y la mejora de los sistemas de gestión energética.

Según esta norma, las revisiones energéticas suministran la información y los datos necesarios para el establecimiento de las líneas base energéticas. Dichas líneas deben expresar mediante una relación matemática: la relación del consumo de energía como función de una serie de variables relevantes. Debe ser un modelo ingenieril, un ratio o un dato de consumo en caso de no existir variables destacadas.

El periodo que se toma para establecer la línea base debe ser representativo en cuanto a las variaciones que se producen (ocupación, producción,…). Este se determina por los datos necesarios para establecer dicha línea.

1.2.5. ISO 50006:2014

La norma ISO 50006 [12] facilita una guía a los organismos del establecimiento, uso y mantenimiento de los indicadores de rendimiento energético (EnPIs) y de las líneas base energéticas (EnBs) como partes del proceso de evaluación de la eficiencia energética.

Esta establece que las líneas base energéticas son referencias que caracterizan y cuantifican el rendimiento energético de una institución durante un determinado periodo de tiempo. Estas líneas también se utilizan para calcular ahorros energéticos a partir de referencias anteriores y posteriores a la mejora implantada. En la Figura 13 se observa la relación entre las líneas base de energía y la eficiencia energética.

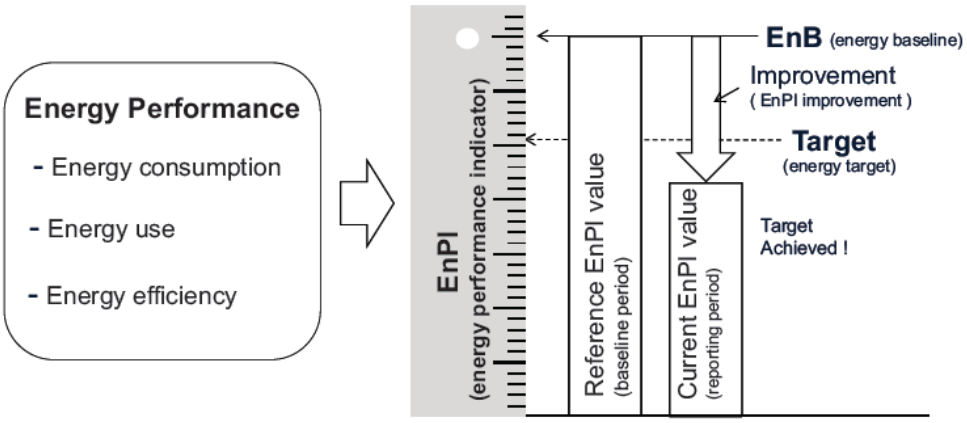

Figura 10. Relación entre rendimiento energético, EnPIs, EnBs y objetivos.

El procedimiento establecido por esta normal para un realizar una buena evaluación de la eficiencia energética se recoge en la Figura 11. Este proceso consta de cinco partes fundamentales:

- Obtención de información relevante del rendimiento energético a partir de revisiones anteriores.
- Identificación de los indicadores de la eficiencia energética.
- Establecimiento de líneas de base energéticas.
- Uso de indicadores y líneas base.
- Mantenimiento y ajuste de indicadores y líneas base.

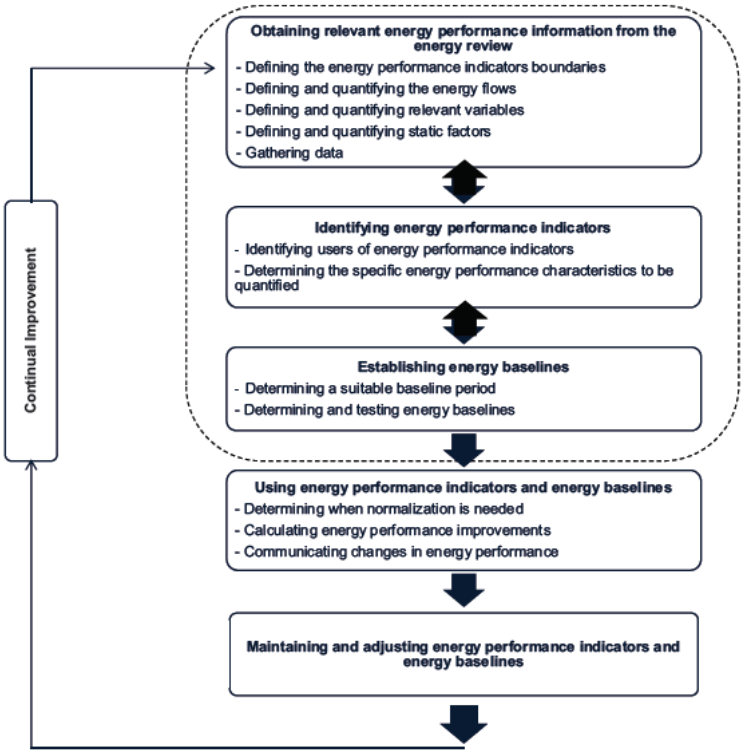

Figura 11. Información general sobre la medición de la eficiencia energética.

1.2.5.1. Obtención de información relevante del rendimiento energético a partir de revisiones anteriores

Se comienza definiendo los límites de los indicadores para lo que se debe tener en cuenta:

- Las responsabilidades organizativas en relación con la gestión de la energía,
- La facilidad de aislar el límite de los indicadores mediante mediciones de la energía y de variables relevantes.
- Los límites de los Sistemas de Gestión de Energía (SGEn)
- El uso significativo de la energía o grupo de usos que la organización designa como prioridad para controlar y mejorar.
- Equipos y procesos que la administración quiera aislar y administrar.

Una vez definidos los límites, se deben fijar y cuantificar los flujos de energía que atraviesan dichos límites. Para ello se utilizan diagramas o mapas energéticos donde se muestran los flujos interiores y los que cruzan los límites. Dependiendo de las necesidades de la organización y de sus SGEn se debe definir otras variables que sean relevantes para definir y evaluar los límites de cada indicador. Para determinar qué variables son relevantes se realiza un análisis de los datos.

Otros datos que se deben tener en cuenta son los estáticos ya que pueden modificar el valor de los límites. Es importante documentar el estado de estos factores durante el establecimiento de los indicadores y de las líneas base. Aunque los factores estáticos no varíen sustancialmente en el periodo de referencia, si las condiciones cambian, éstos podrían cambiar por lo que deben ser revisados.

Por último se ha de hacer una recolección de datos. Los datos a reunir han de ser especificados por la entidad para cada EnPI y su correspondiente EnB. La recogida de datos será periódica y el organismo seleccionará la frecuencia con la que se realizará dicha recolección. Las medidas tomadas y calculadas utilizarán datos recogidos en un periodo de tiempo específico. Antes de proceder al cálculo de EnPIs y EnB, la organización revisará el sistema de medida así como las variables relevantes para asegurar la calidad de los datos. Además, se debe considerar la calibración de los equipos para reducir el riesgo de obtener datos inexactos.

1.2.5.2. Identificación de los indicadores de la eficiencia energética

Para el reconocimiento de los EnPIs, la entidad ha de comprender sus características de consumo tales como la carga base, las cargas variables, la ocupación y el clima entre otros factores. El organismo debe fijar unos objetivos para la planificación del rendimiento energético. Estos objetivos se caracterizarán por los valores de los indicadores. Los principales tipos de EnPIs son:

- Valor de energía medido: consumos.
- Relación de los valores medidos: expresión de la eficiencia energética.
- Modelo estadístico: relación entre el consumo de energía y variables relevantes.
- Modelo basado en la ingeniería: relación entre el consumo de energía y variables relevantes a partir de simulaciones.

Los indicadores deben ser fácilmente comprensibles por los usuarios, adaptándose a la necesidad de los mismos.

1.2.5.3. Establecimiento de líneas de base energéticas

Las EnBs se caracterizan por los valores de los EnPIs durante el periodo de línea base. Los pasos a seguir para establecer las líneas base son:

- Determinar el propósito específico que se utilizará en EnBs.
- Determinar un período de datos adecuado: el periodo de referencia y el de presentación de informes debe ser lo suficientemente largo como para asegurar la variabilidad de los patrones. Son de 12 meses para tener en cuenta la relación entre el consumo y las estaciones.
- Recopilación de datos.

- Determinar y comprobar la EnBs: para determinar las líneas base, deben medirse y calcularse los diferentes indicadores haciendo uso del consumo de energía así como de las variables relevantes. Se ha de probar la validez de la EnB para asegurar que es una referencia apropiada.

1.2.5.4. Uso de indicadores y líneas base energéticos

Se ha de determinar cuándo es necesaria la normalización. La comparación directa de consumo de energía con los períodos de referencia y de notificación (método no normalizado) sólo es válido si no se producen cambios significativos. La normalización se realiza de la siguiente manera:

- Para el caso de una carga base pequeña y una única variable relevante, una sencilla relación entre la energía consumida y dicha variable significativa.
- Para el caso de una carga base grande o varias variables relevantes, se utiliza un modelo que describa dicha relación.

En cuanto al cálculo de las mejoras de eficiencia energética, se evalúan los cambios producidos comparando los valores con las EnBs. Dichos cambios han de ser comunicados.

1.2.5.5. Mantenimiento y ajuste de indicadores y líneas base

La entidad debe asegurarse de que los EnPIs, los límites y EnB siguen siendo adecuados ya que cuando se producen cambios en las instalaciones, procesos o sistemas, puede ocurrir que el uso de la energía, el consumo u otras variables se vean afectados. Para hacer dicha comprobación existen diferentes pruebas:

1. Utilizando modelos estadísticos comparar los valores de referencia con los rangos estadísticos válidos.
2. Identificar cualquier cambio importante en factores estáticos.

Si los valores de la línea base ya no son válidos, se pueden ajustar el período de referencia o ajustarla sin cambiar el período, utilizando varios métodos:

- Backcasting: se utilizan los datos del período para desarrollar un modelo estadístico, y luego se calcula el ejercicio con los datos reales.
- Se toman los datos en condiciones estándar para desarrollar un modelo estadístico y posteriormente se calcula el ejercicio con la energía real y las variables relevantes desde el inicio del estudio y presentación de informes.

También se puede utilizar una combinación de ambos.

1.2.6. ISO 50015

El propósito de esta norma [13]es establecer un conjunto de principios y directrices que se utilizarán para la medición y verificación (M&V) de la eficiencia energética de una entidad o sus componentes. Los M&V tienen como finalidad proporcionar confianza a las partes interesadas sobre los resultados. Deben dirigirse los siguientes principios:

1. La exactitud y la gestión de la incertidumbre apropiada: la incertidumbre de los resultados, incluida la precisión de la medición, debe ser administrado a un nivel apropiado para el propósito de la M&V.
2. La transparencia y la reproducibilidad del proceso de M&V: un proceso M&V debe ser documentado para garantizar la transparencia y la trazabilidad del proceso, contribuyendo así a la confianza de los resultados.
3. La gestión de datos y la planificación de mediciones: la gestión de datos incluye los medios para almacenar, mantener y asegurar los datos. Debe incluir además, información sobre la planificación acerca de la ubicación, la frecuencia, los sensores,… Todo ello irá incorporado en la documentación.
4. La competencia del practicante de M&V: el organismo debe definir los requisitos de competencia ayudando así a la confianza en los resultados.

5. La imparcialidad: el plan de M&V así como los informes del mismo deben contener una declaración sobre la imparcialidad del practicante de M&V.
6. La confidencialidad: cualquier información confidencial necesaria para llevar a cabo la M&V debe ser accesible al practicante de la misma, debiendo documentar la falta de la misma en el plan de M&V en caso de que pueda afectar al resultado.
7. El uso de métodos apropiados: la selección del método de cálculo y método de M&V estarán establecidos en el plan de M&V.

1.2.6.1. Plan de M&V

Existen seis pasos fundamentales del proceso especificados en el plan de M&V (véase Figura 12):

1. Establecer y documentar el plan de M&V, es decir, detallar cada una de las fases que lo forman.
2. Recogida de datos.
3. Verificar la implementación de la acción de mejora de la eficiencia energética (EPIAs)
4. Realizar el análisis de M&V.
5. Informe de resultados y emisión de documentación.
6. Revisar, en caso de que sea necesario, la necesidad de repetir el proceso.

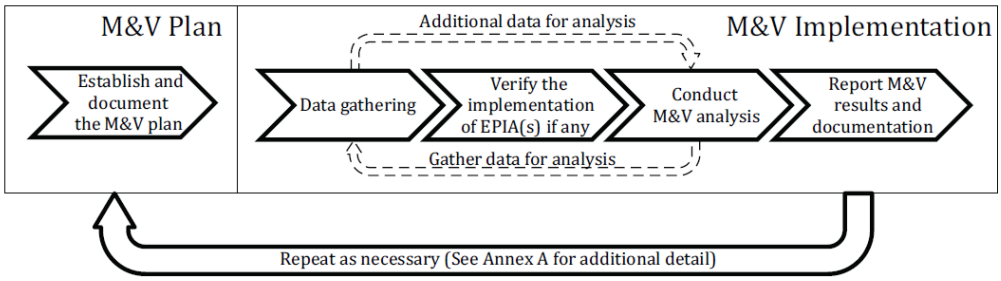

Figura 12. Pasos fundamentales en el proceso de M&V.

Las líneas base energéticas son utilizadas en varias partes del proceso de M&V:

- La acción de mejora de la eficiencia energética que incluye las EnBs como parte de su descripción.
- La elección de los límites de M&V.
- La evaluación preliminar del plan de M&V.
- El análisis de los resultados de M&V.

1.2.6.2. Establecimiento y ajuste de líneas base

Esta norma además, establece y ajusta las líneas base energéticas. Las EnBs deben ser establecidas de acuerdo con las directrices y métodos del M&V, así como los datos utilizados, cuando se quiera mejorar el rendimiento energético a partir de éstas.

Las líneas base deben ser establecidas antes de aplicar cualquier EPIA, salvo que estén disponibles los datos para obtener dicha línea base. Se documentará el establecimiento de las EnBs incluyendo la siguiente información:

- Los datos brutos utilizados para obtener la línea base.
- El período de tiempo especifico utilizado con las condiciones que implique.
- El proceso seguido.
- Los datos procesados.

En el ajuste de EnB, se proporcionarán las condiciones y razones por las que dicho ajuste es necesario y el método empleado para realizarlos. Se incluyen también ajustes no rutinarios donde se incorporan los medios para supervisar que exista la necesidad de dichos ajustes, el procedimiento a seguir en los mismos, así como el método específico a utilizar.

1.3 MINERÍA DE DATOS

La minería de datos (*Data Mining*) busca la extracción de conocimiento implícito, potencialmente útil que permanece oculto en grandes volúmenes de datos con el fin de encontrar patrones y resumirlos de forma que sean fácilmente entendibles por parte de un usuario. Se trata de una tarea interdisciplinar del campo de la informática que combina métodos estadísticos, inteligencia artificial, aprendizaje automático, reconocimiento de patrones y gestión de las bases de datos.

Dichas técnicas resultan interesantes en el campo de estudio de los edificios, intentando buscar en ellas resultados al tratamiento de grandes volúmenes de datos e interpretación de los mismos de forma sencilla con el objetivo de establecer patrones de consumo, usados tanto en gestión energética como en verificación de ahorros.

1.3.1 Reducción de la dimensionalidad

La reducción de la dimensionalidad (*Dimensionality Reduction, DR*) tiene como objetivo encontrar un conjunto de variables que generen un espacio de baja dimension y que conserven, en la medida de lo posible, la información del conjunto de datos original. La reducción de la dimensionalidad busca por un lado, transformar los datos de entrada de tal forma que se generen representaciones que faciliten el reconocimiento de patrones en los datos, y por otro, visualizar dichos datos en un sistema de coordenadas visualizable.

Existen multitud de métodos [14] que permiten transformar un conjunto de datos originales a analizar en un conjunto más reducido, pero que sigue conservando de forma significativa la información que poseen los datos de partida.

Algunos de los campos de aplicación de estas técnicas son:

- Procesado de datos de redes de sensores.
- Procesamiento de imágenes.
- Análisis de datos multivariable.
- Minería de datos.

Las técnicas de reducción de la dimensionalidad se pueden aplicar principalmente a dos tipos de datos. Por un lado, datos espaciales que están constituidos por conjuntos de variables que se pueden asemejar a coordenadas espaciales, lo cual hace que la visualización cuando el número de variables es mayor que tres, sea bastante compleja. Por otro lado, los datos temporales, que incorporan al tipo de datos anteriores información temporal, añadiendo a la visualización su evolución con el tiempo.

Se puede reducir la dimensionalidad de los datos porque generalmente suele contener información redundante, correlaciones o variables que pueden ser desechadas. Por eso el número de variables independientes que pueden describir el sistema, que es lo que se conoce como dimensión intrínseca, es menor que su dimensión nominal.

Una importante utilidad de las técnicas de reducción de la dimensionalidad es la visualización de los datos de manera entendible por el ser humano, ya que estas visualizaciones permiten la detección de patrones de forma rápida y eficiente por medio de la capacidad del sistema visual humano.

Existe un elevado número de técnicas para lograr la reducción de la dimensionalidad, que se pueden agrupar fundamentalmente en dos tipos en función de que traten de preservar la distancia de los datos de entrada o la topología de los mismos.

1.3.1.1 Reducción de la dimensionalidad por preservación de la distancia

1.3.1.1.1 Principal Component Analysis (PCA)

Uno de los métodos más antiguos y conocidos es el Principal Component Analysis (PCA). Este algoritmo genera una colección de variables y_i, combinación lineal de las variables observadas x_i, que tienen la propiedad de poseer una varianza máxima y estar incorreladas. Estas propiedades de los componentes principales permiten explicar, mediante unas pocas variables, la mayor parte de la información que contienen las variables observadas. Dada la sencillez de este método, ha sido ampliamente utilizado, también en la temática referente a la caracterización de consumos energéticos en edificios [15], ya que produce proyecciones que son fáciles de interpretar. Su principal defecto aparece cuando los datos no son lineales ya que tan solo representa una

transformación lineal ortogonal de los datos.

1.3.1.1.2 Multidimensional Scaling (MDS)

Los métodos Multidimensional Scaling (MDS) son un conjunto de métodos más que un procedimento definido. Estos métodos tienen como objetivo generar un conjunto de puntos en un espacio de baja dimensión cuyas distancias mutuas se asemejen lo más posible a las del espacio de origen. El punto de partida de los algoritmos MDS es una matriz formada por las distancias mutuas entre los vectores del espacio de entrada $\{X_K\}$ k=1,…, N.

El objetivo de estos métodos es generar una colección de puntos de un espacio de dimensión inferior, de manera que su matriz de distancias mutuas se asemeje lo más posible a la del espacio de entrada. La matriz de distancias mutuas de un conjunto de puntos define su tipología, es decir, la forma en que estos están dispuestos. La determinación de las proyecciones en estos algoritmos se lleva a cabo a través de la minimización de las funciones de coste, que penalizan disparidades entre las matrices de distancias mutuas, siendo precisamente estas funciones de coste las que marcan las diferencias entre los distintos algoritmos.

MDS constituye un algoritmo de tipo lineal, similar al PCA, que se resuelve como un problema de autovalores. Sin embargo, conserva mejor las distancias largas, lo que preserva mejor la estructura global de los datos. Al ser similar al PCA presenta las mismas ventajas y desventajas. Por comparación el MDS es más flexible que el PCA pero requiero una mayor cantidad de memoria en el procesamiento.

Un ejemplo de uso de dicho método es dentro del mercado económico [16], se emplea el mismo en conjunto con el clustering jerárquico (uso de dendogramas) como herramienta de visualización de datos y creación de grupos característicos y representativos del conjunto. Otro ejemplo de uso es el análisis del conjunto de datos de un supermercado [17] , cuyo objetivo es la predicción del comportamiento del cliente para la toma de decisiones y formación de estrategias.

1.3.1.1.3 Isometric Feature Mapping (Isomap)

El Isometric Feature Mapping es una variante del escalado multidimensional que utiliza distancias geodésicas. El resultado, que busca una optimización global, se puede obtener aplicando MDS lineal a la matriz de distancias mínimas. Este método también puede considerarse un método de aprendizaje de la variedad (manifold learning), ya que considera la forma de los datos del espacio de entrada. Además proporciona una estimación de la dimensión de la misma por medio del número de autovalores distintos de cero encontrados por el algoritmo. Bajo ciertas condiciones, se considera que el Isomap converge a una parametrización correcta de la variedad. La principal ventaja del Isomap es que, a diferencia del PCA y MDS, puede tratar con una gran cantidad de datos, aunque la escasez de datos de entrenamiento, la dependencia de la vecindad y la complejidad del algoritmo dificultan el correcto aprendizaje de la variedad topológica. Esta técnica es muy empleada en tratamiento de imágenes, el paper [18] incluye isomap como herramienta dentro de la técnica de manifold learning.

1.3.1.1.4 Nonlinear Mapping (NLM)

El Nonlinear Mapping (NLM) tiene como propósito reducir la dimensionalidad de un número finito de puntos. Se puede considerar como un tipo de algoritmo MDS que utiliza una función de coste normalizada en función de las distancias euclídeas en el espacio de entrada. La proyección de Sammon intenta preservar las distancias relativas entre los datos, dando un mayor peso a las distancias cortas en el espacio de entrada. De esta forma, se preservan las distancias locales y se busca que la estructura de los datos en el espacio de baja dimensión sea lo más similar posible a la de los datos de entrada. El proceso de minimización de la función de coste se lleva a cabo de forma iterativa, aplicando técnicas estándar como es una variación del método de Newton.

La ventaja de este método es que computacionalmente es muy simple y obtiene buenos resultados para conjuntos de datos no lineales, siempre y cuando no sean muy complejos. Su principal inconveniente es que en algunos casos, el proceso de optimización puede quedarse bloqueado en un mínimo local.

1.3.1.1.5 Geodesic Nonlinear Mapping (GNLM)

El método conocido como Geodesic Nonlinear Mapping (GNLM) es una modificación del NLM cuya función de coste usa distancias geodésicas en lugar de distancias euclídeas, en el espacio de entrada. La principal ventaja de este método es que la proyección es más precisa para datos que presentan un gran plegamiento.

1.3.1.1.6 Curvilinear Component Analysis (CCA)

El Curvilinear Component Analysis (CCA) fue propuesto como una mejora de los mapas autoorganizados. A diferencia de este método, que se verá más adelante, el CCA pertenece a los métodos de preservación de la distancia, estando más relacionado con la proyección de Sammon. Este método minimiza una función de coste basada en distancias entre puntos.

Una de las principales ventajas del CCA es que permite la proyección de nuevos datos sin necesidad de entrenar de nuevo. Esta proyección considera los prototipos como puntos fijos y se aplica una regla de actualización con el fin de mover el nuevo punto a la posición correspondiente. Otra ventaja de este método es que trabaja mejor con estructuras de datos que presentan una elevada curvatura en el espacio de entrada.

1.3.1.1.7 Curvilinear Distance Analysis (CDA)

El algoritmo Curvilinear Distance Analysis (CDA) es una modificación destacable del CCA, el cual usa distancias geodésicas en lugar de curvilíneas. Esto lo asemeja más a los mapas de Kohonen al establecer mejores conexiones con las neuronas vecinas, pero a diferencia de este, la malla no es regular. El CDA mejora el algoritmo anterior, en el sentido de que las distancias son más fiables, pero por el contrario es más difícil calcular la proyección de nuevos puntos y la visualización no es sobre una malla fija y compacta.

1.3.1.2 **Reducción de la dimensionalidad por preservación de la topología**

1.3.1.2.1 Locally Linear Embedding (LLE)

El Locally Linear Embedding (LLE) utiliza *conformal mapping* que es una transformación que trata de preservar los ángulos locales. La idea es reemplazar cada punto de los datos por una combinación lineal de los vecinos. A la hora de hacer la proyección de los datos al espacio de baja dimensión, el LLE asume que los datos tienen una relación geométrica que puede ser preservada mediante traslaciones, rotaciones y escalados al realizar la reducción de la dimensión. Una de las mayores ventajas de este algoritmo es que es simple y puede utilizar la información de interconexión entre todos los puntos, lo que garantiza una convergencia óptima del sistema. Se muesra ejemplo de LLE como herramienta de tratamiento de imágenes [19].

1.3.1.2.2 Laplacian Eigenmap (LE)

El Laplacian Eigenmap (LE) es un método similar al LLE, al considerar solamente los k vecinos más cercanos, pero considera el problema de vecindad de otra manera al utilizar operadores laplacianos para construir el grafo. El LE está basado en la minimización de la distancia entre puntos vecinos, estando esta minimización restringida con el fin de evitar la solución trivial. Esta técncia preserva localmente la topología de los datos y tiene una conexión natural al agrupamiento o clustering, es decir el LE tiende a agrupar la proyección de los datos. Una de las mayores ventajas de este método es la poca cantidad de parámetros que requiere para realizar la proyección, sin embargo, las proyecciones obtenidas por este método suelen ser muy pobres y suele ser más útil para clustering que para reducción de la dimensión.

1.3.1.2.3 Stochastic Neighbor Embedding (SNE)

El Stochastic Neighbor Embedding (SNE) es un método probabilístico para la proyección de los datos, o bien de sus disimilitudes. No trata de preservar la distancia entre puntos, como en el caso de escalado multidimensional, sino las probabilidades de que esos puntos sean vecinos, usando para ello una función gaussiana. De igual forma a las técnicas anteriores en el paper [20] es usado como tratamiento de imágenes.

1.3.1.2.4 Stochastic Neighbor Embedding (t-SNE)

Existe una mejora del SNE, denominada t-Distributed Stochastic Neighbor Embedding la cual utiliza una función simétrica del coste y función t-Student en lugar de la gaussiana para calcular la similuid entre dos puntos en el espacio de baja dimensión. Esto elimina el problema de amontonamiento de los puntos que se puede dar en el SNE al proyectar cuando existen muchos cercanos en el espacio de alta dimensión. Por el contrario, este método no funciona bien cuando la dimensión de proyección es superior a 3, por lo que solo suele ser útil para visualización. Además si la dimensión intrínseca de los datos es elevada tampoco obtiene buenos resultados.

1.3.1.2.5 Mapas autoorganizados

El Self-Organizing Map (SOM) es una red neuronal no supervisada y autoorganizada, basada en un proceso de

aprendizaje competitivo y cooperativo. Realiza una proyección no lineal, ordenada y suave un espacio de entrada multidimensional continuo sobre un espacio de salida discreto de baja dimensión visualizable.

El SOM se puede interpretar como un mapa que preserva la topología de los datos del espacio de entrada en una malla. Las coordenadas de las neuronas del SOM en el mapa reticular de baja dimensión lo convierten en un algoritmo de proyección de los datos del espacio de entrada, además, aproxima la función de densidad de los datos de entrenamiento.

El mapa autoorganizado presenta propiedades que lo hacen idóneo para la visualización y exploración de grandes volúmenes de datos pero uno de sus principales problemas es la ausencia de una función de energía. Esto impide que pueda realizar una minimización por descenso de gradiente para optimizar el algoritmo. Por otra parte, no se puede garantizar la convergencia del algoritmo cuando la dimensión de la malla de salida es mayor de uno.

En el paper [21] se usa la red neuronal como herramienta tanto de tratamiento como para predicción de la demanda eléctrica en un edificio bioclimático a corto plazo (horizonte 60 minutos) obteniendo un error promedio de 11.48%, siendo interesante el estudio de las mismas en gestión energética. También aparece otro ejemplo de uso en [22] certificación energética de edificios, como alternativa rápida y robusta para la predicción de los indicadores de demanda del edificio, considerando que la evaluación manual es exhaustiva y consume demasiado tiempo. El paper [23] utiliza la red neuronal como modelo de predicción de temperatura y humedad interior en un edificio. Los resultados obtenidos fueron buenos y atestiguaron con ello que la red neuronal puede ser usada con fiabilidad para predicción horaria.

1.3.2 Cuantización y clustering de vectores

La cuantización de vectores tiene como objetivo realizar una compresión de los datos en aquellas aplicaciones que toleran cierta distorsión. La idea fundamental detrás de la cuantización es aproximar los vectores de entrada utilizando un número pequeño de vectores prototipo.

El agrupamiento de datos o clustering, es similiar a la cuantización, diferenciándoles únicamente el objetivo, pues este busca agrupaciones interesantes del espacio de entrada. El resultado es directamente una serie de particiones Q de los datos del espacio de entrada, pudiendo identificarse cada región por un vector que representa el centroide del cluster q. Trabajar con los centroides obtenidos con una técnica de clustering es similar a hacerlo con los vectores prototipo en cuantización. El objetivo básico de la cuantización es la representación de los datos mientras que del clustering es su interpretación.

Una gran variedad de algoritmos de clustering [24] han surgido en los últimos años, los cuales se pueden clasificar en:

- Métodos jerárquicos
- Métodos basados en particiones
- Métodos basados en densidad
- Métodos basados en rejillas
- Métodos basados en modelos

1.4 PREDICCIÓN

1.4.1 Series temporales

La predicción de series temporales ha sido estudiada en una gran cantidad de campos, como pueden ser en finanzas para la predicción de los mercados, en la predicción de la transmisión de datos en redes de comunicaciones o en la predicción del consumo de energía eléctrica para la programación de la producción. Alguno de los métodos más utilizados y que más variantes tienen son los basados en técnicas de regresión que abarcan desde técnicas simples de regresión lineal a técnicas más complejas, como el Support Vector Regression (SVR).

1.4.1.1 Modelos paramétricos

Los modelos paramétricos son los métodos de predicción de series temporales más sencillos, realizándose la estimación por medio de una combinación de las muestras anteriores. Son los métodos más ampliamente utilizados en toda clase de campos de aplicación . Existen varios tipos de modelos:

- Modelo simple. Asume que la predicción en el instante t+1 es igual a la muestra en el instante:

$$x(t+1) = x(t)$$

- Modelo autorregresivo (AR). La muestra predicha es proporcional a una serie de N muestras anteriores:

$$x(t+1) = c + \sum_{i=0}^{N-1} x(t-1)\varphi_i + \varepsilon(t)$$

  donde φ_i son los parámetros del modelo, c es una constante y ε es un ruido blanco. Los parámetros se obtienen por medio de una regresión lineal.

- Media móvil (MA). La muestra predicha se calcula por medio de una media móvil del ruido blanco de un proceso. Esta predicción se basa en el hecho de que una muestra está correlacionada con su antecesora por medio del error del proceso:

$$x(t+1) = \sum_{i=0}^{N-1} \varepsilon(t-1)\theta_i$$

- Métodos ARMA. Utilizan una combinación de los métodos anteriores. Existen varios tipos de modelos ARMA, pero el más destacado y utilizado es el *Autoregressive Integrated Moving Average* (ARIMA). El modelo generalmente se refiere como ARIMA (p,d,q) donde p, d y q son los números enteros que indican el orden de la parte autorregresiva, integral y media móvil del modelo.

Dichas técnicas son aplicadas como complemento a la caracterización de consumos energéticos para obtener modelos de gestión energética en edificios [25] destacando además de la aplicación en dicho trabajo de las mismas, la revisión del estado del arte realizada sobre la temática.

1.4.2 Métodos basados en clustering

Estas técnicas se basan en particionar los datos de entrada y generar varios modelos de predicción, en lugar de un único modelo. De dicha forma responden al problema de la estacionalidad de los datos, apareciendo este cuando los datos presentan un comportamiento dependiente del tiempo y de la época del año. En el paper [26] se realiza una comparativa de dos de los principales métodos de cluster usados como técnica de predicción. El modelo de predicción usado en ellos son modelos idénticos al algoritmo de clustering, asigna a tu día un cluster en función de la similitud o distancia a los distintos Clusters obtenidos inicialmente. No obtiene conclusiones contundentes ni aclara la metodología de uso de dichas técnicas en combinación con otros modelos para predicción.

1.4.3 SOM

El SOM no sólo puede utilizarse como método de clustering para agrupar series temporales con la misma dinámica, sino que también puede utilizarse directamente para crear modelos de predicción a partir de series

temporales. Existen muchas aplicaciones que utilizan dicha técnica para predicción.

1.4.4 Técnicas de Manifold Learning

La predicción con estas técnicas se basa en reducir la dimensión de los datos de entrada para generar series temporales con la evolución de los datos originales en el espacio de baja dimensión.

La predicción con estas técnicas se compone de tres pasos:

1. Reducción de la dimensionalidad.
2. Predicción de la trayectoria.
3. Reconstrucción del espacio de alta dimensión.

La mayor ventaja de estos métodos es que permiten predecir simultáneamente gran cantidad de variables minimizando el número de predicciones necesarias, bien sea porque el número de variables se reduce o porque se pueden predecir un mayor número de muestras simultáneas. Al reducir el número de iteraciones se reduce el error de predicción generado. Sin embargo, se produce una acumulación de errores al tener el error de la reducción de la dimensionalidad, el error de predicción propiamente dicho y el error debido a la reconstrucción. Dicha técnica aparece en uso para el tratamiento de imágenes [18] .

2 LÍNEA BASE DE CONSUMOS

2.1 MODELIZACIÓN

Un punto fundamental de la línea base es la relación T-Q (tesis "Metodología de Caracterización Inversa para Edificios" de D. José Sánchez Ramos [27]). El producto el MCS aparece en un entorno destinado a la caracterización de consumos térmicos o eléctricos. El objetivo es obtener un modelo que permita obtener una línea base de la situación de referencia para verificar ahorros.

La filosofía deducida admite que en estos consumos y (t) puede existir una parte debida a excitaciones medibles u (t) (Componente Determinista), y otra no medible v (t) (Componente Estocástica). (Ref.: J. McLellan-Fall 2005).

La siguiente imagen describe las implicaciones que tiene la fase anterior en la caracterización del consumo energético de un sistema, siendo deseable que la componente dominante del modelo sea la determinista (en la figura inferior denotada por sistema).

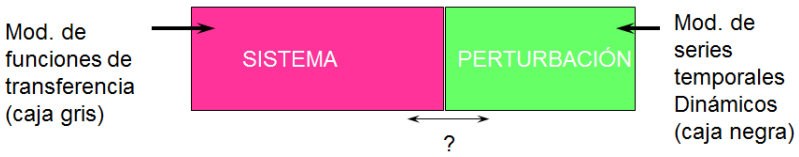

Figura 13. Visión general de los componentes del consumo

La componente determinista debe ser dominante en la variable objetivo, ya que encerrará toda la información física y cierta del sistema a caracterizar. Es por esto que la parte estocástica, la perturbación, debe limitarse, es decir, debe caracterizarse el control y autoaprendizaje del modelo con el tiempo. Ambas partes son los parámetros clave para acotar la validez y aplicabilidad del modelo.

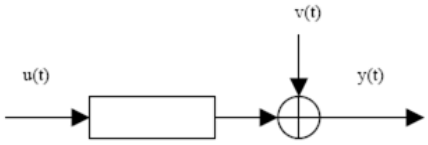

Figura 14. Diagrama de bloques de modelo integrado de consumo

Donde:

- U(t) excitaciones medibles con incertidumbre
- V(t) perturbaciones: invarianza, excitaciones no medibles, sucesos no explicables…
- Y(t) variable objetivo medible

Ahora bien, no todos los consumos responden a una realidad física caracterizable, sino que en muchos casos son implicaciones del comportamiento humano, como el consumo de ascensores del edificio, o de fuerza. Pero en el caso del consumo de climatización, en edificios terciarios controlados de forma centralizada, sí depende principalmente de las condiciones climáticas y las fuentes internas. Es decir el consumo de climatización si suele estar ligado a las cargas térmicas del edificio. Es por este motivo, y en el contexto en el que la tesis se situa, que el objetivo del model integrado es caracerizar mediante la componente determinista la parte física del consumo de climatización (información vinculada a las cargas térmicas del edificio), y con la componente estocástica el resto del consumo no explicable por la componente física caracterizada.

Con respecto a la síntesis realizada del trabajo doctoral, el trabajo que aquí se expone propone sustituir la perturbación por algoritmo de clustering. Este algoritmo tiene como variables de entrada los datos conocidos de meteorología, consumo y calendario laboral, y trabaja en una base diaria.

Por consiguiente, y de forma simplificada se puede entender que el consumo de referencia o línea base puede escribirse como función del resultado del algoritmo de clustering y de las variables climáticas. Este algoritmo devolvería una tipificación de consumos en forma de grupos de días que cumplen una serie de condiciones de

consumo, temperatura exterior, radiación, día de la semana y día festivo/laboral. El consumo clúster tiene una fuerte componente matemática pero también tiene relaciones físicas, por lo que con respecto a la tesis se realiza un mix de la componente determinista y perturbación para la línea base propuesta. Esta resulta:

$$C_{BS}(d) = C_{CLÚSTER}(d) + \Delta C(d)$$

Dónde $C_{CLÚSTER}(d) = F(distancia)$.

Esta distancia a su vez depende del tipo de día, de la temperatura media exterior y de la integral de radiación global horizontal incidente; y está referida a la distancia del punto hasta el centroide de ese clúster. Este centroide queda definido con los valores medios de los días englobados en ese clúster durante el proceso de clustering. Cada día tiene asociado un clúster de referencia, es decir, el clúster cuya distancia al día a estudio es la mínima.

El incremento de consumo debido a la variación climática se obtiene a partir de los valores de referencia que definen el centroide de ese clúster.

$$\Delta C(d) = \sum_{i=0}^{M} a_i \cdot \Delta\overline{T_{EXT}(d-i)} + \sum_{i=0}^{M} b_i \cdot \Delta RAD(d-i) + \sum_{j=1}^{N} d_i \cdot \Delta C(d-j)$$

Agrupando términos, una primera propuesta para la línea base resulta:

$$C_{BS}(d) = \sum_{i=0}^{M} a_i \cdot \overline{T_{EXT}(d-i)} + \sum_{i=0}^{M} b_i \cdot RAD(d-i) + \sum_{i=1}^{M} c_i \cdot C_{CLÚSTER}(d-i) + \sum_{j=1}^{N} d_i \cdot C_{BS}(d-j)$$

Dónde el consumo asociado al tipo de día, consumo de clúster, se introduce como una variable independiente más en el modelo agrupado y por tanto queda vinculado como otro tipo de numerador. Esta nueva variable puede interpretarse como un índice de actividad de ese día.

Es importante destacar que la temperatura media exterior puede ser sustituida por los grados día, pero en este caso hay que fijar una consigna. Es por este motivo que se prefiere trabajar con una temperatura media diaria, aunque tal y como se ha dicho anteriormente ambas variables están relacionadas.

Véase el estudio de determinación del orden y forma del modelo de baseline en el capítulo 3 para completar la información aquí citada.

2.2 CONSUMO CLÚSTER

El consumo de cluster, se define como un consumo asociado al tipo de día. Esta nueva variable, como se ha comentado en la modelización, puede interpretarse como un índice de actividad de ese día.

El consumo de cluster viene definido por tres decisiones importantes:

- Decisión 1-Método de clustering
- Decisión 2-Índices de evaluación
- Decisión 3-Medidas de distancia

En el presente capítulo se describe cada una de ellas con exactitud, mostrando la gran diversidad de posibiladades incluidas en cada una de las decisiones anteriores y se propone un algoritmo que ayuda a llevar a cabo todo el proceso hasta la obtención de la linea base de consumos.

2.2.1 MEDIDAS DE DISTANCIA

Una vez considerado que el objetivo del análisis de cluster consiste en encontrar agrupaciones naturales de un conjunto de individuos de la muestra, es necesario definir qué se entiende por agrupaciones naturales y, por tanto, con arreglo a qué criterio se puede decir que los grupos son más o menos similares.

Dada una matriz de datos mxn tratada como m vectores filas (x_1, x_2, …, x_m) las distintas distancias [28] entre los vectores x_s y x_t se definen a continuación:

2.2.1.1 Distancia euclídean

$$d_{st}^2 = (x_s - x_t)(x_s - x_t)'$$

Obsérvese que la distancia euclidiana es un caso especial de la métrica de minkowski, donde p=2.

2.2.1.2 Distancia seuclídean

Standardized Euclidean Distance

$$d_{st}^2 = (x_s - x_t)V^{-1}(x_s - x_t)'$$

Donde V es la matriz diagonal nxn cuyo j-ésimo elemento de la diagonal es $S(j)^2$, siendo S el vector de desviaciones estándar.

2.2.1.3 Distancia mahalanobis

$$d_{st}^2 = (x_s - x_t)C^{-1}(x_s - x_t)'$$

Donde C es la matriz de covarianza.

2.2.1.4 Distancia cityblock

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - x_{tj}|$$

Obsérvese que la distancia euclidiana es un caso especial de la métrica de minkowski, donde p=1

2.2.1.5 Distancia minkowski

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} |x_{sj} - x_{tj}|^p}$$

Obsérvese que para p=1 la métrica de minkowski es igual a cityblock, para el caso p=2 es igual a la Euclidian y para p=∞ se particulariza en la métrica o distancia de chebychev.

2.2.1.6 **Distancia chebychev**

$$d_{st} = max_j\{|x_{sj} - x_{tj}|\}$$

Obsérvese que la distancia chebychev es un caso especial de la métrica de minkowski, donde p=∞.

2.2.1.7 **Distancia cosine**

$$d_{st} = 1 - \frac{x_s x_t'}{\sqrt{(x_s x_s')(x_t x_t')}}$$

2.2.1.8 **Distancia correlation**

$$d_{st} = 1 - \frac{(x_s - \bar{x_s})(x_t - \bar{x_t})'}{\sqrt{(x_s - \bar{x_s})(x_s - \bar{x_s})'}\sqrt{(x_t - \bar{x_t})(x_t - \bar{x_t})'}}$$

Donde:

$$\bar{x_s} = \frac{1}{n}\sum_j x_{sj}$$

$$\bar{x_t} = \frac{1}{n}\sum_j x_{tj}$$

2.2.1.9 **Distancia hamming**

$$d_{st} = (\#(x_{sj} \neq x_{tj})/n)$$

2.2.1.10 **Distancia jaccard**

$$d_{st} = \frac{\#\left[(x_{sj} \neq x_{tj}) \cap (x_{sj} \neq 0) \cup (x_{tj} \neq 0)\right]}{\#\left[(x_{sj} \neq 0) \cup (x_{tj} \neq 0)\right]}$$

2.2.1.11 **Distancia spearman**

$$d_{st} = 1 - \frac{(r_s - \bar{r_s})(r_t - \bar{r_t})'}{\sqrt{(r_s - \bar{r_s})(r_s - \bar{r_s})'}\sqrt{(r_t - \bar{r_t})(r_t - \bar{r_t})'}}$$

r_{sj} es el rango de x_{sj} tomado para calcular tiedrank.

r_s y r_t son los vectores de rango de coordenadas x_s y x_t.

2.2.2 ÍNDICES DE EVALUACIÓN

Para comparar la eficacia del agrupamiento en función del número de clusters seleccionados para su partición existen una serie de índices que permiten comparar los resultados entre sí y seleccionar el número óptimo.

A continuación se muestran los índices de evaluación más usados.

2.2.2.1 Índice Davies-Bouldin

El índice de Davies Bouldin [29] es una métrica para evaluar el buen funcionamiento de los algoritmos de clustering. La formula del siguiente índice se muestra a continuación:

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Donde n es el número de clusters, c_x denota el centroide del cluster x, σ_x es la distancia media de todos los elementos del cluster x al centroide c_x, y $d(c_i, c_j)$ es la distancia entre los centroides c_i y c_j.

El objetivo de los algoritmos de clustering es producir agrupamientos con baja distancia dentro del mismo cluster, y altas distancias entre los clusters. Por lo tanto, estos algorimos producirán un valor bajo del índice de Davies Bouldin.

El valor $max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right)$ representa el peor caso para el cluster i. La solución óptima es aquella que tiene el índice de Davies Bouldin más bajo.

2.2.2.2 Índice Calinski-Harabasz

El índice Calinski-Harabasz [30] se define como:

$$VRC_k = \frac{SS_B}{SS_W}\, x\, \frac{(N-k)}{(k-1)}$$

Donde SS_B es la varianza entre clusters y SS_W es la varianza entre elementos de un mismo cluster, k es el número de clusters y N el número de observaciones.

Los clusters bien definidos tienen una gran varianza entre clusters SS_B y una pequeña variacion dentro del cluster SS_W. Cuanto mayor sea VRC_k , mejor será la partición de los datos.

Para determinar el número óptimo de grupos hay que maximizar VRC_k con respecto a k. El número óptimo de clusters es la solución con el valor del índice de Calinski-Harabasz más alto.

2.2.2.3 Índice Silhouette

Este índice calcula la anchura del contorno para cada muestra, la media del contorno para cada cluster y el contorno para todo el conjunto de datos. Utilizando este índice, cada grupo puede ser representado por su contorno, el cual se basa en la comparación del empaquetamiento y separación.

El índice de Silhouette [14] viene dado por la siguiente expresión:

$$Sil\ (i) = \frac{a(i) - b(i)}{\max(a(i), b(i))}$$

Donde a(i) es la distancia media del punto i a todos los puntos del mismo cluster y b(i) es el mínimo de las distancias medias entre entre el punto i y todos los puntos de otros clusters.

2.2.3 MÉTODOS JERÁRQUICOS DE ANÁLISIS DE CLUSTERS

2.2.3.1 Introducción

Los llamados métodos jerárquicos [31]tienen por objetivo agrupar clusters para formar uno nuevo o bien para separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se dividen en aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes.

1. Los métodos aglomerativos, también conocidos como ascendentes, comienzan el análisis con tantos grupos como individuos hayas. A partir de estas unidades iniciales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado.
2. Los métodos disociativos, también llamados descendentes, constituyen el proceso inverso al anterior. Comienzan con un conglomerado con todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso, se obtienen tantos grupos como datos han sido tratados.

Los métodos jerárquicos permiten la construcción de un árbol de clasificación, que recibe el nombre de dondograma, en el cual se puede seguir de forma gráfica el procedimiento de unión seguido, mostrando qué grupos se van uniendo, en qué nivel en concreto lo hacen, así como el valor de la medida de asociación entre los grupos cuando éstos se agrupan.

2.2.3.2 Métodos jerárquicos aglomerativos

A continuación se presentan algunas estrategias que pueden ser empleadas a la hora de unir los clústers en las diversas etapas o niveles de procedimiento jerárquico. Ninguno de estos procedimientos proporciona una solución óptima para todos los problemas que se pueden plantear, ya que es posible llegar a distintos resultados según el método elegido. El buen criterio del investigador, el conocimiento del problema planteado, y la experiencia sugerirán el método más adecuado. De esta forma, es conveniente, siempre, usar varios métodos con el fin de contrastar los resultados obtenidos y sacar conclusiones, tanto como si hubiera coincidencias en los resultados obtenidos con métodos distintos como si no las hubiera.

2.2.3.2.1 Estrategia de la distancia minima o similitud máxima

Esta estrategia recibe el nombre en la escritura anglosajona de amalgamamiento simple (single linkage). En este método se considera que la distancia o similitud entre dos Clusters viene dada, respectivamente, por la mínima distancia o máxima similitud entre sus componentes.

Así, si tras efectuar la etapa K-ésima tenemos ya formados n-K Clusters, la distancia entre los Clusters Ci (con ni elementos) y Cj (con nj elementos) sería:

$$d(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_j \in C_j}}{Min}\{d(x_l, x_m)\} \quad l = 1, \dots, n_i \ ; \ m = 1, \dots, n_j$$

Mientras que la similitud, si estuviéramos una medida de tal tipo, entre los dos clusters sería:

$$s(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_j \in C_j}}{Max}\{s(x_l, x_m)\} \quad l = 1, \dots, n_i \ ; \ m = 1, \dots, n_j$$

Con ello, la estrategia seguida en el nivel k+1 será:

1- En el caso de emplear distancias, se unirán los clusters C_i y C_j si:

$$d(C_i, C_j) = \underset{\substack{i1,j1=1,\dots,n-k \\ i1 \neq j1}}{Min}\{d(C_{i1}, C_{j1})\} =$$

$$= \underset{\substack{i1,j1=1,\dots,n-k \\ i1 \neq j1}}{Min}\left\{\underset{\substack{x_l \in C_{i1} \\ x_j \in C_{j1}}}{Min}\{d(x_l, x_m)\}\right\} \quad l = 1, \dots, n_{i1} \ ; \ m = 1, \dots, n_{j1}$$

2- En el caso de emplear similitudes, se unirán los clusters C_i y C_j si:

$$s(C_i, C_j) = Max_{\substack{i1,j1=1,\dots,n-k \\ i1 \neq j1}} \{s(C_{i1}, C_{j1})\} =$$

$$= Max_{\substack{i1,j1=1,\dots,n-k \\ i1 \neq j1}} \left\{ Max_{\substack{x_l \in C_{i1} \\ x_j \in C_{j1}}} \{s(x_l, x_m)\} \right\} \quad l = 1, \dots, n_{i1} \ ; \ m = 1, \dots, n_{j1}$$

Donde, como es natural, sigue la norma general de maximizar similitudes o minimizar distancias.

2.2.3.2.2 Estrategia de la distancia máxima o similitud mínima

En este método, también conocido como amalgamamiento completo (complete linkage), se considera que la distancia o similitud entre dos clusters hay que medirla atendiendo a sus elementos más dispares, o sea, la distancia o similitud entre clusters viene dada, respectivamente, por la máxima distancia o la mínima solicitud entre sus componentes.

Así pues, al igual que la estrategia anterior, si estamos ya en la etapa k-ésima, y por tanto hay ya formado n-k clusters, la distancia y similitud entre los clusters C_i y C_j (con n_i y n_j elementos respectivamente), serán:

$$d(C_i, C_j) = Max_{\substack{x_l \in C_i \\ x_j \in C_j}} \{d(x_l, x_m)\} \quad l = 1, \dots, n_i \ ; \ m = 1, \dots, n_j$$

$$s(C_i, C_j) = Min_{\substack{x_l \in C_i \\ x_j \in C_j}} \{s(x_l, x_m)\} \quad l = 1, \dots, n_i \ ; \ m = 1, \dots, n_j$$

Y con ello, la estrategia seguida en el siguiente nivel k+1, será:

1- En el caso de emplear distancias, se unirán los clusters C_i y C_j si:

$$d(C_i, C_j) = Min_{\substack{i1,j1=1,\dots,n-k \\ i1 \neq j1}} \{d(C_{i1}, C_{j1})\} =$$

$$= Min_{\substack{i1,j1=1,\dots,n-k \\ i1 \neq j1}} \left\{ Max_{\substack{x_l \in C_{i1} \\ x_j \in C_{j1}}} \{d(x_l, x_m)\} \right\} \quad l = 1, \dots, n_{i1} \ ; \ m = 1, \dots, n_{j1}$$

2- En el caso de emplear similitudes, se unirán los clusters C_i y C_j si:

$$s(C_i, C_j) = Max_{\substack{i1,j1=1,\dots,n-k \\ i1 \neq j1}} \{s(C_{i1}, C_{j1})\} =$$

$$= Max_{\substack{i1,j1=1,\dots,n-k \\ i1 \neq j1}} \left\{ Min_{\substack{x_l \in C_{i1} \\ x_j \in C_{j1}}} \{s(x_l, x_m)\} \right\} \quad l = 1, \dots, n_{i1} \ ; \ m = 1, \dots, n_{j1}$$

2.2.3.3 **Estrategia de la distancia, o similitud, promedio no ponderada (Weighted Aritmetic Average)**

En esta estrategia, la distancia o similitud del cluster C_i con el C_j se obtiene con la media aritmética entre la distancia, o similitud, de los componentes de dichos clusters.

Así, si el cluster C_i (con n_i elementos) está compuesto, a su vez, por dos clusters C_{i1} y C_{i2} (con n_{i1} y n_{i2} elementos respectivamente), y el cluster C_j posee n_j elementos, la distancia, o similitud, entre ellos se calcula como:

$$d(C_i, C_j) = \frac{d(C_{i1}, C_j) + d(C_{i2}, C_j)}{2}$$

Notemos que en este método no se tiene en cuenta el tamaña de ninguno de los clusters involucrados en el cálculo, lo cual significa que concede igual importancia a la istancia $d(C_{i1}, C_j)$ que a la distancia $d(C_{i2}, C_j)$.

2.2.3.3.1 Estrategia de la distancia, o similitud, promedio ponderada (Unweighted Aritmetic Average)

Se considera que la distancia, o similitud, entre dos clusters, viene definida por el promedio ponderado de las distancias, o similitudes, de los componentes de un cluster respecto del otro.

Sea dos clusters, C_i y C_j; supongamos que el cluster C_i está formado, a su vez, por otros dos clústeres C_{i1} y C_{i2}, con n_{i1} y n_{i2} elementos respectivamente. Sea n_i=n_{i1}+ n_{i2} el número de elementos de C_i y n_j el número de elementos que componen C_j. Entonces, en términos de distancias (igual puede hacerse para similitudes) , la distancia promedio ponderada sería, notando $x_i \in C_i, x_{i1} \in C_{i1}, x_{i2} \in C_{i2}, x_j \in C_j$

$$d(C_i, C_j) = \frac{1}{(n_{i1} + n_{i2})n_j} \sum_{i=1}^{n_{i1}+n_{i2}} \sum_{j=1}^{n_j} d(x_i, x_j) =$$

$$= \frac{1}{(n_{i1} + n_{i2})n_j} \sum_{i1=1}^{n_{i1}} \sum_{j=1}^{n_j} d(x_{i1}, x_j) + \frac{1}{(n_{i1} + n_{i2})n_j} \sum_{i2=1}^{n_{i2}} \sum_{j=1}^{n_j} d(x_{i2}, x_j) =$$

$$= \frac{n_{i1}}{(n_{i1} + n_{i2})n_{i1}n_j} \sum_{i1=1}^{n_{i1}} \sum_{j=1}^{n_j} d(x_{i1}, x_j) + \frac{n_{i2}}{(n_{i1} + n_{i2})n_{i2}n_j} \sum_{i2=1}^{n_{i2}} \sum_{j=1}^{n_j} d(x_{i2}, x_j) =$$

$$= \frac{n_{i1}}{(n_{i1} + n_{i2})} d(C_{i1}, C_j) + \frac{n_{i2}}{(n_{i1} + n_{i2})} d(C_{i2}, C_j) =$$

$$= \frac{n_{i1}d(C_{i1}, C_j) + n_{i2}d(C_{i2}, C_j)}{(n_{i1} + n_{i2})}$$

Con lo cual la distancia $d(C_i, C_j)$ es el promedio ponderado de las distancias de cada uno de los dos clústers previos, C_{i1} y C_{i2} , con respecto al cluster C_j.

2.2.3.3.2 Método basado en centroides

En estos métodos, la semejanza entre dos clusters viene dada por la semejanza entre sus centroides, esto es, los vectores de medias entre las variables medidas sobre los individuos del cluster.

Entre ellos distinguiremos dos:

1- Método del centroide ponderado, en el que el tamaño de los clusters son considerados a la hora de efectuar los cálculos.

2- Método del centroide no ponderado, o método de la mediana, en el cual el tamaño de los clusters no son considerados.

Véase cada uno de ellos por separado:

1- En cuanto al primero de ellos y centrándonos en la distancia euclídea al cuadrado, supongamos que pretendemos medir la distancia entre los clusters C_j (compuesto por n_j elementos) y C_i (compuesto a su vez por dos clusters, C_{i1} y C_{i2}, con n_{i1} y n_{i2} elementos respectivamente). Sean m^j, m^{i1} y m^{i2} los centroides de los clusters anteriormente citados (obviamente, esos centroides son vectores n dimensionales).

   Así, el centroide del cluster C_i vendrá dado en notación vectorial por:

$$m^i = \frac{n_{i1}m^{i1} + n_{i2}m^{i2}}{n_{i1} + n_{i2}}$$

Cuyas componentes serán:

$$m_l^i = \frac{n_{i1}m_l^{i1} + n_{i2}m_l^{i2}}{n_{i1} + n_{i2}} \quad l = 1, \dots, n$$

Con ello, la distancia euclídea al cuadrado entre los cluster C_i y C_j vendrá dada por:

$$d_2^2(C_i, C_j) = \sum_{l=1}^{n} (m_l^j - m_l^i)^2 =$$

$$= \frac{n_{i1}}{n_{i1} + n_{i2}} d_2^2(C_{i1}, C_j) + \frac{n_{i2}}{n_{i1} + n_{i2}} d_2^2(C_{i2}, C_j) - \frac{n_{i1}n_{i2}}{(n_{i1} + n_{i2})^2} d_2^2(C_{i1}, C_{i2})$$

Nótese que la relación anterior se establece para el caso particular de la distancia euclídea. No obstante, dicha relación se sigue verificando si la distancia empleada viene definida a partir a partir de una norma que proceda de un producto escalar.

2- Una desventaja del procedimiento anterior estriba en que si los tamaños n_{i1} y n_{i2} de los componentes C_i son muy diferentes entre sí, se corre el peligro de que el centroide de dicho cluster, m^i, esté influenciado excesivamente por el componente con tamaño superior y las cualidades del grupo pequeño no se tengan prácticamente en cuenta.

Así la estrategia de la distancia mediana es análoga a la anterior y, por tanto, goza de sus mismas características. Así, si estamos hablando de distancias, la distancia entre el cluster C_i y C_j viene dada por:

$$d(C_i, C_j) = \frac{1}{2}[d(C_{i1}, C_j) + d(C_{i2}, C_j)] - \frac{1}{4}d(C_{i1}, C_{i2})$$

Y si hablamos de similitudes:

$$s(C_i, C_j) = \frac{1}{2}[s(C_{i1}, C_j) + s(C_{i2}, C_j)] + \frac{1}{4}[1 - s(C_{i1}, C_{i2})]$$

Notemos que una característica de los métodos basados en el centroide y sus variantes es que el valor de similaridad o la distancia asociada a los clusters enlazados puede aumentar o disminuir de una etapa a otra. Por ejemplo, cuando la medida es una distancia, la distancia entre los centroides puede ser menor que la de otro par de centroides unidos en una etapa anterior. Esto puede ocurrir ya que los centroides, en cada etapa, pueden cambiar de lugar. Este problema puede llevar a que el resultado del dendograma sea difícil de interpretar.

2.2.3.3.3 Método de Ward

El método de Ward es un procedimiento jerárquico en el cual, en cada etapa, se unen los dos clústers para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del cluster.

Notemos por:

- x_{ij}^k al valor de la j-ésima variable sobre el i-ésimo individuo del k-ésimo cluster, suponiendo que dicho cluster posee n_k individuos.
- m^k al centroide del cluster k, con componentes m_j^k.
- E_k a la suma de cuadrados de los errores del cluster k, o sea, la distancia euclídea al cuadrado entre cada individuo del cluster k a su centroide.

$$E_k = \sum_{i=1}^{n_k}\sum_{j=1}^{n}(x_{ij}^k - m_j^k)^2 = \sum_{i=1}^{n_k}\sum_{j=1}^{n}(x_{ij}^k)^2 - n_k\sum_{j=1}^{n}(m_j^k)^2$$

- E a la suma de los cuadrados de los errores para todos los clusters, o sea, si suponemos que hay h clusters:

$$E = \sum_{k=1}^{h} E_k$$

El proceso comienza con m clusters, cada uno de los cuales está compuesto por un solo individuo, por lo que cada individuo coincide con el centro del cluster y por lo tanto en este primer paso se tendrá que $E_k = 0$ para cada cluster y con ello E=0. El objetivo del método de Ward es encontrar en cada etapa aquellos dos clusters cuya unión proporcione el menor incremento en la suma total de errores, E.

Supongamos ahora que los clusters C_p y C_q se unen resultando un nuevo cluster C_t . Entonces el incremento de E será:

$$\Delta E_{pq} = E_t - E_p - E_q = \frac{n_p n_q}{n_t}\sum_{j=1}^{n}(m_j^p - m_j^q)^2$$

Así el menor incremento de los errores cuadráticos es proporcional a la distancia euclídea al cuadrado de los centroides de los clusters unidos. La suma E no es decreciente y el método, por lo tanto, no presenta los problemas de los métodos del centroide anteriores.

Al igual que en el método anterior se puede demostrar que la relación anterior se sigue verificando para una distancia que venga definida a partir de una norma que proceda de un producto escalar.

2.2.3.4 **Métodos jerárquicos disociativos**

Como se comentó anteriormente, los métodos disociativos, constituyen el proceso inverso a los aglomerativos. Comienzan con un conglomerado que englobe a todos los casos tratados y, a partir del caso inicial, a partir de sucesivas iteraciones, se van formando grupos cada vez menores. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

En cuanto a la clasificación de estos métodos se puede decir que la filosofía de los métodos aglomerativos puede mantenerse para este otro tipo de procedimientos en lo que concierne a la forma de calcular la distancia entre los grupos, si bien, como es lógico, al partir de un grupo único hay que subdividir, se seguirá la estrategia de maximizar las distancias, o minimizar las similaridades, puesto que buscamos ahora los individuos menos similares para separarlos del resto del conglomerado.

Esta clase de métodos son esencialmente de dos tipos:

1- Monetéticos, los cuales dividen los datos sobre la base de un solo atributo y suelen emplearse cuando los datos son de tipo binario.

2- Politéticos, cuyas divisiones se basan en los valores tomados por todas las variables.

Esta clase de procedimiento es bastante menos popular que los ascendentes, pues la literatura sobre ellos no es muy extensa.

2.2.3.5 **Coeficiente de correlación cofenético**

Los métodos jerárquicos imponen una estructura sobre los datos y es necesario considerar si es aceptable o si se producen distorsiones inaceptables en las relaciones originales. El método más usado para verificar este hecho, o sea, para ver la relación entre el dendrograma y la matriz de proximidades original, es el coeficiente de correlación cofenético, el cual es simplemente la correlación entre los $\frac{n(n-1)}{2}$ elementos de la parte superior de la matriz de proximidades observada y los correspondientes en la llamada matriz cofenética, C, cuyos elementos, C_{ij} , se definen como aquellos que determinan la proximidad entre los elementos i y j cuando éstos se unen en

el mismo cluster.

Así, si tras el empleo de varios procedimientos de clusters distintos, éstos conducen a soluciones parecidas, surge de qué método eligiremos como definitivo. La respuesta la da el coeficiente cofenético, ya que aquel método que tenga un coeficiente cofenético más elevado será aquel que presente una menor distorsión en las relaciones originales existentes entre los elementos en estudio.

2.2.3.6 **Número de clusters a determinar**

Con frecuencia, cuando se emplean técnicas de cluster jerárquicas, el investigador no está interesado en la jerarquía completa sino en un subconjunto de particiones obtenidas a partir de ella. Las particiones se obtienen cortando el dendrograma o seleccionando una de las soluciones en la sucesión encajada de clusters que comprende la jerarquía.

Este paso fundamental está entre los problemas que están todavía sin resolver. La técnica usada se basa en cortar el dendrograma de forma subjetiva tras visualizarlo. Obviamente este procedimiento no es nada satisfactorio ya que está generalmente sesgado por la opinión que el investigador posea sobre los datos

2.2.4 MÉTODOS NO JERÁRQUICOS DE ANÁLISIS DE CLUSTERS

2.2.4.1 Introducción

Los métodos jerárquicos, para un conjunto de m individuos, parten de m clusters de un miembro cada uno hasta construir un solo cluster de m miembros (métodos aglomerativos) o viceversa (métodos disociativos).

La idea central de la mayoría de los procedimientos no jerárquicos es elegir alguna partición inicial de individuos y después intercambiar los miembros de estos clusters para obtener una partición mejor.

Los diversos algoritmos [32] existentes se diferencian sobre todo en lo que se entiende por una partición mejor y en los métodos que deben usarse para conseguir mejoras. La idea general de estos métodos es muy similar a la señalada en los algoritmos descendentes en más de un paso empleados en la optimización sin restricciones en programación no lineal. Tales algoritmos empiezan con un punto inicial y generan una secuencia de movimientos de un punto a otro hasta que se encuentra un óptimo local de la función objetivo.

Los métodos estudiados ahora comienzan con una partición inicial de los individuos en grupos o bien con un conjunto de puntos iniciales sobre los cuales pueden formarse los clusters. En muchos casos, la técnica para establecer una partición inicial es parte del algoritmo cluster, aunque estas técnicas usualmente son proporcionadas por sí solas más que como una parte del algoritmo cluster.

2.2.4.2 Métodos basados en particiones

Dado un conjunto de objetos, n se construye k particiones o grupos con k<n. Se asignan los puntos a particiones y se refiena iterativamente el particionado mediante el cambio de ubicación de los objetos.

Cada objeto solo puede pertenecer a una partición y cada partición solo puede tener un objeto.

En general, el criterio para un buen particionado es que los objetos del mismo cluster sean cercanos y los pertenecientes a clusters diferentes sean esencialmente diferentes.

Los más populares son k-medias y k-medianas.

Otros algoritmos usados en dichos métodos son: Neural Gas, CLARA y CLARANS.

2.2.4.3 Métodos basados en densidad

Se agrupan objetos mientras su densidad (número de objetos) en la "vecindad" esté dentro de un cierto umbral.

Los métodos basados en la distancia tienden a funcionar bien con clusters esféricos y mal con clusters de otras formas. Para solucionar este problema otros métodos han desarrollado el concepto de densidad, el cual permite descubrir clusters con formas arbitrarias.

La idea subyacente es hacer un cluster siempre y cuando del cluster exceda de un umbral.

Algoritmos basados en densidad: DBSCAN, DENCLUE

2.2.4.4 Métodos basados en rejillas

Dicho método se basa en dividir el espacio en rejillas a diferentes niveles. Estos métodos segmentan el espacio en un conjunto finite de celdas. Todas las operaciones son ejecutadas sobre celdas.

Si el volumen de datos es grande, estos métodos mantienen tiempos de ejecución moderados, que dependerán más del número de celdas que del número de objetos.

Algoritmos basados en rejillas: STING, CLIQUE

2.2.4.5 Métodos basados en modelos

Método basado en un modelo para cada cluster que mejor ajuste los datos de ese grupo. En este caso, se construyen clusters de funciones de densidad basadas en modelos estadísticos.

La capacidad de interpretación de los datos está limitada por el modelo estadístico utilizado.

Si los objetos quedan fuera de lo que se espera para el modelo, automáticamente se etiquetan como outliers.

Algoritmos basados en modelos: Gaussian Mixture Model, COWEB, Autoclass

2.2.4.6 **K-means**

Es un método clásico de clustering basado en el agrupamiento por partición, dividiendo el conjunto de datos de entrada en un número de grupos determinado. Este algoritmo busca de forma iterativa el número óptimo de grupos que realiza una partición de minima varianza en el espacio de datos de entrada, minimizando la función de coste E:

$$E_{KM} = \sum_{k=1}^{N} \sum_{x_i \in Q_k} \|x_i - q_k\|^2$$

Donde q_k es el centroide del cluster Q_k

Algoritmo k-means

1- Inicializar todos los centroides q_k de forma aleatoria o aplicando conocimiento a priori.

2- Asignar cada dato de entrada x_i al grupo Q_l más próximo, es decir, $x_i \in Q_l$ si

$\|x_i - q_l\| < \|x_i - q_k\|$, i=1,2, … , K $i \neq l$

3- Calcular los centroides, teniendo en cuenta la partición actual. Esto es $q_k = \frac{1}{|Q_k|} \sum_{x_i \in Q_k} x_i$

4- Repetir los pasos 2 y 3 hasta que los centroides dejen de modificarse definitivamente.

Uno de los inconvenientes de este método es que hay que proporcionar de antemano el número de grupos que se quieren obtener. Por ello el algoritmo realiza varias iteraciones, intentando minimizar el error.

Otra de las limitaciones del método es la separación de grupos con formas no convexas. A esto hay que añadir el gran coste computacional cuando el número de datos es elevado.

2.2.4.7 **Neural Gas**

El Neural Gas también es un modelo de agrupamiento por partición. Esta técnica es muy similar al SOM ya que ambas preservan la topología, pero sin la imposición de una estructura de vecindad predefinida, lo que permite trabajar mejor en espacios de entrada complicados.

El Neural Gas permite obtener un número finito K de vectores prototipo, preservando la función de densidad de probabilidad del espacio de entrada. La actualización no se aplica en función de la distancia al vector ganador sino en función del orden de distancia de los vecinos $(q_{i0}, q_{i1}, \dots, q_{ik-1})$ siendo q_{i0} el más cercano al dato x. La adaptación de los vectores prototipo se calcula en cada caso como:

$$q_{ik}(t+1) = q_{ik}(t) + \alpha(t) e^{\frac{-k}{\lambda}} [x(t) - q_{ik}(t)]$$

Donde $\alpha(t)$ es la tasa de aprendizaje, λ es la función de vencindad y el índice k es el orden del vector prototipo q_i con respecto al dato x.

Algoritmo Neural Gas

1- Inicializar los centroides q_i de forma aleatorio o aplicando conocimiento a priori.

2- Calcular el orden de los prototipos con respecto a cada dato de entrada.

3- Aplicar la adaptación de vectores prototipo a cada uno de los datos.

4- Repetir los dos pasos anteriores el número de iteraciones que se haya determinado.

Este método suele obtener mejores resultados de convergencia que el k-means.

2.2.4.8 **DBSCAN**

DBSCAN es un método de clustering basado en densidad, hace crecer regiones con suficiente alta densidad en grupos y descubre grupos con forma arbitraria. Estos grupos están separados por regiones de baja densidad de objetos (ruido).

Los parámetros esenciales del **a**lgoritmo son el radio (ϵ) y el número de puntos mínimo (MinPts).

Se denominan puntos nucleares a aquellos que estan en un $\epsilon - vecindario$ y contienen el mínimo número de puntos.

Los puntos fronterizos son aquellos que tienen menos de MinPts puntos dentro de su vecindario, pero están en la vecindad de un punto nuclear.

Los puntos ruidosos son aquellos que no caen en ninguna de las dos categorías anteriores.

Algoritmo DBSCAN

1- DBSCAN busca clusters comprobando en el $\epsilon - vecindario$ de cada punto.

2- Si en el vecindario de un punto hay más de MinPts, un nuevo cluster con dicho punto como núcleo es creado.

3- DBSCAN iterativamente recoloca los puntos que son directamente alcanzables desde estos objetos núcleo.

4- El proceso termina cuando no se pueden añadir nuevos puntos a ningún cluster.

DBSCAN encuentra clusters no separables linelmente y no depende de las condiciones de inicio. También destacar como ventaja la no nocesidad de imponer un número fijo de clusters a obtener.

Por otro lado dado que DBSCAN está optimizado para que genere clusters de igual densidad, si los objetos forman clusters de diferentes densidades, DBSCAN puede tener dificultades para su localización.

2.2.4.9 **Gaussian Mixture Model (GMM)**

Un modelo de mezcla gaussiano Gaussian Mixture Model (GMM) es una función de densidad de probabilidad representada por una suma de components gausianas.

Un GMM es una suma con pesos de densidades gausianas:

$$p(\vec{x}) = \sum_{i=1}^{M} w_i(N(\overrightarrow{x})\,\mu i, \sum i)$$

Donde $(\overrightarrow{x})$ es un vector D-dimensional de datos wi son los pesos y $(N(\overrightarrow{x})\mu i, \sum i)$ es la densidad gausiana.

Por lo tanto la caracterización se completa con la media, la matriz de covarianza y el peso de cada componente gausiana.

Algoritmo GMM

1- Inicialización. Para cada clase, un vector compuesto de la media y la matriz de covarianzas es construido. Este vector representa las características de la distribución gausiana usada para caracterizar las entidades del conjunto de datos.

2- Estimar la probabilidad de cada elemento de pertenecer a un determinado cluster.

3- Estimar los parámetros de la distribución de probabilidad del próximo ciclo.

4- Ejecutar test de convergencia para ver cuánto ha cambiado el vector de parámetros, y si la tolerancia es menor que el umbral definido el algoritmo se detiene.

2.3 ALGORITMO PROPUESTO DE CLUSTERING

A continuación se describe el algoritmo de clustering propuesto. Como se ha comentado anteriormente, el objetivo es la obtención de grupos de consumo representativos para un año de referencia en cualquier tipo de edificio, dando solución a la parte del modelo denominada $C_{CLÚSTER}$ $(Véase\ 2.1)$. Dichos grupos muestran un consumo tipo para unas determinadas condiciones climáticas (temperatura media exterior, radiación solar global horizontal) y tipo de día. Por lo que en el año de ejecución de la línea base se obtiene el consumo basal del día considerado como el consumo promedio del cluster al que dicho día es asociado.

El primer paso es la elección del método de cluster (no jerárquicos (k-means en nuestro caso) o jerárquicos). Dicha decisión es tomada por el usuario con libertad ya que ambos métodos son igualmente válidos para la clusterización de consumos en edificios. En los siguientes apartados se comentan de forma detallada las etapas a seguir en cada uno de ellos.

2.3.1 K-MEANS

A continuación se muestra el algoritmo k-means propuesto para la obtención de clusters de consumo en edificios. La primera etapa consiste en la preparación de los datos, la segunda en elegir el número de clusters a calcular, en la tercera establecemos la distancia de cálculo para obtener clusters y su posterior ejecución y la última etapa consiste en la ejecución del algoritmo k-means con el fin de obtener los distintos grupos requeridos.

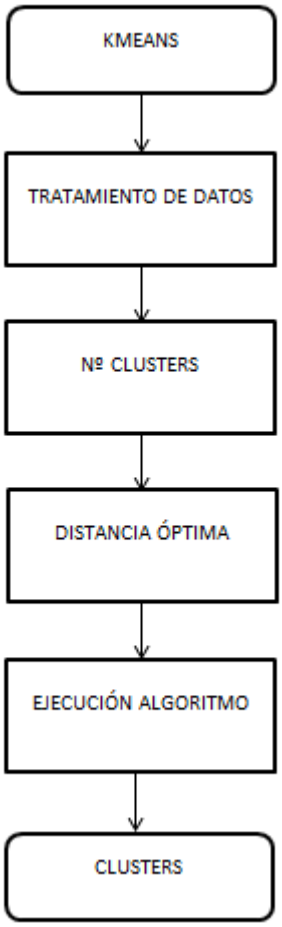

Figura 15. Algoritmo k-means

2.3.1.1 Tratamiento datos

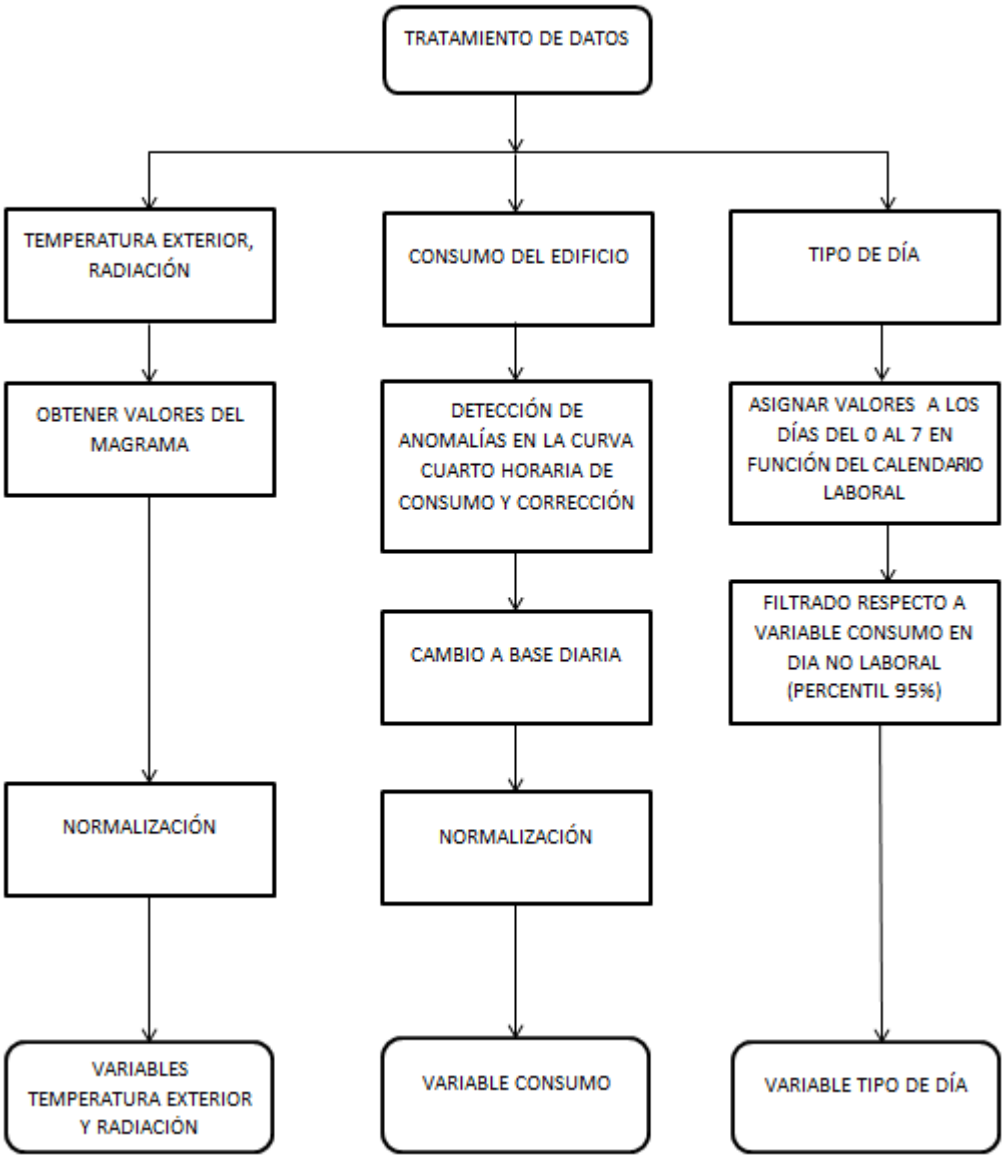

Figura 16. Tratamiento de datos

Los datos que tenemos como variables de entrada (inputs) son la temperatura exterior [ºC], radiación [kWh/m2], consumo del edificio [kWh] y tipo de día (variable numérica de 0 al 7).

Con respecto a la temperatura media exterior y la radiación solar global horizontal, ambas se obtienen del MAGRAMA [33] en base diaria (Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente).

De la variable consumo del edificio se dispone de su curva de consumo cuarto horaria. Dicha curva debe ser analizada con el fin de detectar anomalías y corregirlas. Posteriomente dichos datos se cambian a base diaria.

Estas variables (temperatura exterior, radiación solar y consumo del edificio) deben ser normalizadas antes de considerarse variables de entrada a los algoritmos de cluster, dicho cambio es necesario en variables cuantitativas para suavizar el efecto de la unidad de medida en el cálculo de las distancias.

Con respecto a la variable tipo de día se asigna valores del 0 al 7 en función del calendario laboral correspondiente a la ubicación del edificio en cuestión. El número 0 corresponde a días festivos y del 1 al 7 tenemos los días de la semana (Lunes=1, Martes=2,…, Domingo=7). Posteriormente y teniendo en cuenta el hecho de que exista algún tipo de festividad o actividad del edificio, no señalada en el calendario laboral, pero en el cual el consumo es similar al de un día no laboral, se aplica un filtro de consumo. Dicho filtro se elabora con los días festivos del año de referencia (elegido por el usuario) y se calcula como el percentil 95% de los consumos asociados a dicho día. Es decir, todos los días que tengan un consumo menor o igual a dicho valor serán tipo 0, aunque el calendario laboral en sí no lo especifique. Este filtro será aplicado únicamente a edificios con días de no operación (siendo no aplicable, por ejemplo, a los hospitales, ya que estos presentan una operación continua todo los días del año).

(Véase Estudio 1: Variables de Entrada).

2.3.1.2 **Número de clusters**

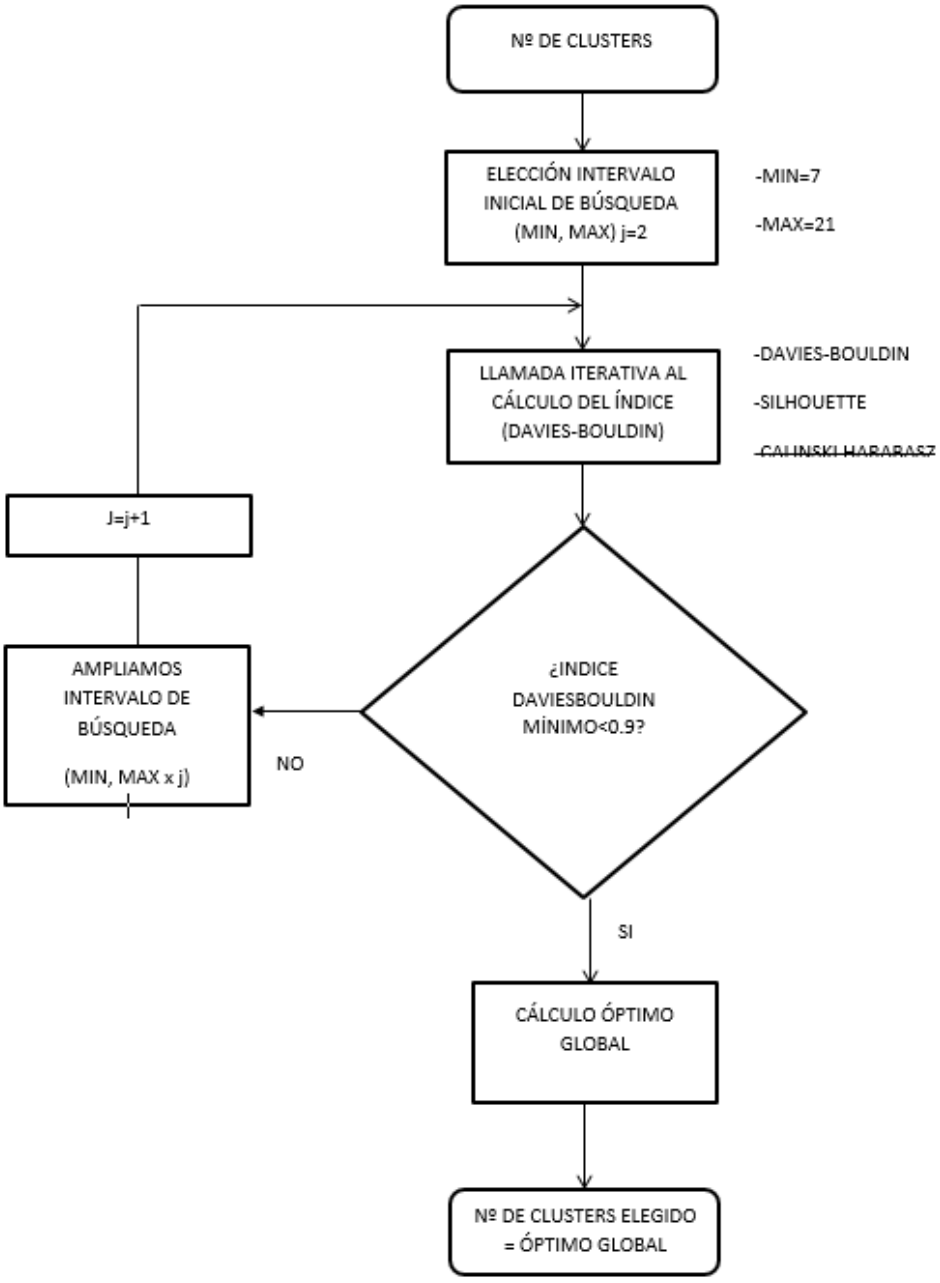

Figura 17. Número de Clusters

El número de Clusters a calcular será obtenido mediante el índice de Davies Bouldin, el cual ha sido tomado por ser el más nombrado y usado en la bibliografía y por la obtención de buenos resultados obtenidos con el mismo. El índice de Silhouette obtiene resultados muy parecidos a Davies Bouldin pero el índice de Calinski Haribasz no, ya que tiende a proponer como número óptimo de Clusters siempre el valor mínimo del intervalo de búsqueda asignado y cuya comprobación de resultados no ha sido satisfactoria, por lo que se descarta del procedimiento.

Para obtener dicho valor partimos de un intervalo de búsqueda de [7,21] cuyo valor mínimo es referido a la diferenciación de los siete días de la semana y valor máximo 21 (correspondiente a tres semanas, una de calefacción, una de refrigeración y una intermedia). Con dicho intervalo de búsqueda se efectúa cien veces la llamada a la función, mediante la cual obtenemos un vector de óptimos, dicha necesidad se debe al propio proceso de optimización (proceso iterativo de consolidación del número de Clusters óptimo). Si el índice de Davies Bouldin mínimo obtenido en dicho vector de llamadas es mayor se procede a ampliar el intervalo de búsqueda hasta que se verifique dicha restricción. El valor 0.9 ha sido tomado tras probar el procedimiento de búsqueda de Clusters en todos los edificios mencionados en el capítulo de aplicación y en todos ellos, un índice de Davies Bouldin mayor a 0.9 muestra resultados de clusters muy heterogéneos.

Una vez cumplida la restricción obtenemos el vector de valores óptimos y de él obtenemos el óptimo global, el cual corresponde al índice de Davies Bouldin más bajo, el cual lleva asociado el número de Clusters óptimo para el caso en estudio.

(Véase Estudio 3: Índices de evaluación del número de clusters).

2.3.1.3 Distancia óptima

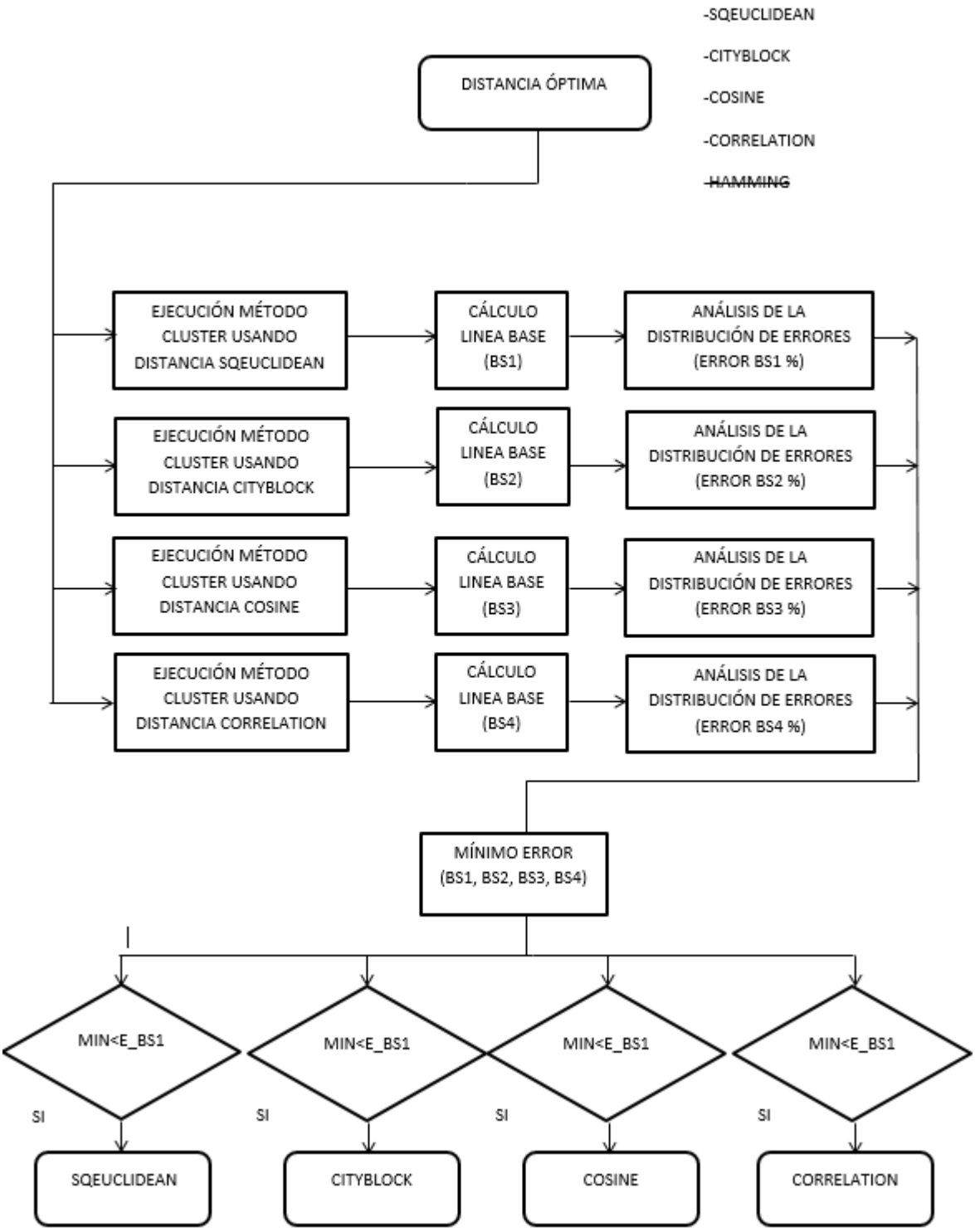

Figura 18. Distancia óptima

La distancia óptima a usar en la obtención de cluster se obtiene de la forma expresada en la figura 18. Las posibilidades de uso son seuclidean, cityblock, cosine, correlation y haming, descartando la última por ser usada únicamente con variables binarias. El uso de una u otra no es claro, por lo que el algoritmo propone evaluar todas las posibilidades, ejecutando el método de cluster para cada caso (una vez fijado el número de Clusters a calcular). Una vez evaluado con las cuatro opciones posibles ejecutamos la línea base con cada uno de ellos para el año de referencia y validación y evaluamos los errores diarios, mensuales y anuales cometidos en cada caso, eligiendo el óptimo en función de dicha evaluación.

(Véase Estudio 4: Influencia de la distancia elegida).

2.3.1.4 Ejecución del algoritmo en año distinto al de referencia

Cuando conocidos la temperatura exterior, radiación solar y tipo de día queremos asignar dicho día a un determinado cluster (calculado anteriormente con el año de referencia). La ejecución se lleva a cabo de la siguiente forma:

1. Con el número de clústers, la distancia elegida y tras la aplicación de k-means al año de referencia se obtienen los centroides y índice de los Clusters calculados.
2. Calculamos la matriz de distancias de cada dato de entrada a los distintos centroides (sin tener en cuenta el consumo).
3. El cluster asociado al nuevo caso será aquel correspondiente a la mínima distancia calculada anteriormente.

2.3.2 MÉTODOS JERÁRQUICOS

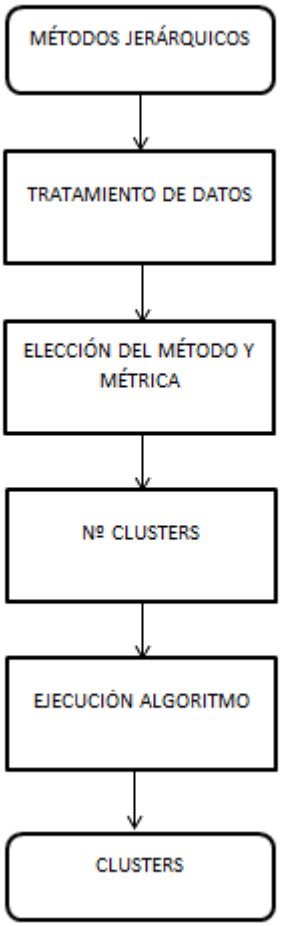

Figura 19. Algoritmo métotos jerárquicos

2.3.2.1 Tratamiento de datos

El tratamiento de datos se lleva a cabo de la misma forma que en k-means (Véase 2.3.1.1).

2.3.2.2 Elección del método y métrica

Para la elección del método y la métrica a usar para la obtención de los Clusters por un procedimiento jerárquico se siguen los siguientes pasos:

1-Cálculo del coeficiente de correlación cofenético para todas las combinaciones de métodos y métricas existentes dentro del grupo de aglomerativos.

2-Eligir los dos casos con mejor índice de correlación cofenético (más cercano a 1), ya que serán aquellos que presentan menor distorsión en las relaciones originales existentes entre los elementos en estudio.

3-Evaluar la distribución de errores de la línea base en el año de referencia y validación.

4-Elección de la distancia mínima en función de la evaluación de los errores (Véase Estudio 5: Comparación K-means Linkage).

2.3.2.3 Número de Clusters

El conjunto de datos se agrupan en conglomerados en función de la distancia de aproximación calculada. Se generan grupos binarios (grupos entre objetos) formados a través de vínculos de distancia y de enlaces de pares de objetos que son muy cercanos. Se crean grupos cada vez más grandes en función de estos pequeños grupos

binarios recién formado con otros objetos hasta que todos los objetos en el conjunto original de datos se encuentran relacionados en lo que denominamos árbol jerárquico o dendograma.

El grupo de árboles binarios es más fácil de entender cuando se ve gráficamente. El dendograma presenta en el eje horizontal los índices de los objetos en el conjunto de datos originales (nodos) y las alturas representan la distancia entre objetos.

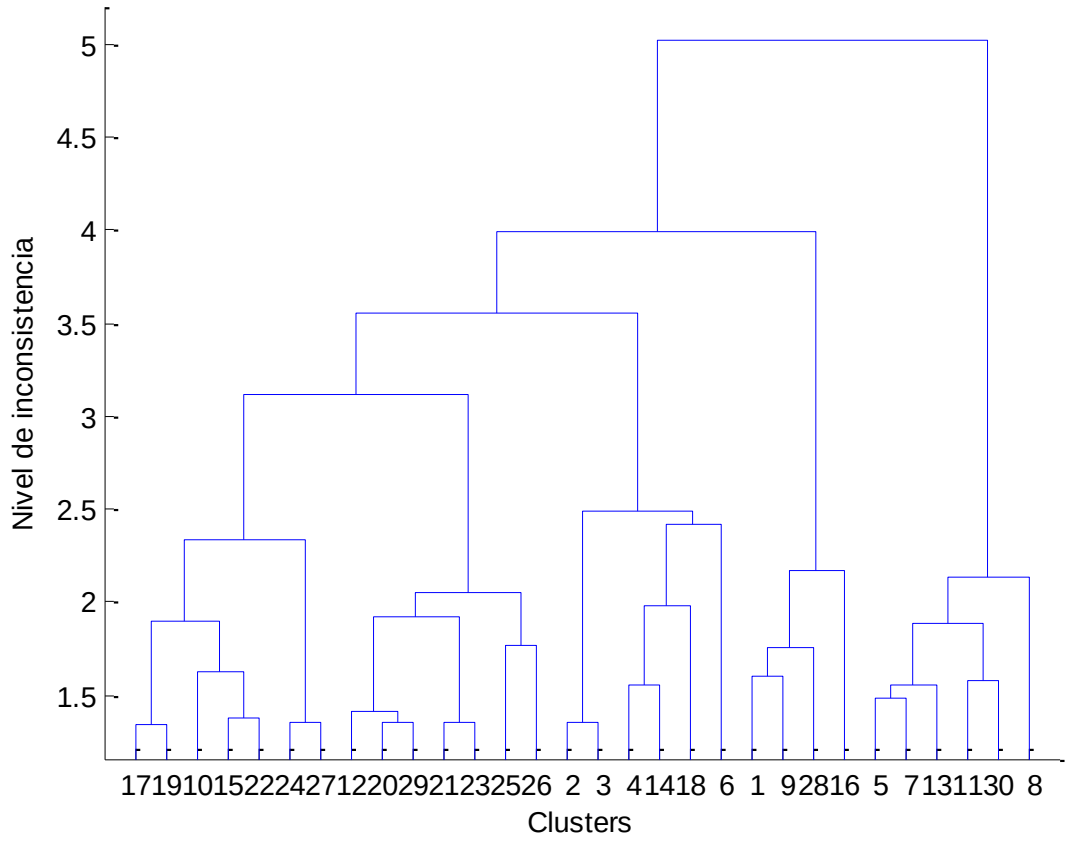

Figura 20. Dendograma

El dendograma (árbol jerárquico) organiza los datos en subcategorías que se van dividiendo en otras hasta llegar al nivel de detalle deseado. Este tipo de representación permite apreciar claramente las relaciones de agrupación entre los datos e incluso entre los mismos grupos, aunque no las relaciones de similitudes o cercanía entre categorías.

La deteminación del número de Clusters se determina trazando una horizontal en el dendograma (Figura 20). La cantidad de Clusters generados serán las intersecciones verticales producidas, en el ejemplo mostrado serían 4 clusters.

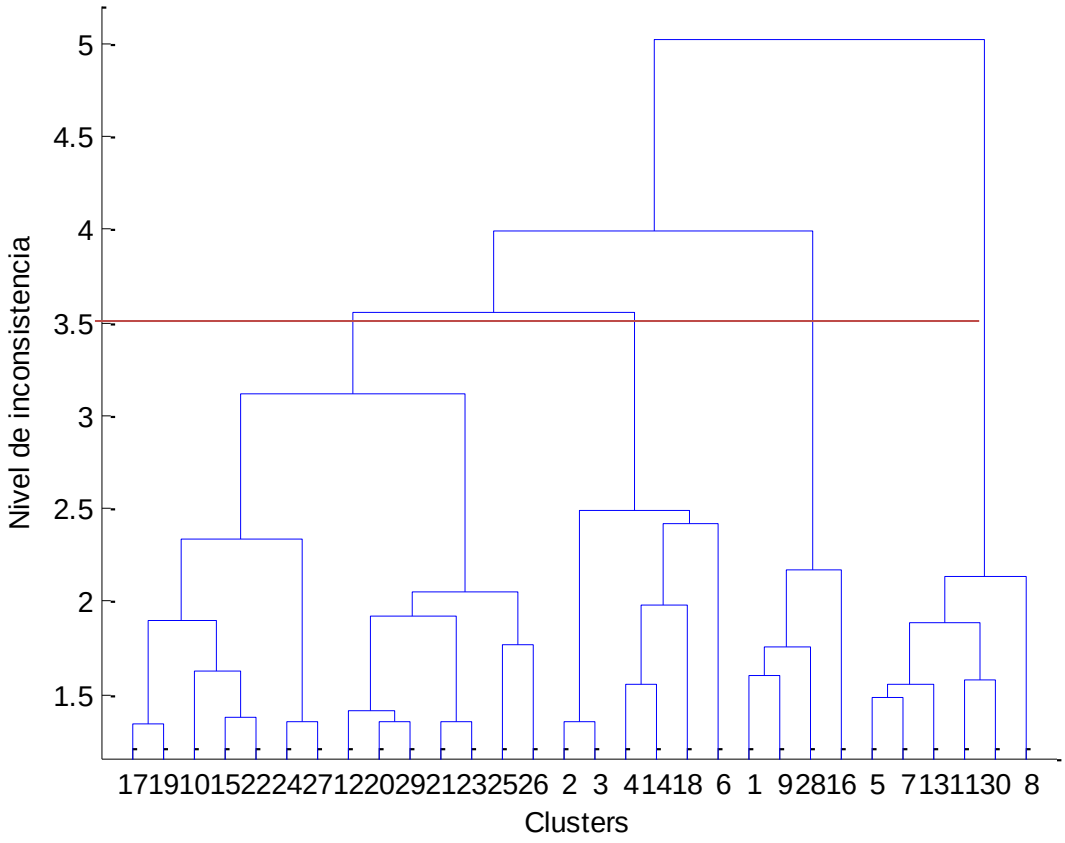

Figura 21. Dendograma corte horizontal 1

Si se baja la línea horizontal se puede observar una mayor cantidad de grupos, donde todas las ramas bajo esta línea conforman nodos diferentes.

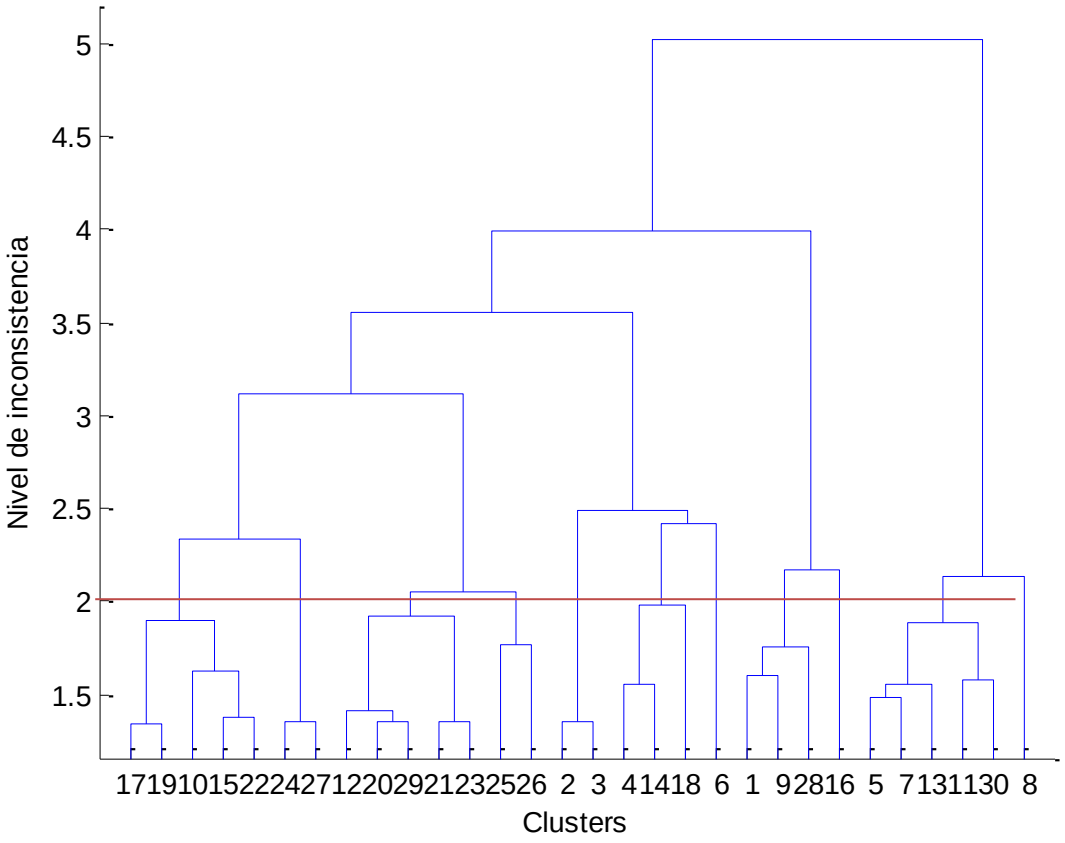

Figura 22. Dendograma corte horizontal 2

La decisión sobre el número de Clusters en estos métodos es un poco subjetiva, especialmente cuando tenemos un gran número de objetos, si seleccionamos pocos, los Clusters resultantes serán heterogéneos y artificiales, mientras que si se seleccionan demasiados, la interpretación de los mismos suele resultar complicada.

Para tomar una decisión sobre el número de Clusters se suele representar los distintos pasos del algoritmo y la distancia a la que se produce la unión, siendo el punto de corte aquel en el que comience a producirse saltos bruscos.

2.3.2.4 Ejecución del algoritmo en un año distinto al de referencia

Cuando conocidos la temperatura exterior, radiación solar y tipo de día queremos asignar dicho día a un determinado cluster (calculado anteriormente con el año de referencia). La ejecución se lleva a cabo de la siguiente forma:

1. Con el número de clústers, la distancia elegida y tras la aplicación de método jerárquico al año de referencia se obtienen los índices de los Clusters calculados.
2. Calculamos los centroides de cada cluster promediando los valores correspondientes en cada grupo obtenido.
3. Calculamos la matriz de distancias de cada dato de entrada a los distintos centroides (sin tener en cuenta el consumo).
4. El cluster asociado al nuevo caso será aquel correspondiente a la mínima distancia calculada anteriormente.

3 ESTUDIOS DESARROLLO DEL PROTOCOLO

3.1 ALCANCE

El capítulo presente es un capítulo de estudios con el cual se obtienen y validan las directrices expuestas en apartados anteriores.

El edificio usado la Agencia Andaluza de la Energía, dicho edificio es una agencia pública empresarial de la Administración de la Junta de Andalucía, y está adscrita a la Consejería de Empleo, Empresa y Comercio, surgió con la finalidad de ser una herramienta puesta al servicio del tejido social, empresarial e institucional andaluz para impulsar el desarrollo energético sostenible de nuestra Comunidad.

Figura 23. Agencia Andaluza de la Energía

Se trata de un edificio que consta de oficinas y despachos, se puede decir que es un edificio típicamente de oficinas. Con un horario cerrado de 9:00 a 14:00, el tener un horario fijo y con una ocupación media, muy similar entre días, estas propiedades ayuda a caracterizar el consumo de este tipo de edificios.

El edificio está situado en c/Isaac Newton, Isla de la Cartuja, Sevilla

Figura 24. Ubicación edificio AAE

3.2 ESTUDIO 1: VARIABLES DE ENTRADA

3.2.1 Descripción

El estudio 1 consiste en la evaluación de la influencia de las variables de entrada, así como la forma de las mismas en la obtención de los clusters de consumo. Para ello se toma el edificio de la Agencia Andaluza de la Energía. El método de cluster usado para dicho análisis es k-means. Fijamos el intervalo de búsqueda del número de clusters de [7,63] (suponiendo mínimo 7 días correspondiente a una semana y 63 considerando las tres estaciones (21 días para calefacción, 21 para refrigeración y 21 para la estación intermedia) y la distancia usada a seuclidean.

3.2.2 Resultados

Caso 1

1- Variables de entrada: consumo del edificio [kWh], temperatura media exterior [ºC], radiación solar global horizontal [kWh/m2], variable binaria (1=laboral, 0=festivo) y variable día de la semana (1 al 7, siendo 1=Lunes, 2=Martes, …, 7=Domingo).

2- Aplicación algoritmo k-means.

El número de clusters obtenidos (mínimo valor del índice de Davies-Bouldin) es igual a 7.

3- Evaluación en el año de referencia. En este apartado es comparado el consumo que tiene el cluster que es asignado a cada día (estimado) y el consumo real (medido) del mismo.

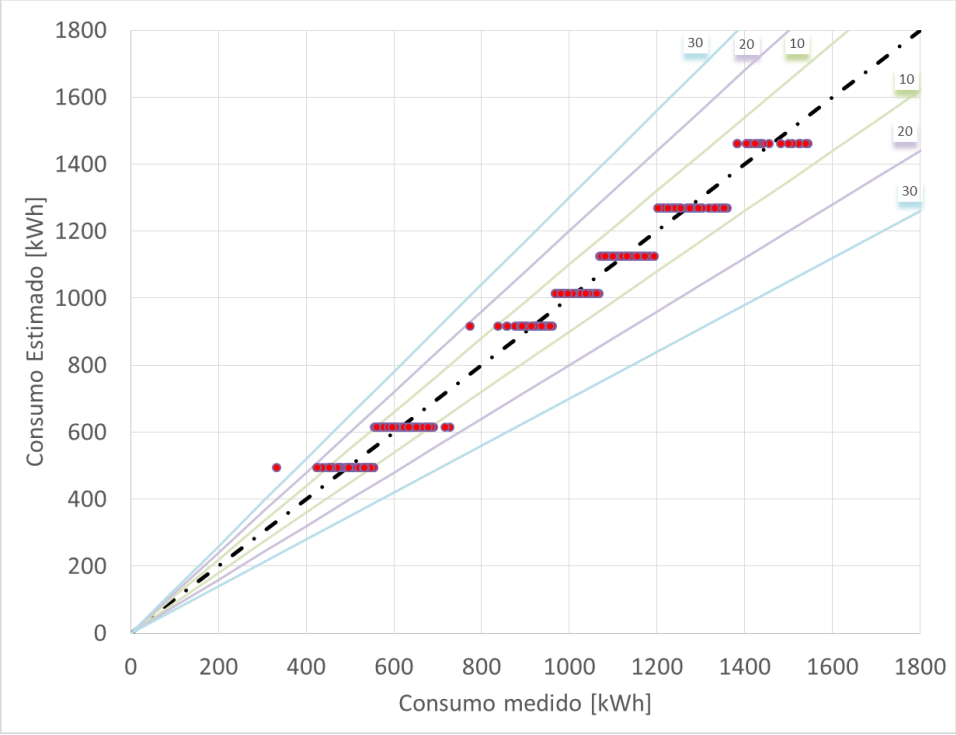

Figura 25. Evaluación Estudio1-Caso1

Error mínimo diario (%)	**0.1**
Error promedio diario (%)	**4.0**
Error máximo diario (%)	**48.9**

Tabla 1. Errores diarios (%) Estudio1-Caso1

4- Ejecución en el año de referencia

Ejecución para el año de referencia. Para ello suponemos desconocido el consumo y asignamos los Clusters a cada día en función de las variables temperatura media exterior, radiación solar y tipo de día, mediante el cálculo de la matriz de distancias. Luego evaluamos el cluster asignado en ejecución y el obtenido inicialmente (teniendo en cuenta el consumo).

Error ejecución (%)	**71.51**

Tabla 2. Error ejecución (%) Estudio1-Caso1

Caso 2

1- Variables de entrada: consumo del edificio [kWh], temperatura media exterior [ºC], radiación solar global horizontal [kWh/m2] (estas tres primeras variables son normalizadas), variable binaria (1=laboral, 0=festivo) y variable día de la semana (1 al 7, siendo 1=Lunes, 2=Martes, …, 7=Domingo).

2- Aplicación algoritmo k-means.

El número de clusters obtenidos (mínimo valor del índice de Davies-Bouldin) es igual a 61.

3- Evaluación en el año de referencia.

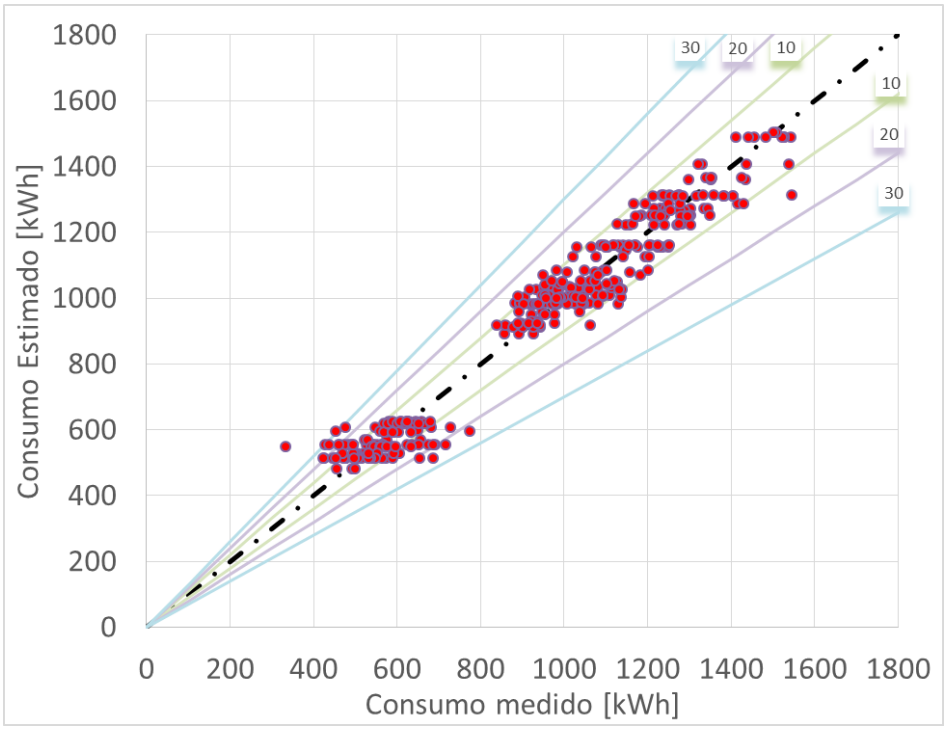

Figura 26. Evaluación Estudio1-Caso2

Error mínimo diario (%)	**0**
Error promedio diario (%)	**5.7**
Error máximo diario (%)	**65.7**

Tabla 3. Errores diarios (%) Estudio1-Caso2

4- Ejecución en el año de referencia

Ejecución para el año de referencia. Para ello suponemos desconocido el consumo y asignamos los Clusters a cada día en función de las variables temperatura media exterior, radiación solar y tipo de día, mediante el cálculo de la matriz de distancias. Luego evaluamos el cluster asignado en ejecución y el obtenido inicialmente (teniendo en cuenta el consumo).

Error ejecución (%)	6.03

Tabla 4. Error ejecución (%) Estudio1-Caso2

Caso 3

1- Variables de entrada: consumo del edificio [kWh], temperatura media exterior [ºC], radiación solar global horizontal [kWh/m2] (estas tres primeras variables son normalizadas), variable tipo de día (0 al 7,donde 0=días festivos por calendario laboral, 1=Lunes, 2=Martes, …, 7=Domingo).

2- Aplicación algoritmo k-means.

El número de clusters obtenidos (mínimo valor del índice de Davies-Bouldin) es igual a 61.

3- Evaluación en el año de referencia.

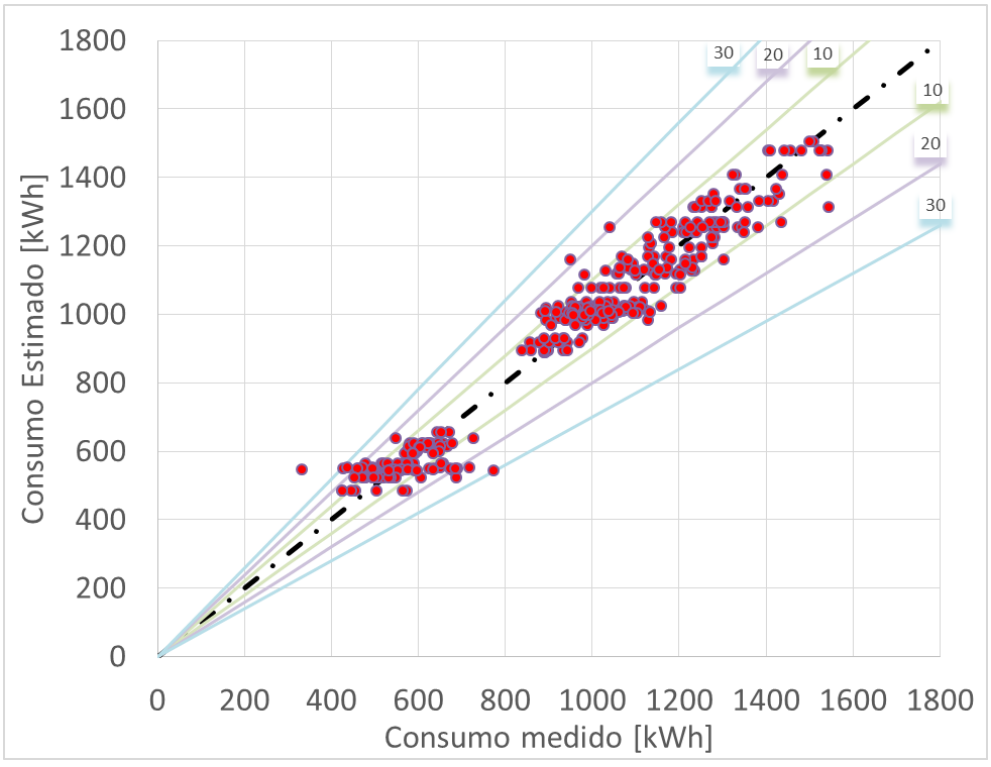

Figura 27. Evaluación Estudio1-Caso3

Error mínimo diario (%)	0
Error promedio diario (%)	5.8
Error máximo diario (%)	64.4

Tabla 5. Errores diarios (%) Estudio1-Caso3

4- Ejecución en el año de referencia

Ejecución para el año de referencia. Para ello suponemos desconocido el consumo y asignamos los Clusters a cada día en función de las variables temperatura media exterior, radiación solar y tipo de día, mediante el cálculo de la matriz de distancias. Luego evaluamos el cluster asignado en ejecución y el obtenido inicialmente (teniendo en cuenta el consumo).

Error ejecución (%)	5.21

Tabla 6. Error ejecución (%) Estudio1-Caso4

Caso 4

1- Variables de entrada: consumo del edificio [kWh], temperatura media exterior [ºC], radiación solar global horizontal [kWh/m2] (estas tres primeras variables son normalizadas), variable tipo de día (0 al 7,donde 0=días festivos por calendario laboral y días de consumo filtrado (percentil 95% de consumo de los días no laborables), 1=Lunes, 2=Martes, …, 7=Domingo).

2- Aplicación algoritmo k-means.

El número de clusters obtenidos (mínimo valor del índice de Davies-Bouldin) es igual a 61.

3- Evaluación en el año de referencia.

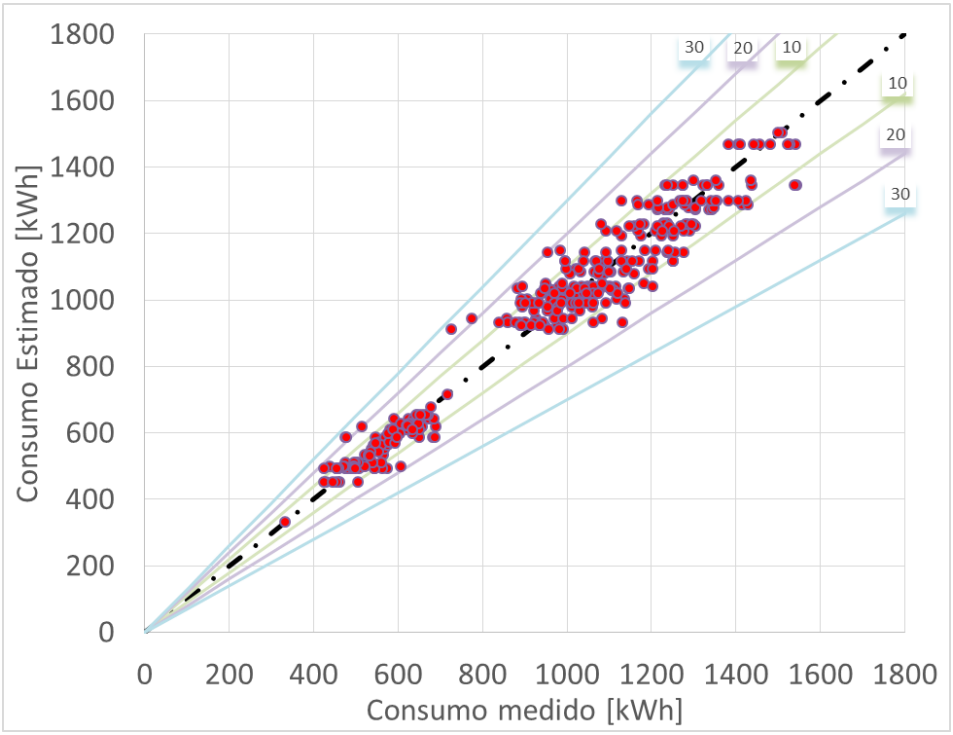

Figura 28. Evaluación Estudio1-Caso4

Error mínimo diario (%)	0
Error promedio diario (%)	5
Error máximo diario (%)	25.7

Tabla 7. Errores diarios (%) Estudio1-Caso4

4- Ejecución en el año de referencia

Ejecución para el año de referencia. Para ello suponemos desconocido el consumo y asignamos los Clusters a cada día en función de las variables temperatura media exterior, radiación solar y tipo de día, mediante el cálculo de la matriz de distancias. Luego evaluamos el cluster asignado en ejecución y el obtenido inicialmente (teniendo en cuenta el consumo).

Error ejecución (%)	8.22

Tabla 8. Error ejecución (%) Estudio1-Caso4

3.2.3 Conclusiones

La conclusión principal del estudio de variables de entrada en la necesidad de normalizar las variables cuantitativas. Véase los errores de ejecución obtenidos, la diferencia básica entre el caso 1 y los demás es la normalización, pues es clara la necesidad de ello en el algoritmo de clustering propuesto anteriormente.

Una vez descartado el caso 1, de los casos 2, 3 y 4 se elige el caso 4 ya que el error máximo diario obtenido en la ejecución es un 60 % menor a los casos 2 y 3.

Por lo tanto las variables de entrada elegidas son las asociadas al caso 4: consumo del edificio [kWh], temperatura media exterior [ºC], radiación solar global horizontal [kWh/m2] (estas tres primeras variables son normalizadas), variable tipo de día (0 al 7,donde 0=días festivos por calendario laboral y días de consumo filtrado (percentil 95% de consumo de los días no laborables), 1=Lunes, 2=Martes, …, 7=Domingo).

	Caso1	**Caso2**	**Caso3**	**Caso4**
Error mínimo diario (%)	0.1	0	0	0
Error promedio diario (%)	4	5.7	5.8	5
Error máximo diario (%)	48.9	65.7	64.4	25.7

Tabla 9. Errores diarios (%) Estudio1-Evaluación año referencia

	Caso1	**Caso2**	**Caso3**	**Caso4**
Error ejecución (%)	71.51	6.03	5.21	8.22

Tabla 10. Errores ejecución (%) Estudio1-Ejecución año referencia

3.3 ESTUDIO 2: DETERMINACIÓN DEL ORDEN DEL MODELO

3.3.1 Descripción

El objetivo de este estudio es pre-establecer el número de numeradores y denominadores apropiados, así como la necesidad de diferenciar las tres grandes estaciones del edificio: refrigeración, calefacción e intermedia. Este objetivo se resuelve utilizando los datos de consumo de 2014 (año de referencia) y 2015 (año de validación) del edificio Agencia Andaluza de la Energía.

Las variables de entrada al algoritmo de clustering son el consumo del edificio [kWh], temperatura media exterior [ºC], radiación solar global horizontal [kWh/m2] (estas tres primeras variables son normalizadas), variable tipo de día (0 al 7,donde 0=días festivos por calendario laboral y días de consumo filtrado (percentil 95% de consumo de los días no laborables), 1=Lunes, 2=Martes, …, 7=Domingo).

El número de Clusters obtenido (óptimo global del vector de óptimos, una vez obtenemos un valor del índice de Davies-Bouldin menor a 0.9, valor obtenido como medida de fiabilidad tras numeras pruebas realizadas en diversos tipos de edificios).

Una vez obtenido el número de Clusters (61), fijando la distancia seuclidean se aplica el algoritmo de clustering al año de referencia (2014), obteniendo con ello los Clusters y sus respectivos centroides, siendo el centroide un vector compuesto por el promedio de los valores incluidos en cada uno de los grupos o Clusters, de las cuatro variables de entrada.

A continuación se expone un ejemplo de cluster típico de calefacción (39) obtenido para el año de referencia y uno típico de refrigeración (11). Como se puede observar los grupos mantienen una similitud clara en los días asociados a cada uno de ellos.

Cluster 11	Consumo [kWh]	Temp Media [ºC]	Radiación [MJ/m2]	Tipo de día
12/05/2014	1434	25.57	7.75	1
16/06/2014	1352	23.4	7.21	1
14/07/2014	1303	26.48	7.47	1
11/08/2014	1215	25.78	7.47	1
18/08/2014	1270	27.34	6.8	1
25/08/2014	1241	27.59	7.02	1
08/09/2014	1145	24.23	5.21	1
15/09/2014	1147	23.64	5.47	1
Centroide	**1263**	**25.50**	**6.80**	**1**

Tabla 11. Resultados cluster 11 (refrigeración) AAE para el año de referencia

Cluster 39	Consumo [kWh]	Temp Media [ºC]	Radiación [MJ/m2]	Tipo de día
02/12/2013	1542	9.43	3.24	1
16/12/2013	1405	11.01	2.71	1
23/12/2013	1410	7.27	2.94	1
30/12/2013	1456	6.17	2.78	1
13/01/2014	1276	11.59	1.79	1
20/01/2014	1526	8.07	3.07	1
27/01/2014	1236	12.77	2.75	1
03/02/2014	1527	7.63	2.12	1
10/02/2014	1523	10.11	4.07	1
17/02/2014	1442	10.46	2.96	1
10/11/2014	1482	11.32	3.22	1
17/11/2014	1382	12.45	3.19	1
Centroide	**1434**	**9.86**	**2.90**	**1**

Tabla 12. Resultados cluster 39 (calefacción) AAE para el año de referencia

A continuación se lleva a cabo la ejecución en el año 2015, para ello procedemos a calcular la matriz de distancias de cada dato de entrada a los distintos centroides sin tener en cuenta el consumo, ya que éste es supuesto desconocido en la ejecución de la línea base. El Cluster asociado a cada día será aquel corresponiente a la mínima distancia calculada, siendo $Text_x$ temperatura del día analizado, $Text_c$ la correspondiente al promedio del cluster, siendo nomenclatura similar a radiación y tipo de día.

$$d = \sqrt{(Text_x - Text_c)^2 + (RAD_x - RAD_c)^2 + (Tipodia_x - Tipodia_c)^2}$$

A continuación se exponen los Clusters 11 y 15 obtenidos en el año de ejecución (2015):

Cluster 11	Consumo [kWh]	Temp Media [ºC]	Radiación [MJ/m2]	Tipo de día
08/06/2015	1405	24.08	19.91	1
29/06/2015	1720	26.91	29.05	1
06/07/2015	1612	29.41	28.02	1
20/07/2015	1553	28.41	28.27	1
27/07/2015	1497	28.39	29.80	1
03/08/2015	1654	27.10	24.78	1
10/08/2015	1465	28.49	23.52	1
17/08/2015	1138	24.03	24.89	1
Centroide	**1506**	**27.10**	**26.03**	**1**

Tabla 13. Resultados cluster 11 (refrigeración) AAE para el año 2015

Cluster 39	Consumo [kWh]	Temp Media [ºC]	Radiación [MJ/m2]	Tipo de día
01/12/2014	1387	13.35	9.49	1
15/12/2014	1615	10.73	10.00	1
22/12/2014	1453	10.87	10.14	1
29/12/2014	1706	7.46	10.89	1
05/01/2015	1402	6.65	7.25	1
12/01/2015	1733	10.84	10.43	1
19/01/2015	1703	9.42	11.25	1
26/01/2015	1736	8.56	13.07	1
02/02/2015	1807	7.91	6.31	1
09/02/2015	1909	9.67	14.72	1
16/02/2015	1603	8.70	5.41	1
23/02/2015	1439	11.57	11.81	1
23/03/2015	1363	12.57	8.83	1
16/11/2015	1246	13.93	11.31	1
23/11/2015	1523	9.40	12.04	1
30/11/2015	1552	12.21	11.09	1
Centroide	**1574**	**10.24**	**10.25**	**1**

Tabla 14. Resultados cluster 39 (calefacción) AAE para el año 2015

Se puede observar como varía el cluster del año 2014 al 2015, esto es debido a que se asignan a cada cluster existente los días más parecidos en sus tres variables de ejecución, pudiendo estar dentro de un cluster un determinado día no parecido pero que es aún menos similar a los restantes creados con el año de referencia.

A continuación se observa la tipificación de días anual para el año 2014 y 2015 mediante un mapa de colores, los cuales se asignan en función del cluster al que pertenecen, el cual tiene asociado un consumo tipo.

Gracias a modificar la distancia calculada, eliminando la influencia del consumo en el cálculo de la misma, es posible obtener dicho mapa en el año de ejecución de línea base (2015).

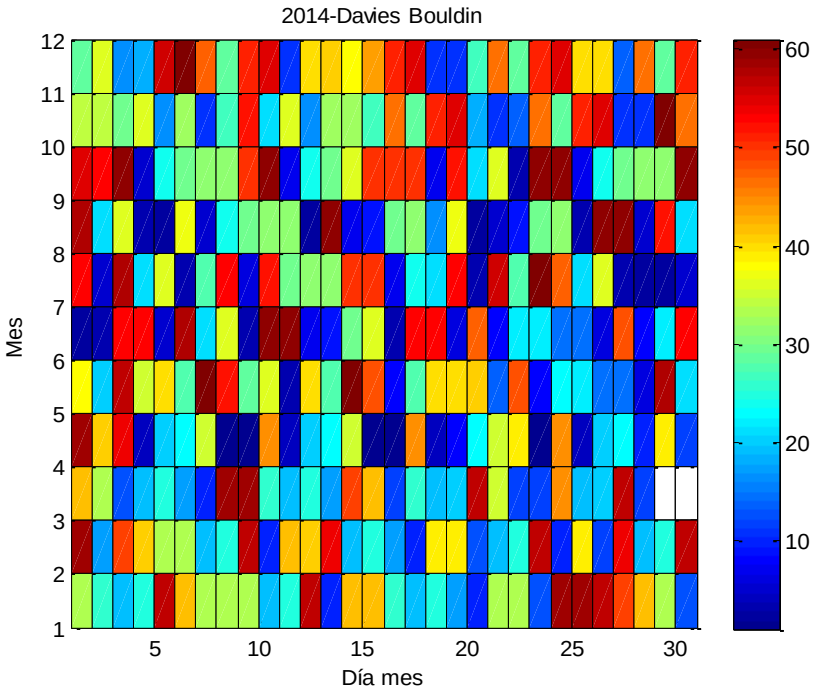

Figura 29. Tipificacicón de días año 2014

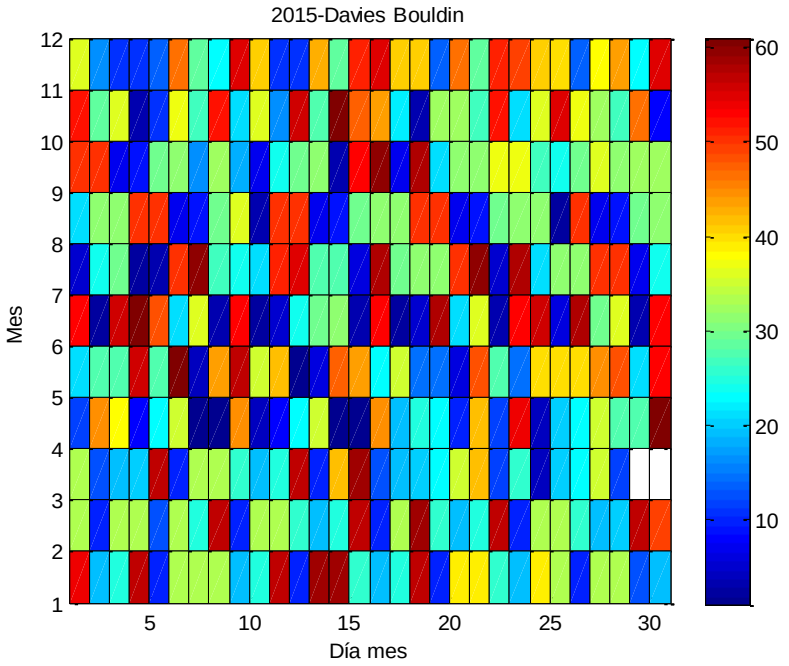

Figura 30. Tipificacicón de días año 2015

3.3.2 Resultados

Análisis de Opción 1-Incremento de Consumo

En primer lugar se obtiene el modelo de incremento de consumo para todo el año de referencia sin buscar estaciones intermedias.

Mod.1 - 1 denominador y 2 numeradores

$$\Delta C(d) = -1.7259 \cdot \Delta\overline{T_{EXT}(d)} + 1.9411 \cdot \Delta\overline{T_{EXT}(d-1)} + 2.0571 \cdot \Delta RAD(d) + 8.4947 \cdot \Delta RAD(d-1) + 0.2284 \cdot \Delta C(d-1)$$

Mod.2 - 1 denominador y 3 numeradores

$$\Delta C(d) = -5.0499 \cdot \Delta\overline{T_{EXT}(d)} - 10.1823 \cdot \Delta\overline{T_{EXT}(d-1)} + 15.1758 \cdot \Delta\overline{T_{EXT}(d-2)} - 5.0499 \cdot \Delta RAD(d) + 29.5711 \cdot \Delta RAD(d-1) + 4.6605 \cdot \Delta RAD(d-2) + 0.3437 \cdot \Delta C(d-1)$$

A primera vista el modelo dos responde con sus coeficientes a la física del problema por los siguientes motivos:

- El denominador toma un valor de 0.3437 lo que se corresponde con una constante de tiempo de 23h para el edificio, lo cual está del orden de lo esperado según los valores de ASHRAE para un edificio ligero.
- El signo asociado a los coeficientes no puede ser analizado por mezclar ambos regímenes calefacción/refrigeración.

Para analizar estos signos se elige como variable climática los grados hora vinculados a una temperatura de consigna de 20ºC, es decir, la integral diaria de la diferencia de temperaturas horaria entre temperatura consigna de 23 ºC y la temperatura exterior, y se toma el mayor valor entre 23 menos la temperatura exterior y temperatura exterior menos 23. De esta forma se toma una decisión sobre si una hora es de calefacción o refrigeración.

Mod.3 - 1 denominador y 3 numeradores (GH)

$$\Delta C(d) = 0.0724 \cdot \Delta GH(d) - 0.2069 \cdot \Delta GH(d-1) + 0.3168 \cdot \Delta GH(d-2) - 4.9154 \cdot \Delta RAD(d) - 1 + 29.5952 \cdot \Delta RAD(d-1) + 4.4132 \cdot \Delta RAD(d-2) + 0.3416 \cdot \Delta C(d-1)$$

En este caso, salvo por el efecto de la radiación, se puede ver cómo la opción con tres numeradores ofrece mayor confianza.

Para esta última opción (Mod.3), sumando al modelo anterior el valor del clúster de cada día, el ajuste resulta:

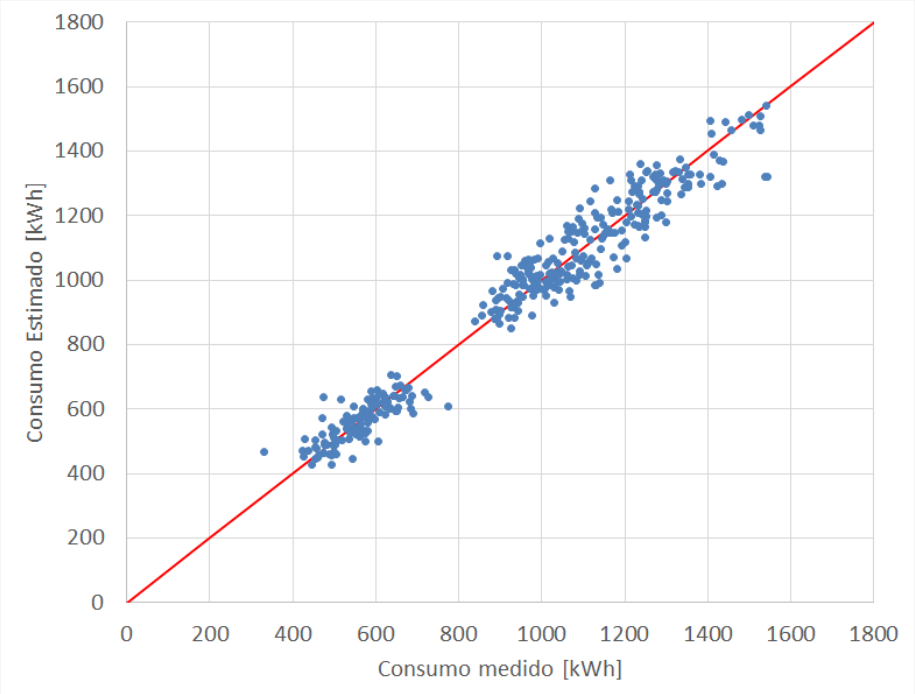

Figura 31. Resultados modelo de incremento de consumo (Mod.3-1d+3n)

Si se analiza la muestra de resultados y los errores cometidos se tiene:

Límite inferior Error [%]	Número casos	% días
5	152	41.6
10	39	10.7
20	6	1.6
30	2	0.5
40	1	0.3
50	0	0.0

Tabla 15. resultados modelo incremento de consumo (Mod.3-1d+3n)

Comparando los resultados de este modelo, con la opción de no usar los grados hora, se tiene:

Límite inferior Error [%]	Número casos	% días
5	156	42.7
10	39	10.7
20	6	1.6
30	2	0.5
40	1	0.3
50	0	0.0

Tabla 16. resultados modelo incremento de consumo (Mod.2 1d+3n)

Los resultados son análogos, por lo que se elegiría la opción temperatura media exterior para eliminar la necesidad de tomar decisiones vinculadas a la consigna del edificio y a las estaciones de refrigeración y calefacción.

	Enero	Febrero	Marzo	Abril	Mayo	Junio
Cons. Medido [kWh]	30948	29602	29962	25076	26952	27972
Mod.1 -Estimación [kWh]	31162	28741	29882	25588	27183	27797
Mod.1 - Errores [%]	0.69	3.00	0.27	2.00	0.85	0.63
Mod.2 -Estimación [kWh]	30891	28497	30080	25440	27233	27915
Mod.2 - Errores [%]	0.19	3.88	0.39	1.43	1.03	0.20
Mod.3 -Estimación [kWh]	30875	28487	30090	25435	27223	27899
Mod.3 - Errores [%]	0.23	3.91	0.42	1.41	1.00	0.26

	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Cons. Medido [kWh]	32260	30007	29127	26690	26444	28463

Mod.1 -Estimación [kWh]	32022	30480	28474	26830	26758	28670
Mod.1 - Errores [%]	0.74	1.55	2.29	0.52	1.18	0.72
Mod.2 -Estimación [kWh]	32101	30660	28477	26696	27097	28492
Mod.2 - Errores [%]	0.49	2.13	2.28	0.02	2.41	0.10
Mod.3 -Estimación [kWh]	32102	30685	28509	26684	27108	28482
Mod.3 - Errores [%]	0.49	2.21	2.17	0.02	2.45	0.07

	Anual
Cons. Medido [kWh]	170512
Mod.1 -Estimación [kWh]	343587
Mod.1 - Errores [%]	0.02
Mod.2 -Estimación [kWh]	343579
Mod.2 - Errores [%]	0.02
Mod.3 -Estimación [kWh]	343580
Mod.3 - Errores [%]	0.02

Tabla 17. Comparación de modelos de la opción incremento de consumo

El modelo 1 ofrece mejores resultados que los otros dos, tanto en los meses candidatos a calefacción como en refrigeración. Es por ello que si solo se atendiera a este criterio, el resultado idóneo sería poner un solo denominador con dos numeradores.

Analisis de Opción 2-Modelo de Consumo

Siguiendo la misma filosofía del caso anterior, se analiza esta opción de modelización y se compara con la mejor del caso anterior. No se muestran resultados del efecto de los grados hora por tener una conclusión idéntica al caso anterior.

Mod.1 - 1 denominador y 2 numeradores

$$C_{BS}(d) = -1.4776 \cdot \overline{T_{EXT}(d)} + 1.1259 \cdot \overline{T_{EXT}(d-1)} + 2.6162 \cdot RAD(d) - 1.8939 \cdot RAD(d-1) + 0.9913 \cdot C_{CLÚSTER}(d) - 0.2239 \cdot C_{CLÚSTER}(d-1) + 0.2355 \cdot C_{BS}(d-1)$$

Mod.2 - 1 denominador y 3 numeradores

$$C_{BS}(d) = -1.5120 \cdot \overline{T_{EXT}(d)} + 2.3284 \cdot \overline{T_{EXT}(d-1)} - 0.6971 \cdot \overline{T_{EXT}(d-2)} + 3.9110 \cdot RAD(d) - 1.2720 \cdot RAD(d-1) - 2.6018 \cdot RAD(d-2) + 0.9894 \cdot C_{CLÚSTER}(d) - 0.2194 \cdot C_{CLÚSTER}(d-1) - 0.0157 \cdot C_{CLÚSTER}(d-2) + 0.2427 \cdot C_{BS}(d-1)$$

La calidad del ajuste con respecto a las opciones anteriores es similar. Por ejemplo el modelo 2 ofrece el siguiente gráfico:

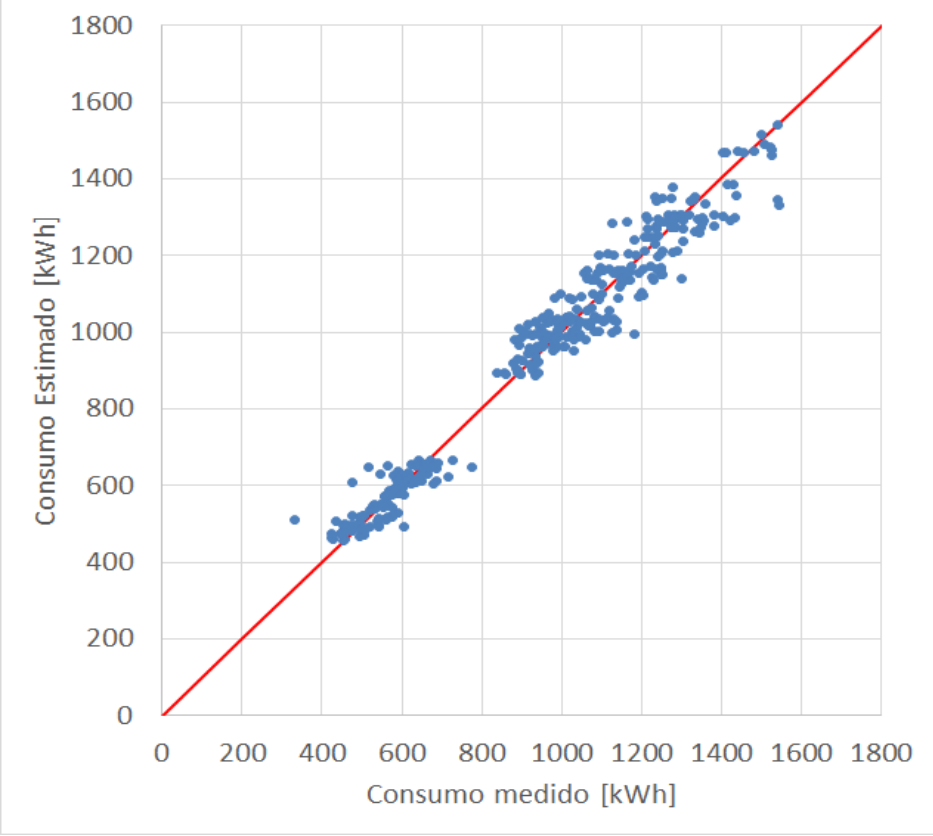

Figura 32. Resultados modelo línea base (Mod.2-1d+3n)

No obstante, cuando se analizan en detalle los resultados se puede ver como en general hay una mejora de casi un 15% en los resultados diarios. Este dato aparece vinculado a la reducción del número de casos con error superior al 5%.

Límite inferior Error [%]	Número casos	% Días
5	134	36.7
10	28	7.7
20	3	0.8
30	1	0.3
40	1	0.3
50	1	0.3

Tabla 18. Resultados modelo línea base (Mod.2 1d+3n)

Si se comparan ambos modelos en la base óptima de explotación final de la línea base, es decir, la base mensual; se puede ver cómo mejoran los errores mensuales, y como el ajuste es aceptable. Las siguientes gráficas y tablas corroboran estos resultados.

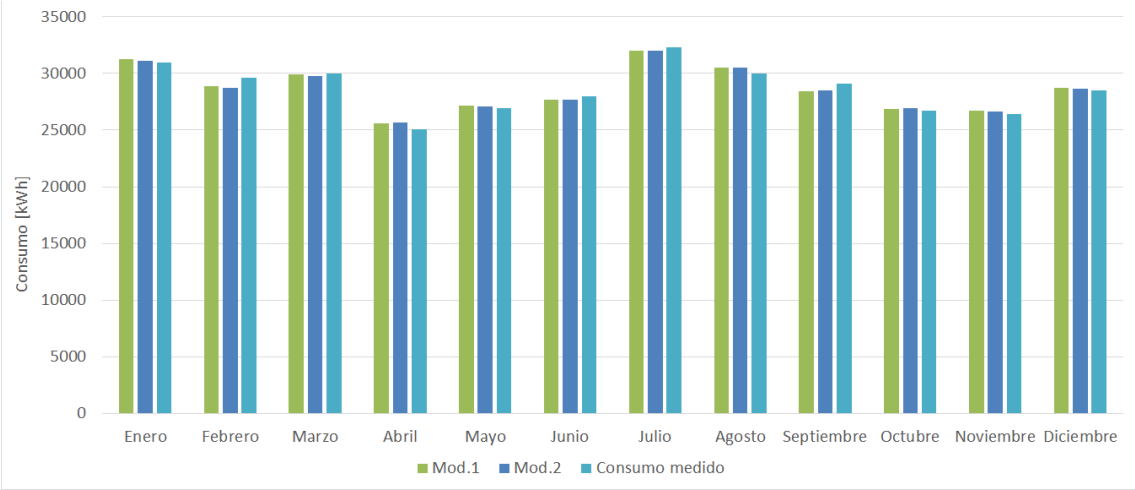

Figura 33. Comparación modelos de línea base

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
Mod.1	0.91	2.61	0.24	2.13	0.68	1.02	0.87	1.56	2.39	0.52	0.87	0.94	0.02
Mod.2	0.61	2.97	0.61	2.20	0.60	1.00	0.84	1.72	2.17	0.77	0.76	0.63	0.05

Tabla 19. Comparación de errores de modelos de línea base

Análisis de Opción 3-Identificación de Estaciones

Esta opción aparece de forma natural al final del estudio de la opción anterior. Es decir, conocida la forma del modelo (1 denominador y 2 numeradores), y su obtención de manera directa a partir del consumo referente al clúster y las excitaciones climáticas; se procede a analizar si es conveniente estimar la estación de calefacción, refrigeración e intermedia para realizar un modelo diferente en cada una de ellas.

La justificación es que el sistema de climatización y la respuesta del edificio en calefacción, refrigeración e intermedias es diferente, por tanto la manera de contemplar esta diferencia es con modelos diferentes (implícita).

Por tanto el modelo de línea base resulta:

$$C_{BS-CAL}(d) = -1.4776 \cdot \Delta GH(d) + 1.1259 \cdot \Delta GH(d-1) + 2.6162 \cdot RAD(d) - 1.8939 \cdot RAD(d-1) + 0.9913 \cdot C_{CLÚSTER}(d) - 0.2239 \cdot C_{CLÚSTER}(d-1) + 0.2355 \cdot C_{BS-CAL}(d-1)$$

$$C_{BS-INT}(d) = -0.0510 \cdot \Delta GH(d) + 0.1607 \cdot \Delta GH(d-1) - 2.1105 \cdot RAD(d) - 2.2880 \cdot RAD(d-1) + 0.9869 \cdot C_{CLÚSTER}(d) - 0.0153 \cdot C_{CLÚSTER}(d-1) + 0.0576 \cdot C_{BS-INT}(d-1)$$

$$C_{BS-REF}(d) = -0.0700 \cdot \Delta GH(d) + 0.030 \cdot \Delta GH(d-1) + 6.0250 \cdot RAD(d) - 3.0868 \cdot RAD(d-1) + 0.9835 \cdot C_{CLÚSTER}(d) - 0.2937 \cdot C_{CLÚSTER}(d-1) + 0.3088 \cdot C_{BS-REF}(d-1)$$

Dónde cada uno de los submodelos anteriores permite en el año de validación tener en cuenta la variación de consumo debido a variaciones climáticas. En el caso de la estación intermedia, se espera que el modelo dependa fuertemente del consumo de clúster y las excitaciones en días anteriores sean la variación del consumo relacionada con el propio error del clúster.

Para realizar este estudio hay que tomar una decisión en cuanto a la duración de la estaciones. El criterio tomado está referido a los grados hora antes definido. Aquellos días en los que los GH con consigna 20 sean dominantes y superiores a 50, se tomarán de calefacción.

Los resultados para los dos años son:

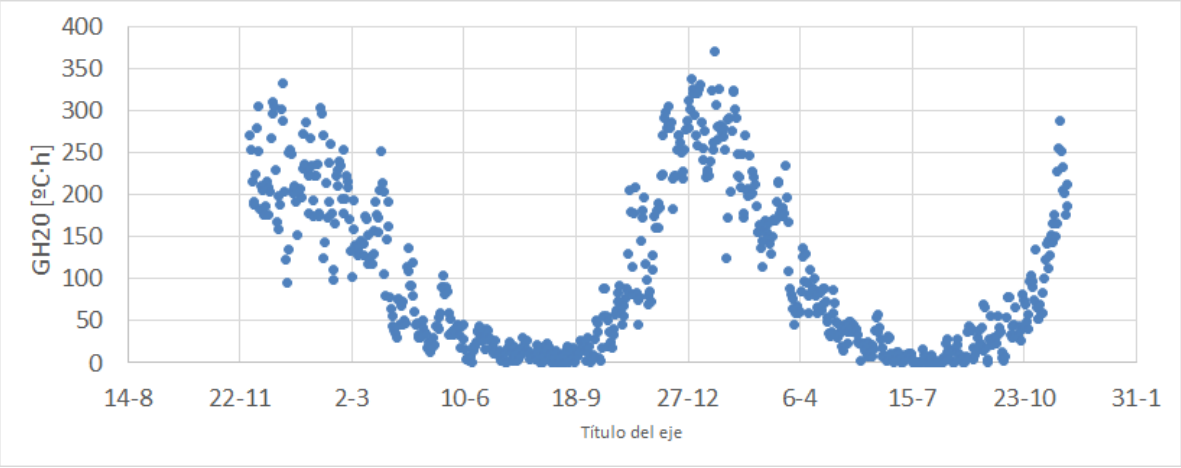

Figura 34. Grados hora 20 para cada uno de los días a estudio

Analizando estas gráficas y los propios clúster de los meses candidatos a intermedios (abril, mayo y octubre), se puede llegar a la conclusión de un límite de 100 ºC·h puede ser válido para considerar la estación de calefacción. Este límite se puede considerar en temperatura media diaria, despejando. Lo que resulta una temperatura media diaria de 15.8ºC.

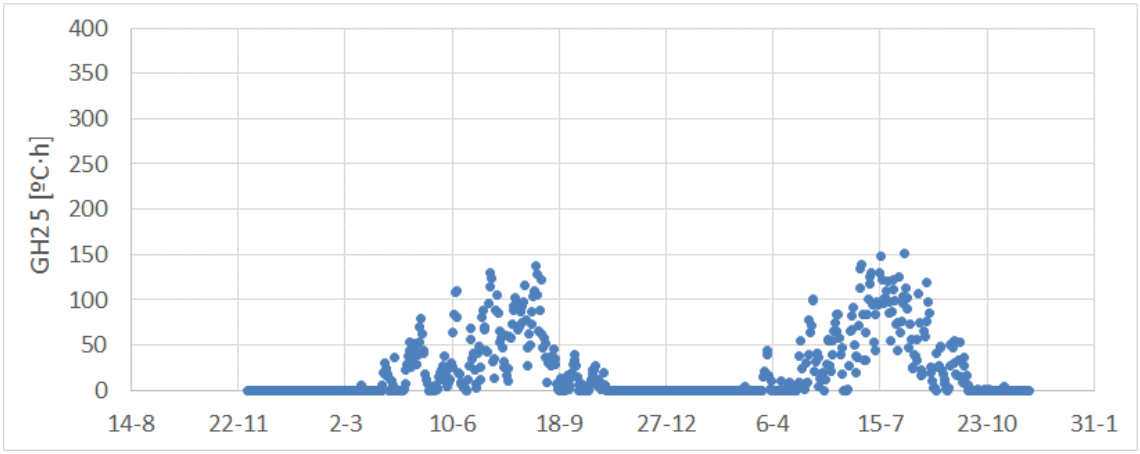

Figura 35. Grados hora 25 para cada uno de los días a estudio

Siguiendo el proceso análogo, se considera refrigeración aquellos días con GH25>25 [ºC·h] y que además tengan un valor de GH20 menor. Hablando en temperatura media diaria mayor o igual a 26ºC.

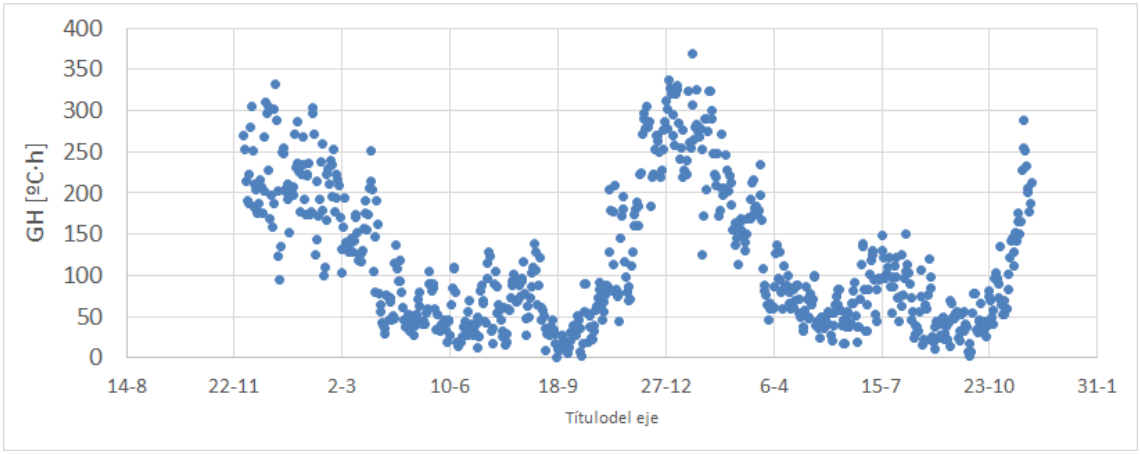

Figura 36. Grados hora neto representativos de cada día

Unificando en una sola variable cada día quedaría representando por el mayor valor de dos anteriores. Sin embargo, la opción que unifica los cambios de estación con más exactitud para el procedimiento ciego matemático es considerar GH 22.5 como 22.5 de consigna menos la temperatura exterior. Resultando:

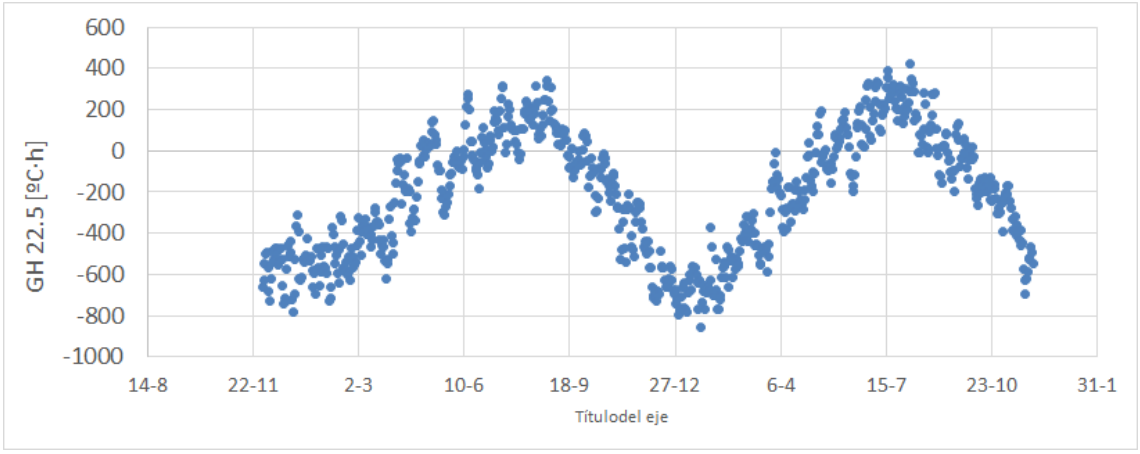

Figura 37. Grados hora 22.5 modelización

En esta nueva consigna resulta que un valor de GH inferior a -160 ºC·h resulta calefacción y mayor a 84 ºC·h para refrigeración. Estos límites son idénticos a los anteriores pero referidos a la nueva base.

Con estos últimos GH son los que se analiza la opción de modelar por estaciones el año. Los resultados aplicando los modelos con los que se comenzó el epígrafe son:

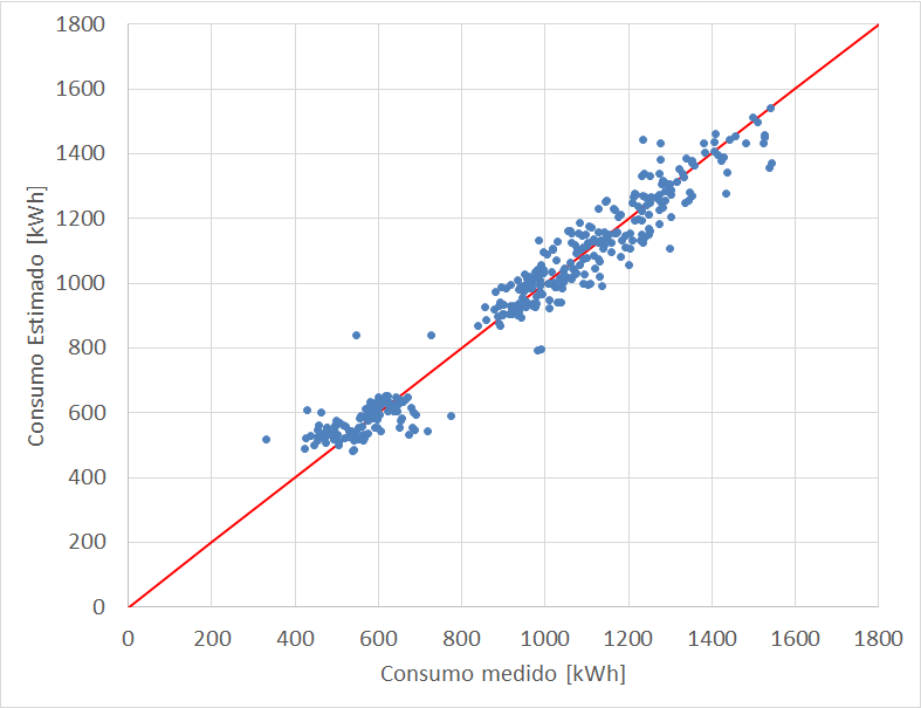

Figura 38. Resultados modelo BS estacional (1d+2n)

El ajuste presenta una calidad a nivel cualitativa, similar a la de las opciones anteriores. Sin embargo el análisis cuantitativo muestra la siguiente distribución de errores:

Límite inferior Error [%]	Número casos	% Días
5	148	40.5
10	34	9.3
20	4	1.1
30	2	0.5
40	1	0.3
50	1	0.3

Tabla 20. Resultados modelo línea base estacional (Mod. 1d+2n)

Esta tabla muestra como los resultados diarios son algo mayores puesto que se ha aumentado en casi un 5% el número de casos con un error superior al 5%.

Finalmente el gráfico mensual confirma que el modelo funciona correctamente.

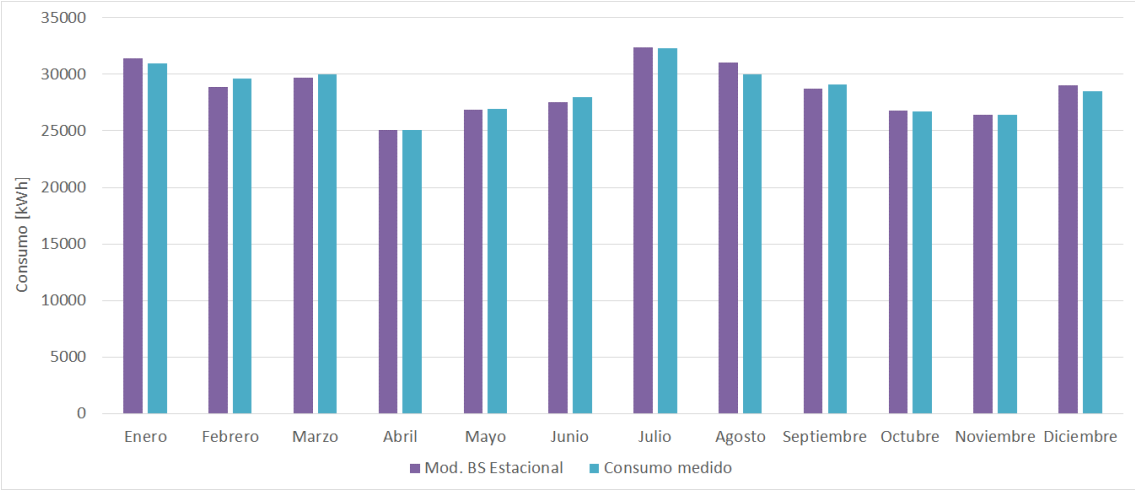

Figura 39. Resultados mensual BS Estacional

Finalmente añadir hasta tres numeradores a este caso no supone una mejora sensible como muestra el análisis de resultados diarios:

Límite inferior Error [%]	Número casos	% Días
5	148	40.5
10	31	8.5
20	5	1.4
30	2	0.5
40	2	0.5
50	0	0.0

Tabla 21. Resultados modelo línea base estacional (Mod. 1d+3n)

3.3.3 Conclusiones

Tras el análisis de las diferentes opciones de modelización, se decide descartar la opción de incremento de consumo por no ofrecer mejores resultados y contemplar un trabajo adicional.

Por consiguiente, se promueve la modelización de la línea base considerando como variables independientes la temperatura exterior, radiación y consumo asociado al clúster del día a estudio. Ahora bien, la temperatura exterior puede aparecer como una temperatura media diaria o como un valor de grados hora/día. En el primer estudio se ha concluido que a nivel de ajuste no supone una mejora cuantitativa sensible para el modelo. No obstante, en el caso de optar por una línea base estacional es conveniente analizar la variable GH y usarla en el propio modelo.

Sin embargo, la decisión final se debe hacer teniendo en cuenta los resultados en el año validación. Es por ello que se procede a la aplicación del clúster.

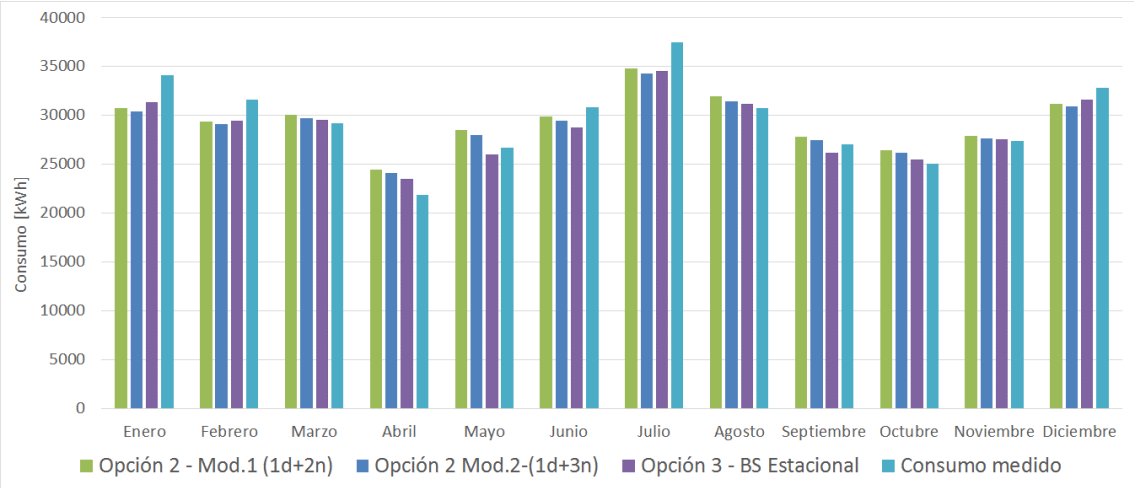

Figura 40. Resultados mensual de la comparación de modelos

Los resultados anuales no presentan una utilidad clara, ya que al igual que los mensuales demuestran que los modelos ofrecen resultados idénticos pero con variaciones diferentes. Por ejemplo el modelo estacional sobrevalora los meses de refrigeración e infravalora los meses intermedios o de cambio. Sin embargo en el resto de meses compensa las desviaciones.

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
Opción 2 - Mod.1 (1d+2n)	5.50	4.53	2.44	12.56	5.50	0.04	4.00	4.30	1.05	1.26	4.38	2.21	1.19
Opción 2 Mod.2-(1d+3n)	12.25	8.91	1.55	9.10	4.54	4.68	9.29	2.44	1.51	4.51	0.93	6.16	1.79
Opción 3 - BS Estacional	8.77	7.33	1.20	6.75	2.64	7.23	8.66	1.56	3.33	1.68	0.74	3.97	2.82

Tabla 22. Comparación de errores de modelos de línea base

Si analizamos los dos mejores modelos, que además se corresponden con las dos vías de modelización propuestas resulta:

Límite inferior Error [%]	Opción 3 - BS Estacional		Opción 2 - Mod.2-(1d+3n)	
	Número casos	% Días	Número casos	% Días
5	281	77.0	292	80.0
10	189	51.8	200	54.8
20	69	18.9	63	17.3
30	16	4.4	17	4.7
40	10	2.7	10	2.7
50	9	2.5	5	1.4

Tabla 23. Análisis de errores diarios (valor promedio 12%)

Ambos modelos ofrecen ajustes similares, pero se puede ver como a nivel diario la opción 2 mejora sensiblemente los resultados de la opción 3. Es por este motivo que en la opción 3 se añade un numerador más para realizar se podía pensar que el óptimo podría ser añadir un numerador a la opción 3.

La conclusión es clara, añadir un numerador mejora los resultados diarios, pero se empeoran los resultados mensuales. No obstante, como la base de trabajo es la diaria y no se quiere promocionar la compensación de errores, se concluye que la mejor opción es la segunda con 3 numeradores. Aunque la opción mixta 2 y 3 (BS Estacional con 3 numeradores) aparece como una opción mejorable si se añaden más datos que permitan mejorar la definición de las estaciones.

3.4 ESTUDIO 3: ÍNDICES DE EVALUACIÓN DEL NÚMERO DE CLUSTERS

3.4.1 Descripción

El objetivo de este estudio es conocer la influencia del índice elegido para el cálculo del número óptimo de clúster, y como ambas dos variables afectan al resultado final de la estimación de la línea base.

Para ello se va a tomar el edificio de la agencia andaluza con 2014 como año de referencia y los modelos anteriores (1 denominador y 3 numeradores, con y sin distinción de estaciones) para estimar 2014 y 2015 como año de validación. La conclusión esperada es el índice por defecto a usar en el algoritmo para la tipología de datos con las que se trabaja.

Tomando K-means, se evalúa la eficiencia de cada uno de los siguientes índices:

- Calinski Harabasz
- Davies Bouldin
- Silhouette

Los índices elegidos son implementados en el algoritmo de determinación del número de clúster y la clasificación de los días en función de los mismos, para el método K-means elegido.

3.4.2 Resultados

Los resultados referentes al índice de Davies Bouldin son los que aparecen en el estudio anterior. De este mismo se adopta que la modelización se realiza con 1 denominador y 3 numeradores, pudiendo descomponerse el año en tres estaciones. Esto último hace que aparezcan tres modelos vinculado al consumo en función de la estación de trabajo. Para elegir las estaciones, tal y como se ha explicado en el estudio anterior, se toman unos límites referidos a los grados hora.

Los resultados de Silhouette (60 clúster) para el año de referencia (2014):

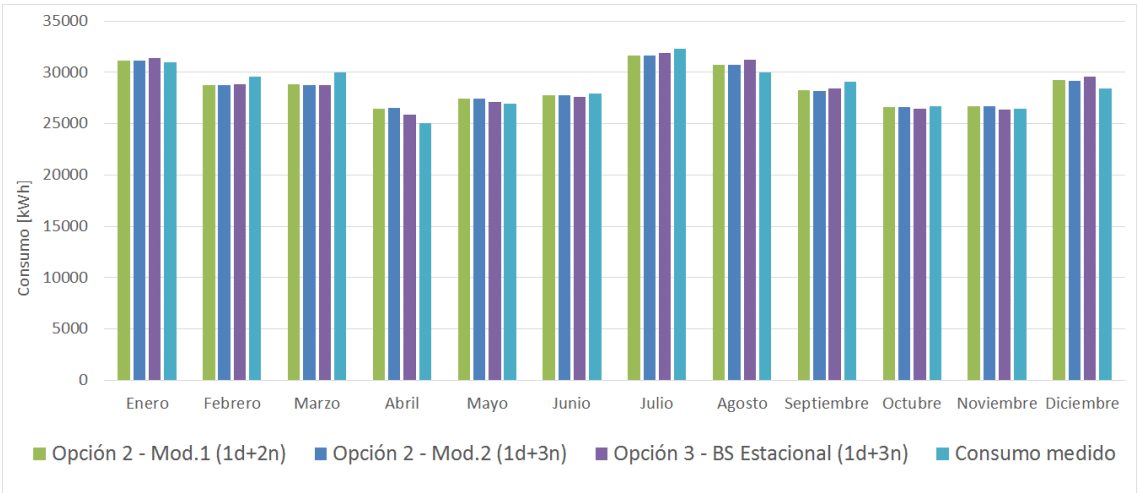

Figura 41. Resultados mensual de la comparación de modelos (Índice de Silhouette – año referencia)

Y para el año de validación:

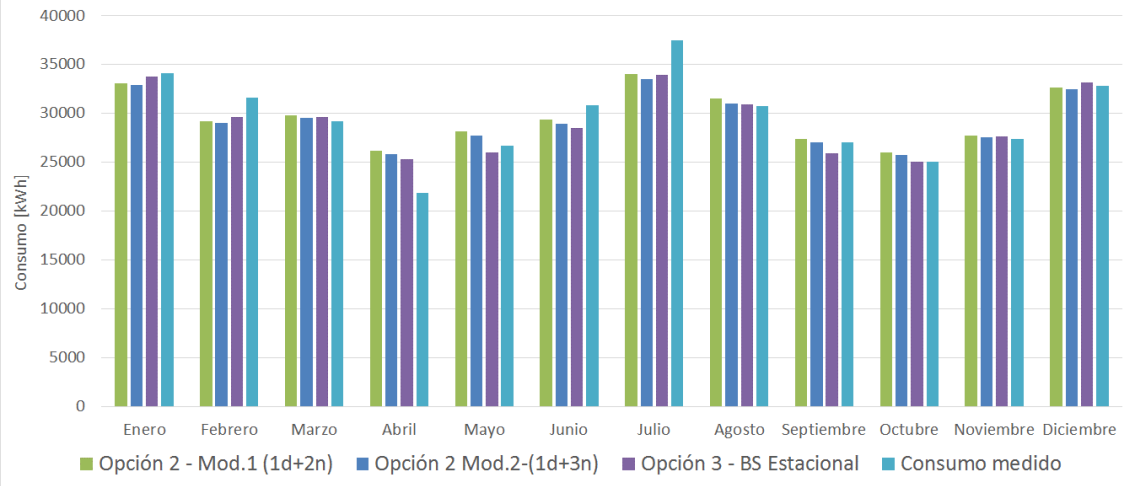

Figura 42. Resultados mensual de la comparación de modelos (Índice de Silhouette – año validación)

Los resultados son análogos al índice de Davies Bouldin, incluso en el número de clúster óptimo. Sin embargo la agrupación que se realiza internamente demuestra que el índice de Silhouette no tiene la misma consistencia que Davies Bouldin para los datos que se están tratando. Véase los errores [%] mensuales y diarios sobre el año de validación para ratificar estas conclusiones:

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
Opción 2 - Mod.1 (1d+2n)	5.50	4.53	2.44	12.56	5.50	0.04	4.00	4.30	1.05	1.26	4.38	2.21	1.19
Opción 2 Mod.2-(1d+3n)	3.77	8.98	1.15	15.31	3.58	6.32	11.87	1.05	0.07	2.87	0.57	1.12	0.99
Opción 3 - BS Estacional	1.04	7.00	1.33	13.61	2.90	8.16	10.53	0.57	4.29	0.20	0.83	1.02	1.59

Tabla 24. Análisis de errores mensuales

Límite inferior Error [%]	Opción 3 - BS Estacional		Opción 2 - Mod.2-(1d+3n)	
	Número casos	% Días	Número casos	%
5	264	72.3	278	76.2
10	183	50.1	201	55.1
20	79	21.6	74	20.3
30	29	7.9	25	6.8
40	12	3.3	12	3.3
50	8	2.2	5	1.4

Tabla 25. Análisis de errores diarios (valor promedio apróx. 15%)

Los resultados no son sensibles al cambio de índice pero demuestran que el índice de Davies Bouldin sigue garantizando mejores resultados para la tipología de variables con las que se trabaja. En el algoritmo de clustering se han dejado los tres índices por la necesidad de probar el algoritmo de manera masiva y tomar una

decisión en cuanto a este tipo de restricciones.

En cambio cuando se usa el índice de Calinski Harabasz, el número óptimo desciende hasta 7, lo que en el ajuste conlleva este tipo de resultados:

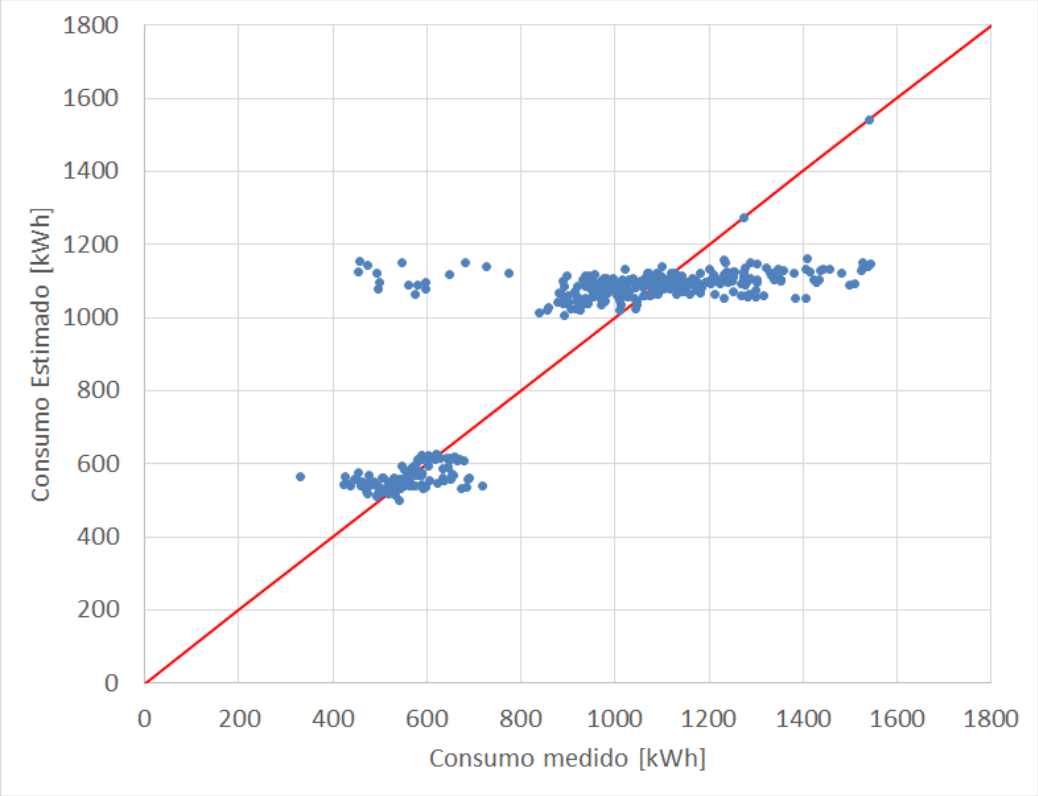

Figura 43. Resultados modelo BS estacional (1d+3n) – año de referencia – Calinski

El ajuste para el año de referencia demuestra que usando esta solución de clustering el modelo pierde toda la correlación con los datos.

Se comprueba que aunque la literatura es ambigua con este índice, su utilidad para el tipo de variables a estudio es nula. Esto es así porque se recomienda el uso de variables enteras / binarias, y en los casos a estudio se tienen variables reales.

3.4.3 Conclusiones

Tras el análisis sobre el edificio de referencia, y la verificación en otros dos edificios más, se puede concluir que el índice de Davies Bouldin para métodos no jerárquicos es el aconsejado por el algoritmo para la creación de líneas base en edificios terciarios. Se toma por tanto como solución por defecto.

El algoritmo necesita su validación masiva para ratificar esta conclusión, ya que el índice fija el número óptimo de clúster y su agrupación para optimizar el valor del ajuste. Es por ello que el valor del índice depende fuertemente de la posición de los centroides y las distancias.

3.5 ESTUDIO 4: INFLUENCIA DE LA DISTANCIA ELEGIDA

3.5.1 Descripción

Tomando K-means, el índice Davies Bouldin, se evalúa la importancia de la elección de la distancia, tanto a nivel del resultado del algoritmo de clustering como de los ajustes de la línea base. El modelo de ajuste se toma el modelo 2 de la opción 2, es decir, 1 denominador y 3 numeradores sin distinguir estaciones.

Como ha sido comentado en el capítulo anterior, el algoritmo de clustering necesita saber qué distancia usar para obtener, una vez fijado el número de clusters, el índice de cluster correspondiente a cada día.

El presente estudio dará como producto la distancia aconsejable a elegir en, una vez fijado el índice de evaluación del número de clusters, y de la tipología de datos que se tienen.

Las cuatro posibilidades de estudio para k-means son las siguientes distancias:

-seuclidean

-cityblock

-cosine

-correlation

(Véase la definición y diferencias entre las mismas en el capítulo 2. Apartado 2.2.1).

La distancia es una variable importante en el clustering, siendo el modo de obtener los distintos grupos en el año de referencia y de asociar los nuevos días (definidos por temperatura media exterior, radiación solar global horizontal y tipo de día) a los grupos ya existentes en el año de ejecución (validación).

3.5.2 Resultados

En primer lugar se analizan los resultados referentes al año de referencia. En ellos se puede observar como las 4 distancias ofrecen buenos resultados. Aunque hay que destacar que correlation es el que mayores errores comete.

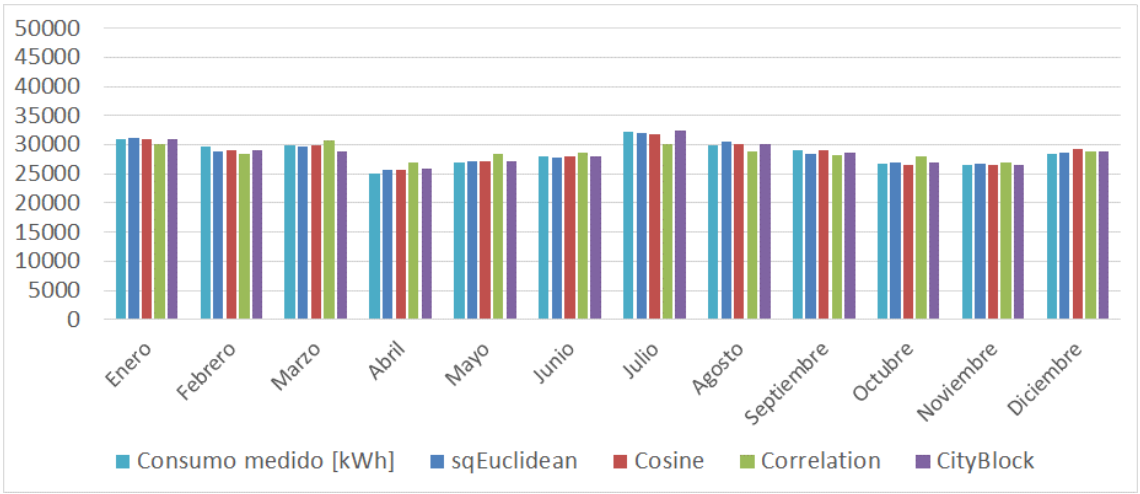

Figura 44. Resultados mensual de la comparación de modelos (año referencia)

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
sEuclidean	0.61	2.97	0.61	2.20	0.60	1.00	0.84	1.72	2.17	0.77	0.76	0.63	0.05
Cosine	0.30	2.11	0.24	2.42	0.45	0.08	1.52	0.22	0.45	0.50	0.63	2.49	0.10
Correlation	2.56	4.15	2.26	6.72	5.15	2.63	7.16	3.94	2.94	4.53	2.16	1.07	0.23
CityBlock	0.25	1.80	3.64	3.49	0.43	0.24	0.24	0.57	1.72	0.97	0.41	1.68	0.02

Tabla 26. Análisis de errores mensuales (año referencia)

Por tanto sobre el año de referencia los resultados no son concluyentes, aunque si se observa la distribución de errores diarios se puede ver como correlation estaría descartado.

	sEuclidean		Cosine		Correlation		CityBlock	
Límite de error [%]	**Nº Casos**	**% de casos**	**Nº Casos**	**% de casos**	**Nº Casos**	**% de casos**	**Nº Casos**	**% de casos**
5	134	36.71	147	40.27	269	73.70	131	35.89
10	28	7.67	39	10.68	179	49.04	32	8.77
20	3	0.82	10	2.74	63	17.26	5	1.37
30	1	0.27	1	0.27	29	7.95	2	0.55
40	1	0.27	0	0.00	22	6.03	2	0.55
50	1	0.27	0	0.00	13	3.56	2	0.55

Tabla 27. Análisis diarios

La distancia "correlation" quedaría descartada para este tipo de aplicaciones. Las otras tres competirían en el año de validación. No obstante, la opción Cosine mueve la distribución de errores a la izquierda, es decir a una franja menor de 30%.

Si se analiza el año de validación, se tiene:

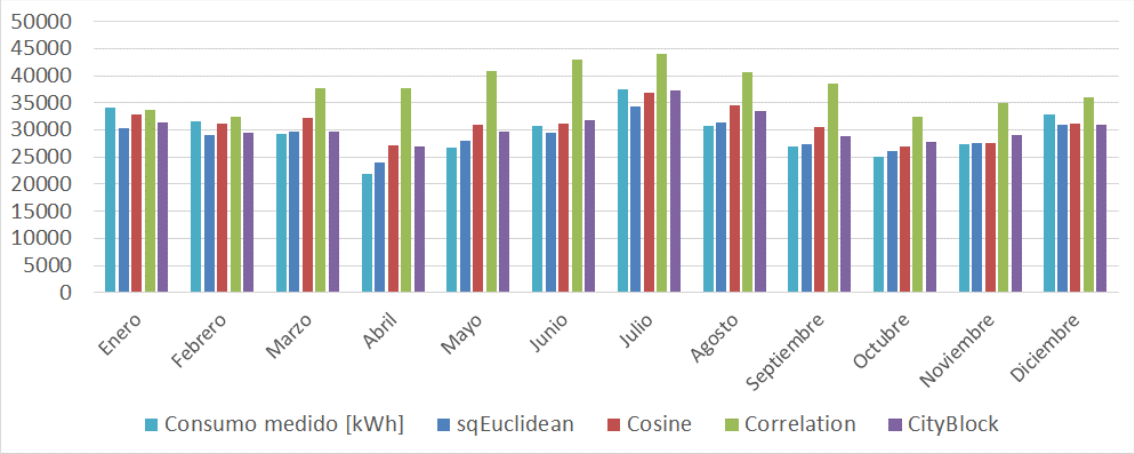

Figura 45. Resultados mensual de la comparación de modelos (año validación)

Correlation muestra que no es aplicable a la tipología de datos. Y aunque los errores mensuales se compensan y el error anual es aceptable en todos los casos, la distribución de errores diarios muestra como sEuclidean sería la opción más favorable, y la elegida por defecto. A su vez Cityblock sería la segunda opción a estudio.

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
sEuclidean	12.25	8.91	1.55	9.10	4.54	4.68	9.29	2.44	1.51	4.51	0.93	6.16	1.79
Cosine	4.08	1.29	9.36	19.51	13.88	1.53	1.56	11.21	11.21	6.92	0.99	4.99	4.99
Correlation	1.16	2.15	22.64	41.84	34.55	28.47	14.91	24.32	29.82	22.80	21.62	8.74	21.46
CityBlock	8.67	7.13	1.93	18.87	9.79	3.38	0.47	8.55	6.49	10.22	5.59	5.83	3.28

Tabla 28. Análisis de errores mensuales (año validación)

	sEuclidean		Cosine		Correlation		CityBlock	
Límite de error [%]	**Nº Casos**	**% de casos**	**Nº Casos**	**% de casos**	**Nº Casos**	**% de casos**	**Nº Casos**	**% de casos**
5	292	80.00	277	75.89	339	92.88	295	80.82
10	200	54.79	224	61.37	321	87.95	225	61.64
20	63	17.26	138	37.81	274	75.07	110	30.14
30	17	4.66	91	24.93	216	59.18	36	9.86
40	10	2.74	49	13.42	176	48.22	13	3.56
50	5	1.37	11	3.01	148	40.55	6	1.64

Tabla 29. Análisis de errores diarios (valor promedio apróx. 40%)

3.5.3 Conclusiones

La distancia juega un papel importante en el algoritmo de clustering propuesto y a su vez en la modelización de la línea base. Es por ello que el propio algoritmo analiza 4 opciones, de las cuales 1 de ellas queda descartada.

De las tres opciones preseleccionadas, hay que comentar que en los edificios probados la que mejores resultados ofrece en validación, tanto diaria como mensual es la sEuclidean.

3.6 ESTUDIO 5: COMPARACIÓN K-MEANS Y LINKAGE

3.6.1 Descripción y resultados

Para comparar los dos grandes métodos, kmeans y linkage, se van a usar las dos distancias con mejores resultados en valor promedio (cosine y sEuclidean). Sobre el año de referencia se obtiene el modelo 2 de la opción 2 (1d+3n) sin distinción de estaciones. Los resultados mensuales son los que aparecen en la siguiente gráficas.

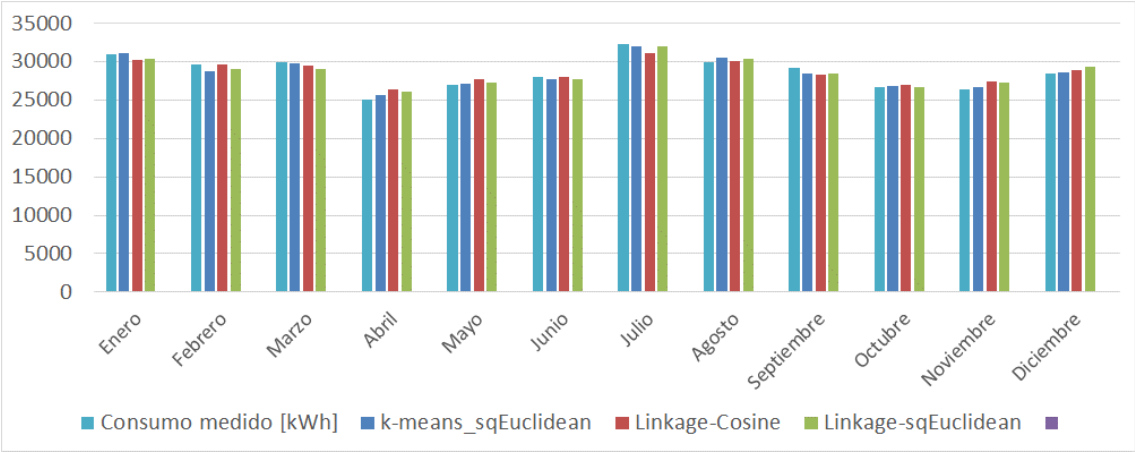

Figura 46. Resultados mensual de la comparación de modelos (año referencia)

Analizando los valores de los errores se puede ver como linkage también es un método aceptable, el problema es que menos automático que el método kmeans.

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
k-means_sEuclidean	0.61	2.97	0.61	2.20	0.60	1.00	0.84	1.72	2.17	0.77	0.76	0.63	0.05
Linkage-Cosine	2.18	0.29	1.68	4.75	2.76	0.01	3.79	0.34	3.10	1.07	3.73	1.42	0.21
Linkage-sEuclidean	1.84	2.08	3.31	3.98	1.41	0.68	1.01	1.17	2.31	0.14	3.01	2.84	0.03

Tabla 30. Análisis de errores mensuales (año referencia)

Finalmente la base de trabajo, base diaria, muestra como la distancia sEuclidean es la que mejores resultados ofrece también en Linkage

	k-means_sEuclidean		Linkage-Cosine		Linkage-sEuclidean	
Límite de error [%]	**Nº Casos**	**% de casos**	**Nº Casos**	**% de casos**	**Nº Casos**	**% de casos**
5	134	36.71	202	55.34	163	44.66
10	28	7.67	97	26.58	66	18.08
20	3	0.82	17	4.66	6	1.64
30	1	0.27	2	0.55	0	0.00
40	1	0.27	2	0.55	0	0.00
50	1	0.27	0	0.00	0	0.00

Tabla 31. Análisis diarios

A continuación se exponen los mapas de colores de clusters obtenidos con los diferentes algoritmos

Algoritmo 1: Linkage, nivel de inconsistencia I, cosine + Modelo 2 Opción 2 (1d+3n) sin uso de estaciones.

Algoritmo 2: K-means, Davis Bouldin, sEuclidean + Modelo 2 Opción 2 (1d+3n) sin uso de estaciones.

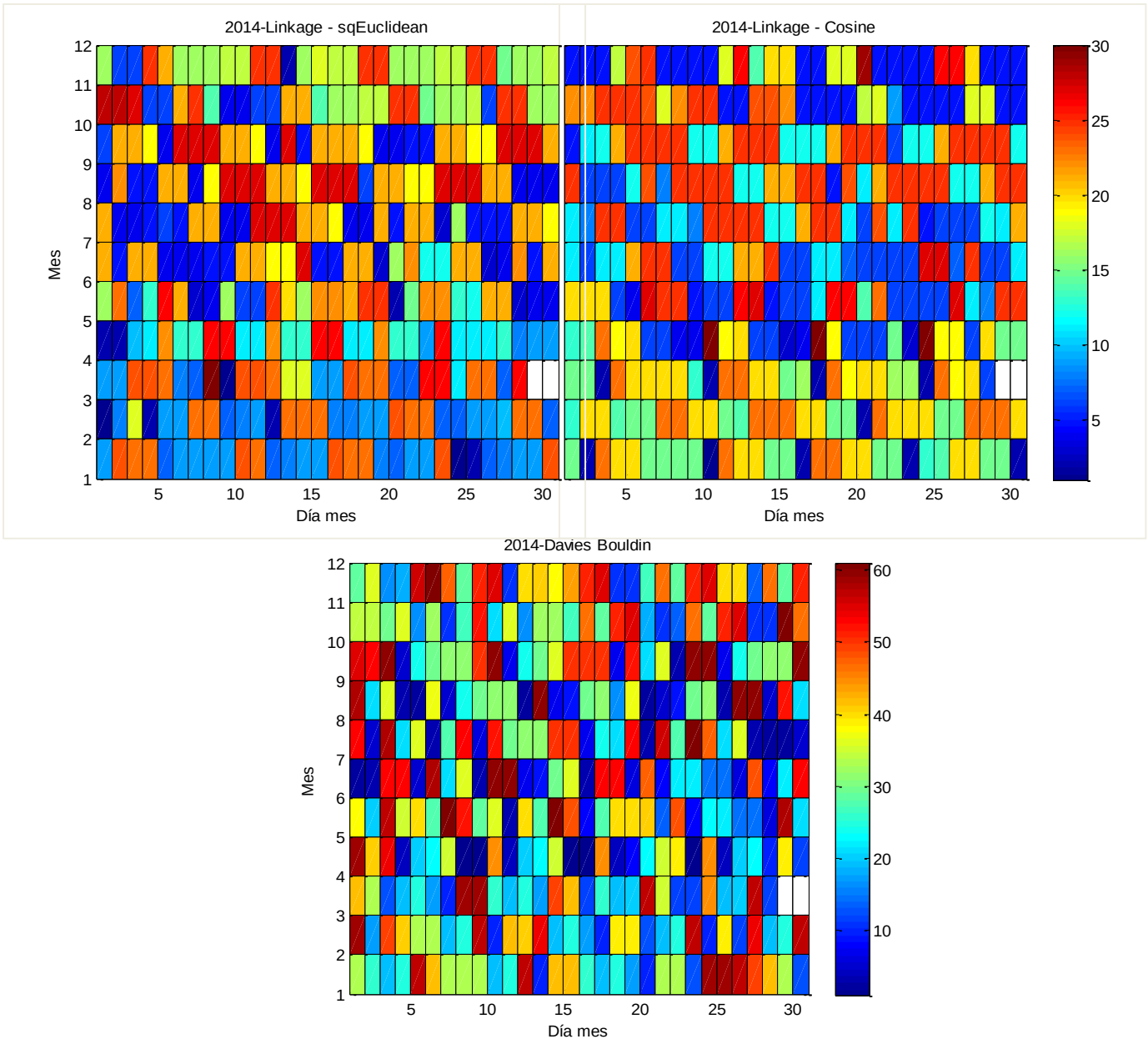

Figura 47. Comparación k-means Linkage mapa de colores

3.6.2 Conclusiones

Una vez elegidos las mejores duplas linkage+distancia, se debe proceder a la obtención del modelo de línea base. Este proceso devuelve cuál es el algoritmo completo óptimo en función del edificio.

Los resultados anteriores muestran que no hay grandes diferencias, tanto para este edificio como para el resto analizados, entre métodos jerárquicos y no jerárquicos. No obstante, los métodos jerárquicos requieren la decisión por parte del usuario del corte en el dendrograma. Este corte requiere un proceso de automatización que no es trivial, ya que la mayoría de las herramientas aproximan el corte recomendado. Habría que analizar si el coeficiente de correlación confenético y el nivel de inconsistencia I (una especie de medida de distancia), ver su tendencia y cómo obtener un criterio de corte.

Esto último se entiende como una línea futura a analizar y trabajar, pero aun así, el método no jerárquico kmeans ofrece las suficientes posibilidades y resultados como para convertirse en el candidato principal para el algoritmo global de obtención de líneas base.

3.7 ESTUDIO 6: INFLUENCIA DEL AÑO DE REFERENCIA

3.7.1 Descripción

Tomando K-means, el índice Davies Bouldin, se evalúa la importancia de la elección de la referenica, tanto a nivel del resultado del algoritmo de clustering como de los ajustes de la línea base. El modelo de ajuste se toma el modelo 2 de la opción 2, es decir, 1 denominador y 3 numeradores sin distinguir estaciones.

3.7.2 Resultados

Este estudio persigue demostrar la importancia del año de referencia y de la buena elección de un año de validación. Si el año de validación, como muestra la siguiente gráfica, no presenta una dependencia igual que el año de referencia en la parte no medida (comportamiento del usuario), significaría que no es un año de validación apropiado, si no que debería ser tomado como una nueva referencia.

Por ejemplo en el edificio audiencia provincial, aunque el nivel de consumo es parecido en ambos años, se puede ver que las relaciones entre consumo, día de operación y clima son diferentes. Esto significa que el comportamiento del usuario puede haber cambiado y el peso de este comportamiento en el consumo es bastante relevante.

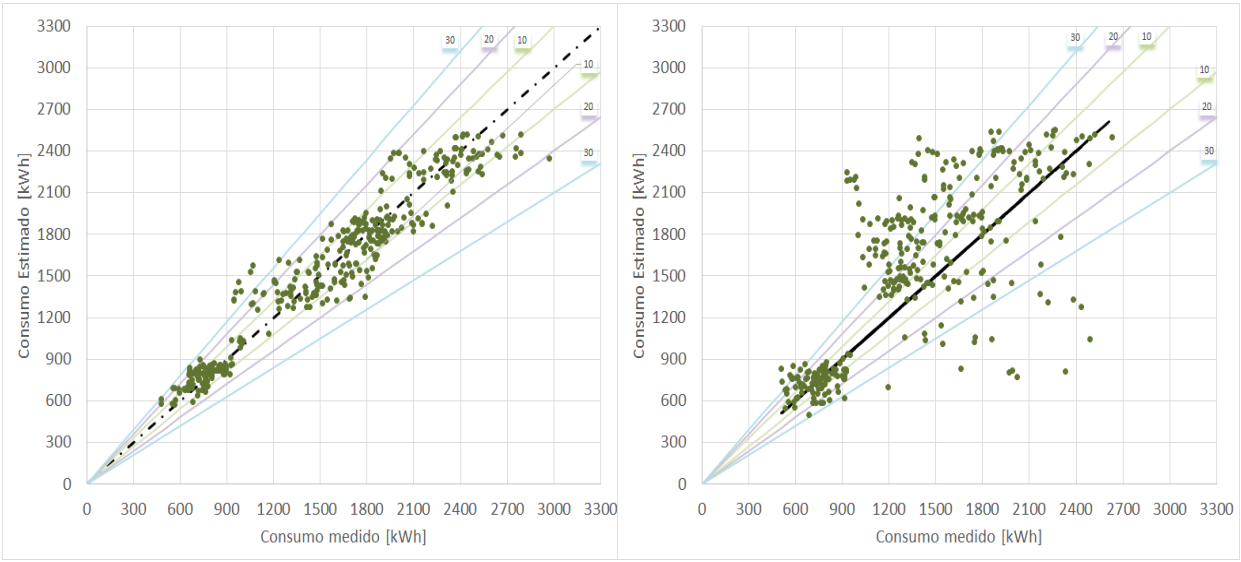

Figura 48. Ejecución del modelo en el edificio Audiencia Provincial (año referencia– año validación)

El resultado anterior se conoce porque al adjudicar cluster a 2015 sin hacer uso del consumo aparecen unas diferencias que por ejemplo en el Edificio Agencia andaluza no se observan. Esto se debe a que el comportamiento del usuario en este edificio tiene una importancia mayor que en el edificio usando para el resto de estudios. Es decir, el algoritmo de clustering tiene en cuenta al usuario de manera implícita al definir los centroides del consumo, el número de clúster y los componentes de ese clúster. Es por ello que si el usuario certificara que el año 2015 es de referencia, el consumo estaría en condiciones de referencia y podría ser usado para la tipificación de días. Los resultados que se tendría en ese caso se muestran en la siguiente figura:

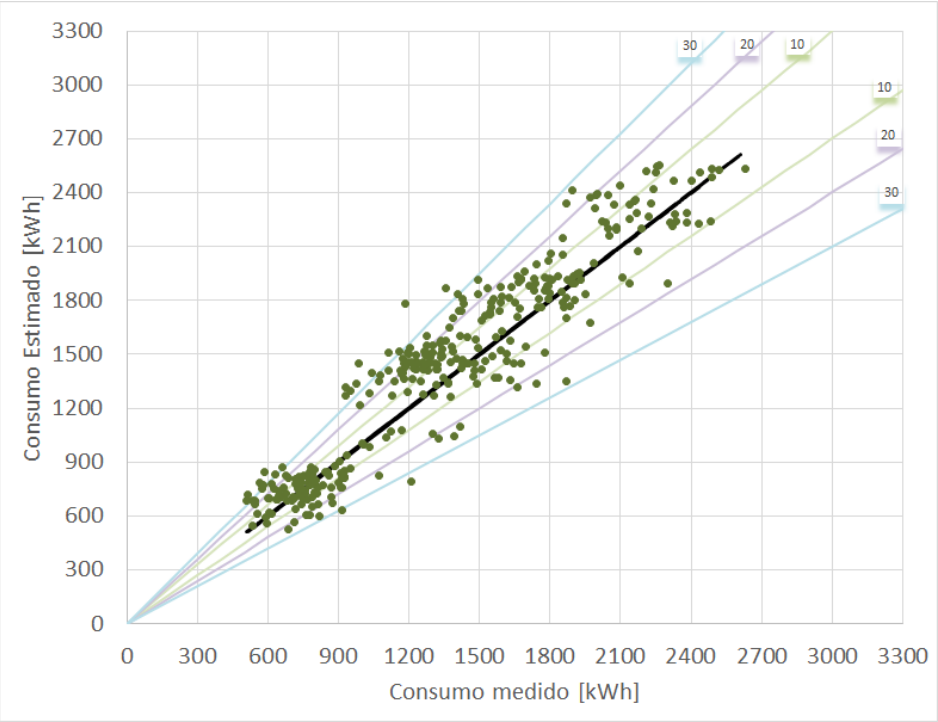

Figura 49. Ejecución del modelo en el edificio Audiencia Provincial con 2015 en referencia

Los resultados son idénticos a los que se tienen en el edificio Agencia Andaluza para el año de validación. Sin embargo, en la Agencia Andaluza no se ha tenido en cuenta el consumo del año de validación para la selección del clúster que se asocia a cada día. Este hecho es el que prueba que dependiendo del edificio el peso en la tipificación de días del usuario puede ser tan importante que haga pensar en intentar descomponerlo para poder tenerlo en cuenta de manera explícita.

La verificación de lo anterior, es decir, de la importancia del usuario en el consumo queda probada cuando se tipifican los días del año de referencia sin usar el valor del consumo en esos días. Si la tipificación resulta idéntica a la obtenida con el consumo quiere decir que la influencia del usuario no es acusada lo suficientemente acusada como para tenerla de forma explícita.

Por tanto, en el año 2015 ha cambiado el comportamiento del usuario y el peso del mismo en el consumo del edificio, de ahí los resultados que se tenían en la primera gráfica. Sin embargo, si se considera que el comportamiento del usuario es el de referencia y se usa el consumo del año 2015 para la tipificación resulta que el ajuste del modelo es de una calidad idéntica al de referencia. Esto significa que el peso del usuario en el consumo sigue siendo el mismo pero su comportamiento ha cambiado, este dato se observa en el consumo de los fines de semana que en año 2015 son sensiblemente mayores.

Ahora bien el óptimo en este caso, sería considerar 2014 y 2015 en la fase inicial del algoritmo, de tal forma que el clustering se realice con ambos años y se pueda caracterizar de forma implícita con mayor exactitud el peso del usuario.

Estas diferencias son del mismo orden de las que aparecen cuando se usa el algoritmo de clustering específico de un edificio en otro totalmente diferente con un clima diferente. Para probar esta afirmación, y con ella, la muestra de que un año en unas condiciones diferentes a las de referencia puede ser considerado como los datos de un edificio diferente, se presentan los resultados del clustering del CHARE de Ecija aplicado al CHARE de Utrera.

El modelo de BS de Ecija para el año de referencia ofrece los siguientes resultados:

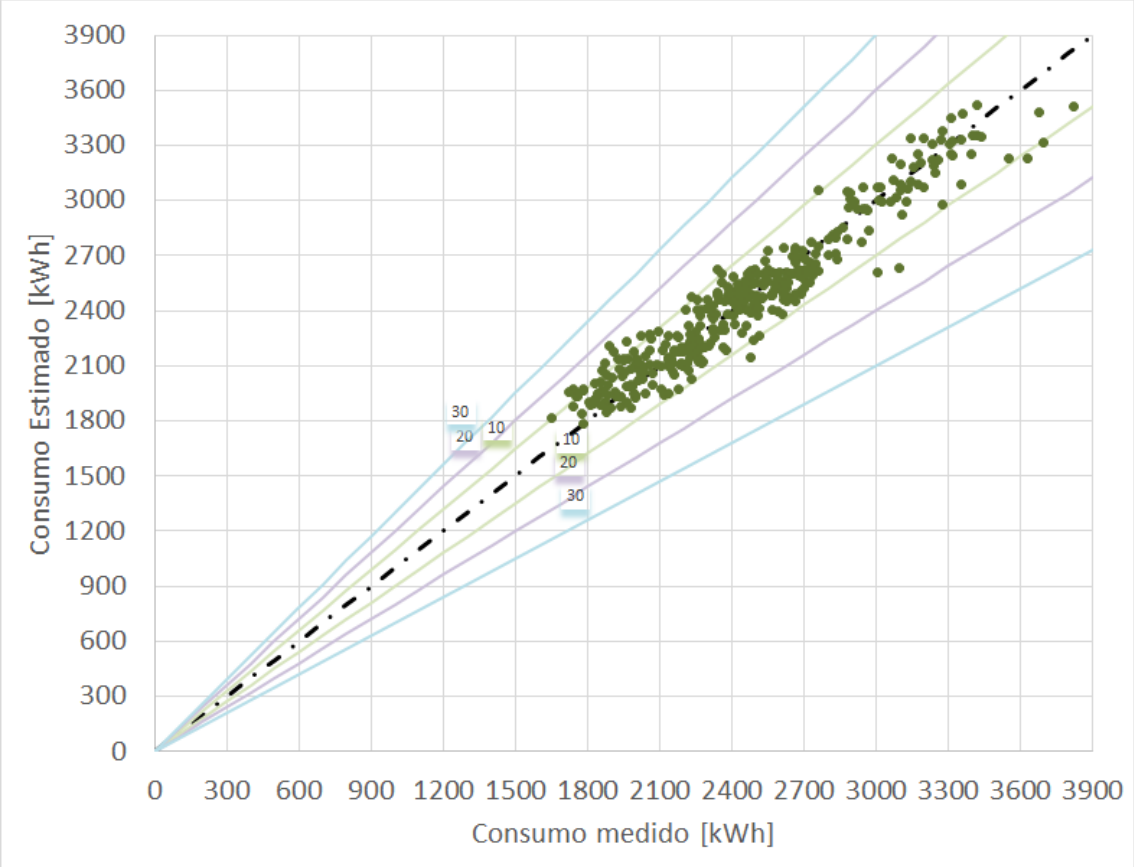

Figura 50. Modelo de Base Line de Écija para el año de referencia de Écija

Si el modelo anterior se aplica al supuesto año de validación resulta:

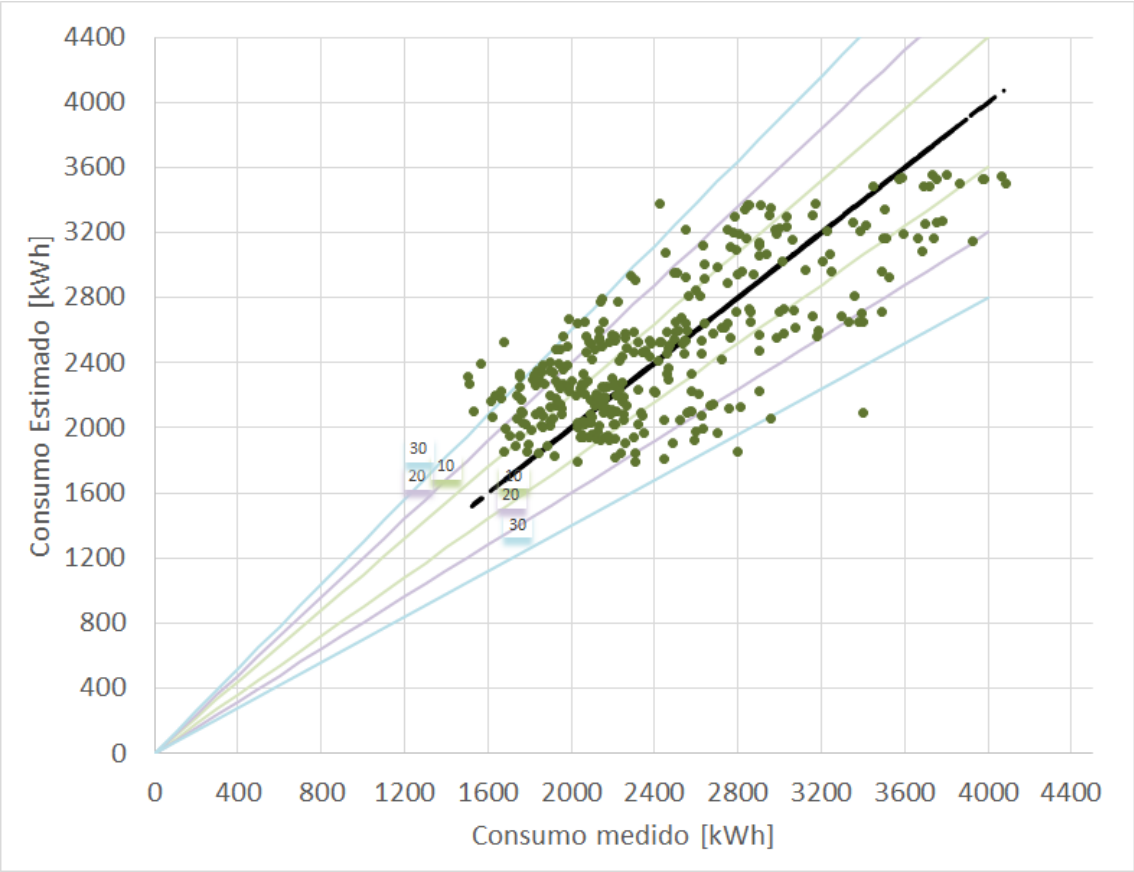

Figura 51. Modelo de Base Line de Écija para el año de validación de Écija

Y ahora se compara el resultado de aplicar el modelo anterior a otro clima y otro edificio de igual uso, en este caso al CHARE de Utrera para el año 2014.

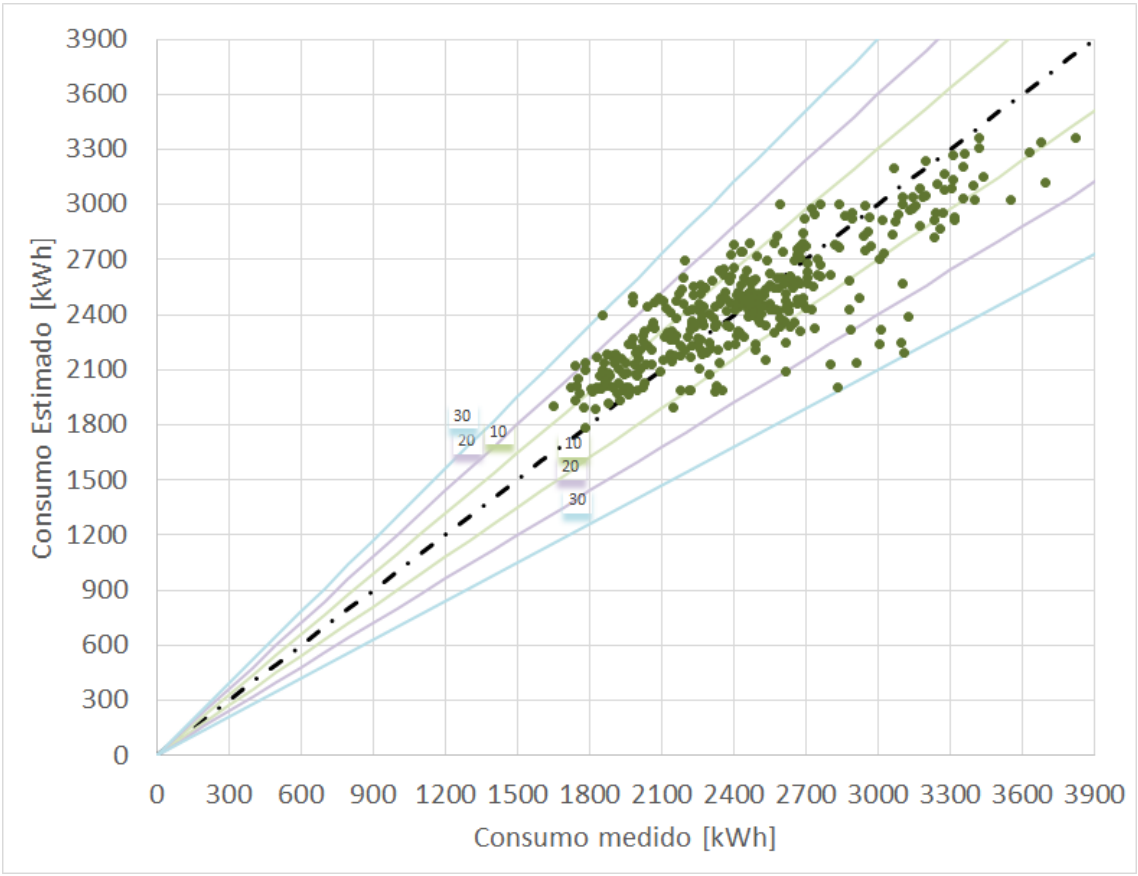

Figura 52. Modelo de Base Line de Écija para el año de referencia de Utrera

Se observa como este gráfico se parece al gráfico anterior, con lo que se demuestra que un año en condiciones diferentes a las de referencia devuelve resultados iguales o peores que la aplicación del modelo de un edificio a otro edificio en otras condiciones climáticas.

3.7.3 Conclusiones

Las variables no medidas están implícitas en el modelo, tanto en la tipificación de días donde cada día tiene asociado un consumo de referencia como en los coeficientes del modelo.

Se ha demostrado que si el año de validación tiene unas condiciones idénticas a la referencia, es decir, condiciones de referencia, los resultados son los esperables; tal y como aparece en el edificio de la agencia andaluza. En cambio cuando el edificio de un año a otro sufre grandes cambios, como la audiencia provincial, se requiere decidir el periodo de validación (2014, 2015 o 2014+2015) para que la línea base tenga sentido.

En tal caso el propio sentido del procedimiento queda demostrado cuando el cliente certifica que 2015 está en condiciones de referencia y el consumo puede ser introducido en el algoritmo como entrada para tipificar días por estar en condiciones de referencia. En ese caso la calidad de los resultados es aceptable.

Un mal año de validación es idéntico a aplicar el modelo de un edificio en otro.

4 APLICACIONES

Las distintas soluciones obtenidas son tomada como modelo de Base Line: Kmeans, Davies Bouldin, seuclidean y modelo opción 2, es decir, un denominador con tres numeradores y sin hacer distinción de estaciones.

4.1 EDIFICIO 2: SEDE JUDICIAL

La Sede Judicial de Sevilla se compone de dos edificios, Audiencia Provincial y Juzgados, ambos de características constructivas y localización muy similares, situados junto al Prado de San Sebastián, Sevilla. Sin embargo, al tener actividades diferentes los trataremos de forma independiente.

Figura 53. Ubicación edificios de la sede judicial

4.1.1 EDIFICIO 2.1: AUDIENCIA PROVINCIAL

4.1.1.1 DESCRIPCIÓN

Las Audiencias Provinciales son órganos jurisdiccionales colegiados, de demarcación provincial y con sede en la capital de la provincia, en este caso Sevilla, con competencia en los órdenes jurisdiccionales civil y penal.

La superficie construida total es de 12000 m², distribuida en las siguientes dependencias:

- Sótano: Donde se encuentran los calabozos, sala de calderas, archivos y despacho de investigadores y archivero.
- Plantas de la primera a la sexta: Distintas secretarías, despachos de presidente, magistrados, jueces, fiscales, secretarios, salas de plenos, de vistas,..
- Planta séptima: Viviendas del presidente y del fiscal de la audiencia.

La dotación personal es del orden de 350 personas.

Como podemos observar al tener múltiples actividades dentro de su competencia, no tiene un horario ni una ocupación diaria fijos, esto no ayudará a la caracterización del consumo.

Figura 54. Fachada edificio Audiencia Provincial

4.1.1.2 LÍNEA BASE

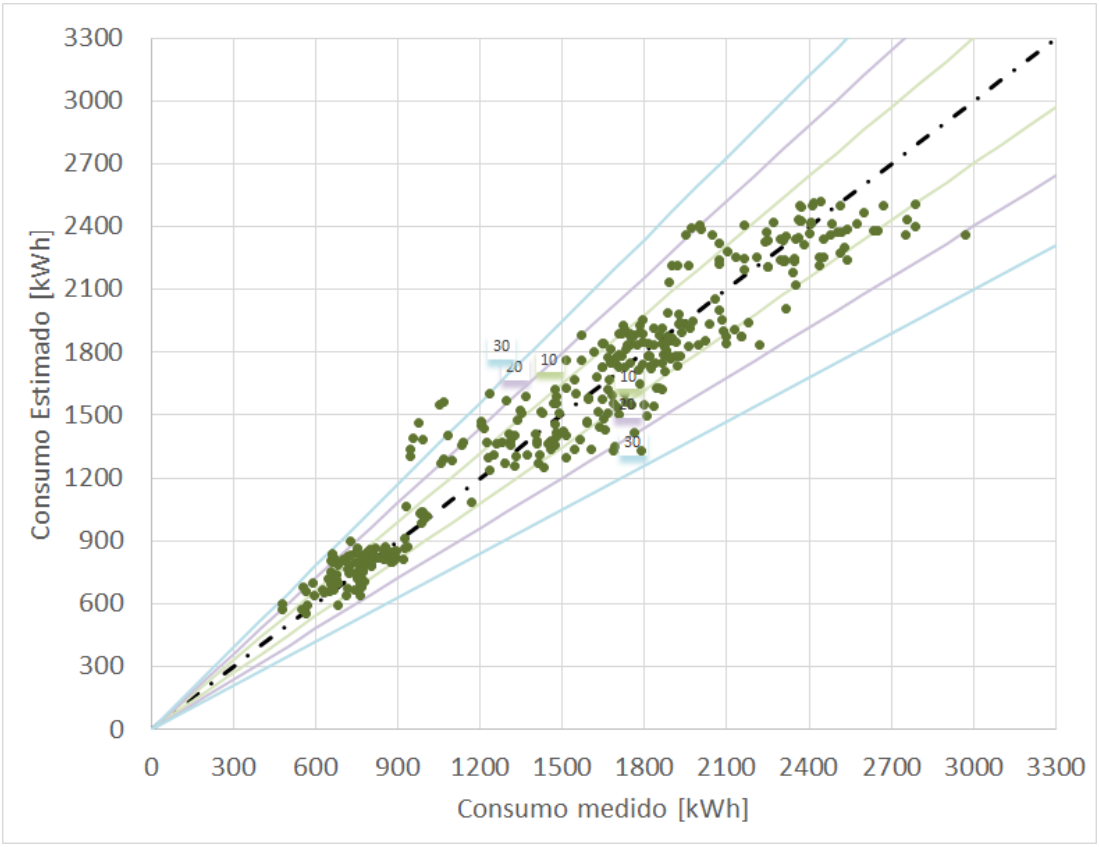

Figura 55. Modelo BS año referencia 2014 Audiencia Provincial

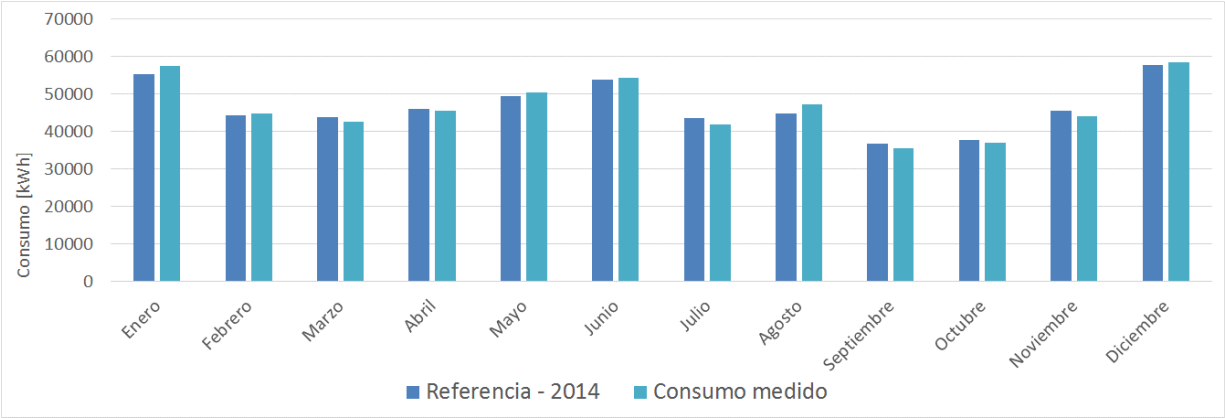

Figura 56. Evaluación consumos modelo BS año referencia 2014 Audiencia Provincial

Error promedio diario [%]		8.4
Límite de error [%]	**Nº Casos**	**% de casos**
5	228	62.5
10	112	30.7
20	23	6.3
30	9	2.5
40	2	0.5
50	0	0.0

Tabla 32. Evaluación errores modelo BS año referencia 2014 Audiencia Provincial

VALIDACIÓN 2015 (suponiendo que en dicho año el edificio se encuentra en condiciones de referencia).

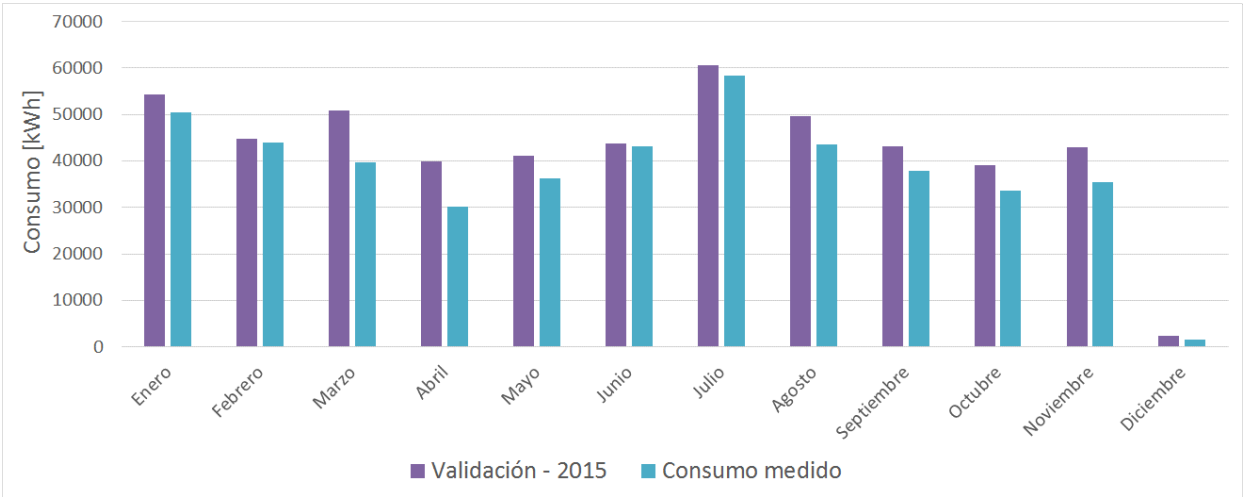

Figura 57. Evaluación consumos modelo BS año validación 2015 Audiencia Provincial

Error promedio diario [%]		24.0
Límite de error [%]	Nº Casos	% de casos
5	283	77.5
10	239	65.5
20	155	42.5
30	101	27.7
40	65	17.8
50	39	10.7

Tabla 33. Evaluación errores modelo BS año validación 2015 Audiencia Provincial

Tabla 34. Comparativa de errores modelo BS Audiencia Provincial

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
Audiencia Provincial-2014	4.08	1.13	3.14	1.26	1.84	0.76	4.35	5.15	3.48	1.94	2.87	1.25	0.01
Audiencia Provincial-2015	6.80	1.70	21.60	24.11	11.51	1.48	3.83	12.34	11.79	13.74	17.29	32.54	11.22

4.1.2 EDIFICIO 2.2: JUZGADOS

4.1.2.1 DESCRIPCIÓN

Los Juzgados ofrecen la primera respuesta a los ciudadanos en los conflictos que pudieran suscitarse, tanto civil como penal. La superficie construida del edificio es de 9700 m². Donde las primeras estancias son las siguientes:

- Sótano: En ella se encuentran los archivos, consultas de médicos, local sindical, calabozos, cuarto del cuadro eléctrico y aseos.
- Planta baja: Juzgados de incidencias, despacho de fiscal, juez, secretarios, oficina y despacho del Sr. Juez de vigilancia penitenciaria, despachos de Sres. Secretarios, dormitorios, registro general, registro civil, sección gubernativa y sus dependencias, decano, policía judicial, archivos y peritos, averiguaciones patrimoniales, aseos, estanco y oficina de correos.
- Planta primera, segunda y tercera: Secretarias civiles, secretarias criminales, despachos de Sres. jueces, fiscales, Sres. Magistrados, Sres. secretarios, salas de vistas, mutualidad, agentes, pequeños archivos y aseos.

Figura 58. Edificio Juzgados

La dotación del personal es de aproximadamente 500 personas. El centro funciona de 8 a 15 horas todos los días del año, excepto fines de semana y festivos. Por la tarde, los juzgados de guardia, dentro de su horario de 24 horas, y el servicio de limpieza de 14 a 22 horas.

Como podemos observar este tipo de edificios no representa uno típicamente de oficinas, no tiene un horario fijo y los que es más destacado la ocupación no es constante, es muy variable, dependiendo de casos excepcionales. Es difícil poder llegar a caracterizar correctamente el consumo en estos casos.

Por eso hemos trabajado con dos tipos de edificios muy diferentes, que dicho contraste nos puede ayudar a encontrar situaciones distintas en el proyecto.

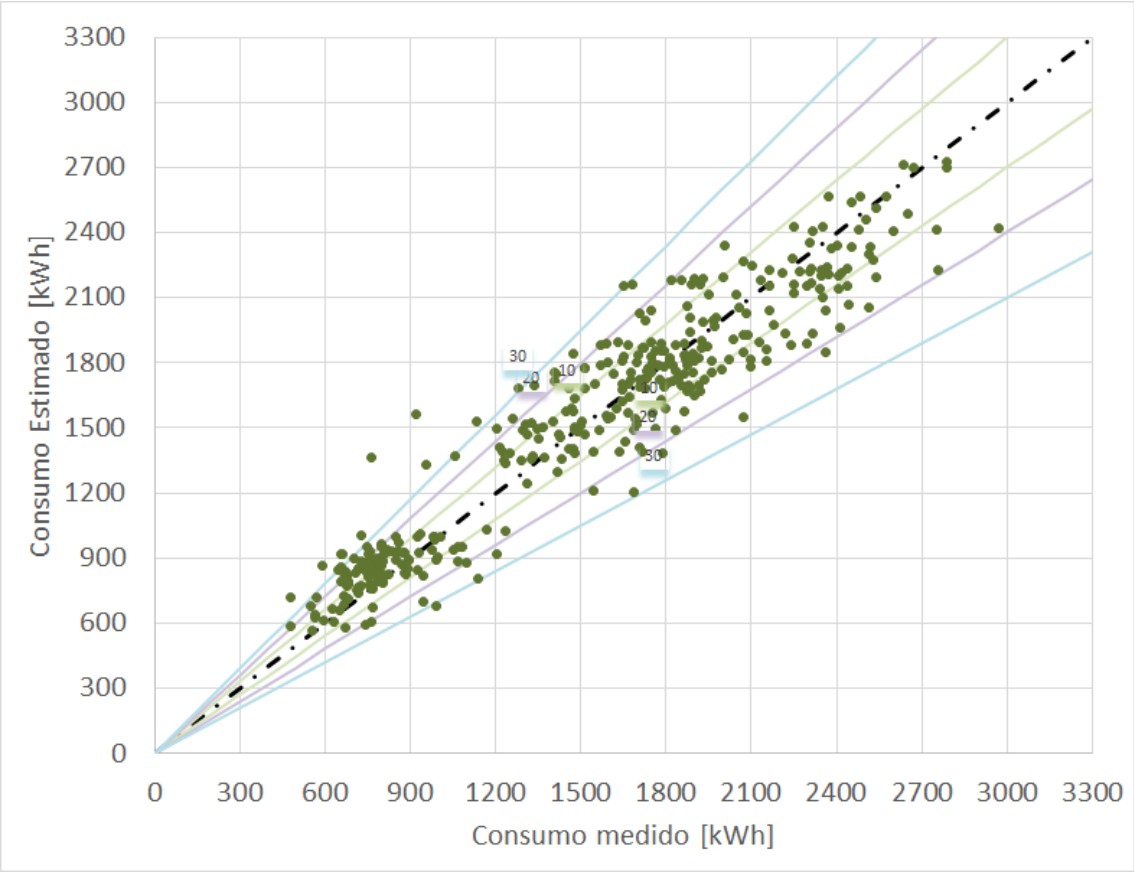

Figura 59. Modelo BS año referencia 2014 Juzgados

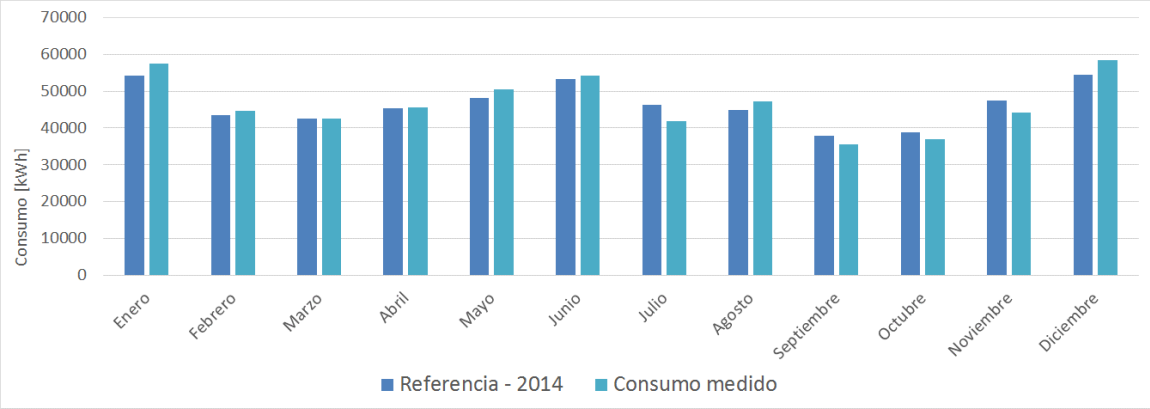

Figura 60. Evaluación consumos modelo BS año referencia 2014 Juzgados

Error promedio diario [%]		10.5
Límite de error [%]	**Nº Casos**	**% de casos**
5	250	68.5
10	149	40.8
20	50	13.7
30	12	3.3
40	5	1.4
50	2	0.5

Tabla 35. Evaluación errores modelo BS año referencia 2014 Juzgados

VALIDACIÓN 2015 (suponiendo que en dicho año el edificio se encuentra en condiciones de referencia).

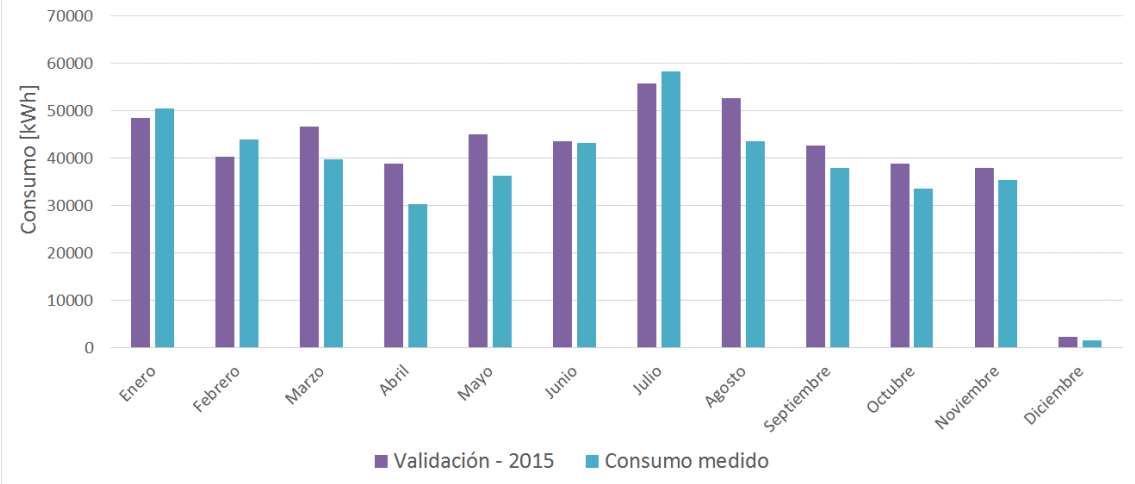

Figura 61. Evaluación consumos modelo BS año validación 2015 Juzgados

Error promedio diario [%]		28.8
Límite de error [%]	**Nº Casos**	**% de casos**
5	283	77.5
10	245	67.1
20	171	46.8
30	123	33.7
40	85	23.3
50	52	14.2

Tabla 36. Evaluación errores modelo BS año validación 2015 Juzgados

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
Juzgados- 2014	5.68	3.11	0.41	0.15	4.47	1.78	9.89	5.32	6.36	4.54	6.91	7.49	0.33
Juzgados-2015	4.20	9.16	14.87	22.04	19.46	0.69	4.63	17.30	10.96	13.35	6.76	31.33	7.81

Tabla 37. Comparativa de errores modelo BS Juzgados

4.2 EDIFICIO 4: HOSPITAL ÉCIJA

4.2.1 DESCRIPCIÓN

El Hospital de Alta Resolución de Écija se encuentra en la localidad de Écija, Sevilla, concretamente en la avenida Doctor Sánchez Malo, s/n, CP 41400. Se trata de un complejo hospitalario. Se adjunta documentación gráfica donde se observa dicha localización.

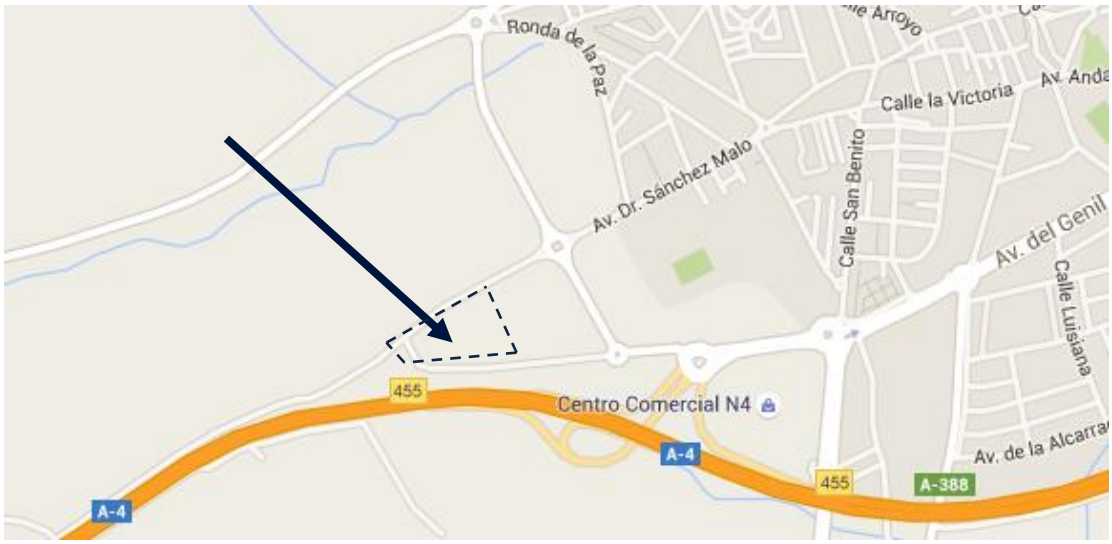

Figura 62. Ubicación hospital Écija

El Hospital de Alta Resolución de Écija, en adelante Hospital de Écija, es un edificio construido en el año 1961. La última reforma de importancia fue realizada en el año 2006.

El Hospital posee una superficie construida de 6118 m^2, con una superficie útil de 5690 m^2.

Figura 63. Hospital Écija

El edificio dispone de planta baja, primera planta, segunda planta y azotea. En la planta baja se encuentran diferentes consultas, salas de urgencia, quirófanos, cocina, almacenes, etc. En la primera planta se encuentran diversos laboratorios, salas de cirugía general, despachos y salas de reuniones. En la segunda planta se

encuentran diferentes habitaciones. En la azotea se encuentran diversos despachos.

4.2.2 LÍNEA BASE

Solución obtenida tomada como modelo de Base Line: Kmeans, Davies Bouldin, seuclidean y modelo opción 2, es decir, un denominador con tres numeradores y sin hacer distinción de estaciones.

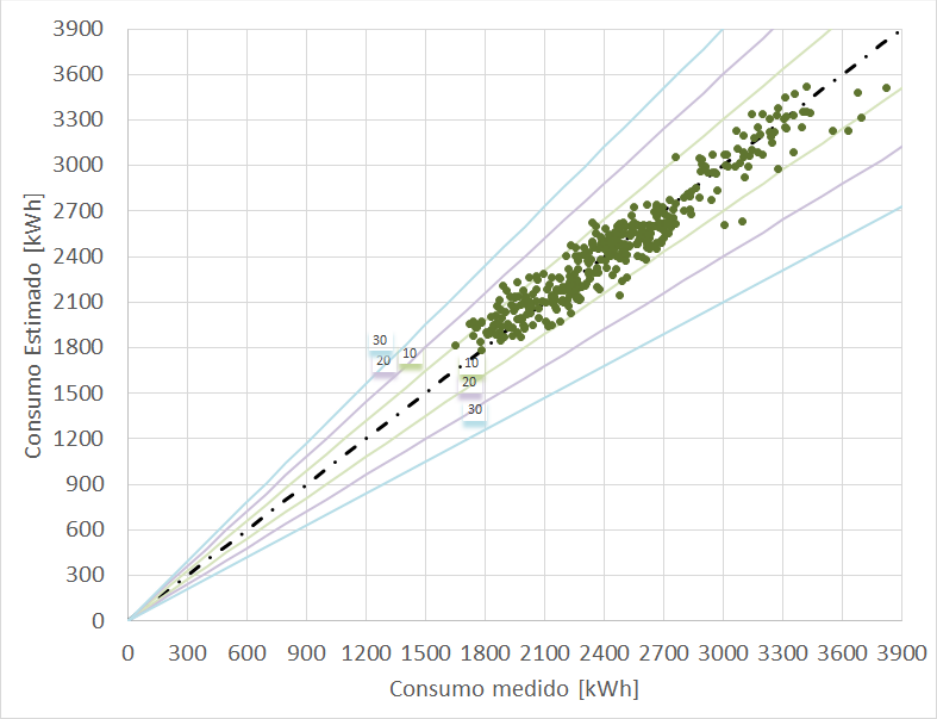

Figura 64. Modelo BS año referencia 2014 Écija

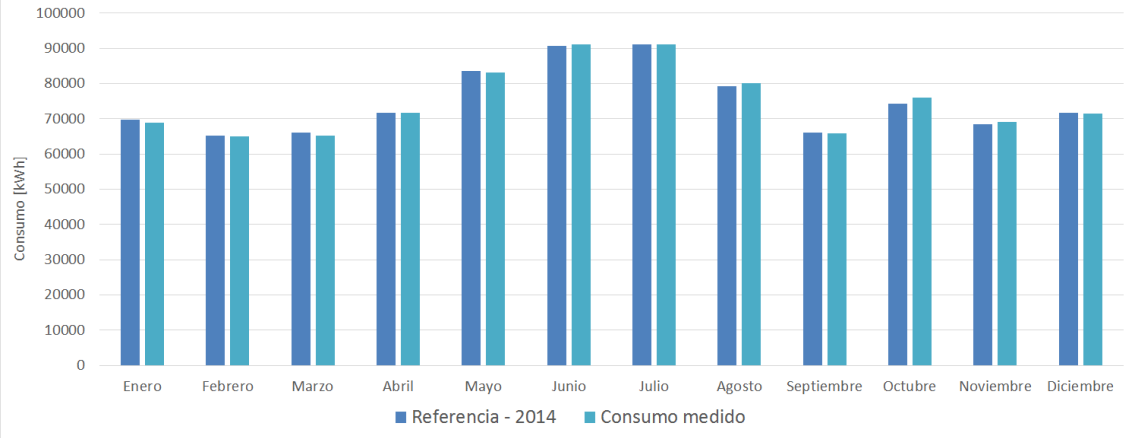

Figura 65. Evaluación consumos modelo BS año referencia 2014 Écija

Error promedio diario [%]	3.9

Límite de error [%]	Nº Casos	% de casos
5	102	27.9
10	20	5.5
20	0	0.0
30	0	0.0
40	0	0.0
50	0	0.0

Tabla 38. Evaluación errores modelo BS año referencia 2014 Écija

VALIDACIÓN 2015 (suponiendo que en dicho año el edificio se encuentra en condiciones de referencia).

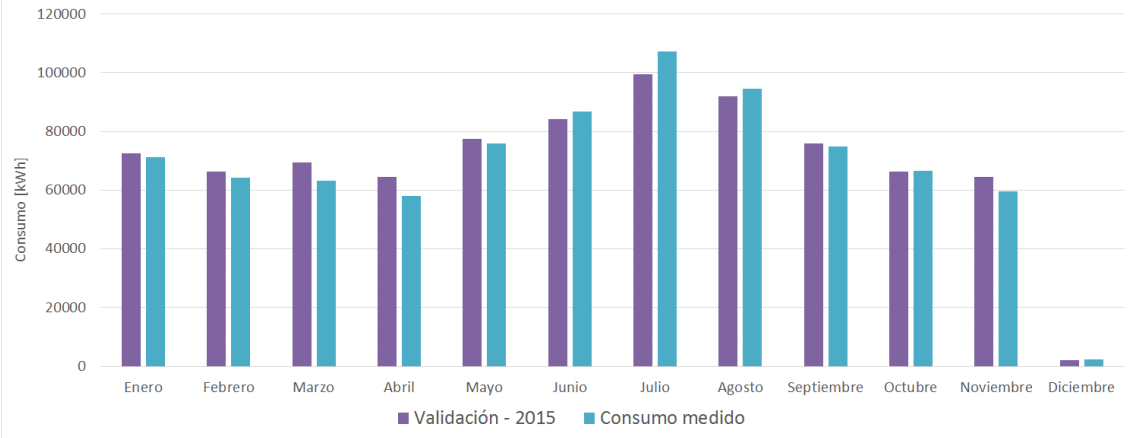

Figura 66. Evaluación consumos modelo BS año validación 2015 Écija

Error promedio [%]	12.9

Límite de error [%]	Nº Casos	% de casos
5	257	70.4
10	177	48.5
20	69	18.9
30	17	4.7
40	4	1.1
50	4	1.1

Tabla 39. Evaluación errores modelo BS año validación 2015 Écija

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
Ecija-2014	1.45	0.38	1.36	0.05	0.65	0.47	0.02	0.98	0.41	2.18	0.86	0.58	0.00
Ecija-2015	1.97	3.13	9.22	9.92	2.03	2.99	7.58	2.75	1.34	0.50	7.77	6.00	1.30

Tabla 40. Comparativa de errores modelo BS Écija

4.3 EDIFICIO 5: HOSPITAL SIERRA NORTE

4.3.1 DESCRIPCIÓN

El Hospital de Alta Resolución de Sierra Norte se encuentra en Constantina, Sevilla, concretamente en la calle Doctor Larrauri s/n, CP 41450. Se trata de un complejo hospitalario. Se adjunta documentación gráfica donde se observa dicha localización.

Figura 67. Ubicación hospital Sierra Norte

El Hospital de Alta Resolución de Sierra Norte, en adelante Hospital de Sierra Norte, es un edificio construido en el año 2007. No se ha realizado ninguna reforma de importancia desde su construcción.

El hospital posee una superficie construida de 6700 m^2, con una superficie útil de 2037.05 m^2.

Figura 68. Hospital Sierra Norte

El edificio dispone de planta baja, primera planta, segunda planta, planta -1 y planta -2. En la planta -2 se

encuentran los laboratorios, unos aseos, un almacén y la oficina del almacén. En la planta -1 se encuentra un área de fisioterapia, vestuarios, almacenes y aseos. En la planta baja se encuentran diferentes consultas, zona de urgencias, aseos, la recepción, salas de espera, etc. En la primera planta se encuentran diversas consultas, quirófanos, aseos, almacenes y áreas de descanso. En la segunda planta se encuentran diferentes almacenes y aseos.

4.3.2 LÍNEA BASE

Solución obtenida tomada como modelo de Base Line: Kmeans, Davies Bouldin, seuclidean y modelo opción 2, es decir, un denominador con tres numeradores y sin hacer distinción de estaciones.

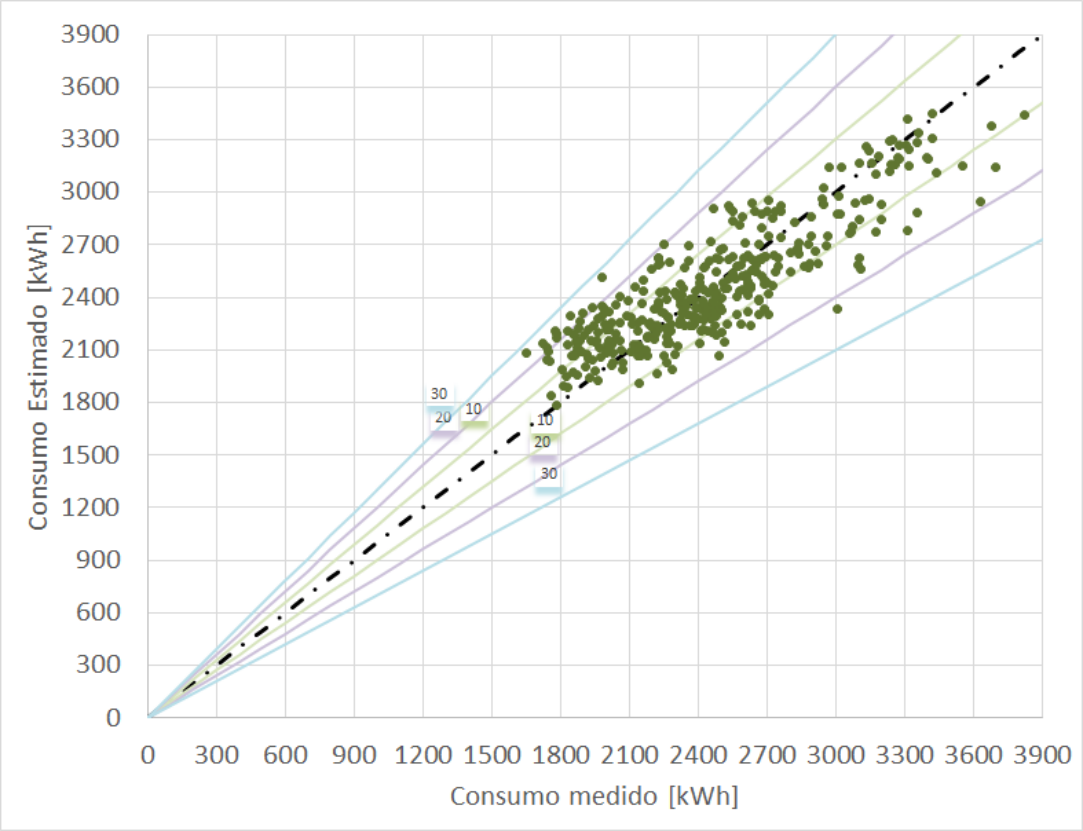

Figura 69. Modelo BS año referencia 2014 Sierra Norte

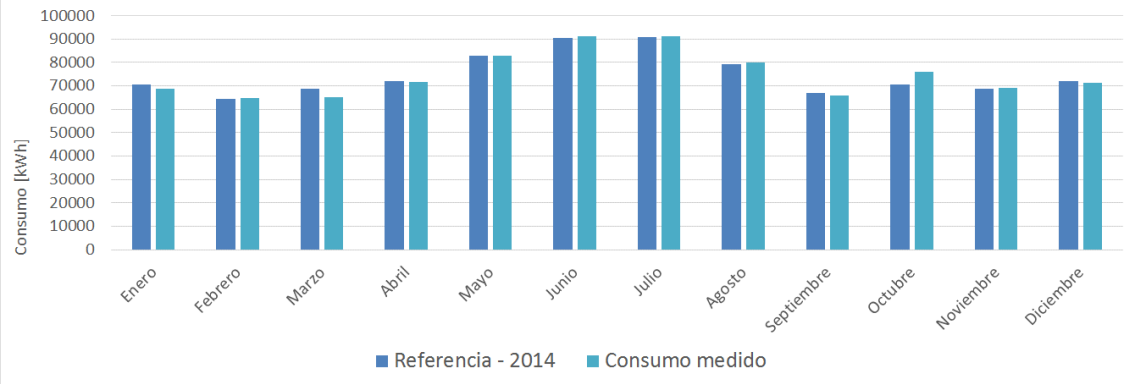

Figura 70. Evaluación consumos modelo BS año referencia 2014 Sierra Norte

Error promedio diario [%]		7.1
Límite de error [%]	**Nº Casos**	**% de casos**
5	212	58.1
10	99	27.1
20	12	3.3
30	0	0.0
40	0	0.0
50	0	0.0

Tabla 41. Evaluación errores modelo BS año referencia 2014 Sierra Norte

VALIDACIÓN 2015 (suponiendo que en dicho año el edificio se encuentra en condiciones de referencia).

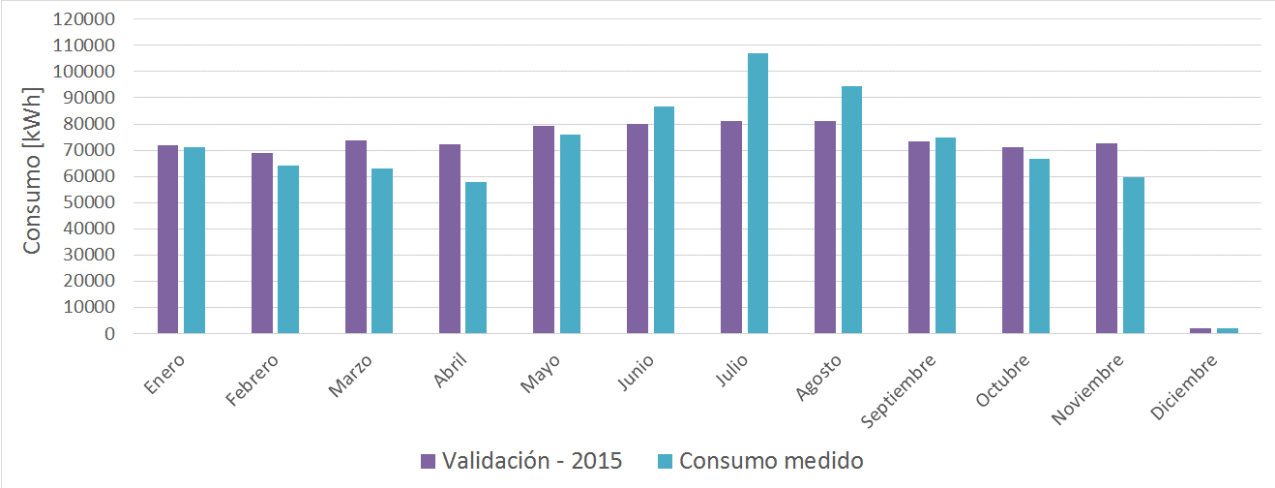

Figura 71. Evaluación consumos modelo BS año validación 2015 Sierra Norte

Error promedio diario [%]		19.2
Límite de error [%]	Nº Casos	% de casos
5	268	73.4
10	213	58.4
20	127	34.8
30	75	20.5
40	29	7.9
50	16	4.4

Tabla 42. Evaluación errores modelo BS año validación 2015 Sierra Norte

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
Sierra Norte-2014	2.30	0.85	5.13	0.45	0.37	0.60	0.17	0.81	1.80	7.83	0.42	0.78	0.08
Sierra Norte-2015	1.15	6.89	14.35	19.84	4.37	8.72	32.15	16.32	2.22	6.52	17.85	3.02	0.43

Tabla 43. Comparativa de errores modelo BS Sierra Norte

4.4 EDIFICIO 6: HOSPITAL UTRERA

4.4.1 DESCRIPCIÓN

El HOSPITAL DE ALTA RESOLUCIÓN DE UTRERA se encuentra en la localidad de Utrera, Sevilla, concretamente en la avenida Brigadas Internacionales s/n, CP 41710. Se adjunta documentación gráfica donde se observa dicha localización.

Figura 72. Ubicación hospital Utrera

El Hospital de Alta Resolución de Utrera es un edificio construido en el año 2005. Las distintas instalaciones se han ido inaugurando a lo largo del año 2007 hasta entrar en pleno funcionamiento el pasado año.

Figura 73. Hospital Utrera

El Hospital está compuesto por cinco módulos unidos entre ellos por pasillos. El edificio dispone de una planta sótano, una planta baja, primera planta y azotea.

En la planta sótano se encuentran despachos y salas técnicas, cocina, lavandería, vestuarios y aparcamiento privado.

En la planta baja se encuentran diversas oficinas y despachos de administración del hospital, la recepción, salas de consultas externas, una zona para urgencias, salas de radioanálisis, etc. En la primera planta se encuentra la hospitalización polivalente, la farmacia, un área quirúrgica y diversas salas de espera. En la azotea se encuentran instaladas diferentes plantas enfriadoras para la climatización del recinto y placas solares térmicas para ACS. En el exterior se encuentra un helipuerto y diferentes accesos en zona de urgencias y entrada principal del hospital.

4.4.2 LÍNEA BASE

Solución obtenida tomada como modelo de Base Line: Kmeans, Davies Bouldin, seuclidean y modelo opción 2, es decir, un denominador con tres numeradores y sin hacer distinción de estaciones.

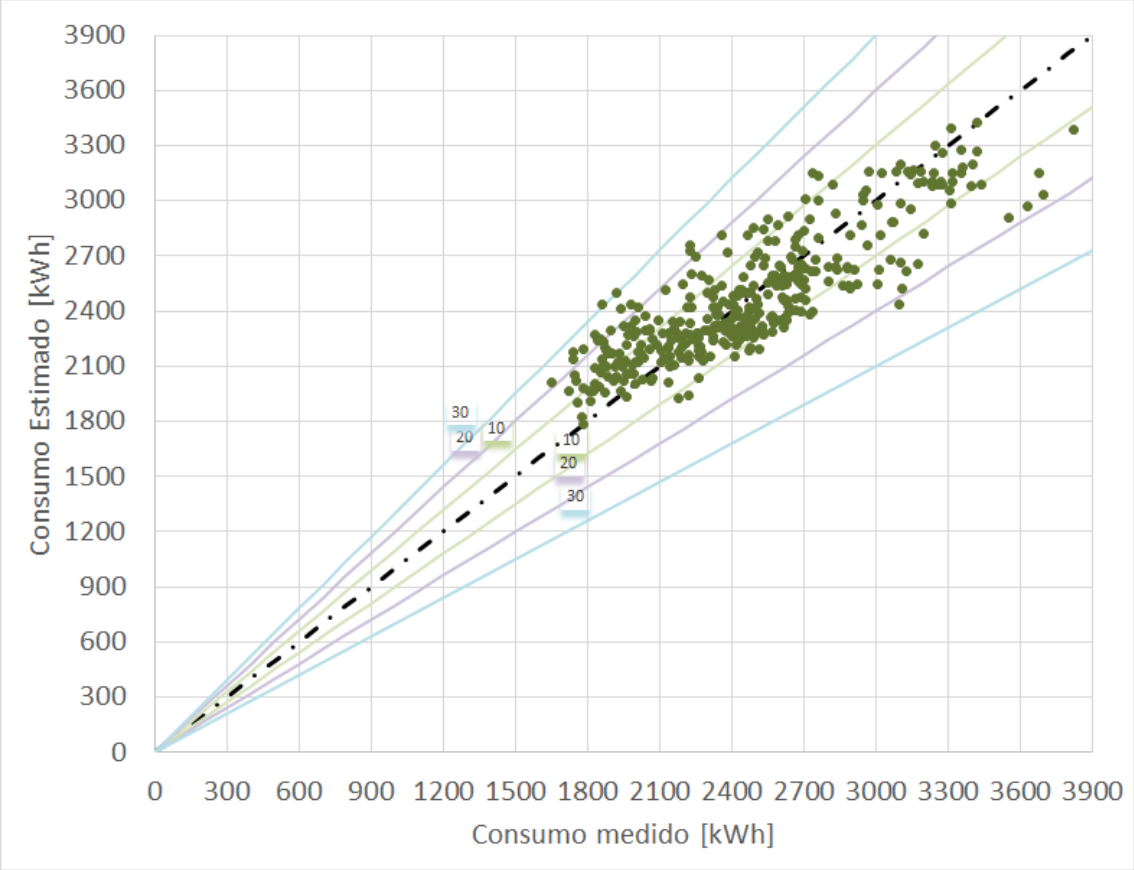

Figura 74. Modelo BS año referencia 2014 Utrera

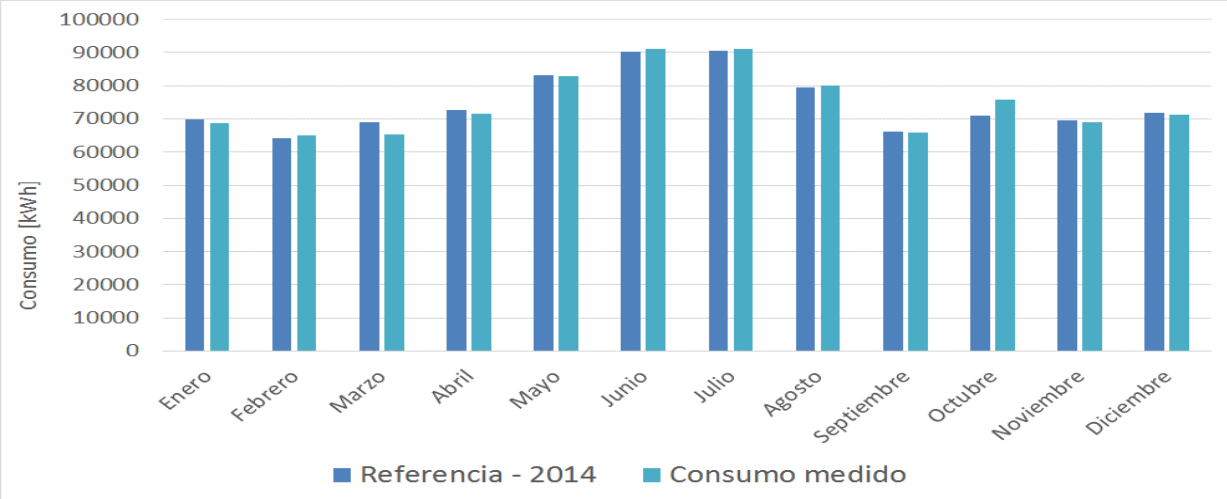

Figura 75. Evaluación consumos modelo BS año referencia 2014 Utrera

Error promedio diario [%]		7.0
Límite de error [%]	**Nº Casos**	**% de casos**
5	205	56.2
10	92	25.2
20	14	3.8
30	0	0.0
40	0	0.0
50	0	0.0

Tabla 44. Evaluación errores modelo BS año referencia 2014 Utrera

VALIDACIÓN 2015 (suponiendo que en dicho año el edificio se encuentra en condiciones de referencia).

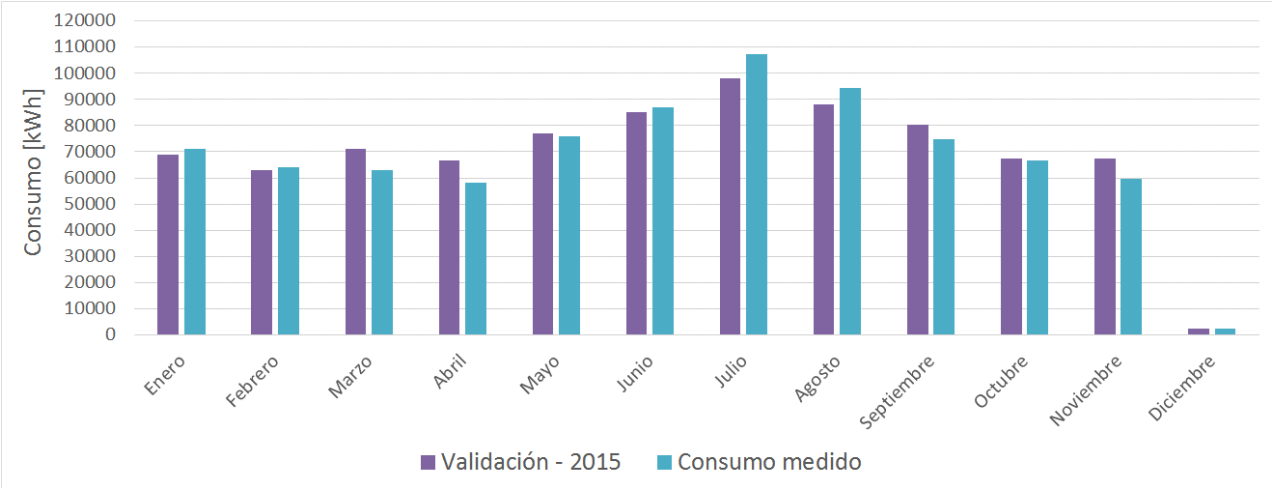

Figura 76. Evaluación consumos modelo BS año validación 2015 Utrera

Error promedio diario [%]		13.2
Límite de error [%]	**Nº Casos**	**% de casos**
5	250	68.5
10	169	46.3
20	73	20.0
30	21	5.8
40	11	3.0
50	4	1.1

Tabla 45. Evaluación errores modelo BS año validación 2015 Utrera

Error [%]	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Anual
Utrera-2014	1.47	1.06	5.38	1.48	0.11	1.00	0.64	0.64	0.48	6.75	0.82	0.63	0.02
Utrera-2015	3.06	1.63	11.32	12.84	1.33	1.91	9.09	7.22	6.60	1.27	11.69	4.49	1.40

Tabla 46. Comparativa de errores modelo BS Utrera

5 CONCLUSIONES

Las principales conclusiones y líneas futuras del trabajo son:

- Se ha desarrollado un algoritmo desde cero, analizando la manera de combinar el clustering de consumos con la caracterización inversa del grupo de investigación. Además este algoritmo tiene presentes varios criterios de decisión que le permiten tener autonomía y adaptarse a la cantidad/calidad de los datos medidos.
- Analizados los métodos de clustering más usados en temas energéticos, aparece una clara separación entre los métodos jerárquicos y no jerárquicos. Se han descartado los métodos basados en redes neuronales por no ser recomendables en la tipología de datos que se tienen.
- Se ha programado el algoritmo en un entorno comercial y matemático para centrar el esfuerzo en dar solución a la metodología. Ahora bien, teniendo está desarrollada se han recopilado las posibles librerías y herramientas para la programación del algoritmo en un entorno explotable por software propios.
- Se ha probado el método y el algoritmo en varios tipos de edificios con diferentes usos, ofreciendo unos resultados de calidad y aceptables.
- Además de su explotación como línea base de consumos en una base de tiempo diaria, se ha trabajado en el proyecto en cómo explotar el procedimiento para predicción de consumos. De esta la estimación diaria se le convierte en la entrada de una estimación en una base de tiempo menor, por ejemplo 15min.
- Por último, este procedimiento se convierte en una herramienta capaz de sustituir a herramientas de simulación energética de edificios como un dato de entrada en condiciones reales de operación y uso. Pudiendo convertirse la entrada para software de optimización de diseño u operación de elementos generadores, consumidores y almacenadores de energía eléctrica y térmica.

REFERENCIAS

[1] Madrid: Centro para el Desarrollo Tecnológico Industrial, "H2020," 2014.

[2] A. L. Comité, E. Y. Social, and E. Y. Al, "Comunicación De La Comisión Al Parlamento Europeo, Concejo, Comité Económico Y Social Europeo Y Al Comité De Las Regiones," p. 15, 2016.

[3] ISO, "Energy management systems. Requirements with guidance for use (ISO 50001:2011)," 2011.

[4] EVO, "International Performance Measurement and Verification Protocol," vol. 1, no. September, 1999.

[5] EVO, "Home - Efficiency Valuation Organization." p. Evo-world.org [online], 2016.

[6] DEXMA, "ISO 50001. ¿Qué es y cómo la implemento en mi empresa?," 2016.

[7] J. M. Carretero Peña, Antonio and García Sánchez, "Gestión de la eficiencia energética [Madrid] AENOR," 2012.

[8] D. C. Washington, "M&V Guidelines," 2015.

[9] ISO, "Energy audits, Draft BS ISO 50002," 2013.

[10] ISO, "ISO50003 - Energy management systems : Requirements for bodies providing audit and certification of energy management systems," *Iso*, no. 0, 2014.

[11] ISO, "ISO 50004:2014 - Energy management systems - Guidance for the implementation. maintenance an d improvement of an energy management system," pp. 1–45, 2014.

[12] ISO 50006, "Energy management systems — Measuring energy performance using energy baselines (EnB) and energy performance indicators (EnPI) — General principles and guidance," vol. 44, no. 0, 2014.

[13] I. 50015, "Energy management systems — Measurement and verification of energy performance of organizations — General principles and guidance," vol. 2014, 2014.

[14] A. Morán, "Análisis y predicción de perfiles de consumo energético en edificios públicos mediante técnicas de minería de datos," 2012.

[15] A. Folch-Fortuny, F. Arteaga, and A. Ferrer, "PCA model building with missing data: New proposals and a comparative study," *Chemom. Intell. Lab. Syst.*, vol. 146, pp. 77–88, 2015.

[16] H. Esmalifalak, A. I. Ajirlou, S. P. Behrouz, and M. Esmalifalak, "(Dis)integration levels across global stock markets: A multidimensional scaling and cluster analysis," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8393–8402, 2015.

[17] I. Cil, "Consumption universes based supermarket layout through association rule mining and multidimensional scaling," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8611–8625, 2012.

[18] G. F. a Zhong Zhao a Jiehua Zhu b, Qi Shen, "Manifold learning: Dimensionality reduction and high dimensional data reconstruction via dictionary learning," *Neurocomputing*, vol. 216, pp. 268–285, 2016.

[19] R. Hettiarachchi and J. F. Peters, "Multi-manifold LLE learning in pattern recognition," *Pattern Recognit.*, vol. 48, no. 9, pp. 2947–2960, 2015.

[20] J. A. Lee, D. H. Peluffo-Ordóñez, and M. Verleysen, "Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure," *Neurocomputing*, vol. 169, pp. 246–261, 2015.

[21] R. Mena, F. Rodríguez, M. Castilla, and M. R. Arahal, "A prediction model based on neural networks for the energy consumption of a bioclimatic building," *Energy Build.*, vol. 82, pp. 142–155, 2014.

[22] F. Khayatian, L. Sarto, and G. Dall'O', "Application of neural networks for evaluating energy performance certificates of residential buildings," *Energy Build.*, vol. 125, pp. 45–54, 2016.

[23] L. Mba, P. Meukam, and A. Kemajou, "Application of artificial neural network for predicting hourly indoor air temperature and relative humidity in modern building in humid region," *Energy Build.*, vol. 121, pp. 32–42, 2016.

[24] E. M. y J. González, “Técnicas de clustering.” 2010.

[25] M.Guerrero, “Modelización y Caracterización de Consumos Energéticos en Edificios,” 2015.

[26] D. Hsu, “Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data,” *Appl. Energy*, vol. 160, pp. 153–163, 2015.

[27] J. Sánchez Ramos, “Metodología Aplicada de Caracterización Térmica Inversa para Edificios,” 2015.

[28] Matlab R2014a Matworks, “Distance,” 2016.

[29] M. Cárdenas-montes, “Índice Davies-Bouldin,” p. 16615, 2011.

[30] V. R. Criterion, “Application of Variance Ratio Criterion (VRC) by Calinski,” no. 1974, pp. 1–6, 1985.

[31] J. Á. Gallardo San Salvador, “Métodos Jerárquicos de Análisis Multivariante,” 1994.

[32] J. Á. Gallardo San Salvador, “Métodos no Jerárquicos de Análisis,” 1971.

[33] “Eportal.magrama.gob.es [online],” 2016.