Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Estadística

Descubrimiento de conocimiento (clustering) en el consumo energético en el edificio inteligente LUCIA

Autora:

Cristina García Sánchez

Tutor:

Dr. José Belarmino Pulido Junquera

Enero 2023

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El significativo aumento de la demanda de energía unido a un incremento de las emisiones de gases de efecto invernadero (GEI) como el CO_2, ha provocado la concienciación sobre el impacto medio ambiental no sólo de la población, sino también de la UE que ha promulgado leyes sobre edificación sostenible para reducir lo citado anteriormente, ya que se trata de uno de los principales sectores que más energía consume y más residuos y emisiones genera.

En este contexto surgen los llamados nZEB o edificios con consumo de energía próximo a cero y los edificios CERO CO_2 que emplean fuentes renovables y tecnologías innovadoras tales como intercambiadores tierra-aire o EAHX. Estos edificios suelen necesitar de avanzados sistemas de control y múltiples sensores por lo que se les suele llamar “edificios inteligentes”.

Uno de estos llamados edificios inteligentes es el edificio de investigación LUCIA (acrónimo de Lanzadera Universitaria de Centros de Investigación Aplicada) de la UVa que ha sido galardonado con numerosos premios como consecuencia de su impacto mínimo sobre el medio ambiente.

En este trabajo se comprueba si este edificio inteligente tiene consumos de energía realmente independientes de las condiciones del exterior. Para ello, se generan clústeres con los datos climatológicos disponibles y se comparan con patrones de consumo de otras secciones y con los protocolos de gestión del edificio.

Palabras clave: GEI, nZEB, CERO CO_2, EAHX, LUCIA, clúster, edificio inteligente, consumo de energía.

ABSTRACT

The significant increase in energy demand coupled with an increase in greenhouse gas (GHG) emissions such as CO_2, has raised awareness of the environmental impact not only of the population, but also of the EU, which has enacted laws on sustainable building to reduce the increase, as it is one of the main sectors that consumes the most energy and generates the most waste and emissions.

In this context, the so-called nZEB or near-zero energy buildings and ZERO CO_2 buildings, using renewable sources and innovative technologies such as earth-air exchangers or EAHX are emerging. These buildings often require advanced control systems and multiple sensors, which is why they are often referred to as "smart buildings".

One of these so-called "smart buildings" is the LUCIA research building (acronym for “Lanzadera Universitaria de Centros de Investigación Aplicada”) at UVa, which has won numerous awards for its minimal impact on the environment.

This work tests whether the LUCIA intelligent building has really energy consumptions independent of external conditions. To do this, clústeres are generated with available climatological data and compared with consumption patterns in other sections and to the control patterns used in the building.

Keywords: GHGs, nZEB, ZERO CO_2, EAHX, LUCIA, cluster, smart building, energy consumption.

ACRÓNIMOS Y ABREVIATURAS

EAHX: Intercambiadores Tierra-Aire

GEI: Gases de Efecto Invernadero

LUCIA: Lanzadera Universitaria de Centros de Investigación Aplicada

NZEB: Edificio con Consumo de Energía Próximo a Cero

SCADA: Supervisión Control y Adquisición de Datos

UTA: Unidad de Tratamiento de Aire

Capítulo 1. Introducción

1.1 Contexto

En la actualidad se está reflexionando considerablemente acerca del impacto negativo que ha provocado y está provocando la actividad del ser humano sobre el medio ambiente, especialmente a partir del siglo XIX en adelante y debido, en parte, a la quema de combustibles fósiles que produce numerosos gases de efecto invernadero (GEI). Por lo general, al hablar de gases de efecto invernadero lo primero que se suele venir a la mente son las emisiones de los vehículos, pero el sector de la edificación constituye también uno de los principales emisores de este tipo de gases.

Debido a las devastadoras consecuencias que esto puede desencadenar (sequías, temperaturas extremas tanto muy cálidas como muy frías o deshielo de los polos, entre otras muchas) se están llevando a cabo numerosas políticas para frenar estos efectos. Una de estas medidas son los llamados NZEB o edificios próximos a cero-energía [1]. Su objetivo consiste en reducir todo lo posible la energía consumida hasta alcanzar un punto en el que dicho consumo sea prácticamente cero, gracias, en parte, al uso de fuentes de energía renovables. Un ejemplo de NZEB es el edificio LUCIA (Lanzadera Universitaria de Centros de Investigación Aplicada) situado en el campus Miguel Delibes de la Universidad de Valladolid.

Por otra parte, también es necesario contextualizar el proyecto en una época en la que se generan cientos de miles de datos en cada instante. Esta "revolución" de los datos ha dado lugar a lo que se conoce como Big Data y aunque existen numerosas definiciones, casi todas ellas consideran tres atributos como la base para definir qué es el Big Data [2]: volumen, velocidad y variedad.

El **volumen** [3] en el contexto del Big Data hace referencia al espacio de almacenamiento necesario para poder registrar los datos de interés con los que se quiere trabajar, pero también se refiere a la magnitud de estos [4]. Si bien es cierto que en la actualidad se habla de Big Data para referirse a tamaños de datos expresados en terabytes y petabytes, no existe un umbral específico que establezca si se trata o no de Big Data [5]. Además, lo que hoy se considera Big Data, en un futuro cercano podría dejar de serlo debido a que las capacidades de almacenamiento aumentarán, y, en consecuencia, será posible capturar conjuntos de datos aún mayores y con más facilidad. Según E. Dumbill [6] se habla de Big Data cuando "se supera la capacidad de procesamiento de los sistemas de bases de datos convencionales" y, en consecuencia, surge la necesidad de elegir "una forma alternativa de procesarlos".

Con la **velocidad** se hace referencia a la rapidez con la que se generan, procesan y analizan los datos [7]. Un claro ejemplo son las redes sociales ya que son uno de los principales factores que han contribuido de manera exacerbada al incremento en la rapidez con la que se generan los datos. Sin embargo, la velocidad también abarca otros factores como por ejemplo la respuesta de una página web o de una aplicación.

Por último, la **variedad** [7] hace referencia a la creciente pluralidad en los tipos de datos que se generan en la actualidad.

Por tanto y, en resumidas cuentas, se puede definir el Big Data como un conjunto de datos de gran volumen caracterizado por la elevada velocidad a la que se generan y por la variedad de los mismos y que se exigen que las necesidades de almacenamiento y cómputo sean muy superiores a las de un sistema convencional.

1.2 Motivación

Existen varios motivos que me han llevado a querer elegir y realizar este trabajo. El primero de ellos es mi interés por profundizar en el aprendizaje no supervisado, ya que, aunque antes de hacer este proyecto tenía conocimientos previos sobre este tema, había numerosos aspectos que desconocía. Además, nunca había llegado a aplicar este tipo de aprendizaje de manera tan directa al mundo real.

Por otra parte, me resulta de gran interés el hecho de que la temática tenga relación con un edificio de la propia Universidad que pretende reducir al mínimo tanto el consumo energético como sus emisiones. De hecho, este tipo de iniciativas son cada vez más necesarias para tratar de disminuir los efectos del cambio climático, especialmente en el sector de la edificación ya que es considerado uno de los principales emisores de gases de efecto invernadero (GEI).

1.3 Metodología

El proyecto se ha desarrollado siguiendo la metodología CRISP-DM o Cross Industry Standard Process for Data Mining [8] que es una de las más utilizadas para planificar y llevar a cabo proyectos relacionados con la minería de datos. Esta metodología se desarrolla en seis fases y aunque en algunos casos existe la posibilidad de no seguir un orden lineal, en este trabajo se siguen las fases de manera lineal.

Primera fase: comprensión de los objetivos del proyecto

La primera fase sirve para conocer en qué consiste el trabajo, para lo cual es fundamental especificar claramente cuál es el objetivo que se persigue en el proyecto. Una vez que dicho objetivo se ha fijado y está claro, el siguiente paso es entender el fundamento teórico y las posibles técnicas que se pueden aplicar para desarrollar dicho proyecto y alcanzar el objetivo propuesto. En este paso se comienza explicando los dos tipos de datos que nos podemos encontrar a la hora de hacer un análisis (etiquetados y no etiquetados) y se continúa explicando las técnicas de aprendizaje automático existentes (aprendizaje supervisado, no supervisado, semisupervisado y por refuerzo) haciendo hincapié en el no supervisado y, dentro de éste, en el análisis de clúster ya que se aplica directamente en el proyecto.

Por otra parte, es fundamental comprender el caso de estudio y es en este punto en el que se analiza en profundidad el edificio LUCIA, especificando qué es, de qué partes consta y a qué se debe su eficiencia.

Segunda fase: análisis de los datos y selección de las características de interés

La segunda fase está dedicada a realizar un primer análisis inicial de los datos sin procesar o *raw data* para saber sus estadísticos básicos y posibles valores atípicos o *outliers*.

Después se procede a filtrarlos, depurarlos y transformarlos a los formatos adecuados para poder extraer la mayor cantidad de información posible de ellos.

A continuación, y ya centrados en el proceso de estudio, se procede a seleccionar las características de interés que se pueden dividir en dos grupos:

1) El primero de ellos hace referencia a datos externos al proceso, concretamente a la humedad y temperatura relativa.

2) El segundo grupo se refiere a los datos del proceso y aunque son muchos y muy variados, los que interesan en este proyecto son las variables que intervienen en el proceso de refrigeración del edificio.

Tercera fase: modelado

Una vez que se han seleccionado las características sobre las cuáles se va a hacer el estudio, se procede a aplicar las técnicas del análisis de clúster.

Cuarta fase: discusión de los resultados

Como consecuencia del modelado, se obtienen los resultados. Es en esta fase donde se analizan y seleccionan los mejores. Además, se razona sobre estos resultados.

Quinta fase: discusión y conclusiones

Finalmente, se discuten los resultados obtenidos y se extraen conclusiones.

Por último, se plantean posibles tareas a realizar como trabajo futuro.

1.4 Objetivos

Con la realización de este proyecto se pretende conseguir un único objetivo: comprobar si es posible diferenciar distintos modos de funcionamiento en el edificio inteligente LUCIA de la Universidad de Valladolid mediante la aplicación de técnicas de aprendizaje no supervisado (análisis de clúster) utilizando datos ambientales de temperatura y humedad, externos al edificio, y datos de funcionamiento de los sistemas de refrigeración.

Para conseguir dicho objetivo se desarrollan varias tareas:

- Comprender los datos disponibles del edificio LUCIA de la UVa, estudiando cómo funciona, de qué partes consta y qué datos del sistema se monitorizan: con qué frecuencia se recogen y cómo se almacenan.
- Estudiar los datos externos del proceso, concretamente la humedad y la temperatura relativa.
- Seleccionar las variables más representativas que intervienen en el proceso de refrigeración del edificio.
- Aplicación de técnicas de clustering sobre los datos externos y sobre los datos de consumo para descubrir posibles patrones de comportamiento.
- Analizar y comparar los resultados obtenidos para tratar de descubrir una posible relación entre los datos externos de temperatura y humedad y los datos de consumo en la refrigeración del edificio.

1.5 Organización de la memoria

Capítulo 1: es el apartado actual. Se contextualiza el proyecto, se explica cuáles son las motivaciones de este, la metodología que se va a aplicar, cuál es el objetivo que se persigue con la realización del trabajo y cómo se estructura la memoria.

Capítulo 2: en primer lugar, se expone el fundamento teórico y las técnicas y métricas para clustering que se van a aplicar y, por último, se detallan las herramientas tecnológicas necesarias para la realización del proyecto.

Capítulo 3: en este apartado se profundiza en el caso de estudio de LUCIA y se procede a comprender cómo funciona el sistema de adquisición de datos del edificio. Además, se hace un primer análisis y exploración de los datos sin procesar y se explican los correspondientes a los sistemas de refrigeración del edificio.

Capítulo 4: se aplican las técnicas necesarias para seleccionar las características de interés.

Capítulo 5: este punto está dedicado a la creación de los clústeres, incluyendo la creación de los modelos y la discusión de los resultados.

Capítulo 6: para concluir, en el sexto y último capítulo se discuten los resultados obtenidos y se presentan las conclusiones que se derivan de la realización del proyecto. Finalmente, se plantean posibles tareas a realizar como trabajo futuro.

Capítulo 7: Bibliografía

ANEXOS

Capítulo 2. Fundamentos y Tecnologías necesarias

2.1 Técnicas de aprendizaje automático

Cada día se producen miles de millones de datos en el mundo, pero por sí solos carecerían de interés si no fueran tratados y analizados correctamente. Aquí es donde cobra real importancia el aprendizaje automático o "machine learning" (ML) que es una rama de la inteligencia artificial (IA) que utiliza algoritmos y métodos estadísticos para predecir un valor numérico (regresión), la clase a la que pertenece una observación (clasificación) o para descubrir patrones de comportamiento en los datos (clustering) y crear modelos que se utilizan para hacer predicciones. La capacidad de predicción será mejor cuantos más y más variados sean los datos disponibles y más entrenados estén los algoritmos.

Es necesario definir qué es una muestra o "instance" que puede describirse como la representación de un determinado objeto. Para ello, a menudo se emplea para representarla un vector d-dimensional $a = \{a_1, a_2, \ldots a_d\} \in \mathbb{R}^d$ donde cada dimensión se identifica como una característica o *"feature"*.

Otro término de interés es etiqueta o *"label",* suele representarse como y y es la predicción para una instancia x . Una etiqueta está formada por un número finito de valores, cada uno de los cuales se conoce con el nombre de clase o *"class".* Si una etiqueta solo cuenta con dos clases, por lo general se suele codificar como $y \in \{-1,1\}$ o $y \in \{0,1\}$. En relación con las etiquetas, los datos se pueden dividir en dos grandes grupos:

- Datos etiquetados o *"labeled data"* [9] son aquellos que están sujetos a un conocimiento previo. La idea fundamental es que las observaciones con etiquetas se utilizan como ejemplos previos para hacer que los algoritmos aprendan. Este tipo de datos se usa en técnicas de aprendizaje supervisado que se explicarán a continuación.
- Datos no etiquetados o *"unlabeled data"* [9] son aquellos que se obtienen mediante observación y para los que se desconoce su clase. Por ejemplo, fotos, artículos de noticias o tweets. Este tipo de datos se usa en técnicas de aprendizaje no supervisado que se explicarán a continuación.

En numerosas ocasiones se integran conjuntamente tanto datos etiquetados como datos de tipo no etiquetados con el fin de crear modelos más precisos. Esto es propio del aprendizaje semi-supervisado que se verá más adelante.

En el caso de este proyecto, los datos de los que se dispone son de la segunda clase, ya que de momento no existen ningún tipo de conocimiento previo, solo se ha recogido información acerca del consumo energético de LUCIA y de las condiciones meteorológicas externas.

Como se ha avanzado anteriormente, existen varios tipos de técnicas de aprendizaje automático:

2.1.1 Aprendizaje supervisado o "supervised learning"

Los datos utilizados son de tipo etiquetados y la idea es la siguiente [10]: dada una serie de datos de entrada o instancias que se pueden denotar como $x = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^n$ y dados una serie de resultados deseados que se pueden denotar como $y_1, y_2, \ldots$, se persigue que el algoritmo aprenda a partir de ellos para que, cuando se proporcione un nuevo dato cuya salida y es desconocida, sea capaz de dar la respuesta y correcta.

Los objetivos habituales de estas técnicas de aprendizaje son hacer previsiones (salida de tipo numérico) o clasificar dichos datos en grupos (salida de tipo etiqueta). Por lo tanto, se pueden distinguir dos grandes subconjuntos:

a. Regresión: el algoritmo se utiliza para relacionar un cierto número de variables y una variable objetivo o *"target"*. El tipo de regresión más utilizado es la regresión lineal [11], que es una modelización entre una variable dependiente continua a la que se le denomina respuesta o *"target"* y una o más variables predictoras independientes. La finalidad es, por lo tanto, entrenar el modelo a partir del conjunto de datos etiquetados para posteriormente poder predecir el target en función de las n-variables predictoras. El modelo para una regresión lineal múltiple que implica la combinación lineal de las variables independientes de entrada se puede formalizar como sigue:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

donde:

- y es la variable respuesta o "target" que se quiere predecir.
- β_0 es el *"intercept"* que representa el valor de la variable respuesta cuando todos los predictores son 0.
- β_i es la pendiente para la variable i. Representa el cambio promedio en la variable respuesta y cuando se aumenta una unidad el valor de x_i y el resto de las predictoras no varían.
- x_i valores de las variables predictoras.
- ε es la variable aleatoria independiente con distribución $N(0, \sigma)$que es la componente aleatoria (error experimental).

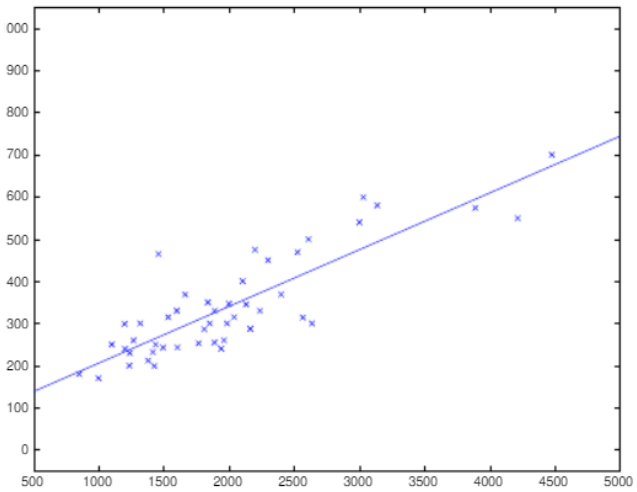


Figura 1: Regresión lineal [12]

Dentro de la regresión también hay que destacar la regresión logística [12] que predice la probabilidad ajustando los datos a una función logística que se define como sigue:

$$h_\theta(x) = g(\theta^T x)$$

Donde la función sigmoide $g(x)$ se define como $g(x) = \frac{1}{1+e^{-x}}$

Gráficamente tiene la forma que se muestra a continuación (véase Figura 2).

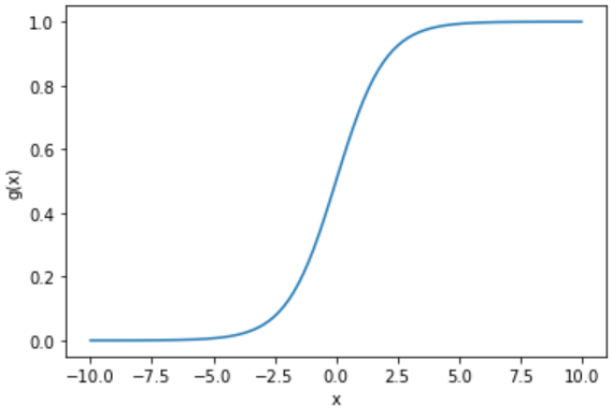


Figura 2: Función logística o función sigmoide

b. Clasificación [13]: el objetivo es obtener un modelo de manera que cada par de entrada-salida se clasifique correctamente dentro de la clase a la realmente pertenece. Puede ocurrir que el clasificador esté sobre ajustado, lo cual supone un problema ya que, en este caso, el clasificador solo está memorizando los datos y carece de capacidad para generalizar.

   Los árboles de clasificación [13] son los más utilizados dentro de este grupo. Se trata de un nodo raíz que carece de arista entrante del cuál parten todos los demás nodos que tienen una única arista de entrada y que pueden o no, tener aristas de salida. En el primer caso se habla de nodos "internos" y en el segundo, de nodos "hojas". Puede constar tanto de atributos categóricos como numéricos (rango de valores). Cada nodo interno divide el espacio de instancia en dos o más subespacios según una determinada función discreta de los valores de los atributos de entrada y cada hoja se asigna a una clase que representa el valor objetivo más apropiado.

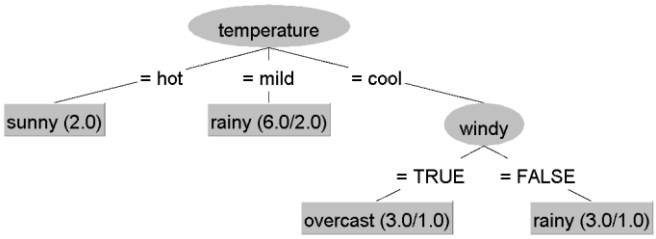


Figura 3: Árbol de clasificación

Otro ejemplo típico de clasificación son las máquinas de vectores soporte o *"support vector machines"* (SVM) [14] aunque este algoritmo también se utiliza en regresión. El objetivo es construir el hiperplano o los hiperplanos que permitan clasificar las entradas en espacio de alta dimensión maximizando el margen entre el hiperplano y dichos vectores soporte.

El SVM más conocido es el lineal, es decir, si los datos son linealmente separables (véanse Figura 4 y Figura 5), existen infinitos hiperplanos que pueden separarlos, pero se trata de encontrar el hiperplano que mejor los separa, o lo que es lo mismo, el hiperplano en el cual el margen entre estas clases es máximo (tal y como se puede ver en la Figura 6). Dicho hiperplano lineal tiene la siguiente expresión: $\{x : \beta_0 + \beta x_i' = 0\}$ donde $x_i \in \mathbb{R}^2$ y las dos clases posibles de la clasificación. La clase y va a depender de las siguientes desigualdades:

$$\beta_0 + \beta_x' > 1 \Rightarrow y = 1$$

$$\beta_0 + \beta_x' < 1 \Rightarrow y = -1$$

En la Figura 4 se observa que no existe un hiperplano lineal que pueda separar a los individuos de la clase 1 (cuadrados) en un grupo y los individuos de la segunda clase (círculos) en otro. Sin embargo, esto no significa que no sean separables, ya que los datos se pueden llevar a dimensiones mayores donde sí sean separables.

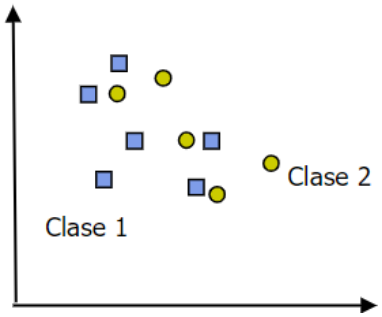


Figura 4: Ejemplo de datos linealmente no separables [14]

En la Figura 5 se aprecia que los datos se pueden separar linealmente en dos clases diferentes y hay infinitos hiperplanos que pueden hacerlo. Además, no habría errores de clasificación.

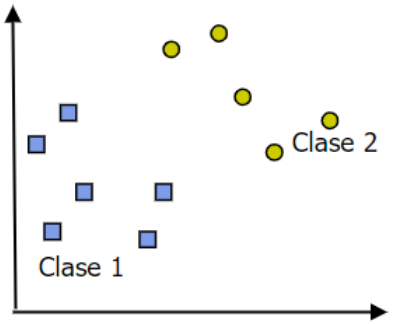


Figura 5: Ejemplo de datos linealmente separables [14]

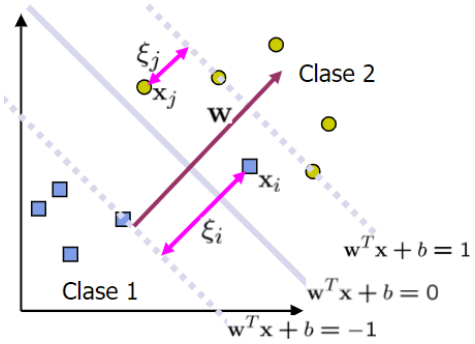


Figura 6: SVM lineal. Clasificador óptimo [14]

2.1.2 Aprendizaje no supervisado o "unsupervised learning"

Este tipo de aprendizaje puede entenderse como una forma de encontrar patrones en los datos, los cuales son de tipo no etiquetado. Un ejemplo clásico es el análisis de clúster.

En el caso de este proyecto, los datos de los que se dispone son de tipo no etiquetado. Además, no existe ningún conocimiento previo y el objetivo es encontrar patrones o modos de funcionamiento ocultos que no se hayan considerado previamente. Todo esto justifica que se vaya a aplicar análisis no supervisado y, concretamente, análisis de clúster. Debido a la importancia de este apartado, se

desarrollará en profundidad en la sección *2.2 Técnicas y métricas para el clustering*, tras explicar las distintas técnicas de aprendizaje automático existentes.

2.1.3 Aprendizaje semi-supervisado o "semi-supervised learning"

Este tercer tipo de aprendizaje [15] realmente es una mezcla de los dos anteriores ya que utiliza tanto datos etiquetados como no etiquetados, es decir, es una mezcla entre el aprendizaje supervisado y el no supervisado, y, en consecuencia, pretende evitar las desventajas de cada uno de estos tipos de aprendizaje. El principal inconveniente del primero es que necesita una gran cantidad de datos de entrenamiento, lo cual se traduce en elevados costos tanto de almacenamiento como de tiempo. En cuanto al segundo, al basarse en la similitud de los datos, no es capaz de agrupar datos que sean totalmente desconocidos o diferentes.

Dentro del aprendizaje semi-supervisado se distinguen dos tipos diferentes: por una parte, se encuentra la clasificación semi-supervisada y por otra, el clustering restringido.

a) Clasificación semi-supervisada [16]
   Al combinar los dos tipos de datos, se reduce el número de datos etiquetados, reduciendo a su vez los costos en almacenamiento y tiempo. A menudo se emplean más datos no etiquetados que etiquetados, pero esto puede variar en función del tipo de estudio que se esté llevando a cabo.
   El objetivo, al igual que en la clasificación supervisada, es crear un clasificador a partir de un conjunto de datos de entrenamiento, pero en este caso con datos etiquetados y no etiquetados, de manera que se pretenden obtener mejores resultados que empleando solo datos del primer tipo.
b) Clustering restringido o semi-supervisado [16]
   El clustering semi-supervisado es una extensión del clustering no supervisado, por lo que el objetivo sigue siendo particionar los objetos en k clústeres atendiendo a dos criterios: debe haber la mayor homogeneidad posible dentro de los clústeres y la mayor heterogeneidad entre ellos. Además, hay que tener en cuenta que puede haber ciertos valores atípicos o outliers. La diferencia fundamental entre el clustering semi-supervisado y el no supervisado radica en que en el primero casi todos los datos son de tipo no etiquetados, pero hay un conjunto de muestras ya etiquetadas asociadas a cierto conocimiento del dominio. Sin embargo, en el no supervisado toda la información de la que se dispone es no etiquetada.

Un ejemplo de técnica semi-supervisada es el "pseudo-labelling" [17] que consiste en utilizar los datos etiquetados para entrenar al modelo como si se tratara de aprendizaje supervisado. Posteriormente, se utiliza el modelo ya entrenado para predecir las etiquetas del resto de datos. Por último, se utiliza todo el conjunto, tantos los etiquetados como los pseudo-etiquetados para entrenar el modelo.

2.1.4 Aprendizaje por refuerzo o "reinforcement learning"

En el aprendizaje por refuerzo [18] se proporciona una serie de datos de entrada, pero no las salidas correctas.

Para entender el funcionamiento del aprendizaje por refuerzo hay que considerar dos componentes:

a. El agente que es el modelo que se quiere entrenar.

b. El ambiente o entorno que contiene una serie de estados discretos y que se espera que no sea determinista [1].

Ambos componentes están conectados y entre ellos se produce la siguiente relación: el agente lleva a cabo una acción que hace que el entorno cambie su estado y emita una recompensa o una penalización. El objetivo es crear una regla que maximice el beneficio o minimice las penalizaciones.

La principal diferencia entre el aprendizaje por refuerzo y el supervisado radica en que en este último se forman pares entrada-salida, es decir, dada una determinada entrada, se espera que se produzca la salida. Sin embargo, en el aprendizaje por refuerzo, el agente se basa en su experiencia previa para tomar decisiones y recibe una recompensa o castigo de inmediato, lo que le permite seguir aprendiendo.

Un ejemplo de este tipo de aprendizaje es el Q-Learning [19]. La explicación de este algoritmo es la siguiente:

1. Se inicializa a cero una tabla llamada *Q-tabla* de dimensión *n x m,* donde *n* hace referencia al número de estados posibles y *m* al número de acciones.
2. Se elige y se lleva a cabo una acción. Como inicialmente la Q-tabla tiene valores 0, la primera decisión se toma de manera aleatoria ya que no se dispone de ningún tipo de información.
3. El agente explora con el fin de obtener estimaciones más precisas de los Q-valores y va actualizando la Q-tabla, es decir, va analizando qué estados se obtienen al realizar diferentes acciones.
4. El agente explora la Q-tabla para conseguir las maximizar las recompensas.

2.2 Técnicas y métricas para el clustering

El análisis de clúster [20] es una técnica de aprendizaje no supervisado que consiste en dividir el conjunto de datos en grupos o clústeres en función de su grado de similitud, de manera que los individuos que pertenecen al mismo subconjunto son más similares entre sí y presentan más disimilitudes con el resto. La importancia del clustering radica en que encuentra agrupaciones entre los datos de tipo no etiquetado.

Existen diferentes tipos de algoritmos en función de cómo se construyen: clustering no jerárquico y jerárquico.

2.2.1 Clustering no jerárquico

Este tipo de métodos presenta cierta complejidad debido a que hay que fijar el número de clústeres antes de realizar el análisis. Una de las principales ventajas es que no necesita calcular una matriz de distancias, lo cual permite ahorrar tiempo y memoria, especialmente cuando se tienen grandes volúmenes de datos. Dentro de este apartado destacan:

[1] En el aprendizaje por refuerzo el entorno juega un papel fundamental y generalmente, se espera que, aunque se realice la misma acción en el mismo estado, pero en momentos de tiempo distintos, los resultados de los estados posteriores que se obtienen podrían ser diferentes.

-Kmedias [21]

Es el método no jerárquico más famoso y se utiliza cuando las variables de los datos son de tipo cuantitativo. El objetivo es encontrar k centros óptimos que se van a denotar como $m_1^*, m_2^*, \dots, m_k^*$, de manera que la suma de cuadrados de las distancias entre los individuos de cada grupo sea mínima.

El procedimiento se realiza en dos pasos:

1. Paso de asignación donde se particionan las observaciones y se asignan a los centroides más cercanos, es decir

$$\text{Cluster } J = \{x_i : \|x_i - m_J^*\|^2 \leq \|x_i - m_j^*\|^2 \text{ para } j \neq J\}$$

2. Paso de actualización en el que se calculan los nuevos centroides mediante el cálculo de las medias de las observaciones que forman cada partición.

$$m_{i=}\frac{1}{|J|}\sum_{x_j \in J} x_j$$

Estos dos pasos se repiten iterativamente hasta que el algoritmo converge, es decir, hasta que las asignaciones a las particiones son iguales o prácticamente no varían.

-K-medioides [21]

El procedimiento es prácticamente igual que en las k-medias, exceptuando que, en lugar de calcular las medias de cada partición, se calculan las medianas, por lo que tiende a ser más robusto frente a la presencia de valores atípicos. Sin embargo, si no hay outliers ambos métodos darán lugar a resultados similares.

En cualquiera de los dos métodos anteriores surge la necesidad de definir la disimilitud o, lo que es lo mismo, las diferencias entre dos individuos que se explican con las distancias: si tenemos medidas en n variables para los individuos $p = (p_1, p_2, \dots, p_n)$, $q = (q_1, q_2, \dots, q_n)$:

- La **distancia euclídea** [22] es la raíz cuadrada de la suma cuadrada de la distancia (diferencia) entre cada variable, es decir:

$$d_{(p,q)} = \sqrt{(p_1 - q_1)^2 + \cdots + (p_{n-}q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

- La **distancia de Manhattan** [22] es igual a la suma de las distancias absolutas de cada variable, es decir:

$$d(p,q) = |p_1 - q_1| + \cdots + |p_n - q_n| = \sqrt{\sum_{i=1}^{n}|p_i - q_i|}$$

Para identificar el número óptimo de clústeres k, son necesarias otras métricas como el método del codo o Silhouette:

1. El método del codo [23]

La clave de este método está en la inercia, es decir en la suma de cuadrados del error (SCE) que se calcula como sigue:

$$SCE = \sum_{k-1}^{k}\sum_{x_i \epsilon S_k} ||x_i - m_k^*||^2$$

y es la suma de la distancia euclídea de cada punto al centroide más cercano. El método consiste en ir incrementando el valor de k (generalmente hasta llegar a 10) y por cada valor de k se calcula la SCE. Llegará un punto en el que el número de errores no cambia significativamente entre un clúster y otro, por lo que se habrá encontrado el valor óptimo de k. Gráficamente cuando el valor del error disminuye

drásticamente se aprecia que se forma un ángulo menor y se forma el "codo" que permite identificar el número óptimo de clústeres.

2. Método de Silhouette [24]

El método de Silhouette que propone Rousseau [24] calcula el número óptimo de clústeres a partir del coeficiente de Silhouette que se calcula como sigue:

$$s(i) = \frac{b(i)-a(i)}{máx\{a(i),b(i)\}} \; tal\; que \; -1 \leq s(i) \leq 1$$

$donde\; b(i)$ es la distancia media mínima de un punto a un clúster cualquiera (distinto al que pertenece) y $a(i)$ se refiere también a la distancia media, pero en este caso entre un punto y todos los puntos del mismo clúster.

Un valor de este coeficiente cercano a 1 indica que los datos están correctamente agrupados y, por tanto, el número de clústeres es adecuado y viceversa. Por lo que el método consiste en ir aumentando el valor de k (generalmente hasta llegar hasta 10) y encontrar aquel que tenga el valor de s(i) más cercano a 1.

3. Algoritmo DBSCAN [25]

La lógica de este algoritmo es la siguiente [25]: dados dos parámetros de entrada v (distancia máxima de vecindad) w (nº mínimo de puntos para formar un clúster), se selecciona aleatoriamente un punto y, si no ha sido seleccionado anteriormente, se encuentran sus puntos vecinos atendiendo a v. Si dicho número es menor que w, el punto es marcado como atípico u *"outlier".* En caso contrario se marca como "punto denso". El proceso se repite iterativamente hasta que se han visitado todos los puntos y se han asignado a un clúster o bien se han marcado como atípicos.

2.2.2 Clustering jerárquico

El objetivo de este algoritmo [26] es formar grupos de elementos que sean lo más homogéneos posibles entre sí y lo más heterogéneos con respecto a los elementos del resto de grupos en función de algún criterio de disimilitud. La principal característica de este tipo de agrupación es que tiene la capacidad de fijar por sí mismo el número de clústeres. Sin embargo, una gran desventaja es que es muy sensible a valores atípicos o *"outliers".*

Dentro del clúster jerárquico se distinguen dos tipos de estrategias: aglomerativas y divisivas.

a) Aglomerativas o "bottom-up" [27]

Es la estrategia que más se utilizada de las dos. Se parte de elementos individuales y se van agrupando hasta obtener el número de clústeres deseado, es decir, se fusionan subconjuntos $X1,\; X2$ iterativamente hasta que se llega a la raíz, verificándose $X1 \cap X2 = \emptyset, X1 \cup X2 = X$. Los métodos que se emplean en clustering aglomerativo para tomar la decisión de agrupar o no clústeres son fundamentalmente tres:

- **Single Linkage** [28]

  Al simple linkage o vinculación simple también se le conoce con el nombre del método del vecino más cercano. El razonamiento es el siguiente: en cada paso se combinan los dos clústeres que contienen el par de elementos más próximo que aún no pertenece al mismo clúster, es decir:

$$\delta(A,\, B) = \min_{p\in A;\; q\in B} \; d(p,\, q).$$

Con otras palabras, se selecciona la distancia mínima entre un caso del primer clúster y un caso del segundo.

La complejidad computacional en single linkage es de $O(n^2)$.

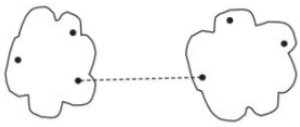


Figura 7: Single Linkage [28]

- **Complete Linkage** [28]

  El razonamiento es muy similar al anterior, pero en este caso la distancia entre los clústeres es la distancia entre los dos pares de elementos (pertenecientes a diferentes clústeres) que están más alejados entre sí. La distancia más corta de todas esas distancias máximas es la que provoca la fusión de los clústeres.

$$\delta(A,\, B) \;= \max_{p \in A;\, q \in B} \; d(p,\, q)$$

  Una de las principales desventajas es su gran sensibilidad a la presencia de datos atípicos. La complejidad computacional en complete linkage es de $O(n^2 \log n)$.

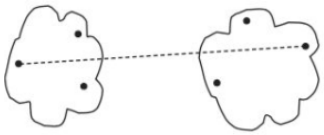


Figura 8: Complete Linkage [28]

- **Average linkage** [28]

  En este caso la distancia entre los clústeres se calcula como la media de las distancias de cada par de elementos. La complejidad computacional de average linkage es de $O(n^2 \log n)$.

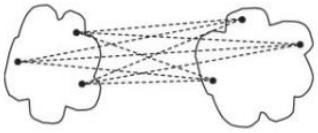


Figura 9: Average Linkage [28]

El clustering jerárquico aglomerativo se suele representar mediante dendrogramas (véase la Figura 10), que se analizan de abajo a arriba. En la Figura 10 se observa que se ha utilizado una partición de dos conglomerados: las cinco observaciones en verde forman el primer conglomerado y las otras cinco que están en naranja, el segundo. Interpretando los dendogramas en función de la altura, también se puede decir que tanto en el single como en el complete linkage, las observaciones más correlacionadas son las que están representadas en verdes y que se encuentran más a la derecha. En el caso del average linkage, las más correlacionadas son las dos naranjas de la derecha.
Este pequeño ejemplo sirve también para mostrar que los resultados pueden variar en función de la distancia y del linkage que se empleen.

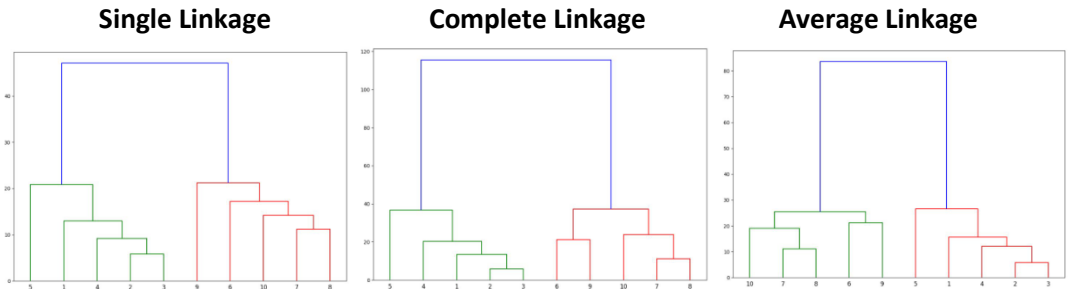


Figura 10: Dendogramas [28]

b) Divisivas o "top-down" [27]

Se parte de un único conjunto formado por todos los elementos y se va subdividiendo en función de un índice de disimilitud, es decir, se parte desde la raíz que contiene todo el conjunto de datos (X) y de manera recursiva, divide dicho conjunto en dos nodos hijos ($X1$, $X2$), de manera que la intersección de elementos de ambos es el conjunto vacío ($X1 \cap X2 = \emptyset$) y su unión es el conjunto del que se ha partido, es decir, $X1 \cup X2 = X$.

2.3 Reducción de la dimensionalidad

Muchos problemas de la vida real necesitan las técnicas de aprendizaje automático que se han citado anteriormente para encontrar una solución. Sin embargo, tanto en estos casos como en otros en los que hay que emplear otros métodos no se pueden aplicar directamente las técnicas necesarias para encontrar una solución o no se debería hacerlo si el número de variables que contienen es muy elevado porque los resultados que se alcanzan pueden ser incorrectos debido a que trabajar con grandes cantidades de variables da lugar a numerosos problemas, entre los que destacan la multicolinealidad y el sobreajuste.

La multicolinealidad se produce cuando las variables explicativas de un modelo de regresión están altamente correlacionadas entre sí y, en realidad, deberían de ser independientes.

Otro de los problemas es que un número elevado de variables implica un número elevado de dimensiones, lo cual también supone un problema en la visualización de los datos ya que no es factible representar más de tres dimensiones ni, por tanto, más de tres variables, y un problema de complejidad (espacio y tiempo) de los algoritmos.

Para evitar los problemas anteriores, cuando se dispone de un elevado número de variables, se emplean técnicas de reducción de la dimensionalidad entre las que destaca fundamentalmente Análisis en Componentes Principales (ACP) y el coeficiente de correlación de Pearson.

2.3.1 Análisis en componentes principales (ACP)

El análisis en componentes principales [30] es una técnica de reducción de la dimensionalidad que consiste en encontrar un subgrupo de variables que son combinaciones lineales (c.l.) de las variables iniciales de manera que sean capaces de recoger la mayor parte de la información disponible. Es decir,

se trata de buscar la c.l. del menor número posible de variables que maximiza la variabilidad. Estas variables tienen que estar incorrelacionadas.

Dadas t variables iniciales, el objetivo es calcular los t autovectores y los t autovalores $\lambda_{1,\dots,}\lambda_t$ correspondientes ya que, cada uno de estos indica la cantidad de información recogida por la proyección sobre cada autovector. Por teoría [31] se sabe que

$$\sum_{i=1}^{t} \lambda_i \; = \; inercia\, total$$

por lo que, si se deciden seleccionar h autovectores, se explicará $e \; = \frac{\sum_{i=1}^{h} \lambda_i}{\sum_{i=1}^{t} \lambda_i}$ * 100 % de la inercia total. Como el objetivo es reducir el número de dimensiones perdiendo la menor cantidad de información posible, se seleccionarán los autovectores que más información recojan que serán los asociados a los autovalores más elevados, pero ¿cuántos se deben seleccionar? Los criterios más habituales son:

- El método del codo (explicado anteriormente).
- Conseguir superar cierto porcentaje de inercia explicada, por ejemplo, 75%.
- Seleccionar las componentes que expliquen una inercia superior al promedio de los autovalores.[2]

2.3.2 Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson [32] también es considerado un método de reducción de la dimensionalidad ya que se basa en la selección de características o "features" en función del nivel de correlación de las mismas, eliminando aquellas que estén altamente correlacionadas y se calcula como el cociente de la covarianza de x , y entre el producto de la desviación típica de x por la desviación típica de y , es decir:

$$\rho(x, y) \; = \; \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \; -1 \leq \rho \leq 1$$

Donde:

- ρ = 0 indica ausencia de correlación entre las variables
- ρ < 0 indica una correlación negativa entre las variables, es decir, si el valor de una aumenta el doble, el triple, ... el valor de la otra disminuye el doble, el triple, … y viceversa
- ρ > 0 indica una correlación positiva entre las variables, es decir, si el valor de una aumenta el doble, el triple, ... el valor de la otra aumenta el doble, el triple, …. De manera análoga si disminuye.
- Cuanto más próximo a 1 sea $|\rho|$, más fuerte será la correlación entre las variables y viceversa.

[2] Si el análisis es normado, es decir, si se verifica que $\sum_{i=1}^{t} \lambda_i \; = t \; = \; inercial\, total$, se seleccionarán aquellos autovalores mayores que 1.

2.4 Normalización vs estandarización

Dado un conjunto de datos, en la mayoría de los casos surge la necesidad de normalizarlos o estandarizarlos para obtener resultados fiables al aplicar algoritmos, métodos o manipularlos en general.

Se debe normalizar [3][33] un conjunto de datos cuando existan diferencias importantes en los rangos de valores de las variables. Por lo general, el rango en el que se encuentran los datos normalizados suele ser [0, 1], pero también es frecuente encontrar el rango [-1, 1]. En ambos casos, la idea es la misma (véase Figura 11 donde se ilustra claramente el concepto) y el cálculo es muy similar:

- $x' = \frac{x - min}{max - min}$ para obtener datos dentro del rango [0, 1]
- $x' = 2 \cdot \left(\frac{x - min}{max - min}\right) - 1$ para obtener datos dentro rango [-1, 1]

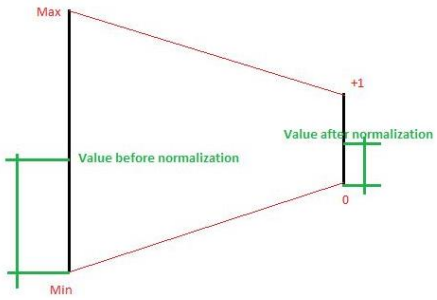


Figura 11: Normalización [33]

En cuanto a la estandarización [33], permite ajustar los datos transformándolos en una $N(0,1)$. Para ello:

- $x' = \frac{x - \overline{x}}{\sigma}$

2.5 Herramientas tecnológicas

Las técnicas de análisis de clúster se han llevado a la práctica utilizando Python [29] como lenguaje de programación que se caracteriza por ser un lenguaje de alto nivel y dinámico que hace hincapié en la legibilidad del código. Es fundamental en machine learning y en ocasiones se puede usar para trabajar de manera más eficiente que con otros lenguajes. Consta de gran variedad de bibliotecas. Las que se han usado en este proyecto son las siguientes:

- **Numpy:** se utiliza fundamentalmente para crear y trabajar con vectores o *"arrays"* y matrices de elevadas dimensiones. Además, incluye numerosas funciones matemáticas, estadísticas e incluso lógicas.
- **Pandas:** define nuevas estructuras de datos como *"Series"* (de una dimensión, similares a los arrays, de tamaño inmutable y cuyos elementos son del mismo tipo) o *"DataFrames"* (de dos dimensiones, forma de tabla y cuyas columnas son de tipo *"Serie"*). Es una de las bibliotecas

[3] Python permite normalizar los datos directamente con la función MinMaxScaler que se encuentra dentro de la librería scikit-learn.

que más se ha utilizado en este caso debido a que permite seleccionar filas y columnas tanto por nombres como por índices y permite reordenarlas, combinarlas o eliminarlas, lo cual facilita considerablemente la manipulación de los datos.

Por último, hay que destacar *"groupby"* y las funciones de agregación. Con *"groupby"* se pueden agrupar columnas en función de una característica, lo cual ha resultado muy útil ya que nos ha permitido agrupar distintos valores por semanas. En cuanto a las funciones de agregación ("agg"), se han aplicado sobre los grupos que se han mencionado anteriormente y se ha podido calcular en cada uno de dichos grupos la media, el mínimo y el máximo.

- **Scikit-learn:** muy útil en el ámbito de aprendizaje automático. En este proyecto se ha utilizado para aplicar los algoritmos de clustering, reducción de la dimensionalidad y normalización de los datos.
- **Matplotlib:** biblioteca de Python enfocada a las representaciones gráficas con diversas funcionalidades como fijar títulos, colores e incluso número de píxeles. Concretamente se ha utilizado para representar las distintas variables objeto de estudio.
- **Plotly:** biblioteca de Python que se emplea para la visualización interactiva de datos.

El entorno en el que se ha desarrollado la parte de programación es Jupyter notebook que me ha permitido tener gran orden en el desarrollo de la lógica ya que el código se estructura en celdas independientes y se pueden obtener los resultados de la ejecución de cada una de ellas. Además, Jupyter Notebook permite exportar a formatos como Python, HTML, PDF o Markdown, entre otros.

Capítulo 3. Descripción del edificio LUCIA y de los datos disponibles

3.1 Edificio LUCIA

El edificio LUCIA de la Universidad de Valladolid [34] es un edificio de consumo energético próximo a cero-energía, NZEB, y nace como consecuencia de la necesidad de reducir al mínimo el impacto medio ambiental y la demanda de energía del sector de la edificación. Se caracteriza por ser un laboratorio donde se aplican técnicas de eficiencia energética renovables teniendo en cuenta el clima del lugar en el que se encuentra: veranos secos y cálidos e inviernos con gélidas temperaturas.

Partes de LUCIA

Para cumplir los objetivos que caracterizan a un NZEB se aplican estrategias tanto pasivas como activas [35].

Dentro del primer grupo y, concretamente, en el caso de LUCIA cabe destacar:

- Fachada en forma de zigzag. Su objetivo no es estético, sino que con esto se consigue que todas las ventanas se reorienten al sur y, en consecuencia, se maximizan las ganancias solares durante el invierno y se produce sombra durante el verano.
- Diseño bioclimático con gran capacidad de aislamiento térmico.
- Grandes módulos acristalados o lucernarios que permiten la entrada de luz natural filtrando las radiaciones solares dañinas.
- Aparcamiento al aire libre rodeado de vegetación autóctona que apenas necesita cuidados para su mantenimiento. Al no estar completamente cubierto tiene la ventaja de disponer de luz natural, así como de ventilación constante.
- El EAHX o intercambiador de calor tierra-aire también se encuentra integrado en el nZEB y consiste en varios tubos enterrados a varios metros que hacen posible el intercambio de energía entre el terreno arcilloso y el edificio. Su funcionamiento es posible gracias a la propiedad de inercia térmica [36] del suelo que puede definirse como la capacidad que tiene la masa (en este caso el suelo) de conservar la energía térmica recibida (el calor) y liberarla lentamente. Como consecuencia de esta propiedad y debido a que en los meses de invierno el aire exterior es más frío que la temperatura que hay en los tubos, el aire del exterior se calienta circulando por dichos tubos subterráneos y llega al edificio calentando el mismo. En verano ocurre todo lo contrario, por lo que sirve para enfriarlo.

  Si bien es cierto que el EAHX es muy eficiente, también hay que destacar la técnica del free-cooling [37] que es capaz de refrigerar el edificio utilizando el aire del exterior cuando la temperatura de éste es inferior a la del interior del edificio. Esto se consigue gracias a una serie de compuertas y ventiladores que hacen posible dicho intercambio.

  Además, el free-cooling tiene la ventaja de que puede mejorar considerablemente la calidad del aire del interior del edificio.

  El funcionamiento del free-cooling es posible gracias a la Unidad de Tratamiento de Aire o UTA que es capaz de proporcionar una ventilación total de aire exterior.

  El free-cooling se suele utilizar cuando el EAHX no es suficiente para alcanzar el confort de temperatura.

Dentro de las estrategias activas cabe resaltar:

- Energía solar fotovoltaica con la que se cubre gran parte de la demanda eléctrica.
- Caldera de biomasa que permite satisfacer parcialmente la demanda de calefacción, la cual necesita una cantidad superior de combustible en los meses más fríos.

Eficiencia

Los sistemas renovables instalados en LUCIA consiguen disminuir considerablemente las emisiones de CO_2 en relación con los combustibles convencionales.

Gracias a las instalaciones fotovoltaicas se consigue que el consumo de energía de iluminación del edificio se encuentre considerablemente por debajo del que tienen otros similares.

Por lo que a la caldera de biomasa respecta, apenas produce emisiones de CO_2. Existe una relación directa entre el consumo de este combustible y la demanda de calefacción: el consumo será mayor en los meses más fríos y viceversa.

El periodo del año en el que se consigue un mayor ahorro de energía con el EAHX es en los meses de primavera y disminuye en verano y otoño.

La utilización en conjunto de estas instalaciones no sólo consigue una reducción del impacto medio ambiental, también permite un ahorro considerable de tipo económico.

3.2 El sistema de adquisición de datos

En este tipo de edificios es indispensable disponer de la información relativa a la energía que se consume, los datos climatológicos y gestionar de manera óptima los parámetros energéticos utilizables. Todo esto se consigue gracias a un sistema de control y gestión llamado DESIGO fabricado por SIEMENS® [38] y que se basa en un software de control y adquisición de datos de tipo SCADA de DESIGO. Una de las principales características es que permite obtener datos acerca de las diferentes partes del edificio en tiempo real y se manipulan y analizan para conocer tanto posibles fallos del sistema como las necesidades de éste. Además, gracias a este registro periódico de datos, se pueden hacer previsiones de la futura demanda de energía y, en consecuencia, suministrarla de manera eficiente.

Los datos iniciales de los que se parten en este proyecto son datos sin procesar, es decir, raw data, por lo que inicialmente va a ser necesario tratarlos y depurarlos. Concretamente, el archivo de datos consta de 114582 filas y 624 columnas y, tras importarlos como DataFrame en Jupyter Notebook y hacer una primera visualización de las primeras filas se obtiene que cada columna hace referencia o bien a una medición o bien al momento (fecha y hora exactas) en la que se produce dicha medición (véase Figura 12).

	SeVar`Atic`TRAt	date	SeVar`CaudAgu`AguACS`Contador`Actual	date.1	SeVar`CaudAgu`AguACS`ContTot1	date.2
0	15.0	2019-01-01 00:23:05	3500.0	2019-01-01 00:17:44	273360.0	2019-01-01 00:17:45
1	15.0	2019-01-01 00:38:05	3500.0	2019-01-01 00:32:44	273360.0	2019-01-01 00:32:44
2	14.9	2019-01-01 00:53:06	3500.0	2019-01-01 00:47:45	273360.0	2019-01-01 00:47:45
3	14.9	2019-01-01 01:08:06	3500.0	2019-01-01 01:02:45	273360.0	2019-01-01 01:02:45
4	14.9	2019-01-01 01:23:06	3500.0	2019-01-01 01:17:45	273360.0	2019-01-01 01:17:45

Figura 12: Primeras filas y columnas de los datos iniciales

Sin embargo, en este paso se observa que no todos los datos son de utilidad (véase Figura 13) porque en el nombre de las últimas columnas parece que hay datos y el resto de las filas de dichas columnas son NaN, es decir, no son números. Se procede a detectar a partir de qué columna empieza a suceder esto y se obtiene que es de la columna 307 en adelante. Se sospecha que lo que ha podido ocurrir es que todos esos valores en verdad deberían haberse introducido como una fila, ya que justo son 306 valores que coincide con el número de columnas (306), pero como no hay se eliminan todas ya que no van a aportar información al estudio y a partir de este momento el conjunto de datos estará formado por 306 columnas, de las cuales la mitad hacen referencia a mediciones de distinto tipo y partes del edificio y la otra mitad al momento en el que se ha recogido dicha medición.

...	5.1	2019-01-05 17:28:24	2199.3	2019-01-01 17:29:07.21	23.73	2019-01-01 17:29:07.22	0.10	2019-01-01 17:29:09	1	2019-01-01 17:29:09.1
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figura 13: Últimas filas y columnas de los datos iniciales

Tras terminar de hacer el primer análisis sobre los datos sin procesar se obtienen las siguientes conclusiones:

- Se han monitorizado 153 variables.
- Por cada variable monitorizada hay una columna asociada que recoge la fecha en la que se produce cada registro.
- Concretamente se han registrado 17747328 datos: la mitad hacen referencia a las mediciones de las variables y la otra mitad a la fecha correspondiente.
- El primer registro monitorizado es del 01/01/2019 y el último es del 22/05/2021, por lo que han transcurrido 2 años, 4 meses y 21 días (872 días) entre la primera monitorización y la última.
- La primera monitorización para todas las variables es el 01/01/2019, pero la última varía de unas variables a otras.
- La variable para la cuál hay mas registros es *CondExt`TOa* que hace referencia a la temperatura del aire exterior con un total de 114583 registros y el último de ellos fue tomado el 21/05/2021.
- La variable que cuenta con menos registros es *AInt`P1`AnlzRed12`AI65* que se refiere a la energía activa en el área de integraciones en la Planta Primera`P1B5G`, con un total de 3343 registros y el último de ellos fue tomado el 21/05/2021.
- De los dos puntos anteriores se concluye que hay fluctuaciones en las frecuencias con las que se producen los registros entre las variables.
- Como consecuencia del punto anterior, surge la duda de si también hay fluctuaciones en las frencuencias de los registros dentro de cada variable y finalmente se concluye que ocurre en varias ocasiones, pasando por ejemplo de un intervalo de 15’ a otro de 1’ o incluso de 30’.
- El nombre de las variables sigue cierto patrón en función del sistema al que pertenecen. Este último apartado se amplía a continuación.

Las variables se pueden clasificar en seis grupos en función del patrón del nombre que tienen y, en consecuencia, del sistema al que pertenecen:

- Las variables que hacen referencia al área de producción se identifican con “AProd” y miden tanto la refrigeración como el enfriamiento de circuitos primarios y secundarios del edifico como la calefacción de circuitos secundarios (agua caliente y fancoils fundamentalmente).
- Las variables pertenecientes al área de unidades terminales se identifican con “AUT” y miden las temperaturas registradas en las diferentes partes de LUCIA y las demandas de refrigeración y calefacción las plantas baja, primera y segunda.
- Aquellas que se identifican por “SeVar” aluden a señales varias, esencialmente a las mediciones de los caudales del agua (contadores de agua general, climatización o reciclada entre otras).
- Las variables “ACl`UTA01” hacen alusión al área de climatización (“ACl”). Además, todas ellas forman parte de la Unidad de Tratamiento del Aire (de ahí el “UTA”) y en todas ellas se trata de aire primario (“01”). Los registros varían desde humedades, a temperaturas e incluso presión.
- Las variables que comienzan por “AInt” forman parte de las mediciones del área de integraciones y aluden tanto a la energía (activa o acumulada) como a la potencia de las diferentes estancias de las plantas baja, primera, segunda, sótano y los contadores de absorción de calor o enfriamiento.
- Por último, hay dos variables que se refieren a las condiciones exteriores (“CondExt”) de temperatura y humedad que se analizarán en profundidad en la sección 4.1 *Datos relativos a las condiciones externas de humedad y temperatura*.

Capítulo 4. Análisis de los datos y selección de los mismos

4.1 Datos relativos a las condiciones externas de humedad y temperatura

Después de realizar un primer análisis general sobre los datos sin procesar de todo el *dataset*, pasamos a centrarnos en aquellos que hacen referencia a las condiciones externas. Para ello se seleccionan tanto las columnas con los registros de humedad y temperatura (variables “CondExt`HuOa” y "CondExt`TOa" respectivamente) como la fecha[4] en la que se han recogido dichas mediciones (variables “date.12” y "date.13") y se crean los dataframes de “humedad” y “temperatura” que contienen dicha información con las columnas renombradas para que sus nombres sean más descriptivos como se aprecia en la Tabla 1 y Tabla 2.

	humedad_ext	fecha
0	92.5	2019-01-01 00:23:05
1	90.0	2019-01-01 00:38:05
2	90.3	2019-01-01 00:53:06
3	91.6	2019-01-01 01:08:06
4	91.1	2019-01-01 01:23:06
...	...	...
87257	45.3	2021-05-21 22:46:04
87258	45.8	2021-05-21 23:01:04
87259	46.2	2021-05-21 23:16:04
87260	47.3	2021-05-21 23:31:04
87261	48.6	2021-05-21 23:46:04

Tabla 1: Datos sin procesar de la humedad relativa

	temperatura_ext	fecha
0	1.1	2019-01-01 00:19:14
1	1.5	2019-01-01 00:30:23
2	1.4	2019-01-01 00:41:32
3	1.1	2019-01-01 00:52:42
4	1.1	2019-01-01 01:03:51
...	...	...
114577	16.4	2021-05-21 23:11:04
114578	16.2	2021-05-21 23:21:04
114579	16.0	2021-05-21 23:31:04
114580	15.8	2021-05-21 23:41:04
114581	15.6	2021-05-21 23:51:05

Tabla 2: Datos sin procesar de la temperatura relativa

Una observación es que no existen registros completos para el año 2021 (el último día registrado es el 21 de mayo de dicho año), por lo que se decide hacer el estudio sólo con los datos de 2019 y 2020.

Otra observación es el hecho de que el número de registros que se tienen de humedad (87261) difiere de los que se tienen de temperatura (114581) lo cual se debe a que la frecuencia con la que se realizan las mediciones de esta última variable es menor que la frecuencia con la que se realizan las de la

[4] El formato de la fecha se ha pasado al tipo de dato “datetime” con el formato "%Y-%m-%d %H:%M:%S" que hace referencia a “Año-mes-día hora:minutos:segundos” para poder trabajar correctamente en Python.

humedad. Si se uniera la información de ambas variables en un único dataframe, se perdería la diferencia de datos de la temperatura, por lo que de momento se trabaja con dos dataframes: uno contiene la información relativa a la temperatura y otro a la humedad. Sin embargo, más adelante la diferencia de registros de humedad y temperatura no a va a ser un problema ya que se va a trabajar con la media de los datos agrupados por semanas[5] lo cual se consigue con los comandos "groupby" y "dt.isocalendar().week" y una función de agregación con la que se calcula la media ("agg" y "mean").

Una vez que se tienen sólo los datos externos correspondientes a los años 2019 y 2020, se separan por un lado los de 2019 y, por otro, los de 2020, ya que, el análisis de clúster se va a hacer de cada año.

A continuación, se describen los estadísticos básicos tanto de la humedad como de la temperatura de ambos años respectivamente (véanse Tabla 3 y Tabla 4) y, de nuevo, los registros de temperatura son mayores que los de la humedad tanto en 2019 como en 2020. En cuanto a los valores mínimos y máximos, se aprecia que hay coherencia ya que para el caso de la humedad son 0% y 100% respectivamente (no tienen sentido valores negativos ni por encima de 100) y para la temperatura se tienen –5. 5ºC en 2019, -3.5ºC en 2020 de mínima y 39.3ºC de máxima en ambos casos. Aunque estos valores no son los más usuales en Valladolid, pueden alcanzarse en días muy fríos de invierno o muy calurosos de verano.

Estadísticos	Valor humedad	Valor temperatura
número	36697.00	47771.00
media	59.81	14.56
desv. típica	25.72	8.60
min	4.80	-5.50
25%	38.60	8.30
50%	61.80	13.40
75%	79.30	20.00
max	100.00	39.30

Tabla 3: Estadísticos para humedad y tº año 2019

Estadísticos	Valor humedad	Valor temperatura
número	36510.00	47708.00
media	67.28	14.73
desv. típica	26.70	8.28
min	0.00	-3.50
25%	46.70	9.00
50%	70.60	13.50
75%	92.20	19.20
max	100.00	39.30

Tabla 4: Estadísticos para humedad y tº año 2020

A continuación, se realiza un estudio de la normalidad de los datos:

- Estudio de la normalidad en los datos de temperatura en ambos años por separado:

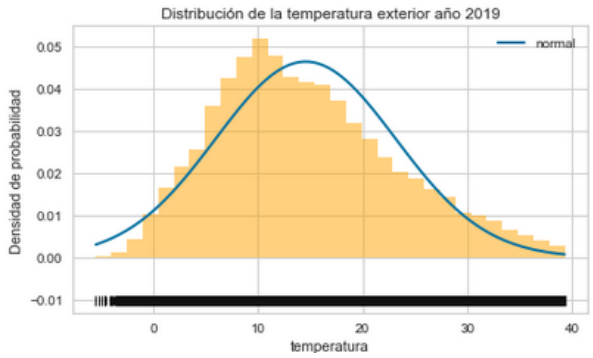


Figura 14: Distribución de la temperatura año 2019

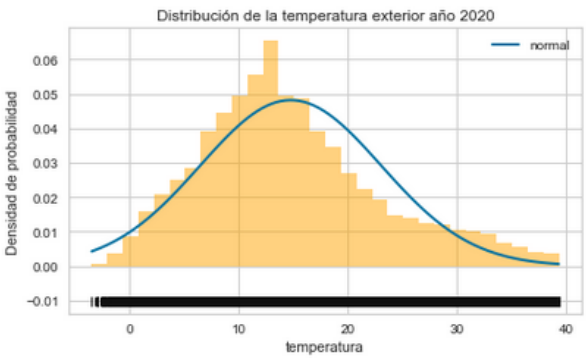


Figura 15: Distribución de la temperatura año 2020

[5] Tras agrupar los datos por semana y calcular la media de cada una de ellas, se han obtenido 53 semanas en lugar de 52, el motivo es que el año 2020 es bisiesto

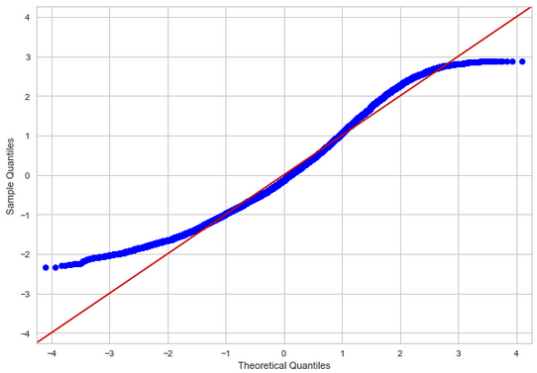


Figura 16:QQ-Plot temperatura año 2019

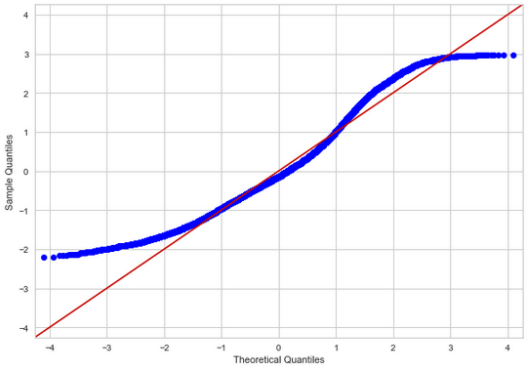


Figura 17: QQ-plot temperatura año 2020

En los gráficos de distribución vemos que los datos no se ajustan correctamente a la distribución normal. En cuanto a los QQ-plot, en ninguno de los dos casos los puntos se encuentran perfectamente en línea recta, sino que se dispersan significativamente, entonces no podemos concluir una relación entre los ejes y, en consecuencia, se concluye que los valores no se distribuyen normalmente.

Aunque gráficamente ya se puede concluir que no existe normalidad, se realiza el test de Shapiro Wilks para ambos años donde se contrastan las siguientes hipótesis:

H0: los datos siguen una distribución Normal

H1: los datos no siguen una distribución Normal

Se obtiene un p-valor = 0.000 < 0.05 tanto en 2019 como en 2020, por lo que se rechaza la hipótesis nula en favor de la alternativa, es decir, se rechaza que los datos se distribuyen según una Normal. Sin embargo, se obtiene el aviso de que "el p-valor puede no ser preciso para N >5000", por lo que se hace el test D'Agostino y se llega a la misma conclusión.

- Estudio de la normalidad en los datos de humedad en ambos años por separado:

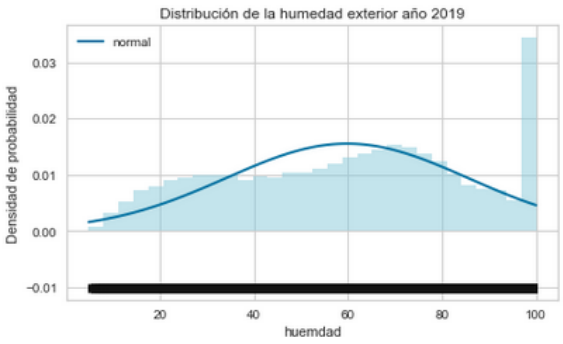


Figura 18: Distribución de la humedad exterior 2019

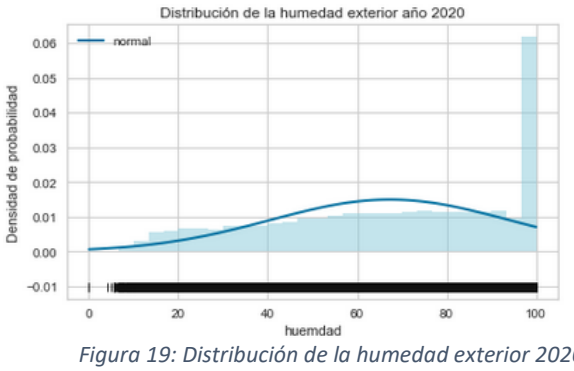


Figura 19: Distribución de la humedad exterior 2020

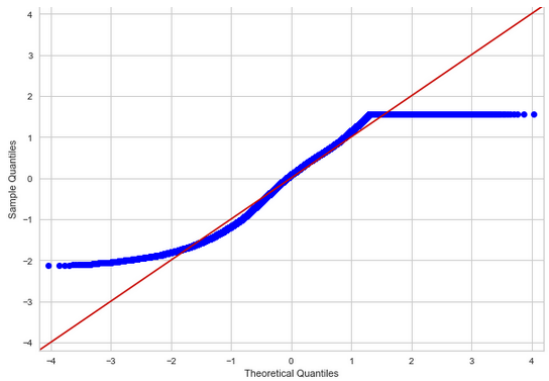


Figura 20: QQ-Plot de la humedad año 2019

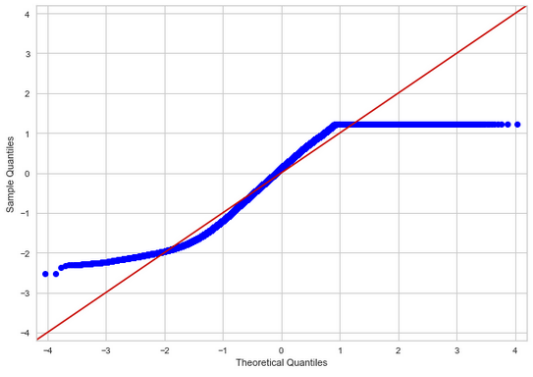


Figura 21: QQ-Plot de la humedad año 2020

En este caso también se puede apreciar la falta de normalidad gráficamente. Realizando el test D'Agostino y obtener un pvalor = 0 < 0.05, se rechaza la hipótesis de normalidad y se concluye que no se puede afirmar que los datos sigan una distribución normal, por lo que se procede a agrupar los datos por semanas, calcular la media de cada una de ellas y normalizar los datos (explicación teórica sección 2.4 *Normalización vs estandarización*). En la Figura 22 y Figura 23 aparecen representadas las dispersiones de la temperatura y la humedad media por semanas y por años. Se observa que la dispersión varía ligeramente entre un año y otro.

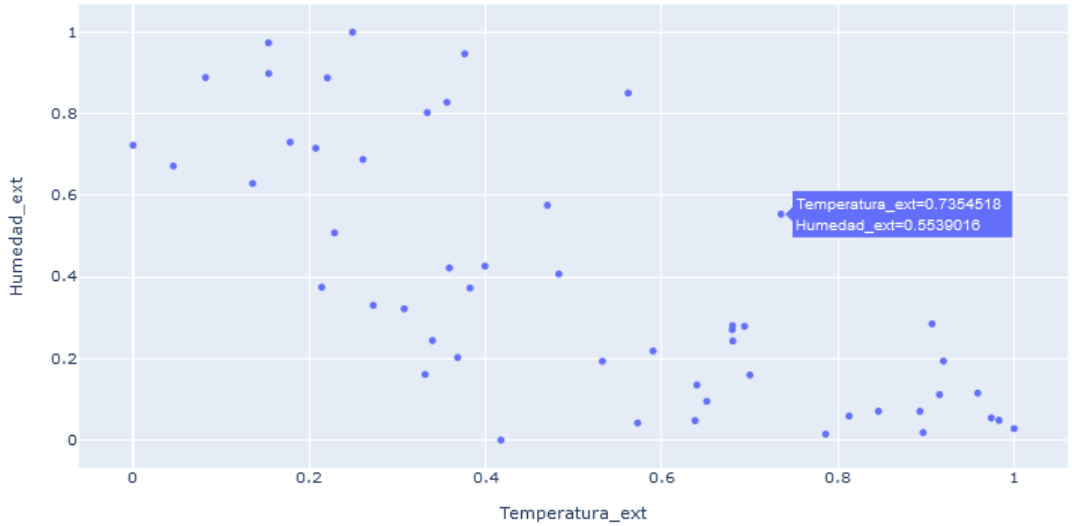


Figura 22: Dispersión datos tª y humedad 2019

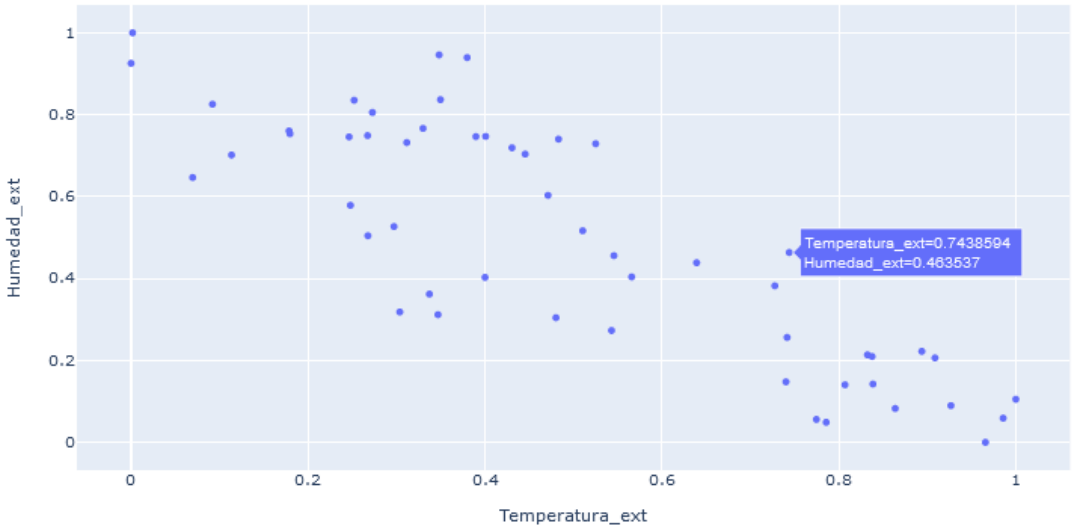


Figura 23: Dispersión datos tª y humedad 2020

Otros gráficos interesantes de cara al estudio son los diagramas lineales de ambas variables sobre un mismo plano y con las variables ya normalizadas, tal y como se observa en la Figura 24 y en la Figura 25, donde en el eje x se representan las semanas y en el eje y los diferentes valores de humedad y temperatura normalizados:

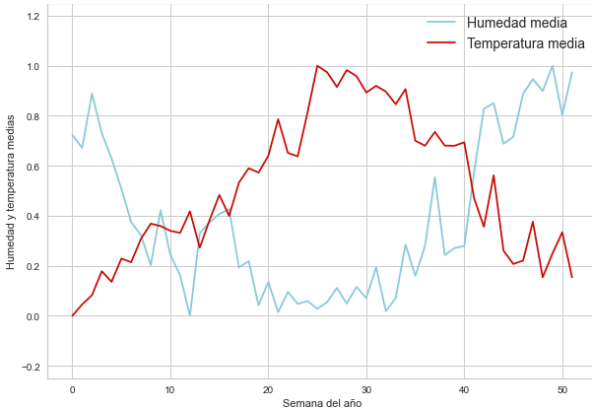


Figura 24: Diagrama lineal de humedad y tª 2019

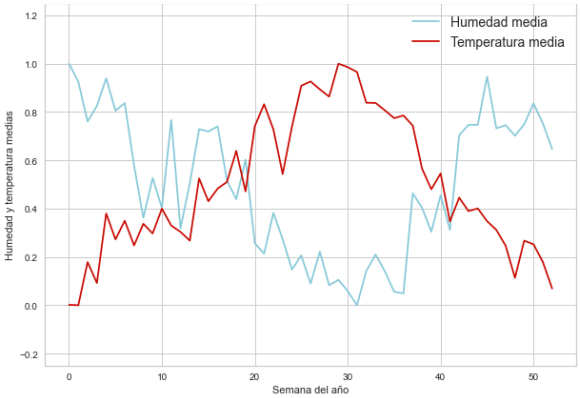


Figura 25: Diagrama lineal de humedad y tª 2020

A partir del gráfico anterior, se puede empezar a sospechar que pueden existir al menos tres grupos de comportamiento, sin embargo, será en el Capítulo 5. *Clustering* donde se creen los modelos y se discutan los resultados.

El siguiente paso consiste en analizar la posible existencia de valores atípicos en el conjunto de datos, ya que su presencia puede afectar negativamente a las conclusiones que se obtengan tras la aplicación de los métodos. Para detectarlos se ha decidido representar un diagrama de caja para cada variable (véanse Figura 26, Figura 27, Figura 28 y Figura 29) y se observa que no hay puntos representados fuera de los bigotes, por lo que gráficamente se puede concluir que no existen valores anormales que difieran del resto de la muestra y, por tanto, no es necesario eliminar ninguna observación más para realizar el análisis.

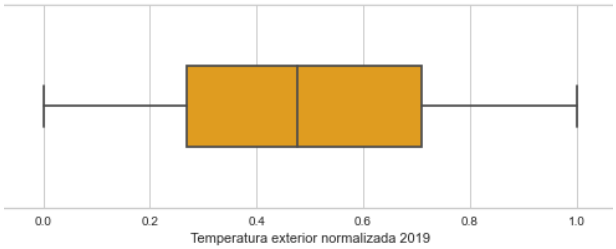


Figura 26: Diagrama de caja de temperatura 2019

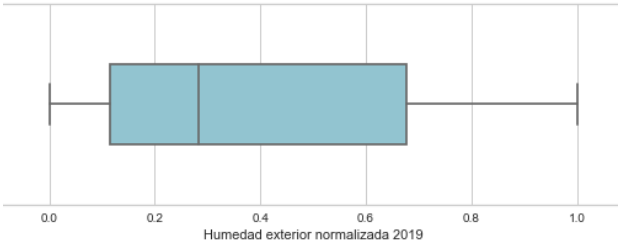


Figura 27: Diagrama de caja de humedad 2019

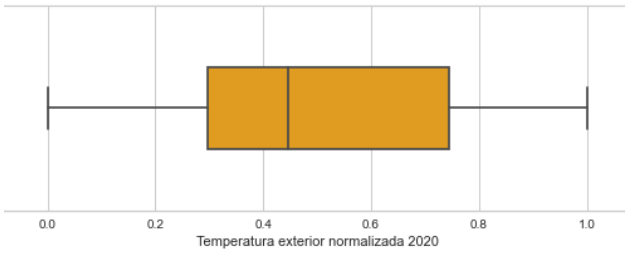


Figura 28: Diagrama de caja de temperatura 2020

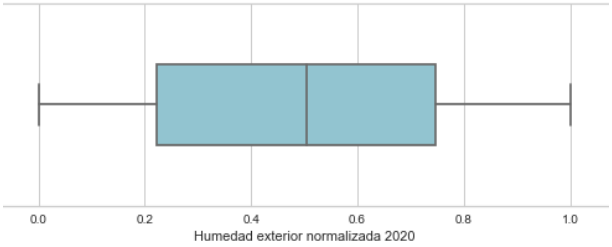


Figura 29: Diagrama de caja de humedad 2020

4.2 Datos del sistema de refrigeración

Por otra parte, tenemos los datos que hacen referencia al consumo energético asociados a la producción de refrigeración del edifico[6]. La refrigeración del edificio LUCIA se lleva a cabo gracias a dos mecanismos, el primero de ellos es una enfriadora eléctrica convencional que consume energía eléctrica para producir el agua fría que se emplea en el sistema de distribución de frío del edificio. Esta variable en DESIGO aparece denominada como "AInt'PS'AnlzRed22'AI65" y se ha renombrado como "red22" para abreviar y facilitar su manipulación.

El segundo mecanismo es un sistema de refrigeración a través de una máquina de absorción que emplea tanto agua caliente como electricidad para su ciclo frigorífico. La electricidad se utiliza en la enfriadora por absorción y en una torre de refrigeración que también consume electricidad y que es necesaria ya que sirve para disipar el calor producido en el ciclo. A la electricidad consumida por absorción "OP01'AInt'PS'AnlzRed23'AI65" y a la energía consumida por la torre de refrigeración "OP01'AInt'PS'AnlzRed24'AI65" se las ha renombrado, respectivamente, como "red23" y "red24". Por último, el agua caliente, que en DESIGO se identifica como "AInt'MtrH1'CumEg", en el desarrollo de este proyecto se ha renombrado como "cumeg".

Hay que mencionar también que, por cada variable de consumo, además de existir una que expresa el momento en el que se toma el registro (fecha), existe otra que se llama "tag de tipo" y que indica si la medición se ha llevado a cabo correctamente (tag de tipo = bien) o si, por el contrario, ha habido algún problema (tag de tipo = Mal, alarma o tag de tipo = Bien, hora cambiada).

Una vez que se han obtenido los datos de DESIGO y se han importado, el siguiente paso es renombrar las variables como se ha explicado y transformar los registros de las fechas para que tengan un formato correcto que permita su tratamiento en Python. En la Figura 30 se puede observar que inicialmente se cuenta con 72053 filas y 12 columnas, sin embargo, ya se aprecian valores sin informar (NaN y NaT) e incluso negativos y algunos registros que no tienen un tag de tipo correcto.

	fecha_red22	red22	tag_22	fecha_red23	red23	tag_23	fecha_red24	red24	tag_24	fecha_cumeg	cumeg	tag_cumeg
0	2019-01-01 13:44:00	-2.0	Bien, Hora cambiada	2019-01-01 13:44:00	-2.0	Bien, Hora cambiada	2019-01-01 13:44:00	-2.0	Bien, Hora cambiada	2019-01-01 00:02:00	38480.0	Bien
1	2019-01-01 17:29:00	190846.3	Bien	2019-01-01 17:29:00	1933.7	Bien	2019-01-01 17:29:00	6204.8	Bien	2019-01-01 00:17:00	38480.0	Bien
2	2019-01-01 19:40:00	190846.3	Bien, Hora cambiada	2019-01-01 19:40:00	1933.7	Bien, Hora cambiada	2019-01-01 19:40:00	6204.8	Bien, Hora cambiada	2019-01-01 00:32:00	38480.0	Bien
3	2019-01-02 04:05:00	190846.3	Bien, Hora cambiada	2019-01-02 04:05:00	1933.7	Bien, Hora cambiada	2019-01-02 04:05:00	6204.8	Bien, Hora cambiada	2019-01-01 00:47:00	38480.0	Bien
4	2019-01-02 04:50:00	190846.3	Bien, Hora cambiada	2019-01-02 04:50:00	1933.7	Bien, Hora cambiada	2019-01-02 04:50:00	6204.8	Bien, Hora cambiada	2019-01-01 01:02:00	38480.0	Bien
...	...	...	...	...	...	...	...	...	...	...	...	...
72049	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	Mal, Alarma
72050	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	Mal, Alarma
72051	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	Mal, Alarma
72052	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	Mal, Alarma
72053	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	NaN	NaT	NaN	Mal, Alarma

Figura 30: Datos del consumo de energía sin procesar

A continuación, se crean cuatro dataframes, cada uno de los cuales contiene toda la información asociada a cada variable de consumo (consumo en KWh, fecha del registro y el tag). Una observación que destacar es que red22, red23, red24 y cumeg miden el consumo acumulado, es decir, miden lo que se ha consumido hasta la fecha indicada, por lo que no tienen sentido valores negativos, nulos o

[6] Si bien es cierto que el *dataset* original cuenta con mucha información acerca de distintas variables, no aparecen registradas todas las que intervienen en la refrigeración del edificio, por lo que se optó por extraerlas de DESIGO directamente.

NaN (equivale a NaT en el caso de fechas), así que se eliminan del estudio y, además, se proceden a seleccionar únicamente aquellos registros cuyo tag de tipo sea estrictamente igual a bien, para evitar posibles fallos en el análisis.

Lo siguiente que se va a hacer es realizar la representación gráfica de cada variable para ver si existen valores anómalos (véase Figura 31).

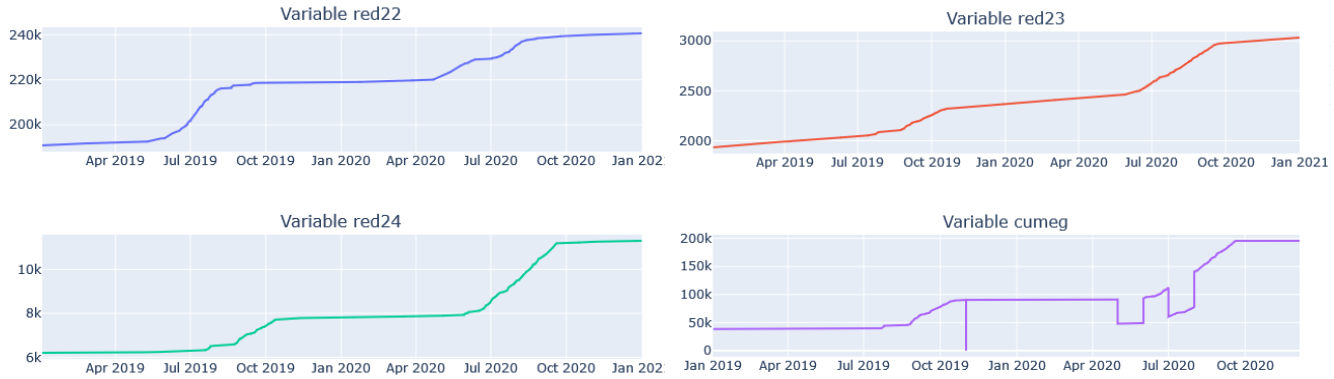


Figura 31: Consumo diario acumulado de todas las variables

En la figura anterior se aprecia que, para todas las variables, el consumo es, en mayor o menor medida, ascendente, salvo para la variable cumeg. Al ser un gráfico interactivo, se puede ver fácilmente en qué momentos se producen esos descensos que no deberían ocurrir (véase Figura 32).

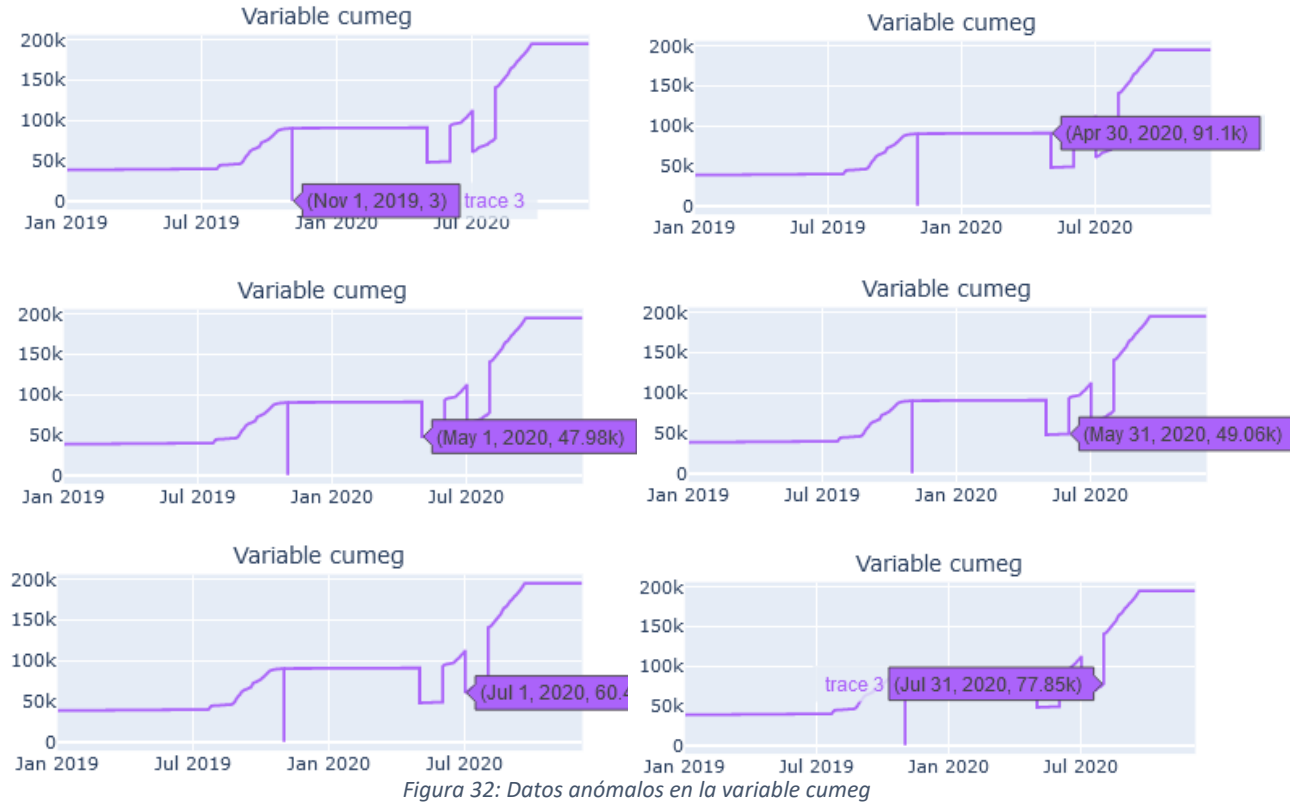


Figura 32: Datos anómalos en la variable cumeg

El 1 de noviembre de 2019, el consumo es 3 KWh, lo cual no tiene sentido, ya que, como se ha comentado anteriormente, el consumo es acumulativo, por lo que se procede a eliminar este registro del estudio. Por otra parte, en la figura también se aprecia que el 30 de abril de 2020 el consumo era 91.1 KWh, pero el día siguiente desciende a 47.98 KWh y el 31 de este mismo mes asciende ligeramente hasta los 49.06 KWh. Algo similar ocurre en el mes de julio. Se asume que es un fallo en el sensor y se deciden trasladar los datos de los meses de mayo y junio tal y como se aprecia en la

Figura 33. Haciendo dicha traslación, se corrige el problema de la disminución del consumo acumulado en los dos meses citados y el estudio no se verá afectado porque los datos con los que se va a trabajar posteriormente son consumos semanales, no acumulados.

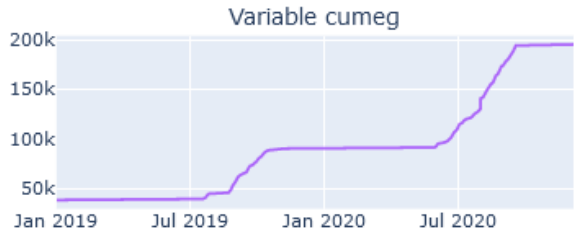


Figura 33: Consumo acumulado corregido de la variable cumeg

Seguidamente, se procede a detectar la presencia de más posibles outliers sobre las columnas que miden el consumo, para ello, se aplica Z-score sobre estos datos con el objetivo de ver cuántas desviaciones estándar se aparta cada registro de la media y, todos aquellos que sean menores que -3 o mayores que 3, en lugar de ser eliminados del estudio, van a ser interpolados con la función "interpolate()" de Python.

En el apartado anterior, los datos relativos a la temperatura y humedad fueron agrupados por semanas y separados por años. Para los datos de consumo, se procede de la misma manera. Sin embargo, surge el problema de que no existe información para las semanas 12 y 20 de red22, red23 y red24 en el año 2019, por lo que se decide interpolar estos valores. Para ello, primero es necesario crear dos nuevas filas en el dataframe, indicando únicamente week = 12 para la primera y week = 20 para la segunda, el resto de los valores de la fila se introducen como nulos y es la función interpolate() de nuevo la que se encarga de obtener dichos valores. Para la variable cumeg no existe este problema para el año 2019, sin embargo, para el 2020, no hay registros que tengan un tag de tipo estrictamente igual a bien para las 4 últimas semanas del año[7] de cumeg, por lo que se aplica la misma lógica que se ha explicado para las variables red22, red23 y red24.

Como se ha mencionado anteriormente, la información disponible hace referencia al consumo acumulado, sin embargo, lo que interesa realmente es saber cuánto se ha consumido semanalmente. Para obtener esta información, se calcula el consumo máximo y el mínimo de cada semana y, posteriormente, se calcula la diferencia obteniendo así el consumo semanal. El motivo de calcularlo de esta manera se debe a que, al tener consumos acumulados, el consumo máximo hace referencia al último día de la semana y, el mínimo, al primero. Este razonamiento se aplica a cada variable y a cada año. Finalmente se crean dos nuevos dataframes llamados consumo_2019 y consumo_2020 que contienen la información relativa al consumo de cada variable por semanas y para cada año.

En la Figura 34 y Figura 35 aparecen, respectivamente, las cinco primeras líneas de los dataframes consumo_2019 y consumo_2020.

[7] El año 2020 es bisiesto, por lo que tiene 53 semanas. En este caso, al mencionar que no hay registros con tipo correcto para las últimas 4 semanas del año, se está haciendo referencia a las semanas 50-53.

week	dif_red22	dif_red23	dif_red24	dif_cumeg
1.0	118.90	3.30	0.00	170.0
2.0	56.30	2.60	0.00	80.0
3.0	96.50	3.90	0.00	0.0
4.0	66.70	3.30	1.60	0.0
5.0	65.00	3.30	1.30	10.0

Figura 34: Consumo semanal año 2019

week	dif_red22	dif_red23	dif_red24	dif_cumeg
1	52.2	4.6	4.4	70.0
2	143.1	4.6	4.6	10.0
3	89.1	4.5	3.5	10.0
4	93.4	4.6	3.5	0.0
5	73.0	4.7	4.3	10.0

Figura 35: Consumo semanal año 2020

Resulta de interés graficar el consumo semanal de cada variable por año, lo cual se puede ver en la Figura 36 y en la Figura 37.

Figura 36: Consumo semanal de cada variable 2019

En la Figura 36 se aprecia que la variable que mide el consumo energético de la enfriadora eléctrica convencional (red22) tiene un comportamiento ligeramente distinto al de las tres variables que intervienen en el consumo de la enfriadora por absorción (red23, red24 y cumeg), ya que, aunque todas presentan un pico entorno a la semana 28, el descenso de las tres últimas variables es mucho mayor que el que experimenta red22, que es más progresivo. Todas las variables presentan otros dos picos alrededor de las semanas 34 y 36, con un posterior descenso en el consumo. Sin embargo, las tres últimas variables presentan, además, un pico llano aproximadamente en la semana 38, lo cual no ocurre en red22.

Figura 37: Consumo semanal de cada variable 2020

En la Figura 37 se observa que en este caso la variable que se comporta diferente a las demás es cumeg que presenta tres picos de consumo en las semanas 18, 27 y 31. Red22 experimenta un descenso significativo en la semana 25 y red23 y red24 lo experimentan un par de semanas después. Claramente son red23 y red24 las variables que tiene un comportamiento más similar en el año 2020.

Al comparar el comportamiento de la misma variable en cada año también se aprecian ligeras diferencias, lo cual sirve para reforzar la decisión de analizar los datos separados por años.

En el ANEXO III – *Comparación gráfica del consumo semanal de cada variable años 2019 y 2020* aparecen representadas las mismas gráficas que en la Figura 36 y la Figura 37, pero más ampliadas, lo cual permite descubrir que, aunque cumeg hace referencia al agua caliente que se consume en la máquina de absorción, vemos que hay muchas semanas donde hay consumos pequeños en los sensores red23 y red24, y sin embargo en esas semanas cumeg vale 0. Sin embargo, no se trata de errores en las mediciones, sino que esos consumos pequeños están asociados a tareas de la máquina o sobre la máquina para su correcto mantenimiento y disponibilidad (arranques programados sin demanda y consumos de elementos de los equipos que permiten garantizar su disponibilidad), por lo que se van a considerar de cara al análisis.

El siguiente paso es estandarizar los datos de consumo de 2019 y 2020 (justificación teórica sección 2.4 *Normalización vs estandarización*), habiendo eliminado previamente la columna "week" y estudiar la correlación entre las variables de cada uno de los años.

Para el 2019, la variable red24 presenta una elevadísima correlación con red23 (véase Figura 38), por lo que se decide eliminar esta primera del estudio. Tras eliminar dicha variable, la correlación entre las variables no es alta (véase Figura 39), por lo que se decide hacer el análisis de clúster con las tres variables restantes para el año 2019.

	dif_red22	dif_red23	dif_red24	dif_cumeg
dif_red22	1.00	0.10	0.13	-0.03
dif_red23	0.10	1.00	0.99	0.53
dif_red24	0.13	0.99	1.00	0.47
dif_cumeg	-0.03	0.53	0.47	1.00

Figura 38: Correlaciones consumo sin red24 2019

	dif_red22	dif_red23	dif_cumeg
dif_red22	1.00	0.10	-0.03
dif_red23	0.10	1.00	0.53
dif_cumeg	-0.03	0.53	1.00

Figura 39: Correlaciones consumo 2020

Con los datos del año 2020, ocurre algo similar, la correlación entre las variables red23 y red24 es muy elevada (véase Figura 40), por lo que se elimina también del estudio y, tras calcular de nuevo las correlaciones, se siguen obteniendo valores altos para la variable dif_cumeg (véase Figura 41). A pesar de la elevada correlación, esta variable no va a ser eliminada del estudio por dos motivos:

a) Es necesaria para poder comparar los resultados que se obtengan con los del año 2019.
b) Realmente el estudio sólo se va a realizar con tres variables que no es un número elevado, por lo que no es tan importante eliminarlas.

	dif_red22	dif_red23	dif_red24	dif_cumeg
dif_red22	1.00	0.32	0.34	0.23
dif_red23	0.32	1.00	0.99	0.98
dif_red24	0.34	0.99	1.00	0.97
dif_cumeg	0.23	0.98	0.97	1.00

Figura 40: Correlaciones consumo 2020

	dif_red22	dif_red23	dif_cumeg
dif_red22	1.00	0.32	0.23
dif_red23	0.32	1.00	0.98
dif_cumeg	0.23	0.98	1.00

Figura 41: Correlaciones consumo sin red24 2020

Capítulo 5. Clustering

5.1 Clustering sobre datos externos

A raíz de los gráficos de la Figura 24 y la Figura 25, se empezó a sospechar que pueden existir al menos tres grupos de comportamiento en los datos de temperatura y humedad, para confirmarlo o desmentirlo se van a aplicar las técnicas de clustering explicadas en la sección *2.2 Técnicas y métricas para el clustering*.

En primer lugar, se va a realizar clustering no jerárquico sobre los datos de humedad y temperatura que han sido depurados en la sección anterior (*4.1 Datos relativos a las condiciones externas de humedad y temperatura*). El primer paso consiste en obtener el número óptimo de clústeres k y, para determinarlo, se va a aplicar el método del codo con el concepto de inercia y el de Silhouette. Gráficamente, para el año 2019 (véase Figura 43) parece que la agrupación óptima de los datos en dos clústeres (k=2) podría ser la óptima ya que es en 2 cuando se aprecia que el valor del error disminuye drásticamente y se forma el "codo" en el gráfico. Para el año 2020 (véase Figura 42) se sospecha que el óptimo es k=3.

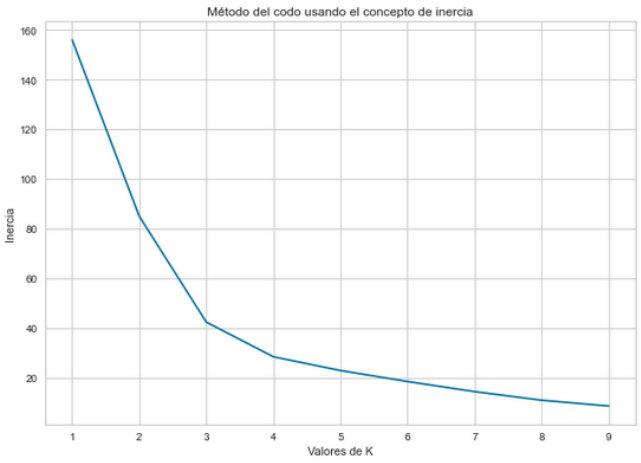


Figura 42: Método del codo datos 2020

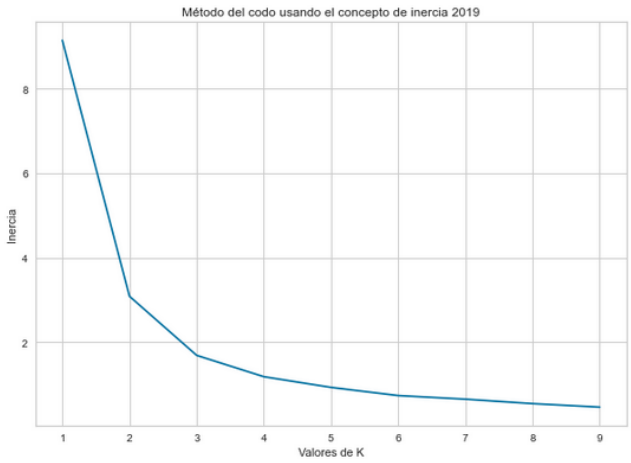


Figura 43: Método del codo datos 2019

Se calcula también el coeficiente de Silhouette (véase Tabla 5) para tener una segunda visión en la elección del número de clústeres y se observa que, en realidad, el coeficiente más alto de Silhouette se obtiene con k = 2 (0.531 en el caso del año 2019 y 0.576 en 2020), sin embargo, con k = 3 se obtiene un resultado muy similar (0.498 y 0.516 respectivamente).

Inercia (con elbow)			**Coeficiente de Silhouette**		
k	**2019**	**2020**	**k**	**2019**	**2020**
2	3.083	2.455	2	0.531	0.576
3	1.682	1.342	3	0.498	0.516
4	1.177	0.999	4	0.486	0.464
5	0.923	0.739	5	0.435	0.471
6	0.729	0.609	6	0.457	0.452
7	0.644	0.514	7	0.463	0.475
8	0.538	0.428	8	0.448	0.378
9	0.455	0.339	9	0.416	0.411

Tabla 5: Inercia con método del codo y coeficiente de Silhouette condiciones externas

Se va a aplicar un último método para determinar el k óptimo: el algoritmo DBSCAN utilizando como parámetros *eps* = 0.3 (distancia máxima de vecindad) y *min_samples* = 6 (número mínimo de puntos para formar una agrupación). El resultado para el año 2019 es que cada punto tiene asignada una de las tres etiquetas siguientes: {0, 1, 2}. Concretamente, hay 15 semanas a las que se las asigna la etiqueta 0; 14 a las que se las asigna la 1 y 23 que son asignadas a la 2. Para el año 2020 también se obtienen tres grupos, pero el número de individuos que pertenece a cada uno varía ligeramente: existen 23 elementos asignados al clúster etiquetado como 0, 15 asignados al 1 y otros 15 al 2.

En consecuencia, se va a representar el resultado de los algoritmos DBSCAN y K-Means con k = 3. En caso de que los resultados obtenidos carezcan de sentido o la agrupación no sea clara, se estudiará si se deben elegir 2 o 4 grupos en lugar de 3.

Tras representar gráficamente el resultado del algoritmo K-Means con k=3 y DBSCAN con los datos relativos al año 2019, se obtiene que hay tres grupos distintos tal y como se puede observar en la Figura 44:

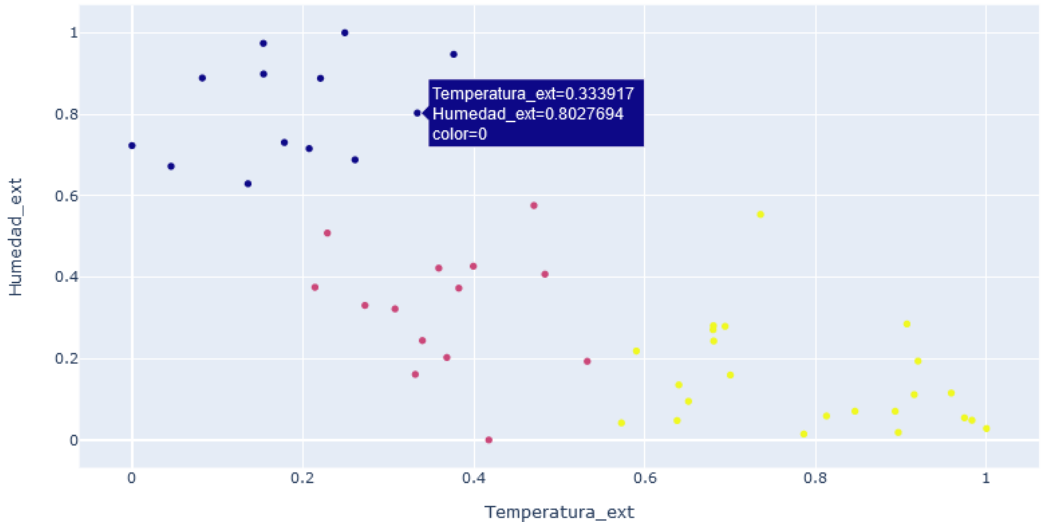


Figura 44: Gráfico interactivo clustering datos 2019

En el caso del año 2020 también se distinguen tres grupos como se puede ver en la figura siguiente:

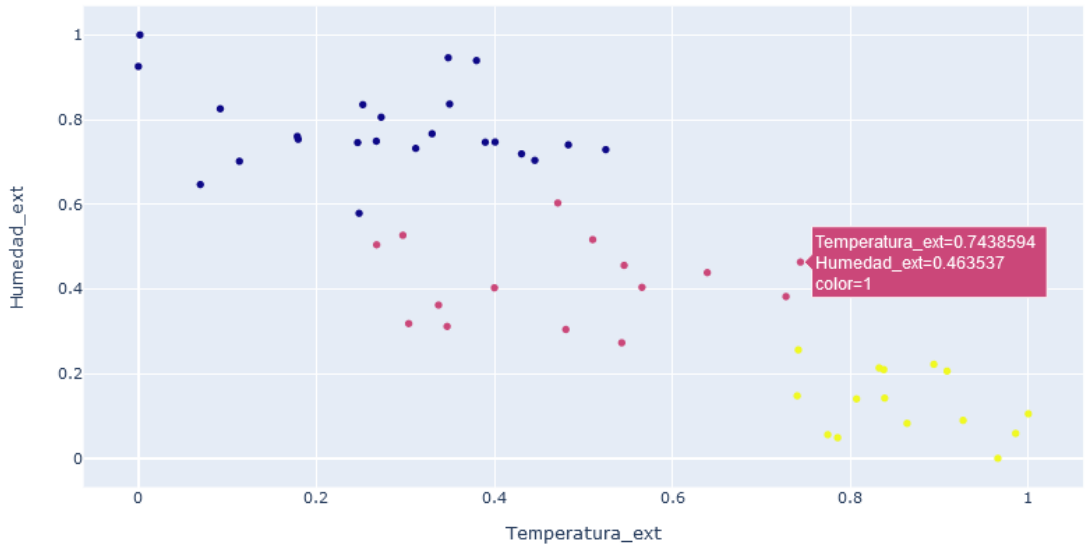


Figura 45: Gráfico interactivo clustering datos 2020

Anteriormente, al observar la Figura 24 y en la Figura 25 se sospechó de la posible presencia de 3 modelos de comportamiento y tras realizar análisis de clúster no jerárquico (véanse Figura 44 y Figura 45) se ha concluido que los datos de humedad y temperatura se pueden agrupar en 3 grupos tanto en 2019 como en 2020. En ambos casos, el primer grupo aparece representado en color azul (etiqueta clúster = 0) y se caracteriza por tener unos valores de humedad muy altos y unos de temperatura muy bajos. El segundo grupo es el que aparece en fucsia (etiqueta clúster = 1) y tiene unos valores intermedios tanto para la variable de humedad como para la de temperatura. Por último, tenemos los puntos amarillos de la que conforman la tercera agrupación (etiqueta clúster = 2). Este tercer grupo se caracteriza por tener valores bajos de humedad y altos de temperatura.

Se decide realizar también un gráfico 3D interactivo donde se representa las semanas del año a las que hacen referencia cada uno de los clústeres (véanse Figura 46 y Figura 47) y ver si existe algún tipo de relación entre los grupos que hemos obtenido y las semanas del año.

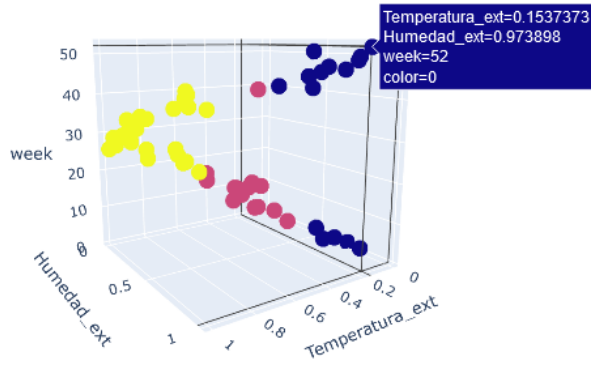


Figura 46: Gráfico interactivo clústeres datos externos 2019

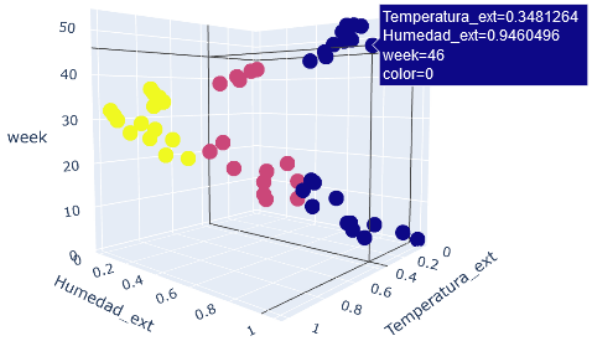


Figura 47: Gráfico interactivo clústeres datos externos 2020

En los gráficos anteriores, los datos se mantienen agrupados y muestran que, en las primeras y últimas semanas del año (puntos representados en azul), están asociadas al clúster 0 que representa temperaturas bajas y humedades altas. El clúster 1 está asociado con semanas que corresponden a un periodo transitorio entre temperaturas altas y humedades bajas o viceversa. Por último, las semanas correspondientes a los meses de verano están asociadas al clúster 2 que se caracteriza por tener unos valores de temperatura altos y humedades bajas.

A continuación, volvemos a representar la humedad y la temperatura media por semanas tal y como se hizo en la Figura 24 y en la Figura 25, pero esta vez, el color de cada semana va a corresponder con el clúster al que está asociado. Para los años 2019 y 2020 se obtiene:

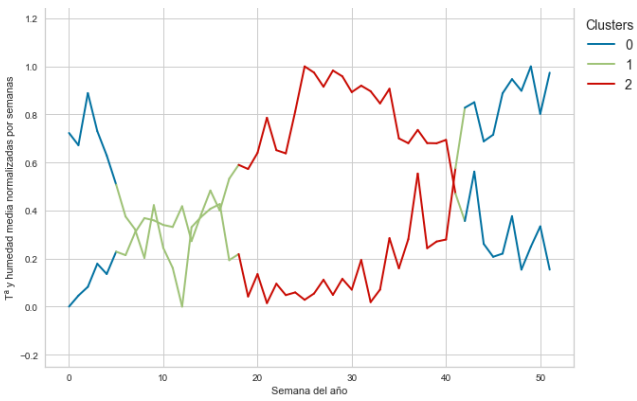


Figura 48: Distribución hª y tª por semanas y clústeres 2019

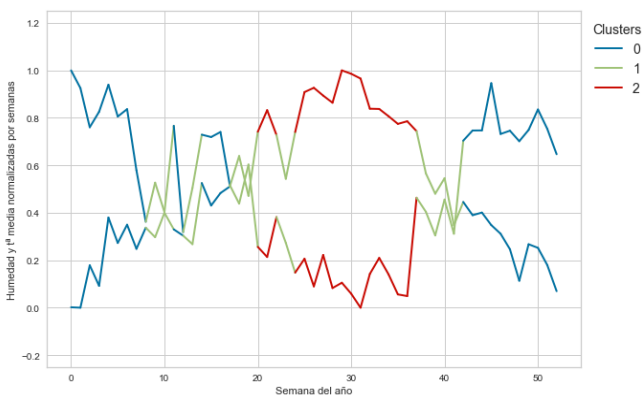


Figura 49: Distribución hª y tª por semanas y clústeres 2020

También se ha realizado clustering jerárquico sobre el conjunto de datos de humedad y temperatura utilizando single linkage, complete linkage y average linkage (explicación teórica en la sección 2.2.2 Clustering jerárquico) y se han obtenido, al igual que con clustering no jerárquico, tres grupos que aparecen representados en tres colores distintos: verde, rojo y morado. Para ver los dendogramas que han resultado del clustering jerárquico aglomerativo del año 2019 y del año 2020 consultar ANEXO II – *Resultados de aplicar clustering jerárquico sobre las variables de condiciones externas*.

5.2 Clustering sobre datos de consumo energético

La lógica que se va a aplicar para realizar clustering sobre los datos relativos al consumo energético asociados a la refrigeración del edificio es análoga a la que se ha seguido en la sección anterior con los datos de humedad y temperatura.

En primer lugar, es necesario determinar el número de clústeres con el cuál se debe aplicar K-Means, tanto para el año 2019 como para el 2020. Con el método del codo y observando la Figura 50 y la Figura 51 se podría justificar utilizar 3 o 4 clústeres.

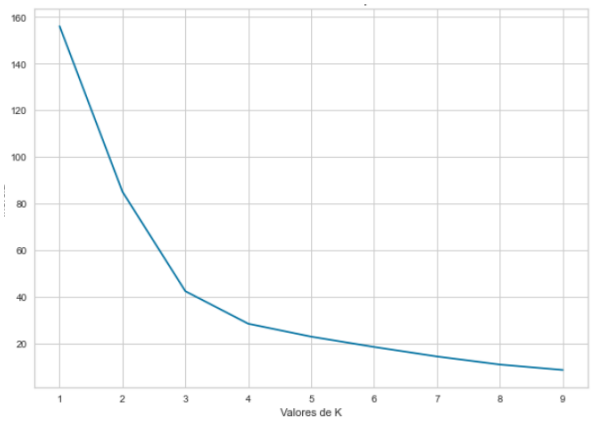


Figura 50: Método del codo consumo energía 2019

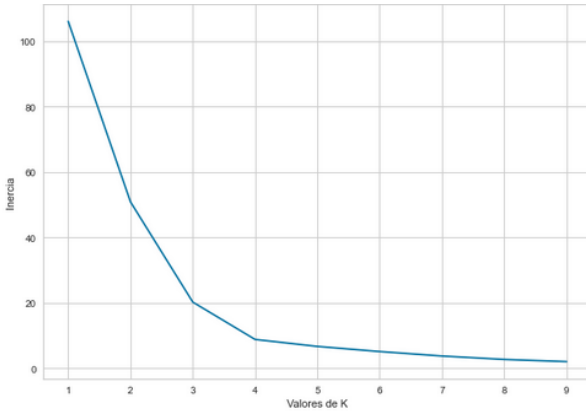


Figura 51: Método del codo consumo energía 2020

Para tomar la decisión más adecuada, se decide calcular también el coeficiente de Silhouette (véase Tabla 6) para tener una segunda visión en la elección del número de clústeres y se observa que el coeficiente más alto de Silhouette se obtiene con k = 4 (0.702 en el caso del año 2019 y 0.789 en 2020).

Inercia[8] (con elbow)			**Coeficiente de Silhouette**		
k	**2019**	**2020**	**k**	**2019**	**2020**
2	84.955	50.872	2	0.633	0.627
3	42.280	20.265	3	0.699	0.747
4	28.326	8.868	4	0.702	0.789
5	22.787	6.735	5	0.695	0.770
6	18.365	5165	6	0.698	0.725
7	14.283	3.804	7	0.690	0.724
8	10.823	2.759	8	0.640	0.730
9	8.489	2.094	9	0.634	0.728

Tabla 6: Inercia con método del codo y coeficiente de Silhouette del consumo energético

Como la elección del número óptimo de clústeres no está del todo clara en este caso, se decide aplicar k-means con k = 3 y con k = 4 y representar gráficamente el resultado para ver cuál de las dos permite

[8] La inercia es la suma de las distancias al cuadrado de cada punto del clúster a su centroide.

agrupar mejor los datos. Tanto para el año 2019 como para el año 2020, se aplica k-means con las variables red22, red23 y cumeg, ya que red24 fue eliminada del estudio debido a su alta correlación.

El resultado gráfico de la aplicación del algoritmo con k = 3 es el siguiente:

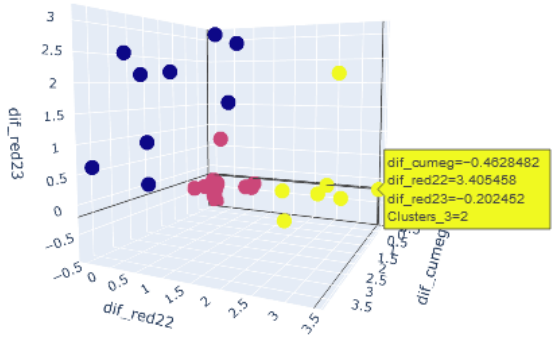


Figura 52: Gráfico 3D interactivo 3 clústeres consumo 2019

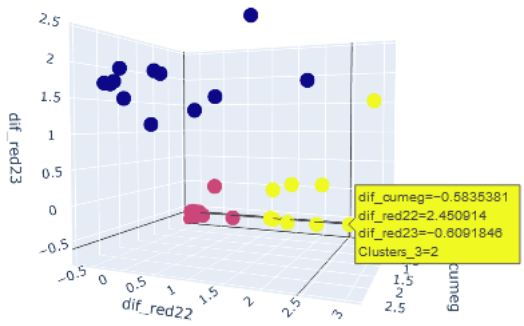


Figura 53:Gráfico 3D interactivo 3 clústeres consumo 2020

El resultado gráfico de la aplicación del algoritmo con k = 4 es el siguiente:

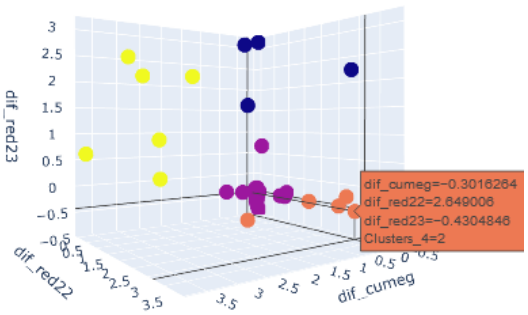


Figura 54: Gráfico 3D interactivo 4 clústeres consumo 2019

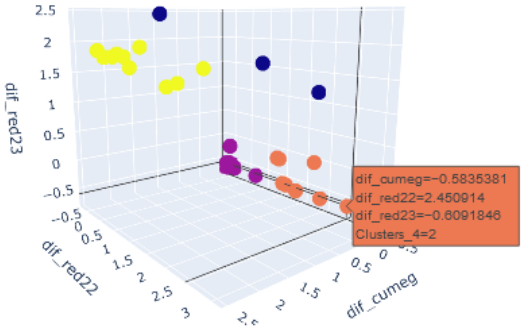


Figura 55: Gráfico 3D interactivo 4 clústeres consumo 2020

Anteriormente se ha visto que el método del codo admite tanto k = 3 como k = 4 y el de Silhouette tiene un valor muy próximo para 3 y 4 tanto para el año 2019 como para el año 2020. Gráficamente, en la Figura 52 y en la Figura 53 se observa que con k = 3 los datos se pueden separar e identificar. Además, facilitará la posterior comparación, por tanto, se elige 3 como número de clústeres. En caso de que los resultados no sean coherentes o no permitan sacar conclusiones, se cambiará la selección del parámetro y se repetirá el análisis de nuevo.

Para el año 2019, en la Figura 52 se aprecia que existen tres grupos diferenciados, cada uno de los cuales está representado en un color distinto. El grupo en color amarillo está etiquetado con el número 2 y se caracteriza por tener unos valores muy altos de la variable red22 y muy bajos de las variables red23 y cumeg. Como red22 hace referencia a la enfriadora número 1 y las otras dos a la enfriadora número 2, lo que lo que realmente está sucediendo es que la enfriadora 1 está en funcionamiento. El clúster 0, representado en azul, se caracteriza justo, por lo contrario, ya que tiene valores altos de consumo de red23 y cumeg y prácticamente nulos de la variable red22, es decir, se caracteriza por representar que la enfriadora 2 está activa. El grupo 1 es de color fucsia y tiene, en general, consumos prácticamente nulos o muy bajos de energía, lo que significa que ninguna de las dos enfriadoras está en funcionamiento. Esos pequeños consumos que se dan en algunas semanas cuando la enfriadora correspondiente no está funcionando, no son sinónimo de que está activa, sino que son necesarios para garantizar el correcto funcionamiento y mantenimiento del propio sistema de refrigeración.

Para poder explicar la relación entre los clústeres obtenidos y las semanas a las que están asociados, se ha añadido una columna con esta información al dataframe que contiene las variables de consumo del 2019 y los clústeres asignados. Véase ANEXO IV - Variables de consumo y clústeres asignados apartado IV.i. De este anexo se ha extraído la información que aparece en la Tabla 7:

Clúster	Significado	Semanas
1	Consumos nulos de energía	1-24, 32, 43-52
2	Enfriadora 1 activa	25-31
0	Enfriadora 2 activa	33-41

Tabla 7: Relación clústeres consumo y semanas año 2019

Las semanas a las que hacen referencia los datos del clúster 1 (consumo prácticamente nulo de energía) comprenden de la 1 a la 24, es decir, desde enero hasta la 2ª semana de junio, y de la 43 a la 52, es decir, de la 4ª semana de octubre hasta final de diciembre. En la semana 32 realmente hay un pequeño consumo de energía de la variable red23, pero no parece ser lo suficientemente grande para que se asigne al clúster 0. Exceptuando esto, los resultados son coherentes, ya que en los meses más fríos del año el edificio no está consumiendo energía para producir refrigeración. El clúster 2 (solo la enfriadora 1 está activa) está asociado a las semanas 25-31, es decir, a las dos últimas semanas de junio hasta la primera de agosto. Por último, las semanas 33-41 están asociadas con el clúster 0 caracterizado por el funcionamiento único de la enfriadora 2. Dichas semanas comprenden desde la 3ª de agosto a la 1ª de octubre.

Para el año 2020, en la Figura 53 se aprecia que existen tres grupos diferenciados y que se obtienen los mismos comportamientos que en 2019:

- Clúster 1 en color fucsia, consumos prácticamente nulos de energía.
- Clúster 2 representado en amarillo y funcionamiento único de la enfriadora 1.
- Clúster 0 en azul, funcionamiento único de la enfriadora 2.

Para poder explicar la relación entre los clústeres obtenidos y las semanas a las que están asociados, se ha añadido una columna con esta información al dataframe que contiene las variables de consumo del 2020 y los clústeres asignados. Véase ANEXO IV - *Variables de consumo y clústeres asignados* apartado IV.ii. De este anexo se ha extraído la información que aparece en la Tabla 8:

Clúster	Significado	Semanas
1	Consumos nulos de energía	1-16, 24, 39-53
2	Enfriadora 1 activa	17-23, 29, 31
0	Enfriadora 2 activa	25-28, 30, 32-38

Tabla 8: Relación clústeres consumo y semanas año 2020

El clúster 1 (consumos prácticamente nulos de energía) está asociado a las semanas 1-16, es decir, de enero a la 3ª semana de abril, y 39-53, es decir, de la 4ª semana de septiembre hasta finales de diciembre. En la semana 24 realmente hay un pequeño consumo de energía de la variable red22, pero no parece ser lo suficientemente grande para que se asigne al clúster 2. El clúster 2 (funcionamiento único de la enfriadora 1) está asociado a las semanas 17-23, es decir, de la 4ª semana de abril a la 1ª de junio, y a las semanas 29 y 31 que corresponden, respectivamente, a la 3ª y 5ª de julio. El clúster 0 (funcionamiento único de la enfriadora 2) se asocia a las semanas 25-28 y 30, es decir, de la 3ª de junio a la 2ª y 4ª semana de julio, y de la 32-38, es decir, de la 2ª de agosto hasta la 3ª de septiembre.

5.3 Relaciones entre clústeres de condiciones externas y de consumo

Por último, se van a comparar los clústeres de condiciones externas y de consumo que se han asignado a cada una de las semanas para cada año. Para ello se han creado dos tablas, una para el 2019 y otra para el 2020, que contienen la semana, el clúster de humedad y temperatura y el clúster de condiciones externas asignados. El objetivo es poder comparar qué clústeres de consumo tienen asignadas las semanas asociadas a los clústeres 0, 1 y 2 de humedad y temperatura. Véase *ANEXO V – Relación entre clústeres de condiciones externas y consumo* para consultar las tablas completas. A partir de la información de este anexo se han realizado las siguientes tablas:

Week	Clúster H&T	Clúster consumo
1, 2, 3, 4, 5, 43, 44, 45, 46, 47, 48, 49, 50, 51 y 52	0	1
6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 y 42	1	1
33, 34, 35, 36, 37, 38, 39, 40 y 41	2	0
19, 20, 21, 22, 23, 24 y 32	2	1
25, 26, 27, 28, 29, 30 y 31	2	2

Tabla 9: Clústeres asociados a cada semana 2019

Week	Clúster H&T	Clúster consumo
1, 2, 3, 4, 5, 6, 7, 8, 12, 15, 16, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52 y 53	0	1
17	0	2
38	1	0
9, 10, 11, 13, 14, 24, 39, 40, 41 y 42	1	1
18, 19, 20 y 23	1	2
25, 26, 27, 28, 30, 32, 33, 34, 35, 36 y 37	2	0
21, 22, 29 y 31	2	2

Tabla 10: Clústeres asociados a cada semana 2020

Tanto para el año 2019 como para el año 2020 se han obtenido 3 clústeres de condiciones externas y 3 clústeres de consumo cuyos comportamientos se recogen brevemenete en la Tabla 11:

Clúster	**0**	**1**	**2**
Condiciones externas	Humedad alta Temperatura baja	Humedad media Temperatura media	Humedad baja Temperatura alta
Consumo	Enfriadora 2 activa	Nulo de energía	Enfriadora 1 activa

Tabla 11: Resumen significado de pertenencia a cada clúster

A partir de la información de las tablas anteriores se observa que en 2019 todas las semanas asignadas al clúster 0 de condiciones externas, tienen asignadas el clúster 1 de consumo, es decir, no hay refrigeración en las semanas más frías. El comportamiento en 2020 es análogo, salvo por la semana 17 en la que la enfriadora 1 está activa.

Todas las semanas asociadas al clúster 1 de condiciones externas en 2019, están asociadas al clúster 1 de consumo, es decir, cuando la humedad y temperatura son medias, no es necesaria la refrigeración del edificio. En el año 2020 ocurre lo mismo con la mayoría de las semanas, pero hay pequeñas excepciones para algunas semanas en las que está activa alguna enfriadora. El comportamiento para humedad y temperaturas medias puede ser una evidencia de la capacidad aislante que tiene el edificio.

Para las semanas con temperaturas más cálidas del año, tanto en el año 2019 como en 2020, se observa que la mayor parte de ellas están asociadas con algún clúster que representa que una de las enfriadoras está activa. Sin embargo, en algunas semanas cálidas del 2019 no hay refrigeración. De nuevo esto podría deberse a la eficiencia del aislamiento del edificio. En el año 2020 también existen algunas excepciones, pero hay que tener en cuenta que las condiciones de consumo de las semanas de marzo a junio fueron condicionadas por la pandemia, con lo cual sería necesario analizar los datos de todo 2021 para sacar conclusiones más sólidas.

Del análisis anterior, se puede observar que existen 3 zonas: dos correspondientes al invierno y al verano y una de transición que corresponde a las estaciones de otoño y primavera. Tras haber estudiado la relación entre dichas zonas y los comportamientos en el consumo, se deriva que las condiciones ambientales condicionan, en cierta medida, el funcionamiento del sistema de refrigeración del edificio. Aunque existen algunos comportamientos que no se ajustan bien, podrían deberse a variaciones climatológicas puntuales o al funcionamiento elegido por los gestores de LUCIA en estas semanas.

Capítulo 6. Conclusiones y trabajo futuro

Con el desarrollo de este proyecto se han cubierto todas las tareas planteadas inicialmente para conseguir el objetivo único del estudio: comprobar si es posible identificar distintos modos de comportamiento en el edificio LUCIA utilizando datos externos de temperatura y humedad y datos de los sistemas de refrigeración del edificio de los años 2019 y 2020.

En este trabajo se ha aplicado la metodología CRISP para analizar dos tipos de datos: humedad y temperatura, asociados a condiciones externas, y consumos de electricidad y agua caliente, asociados al sistema de refrigeración de LUCIA. Tras analizar y procesar los datos en bruto, se han transformado y se ha usado la metodología de creación de clústeres que se ha aplicado a ambos tipos de datos, obteniéndose un conjunto de clústeres que se han comparado, es decir, se han relacionado los clústeres de condiciones externas con los de consumo, todos ellos asignados a las distintas semanas del año. Se ha observado que, tanto en 2019 como en 2020, las semanas asignadas al clúster de condiciones que representa temperaturas bajas y humedad alta, están asociadas a su vez al clúster de consumo que indica que no se consume energía para refrigerar el edificio. Todas las semanas asignadas al clúster con temperaturas y humedades medias en 2019, están también asignadas al clúster de consumo nulo de energía, aunque en 2020 existe una pequeña variación de semanas asignadas a algún clúster de consumo de energía no nulo, esto no supone una diferencia representativa. Por último, las semanas asignadas al clúster de temperatura alta y humedad baja, están asignadas mayoritariamente a clústeres en los que el sistema de refrigeración está activo, aunque, de forma similar al clúster anterior, existe una pequeña variación de semanas entre ambos años, que no supone una diferencia representativa.

Del análisis realizado se observa que hay una relación entre las condiciones externas y el funcionamiento del sistema de refrigeración, ya que, durante las semanas más calurosas, el sistema de refrigeración está generalmente en funcionamiento y en las más frías e incluso templadas no se consume energía para refrigerarlo. También se observan claramente tres modos de funcionamiento en los sistemas en función de las condiciones externas que podrían recomendar acciones de control distintas a las actuales, donde solo hay dos modos.

Como trabajo futuro se puede plantear realizar el mismo estudio, pero con datos relativos a más años para ver cómo evolucionan tanto las condiciones externas como el consumo y las relaciones entre ambos. Otros trabajos futuros podrían ser realizar un estudio similar, pero empleando los datos del sistema de calefacción de LUCIA o con los datos de un edificio de la Universidad que no sea inteligente y analizar si los consumos de energía derivados de la refrigeración y calefacción del edificio se comportan de manera similar a los de LUCIA y si las relaciones con las condiciones externas son más o menos evidentes.

Capítulo 7. Bibliografía

[1] Mohamed, A., Hasan, A., & Sirén, K. (2014). Fulfillment of net-zero energy building (NZEB) with four metrics in a single-family house with different heating alternatives. Applied Energy, 114, 385-399.

URL:https://doi.org/10.1016/j.apenergy.2013.09.065

[2] Bedi, P., Jindal, V., & Gautam, A. (2014, September). Beginning with big data simplified. In 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC) (pp. 1-7). IEEE. URL:Beginning with big data simplified | IEEE Conference Publication | IEEE Xplore

[3] Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), 2053951716631130.

URL: https://doi.org/10.1177/2053951716631130

[4] Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013, May). Addressing big data issues in scientific data infrastructure. In 2013 International Conference on Collaboration Technologies and Systems (CTS) (pp. 48-55). IEEE.

URL: https://ieeexplore.ieee.org/document/6567203

[5] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

URL: http://www.sciencedirect.com/science/article/pii/S026840121 4001066

[6] Dumbill, E. (2013). Making Sense of Big Data. Big data, 1 1, 1-2.

URL: https://doi.org/10.1089/big.2012.1503

[7] Camargo-Vega, J. J., Camargo-Ortega, J. F., & Joyanes-Aguilar, L. (2015). Conociendo big data. Revista Facultad de Ingeniería, 24(38), 63-77.

URL: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-11292015000100006

[8] CRISP-DM URL: CRISP-DM: La metodología para poner orden en los proyectos - Sngular Fecha de último acceso: 18-08-2022

[9] Blum, A., & Mitchell, T. (1998, July). Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational learning theory (pp. 92-100).

URL: https://dl.acm.org/doi/pdf/10.1145/279943.279962

[10] Kothari, R., & Jain, V. (2002, May). Learning from labeled and unlabeled data. In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290) (Vol. 3, pp. 2803-2808). IEEE.

URL:https://ieeexplore.ieee.org/abstract/document/1007592?casa_token=n2cZzI0q5YAAAAA:1FqqKnalC96hjXYQQHemOhUsnkNLk0u1UP5sYBCiAp87thycHzWJJPybaer4KRSknMn5Qyz

[11] Hastie, T., Tibshirani, R., Friedman, J. (2009). Overview of Supervised Learning. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY.

URL:https://doi.org/10.1007/978-0-387-84858-7_2

[12] Andrew Ng. (2012) CS229 Lecture notes Machine Learning - Supervised learning. Standford University.

URL:https://www.studocu.com/en-us/document/stanford-university/machine-learning/cs229-notes-1-machine-learning-by-andrew/1403919

[13] Rokach, L., Maimon, O. (2009). Classification Trees. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA.

URL:https://doi.org/10.1007/978-0-387-09823-4_9

[14] Betancourt, G. A. (2005). Las máquinas de soporte vectorial (SVMs). *Scientia et technica*, 1(27).

URL:https://revistas.utp.edu.co/index.php/revistaciencia/article/view/6895

[15] Reddy, Y. C. A. P., Viswanath, P., & Reddy, B. E. (2018). Semi-supervised learning: A brief review. International Journal of Engineering & Technology, 7(1.8), 81.

URL:https://www.researchgate.net/profile/EswaraB/publication/324050146_Semisupervised_learning_a_brief_review/links/5b5c02f8458515c4b24e2b15/Semi-supervised-learning-a-brief-review.pdf

[16] Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 3(1), 1-130.

URL: https://www.morganclaypool.com/doi/abs/10.2200/S00196ED1V01Y200906AIM006

[17] Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020, July). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

URL:https://ieeexplore.ieee.org/abstract/document/9207304casa_token=l7x65TmHJlgAAAAA:9NwnE3LHNho1IdjBYrJSizU7HGwvN7MlKzM3Pkx2VUifTZ2bfmHsycTI6uAMBRJ2cbfXBw7OI45I

[18] Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285.

URL:https://www.jair.org/index.php/jair/article/view/10166/24110

[19] Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3), 279-292.

URL: https://doi.org/10.1007/BF00992698

[20] Cover, T. M., Diday, E., Rosenfeld, A., Simon, J. C., Wagner, T. J., Weszka, J. S., & Wolf, J. J.

(1976). Digital pattern recognition (pp. 52-53). K. S. Fu (Ed.). Berlin: Springer-Verlag.

URL: https://link.springer.com/chapter/10.1007/978-3-642-96303-2_3

[21] García Escudero L. and Fernández Temprano, M., (2020). Clasificación no supervisada: análisis de clúster.

URL:https://cursoanterior4.campusvirtual.uva.es/2020_2021/pluginfile.php/1418091/mod_resource/content/2/anda2.pdf

[22] Sasirekha, K., & Baby, P. (2013). Agglomerative hierarchical clustering Algorithm-A. International Journal of Scientific and Research Publications, 83(3), 83.

URL:https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.1033&rep=rep1&type=pdf

[23] Marutho, D., Handaka, S. H., & Wijaya, E. (2018, September). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In 2018 International Seminar on Application for Technology of Information and Communication (pp. 533-538). IEEE.

URL:https://ieeexplore.ieee.org/document/8549751

[24] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of clúster analysis. Journal of Computational and Applied Mathematics. Vol.20, pp.53-65.

URL:https://archive.org/details/sim_journal-of-computational-and-applied-mathematics_1987-11_20/page/52/mode/2up

[25] Kumar, K. M., & Reddy, A. R. M. (2016). A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. Pattern Recognition, 58, 39-48.

URL: https://doi.org/10.1016/j.patcog.2016.03.008

[26] Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E., & Malley, J. D. (2011). Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. Genetic epidemiology, 35(S1), S5-S11.

URL:https://onlinelibrary.wiley.com/doi/epdf/10.1002/gepi.20642

[27] Day, W. H., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. Journal of classification, 1(1), 7 24.

URL:https://link.springer.com/article/10.1007/bf01890115

[28] Jarman, A. M. (2020). Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. Georgia Southern University.

URL:https://www.researchgate.net/profile/AngurJarman/publication/339443595_Hierarchical_Clúster_Analysis_Comparison_of_Single_linkageComplete_linkage_Average_linkage_and_Centroid_Linkage_Method/links/5e52f582a6fdcc2f8f5d6f2e/Hierarchical-Clúster-Analysis-Comparison-of-Single-linkage-Complete-linkage-Average-linkage-and-Centroid-Linkage-Method.pdf

[29] Anaconda-Python URL: https://anaconda.org/anaconda/python Fecha de último acceso: 18-07-2022.

[30] Abdi, H., & Williams, L. J. (2010). Principal Component Analysis. Wiley interdisciplinary reviews: Computational Statistics, 2(4), 433-459.

URL:https://doi.org/10.1002/wics.101

[31] García Escudero, L. and Fernández Temprano, M., (2020). Análisis en componentes principales.

URL:https://cursoanterior4.campusvirtual.uva.es/2020_2021/pluginfile.php/1418063/mod_resource/content/5/anda2019.pdf

[32] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1-4). Springer, Berlin, Heidelberg.

URL:https://link.springer.com/content/pdf/10.1007/978-3-642-00296-0_5.pdf

[33] Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: a technical report. Machine Learning Technical Reports, 1(1), 1-6.

URL:https://docs.google.com/document/d/1x0A1nUz1WWtMCZb5oVzF0SVMY7a_58KQulqQVT8LaVA/edit#

[34] Rey-Hernández, J., Lorenzo-González, S., San José-Alonso (2018). Design and Characterization of a Geothermal Earth-Air Heat Exchager (EAHX), Applied to NZEB Building.

URL:https://iifiir.org/en/fridoc/design-and-characterization-of-a-geothermal-earth-air-heat-exchanger-33753

[35] Rey-Hernández, J., Velasco-Gómez, E., San José-Alonso, J., Tejero-González, A., González-González, S., & Rey-Martínez, F. (2018). Monitoring Data Study of the Performance of Renewable Energy Systems in a Near Zero Energy Building in Spain: A Case Study. Energies, 11(11), 2979.

URL:https://doi.org/10.3390/en11112979

[36] Turégano, J. A., Hernández, M. A., & García, F. (2003). La inercia térmica de los edificios y su incidencia en las condiciones de confort como refuerzo de los aportes solares de carácter pasivo. Zaragoza, España.

URL:http://dx.doi.org/10.3390/en11112979

[37] Lee, K. P., & Chen, H. L. (2013). Analysis of energy saving potential of air-side free cooling for data centers in worldwide climate zones. Energy and Buildings, 64, 103-112.

URL:https://www.sciencedirect.com/science/article/abs/pii/S0378778813002491

[38] SIEMENS DESIGO URL: Sistema Desigo | Automatización de edificios y sistemas de control | Siemens Spain Fecha de último acceso: 29/10/2022

ANEXO I - Condiciones externas y clústeres

I.i) Resultados 2019

Temperatura_ext	Humedad_ext	Clusters	week
0.000000	0.722876	0	1
0.045725	0.671942	0	2
0.082253	0.888869	0	3
0.178290	0.730446	0	4
0.135711	0.629220	0	5
0.228647	0.508137	1	6
0.214111	0.374970	1	7
0.307784	0.321956	1	8
0.368427	0.202616	1	9
0.358982	0.422057	1	10
0.339859	0.244278	1	11
0.331518	0.161211	1	12
0.417580	0.000000	1	13
0.272612	0.330571	1	14
0.382461	0.372858	1	15
0.483252	0.407103	1	16
0.399451	0.426735	1	17
0.532742	0.193279	1	18
0.590227	0.218596	2	19
0.572769	0.042011	2	20
0.639921	0.135276	2	21
0.786151	0.014694	2	22
0.651185	0.095039	2	23
0.637896	0.047761	2	24
0.812644	0.059017	2	25
1.000000	0.028183	2	26

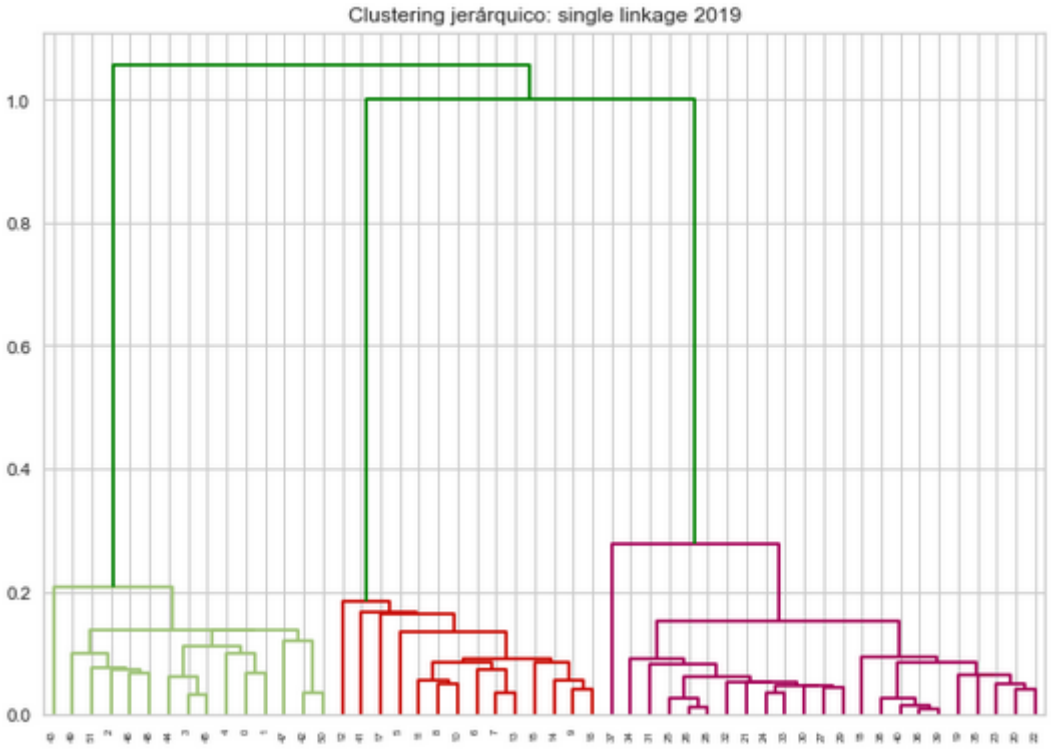
Temperatura_ext	Humedad_ext	Clusters	week
0.974129	0.054512	2	27
0.915330	0.111506	2	28
0.982895	0.048599	2	29
0.958696	0.115292	2	30
0.893071	0.070643	2	31
0.919800	0.193853	2	32
0.896686	0.018653	2	33
0.846020	0.070808	2	34
0.906817	0.284932	2	35
0.700178	0.159553	2	36
0.680390	0.280661	2	37
0.735452	0.553902	2	38
0.680723	0.243114	2	39
0.679993	0.270852	2	40
0.694052	0.279116	2	41
0.470347	0.575750	1	42
0.356242	0.828140	0	43
0.562026	0.850723	0	44
0.260951	0.688047	0	45
0.207308	0.715563	0	46
0.220534	0.888069	0	47
0.376431	0.946936	0	48
0.153986	0.898670	0	49
0.249082	1.000000	0	50
0.333917	0.802769	0	51
0.153737	0.973898	0	52

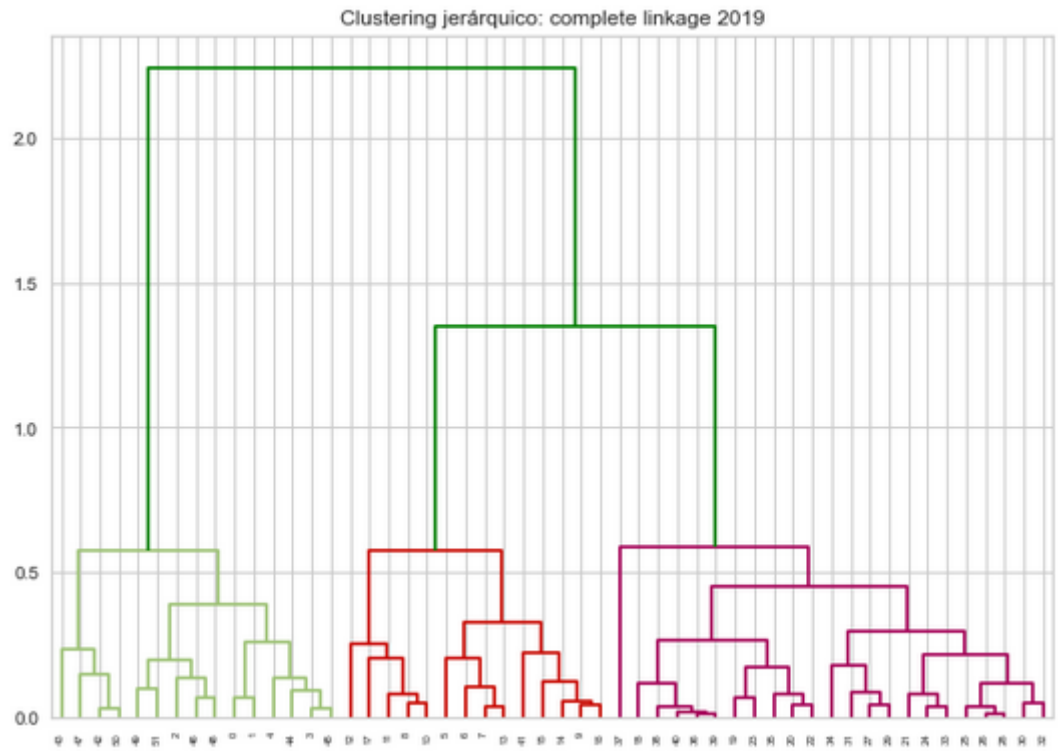
I.ii) Resultados 2020

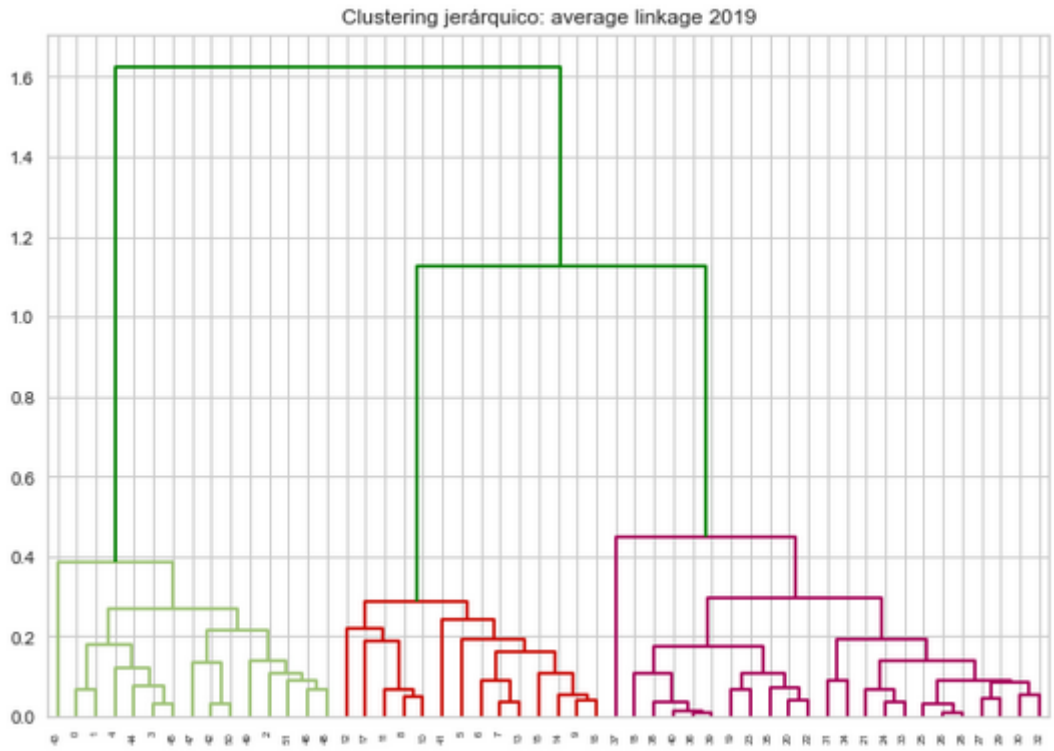
Temperatura_ext	Humedad_ext	Clusters	week
0.001738	1.000000	0	1
0.000000	0.925448	0	2
0.178547	0.760355	0	3
0.092019	0.825595	0	4
0.379853	0.939480	0	5
0.272767	0.805588	0	6
0.349661	0.836683	0	7
0.247968	0.578802	0	8
0.337204	0.362065	1	9
0.297173	0.526846	1	10
0.400040	0.402816	1	11
0.329972	0.766553	0	12
0.303727	0.318277	1	13
0.267705	0.504465	1	14
0.525148	0.729130	0	15
0.430490	0.719146	0	16
0.483068	0.740397	0	17
0.510445	0.516654	1	18
0.639224	0.438639	1	19
0.471221	0.603119	1	20
0.741488	0.256215	2	21
0.832264	0.213657	2	22
0.727601	0.382110	1	23
0.543074	0.272984	1	24
0.740194	0.147695	2	25
0.908589	0.206148	2	26
0.926630	0.089728	2	27
0.893710	0.222091	2	28
0.863668	0.082758	2	29
1.000000	0.105148	2	30
0.985647	0.058961	2	31
0.965654	0.000000	2	32
0.838450	0.142336	2	33
0.837474	0.209642	2	34
0.806884	0.140468	2	35
0.774558	0.055944	2	36
0.785646	0.048948	2	37
0.743859	0.463537	1	38
0.565784	0.403937	1	39
0.480241	0.304466	1	40
0.545755	0.455770	1	41
0.346866	0.311671	1	42
0.445431	0.703784	0	43
0.389744	0.746642	0	44
0.400778	0.747054	0	45
0.348126	0.946050	0	46
0.311594	0.732036	0	47
0.246391	0.745810	0	48
0.113548	0.701592	0	49
0.267364	0.749050	0	50
0.252152	0.835297	0	51
0.179564	0.753545	0	52
0.069616	0.646735	0	53

ANEXO II – Resultados de aplicar clustering jerárquico sobre las variables de condiciones externas

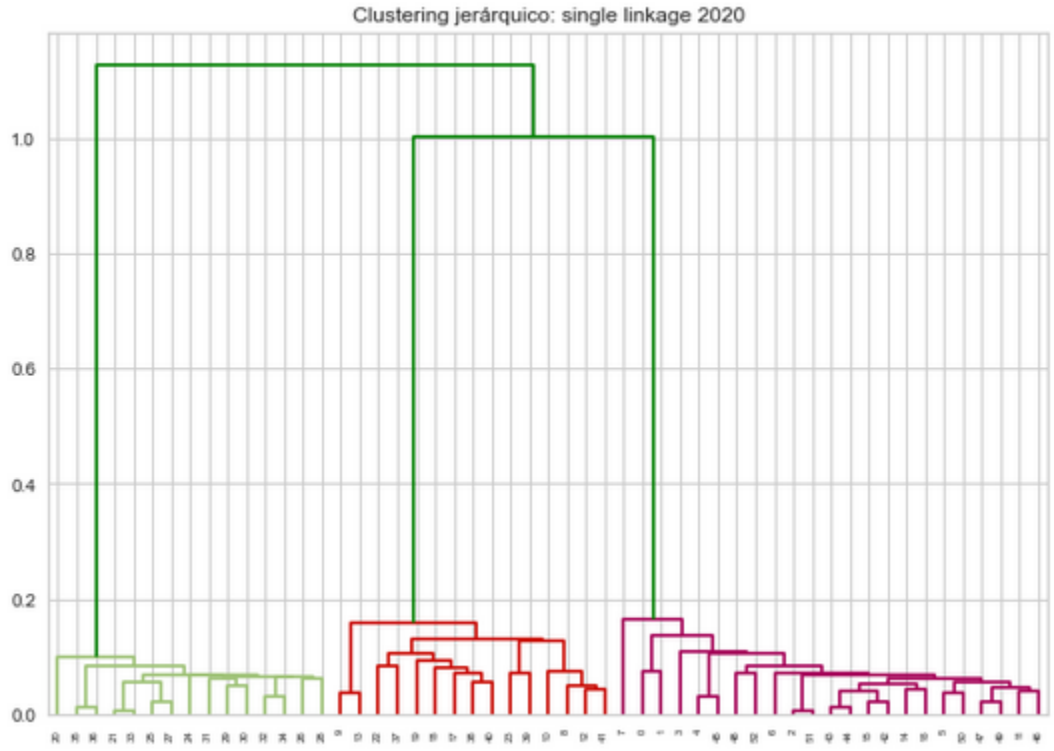
II.i) Resultados 2019

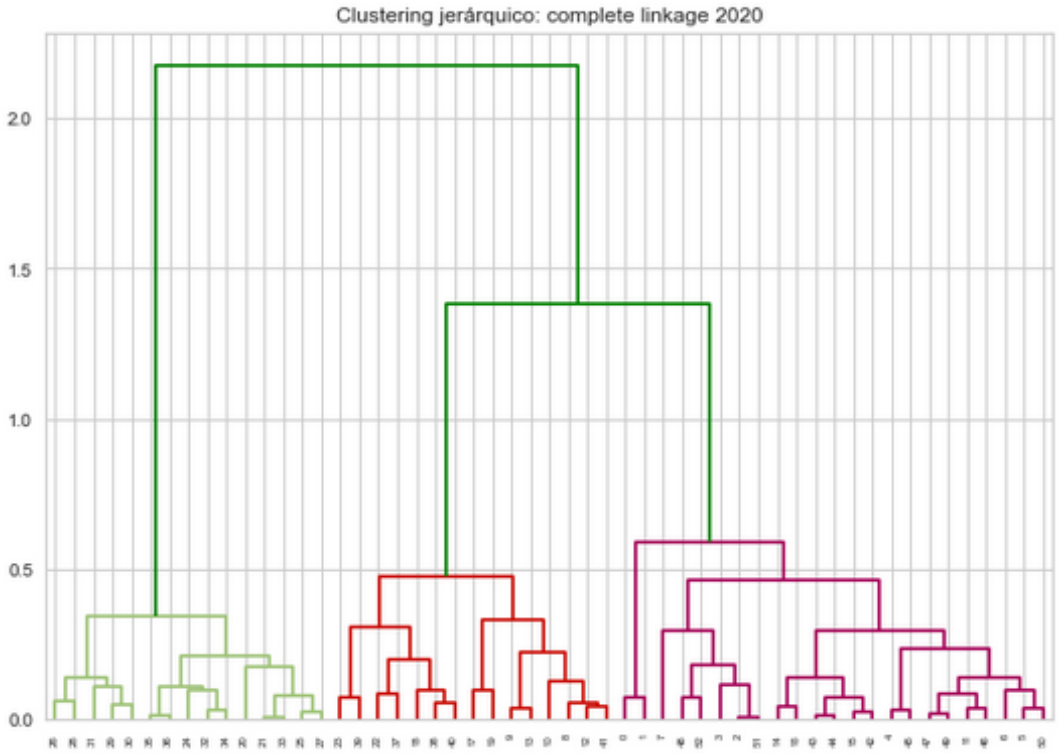


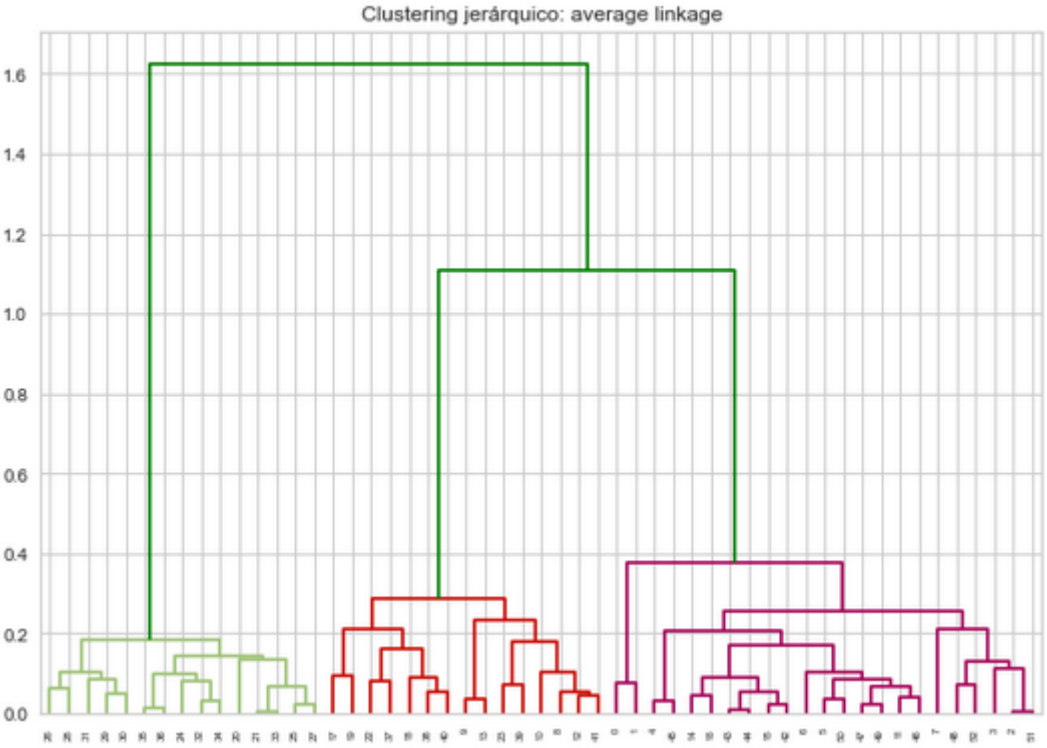




II.ii) Resultados 2020

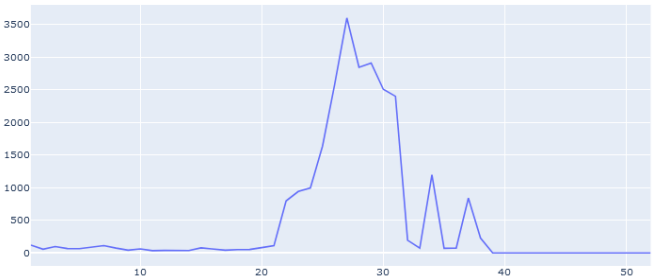



Clustering jerárquico: complete linkage 2020
2.0
1.5
1.0
0.5
0.0

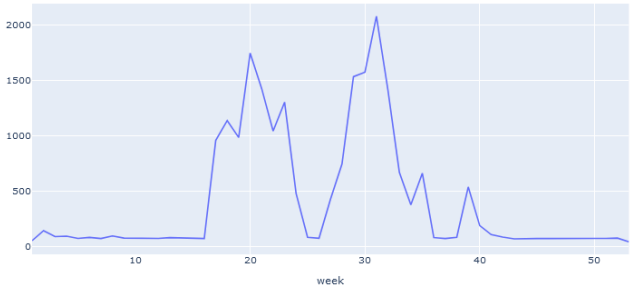

Clustering jerárquico: average linkage
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

ANEXO III – Comparación gráfica del consumo semanal de cada variable años 2019 y 2020

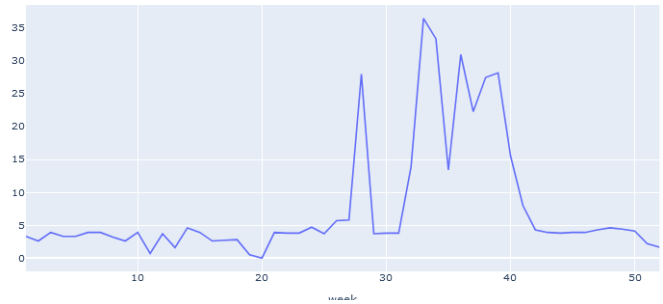
Consumo de la variable red22 año 2019



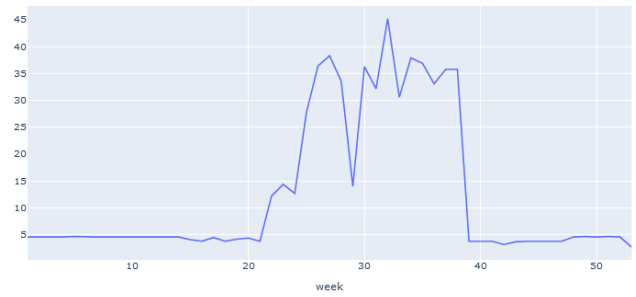
Consumo de la variable red22 año 2020



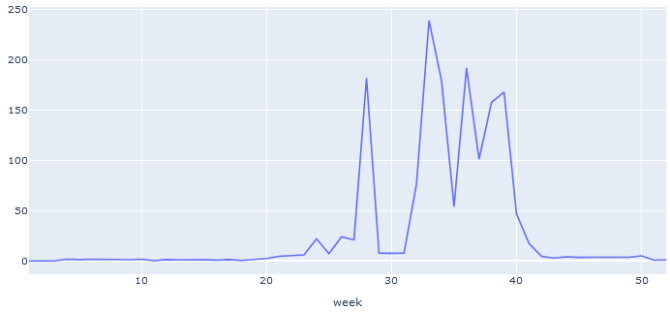
Consumo de la variable red23 año 2019



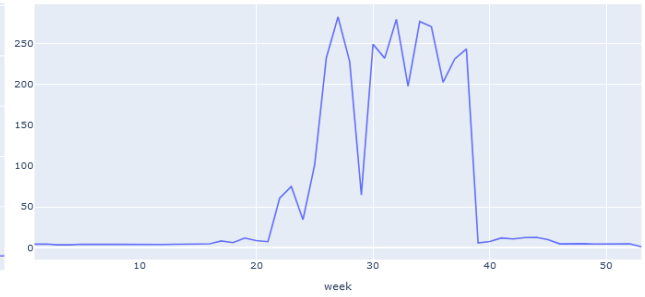
Consumo de la variable red23 año 2020



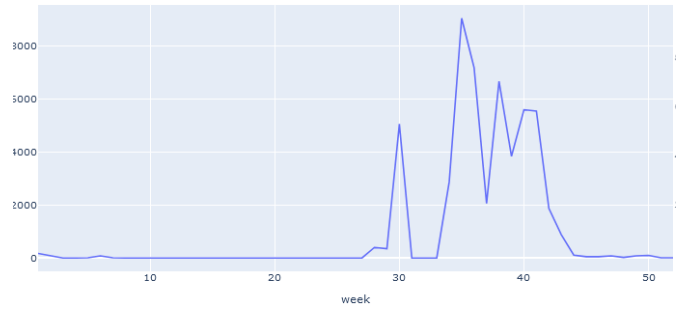
Consumo de la variable red24 año 2019



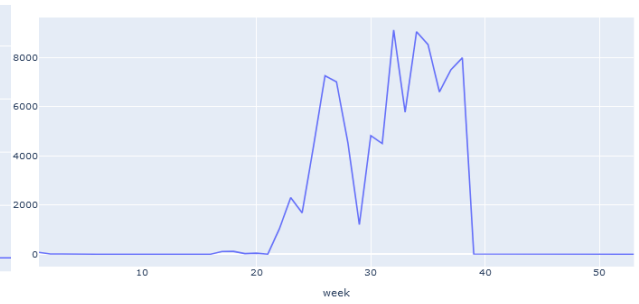
Consumo de la variable red24 año 2020



Consumo de la variable cumeg año 2019



Consumo de la variable cumeg año 2020



ANEXO IV - Variables de consumo y clústeres asignados

IV.i) Resultados 2019 con k=3

dif_red22	dif_red23	dif_cumeg	Clusters_3	week
-0.401729	-0.473919	-0.384540	1	1.0
-0.470169	-0.549930	-0.425998	1	2.0
-0.426219	-0.408767	-0.462848	1	3.0
-0.458799	-0.473919	-0.462848	1	4.0
-0.460657	-0.473919	-0.458242	1	5.0
-0.433872	-0.408767	-0.425998	1	6.0
-0.410584	-0.408767	-0.458242	1	7.0
-0.451364	-0.484778	-0.462848	1	8.0
-0.487006	-0.549930	-0.462848	1	9.0
-0.464375	-0.408767	-0.462848	1	10.0
-0.494222	-0.756245	-0.462848	1	11.0
-0.490395	-0.430485	-0.462848	1	12.0
-0.486569	-0.658517	-0.462848	1	13.0
-0.492144	-0.332756	-0.462848	1	14.0
-0.447647	-0.408767	-0.462848	1	15.0
-0.464265	-0.549930	-0.462848	1	16.0
-0.487006	-0.539072	-0.458242	1	17.0
-0.476073	-0.528213	-0.462848	1	18.0
-0.475308	-0.777963	-0.462848	1	19.0
-0.443547	-0.832256	-0.462848	1	20.0
-0.411787	-0.408767	-0.462848	1	21.0
0.338433	-0.419626	-0.462848	1	22.0
0.496852	-0.419626	-0.462848	1	23.0
0.555999	-0.321898	-0.462848	1	24.0
1.258880	-0.430485	-0.462848	2	25.0
2.295981	-0.213311	-0.462848	2	26.0

dif_red22	dif_red23	dif_cumeg	Clusters_3	week
3.405458	-0.202452	-0.462848	2	27.0
2.578380	2.208178	-0.278595	2	28.0
2.649006	-0.430485	-0.301626	2	29.0
2.210157	-0.419626	1.867958	2	30.0
2.092846	-0.419626	-0.462848	2	31.0
-0.320169	0.677102	-0.462848	1	32.0
-0.453770	3.131167	-0.462848	0	33.0
0.775424	2.794548	0.859170	0	34.0
-0.454644	0.622809	3.705886	0	35.0
-0.454207	2.533939	2.844501	0	36.0
0.386210	1.589233	0.486057	0	37.0
-0.283106	2.153885	2.609578	0	38.0
-0.531722	2.229896	1.305985	0	39.0
-0.531722	0.861700	2.116700	0	40.0
-0.531722	0.036439	2.093668	0	41.0
-0.531722	-0.365332	0.398537	1	42.0
-0.531722	-0.408767	-0.057491	1	43.0
-0.531722	-0.419626	-0.412178	1	44.0
-0.531722	-0.408767	-0.439817	1	45.0
-0.531722	-0.408767	-0.439817	1	46.0
-0.531612	-0.365332	-0.425998	1	47.0
-0.531722	-0.332756	-0.453636	1	48.0
-0.531722	-0.354474	-0.425998	1	49.0
-0.531722	-0.387050	-0.416785	1	50.0
-0.531722	-0.593365	-0.458242	1	51.0
-0.531722	-0.653088	-0.458242	1	52.0

IV.ii) Resultados 2020 con k = 3

dif_red22	dif_red23	dif_cumeg	Clusters_3	week
-0.665370	-0.594240	-0.573411	1	1
-0.498120	-0.594240	-0.593666	1	2
-0.597476	-0.601713	-0.593666	1	3
-0.589565	-0.594240	-0.597041	1	4
-0.627099	-0.586768	-0.593666	1	5
-0.611276	-0.594240	-0.597041	1	6
-0.628571	-0.594240	-0.597041	1	7
-0.586621	-0.594240	-0.597041	1	8
-0.622315	-0.594240	-0.597041	1	9
-0.625627	-0.601713	-0.597041	1	10
-0.621947	-0.594240	-0.597041	1	11
-0.627835	-0.594240	-0.593666	1	12
-0.613852	-0.601713	-0.597041	1	13
-0.617715	-0.594240	-0.593666	1	14
-0.628387	-0.631601	-0.597041	1	15
-0.629307	-0.654017	-0.597041	1	16
1.000500	-0.601713	-0.559907	2	17
1.330951	-0.654017	-0.558895	2	18
1.048522	-0.624129	-0.590290	2	19
2.450914	-0.609185	-0.583538	2	20
1.863426	-0.654017	-0.597041	2	21
1.157446	-0.026364	-0.252708	2	22
1.636563	0.138022	0.179396	2	23
0.121567	0.010997	-0.029905	1	24
-0.609436	1.131806	0.891693	0	25
-0.624339	1.781875	1.853800	0	26
0.024787	1.923845	1.770080	0	27
0.607860	1.572658	0.932202	0	28
2.059194	0.108133	-0.191944	2	29
2.134263	1.774403	1.030101	0	30
3.064346	1.468049	0.922075	2	31
1.841715	2.439417	2.481701	0	32
0.468945	1.348496	1.354179	0	33
-0.069233	1.893956	2.458070	0	34
0.456066	1.819236	2.282528	0	35
-0.613484	1.535297	1.630996	0	36
-0.629859	1.737043	1.934819	0	37
-0.609068	1.737043	2.100234	0	38
0.227547	-0.654017	-0.593666	1	39
-0.414772	-0.654017	-0.593666	1	40
-0.564726	-0.654017	-0.597041	1	41
-0.605940	-0.698849	-0.597041	1	42
-0.635011	-0.661489	-0.597041	1	43
-0.629675	-0.654017	-0.593666	1	44
-0.630043	-0.654017	-0.597041	1	45
-0.629675	-0.654017	-0.597041	1	46
-0.627651	-0.654017	-0.597041	1	47
-0.624523	-0.594240	-0.597041	1	48
-0.623419	-0.586768	-0.597041	1	49
-0.628203	-0.594240	-0.597041	1	50
-0.628571	-0.586768	-0.597041	1	51
-0.622499	-0.594240	-0.597041	1	52
-0.683401	-0.736210	-0.597041	1	53

ANEXO V – Relación entre clústeres de condiciones externas y consumo

V.i) Resultados 2019

week	Cluster_H&T	Cluster_3
1.0	0	1
2.0	0	1
3.0	0	1
4.0	0	1
5.0	0	1
6.0	1	1
7.0	1	1
8.0	1	1
9.0	1	1
10.0	1	1
11.0	1	1
12.0	1	1
13.0	1	1
14.0	1	1
15.0	1	1
16.0	1	1
17.0	1	1
18.0	1	1
19.0	2	1
20.0	2	1
21.0	2	1
22.0	2	1
23.0	2	1
24.0	2	1
25.0	2	2
26.0	2	2

week	Cluster_H&T	Cluster_3
27.0	2	2
28.0	2	2
29.0	2	2
30.0	2	2
31.0	2	2
32.0	2	1
33.0	2	0
34.0	2	0
35.0	2	0
36.0	2	0
37.0	2	0
38.0	2	0
39.0	2	0
40.0	2	0
41.0	2	0
42.0	1	1
43.0	0	1
44.0	0	1
45.0	0	1
46.0	0	1
47.0	0	1
48.0	0	1
49.0	0	1
50.0	0	1
51.0	0	1
52.0	0	1

Desglose por clústeres:

Week	Clúster H&T	Clúster consumo
6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 y 42	1	1
25, 26, 27, 28, 29, 30 y 31	2	2
1, 2, 3, 4, 5, 43, 44, 45, 46, 47, 48, 49, 50, 51 y 52	0	1
33, 34, 35, 36, 37, 38, 39, 40 y 41	2	0
19, 20, 21, 22, 23, 24 y 32	2	1

V.ii) Resultados 2020

week	Cluster_H&T	Cluster_consumo
1	0	1
2	0	1
3	0	1
4	0	1
5	0	1
6	0	1
7	0	1
8	0	1
9	1	1
10	1	1
11	1	1
12	0	1
13	1	1
14	1	1
15	0	1
16	0	1
17	0	2
18	1	2
19	1	2
20	1	2
21	2	2
22	2	2
23	1	2
24	1	1
25	2	3
26	2	3
27	2	3
28	2	3
29	2	2
30	2	0
31	2	0
32	2	0
33	2	3
34	2	3
35	2	3
36	2	3
37	2	3
38	1	3
39	1	1
40	1	1
41	1	1
42	1	1
43	0	1
44	0	1
45	0	1
46	0	1
47	0	1
48	0	1
49	0	1
50	0	1
51	0	1
52	0	1
53	0	1

Desglose por clústeres:

Week	Clúster H&T	Clúster consumo
9, 10, 11, 13, 14, 24, 39, 40, 41 y 42	1	1
21, 22, 29 y 31	2	2
1, 2, 3, 4, 5, 6, 7, 8, 12, 15, 16, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52 y 53	0	1
17	0	2
38	1	0
18, 19, 20 y 23	1	2
25, 26, 27, 28, 30, 32, 33, 34, 35, 36 y 37	2	0