UNIVERSIDAD INTERNACIONAL DE LA RIOJA

**Universidad Internacional de La Rioja (UNIR)**

**Escuela de Ingeniería**

**Máster Análisis y Visualización de Datos Masivos / Visual Analytics and Big Data**

# Detección de anomalías en series temporales multivariantes

***Trabajo Fin de Máster***

Presentado por: ***González San Francisco, Alberto***
Director: ***Alcaide Villar, Daniel***

Ciudad: *Madrid, España*
Fecha: 27/07/2017

# *ÍNDICE DE CONTENIDO*

**RESUMEN** ..... **6**

**ABSTRACT** ..... **8**

**1. INTRODUCCIÓN** ..... **9**

**1.1 Motivación** ..... **10**

**1.2 Planteamiento de trabajo** ..... **11**

**1.3 Estructura de capítulos** ..... **11**

**2 ESTADO DEL ARTE** ..... **12**

**2.1 Situación actual** ..... **12**

**2.2 ¿Qué son las anomalías?** ..... **13**

**2.3 Taxonomía de las anomalías** ..... **14**

2.3.1 Naturaleza de los datos de entrada ..... 15

2.3.2 Tipos de outliers ..... 15

2.3.3 Etiquetas en los datos ..... 17

2.3.4 Presentación de anomalías ..... 19

**2.4 Aplicaciones** ..... **20**

2.4.1 Intrusión en sistemas informáticos ..... 20

2.4.2 Fraude ..... 21

2.4.3 Sector Industrial ..... 22

2.4.4 Otros ámbitos de aplicación ..... 22

**3 CLUSTERING JERÁRQUICO** ..... **23**

**3.1 Introducción** ..... **23**

**3.2 Clustering y detección de anomalías** ..... **24**

**3.3 Clustering jerárquico** ..... **24**

**3.4 Medidas de la distancia en Métodos Algorítmicos** ..... **26**

**3.5 Ventajas e Inconvenientes del clustering jerárquico** ..... **28**

**3.6 Fases del proceso de Clustering** ..... **29**

**3.7 Implementación en R** ..... **31**

3.7.1 Ejemplo ..... 31

**4 DENDROGRAMAS ..... 33**
**4.1 Definición ..... 33**
**4.2 ¿Cómo se interpreta un Dendrograma? ..... 33**
4.2.1 Análisis e Interpretación de un dendrograma ..... 34
4.2.2 Disposición de los elementos ..... 34
4.2.3 Ejemplo de Análisis ..... 35
**5 DTW ..... 36**
**5.1 Introducción ..... 36**
**5.2 Ventajas del algoritmo DTW ..... 36**
**5.3 DTW vs Distancia Euclídea ..... 37**
**5.4 Ejemplo ..... 39**
**5.5 Medida de similitud DTW ..... 40**
**5.6 Cálculo de Trayectorias ..... 41**
**5.7 Complejidad del algoritmo ..... 43**
**5.8 Algoritmos de aproximación del método DTW ..... 43**
**5.9 Restricciones en las trayectorias ..... 44**
**6 CASO PRÁCTICO: MEDIDAS DE SENSORES ..... 45**
**6.1 Introducción ..... 45**
**6.2 Hipótesis de Trabajo ..... 46**
6.2.1 Fases del proceso ..... 46
6.2.2 Análisis de ficheros ..... 47
6.2.3 Formato de las columnas y mapas del edificio ..... 48
**6.3 Análisis ..... 50**
6.3.1 Ejemplo: concentración Hazium ..... 51
6.3.2 Hazium ..... 58
6.3.3 Temperatura ..... 61
6.3.4 Concentración $CO_2$ ..... 62
6.3.5 Demanda eléctrica total ..... 64
6.3.6 Patrón de comportamiento similar ..... 65
**7 EVALUACIÓN DE LA METODOLOGÍA ..... 66**
**7.1 Análisis de varias variables en un mismo período de tiempo ..... 66**
**7.2 Análisis del mismo sensor a lo largo de los días analizados ..... 67**
**7.3 Análisis de otros conjuntos de datos ..... 69**
**8 DISCUSIÓN Y CONCLUSIONES ..... 71**

**8.1 Discusión ..... 71**
**8.2 Conclusiones ..... 72**
**8.3 Trabajo futuro ..... 73**
**9 ANEXO ..... 74**
**9.1 Nombre de Columnas ..... 74**
**9.2 Descripción de Columnas ..... 75**
**9.3 Comandos en R ..... 77**
**9.4 Tipos de Gráficas ..... 79**
**10 BIBLIOGRAFÍA ..... 80**

# Resumen

En estas últimas décadas, el concepto de ***Data Mining***[1] ha adquirido un papel determinante, independientemente del sector al que pertenezca una empresa, dado que, mediante el análisis y tratamiento de los datos que éstas recaban de forma masiva a diario, se logra transformar un "simple" conjunto de datos recolectados por sus Sistemas de Información, en puro conocimiento.

Por poner algunas cifras reales que nos facilite la comprensión del avance acontecido en este aspecto, desde 2014 los entornos empresariales han ido incrementando la inversión en Big Data y Analytics de manera sustancial, de tal forma que según el informe publicado por *IDG* [2], un 83% de las grandes empresas y un hasta un 63% de las pymes han invertido o tienen planificada una estrategia de inversión en este tipo de iniciativas.

Dicho esto, discernir entre los datos que tienen un comportamiento considerado como "normal" y aquellos que se presentan como atípicos, puede resultar interesante, e incluso crucial, dependiendo del tipo de datos que estemos analizando. En las técnicas de clustering, la detección de anomalías surge de forma bastante natural, debido fundamentalmente a la agrupación en clústeres que se genera, siempre basada en su similitud (distancia), razón por la que se ha optado por utilizarlo a lo largo del trabajo presentado en las siguientes páginas.

**Nota** - Los ficheros y scripts generados necesarios para la construcción de este documento se han subido a la URL pública: https://github.com/agsf111/TFM.git

1 En español, se traduce como Minería de datos, que se define de forma genérica como la exploración y tratamiento de los datos, una de las etapas de análisis previo que forman parte del KDD (Knowledge Discovery in Databases); es un campo de la estadística y las ciencias de la computación cuyo objetivo es intentar descubrir patrones en grandes volúmenes de conjuntos de datos.

2 El Informe referido de *IDG* se puede consultar en la URL:

*https://www.idgenterprise.com/resource/research/2015-big-data-and-analytics-survey/*

# Abstract

In the last decades, Data mining concept has acquired a decisive role, independently of the sector to which a company belongs, since, through the analysis and treatment of the data that they collect on a massive daily basis, it is possible to transform a "Simple" set of data collected by their Information Systems, in pure knowledge.

To put some real figures to facilitate the understanding of the progress made in this regard, since 2014 business environments have been increasing investment in Big Data and Analytics in a substantial way, so that according to the report published by IDG 2, an 83 % of large companies and up to 63% of SMEs have invested or have planned an investment strategy in this type of initiatives.

That said, discerning between data that have behavior considered "normal" and those that are presented as atypical, may be interesting, and even crucial, depending on the type of data we are analyzing. In Clustering techniques, the detection of Anomalies arises quite naturally, due mainly to clustering in Clusters that is generated, always based on their similarity (distance), reason why it has been chosen to use it throughout the work Presented in the following pages.

---

1 Data mining is defined as the exploration and treatment of the data, one of the stages of previous analysis that are part of the Knowledge Discovery in Databases (KDD); Is a field of statistics and computer science whose aim is to try to discover patterns in large volumes of data sets.

2 The referred Report of IDG can be consulted in the URL:

*https://www.idgenterprise.com/resource/research/2015-big-data-and-analytics-survey/*

# 1. Introducción

Hoy en día, el volumen de datos recabados por los sistemas informatizados es una fuente de información tan amplia que, realmente, no somos capaces de tratarla de un modo eficiente. Dicha información puede ser muy valiosa para las empresas en cualquier escenario imaginable, mayor incluso de la que se pueda suponer a priori.

Dicho esto, se plantea el reto de cómo obtener, no una colección de datos, sino información y conocimiento a partir de ellos, para poder aplicar el procedimiento definido que corresponda en cada caso.

### *Datos de origen*

Todos los años se celebra una competición a escala mundial creado por la asociación *VAST* (Visual Analytics Science and Technology), cuyo objetivo es fundamentalmente el avance en el campo del Análisis Visual.

Para ello, se crean varios retos (VAST Challenges) destinados a que los científicos de datos comprendan a través de un caso práctico, cómo la transformación de los datos, así como su visualización e interacción puede transformar un desarrollo de software en una herramienta potente capaz de realizar ciertas tareas con esa información procesada (benchmarks), e incluso, llegar a suponer un avance real en la resolución de problemas más complejos.

Los datos analizados en este documento pertenecen a uno de estos retos, más concretamente a la competición del año 2016. Veamos en qué consiste para tener una visión global del problema planteado:

Una empresa llamada GAStech (ficticia), a raíz de un problema de seguridad de los empleados acontecido en 2014, en el que hubo un caso de secuestro de empleados, decide mejorar las instalaciones de sus oficinas; decide construir un nuevo edificio de tres plantas cerca de su localización anterior que cumple con los estándares más estrictos en materia de eficiencia energética y ha sido equipado con todo tipo de sensores que miden distintos parámetros del propio edificio, como pueden ser la temperatura o la concentración por zonas de distintas sustancias químicas. A esto, se suma que los empleados tienen que llevar tarjetas de proximidad, con el fin de evitar otro caso como el de 2014.

El reto plantea que, como expertos en *Visual Analytics*, se lleve a cabo un análisis exhaustivo de los datos recolectados por estos sensores en un período de dos semanas, entre los días 31 de mayo y 13 de Junio, con el fin de detectar patrones de comportamiento que posibilite la identificación de posibles fallos operacionales en el edificio, así como problemas de seguridad que pongan en riesgo al personal que trabaja en él.

Para más información, se puede consultar *VAST Challenge 2016: Mini-Challenge 2*.

## 1.1 Motivación

Los valores presentados por los sensores, ubicados en zonas estratégicas a lo largo de las tres plantas del edificio, una vez son tratados, analizados e interpretados, nos pueden aportar ciertas pistas sobre si el funcionamiento es el esperado o no, lo que se traduce en la detección de posibles anomalías (outliers), datos atípicos que pueden influir en fallos de tipo operativo o incluso, en fallos de seguridad en relación con el personal que trabaja en él.

En este proyecto proponemos la combinación de técnicas de *Clustering Jerárquico* basado en la distancia *DTW* con otras técnicas de análisis visual.

Esto haría factible esa transformación de datos en conocimiento a la que hacía referencia anteriormente, uno de los mayores valores con que una empresa puede contar hoy en día.

La solución presentada en este documento se ha desarrollado para unos datos específicos, pero se podría generalizar a otros conjuntos de datos similares, por lo que se puede convertir en una metodología general.

La dificultad de este proyecto radica en el volumen de datos analizados, lo que se traduce en que detectar anomalías no siempre es algo trivial o sencillo. La razón de decantarnos por técnicas como el clustering ha sido precisamente su capacidad para resumir esos datos.

La motivación principal de este trabajo ha sido en todo momento crear una metodología con la que poder detectar anomalías en series temporales multivariantes, lo que nos permite, una vez aislados, adquirir un conocimiento sobre los mismos, con el que se contribuirá a la toma de decisiones.

## 1.2 Planteamiento de trabajo

En todo momento, el planteamiento del trabajo ha sido crear una metodología que permita aplicar una técnica de análisis de datos, con el fin de transformar la simple recogida de datos en información valiosa. Como es lógico, dicha metodología añade valor, dado que podría ser aplicada en otros escenarios diferentes.

El tratamiento de los datos se ha llevado a cabo en un entorno de programación R. Se ha elegido este lenguaje por su enfoque hacia el análisis estadístico, dado que permite cargar diferentes paquetes con funcionalidades de cálculo y gráficas. Mediante el uso de librerías específicas de Data Mining (minería de datos), junto con técnicas de Clustering Jerárquico y de visualización a través de dendrogramas, nos ha permitido extraer y analizar dicha información.

## 1.3 Estructura de capítulos

Los capítulos que componen el presente documento explican de forma detallada los conceptos en los que se ha basado la Metodología definida en el trabajo, cuyo objetivo es permitir la detección de anomalías en los datos analizados.

En el capítulo 2 se desarrolla el concepto de detección de anomalías, qué tipos existen, y se presentan algunos ejemplos significativos de cómo se aplican a diferentes sectores. En el *capítulo 3* se explica el método de *Clustering Jerárquico* que permite la detección de grupos de elementos que sean similares (clústeres). En el *capítulo 4* se describe la lectura e interpretación de los *dendrogramas*, un tipo de representación gráfica que nos facilita la interpretación de los grupos, diferenciando aquellos datos que presentan características similares. En el *capítulo 5* se analizan las características principales del algoritmo *DTW*, utilizado como medida de distancia entre elementos. Todas estas técnicas se han aplicado en un entorno de *programación en R*.

En el *capítulo 6* se aplican los conceptos presentados en el *reto* que nos proponen solucionar (VAST Challenge). La *evaluación de la metodología* propuesta a tal efecto se describe a lo largo del *capítulo 7*. Por último, en el *capítulo 8* se exponen las *conclusiones* que se derivan del estudio realizado.

# 2 Estado del Arte

## 2.1 Situación actual

En las últimas décadas se ha experimentado un cambio drástico en el volumen de datos que se generan a diario en los Sistemas de Información, registrando todo tipo de conceptos. Tanto el hardware como el software de gestión y control han evolucionado de una forma exponencial desde los años 80, y con ello, se ha alcanzado un procesamiento lo más heterogéneo posible en lo que se refiere a orígenes, formatos o tratamiento de esos datos.


**Figura 1. Pirámide del conocimiento**

Con independencia del sector al que nos refiramos, este volumen escapa con creces la capacidad que se tiene en la actualidad de recolectar, almacenar, y sobre todo, de comprender esos datos, lo que ha hecho que el *Data Mining* (minería de datos) se haya posicionado como una solución factible para la cada vez más necesaria transformación de datos en conocimiento, cuya pirámide se representa en la Figura 1.

Se debe buscar cómo automatizar este proceso de transformación, para de esta forma, en un tiempo aceptable, adquirir un conocimiento que se pueda aplicar en la toma de decisiones. Es en este escenario, donde surge el concepto de *BI* (Business Intelligence o Inteligencia de Negocios).

Como es sabido, los datos almacenados de forma individual no se pueden considerar información; Los datos sólo toman relevancia, y se pueden convertir en conocimiento, cuando son interpretados en un contexto determinado, obteniéndose tras un análisis detallado unos patrones "de comportamiento" (modelos) que sean fiables desde el punto de vista estadístico. Estos patrones detectados pueden no ser obvios a priori, y pueden ser totalmente desconocidos antes de llevar a cabo dicho análisis.

Es aquí donde toma enorme importancia el concepto de detección de anomalías, que en el contexto del Data Mining, se puede definir como el proceso de identificación de aquellos datos, eventos, elementos de un grupo u observaciones que no se ajustan a un patrón

esperado, es decir, datos que presentan un "comportamiento" diferente a lo que se considera "normal" dentro de un conjunto de datos [3]

Por ejemplo, la detección de fraude se basa en poder crear unos patrones que permitan definir, mediante perfiles, el comportamiento que se considera "normal" por parte de los usuarios, con el fin de detectar como anomalía todo lo que no pertenezca a dicho patrón, es decir, aquel comportamiento que se pueda considerar atípico.

## 2.2 ¿Qué son las anomalías?

El concepto de ***detección de anomalías*** se refiere a la técnica mediante la cual, se busca dentro de un conjunto de datos, aquellos patrones de comportamiento que no se ajustan al comportamiento esperado. Normalmente nos referimos a estos datos, por citar sólo algunos, como datos atípicos, outliers, excepciones, anomalías o particularidades. De todos estos conceptos, en escenarios de Data Analysis, normalmente se puede intercambiar el uso que se hace entre conceptos como outliers y anomalías.

Desde el siglo XIX, la rama de la Estadística se ha interesado por el estudio detallado del análisis estadístico de detección de anomalías (*estudio de Glaisher*, 1872), lo que ha propiciado que se hayan desarrollado diferentes técnicas. Algunas, se han desarrollado de forma explícita para un campo de aplicación específico.

Implica diferenciar entre anomalía y ruido, de tal forma, que este segundo concepto hace referencia a todo dato que no es interesante para en analista, y que, por tanto, debe tratarse de forma previa al análisis, ignorándolos, o incluso, eliminándolos del conjunto de datos de entrada que sea objeto de estudio.

El interés real que tiene la detección de anomalías reside en el hecho de que se pasa de un análisis de los datos teórico a una interpretación real, con un significado a veces crítico que se puede aplicar en un amplio rango de aplicaciones prácticas. Dependiendo de qué significado tome este análisis, podremos llevar a cabo una toma de decisiones en los diferentes contextos que se estén tratando.

[3] Página web **http://copro.com.ar/Deteccion_de_anomalias.html**

## 2.3 Taxonomía de las anomalías

Podemos suponer a priori que en líneas generales, puede resultar bastante trivial diferenciar una región de comportamiento "normal" de una que no lo sea, pero la realidad es muy diferente. Existen diversos factores que convierte casi en un reto diferenciar la frontera entre ambos conceptos: [4]

- Cuando las anomalías surgen como resultado de acciones maliciosas, se ven enmascaradas porque se intenta hacer que aparezcan como datos normales. Esto implica una dificultad añadida a cómo definir qué se considera comportamiento normal.

- En muchos dominios, lo que se considera comportamiento normal puede no ser suficientemente representativo en el futuro.

- Dependiendo del campo de aplicación que se analice, un comportamiento considerado como normal en uno de ellos, puede tratarse como anomalía en otros. Esto implica que la técnica desarrollada para la detección de anomalías, se presente como eficaz o no, dependiendo del entorno analizado.

  Por ejemplo, por el riesgo que puede conllevar, una pequeña desviación de los datos en el campo de la medicina puede tratarse como anomalía, mientras que en un sector de mercados de stock (Retail), esa misma desviación podría considerarse dentro de la normalidad.

- Como es obvio, puede ser un problema utilizar como datos de entrenamiento, aquellos datos que sean anomalías y que no sean tratados como tal.

- A veces, es difícil distinguir entre datos anómalos y aquellos tratados como ruido. Esto implica una dificultad añadida a la hora de su tratamiento y borrado.

[4] Página web
http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf [1.2]

La resolución del problema de detección de anomalías presenta diferentes formulaciones, que son específicas para cada caso. Depende de diferentes factores, como pueden ser la naturaleza de los datos de entrada, la capacidad de ser categorizados o las condiciones derivadas del entorno en que se aplica el análisis. En las siguientes secciones detallaremos estos factores.

### 2.3.1 Naturaleza de los datos de entrada

Es uno de los aspectos claves en las técnicas de detección de anomalías. Generalmente, los datos de entrada consisten en una colección de objetos (instancias), donde cada instancia se puede describir con una serie de atributos (variables), que a su vez, pueden ser de diferentes tipos: binarios, categóricos o continuos. Cada instancia puede estar formada por un solo atributo (univariante) o por varios (multivariante), ya sean éstos, del mismo o de diferente tipo. La naturaleza de estos atributos va a determinar cómo se van a aplicar las diferentes técnicas de detección. Por ejemplo, hay técnicas de clasificación o basadas en parámetros estadísticos que no son aplicables en todos los casos.

Los datos de entrada también se pueden clasificar de acuerdo a la relación existente entre las instancias. Por defecto, la mayoría de las técnicas de detección de anomalías aceptan como hipótesis inicial que no existe dicha relación, pero en otros casos, como por ejemplo en series temporales o en secuencias del genoma, sí puede tener mucha relevancia.

### 2.3.2 Tipos de outliers

Atendiendo a su composición y su relación con el resto de los datos, los outliers pueden pertenecer a uno de estos tres grandes grupos:

- **Outlier Puntual**

Si se tiene un conjunto de datos, como se muestra en la Figura 2, que de acuerdo a unas características determinadas, se pueden observar grupos homogéneos (áreas $N_1$ y $N_2$), se considera anomalías o outliers a todos los elementos que se salgan de esas zonas (normales), como son en este caso los puntos $O_1$, $O_2$ y $O_3$.

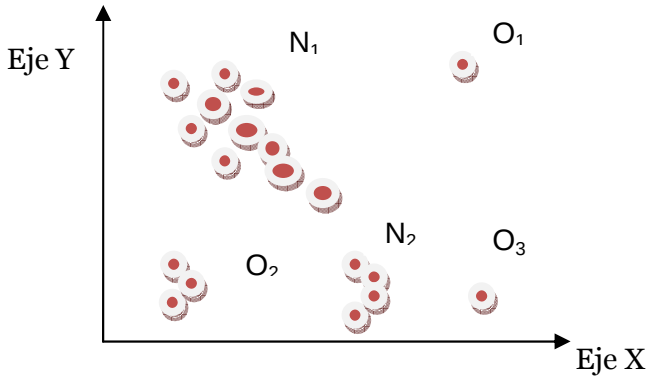



**Figura 2. Outlier puntual**

Si aplicamos la detección de anomalías (outliers) en un caso real, como pueden ser las *transacciones bancarias*, como es lógico, va a depender de la variable que se tome como referencia para el conjunto de datos.

Pongamos por caso que tomamos el dinero gastado por una persona como variable a analizar; detectaremos como anomalía (posibilidad de fraude), cualquier transacción que proceda de esa misma persona y muestre un dinero gastado que sea muy elevado comparado a lo que gasta de forma habitual, movimientos estos últimos, que serán considerados dentro de comportamiento "normal".

- **Outlier Contextual**

Este tipo de outlier sólo se comporta como un dato anómalo en un contexto determinado. Son casos individuales de los datos, pero a diferencia del caso anterior, pueden considerarse datos normales en un contexto diferente.

Volviendo al caso de la detección de fraude bancario, podríamos poner un ejemplo en el análisis del uso de las tarjetas de crédito: Si consideramos como comportamiento normal el gasto medio de una persona en una gasolinera de 50 euros, y un gasto superior, digamos 300 euros, en una joyería, ambos forman parte de los datos que se toman como normales en un contexto determinado.

Ahora bien, ¿Qué sucede si esa misma persona realiza un pago con la tarjeta de 300 euros en la gasolinera?... Eso ya formaría parte de un gasto en un contexto que, aunque se trata de la misma cantidad (300 euros), es muy superior al que gasta normalmente en las gasolineras, por lo que se podría definir como outlier.

- **Outlier Colectivo**

Se definen como outliers sólo cuando se observan los datos en forma global, mientras que de forma separada no se pueden considerar datos anómalos por sí mismos.

Como ejemplo, se puede observar un resultado de un electrocardiograma (Figura 3), donde el valor constante por sí mismo no sería outlier si no se analiza de forma conjunta.

**Figura 3. Electrocardiograma**

### 2.3.3 Etiquetas en los datos

Las etiquetas asociadas a las instancias de datos denotan si un dato se considera normal o, por el contrario, puede presentar un comportamiento anómalo. El problema, radica en que tener todos los datos "etiquetados" es un proceso que resulta muy costoso, ya que es totalmente manual y tiene que ser realizado por un experto, lo que conlleva un esfuerzo muy significativo.

Normalmente es más difícil tener categorizados todos los posibles comportamientos anómalos existentes, frente a los comportamientos normales se trata de un proceso dinámico, por lo que podría darse la casuística de generarse nuevos datos anómalos, que no estarían etiquetados.

Hay que pensar que tiene un impacto en aplicaciones reales: por ejemplo, en el ámbito del tráfico aéreo y su seguridad, una anomalía podría desembocar en una tragedia a gran escala, de ahí que se consideren eventos raros.

Dependiendo de a qué tipo de datos le aplican las etiquetas, las técnicas de detección de anomalías (*Data mining concepts and techniques,* 2012) pueden operar en uno de los siguientes modos:

- **Detección de anomalía Supervisada**

  Se asume que se han podido catalogar todas las instancias (de entrenamiento), tanto las clases normales como las anómalas.

  Una forma de abordar este tipo de escenario es llevando a cabo un modelo predictivo para instancias tipo normal vs anómalo. En caso de encontrar un objeto no categorizado, se compara con el modelo y se determina a qué grupo pertenece.

  Este modelo de detección presenta dos problemas, fundamentalmente: por un lado, las anomalías, normalmente, se dan en mucho menor número que las instancias catalogadas como normales; Y por otro lado, encontrar una etiqueta que sea representativa, en especial para las anomalías, puede suponer un auténtico reto.

- **Detección de anomalía No supervisada**

  Este modelo no requiere datos de entrenamiento y se asume que las instancias normales son más numerosas que las anomalías en los datos de test. Es el caso más común.

  La metodología que se presenta en este trabajo de TFM pertenece a esta categoría.

- **Detección de anomalía Semisupervisada**

  En las técnicas que operan en este modo, se asume que sólo se han etiquetado los datos para las clases normales. La típica solución que se aplica en estos casos, suele ser construir un modelo para las clases que se correspondan con un comportamiento normal y usar dicho modelo para identificar anomalías en los datos de prueba.

  Los modelos semisupervisados se pueden adaptar a utilizarse como modelos no supervisados utilizando como datos de entrenamiento, datos no etiquetados. En este caso, se asume que en los datos de prueba existen pocas anomalías, tratándose de técnicas muy robustas.

### 2.3.4 Presentación de anomalías

Uno de los aspectos que son importantes en las técnicas de detección de anomalías es la forma en la que se informa. Pueden ser de dos tipos:

- *Score* (Puntuación) - Mediante técnicas de puntuación, se asigna un valor a cada instancia que irá en función del grado en el que se considera anomalía. Un analista de datos puede entonces analizar, o bien, aquellos pocos valores que sean los "top", o bien, analizar sólo los que sobrepasen un umbral definido.

- *Labels* (Etiquetas) – Se asigna mediante diferentes técnicas una categorización; cada instancia de prueba es etiquetada como dato normal o anómalo. Actúa de modo binario, por lo que no existe la posibilidad de tratar sólo un subconjunto de datos, que sobrepasen cierto umbral, por ejemplo, como se vio en el anterior caso (podría hacerse algo similar basado en parámetros).

## 2.4 Aplicaciones

Las técnicas de detección de anomalías se aplican hoy en día en una amplia variedad de aplicaciones prácticas, ya sea en la fase de pre-procesamiento de datos o en la fase de análisis final, aplicado a todos los sectores de negocio.

Algunas aplicaciones de técnicas de detección de anomalías son:

### 2.4.1 Intrusión en sistemas informáticos

Este tipo de detección es interesante desde el punto de vista de seguridad en los sistemas informáticos. Es un comportamiento que difiere de lo que se considera normal, por lo que debe ser tratado como una amenaza, y más cuando se tiene en cuenta el volumen de datos que normalmente se maneja en este tipo de escenario, donde por regla general, juega un papel fundamental el streaming, por lo que se requiere un análisis online de los datos.

Denning en 1987 realiza por primera vez una clasificación de este tipo de detección de intrusos en dos grandes categorías: [5]

- *Host-based*

  Se realizan llamadas al sistema operativo para propagar el comportamiento anómalo, ya sea un programa intrusivo (normalmente malicioso), un cambio de comportamiento no autorizado, o bien, cualquier tipo de violación de políticas del sistema.

- *Network-based*

  Se realizan intrusiones a nivel de red. Normalmente se contemplan en este grupo todo tipo de acciones llevadas a cabo por hackers, cuyo objetivo es entrar en una red de forma no autorizada para así, poder extraer información restringida.

[5] Página web
http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf

### 2.4.2 Fraude

Se entiende por detección de fraude a la detección de un comportamiento criminal por el que se ven afectadas entidades comerciales tales como bancos, agencias de seguros, operadoras telefónicas, etc.

En este caso, el usuario que realiza dicha acción puede ser un cliente real de la propia organización, o bien, no serlo, en cuyo caso se habla de suplantación de identidad. El fraude ocurre cuando este tipo de usuario hace uso de los recursos de la organización de una manera que no ha sido nunca autorizada.

La técnica de detección de este tipo de anomalía la plantearon en 1999 (Fawcett y Provost), definiendo lo que se conoce como Activity Monitoring, una metodología por la cual la mejor forma de abordar este problema es definir un perfil por usuario y monitorizar todos esos perfiles, de manera que todo lo que difiera de un patrón de comportamiento normal, pase a considerarse una anomalía.

Dado el impacto económico que puede tener, con pérdidas muy sustanciales, el objetivo principal de las organizaciones es detectar de forma inmediata este tipo de anomalía (fraude).

Algunos ejemplos pertenecientes a esta categoría pueden ser fraudes de diferentes tipos:

- Relacionado con el uso de tarjetas de crédito.

- Las aseguradoras, donde una de las que mayor impacto económico tiene es el sector automovilístico

- En el mercado de valores, el fenómeno que se conoce como *Insider Trading*, donde se realizan operaciones en base a informaciones confidenciales adquiridas de forma previa a hacerse públicas, con la ventaja obvia que conlleva. Existen diferentes tipos de información.

### 2.4.3 Sector Industrial

El continuo uso de maquinaria en el sector industrial, hace que se pueda ver afectada o dañada cierta maquinaria e instrumentación. Se conoce como datos de sensores, ya que se lleva a cabo medidas continuas por parte de sensores que están monitorizados tipo 24x7. Dado su impacto económico, tiene que detectarse con suficiente antelación para no sufrir pérdidas en la medida de lo posible.

Se detectan dos tipos de anomalías en esta categoría: Unidades mecánicas (motores, turbinas, etc.) y defectos estructurales (tensión en el fuselaje de un avión o grietas en una viga).

### 2.4.4 Otros ámbitos de aplicación

Existen otros escenarios en los que se utiliza en la actualidad la técnica de detección de anomalías, entre los que, por citar sólo unos ejemplos, encontramos: en el ámbito económico, el fraude en subastas; en ámbitos más tecnológicos, se emplea en la detección de intrusión en redes sociales, procesamiento de imágenes o de texto, o comportamientos programados de robots; Por último, en un ámbito más científico, se aplica en la detección de células cancerígenas, anomalías cromosómicas o en la detección de enfermedades congénitas. Comentar al respecto, que el ámbito de la sanidad y salud pública es un área crítica, dado que normalmente, se ven involucrados ciertos datos de los pacientes, tratados como resultados definitivos.

# 3 Clustering jerárquico

## 3.1 Introducción

La técnica de *Clustering* es una técnica de análisis de datos que permite la agrupación de un conjunto de datos (dataset) en diferentes subconjuntos, identificados como ***Clústeres***.

La característica principal de esta técnica reside en que los elementos que forman parte de un mismo clúster son similares entre sí, pero diferentes a aquellos elementos que forman parte de otro clúster. El conjunto de los diferentes clústeres que surgen como resultado de dicho análisis, es lo que se conoce como ***Clustering***.

La técnica de Clustering ha encontrado su uso en numerosas aplicaciones prácticas, entre las que se encuentran disciplinas relacionadas con el *Marketing* o el *Business Intelligence (BI)*; dentro del campo científico, se aplica de forma frecuente en *medicina* y *biología*. Igualmente, tiene un uso muy extendido en *patrones de búsqueda en la web* o en *técnicas de reconocimiento*, ya sea de patrones de *imagen* o de *voz*.

Dado que el agrupamiento en realidad lo lleva a cabo el algoritmo que se esté utilizando, nada impide que se puedan generar diferentes agrupaciones a partir de un mismo conjunto de datos de entrada; este tipo de técnicas entran en la categoría de ***aprendizaje no supervisado*** [6], si nos basamos en el hecho de que permite descubrir grupos o asociaciones totalmente desconocidas en fases previas al análisis.[7]

[6] **Aprendizaje no supervisado** es un método de aprendizaje automático donde un modelo es ajustado a las observaciones. Se distingue del aprendizaje supervisado por el hecho de que no hay un conocimiento a priori. En el aprendizaje no supervisado, un conjunto de datos de objetos de entrada es tratado. Así, el aprendizaje no supervisado típicamente trata los objetos de entrada como un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos.
*Wikipedia.- Recuperado el 12 de Julio, 2017 - https://es.wikipedia.org/wiki/Aprendizaje_no_supervisado*

[7] Dicha técnica recibe varios nombres: en ciertos entornos, se denomina ***Clasificación automática***, dado que un conjunto de datos que presenta estas características - iguales entre sí y diferentes a los datos de otro clúster- se puede entender como una *clase*. En otros escenarios, se denomina ***Segmentación***, en base a cómo, a partir de un conjunto grande de datos, se producen particiones de acuerdo a su similitud.

La metodología desarrollada en el presente documento, creada con la finalidad de permitir la detección de anomalías, se basa en el uso de la técnica de ***clustering jerárquico*** (aprendizaje no supervisado), por lo que será la técnica que se explique más detalladamente a lo largo de nuestro análisis en las correspondientes secciones.

## 3.2 Clustering y detección de anomalías

En los últimos años, la técnica de Clustering, unida a la detección de anomalías, se ha convertido en una herramienta muy potente a la hora de detectar grupos con comportamiento diferentes al resto (outliers)[8]. Dicha unión de técnicas, posibilita esa transformación de datos en conocimiento a la que hacía referencia anteriormente.

La técnica de Clustering se basa en la medida de similitud entre valores de ciertos atributos que presentan los datos analizados. En líneas generales, tiene por objetivo maximizar la similitud entre las instancias que forman cada clúster, a la vez que intenta minimizar la similitud entre clusters. Se ordenan los objetos de acuerdo a diferentes niveles de similitud.

Ese tipo de agrupamiento hace posible su uso como técnica de detección de anomalías, dado que permite la identificación de los clústeres que presentan menos similitud, y con ello, analizar si se trata o no de una anomalía.

## 3.3 Clustering jerárquico

Un método de Clustering jerárquico representa los datos agrupados en forma jerarquizada, también denominado "árbol" de clústeres, lo cual es muy útil a la hora de obtener una representación de la evolución de los clústeres, lo que facilita la visualización de esos grupos.

Es un método que, por su definición, es muy sensible a la presencia de outliers (anomalías), lo que se traduce en que las conclusiones que se obtienen tras analizar los clústeres generados se deben de estudiar de forma detallada y contextualizada.

[8] http://ceur-ws.org/Vol-558/Art_8.pdf

Aunque la técnica de Clustering jerárquico es una técnica de agrupación y, como tal, no se ha creado como técnica específica de detección de anomalías, la metodología que se desarrolla a lo largo de este trabajo se basa en transformar su uso con el fin de poder encontrar datos que presenten un comportamiento diferente, analizados posteriormente como posibles anomalías.

El método de Clustering Jerárquico se clasifica en dos grandes categorías: ***aglomerativo*** y ***divisivo***.

En el método de agrupamiento jerárquico aglomerativo, utiliza una estrategia "bottom-up", es decir, la descomposición jerárquica de elementos se lleva a cabo de una forma ascendente (fusión), de abajo a arriba.

Típicamente, cada elemento forma su propio clúster, con el objetivo final de agrupar de forma iterativa los elementos en clusters que sean más grandes cada vez, hasta que, o bien se cumpla una determinada condición, o bien, todos los objetos formen parte de un único clúster, en cuyo caso, se define como clúster raíz de la jerarquía obtenida.

La fase de fusión de los clústeres se basa en una o varias medidas de similitud para encontrar los dos clústeres más próximos entre sí, de manera que se fusionen para formar un nuevo clúster único. Este proceso combina los clústeres por iteración, por lo que, dado que cada clúster tiene por lo menos un elemento, es un método que requiere como mínimo N iteraciones.

El método divisivo, al contrario que en el anterior caso, emplea una estrategia Top-down, es decir de arriba hacia abajo, descendente: en un primer momento, todos los objetos pertenecen a un único clúster (raíz), y se va dividiendo en sucesivos clústeres cada vez más pequeños de forma recursiva, continuando el proceso hasta que, o bien, se cumple una condición determinada, o bien, que cada clúster esté formado por un único elemento o que todos los elementos que forman parte de cada clúster sean lo suficientemente similares entre sí.

En la técnica de clustering jerárquico -aglomerativo o divisivo- el usuario siempre debe definir el umbral de similitud, el cual, define el número de clústeres que se forman.

## 3.4 Medidas de la distancia en Métodos Algorítmicos

Independientemente del método algorítmico por el que se opte, sea aglomerado o divisivo, el principal problema a resolver es calcular la distancia entre dos grupos cualesquiera, dado que la agrupación, como hemos visto, se lleva a cabo en base a esa distancia. Así pues, la siguiente pregunta que surge es: ¿Cómo calculamos la distancia entre dos clusters?

En líneas generales, se definen 4 tipos de distancias (***Linkage measures***):

- **Maximum distance**: $dist_{max}(C_i, C_j) = \max(|p - p'|)$
- **Minimum distance**: $dist_{min}(C_i, C_j) = \min(|p - p'|)$
- **Average distance**: $dist_{avg}(C_i, C_j) = 1/n_i n_j \sum |p - p'|$
- **Mean distance**: $dist_{mean}(C_i, C_j/) = |m_i - m_j|$

En estas fórmulas, se asumen como condiciones, que $|p - p'|$ es la distancia entre dos objetos o puntos p y $p_o$, $m_i$ es la media del clúster $C_i$ y $n_i$ es el número de objetos que pertenecen a dicho clúster (se trata de igual forma para el clúster $C_j$ con $n_j$ elementos).

Atendiendo al uso de un tipo de medida u otro, tal y como se representa en la Figura 4, podemos clasificar los algoritmos de Clustering en:

- **Nearest-neighbor:** En los casos en los que el algoritmo usa la medida de distancia mínima $-d_{min}(Ci, Cj)$ – para calcular la distancia entre clústeres.

  Si el proceso finaliza cuando la distancia entre los clústeres más cercanos supera un umbral definido por el usuario, se denomina algoritmo **Single-link** o de enlace único. Esta técnica recibe el nombre de MIN.

- **Farthest-neighbor** - En los casos en los que el algoritmo usa como medida de distancia entre clusters la distancia máxima, $d_{max}(Ci, Cj)$.

  A diferencia del caso anterior, si el proceso de clustering finaliza cuando la distancia máxima entre los clústeres más cercanos supera un valor definido, es lo que se conoce como algoritmo **Complete-link** o de **enlace completo**. Esta técnica recibe el nombre de *MAX*.

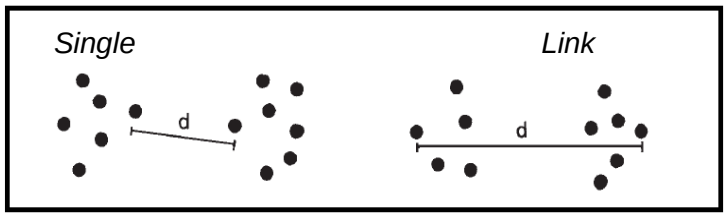



**Figura 4. Cálculo de las distancias entre dos clústeres**

Por otro lado, la **distancia media** $dis_{avg.}(C_i, C_j)$ se define como la media de las distancias entre todos los pares de elementos de los clústeres, tal y como se indica en la Figura 5.

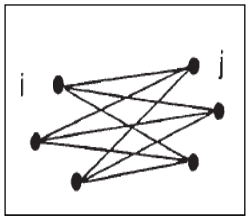



**Figura 5. Distancia media**

La **distancia mean** $dist_{mean}(C_i, C_j)$ se define a partir de la distancia media en cada clúster o centroide, tal y como se representa en la Figura 6.

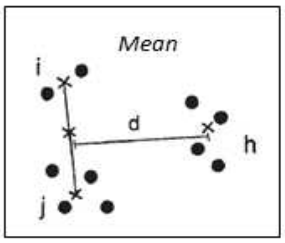



**Figura 6. Mean**

En líneas generales, la calidad de los clústeres que se van generando en las sucesivas iteraciones, es mayor, si son lo bastante compactos.

Se debe pensar en el hecho de que si un método de clustering presenta una baja calidad, se puede traducir en que los clusters generados no encuentren un sentido práctico, lo que provocaría llegar a conclusiones no del todo fiables o incluso falsas.

Con los algoritmos tipo *Nearest-neighbor*, se tiende normalmente a minimizar el diámetro de los clusters de elementos en cada iteración.

## 3.5 Ventajas e Inconvenientes del clustering jerárquico

Todo algoritmo tiene sus ventajas y sus desventajas respecto a otros; como se suele decir, no hay algoritmo perfecto para todos los ámbitos. Esto nos lleva a presentar las ventajas e inconvenientes más destacadas en el caso de Clustering jerárquico:

*Ventajas*

- **Es flexible**. Se puede utilizar cualquier medida de distancia como métrica.
- **Resultado independiente del número de clústeres.** No requiere realizar inferencias sobre el número de clústeres generados.
- **Dendrogramas.** Permite representar los sucesivos clústeres en forma de árbol.

Inconvenientes

- **Es sensible a anomalías (outliers)** -datos que presenten valores atípicos (ruido) en los casos en los que se decide realizar mediciones basadas en las distancias mínimas o máximas.

Una posible solución intermedia entre ambas distancias, es optar por el uso de la distancia media, método que no presenta dichos inconvenientes y que presenta como principales ventajas, que es muy fácil de calcular, además de permitir tratar datos, tanto de tipo categórico, como de tipo numérico. [9]

- ***Es un proceso que puede resultar muy costoso***.

Desde el punto de vista de los requisitos de computación, para asignar un clúster, se requiere recorrer toda la matriz de datos, calculando su similitud (distancia), lo que puede exigir un elevado consumo de computación.

[9] Los casos en los que se usa la media como medición de la distancia, para datos de tipo categóricos no siempre va a ser factible definir el vector medio.

No obstante, se debe de aclarar que, respecto al método aglomerativo, se acepta como hecho contrastado, que una de las mayores ventajas que presenta es su relativo bajo coste computacional frente a otros métodos jerárquicos, como por ejemplo, pueden ser los métodos divisivos, lo que en la práctica se traduce en una mayor velocidad de procesado, debido en gran parte a la menor complejidad que presenta.

- ***Los clústeres no siempre son fáciles de visualizar***.

En general, para un dataset grande, genera un árbol (dendrograma) que puede resultar muy complejo y difícil de visualizar, lo que, como es lógico no facilita la tarea de análisis posterior, pudiendo generar confusión.

- ***Es un modelo sensible a los outliers***.

El problema radica en el número de clusters identificados; para aquellos datos que presentan cierta similitud es fiable porque los agrupa en un clúster común, pero no se comporta del todo bien con datos que no tienen esa similitud, por lo que estos "outliers" los suele asociar en un clúster de forma aislada.

Por regla general, es un algoritmo que funciona mucho mejor cuando el dataset está formado por unos pocos cientos de datos por agrupar.

## 3.6 Fases del proceso de Clustering

De un modo general, el proceso de Clustering empieza definiendo un cluster por cada uno de los datos de entrada, y en las sucesivas etapas del algoritmo, los va agrupando, unificando los dos clusters que se consideran más cercanos, en un nuevo único clúster, basándose en esta distancia. De una forma más detallada, las distintas fases que forman el proceso completo de Clustering son:

- En una primera fase, se contempla el escenario inicial, en el que se definen aquellos parámetros de inicio que corresponden a los datos de entrada que queremos analizar (dataset).

Estos datos, lógicamente, están sin agrupar en ningún tipo de clúster. Si definimos L(k) como el nivel del clúster K y le asignamos una secuencia m al número de agrupaciones, la condición inicial sería:

Nivel $L(0) = 0$ *(Nivel inicial = 0)*
*Número de secuencia,* $m = 0$ *siendo* $m = 0, 1 \ldots (n-1)$

- Acto seguido, en la segunda fase, se realiza una búsqueda de qué pares de clústeres son más similares, aplicando el modo elegido para calcular la distancia. En nuestro caso, sería la distancia media (mean): dados 2 clústeres A y B, se define como:

$$d[A, B] = mean\ (\ d[(i),(j)]\ )$$

- Se incrementa el número de secuencia en una unidad (de m a *m +1), y unificamos esos dos clústeres en un único cluster nuevo*. Se forma el clustering *m y se define el nivel de clustering como:*

$$L(m) = d[(A), (B)]$$

- En la siguiente fase, se actualiza la matriz de similitud, lo que conlleva un borrado de las filas y columnas que corresponden a los clusters A y B, y añadiendo como filas y columnas el nuevo clúster generado en la anterior fase.

  Ahora la proximidad se definiría $d[(k), (A,B)] = mean\ d[(k),(A)], d[(k),(B)]$

- Llegados a la siguiente fase, se pueden dar varios escenarios:

  - Todos los objetos están agrupados en clúster único, en cuyo caso, se daría por finalizado el proceso de clustering.

  - Se ha alcanzado el umbral definido de número de clusters que queremos que se formen, en cuyo caso, al igual que en el caso anterior, daríamos por finalizado el proceso.

  - No se cumplen ninguno de los dos escenarios anteriores, en cuyo caso, repetiríamos los pasos anteriores desde la fase 2.

## 3.7 Implementación en R

En esta sección vamos a explicar cómo se realizaría el proceso de clustering jerárquico con el paquete estadístico R. Para su ilustración se utilizarán los datos del caso práctico utilizado en este TFM, en concreto, el código correspondiente a la concentración de Hazium.

Para el tratamiento de los datos recabados por los sensores temporales (fichero CSV), se ha optado por el uso de la medida de distancia ***DTW (Dynamic Time Warping).*** Más en concreto, se ha decidido utilizar el paquete del mismo nombre *-dtw-* que está totalmente integrado en el método ***DIST***. Como resultado de dicho algoritmo, se genera una matriz de similitud basada en distancias.

Por último, respecto al proceso de Clustering Jerárquico, su implementación en R se ha llevado a cabo mediante el uso del paquete ***hclust***, lo que unido a funciones de visualización de datos específicas como *ggplot2*, entre otras, ha permitido de una forma sencilla y práctica, la representación de los resultados en forma de dendrogramas.

### 3.7.1 Ejemplo

```
setwd=("<dir>")

install.packages("dtw")
library(dtw)

pr1 <- read.csv(file="<dir>/f1z8a-MC2.csv", header=T, sep = ",")
pr2 <- (pr1[,c(1,2)])
write.csv(pr2, file = "<dir>/pr1.csv", row.names=FALSE)

pr2$DateTime = as.POSIXct(pr2$Date.Time, format="%Y-%m-%d %H:%M:%S",tz="CET")

pr2 <- (pr2[,c(3,2)])

class(pr2[[1]])
class(pr2[[2]])

install.packages("reshape")
library(reshape)
```

```
pr2$hour =strftime(pr2$DateTime, format = "%H:%M")
pr2$DateTime<- as.numeric(format(pr2$DateTime, format="%d"))
pr2 <- cast(pr2,DateTime~hour, value='F_1_Z_8A..Hazium.Concentration')

res =dist(pr2, method="DTW")
clusters = hclust(res)
plot(clusters, main = "Clustering F_1_Z_8A - Hazium Concentration", xlab="Clustering Technique")
```

Como resultado se obtiene el siguiente dendrograma:

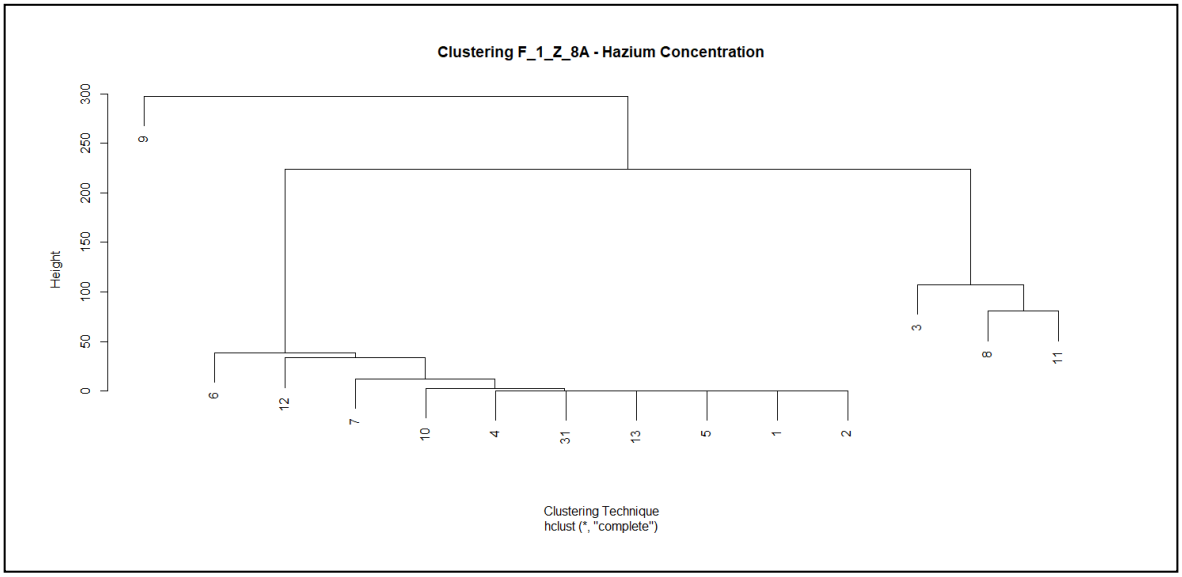



**Figura 7. Dendrograma concentración Hazium F_1_Z_8a**

# 4 Dendrogramas

## 4.1 Definición

Un *Dendrograma* es un diagrama tipo árbol que permite representar de forma gráfica la relación de grupos similares (clusters), derivados, en el ejemplo representado, de la aplicación del algoritmo de *Clustering jerárquico*.

## 4.2 ¿Cómo se interpreta un Dendrograma?

Para explicar cómo se realiza una lectura correcta de un dendrograma, en esta sección se utiliza un ejemplo extraído de los casos analizados; en concreto, se representan los grupos generados a partir de los valores medidos por los sensores ubicados en la tercera planta, en la zona 1 del edificio (F_3 – Z_1).

Estos datos corresponden a la potencia de luz consumida a lo largo de las dos semanas en las que se han recabado dichos datos, un indicador claro de la actividad en el edificio, a partir del cual, se pueden extraer diferentes conclusiones al respecto. (Para ver en más detalle, consultar capítulo de conclusiones)

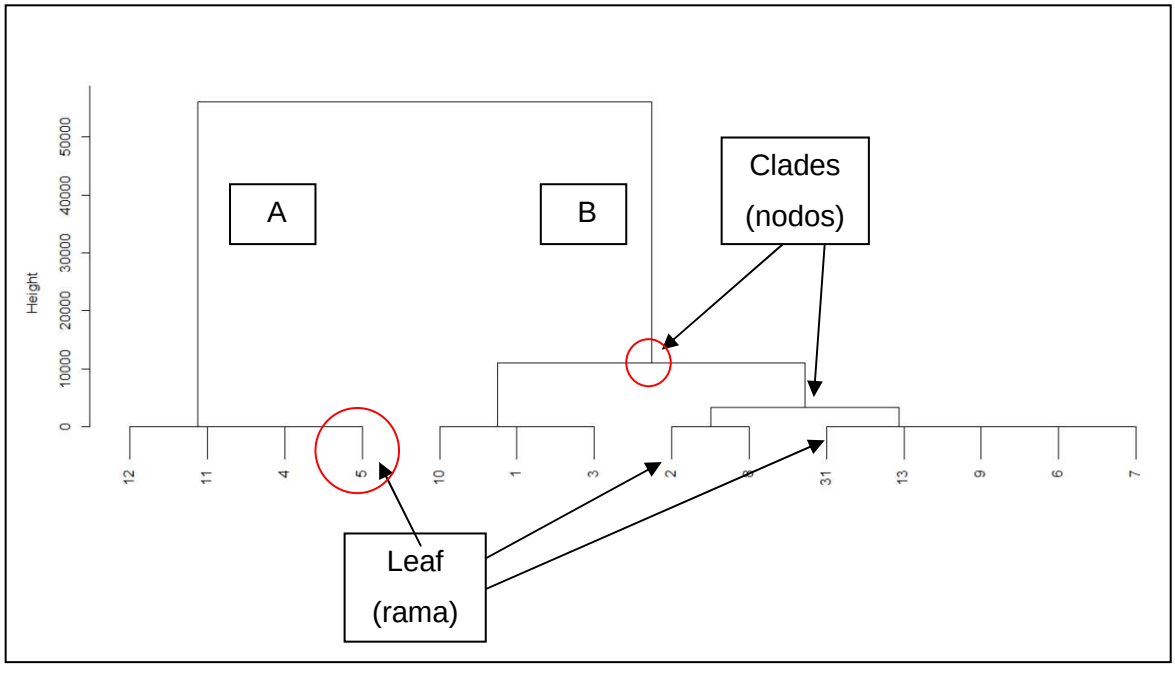



**Figura 8. Demanda de potencia eléctrica en la planta 3, zona 1**

### 4.2.1 Análisis e Interpretación de un dendrograma

Como se puede observar en la Figura 8, un dendrograma está formado básicamente por *Nodos* y *Ramas*, donde los nodos pueden compartir una única rama o varias, no existiendo un límite real en el número de subdivisiones que puedan aparecer.

### 4.2.2 Disposición de los elementos

Un dendrograma se puede interpretar de dos formas diferentes, dependiendo del objetivo a conseguir: si se quiere identificar grupos genéricos a gran escala, o si el objetivo es analizar la similitud entre diferentes secciones individuales.

En el primer caso, se inicia la lectura del dendrograma de forma descendente, de arriba a abajo (top-down), identificando en primer lugar aquellos puntos de ramificación posicionados en las partes superiores de nuestro árbol resultante.

En el segundo caso, al que pertenece nuestro ejemplo de esta sección, realizamos la lectura de abajo a arriba (bottom-up) identificando primero los nodos que nos vamos encontrando, y así, unirlos progresivamente, a medida que se avanza por la estructura en la lectura ascendente.

Volviendo a la figura 8, este tipo de representación nos indica la similitud entre los valores obtenidos por los sensores por día, en base a la disposición que se tiene en el dendrograma.

En líneas generales, la interpretación de un dendrograma se basa en ciertas reglas:

- La disposición de los nodos constituidos indican la similitud entre los elementos.

- La altura que presentan los puntos de unión (ramificación) indica lo semejantes o diferentes son las ramas, de tal forma que a mayor altura, mayor diferencia presentan entre sí.

- La orientación horizontal de los dendrogramas es irrelevante. Se podría representar de forma simétrica respecto al eje vertical, y los resultados o conclusiones no variarían.

- Particularizando para el caso que se está analizando, cada rama obtenida en el dendrograma pertenece a los días en los que se recogieron los datos a analizar.

- Se puede afirmar que el dendrograma es una buena herramienta de apoyo **para los casos de detección de anomalías**, objetivo principal de la metodología desarrollada a lo largo de este documento. Además de las conclusiones que se pueden deducir de su análisis, nos puede dar pistas sobre en qué conjunto de datos deberíamos de centrar nuestro estudio de forma más detallada. Permite contrastar los resultados con otras hipótesis o métodos utilizados de forma paralela.

### 4.2.3 Ejemplo de Análisis

Como ejemplo, si analizamos el dendrograma de la Figura 8 (lectura bottom-up), podemos llegar a varias conclusiones, entre las que comento sólo las más relevantes:

- Se observan dos grandes grupos (clústeres), que podemos asignar, por ejemplo como A y B.

- El grupo A corresponde a los días 4, 5, 11 y 12, mientras que el grupo B lo forma el resto de días, dentro de los 14 días analizados.

- Se puede interpretar, como es el caso, que el grupo A se corresponde con los fines de semana, y el grupo B son los días de diario pertenecientes a las dos semanas analizadas.

- Se puede afirmar que los elementos del clúster A presentan mayor similitud entre sí que con respecto a los elementos que conforman el clúster B.

- El criterio de similitud aplicado en los puntos anteriores se puede aplicar a todos los elementos y nodos que se han formado. Por ejemplo, en el clúster B, se observa que los días 10, 1 y 3 presentan mayor similitud que el resto de días dentro de dicho clúster.

# 5 DTW

## 5.1 Introducción

*Dynamic Time Warping (DTW)* se creó en los años 60, si bien, extendió su uso y se popularizó a partir de los años 70. Es una de las métricas más utilizadas en la actualidad.

Con la tecnología existente en la época de su creación, no había posibilidad de hacer uso de todo su potencial, por lo que no ha sido hasta esta última década, con sistemas más modernos, cuando ha sufrido una verdadera transformación, evolucionando de tal forma que, aunque en los primeros años estaba principalmente enfocado a técnicas de reconocimiento de voz, hoy en día es utilizado en todo tipo de área relacionada con el Data Mining y entornos de Big Data.

En definitiva, DTW se centra en alinear dos secuencias (temporales), con el fin de generar la medida de distancia más representativa de su diferencia total. Este cálculo se ha convertido en el más utilizado en ciertos campos de investigación, como pueden ser la Medicina, la Biología o la Astrofísica, por poner sólo varios ejemplos.

En la última década, se ha dado mucha importancia al tratamiento de datos, con conceptos como clustering o clasificación, dos de las técnicas más utilizadas en el campo del Data Mining.

## 5.2 Ventajas del algoritmo DTW

Hace pocos años, cuando se planteaba un problema de este tipo, estaba generalizado la aplicación de la distancia euclídea como métrica, pero presentaba un claro inconveniente: su sensibilidad a la distorsión del eje del tiempo. Este hecho, era conocido ya por entonces, lo que favoreció que se optara por utilizar una medida de distancia diferente que proporcionara una solución directa a la problemática mencionada.

Las principales ventajas del algoritmo DTW son:

- Mayor flexibilidad. Admite el uso de series en las que el eje distorsionado puede ser diferente al tiempo, como, por ejemplo, la aplicación del reconocimiento de formas, caso en el que la dimensión afectada es el ángulo.

- Comparación de dos series temporales que sean similares pero que se encuentren fuera de fase, permitiendo a su vez, el alineamiento de ambas; se distorsionan de una forma *no lineal* con el objetivo de hacerlas coincidir.

- Tratamiento de deformaciones de tiempo que se ya velocidades diferentes asociadas con los datos dependientes del tiempo.

## 5.3 DTW vs Distancia Euclídea

Para poder comparar series temporales, hay que definir distancia, $D(x_1 , x_2)$, donde $x_1 , x_2$ son vectores, que cumple las siguientes características:

- ✓ No negativo: $D(x_1 , x_2) \geq 0$
- ✓ Simetría: $D(x_1 , x_2) = D(x_2 , x_1)$
- ✓ Desigualdad triangular: $D(x_1 , x_2) \leq D(x_1 , x_3) + D(x_3 , x_2)$
- ✓ Axioma de coincidencia: $D(x_1 , x_2) = 0$, solo si $x_1 = x_2$

Existen diversas métricas (formas de medir la distancia): distancia Euclídea, distancia Manhattan, distancia máxima, distancia de Minkowski y distancia de Mahalanobis.

No es objetivo de este trabajo desarrollar todas las métricas; sólo nos vamos a centrar en la distancia *Euclídea*, que, para 2 clusters $P_1$ y $P_2$, se define con la expresión:

$$d(P_1,P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Es especialmente potente en los casos en los que el dataset de entrada presenta un volumen cada vez mayor, siendo incluso más efectivo que otras técnicas más complejas.

Sin embargo, no es recomendable su uso en ciertos contextos, dado que presenta una serie de limitaciones, entre las que se encuentran:

- ✓ No compara todos los tipos de series temporales, sino solo aquellas que presentan la misma longitud.

- ✓ No es fiable si se tratan datos atípicos (ruido).

En comparación con la distancia Euclídea, la métrica DTW [10] es mucho más robusto a la hora de realizar el cálculo de la distancia (matriz de similitud). Muestra su verdadero potencial al permitir la comparación de series temporales de diferente longitud, ya que, internamente, sustituye la comparación uno-uno -punto a punto- por comparativa uno-muchos (o viceversa).

Como resultado final, dadas 2 series temporales de longitud N y M, respectivamente, el método DTW trata los datos contenidos en una matriz N x M.

[10] Si nos atenemos a un criterio puramente de su definición, *DTW* no es una métrica al uso, dado que no cumple la característica de *Desigualdad triangular*, sino que calcula la forma más óptima de asociar dos series temporales. Con el objetivo de obtener una medida de similitud entre dichas series, se distorsionan en la escala temporal de una forma no lineal.

## 5.4 Ejemplo

Un ejemplo clásico que se suele presentar para llegar a comprender mejor este algoritmo, es el estudio de dos señales codificadas; Cuando ambas señales están perfectamente alineadas (figura 9), no presenta mayor problema, dado que podemos optar por comparar ambas señales realizando un cálculo de la distancia euclídea entre ambas señales como suma de las diferencias de frecuencia en cada punto a lo largo de ellas.

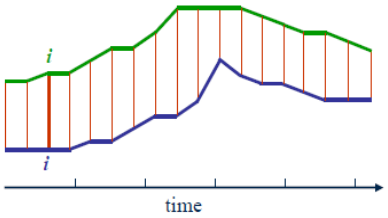



**Figura 9. Señales alineadas**

Pero, en el caso de tener esas mismas señales no alineadas, como en el caso de la figura 10, ¿cómo sabemos qué puntos comparar entre sí?

En el campo de procesamiento de señales, aquí es donde el algoritmo *DTW* presenta su verdadera potencia como algoritmo.

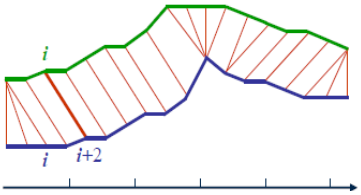



**Figura 10. Señales desacopladas**

Para llevar a cabo dicha comparación, pasa por diferentes fases:

- En una primera fase, se comparan todos los puntos de una señal con cada punto de la segunda señal, y a partir de esta comparativa, genera una matriz.

- En una segunda fase, se lleva a cabo un análisis de dicha matriz, empezando por su esquina inferior izquierda y finalizando en la esquina superior derecha, datos que marcan el principio y el final de las señales analizadas, respectivamente.

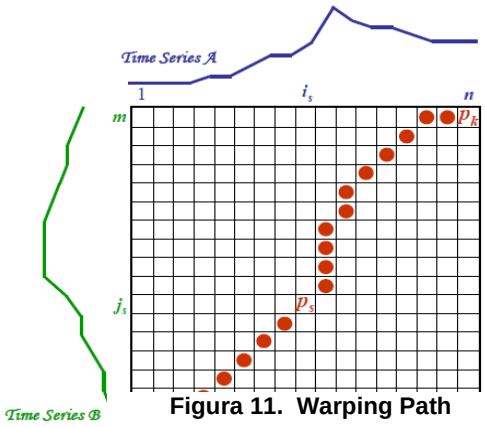



**Figura 11. Warping Path**

- En la siguiente fase, se calcula para cada celda la distancia acumulada: se selecciona la celda vecina en la matriz a la izquierda o debajo con la distancia acumulativa más baja, y se añade este valor a la distancia de la célula focal.

Cuando se completa este proceso, el valor obtenido en la celda superior derecha representa la distancia entre las dos señales, correspondiente a la vía más eficiente en base a la matriz generada, denominado *Warping Path* (Figura 11).

## 5.5 Medida de similitud DTW

Suponemos definidas dos series A y B de longitud I y J respectivamente

$A = a_1 , a_2 , a_3 \ldots a_i \ldots a_I$

$B = b_1 , b_2 , b_3 \ldots bj \ldots b_J$

considerando en la Figura 12 el plano de ejes i , j ciertos alineamientos entre los índices de ambas series, partiendo de trayectorias $F$ del tipo:

$$F = c_1, c_2, \ldots, c_k, \ldots, c_K$$

siendo

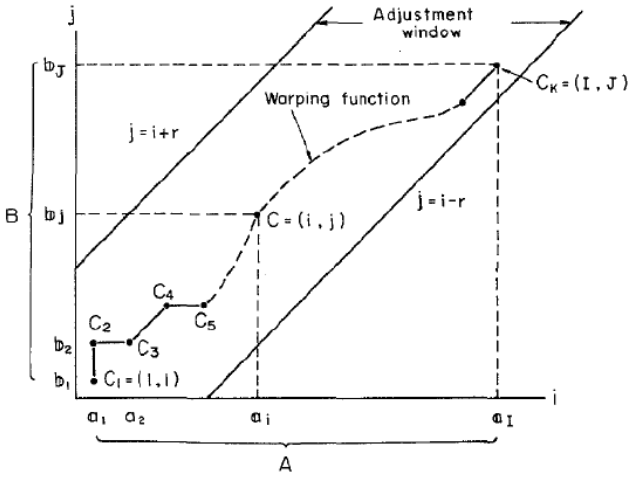



**Figura 12 Warping function**

$$c_k = (i_k, j_k)$$

son puntos de dicho plano, que cumplen la condición: $máx\,(I,J) \leq K \leq I + J$

y que, como veremos, F son trayectorias que deben cumplir ciertas restricciones.

Definimos para cada punto $C_K$, la distancia entre los dos valores de ambas series que fueron alineados como:

$$d_k(c_k) = d_k(i_k, j_k) = \|a_{i_k} - b_{j_k}\|$$

Dada una trayectoria F, se define suma ponderada de estas distancias,

$$S(F) = \sum_{k=1}^{k=K} d(c_k) w_k \; con \; w_k \geq 0$$

Si se eliminan las diferencias debidas fundamentalmente a "deformaciones" en el tiempo, se puede encontrar de manera satisfactoria una trayectoria denominada óptima *Fopt* , cuya suma resultante *S(Fopt)* es la distancia más fiel entre las series A y B anteriormente descrita. Se llega así a la definición de *medida de similitud DTW* entre dos series A y B:

$$DTW(A,B) = \min_F \{S(F)\}$$

## 5.6 Cálculo de Trayectorias

Las posibles trayectorias que se pueden ir calculando en el proceso tienen ciertas condiciones por naturaleza, por lo que una deformación temporal, para posibilitar su alineamiento, siempre va a presentar una serie de restricciones. Se nombran a continuación:

- ***Condiciones de monotonía***: $i_{k-1} \leq i_k \quad y \quad j_{k-1} \leq j_k$

- ***Condiciones de continuidad***: $i_k - i_{k-1} \leq 1 \quad y \quad j_k - j_{k-1} \leq 1$

Las dos condiciones anteriores impactan sobre los valores que pueden ser asignados, De acuerdo a su definición, el punto sólo puede presentar ciertos valores:

$$c_k = \begin{cases} (i_k, j_k + 1) \\ (i_k + 1, j_k + 1) \\ (i_k + 1, j_k) \end{cases}$$

- ***Condiciones de borde***: $i_1 = 1\,, j_1 = 1\,, i_K = I\,, j_K = J$

- ***Condiciones sobre la pendiente***: La pendiente de la trayectoria no debe ser pronunciada (extrema), es decir, durante la comparación de las series, nunca pueden coincidir tramos cortos con tramos largos.

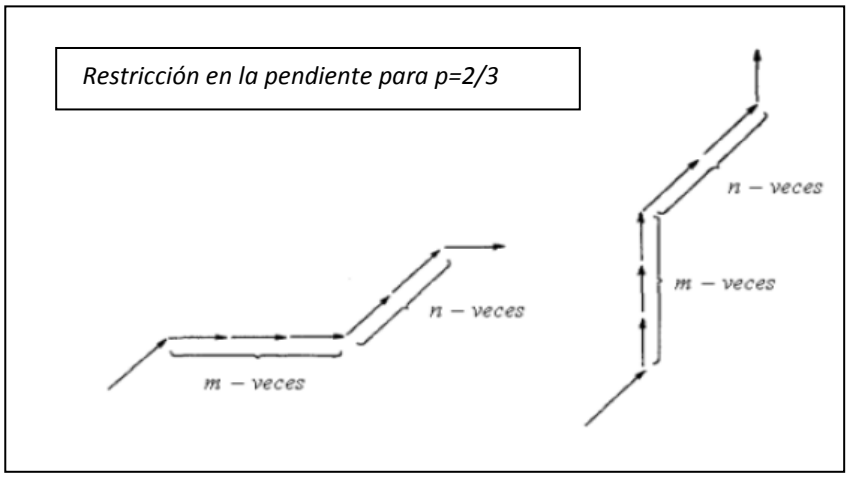



**Figura 13 Restricción en la pendiente**

Como condición (Figura 13) se establece el siguiente criterio: si un punto $C_k$ se desplaza m-veces de forma consecutiva en la dirección de uno de los ejes, implica que no se puede volver a mover en esa dirección, sin antes moverse n-veces a lo largo de la diagonal.

La forma en la que se mide la severidad de la restricción, se basa en el parámetro definido como $\boldsymbol{p=n/m}$ donde, a medida que aumenta el valor de p, mayor es esa restricción.

Los casos más extremos van a ser aquellos en los que se cumpla que p=0, en cuyo caso, se considera que no existe ninguna restricción, o bien, aquel en el que p=∞, que corresponde al hecho de que no se permite ningún tipo de deformación (en realidad, la trayectoria está restringida a moverse por la diagonal).

## 5.7 Complejidad del algoritmo

La complejidad del algoritmo DTW, por cómo se ha construido y definido, es O (I x J), siempre teniendo en cuenta que debe calcularse toda la matriz. En el caso particular de I = J su complejidad pasa a ser O ($I^2$); Esta característica de ser proporcional al cuadrado, resulta ser una desventaja sustancial siempre que se trabaje con series temporales largas. [11]

## 5.8 Algoritmos de aproximación del método DTW

Con el fin de llevar a cabo un intento de mejora del orden de complejidad del algoritmo DTW clásico, existen varias aproximaciones con las que se logra acelerar los cálculos relacionados con la trayectoria DTW.

Estas acciones de mejora pueden tratar de:

- Modificar las restricciones de las trayectorias.
- Modificar la representación de la Serie temporal.
- Simplificar por etapas (técnica FastDTW)

[11] Se estima que una serie temporal se puede considerar larga, si cuenta con un número cercano a mil componentes.

## 5.9 Restricciones en las trayectorias

¿Tiene sentido calcular todos los tipos de trayectoria posibles?

La respuesta, como era lo esperado, es negativa. Tal y como se ha comentado, con estos algoritmos de aproximación, se impide la creación de aquellas trayectorias que implican una deformación temporal que resulte excesiva. Se crea con ello una zona, conocida como *Bandas de Sakoe-Chiba* [12], en la que se puede encontrar el camino más óptimo con una mayor probabilidad.

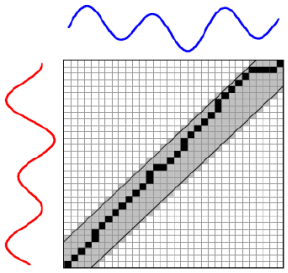


**Figura 14 Bandas de Sakoe-Chiba**

Este tipo de restricciones son del tipo:

$$|i_k - j_k| \leq r$$

Al no tener que calcular la matriz entera, sino sólo los correspondientes a la zona creada, estos algoritmos se mejoran respecto al método original y, como resultado, son bastante más rápidos. Obviamente, las trayectorias elegidas no tienen por qué ser las más óptimas, pero esto no es un problema dichas trayectorias se encuentran cerca de la diagonal.

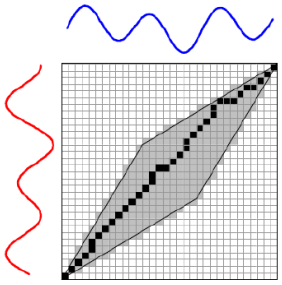


**Figura 15 Paralelogramo de Itakura**

Sólo las distancias DTW encontradas difieren mucho de las reales si las series sufren una deformación significativa.

Otro método que se utiliza con el objetivo de disminuir la complejidad del algoritmo consiste en calcular el camino óptimo DTW sobre un subconjunto reducido de las series.

La forma en la que se puede abordar este proceso es transformar la serie original a intervalos iguales, para acto seguido, definir como representación reducida, la serie que forman los promedios de los valores de la serie original en cada intervalo. [13]

[12] Otra zona utilizada de manera frecuente en los escenarios reales es la de "Paralelogramo de Itakura" (Itakura, 1975).

[13] Otro método utilizado como parte del proceso de mejora del algoritmo DTW, ha sido *FastDTW*. No obstante, el objetivo de este TFM no es analizar detalladamente las distintas tecnologías, sino sólo mencionarlas.

# 6 Caso práctico: medidas de sensores

## 6.1 Introducción

El objetivo de este trabajo, como se ha comentado con anterioridad, es crear una metodología que permita identificar anomalías en los valores de diferentes medidas hechas por los sensores de un edificio de oficinas. Para ello, se ha decidido aplicar la técnica de c*lustering jerárquico* combinado con el análisis visual de los datos originales durante el análisis de los datos.

El caso práctico desarrollado parte de un dataset (datos de entrada)[14] que contiene en formato CSV, diferentes medidas tomadas por sensores en un periodo de tiempo de dos semanas completas (del 31 de mayo al 13 de junio de 2015), a intervalos de tiempo de 5 minutos.

Los parámetros que se miden son de diversa naturaleza:

- **Suministro eléctrico**: se mide la demanda eléctrica total del edificio, la potencia del sistema de aire acondicionado frío/caliente (sistema HVAC), la potencia medida por diversos ventiladores y bobinas del edificio, el sistema de agua caliente, la potencia de diferentes aparatos eléctricos y la potencia de luz de las diferentes zonas.

- **Temperatura**: se mide la temperatura que registra el aire existente junto con la temperatura de retorno del sistema HVAC, la temperatura a la que está el agua en el sistema del agua caliente, la temperatura que tiene el aire que entra en una zona por el sistema de ventilación y la temperatura de cada zona.

- **Sistema HVAC**: se mide el estado on/off del sistema HVAC cuando está programado que esté apagado.

- **Flujos de aire y agua**: se mide el flujo de aire de entrega a las zonas del edificio por parte del sistema HCAV, así como su flujo de retorno al propio sistema y el de salida al exterior, y, por último, el flujo de agua al sistema de agua caliente.

[14] Para acceder a los datos de origen, visite
http://vacommunity.org/2016+VAST+Challenge%3A+MC2

- **Concentración del aire**: se mide la concentración de $CO_2$ en todas las zonas del edificio.
- **Otras medidas**: se miden parámetros como el porcentaje de aire del exterior que entrega el sistema HVAC, la dirección y velocidad del viento en las cercanías del edificio o la programación del punto de enfriamiento/calentamiento de la zona.

El volumen de datos analizados puede considerarse elevado. Para hacernos una idea, diremos que los sensores están distribuidos por diversas zonas a lo largo de las tres plantas de que consta el edificio: en la primera planta, existen hasta 9 zonas definidas en las que se toman diferentes medidas, en la segunda planta hay un total de 16 zonas y, por último, en la tercera planta, existen 13 zonas.

En el dataset de entrada, el volumen de datos analizados se basa en un fichero con formato CSV, que cuenta con un total de 4032 filas y 419 columnas. Las filas corresponden con las medidas tomadas (14 días, las 24 horas del día, cada 5 minutos), mientras que las columnas se corresponden con cada uno de los diferentes sensores distribuidos por las zonas diferenciadas a lo largo de las tres plantas del edificio. Existen, además, otros ficheros CSV con medidas de la concentración de Hazium (sólo tiene dos columnas), así como datos informativos sobre el propio edificio.

## 6.2 Hipótesis de Trabajo

### 6.2.1 Fases del proceso

A continuación, se describen las fases que componen el trabajo llevado a cabo:

En primer lugar, se han tratado los datos de entrada (dataset). Dado el volumen que presentaba el fichero, ha interesado explorar los sensores de forma individual a lo largo de los 14 días de toma de datos, lo que facilitaba el análisis. Estos valores se han agrupado por planta.

Por ejemplo, si queremos analizar la potencia de luz en el edificio, vemos qué columnas del fichero de entrada tienen esos valores y así puedo realizar tareas de Data Mining sobre ellos.

Ha sido de gran utilidad aplicar conocimientos de Shell Scripting en un emulador de entorno UNIX instalado en local en mi PC. El uso de bucles tipo *for* y *while* es una de las formas más eficaces de generalizar los comandos aplicados sobre una columna a todas las columnas que se quieran seleccionar en cada momento.

Todo el trabajo se ha desarrollado en un entorno de programación R, mediante el uso de librerías específicas de Data Mining (reshape, ggplot2 o plotly), así como diversas funciones (cast, melt, hclust, facet_wrap…), cada una con su funcionalidad.

Respecto a la visualización de resultados se ha optado por diferentes representaciones: *gráficos simples*, en los que se representa un parámetro para un día en concreto, o *gráficos compuestos* en un único panel, en los que se representa un mismo parámetro para los 14 días en los que se han recolectado los datos.

Basado en técnicas de Clustering, el análisis se ha basado en la representación en dendrogramas; ambas técnicas han posibilitado la interpretación de los resultados obtenidos (ver capítulo de conclusiones).

### 6.2.2 Análisis de ficheros

Como analistas de datos, se nos ha entregado un conjunto de datos (dataset de entrada) extraído en formato CSV.

Dicho conjunto de datos, no sólo contiene los valores que han medido los sensores ubicados en el edificio de oficinas, sino que sirve como descripción del escenario que se va a analizar; se describe mediante planos las distintas zonas que componen cada planta del edificio, así como la descripción de cada columna, indicando qué parámetro se mide en cada caso.

Con toda esa información, se ha optado por el cruce de ciertas medidas, ya que analizando la relación entre valores a priori inconexos, se puede verificar la interpretación que se haya dado con anterioridad.

Obviamente, por el volumen tratado en los ficheros, dado que el objetivo final es la detección de anomalías, se han estudiado aquellos parámetros que resultaran más convenientes a la hora de obtener conclusiones: la demanda eléctrica total (Total.Electric.Demand.Power), la concentración de $CO_2$ (RETURN.OUTLET.CO2.Concentration), la temperatura (Thermostat.Temp), en relación con la temperatura del sistema HVAC (variables Thermostat.Cooling.Setpoint y Thermostat.Heating.Setpoint) y, por último, la concentración de Hazium (variable Hazium Concentration).

Se ha trabajado bajo la hipótesis de que este tipo de análisis global, va a generar un conocimiento sobre los datos, que es lo que nos va a permitir detectar aquellos comportamientos que resulten atípicos, es decir, la detección de anomalías significativas, que era nuestro objetivo marcado desde el primer momento.

### 6.2.3 Formato de las columnas y mapas del edificio

Los campos tienen un formato concreto de nomenclatura. Es equivalente en todas las columnas del fichero, por lo que para explicarlo, utilizamos como ejemplo el fichero de *Hazium*, una sustancia potencialmente peligrosa en concentraciones elevadas, que exige vigilar su presencia en el edificio.

Los diferentes sensores no están en todas las zonas del edificio, sino que se distribuyen por determinadas áreas, aquellas que se han considerado necesarias para tener un sistema completo de monitorización.

La nomenclatura utilizada nos indica la ubicación del edificio en el que se ha medido esa variable (Hazium en este caso): por ejemplo *F_1_Z_1: Hazium Concentration* significa que el sensor está en la Planta 1, en la zona 1 (F = floor 1, Z = zone) y está midiendo la concentración de Hazium en esa zona determinada.

A partir de la información contenida en los ficheros de mapas, podemos de forma visual, saber dónde está dicho sensor. Para la planta 1 (Figura 16a) sería la zona indicada:

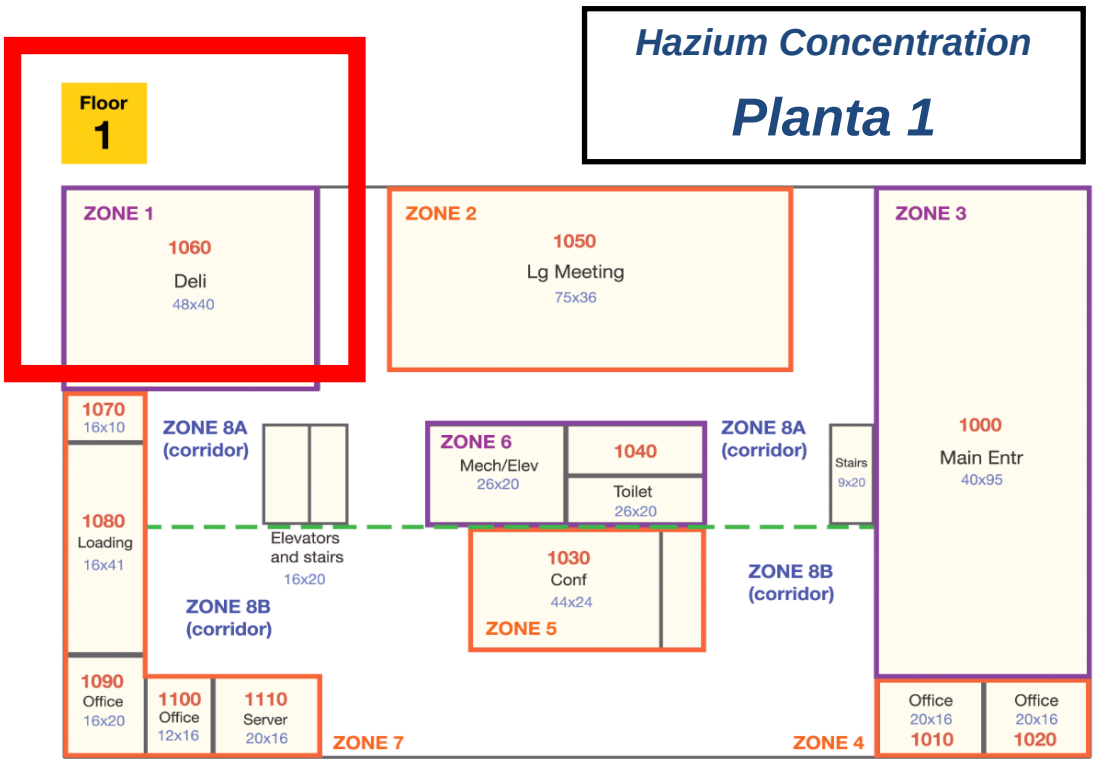



**Figura 16a Mapa de zonas en la planta 1**

La descripción de este parámetro medido por los sensores se describe en la Figura 16b:

| | | |
|---|---|---|
| F_#_Z_#: Hazium Concentration | [ppm] | Concentration of Hazium measured at the zone's return air grille |

**Figura 16b. Formato de parámetros**

## 6.3 Análisis

En el análisis realizado, se han tratado datos recabados por distintos sensores distribuidos en distintas zonas de un edificio de oficinas. Las conclusiones obtenidas a partir de dicho análisis, nos permite definir una metodología concreta que posibilite detectar aquellos datos que pueden ser considerados como valores atípicos.

La metodología definida en este trabajo se basa principalmente en el uso de técnicas de c*lustering jerárquico*, tomando como método de distancia entre elementos la distancia *DTW*.

La visualización de los datos se ha resuelto mediante la generación de dendrogramas, complementado con el estudio puntual mediante gráficos. Ambos modelos permiten que, por parte del usuario, se lleve a cabo un análisis y una verificación del escenario rápida y efectiva.

El análisis detallado de los datos empieza por la visualización directa de los dendrogramas que se han generado. Esta primera fase nos va a permitir detectar de forma visual aquellos grupos de datos (clústeres) que se comportan de una manera atípica, y que por tanto, podrían constituir una posible anomalía.

La segunda fase consiste en verificar dicho comportamiento mediante las correspondientes gráficas obtenidas para esos datos, llevándose a cabo un análisis más específico, de una forma más puntual.

En esta segunda fase, podemos diferenciar dos formas de afrontar el análisis:

- Comparando los datos de diferentes parámetros para una fecha determinada

  Se realiza un análisis de otros parámetros diferentes al analizado de forma aislada y se comparan con el fin de encontrar una posible relación entre ellos.
  De esta forma, podemos suponer que en una fecha determinada, los datos presentan un comportamiento anómalo, tal y como queríamos verificar.

- Comparando los datos del mismo parámetro para los 14 días

  Se analizan los gráficos obtenidos para un parámetro determinado a lo largo de los 14 días que se han analizado en el estudio y se comparan los datos.

El comportamiento más repetido a lo largo de los días se va a definir como normal, mientras que aquellos cambios puntuales que se observen se van a poder considerar datos anómalos.

El objetivo final es encontrar un método con el que confirmar que se cumple nuestra hipótesis inicial de que se trata de una anomalía. Los datos que no son similares a aquellos que se consideran "normales", se presentan como que tienen un comportamiento atípico, y por tanto, una vez se han verificado, pasan a pertenecer al grupo de anomalías.

Como norma general, una metodología es más fiable cuanta mayor capacidad presenta para llegar a conclusiones fiables y verdaderas. Se deben de evitar tanto los *falsos positivos*, en los que un dato se trata como anómalo pero realmente es un comportamiento normal, como los *falsos negativos*, en los que el resultado es tratado como un dato normal, pero en realidad presenta un comportamiento anómalo.

Como es lógico, este último caso es crítico, dado que podría ocurrir algún evento que conlleve un riesgo, incluso para la salud de algún empleado, como se da por ejemplo en el caso de la medida de Hazium, un gas potencialmente peligroso en concentraciones altas, como se constata en los siguientes apartados.

### 6.3.1 Ejemplo: concentración Hazium

En esta sección, dada su criticidad, se explica de forma detallada como se ha llevado a cabo el análisis de los datos para el caso concreto de la concentración de Hazium, análisis a partir del cual, se han obtenido las conclusiones sobre las anomalías encontradas (sección 6.3.2).

Tal y como se ha comentado en apartados anteriores, la primera fase consiste en analizar el dendrograma. En el caso que se está considerando, consiste básicamente en 4 dendrogramas, una por cada zona del edificio que tiene sensores para medir la concentración de este gas. Estos dendrogramas se representan en las figuras 17 a 20:

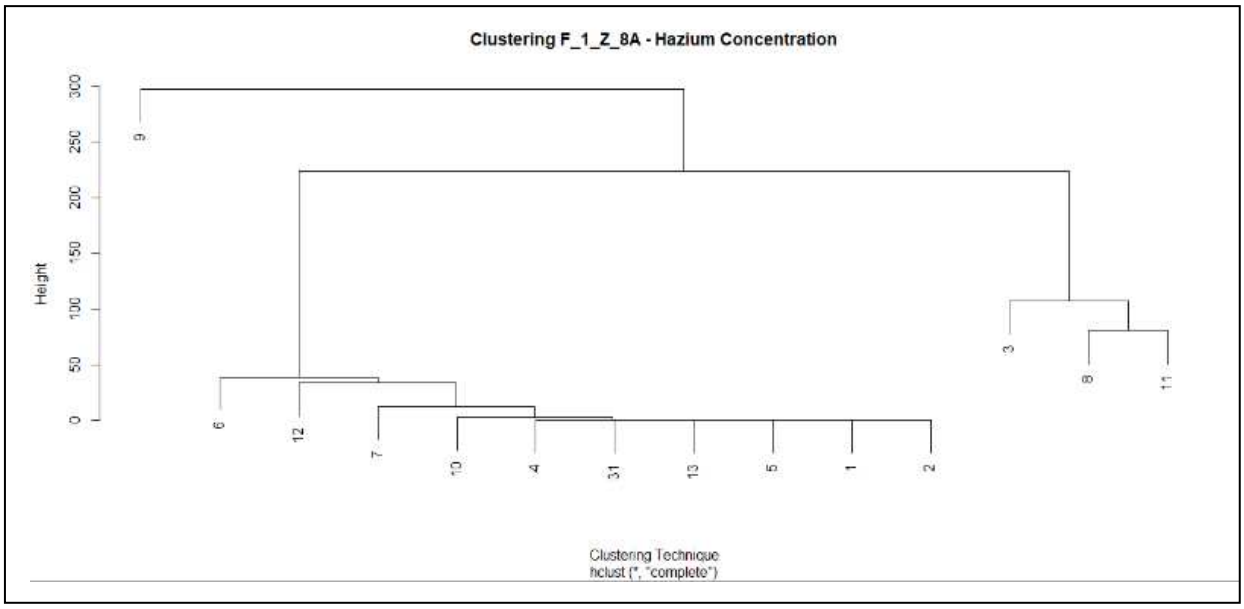



**Figura 17 Dendrograma para la concentración de Hazium en la planta 1, zona 8ª**

En este primer dendrograma (Figura 17) se observan claramente diferenciados 2 clústeres, uno de ellos correspondiente al día 9. Para el resto de días analizados, se diferencian de nuevo 2 subgrupos, entre los que se debe hacer mención a los días 3, 8 y 11. Estos 4 días constituyen en la planta 1 las anomalías potenciales existentes.

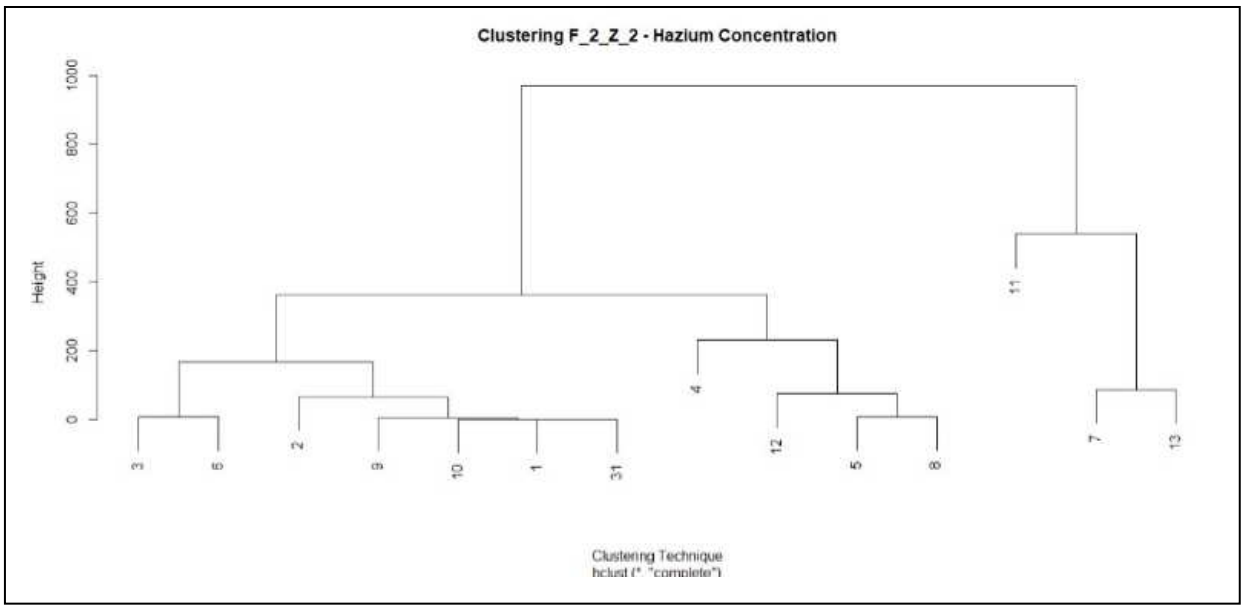



**Figura 18 Dendrograma para la concentración de Hazium en la planta 2, zona 2**

De manera análoga al análisis del dendrograma anterior, podemos llegar a la conclusión de la existencia de 3 grandes grupos en el dendrograma de la Figura 18, si bien, cabe destacar el día 11 como potencial dato anómalo de los sensores ubicados en la planta 2, zona 2.

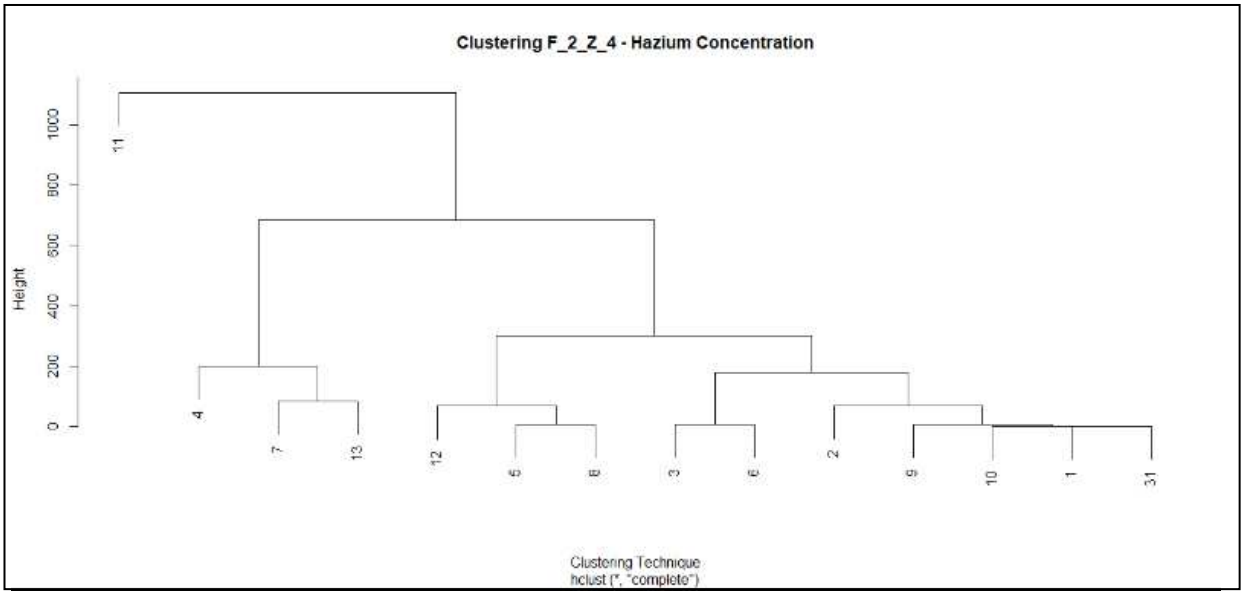



**Figura 19 Dendrograma para la concentración de Hazium en la planta 2, zona 4**

Como se puede ver en la Figura 19, existen 2 grandes clústeres, con 4 subgrupos de elementos en uno de ellos. Destaca el día 11 claramente como potencial anomalía.

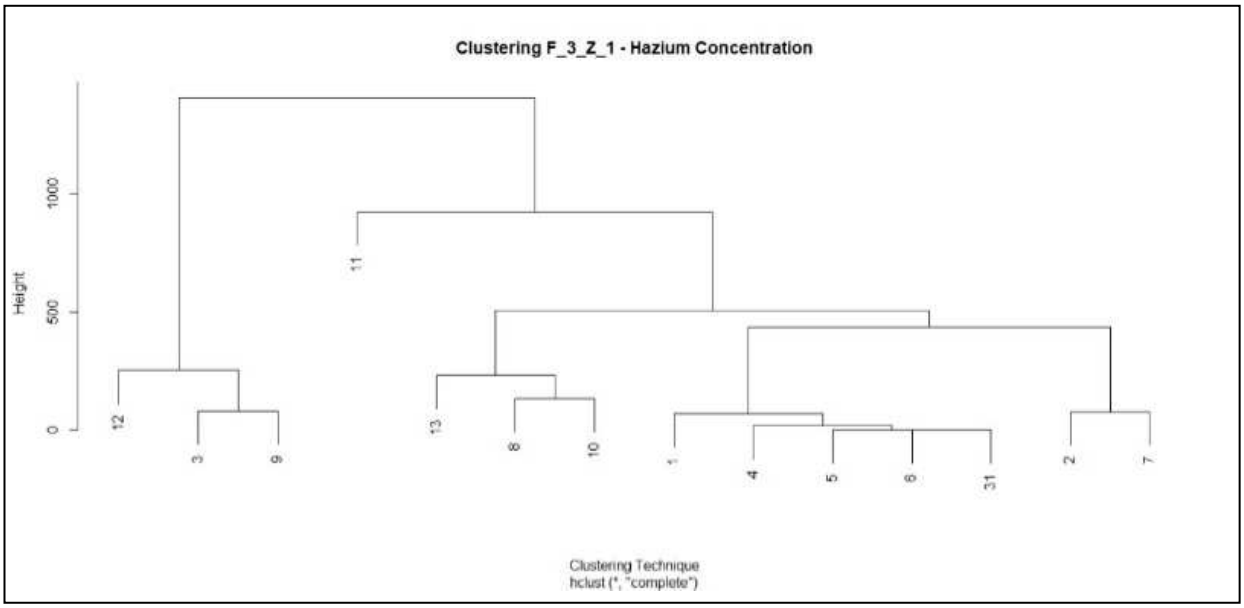



**Figura 20 Dendrograma para la concentración de Hazium en la planta 3, zona 1**

En la Figura 20, para la planta 3, se diferencian 2 grandes clústeres. Cabría destacar de nuevo el día 11, como dato atípico dentro de uno de los subgrupos. El día 12, junto con el día 3 y 9 también deberían de analizarse de forma detallada.

A partir de la interpretación de estos dendrogramas (ver sección 4.2), podemos obtener de forma visual los grupos que son similares (días), generados a partir de la técnica de clustering jerárquico utilizada para nuestro análisis.

Un estudio puntual posterior, en el que se analice de forma detallada un día concreto, junto con otros parámetros que corroboren la teoría, nos permitirá verificar si se trata de una anomalía.

A continuación, se analizan las gráficas (Figura 21 a 24), obtenidas a partir de los datos correspondientes a la concentración de Hazium en cada zona, a lo largo de los 14 días en los que se ha hecho análisis. Esto, permitirá considerar los datos como posibles anomalías, si ese día, se presenta un comportamiento diferente al "normal", el más repetido en la serie temporal.

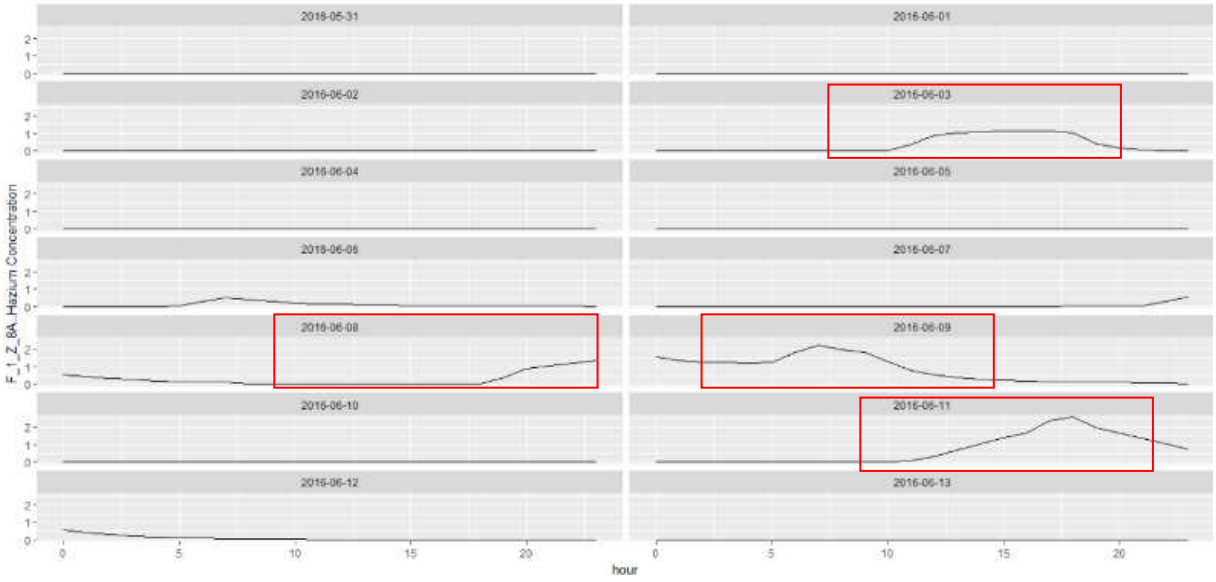



**Figura 21 Panel F1 Z8a**

Si analizamos este panel, se observa claramente como los días que se señalan en las gráficas coinciden con la información obtenida del análisis previo del dendrograma (Figura 17). Se confirma que puede tratarse de una anomalía.

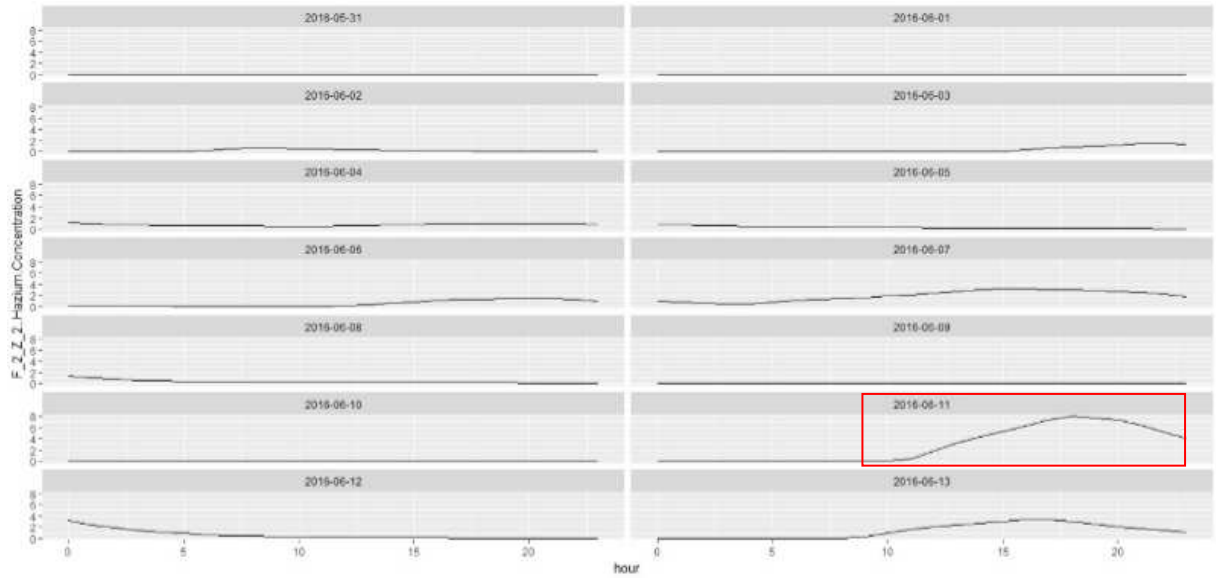



**Figura 22 Panel F2 Z2**

De igual forma, un análisis visual de la Figura 22 nos da información sobre un pico de concentración el día 11, tal y como se dedujo al interpretar el dendrograma de la Figura 18.

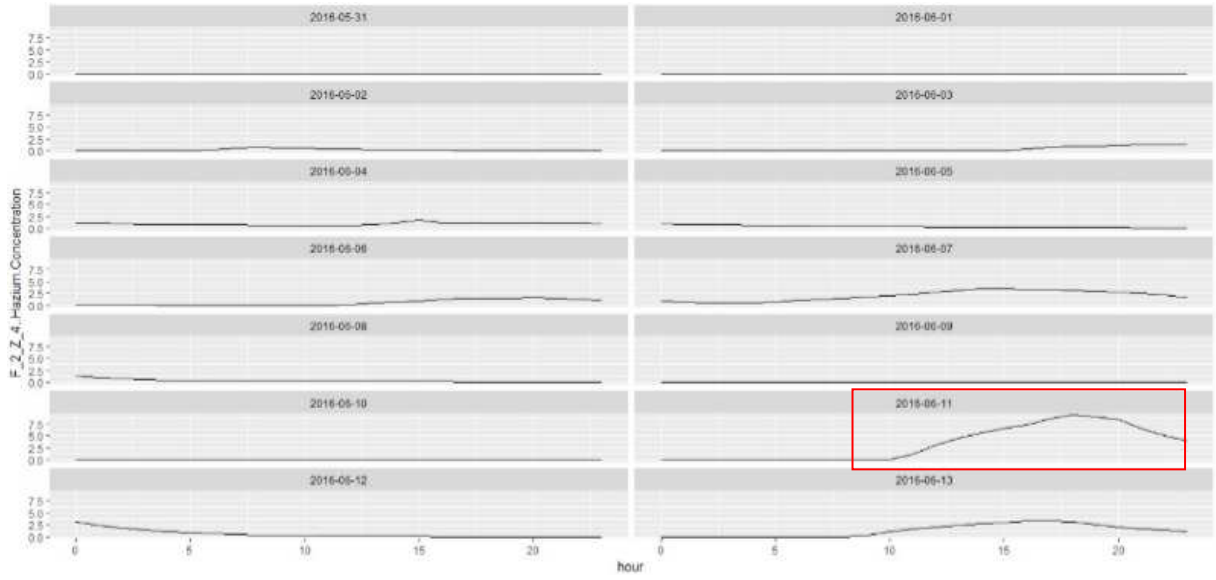



**Figura 23 Panel F2 Z4**

En la planta 2, la zona 4, la situación es similar a la anterior, destacando la concentración medida en el día 11.

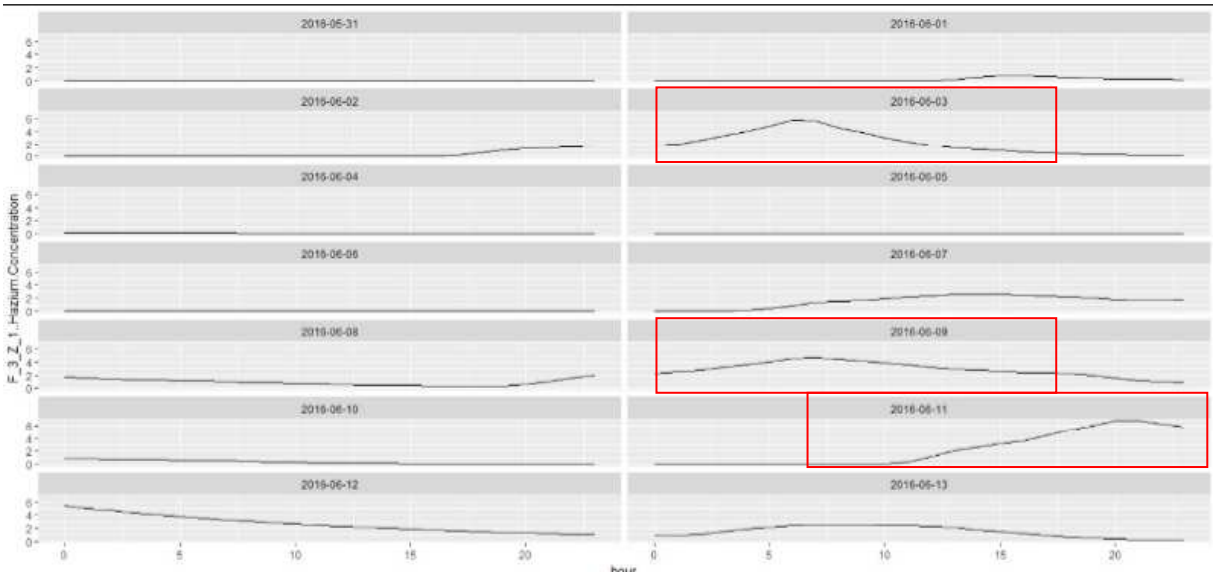



**Figura 24 Panel F3 Z1**

La Figura 24 nos da información coherente con el análisis del dendrograma, como era de esperar: los días 3, 9 y sobretodo, el día 11 se registró una concentración elevada.

Desde un punto de vista práctico, resulta muy útil obtener este tipo de gráfico, dado lo fácil que resulta diferenciar un comportamiento atípico en un día concreto de lo que podemos considerar como comportamiento "normal".

Este estudio, se tendrá que ver avalado con un análisis individual posterior más detallado para ese día "atípico" concreto; por ejemplo, se puede concretar la hora a la que se da la anomalía. Para ello, se tendrá que hacer un estudio de las gráficas individuales (por día), para el día que se está analizando.

Por ejemplo, si detectamos una posible anomalía el día 3 de junio, en la planta 1, en la zona 8a del edificio, que es cuando hemos indicado en las conclusiones (sección 6.3.1) que se detecta el mayor pico de concentración, tendremos que analizar la gráfica de ese día, correspondiente a la Figura 25. Dicha gráfica nos permite ver a qué hora se da el pico:

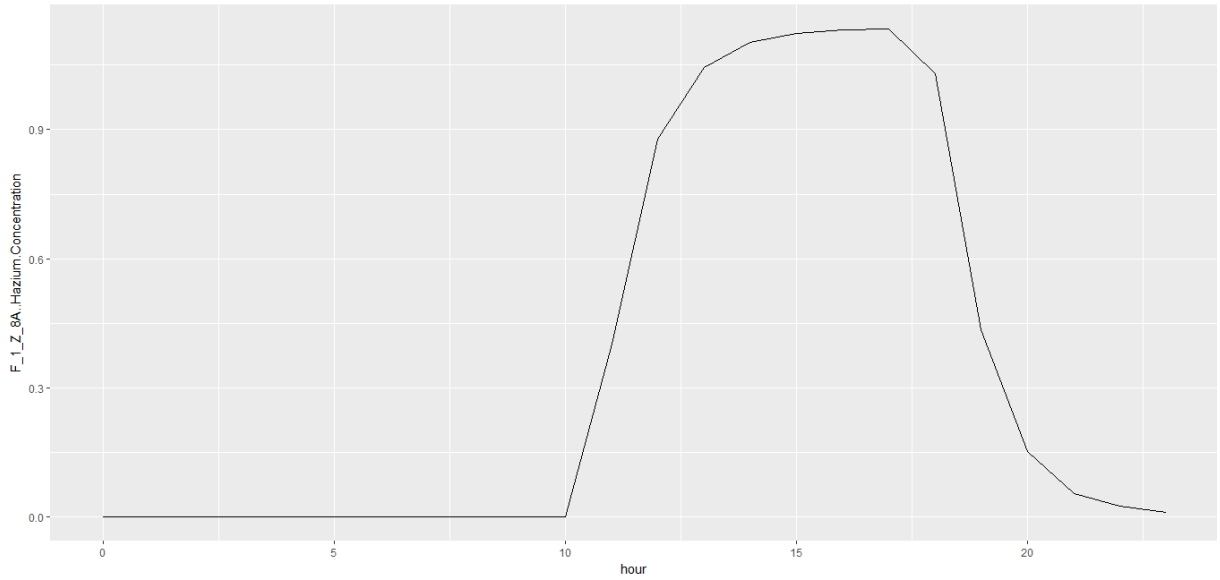



**Figura 25 Gráfica diaria**

El siguiente paso consiste en verificar si el pico registrado es una anomalía o entra dentro de la normalidad. Para ello, tendríamos que analizar si se da a esas horas en otras plantas y zonas del edificio, o por el contrario, si sólo se da en una zona concreta.

Si el parámetro analizado pudiera presentar alguna relación con otros sensores, tendríamos que ver, para ese momento puntual específico, si otro parámetro medido, puede confirmar que es un dato anómalo. En este caso analizado, es un valor independiente, vemos que el pico se da en un despacho concreto (ver sección 6.3.1), por lo que se debe de tomar como anomalía y llegar una solución por el riesgo alto que conlleva este pico de concentración.

### 6.3.2 Hazium

El *Hazium* se describe como un gas peligroso para la salud humana en altas concentraciones. Este hecho, conlleva tener que hacer un análisis lo más exhaustivo posible de las medidas que recaban los sensores del edificio, ya que el riesgo, en caso de no detectarse una anomalía, es muy alto para todo el personal que se encuentre dentro de las zonas afectadas. Puede ser que por su naturaleza, tienda a tener mayor concentración en las plantas altas del edificio.

Los datos que se han analizado durante dos semanas (del 31 de mayo al 13 de junio), nos llevan a concluir que el edificio puede tener ciertos problemas de concentración de este gas:

La concentración de Hazium se analiza en 4 zonas del edificio -una zona en la primera y tercera planta, y dos zonas en la planta segunda. Como en todas las medidas analizadas, se divide el análisis en dos grandes bloques: días laborales y fin de semana.

Durante los días laborales (lunes a viernes) destacan picos de concentración los días 3 y 9 de junio (ver figuras 26 a 29, la sección 1). Parece curioso que este aumento respecto a la media, se da de manera homogénea en todas las plantas del edificio. Además, se detecta un ligero aumento el día 7 de junio, sobre todo en la planta segunda (ver sección 2).

Durante el fin de semana, que corresponde con los días 4, 5, 11 y 12 de Junio (sábados y domingos), llama la atención un pico generalizado el sábado 11 de junio en todas las plantas, como se ilustra en la sección 3.

La máxima concentración durante las dos semanas analizadas se recoge el día 3 de junio en la zona 1 de la planta 3. Este pico sí podría ser peligroso, y más teniendo en cuenta que se da de forma localizada en un despacho, que obtenido a partir de otros datos (ficheros de movimiento), pertenece a un ejecutivo de la empresa.

Se considera que hay que dar prioridad absoluta a estudiar más detalladamente cuál es la causa que lo ha provocado porque podríamos tener dos escenarios posibles: o bien, se trata de un problema de seguridad, o bien, podría tratarse de un problema estructural en el que se hay producido un escape, pero creo que es menos probable porque se habría detectado de forma constante en las zonas colindantes en los días sucesivos, dato que no se observa.

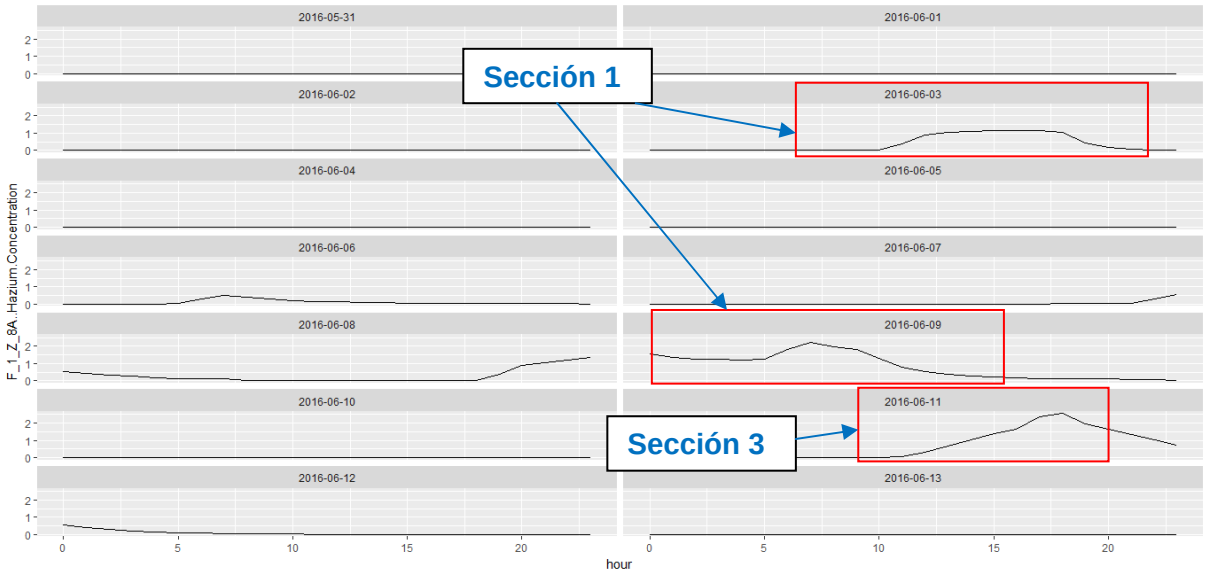



**Figura 26 Concentración Hazium F1_Z8A: análisis visual**

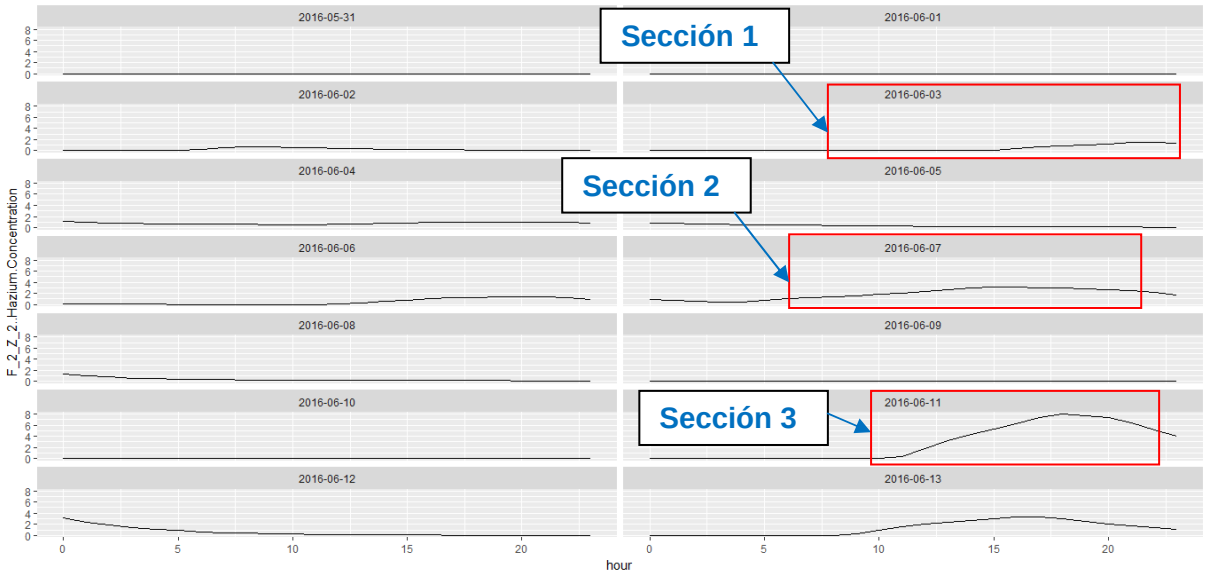



**Figura 27 Concentración Hazium F2_Z2: análisis visual**

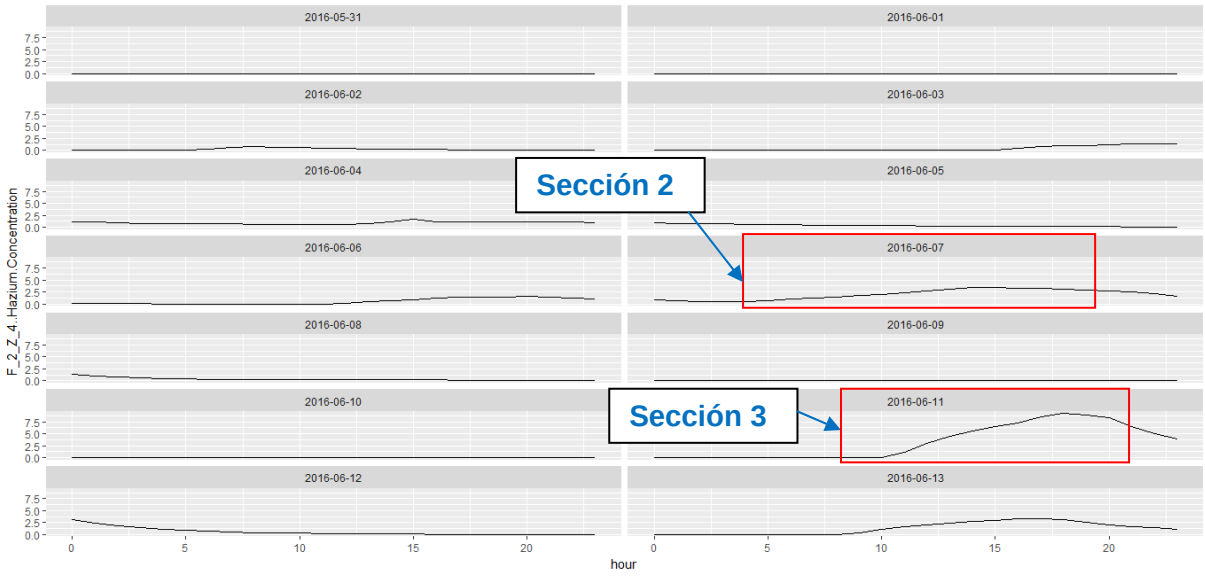



**Figura 28 Concentración Hazium F2_Z4: análisis visual**

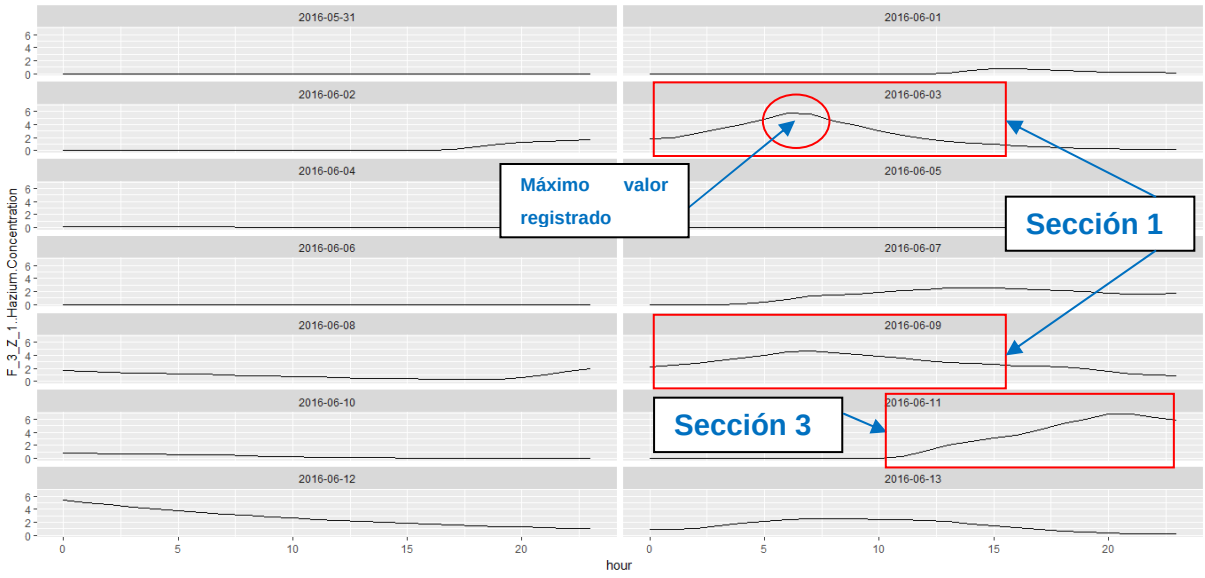



**Figura 29 Concentración Hazium F3_Z1: análisis visual**

### 6.3.3 Temperatura

Respecto a la temperatura, los sensores están distribuidos por numerosas zonas del edificio; en concreto, se toman medidas de 8 zonas en la primera planta, 17 zonas en la segunda planta y 12 en la planta tres.

Analizando detalladamente los datos medidos por los sensores, junto con sus respectivas gráficas, se observa que la temperatura es una variable muy sensible a factores externos, por lo que presenta una variación mayor que otras, hasta tal punto que se da de una forma periódica a lo largo de todos los días, independientemente de la planta o las zonas del edificio que se analicen. Este hecho se ilustra en la Figuras 30.

No obstante, aceptando este hecho, la anomalía que destaca por encima de todas se corresponde con los días 7 y 8 de junio, donde se detecta un aumento considerable de la temperatura de forma generalizada, siempre precedida de un descenso de la misma a primera hora de la mañana (este hecho tiene relación directa con los valores obtenidos para las variables *Thermostat.Heating.Setpoint* y *Thermostat.Cooling.Setpoint*, tal y como se representa en la Figura 31).

El máximo valor se alcanza esos mismos días (por ejemplo, si nos fijamos en la segunda planta, se da el máximo valor a lo largo de las zonas 1 y 9). (ver Figura 31 en color azul)

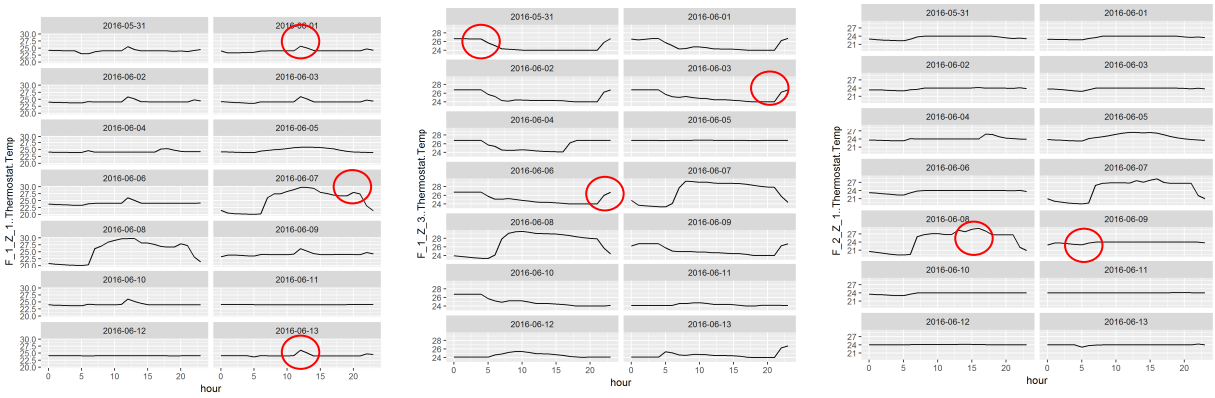



**Figura 30 Temperatura: valores sensibles a factores externos**

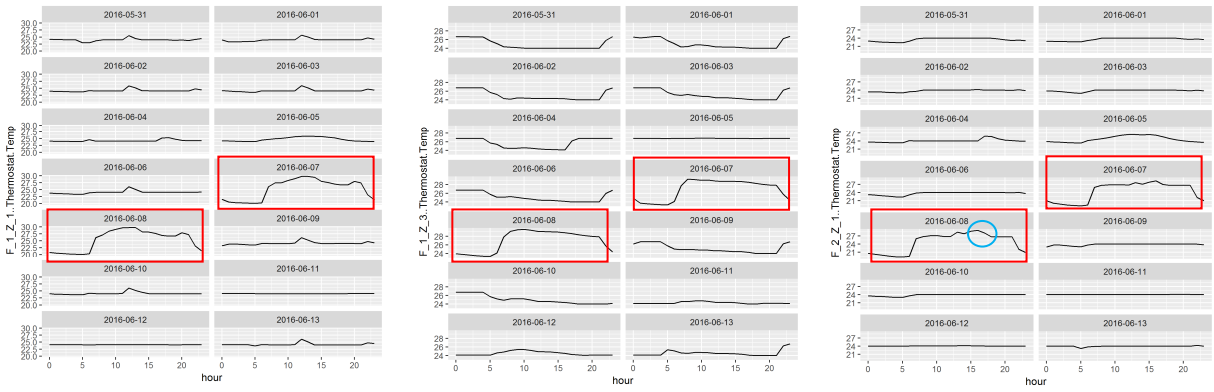



**Figura 31 Temperatura: Aumento en los días 7 y 8 de junio**

### 6.3.4 Concentración $CO_2$

La concentración de $CO_2$ se mide en diferentes sensores distribuidos en las diferentes plantas del edificio, distribuidas en 7 zonas de la primera planta, 18 en la segunda y 12 en la tercera.

En general, podemos decir que la concentración es una variable que fluctúa mucho dependiendo del día que se trate, con más de un pico que se da de una forma puntual a lo largo de diversas zonas, tanto en la planta 1 como en la planta 3 los días 1 y 5 de junio.

En la Figura 32 se representa, a modo de ejemplo donde se observa este comportamiento, los valores medidos en la zona 5 de las plantas 1 y 3.

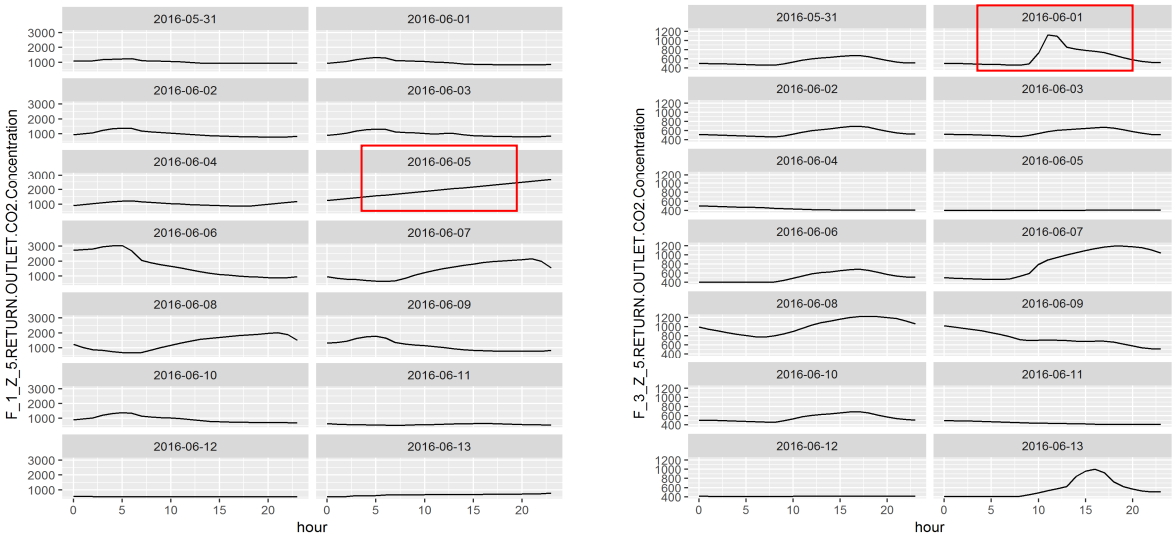



**Figura 32. Concentración $CO_2$ variable. Días 1 y 5 de junio como valores máximos en diferentes zonas**

Ahora bien, si lo analizamos de una forma más global, con los datos y gráficas de todas las zonas, se observa un patrón de comportamiento similar: se detecta una anomalía en los días 7 y 8 de junio, días en los que existe un aumento significativo de la concentración, que siempre se da a última hora de la noche de esos días. Al respecto, analizando las gráficas correspondientes a la variable *VAV.REHEAT.Damper.Position* (ver Figura 34), *que* mide la posición del interruptor de aire (0 ó 1), se observa una relación directa con el hecho cementado: esos dos días, está apagado. Las Figuras 33 muestra dicho comportamiento en dos plantas diferentes:

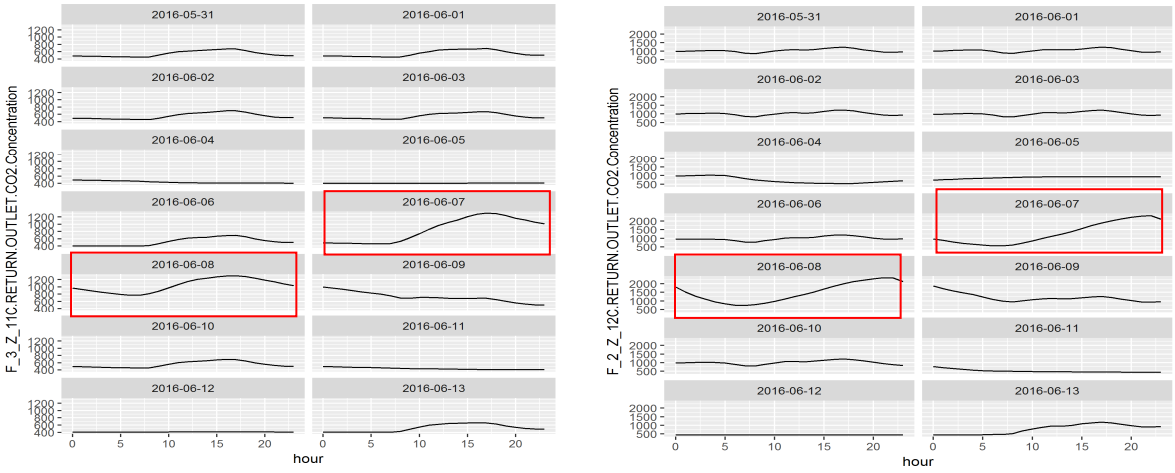



**Figura 33. Concentración $CO_2$ días 7 y 8 de junio en diferentes plantas**

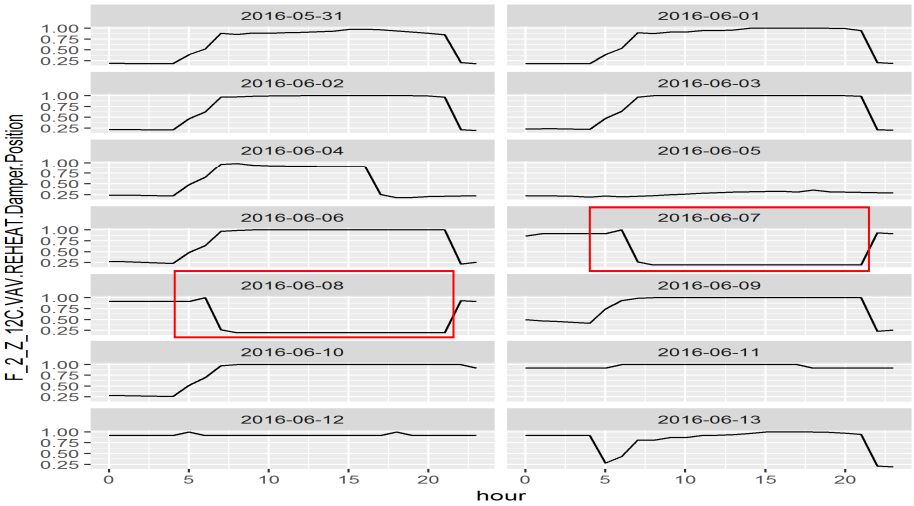



**Figura 34. VAV.REHEAT.Damper.Position - Relación interruptor apagado**

### 6.3.5 Demanda eléctrica total

Esta medida representa la demanda de potencia eléctrica, lo que nos puede dar una idea bastante precisa sobre la actividad en el edificio. Al igual que en casos anteriores, se puede distinguir entre días laborables (lunes a viernes) y fines de semana.

En general, como se muestra en la Figura 35, para todos los días correspondientes a las dos semanas analizadas, se observa un comportamiento bastante similar: mayor demanda eléctrica en las horas de trabajo (aproximadamente de 8 de la mañana a 6 de la tarde), disminuyendo a primera hora del día, así como durante la madrugada, tal y como se representa en la *sección 1*. El fin de semana del 4 y 5 de junio no presenta un valor significativo, lo que se traduce en una baja actividad en el edificio esos dos días, como se representa en la *sección 2*.

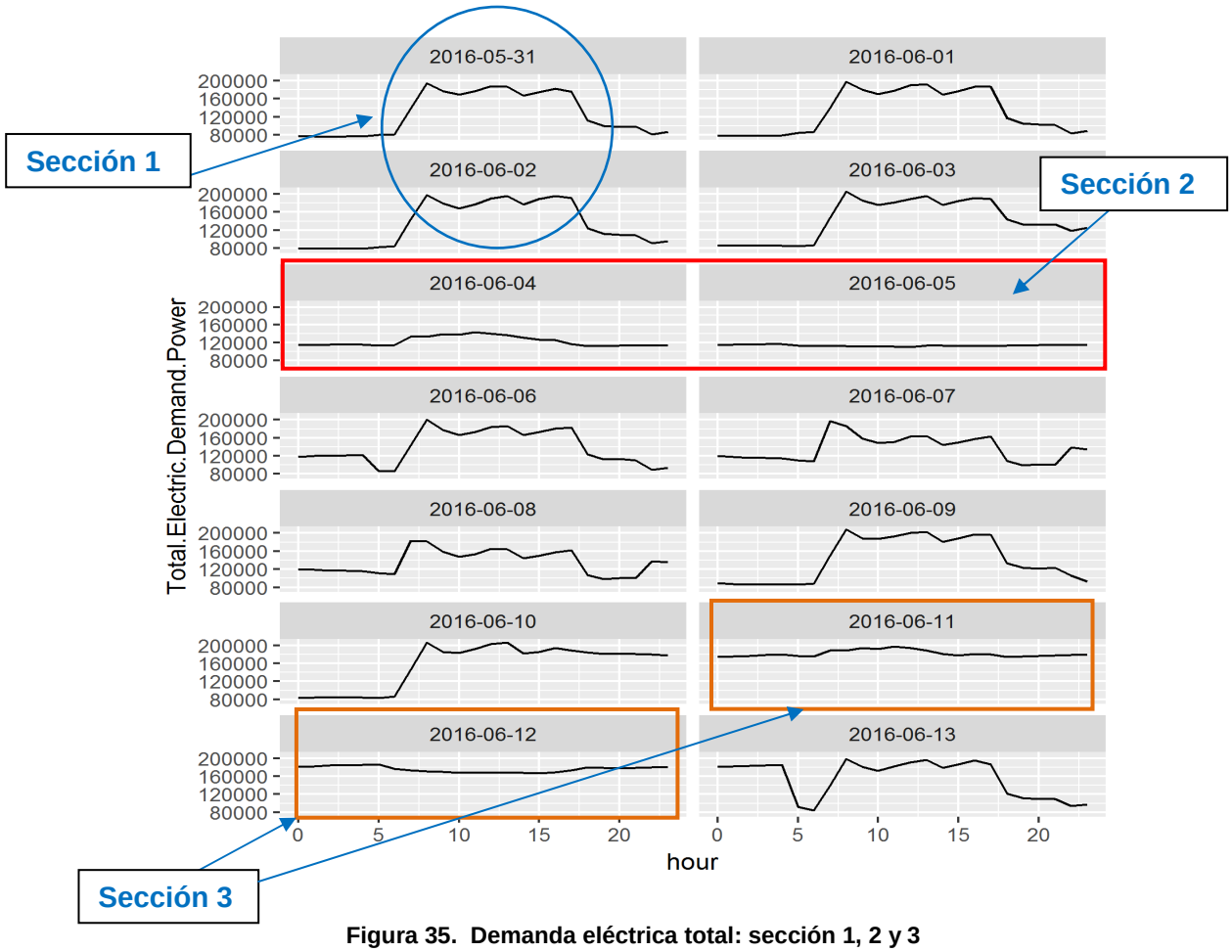



**Figura 35. Demanda eléctrica total: sección 1, 2 y 3**

Dicho esto, donde sí se detecta una clara anomalía en los valores que registran los sensores, es en el fin de semana del 11 y 12 de junio, como se observa en la sección 3; en estos días, hay un aumento significativo de la demanda que se mantiene de forma constante, lo que indica una alta actividad en el edificio, comportamiento que a priori, no era algo esperado.

### 6.3.6 Patrón de comportamiento similar

Si nos fijamos en los valores medidos por los sensores en diferentes medidas, como flujo de aire, concentración de $CO_2$ o temperatura en el edificio, se observa un comportamiento similar en dichas medidas.

Este patrón de comportamiento repetido en las diferentes medidas a lo largo de los días 7 y 8 de junio, se puede interpretar como que se ha dado algún tipo de problema con los sistemas esos días. Este hecho, para la variable de flujo de aire se ilustra en la Figura 36, en la que se representan varias zonas de la planta 1 (se pueden comparar con la figura 35).

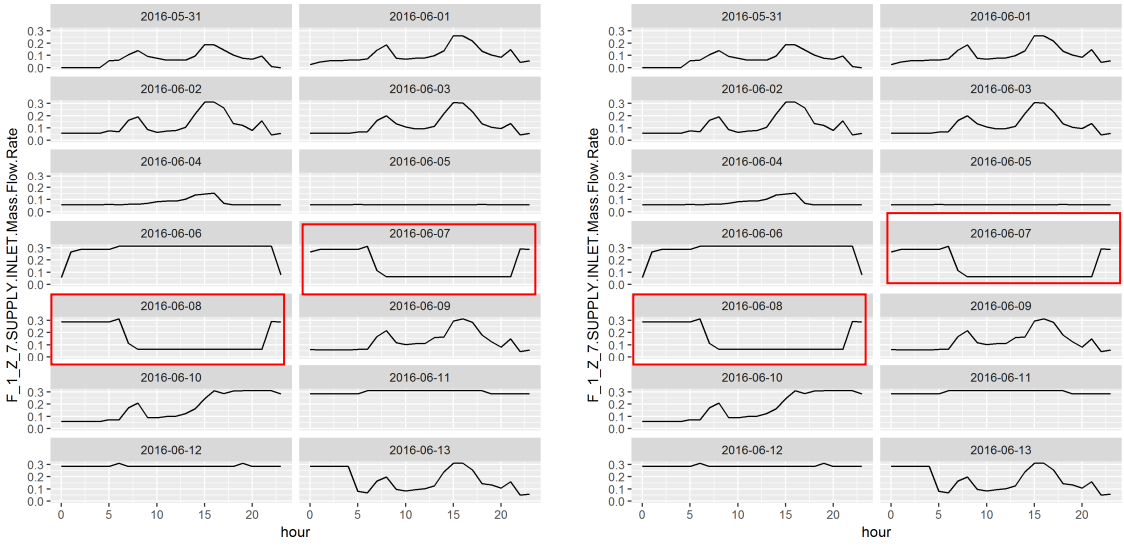



**Figura 36. Comportamiento similar – comparación con la variable flujo de aire**

# 7 Evaluación de la Metodología

La metodología que se ha creado para la detección de anomalías necesita ser validada a partir de la información disponible con el objetivo de poder confirmar la eficiencia de nuestros procesos, así como ratificar las conclusiones extraídas en base al análisis exhaustivo de los datos. En esta sección se presenta el proceso de evaluación.

## 7.1 Análisis de varias variables en un mismo período de tiempo

Si se encuentra una similitud en la tendencia que toman las diferentes variables en un mismo período de tiempo, podemos llegar a evidenciar de algún modo la eficacia del método, dado que podríamos suponer que todas las anomalías recabadas que se han comprobado son coherentes con las conclusiones que se han obtenido.

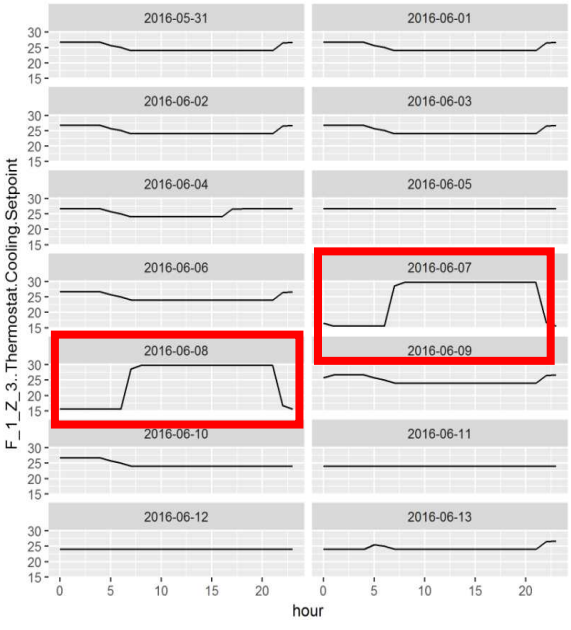



**Figura 37. Comparativa por día: Thermostat Cooling**

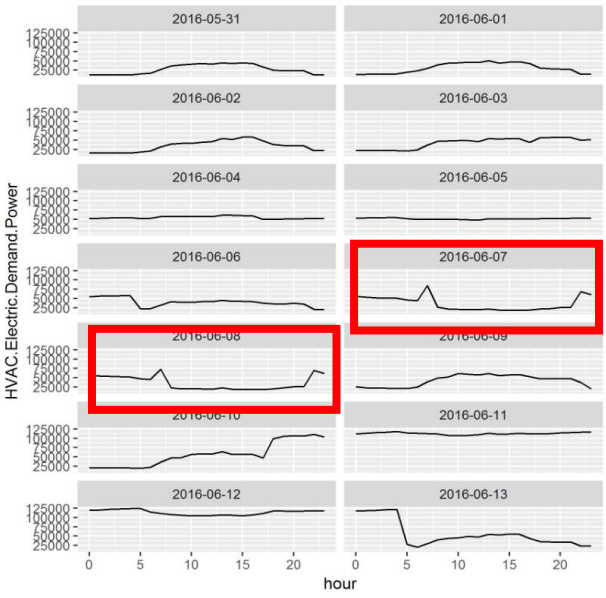



**Figura 38. Comparativa por día: demanda de potencia eléctrica**

Como ejemplo, se ha comparado la temperatura del sistema de frío y la demanda de potencia eléctrica del sistema HVAC, dos variables que, como hipótesis inicial, suponemos que están íntimamente ligadas entre sí. Vamos a comparar los valores (por día) para unos días en concreto, en nuestro caso, los días 7 y 8 de junio, que es donde se detecta una posible anomalía.

El análisis de las gráficas correspondientes debería de corroborar el hecho de que estamos ante unos datos anómalos, como se muestra en las figuras 37 y 38. Se han destacado los días mencionados, con el fin de corroborar su relación: cuando la temperatura del termostato es elevada (unos 30º), la demanda eléctrica disminuye.

Un comportamiento similar existe entre las variables de concentración $CO_2$ y la variable VAV.REHEAT.Damper.Position: si está apagado, la concentración de $CO_2$ aumenta.

## 7.2 Análisis del mismo sensor a lo largo de los días analizados

Con este método, comparamos la misma variable (sensor) y vemos su evolución a lo largo de las dos semanas de recogida de datos.

Como ejemplo, en la Figura 39 se representa la temperatura tomada por uno de los sensores, correspondiente a la variable *F1_Z3_Thermostat.Cooling.Setpoint*, comparando los días 7 y 8 de junio, que es donde se puede predecir un comportamiento anómalo, con el resto de días.

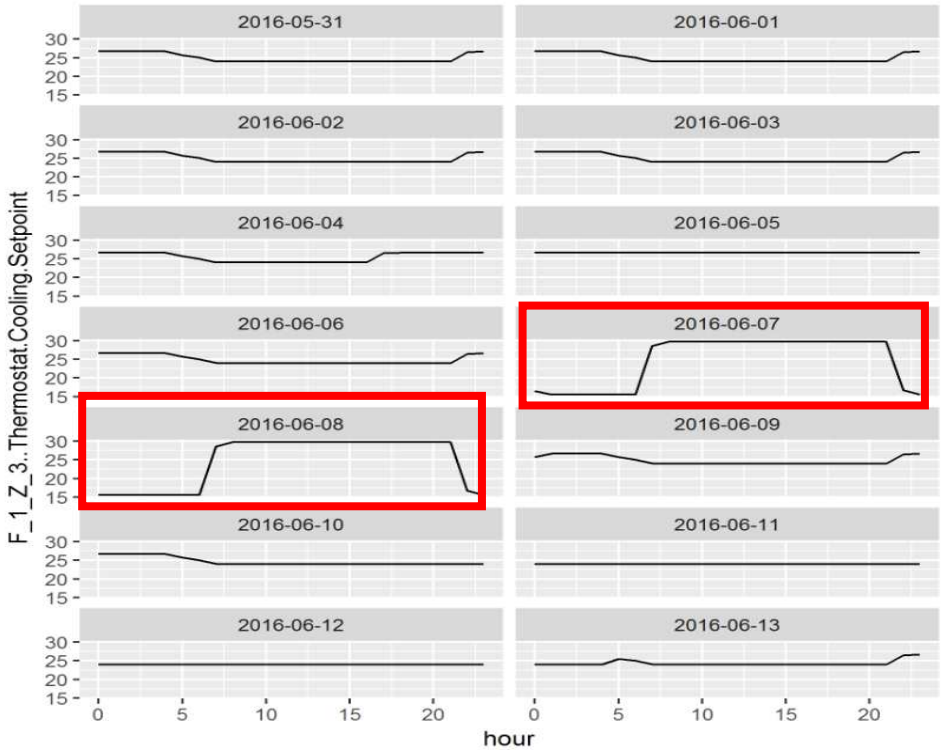



**Figura 39 Comparativa de valores del mismo sensor**

En la Figura 39 se observa un aumento claro de la temperatura los días 7 y 8 de junio.

## 7.3 Análisis de otros conjuntos de datos

En este caso, comparamos los datos analizados en nuestro estudio con datos de otros tipos de sensores de diferente naturaleza, como pueden ser, por ejemplo, sensores de movimiento o de acceso al edificio; estas variables no pertenecen al dataset de nuestro reto, por lo que tendríamos que tener acceso a estos tipos de datos.

La Figura 40 representa los datos obtenidos de los sensores de movimiento fijos en el edificio. Destaca que los fines de semana, a horas centrales de los días, también hay movimiento, lo que se traduce en que hay empleados que han estado trabajando esos días. Se puede observar que el día 4 (sábado) no se han recabado datos.

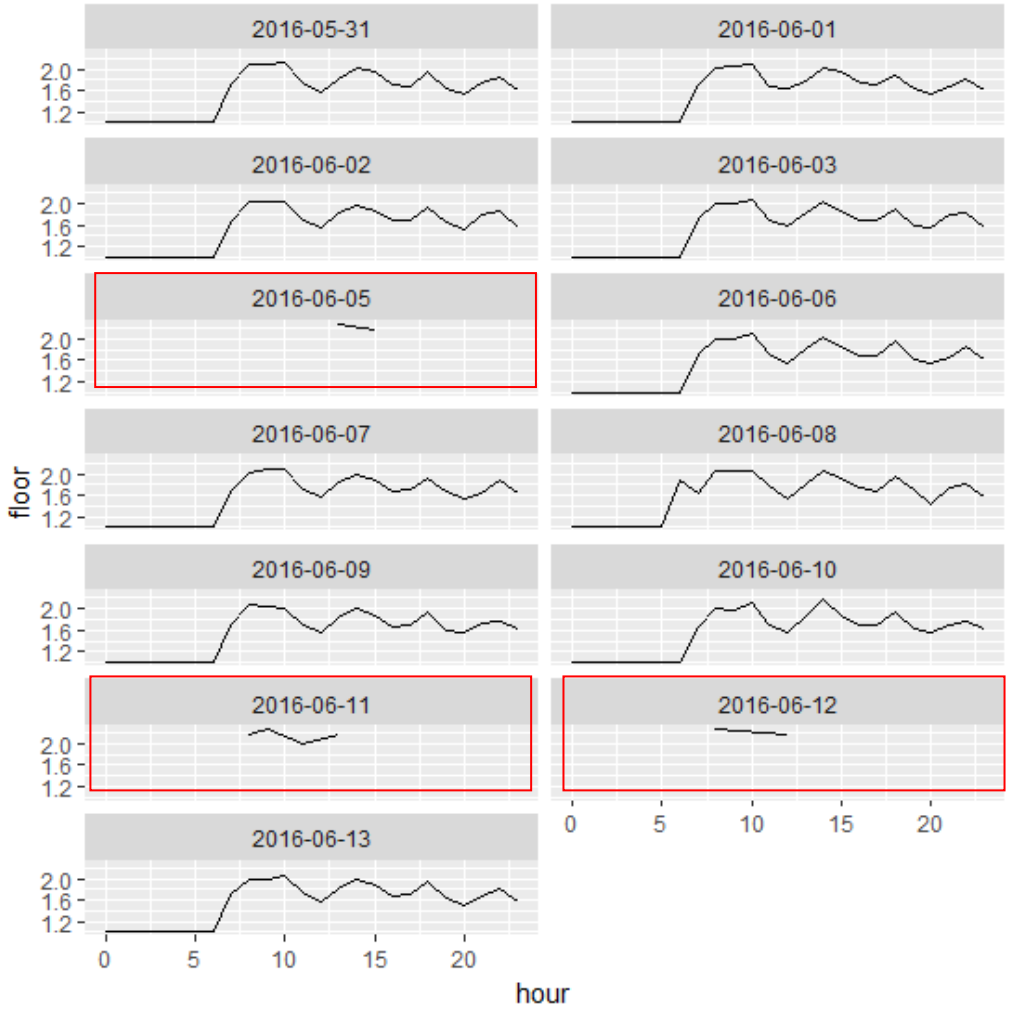



**Figura 40: Sensores de movimiento fijos**

Si relacionamos este movimiento de personal con la demanda eléctrica, se puede ver de forma directa su relación, representada en la Figura 41, en la que se ha optado por destacar el aumento de la demanda de consumo eléctrico, en comparación con un día sin movimiento como es el día 5 de junio:

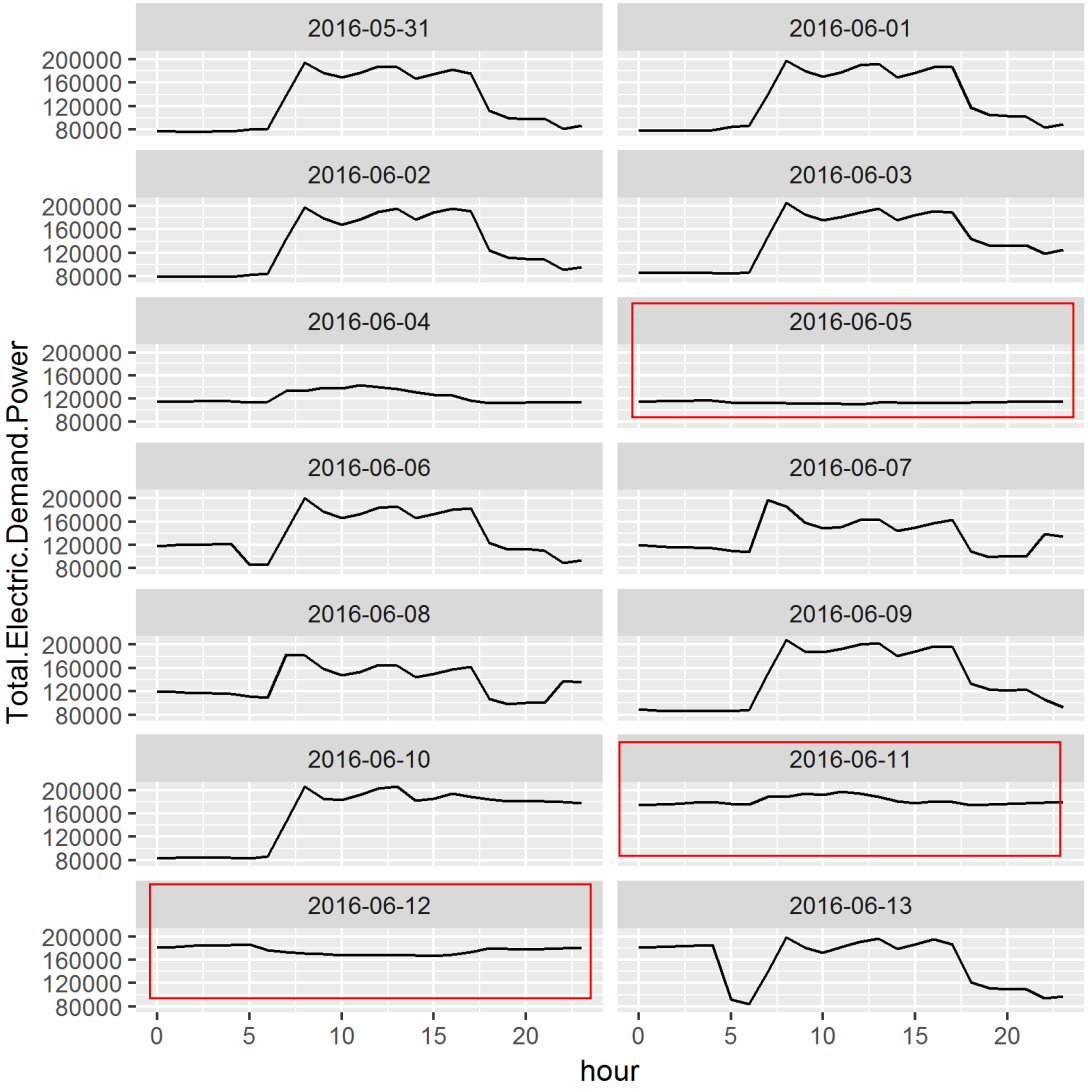



**Figura 41. Demanda de Potencia eléctrica: días 11 y 12 de junio.**

**Comparación con el día 5 de Junio y la Figura 40.**

# 8 Discusión y conclusiones

En esta sección se reflexiona sobre la metodología definida. A modo de conclusión, se presentan las ventajas y desventajas que presentan las técnicas de clustering jerárquico. Por último, se comentan posibles líneas de acción en el futuro:

## 8.1 Discusión

La metodología definida en este trabajo permite la detección de anomalías en series temporales multivariantes. Partiendo de un conjunto de datos de entrada, se analizan mediante técnicas de clustering jerárquico, habiéndose optado por la distancia DTW (Dynamic Time Warping) como métrica de distancia entre elementos.

En líneas generales, se puede afirmar que el algoritmo utilizado en nuestra metodología es un algoritmo simple en comparación con otros métodos de clasificación como pueden ser *Redes Neuronales o Redes Bayesianas*[15]. Además, se trata de un algoritmo bastante flexible: aunque visto como una técnica de agrupación de variables es bastante similar al Análisis factorial presenta una flexibilidad mucho mayor en lo que se refiere a las condiciones que se requieren para su uso[16].

Por otro lado, si se entiende como una técnica de agrupación de casos, resulta bastante semejante al análisis discriminante. A diferencia de éste, a la hora de hacer la "clasificación", no toma como referencia una variable dependiente (grupos de clasificación), sino que se centra en la agrupación de los objetos basándose en la similitud de los casos, por lo que permite detectar un número óptimo de clústeres. No se asume una distribución previa de las variables analizadas.

[15] Página Web
http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf

[16] Las condiciones más relevantes son: no se exige en ningún momento que sea lineal, no se requiere simetría, se permite el uso de variables categóricas y admite varios modelos de estimación en el cálculo de la matriz de distancias

En la técnica de clustering jerárquico destaca la facilidad para el usuario a la hora de interpretar los resultados mediante un primer análisis visual; esto, permite obtener las conclusiones correspondientes de un modo relativamente rápido.

No obstante, debemos ser conscientes de que la técnica de clustering jerárquico presenta también ciertos inconvenientes: sólo permite llevar a cabo un análisis meramente descriptivo, por lo que, en ningún caso se puede considerar realizar con esta técnica un análisis teórico o inferencial. Este es el motivo por el que se usa principalmente como técnica exploratoria, tal y como se ha descrito en los sucesivos capítulos de este documento.

En cuanto a las soluciones que puede ofrecer esta técnica se debe tener muy presente que no son únicas. Estas soluciones, además se apoyan en dendrogramas que pueden llegar a ser muy complejos, lo que puede complicar su visualización en el caso de tratar varias variables.

## 8.2 Conclusiones

La metodología definida ha permitido, como era el objetivo, detectar ciertas anomalías mediante el análisis exhaustivo de un conjunto de datos específicos. No obstante, se ha mostrado como una técnica factible de utilizar en el análisis de otros tipos de datos similares, siempre que se trate de series temporales multivariantes.

Desde un punto de vista más técnico, esta metodología ha permitido profundizar en el conocimiento de los aspectos más teóricos de las técnicas de clustering jerárquico, así como de la métrica DTW. Este concepto ha resultado muy interesante, dada la transformación llevada a cabo como pura técnica de agrupación, que es su principal utilidad, en una técnica de detección de anomalías. Esto, a decir verdad, es algo novedoso y, como ha quedado demostrado, podría valer para acometer otros retos a corto/medio plazo en el mundo del análisis de datos masivos.

## 8.3 Trabajo futuro

La granularidad y el detalle con que se plantea la resolución de un problema es siempre decisión del analista; dicho de otra forma, un mismo problema se puede analizar desde diferentes puntos de vista.

En nuestro caso, se podría haber escogido una línea de trabajo diferente y haber realizado un estudio por hora o por día como unidad mínima existente en el análisis de los datos, como niveles de resolución.

# 9 Anexo

En esta sección se muestran, a modo de ejemplo, algunos comandos ejecutados en R como muestra del proceso llevado a cabo durante el trabajo expuesto, según las necesidades. Además, se anexan ejemplos de tablas en las que se describen los campos que se van a analizar, así como ejemplos de Gráficos y dendrogramas generados en R.

## 9.1 Nombre de Columnas

Comando en R

```
setwd=("<DIR>")
fichero1 <- read.csv(file="<DIR>/bldg-MC2.csv" ,  sep = ",")

names(fichero1)
```

Salida (Ejemplo)

```
 [1] "Date.Time"                                  "Drybulb.Temperature"
 [3] "Water.Heater.Tank.Temperature"              "Water.Heater.Gas.Rate"
 [5] "Supply.Side.Inlet.Mass.Flow.Rate"           "Supply.Side.Inlet.Temperature"
 [7] "Supply.Side.Outlet.Temperature"             "HVAC.Electric.Demand.Power"
 [9] "Total.Electric.Demand.Power"                "Loop.Temp.Schedule"
[11] "Water.Heater.Setpoint"                      "DELI.FAN.Power"
[13] "Pump.Power"                                 "F_1_Z_1..Lights.Power"
[15] "F_1_Z_2..Lights.Power"                      "F_1_Z_3..Lights.Power"
[17] "F_1_Z_4..Lights.Power"                      "F_1_Z_5..Lights.Power"
[19] "F_1_Z_7..Lights.Power"                      "F_1_Z_8A..Lights.Power"
[21] "F_1_Z_8B..Lights.Power"                     "F_1_Z_1..Equipment.Power"
[23] "F_1_Z_2..Equipment.Power"                   "F_1_Z_3..Equipment.Power"
[25] "F_1_Z_4..Equipment.Power"                   "F_1_Z_5..Equipment.Power"
[27] "F_1_Z_7..Equipment.Power"                   "F_1_Z_8A..Equipment.Power"
[29] "F_1_Z_8B..Equipment.Power"                  "F_1_Z_1..Thermostat.Temp"
[31] "F_1_Z_1..Thermostat.Heating.Setpoint"       "F_1_Z_1..Thermostat.Cooling.Setpoint"
[33] "F_1_Z_2..Thermostat.Temp"                   "F_1_Z_2..Thermostat.Heating.Setpoint"
[35] "F_1_Z_2..Thermostat.Cooling.Setpoint"       "F_1_Z_3..Thermostat.Temp"
[37] "F_1_Z_3..Thermostat.Heating.Setpoint"       "F_1_Z_3..Thermostat.Cooling.Setpoint"
[39] "F_1_Z_4..Thermostat.Temp"                   "F_1_Z_4..Thermostat.Heating.Setpoint"
[41] "F_1_Z_4..Thermostat.Cooling.Setpoint"       "F_1_Z_5..Thermostat.Temp"
```

```
[43] "F_1_Z_5..Thermostat.Heating.Setpoint"          "F_1_Z_5..Thermostat.Cooling.Setpoint"
[45] "F_1_Z_7..Thermostat.Temp"                      "F_1_Z_7..Thermostat.Heating.Setpoint"
[47] "F_1_Z_7..Thermostat.Cooling.Setpoint"          "F_1_Z_8A..Thermostat.Temp"
[49] "F_1_Z_8A..Thermostat.Heating.Setpoint"         "F_1_Z_8A..Thermostat.Cooling.Setpoint"
[51] "F_1_Z_8B..Thermostat.Temp"                     "F_1_Z_8B..Thermostat.Heating.Setpoint"
[53] "F_1_Z_8B..Thermostat.Cooling.Setpoint"         "F_1.VAV.Availability.Manager.Night.Cycle.Control.Status"
[55] "F_1_VAV_SYS.SUPPLY.FAN.Fan.Power"              "F_1_BATH_EXHAUST.Fan.Power"
[57] "F_1_Z_1.VAV.REHEAT.Damper.Position"            "F_1_Z_2.VAV.REHEAT.Damper.Position"
[59] "F_1_Z_3.VAV.REHEAT.Damper.Position"            "F_1_Z_4.VAV.REHEAT.Damper.Position"
[61] "F_1_Z_5.VAV.REHEAT.Damper.Position"            "F_1_Z_7.VAV.REHEAT.Damper.Position"
[63] "F_1_Z_8A.VAV.REHEAT.Damper.Position"           "F_1_Z_8B.VAV.REHEAT.Damper.Position"
[65] "F_1_Z_1.REHEAT.COIL.Power"                     "F_1_Z_2.REHEAT.COIL.Power"
[67] "F_1_Z_3.REHEAT.COIL.Power"                     "F_1_Z_4.REHEAT.COIL.Power"
[69] "F_1_Z_5.REHEAT.COIL.Power"                     "F_1_Z_7.REHEAT.COIL.Power"
[71] "F_1_Z_8A.REHEAT.COIL.Power"                    "F_1_Z_8B.REHEAT.COIL.Power"
[73] "F_1_VAV_SYS.HEATING.COIL.Power"                "F_1_VAV_SYS.Outdoor.Air.Flow.Fraction"
```

## 9.2 Descripción de Columnas

Las descripciones de los valores (columnas) encontrados en los ficheros son del tipo:

| Field | Units | Description |
|---|---|---|
| F_#_BATH_EXHAUST:Fan Power | [W] | Power used by the bathroom exhaust fan |
| F_#_VAV_SYS AIR LOOP INLET Mass Flow Rate | [kg/s] | Total flow rate of air returning to the HVAC system from all zones it serves |
| F_#_VAV_SYS AIR LOOP INLET Temperature | [C] | Mixed temperature of air returning to the HVAC system from all zones it serves |
| F_# VAV Availability Manager Night Cycle Control Status | | On/off status of the HVAC system during periods when the system is normally scheduled off. The night cycle manager cycles the HVAC system to maintain night and weekend set point temperatures. |
| F_#_VAV_SYS COOLING COIL Power | [W] | Power used by the HVAC system cooling coil |
| F_#_VAV_SYS HEATING COIL Power | [W] | Power used by the HVAC system heating coil |
| F_#_VAV_SYS SUPPLY FAN OUTLET Mass Flow Rate | [kg/s] | Total flow rate of air delivered by the HVAC system fan to the zones it serves |
| F_#_VAV_SYS SUPPLY FAN OUTLET Temperature | [C] | Temperature of the air exiting the HVAC system fan |
| F_#_VAV_SYS SUPPLY FAN:Fan Power | [W] | Power used by the HVAC system fan |
| F_#_VAV_SYS Outdoor Air Flow Fraction | | Percentage of total air delivered by the HVAC system that is from the outside |
| F_#_VAV_SYS Outdoor Air Mass Flow Rate | [kg/s] | Flow rate of outside air entering the HVAC system |
| COOL Schedule Value | | The supply air temperature set point. Air exiting the HVAC system fan is maintained at this temperature during cooling operation |

| | | |
|---|---|---|
| DELI-FAN Power | [W] | Power used by the deli exhaust fan |
| Drybulb Temperature | [C] | Drybulb temperature of the outside air |
| Wind Direction | [deg] | Direction of wind outside of the building |
| Wind Speed | [m/s] | Speed of wind outside of the building |
| HEAT Schedule Value | | The supply air temperature set point. Air exiting the HVAC system fan is maintained at this temperature during heating operation |
| Pump Power | [W] | Power used by the hot water system pump |
| Water Heater Setpoint | | Water heater set point temperature |
| Water Heater Gas Rate | [W] | Rate at which the water heater burns natural gas |
| Water Heater Tank Temperature | [C] | Temperature of the water inside the hot water heater |
| Loop Temp Schedule | | Temperature set point of the hot water loop. This is the temperature at which hot water is delivered to hot water appliances and fixtures. |
| Supply Side Inlet Mass Flow Rate | [kg/s] | Flow rate of water entering the hot water heater |
| Supply Side Inlet Temperature | [C] | Temperature of the water entering the hot water heater |
| Supply Side Outlet Temperature | [C] | Temperature of the water exiting the hot water heater |
| F_#_Z_# REHEAT COIL Power | [W] | Power used by the zone air supply box reheat coil |
| F_#_Z_# RETURN OUTLET CO2 Concentration | [ppm] | Concentration of C02 measured at the zone's return air grille |
| F_#_Z_# SUPPLY INLET Mass Flow Rate | [kg/s] | Flow rate of the air entering the zone from its air supply box |

| | | |
|---|---|---|
| F_#_Z_# SUPPLY INLET Temperature | [C] | Temperature of the air entering the zone from its air supply box |
| F_#_Z_# VAV REHEAT Damper Position | | Position of the zone's air supply box damper. 1 corresponds to fully open, 0 corresponds to fully closed |
| F_#_Z_#: Equipment Power | [W] | Power used by the electric equipment in the zone |
| F_#_Z_#: Lights Power | [W] | Power used by the lights in the zone |
| F_#_Z_#: Mechanical Ventilation Mass Flow Rate | [kg/s] | Ventilation rate of the zone exhaust fan |
| F_#_Z_#: Thermostat Temp | [C] | Temperature of the air inside the zone |
| F_#_Z_#: Thermostat Cooling Setpoint | [C] | Cooling set point schedule for the zone |
| F_#_Z_#: Thermostat Heating Setpoint | [C] | Heating set point schedule for the zone |
| Total Electric Demand Power | [W] | Total power used by the building |
| HVAC Electric Demand Power | [W] | Total power used by the building's HVAC system including coils, fans and pumps. |

## 9.3 Comandos en R

Shell Script (Generalización a todas las columnas del fichero de entrada)

```
i=1
while [[ $i -le 416 ]]
do

echo "fichero"$i" <- (fichero1[,c(1,"$i")])"
echo "fichero"$i"\$Date.Time = as.POSIXct(fichero"$i"\$Date.Time,
        format=\"%Y-%m-%d %H:%M:%S\",tz=\"CET\")"
echo "fichero"$i"\$hour <- as.numeric(format(fichero"$i"\$Date.Time, format=\"%H\"))"
echo "fichero"$i"\$Date.Time <- format(fichero"$i"\$Date.Time, format=\"%Y-%m-%d %H\")"
echo "fichero <- melt (fichero"$i", c(\"Date.Time\"))"
echo "fichero <-cast (fichero, Date.Time~variable, mean)"

echo "fichero\$Date.Time = as.POSIXct(fichero\$Date.Time, format=\"%Y-%m-%d\",tz=\"CET\")"

echo "sp <- ggplot(data=(fichero), aes_string(x=(colnames(fichero)[3]), y=(colnames(fichero)[2]),
group=\"Date.Time\")) + geom_line()"

echo "sp + facet_wrap( ~ Date.Time, ncol=2)"
echo "p <- ggplotly()"

echo "ggsave(filename=\"myPlot"$i".png\")"

let i=i+1
done > < FICHERO_SALIDA >
```

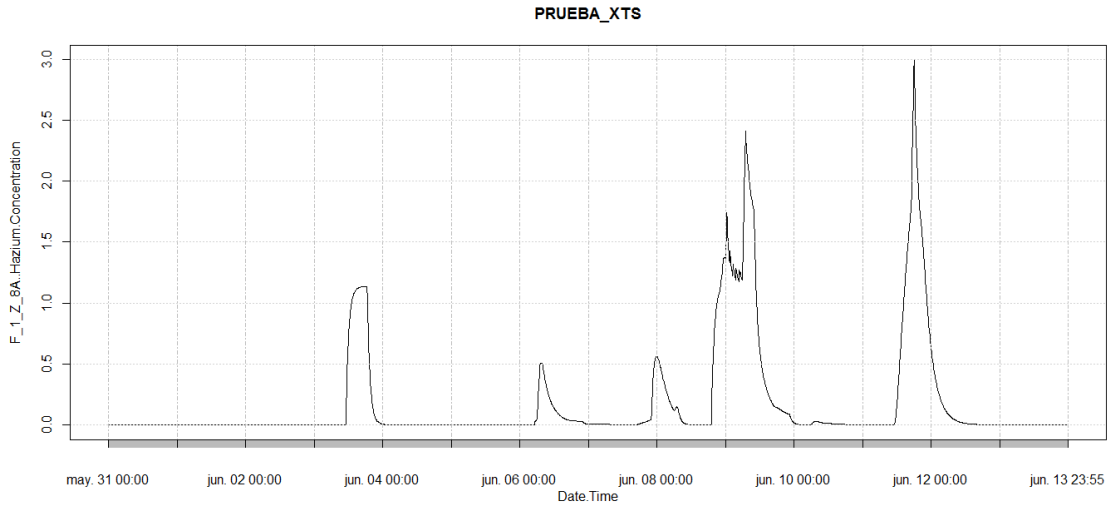
## Ejemplo de Salida (Columna 2)

```
fichero2 <- (fichero1[,c(1,2)])
fichero2$Date.Time = as.POSIXct(fichero2$Date.Time, format="%Y-%m-%d %H:%M:%S",tz="CET")
fichero2$hour <- as.numeric(format(fichero2$Date.Time, format="%H"))
fichero2$Date.Time <- format(fichero2$Date.Time, format="%Y-%m-%d %H")
fichero <- melt (fichero2, c("Date.Time"))
fichero <-cast (fichero, Date.Time~variable, mean)
fichero$Date.Time = as.POSIXct(fichero$Date.Time, format="%Y-%m-%d",tz="CET")
sp <- ggplot(data=(fichero), aes_string(x=(colnames(fichero)[3]), y=(colnames(fichero)[2]),
group="Date.Time")) + geom_line()
sp + facet_wrap( ~ Date.Time, ncol=2)
p <- ggplotly()
ggsave(filename="myPlot2.png")
```

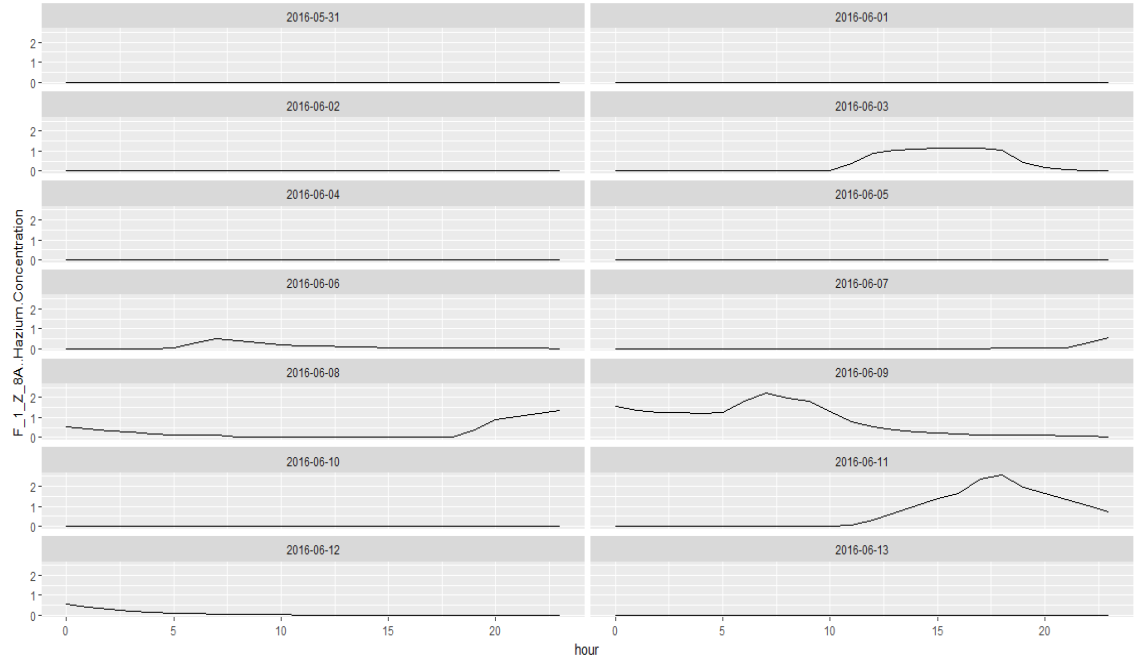
## 9.4 Tipos de Gráficas

En esta sección se presentan los diferentes tipos de gráficos generados en R:

### Gráficas simples (validación puntual)




### Gráficas 2 semanas por zona (comparación)



# 10 Bibliografía

Chang, Winston (2013). *R Graphics Cookbook. O'Reilly*

Crawley, Michael J. (2011). *The R Book*. Wiley.

cucis.ece.northwestern. (s.f.) Obtenido de búsqueda de AnomalyDetection
http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf

*CEUR-WS*. (s.f.) Obtenido de búsqueda de workshop International EIG2009
http://ceur-ws.org/Vol-558/Art_8.pdf

*cucis.ece.northwestern*. (s.f.) Obtenido de búsqueda de Anomaly Detection Survey
http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf
*(Varun Chandola. University of Minnesota,)*

Detección de anomalías. (s.f.) Obtenido de búsqueda de deteccion de anomalias
http://copro.com.ar/Deteccion_de_anomalias.html

*DTW*. (s.f.) Obtenido de búsqueda de DTW
https://www.jstatsoft.org/article/view/v031i07/v31i07.pdf

Kabakoff, Robert (2011). *R in Action*. Manning.

*VAST 2016*. (s.f.). Obtenido de Sitio web que permite descargar los datos de origen:
http://vacommunity.org/2016+VAST+Challenge%3A+MC2

*Wikipedia*. (12 de julio de 2017). *Aprendizaje no supervisado*. Obtenido de https://es.wikipedia.org/wiki/Aprendizaje_no_supervisado

Kamber, Micheline - Pei Jian - Han,Jiawei (2012). *Data Mining: Concepts and Techniques.* MK

*Packages in R:*

*Reshape*. (s.f.) Obtenido de búsqueda de Package Reshape
https://cran.r-project.org/web/packages/reshape/reshape.pdf

*DTW*. (s.f.) Obtenido de búsqueda de Package DTW
https://cran.r-project.org/web/packages/dtw/dtw.pdf

*ggplot2*. (s.f.) Obtenido de búsqueda de Package ggplot2
https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

*dplyr*. (s.f.) Obtenido de búsqueda de Package dplyr
https://cran.r-project.org/web/packages/dplyr/dplyr.pdf